ELSEVIER
NORTH
HOLLAND

VOLUME 3

HANDBOOK OF ALGEBRA

M. HAZEWINKEL
EDITOR

HANDBOOK OF ALGEBRA
VOLUME 3

ELSEVIER
AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK •
OXFORD • PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

HANDBOOK OF ALGEBRA
Volume 3

edited by
M. HAZEWINKEL
CWI, Amsterdam

2003

ELSEVIER
AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK •
OXFORD • PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

ELSEVIER SCIENCE B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2003

Library of Congress Cataloging-in-Publication Data

A catalog record from the Library of Congress has been applied for.

Bitish Library Cataloguing-in-Publication Data

A catalog record from the British Library has been applied for.

ISBN: 0 444 51264 0
ISSN: 1570 7954

⊗ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Printed in The Netherlands.

Preface

Basic philosophy

Algebra, as we know it today, consists of a great many ideas, concepts and results. A reasonable estimate of the number of these different "items" would be somewhere between 50 000 and 200 000. Many of these have been named and many more could (and perhaps should) have a "name", or other convenient designation. Even the nonspecialist is quite likely to encounter most of these, either somewhere in the literature distinguised as a definition or a theorem or to hear about them and feel the need for more information. If this happens, one should be able to find at least something in this Handbook; and hopefully enough to judge whether it is worthwhile to pursue the quest at least. In addition to the primary information references to relevant articles, books or lecture notes should help the reader to complete his understanding.

To make this possible we have provided an index which is more extensive than usual, and not limited to definitions, theorems and the like.

For the purposes of this Handbook, algebra has been defined more or less arbitrarily as the union of the following areas of the Mathematics Subject Classification Scheme:

- 20 (Group theory)
- 19 (K-theory; this will be treated at an intermediate level)
- 18 (Category theory and homological algebra; including some of the uses of category in computer science, often classified somewhere in section 68)
- 17 (Nonassociative rings and algebras; especially Lie algebras)
- 16 (Associative rings and algebras)
- 15 (Linear and multilinear algebra, Matrix theory)
- 13 (Commutative rings and algebras; here there is a fine line to tread between commutative algebras and algebraic geometry; algebraic geometry is definitely not a topic that will be dealt with in this Handbook; there will, hopefully, one day be a separate Handbook on that topic)
- 12 (Field theory and polynomials)
- 11 The part of that also used to be classified under 12 (Algebraic number theory)
- 08 (General algebraic systems)
- 06 (Certain parts; but not topics specific to Boolean algebras as there is a separate three-volume Handbook of Boolean Algebras)

Planning

Originally (1992), we expected to cover the whole field in a systematic way. Volume 1 would be devoted to what is now called Section 1 (see below), Volume 2 to Section 2, and so on. A detailed and comprehensive plan was made in terms of topics which needed to be covered and authors to be invited. That turned out to be an inefficient approach. Different authors have different priorities and to wait for the last contribution to a volume, as planned originally, would have resulted in long delays. Therefore, we have opted for a dynamically evolving plan. This also permits to take new developments into account.

This means that articles are published as they arrive and that the reader will find in this third volume articles from five different sections. The advantages of this scheme are two-fold: accepted articles will be published quickly and the outline of the series can be allowed to evolve as the various volumes are published. Suggestions from readers both as to topics to be covered and authors to be invited are most welcome and will taken into serious consideration.

The list of the sections now looks as follows:

Section 1: Linear algebra. Fields. Algebraic number theory
Section 2: Category theory. Homological and homotopical algebra. Methods from logic (algebraic model theory)
Section 3: Commutative and associative rings and algebras
Section 4: Other algebraic structures. Nonassociative rings and algebras. Commutative and associative rings and algebras with extra structure
Section 5: Groups and semigroups
Section 6: Representations and invariant theory
Section 7: Machine computation. Algorithms. Tables
Section 8: Applied algebra
Section 9: History of algebra

For a more detailed plan (2002 version), the reader is referred to the Outline of the Series following this preface.

The individual chapters

It is not the intention that the handbook as a whole can also be a substitute undergraduate or even graduate, textbook. The treatment of the various topics will be much too dense and professional for that. Basically, the level is graduate and up, and such material as can be found in P.M. Cohn's three volume textbook "Algebra" (Wiley) will, as a rule, be assumed. An important function of the articles in this Handbook is to provide professional mathematicians working in a different area with sufficient information on the topic in question if and when it is needed.

Each chapter combines some of the features of both a graduate-level textbook and a research-level survey. Not all of the ingredients mentioned below will be appropriate in each case, but authors have been asked to include the following:

– Introduction (including motivation and historical remarks)
– Outline of the chapter
– Basic concepts, definitions, and results (proofs or ideas/sketches of the proofs are given when space permits)
– Comments on the relevance of the results, relations to other results, and applications
– Review of the relevant literature; possibly supplemented with the opinion of the author on recent developments and future directions
– Extensive bibliography (several hundred items will not be exceptional)

The future

Of course, ideally, a comprehensive series of books like this should be interactive and have a hypertext structure to make finding material and navigation through it immediate and intuitive. It should also incorporate the various algorithms in implemented form as well as permit a certain amount of dialogue with the reader. Plans for such an interactive, hypertext, CD-Rom-based version certainly exist but the realization is still a nontrivial number of years in the future.

Kvoseliai, July 2003 Michiel Hazewinkel

Kaum nennt man die Dinge beim richtigen Namen,
so verlieren sie ihren gefährlichen Zauber

(You have but to know an object by its proper name
for it to lose its dangerous magic)

E. Canetti

This Page Intentionally Left Blank

Outline of the Series

(as of June 2002)

Philosophy and principles of the Handbook of Algebra

Compared to the outline in Volume 1 this version differs in several aspects.

First, there is a major shift in emphasis away from completeness as far as more elementary material is concerned and towards more emphasis on recent developments and active areas.

Second, the plan is now more dynamic in that there is no longer a fixed list of topics to be covered, determined long in advance. Instead there is a more flexible nonrigid list that can and does change in response to new developments and availability of authors.

The new policy is to work with a dynamic list of topics that should be covered, to arrange these in sections and larger groups according to the major divisions into which algebra falls, and to publish collections of contributions as they become available from the invited authors.

The coding by style below is as follows.

- **Author(s) in bold,** followed by chapter title: articles (chapters) that have been received and are published or ready for publication.
- Chapter title in *italic*: chapters that are being written.
- Chapter title in plain text: topics that should be covered but for which no author has yet been definitely contracted.
- Chapters that are included in Volumes 1–3 have a (x; yy pp.) after them, where ‘x’ is the volume number and ‘yy’ is the number of pages.

Compared to the plan that appeared in Volume 1 the section on “Representation and invariant theory” has been thoroughly revised. The changes of this current version compared to the one in Volume 2 (2000) are relatively minor: mostly the addition of some 5 topics.

Section 1. Linear algebra. Fields. Algebraic number theory

A. Linear Algebra

J.A. Hermida-Alonso, Linear algebra over commutative rings (3; 59 pp.)
Linear inequalities (also involving matrices)
Orderings (partial and total) on vectors and matrices
Positive matrices
Special kinds of matrices such as Toeplitz and Hankel
Integral matrices. Matrices over other rings and fields
Quasideterminants, and determinants over noncommutative fields

B. Linear (In)dependence

J.P.S. Kung, Matroids (1; 28 pp.)

C. Algebras Arising from Vector Spaces

Clifford algebras, related algebras, and applications

D. Fields, Galois Theory, and Algebraic Number Theory

(There is also an article on ordered fields in Section 4)
J.K. Deveney and J.N. Mordeson, Higher derivation Galois theory of inseparable field extensions (1; 34 pp.)
I. Fesenko, Complete discrete valuation fields. Abelian local class field theories (1; 48 pp.)
M. Jarden, Infinite Galois theory (1; 52 pp.)
R. Lidl and H. Niederreiter, Finite fields and their applications (1; 44 pp.)
W. Narkiewicz, Global class field theory (1; 30 pp.)
H. van Tilborg, Finite fields and error correcting codes (1; 28 pp.)
Skew fields and division rings. Brauer group
Topological and valued fields. Valuation theory
Zeta and L-functions of fields and related topics
Structure of Galois modules
Constructive Galois theory (realizations of groups as Galois groups)

E. Nonabelian Class Field Theory and the Langlands Program

(To be arranged in several chapters by Y. Ihara)

F. Generalizations of Fields and Related Objects

U. Hebisch and H.J. Weinert, Semi-rings and semi-fields (1; 38 pp.)
G. Pilz, Near rings and near fields (1; 36 pp.)

Section 2. Category theory. Homological and homotopical algebra. Methods from logic

A. Category Theory

S. MacLane and I. Moerdijk, Topos theory (1; 28 pp.)
R. Street, Categorical structures (1; 50 pp.)

B.I. Plotkin, Algebra, categories and databases (2; 68 pp.)
P.S. Scott, Some aspects of categories in computer science (2; 73 pp.)
E. Manes, Monads of sets (3; 87 pp.)

B. Homological Algebra. Cohomology. Cohomological Methods in Algebra. Homotopical Algebra

J.F. Carlson, The cohomology of groups (1; 30 pp.)
A. Generalov, Relative homological algebra. Cohomology of categories, posets, and coalgebras (1; 28 pp.)
J.F. Jardine, Homotopy and homotopical algebra (1; 32 pp.)
B. Keller, Derived categories and their uses (1; 32 pp.)
A.Ya. Helemskii, Homology for the algebras of analysis (2; 122 pp.)
Galois cohomology
Cohomology of commutative and associative algebras
Cohomology of Lie algebras
Cohomology of group schemes

C. Algebraic K-theory

A. Kuku, Classical algebraic K-theory: the functors K_0, K_1, K_2 (3; 40 pp.)
Algebraic K-theory: the higher K-functors
Grothendieck groups
K_2 and symbols
KK-theory and EXT
Hilbert C^*-modules
Index theory for elliptic operators over C^* algebras
Algebraic K-theory (including the higher K_n)
Simplicial algebraic K-theory
Chern character in algebraic K-theory
Noncommutative differential geometry
K-theory of noncommutative rings
Algebraic L-theory
Cyclic cohomology

D. Model Theoretic Algebra

(See also P.C. Eklof, Whitehead modules, in Section 3B)
M. Prest, Model theory for algebra (3; 28 pp.)
M. Prest, Model theory and modules (3; 27 pp.)
Logical properties of fields and applications
Recursive algebras
Logical properties of Boolean algebras
F.O. Wagner, Stable groups (2; 40 pp.)

E. Rings up to Homotopy

Rings up to homotopy

Section 3. Commutative and associative rings and algebras

A. Commutative Rings and Algebras

(See also C. Faith, Coherent rings and annihilator conditions in matrix and polynomial rings, in Section 3B)

J.P. Lafon, Ideals and modules (1; 24 pp.)

General theory. Radicals, prime ideals etc. Local rings (general). Finiteness and chain conditions

Extensions. Galois theory of rings

Modules with quadratic form

Homological algebra and commutative rings. Ext, Tor, etc. Special properties (p.i.d., factorial, Gorenstein, Cohen–Macauley, Bezout, Fatou, Japanese, excellent, Ore, Prüfer, Dedekind, ... and their interrelations)

D. Popescu, Artin approximation (2; 34 pp.)

Finite commutative rings and algebras (see also Section 3B)

Localization. Local–global theory

Rings associated to combinatorial and partial order structures (straightening laws, Hodge algebras, shellability, ...)

Witt rings, real spectra

R.H. Villareal, Monomial algebras and polyhedral geometry (3; 58 pp.)

B. Associative Rings and Algebras

P.M. Cohn, Polynomial and power series rings. Free algebras, firs and semifirs (1; 30 pp.)

Classification of Artinian algebras and rings

V.K. Kharchenko, Simple, prime, and semi-prime rings (1; 52 pp.)

A. van den Essen, Algebraic microlocalization and modules with regular singularities over filtered rings (1; 28 pp.)

F. Van Oystaeyen, Separable algebras (2; 43 pp.)

K. Yamagata, Frobenius rings (1; 48 pp.)

V.K. Kharchenko, Fixed rings and noncommutative invariant theory (2; 38 pp.)

General theory of associative rings and algebras

Rings of quotients. Noncommutative localization. Torsion theories

von Neumann regular rings

Semi-regular and pi-regular rings

Lattices of submodules

A.A. Tuganbaev, Modules with distributive submodule lattice (2; 16 pp.)

A.A. Tuganbaev, Serial and distributive modules and rings (2; 19 pp.)

PI rings

Generalized identities

Endomorphism rings, rings of linear transformations, matrix rings

Homological classification of (noncommutative) rings

S.K. Sehgal, Group rings (3; 87 pp.)

Dimension theory

A. Facchini, The Krull–Schmidt theorem (3; 41 pp.)
Duality. Morita-duality
Commutants of differential operators
E.E. Enochs, Flat covers (3; 14 pp.)
C. Faith, Coherent rings and annihilator conditions in matrix and polynomial rings (3; 30 pp.)
Rings of differential operators
Graded and filtered rings and modules (also commutative)
P.C. Eklof, Whitehead modules (3; 25 pp.)
Goldie's theorem, Noetherian rings and related rings
Sheaves in ring theory
A.A. Tuganbaev, Modules with the exchange property and exchange rings (2; 19 pp.)
Finite associative rings (see also Section 3A)
T.Y. Lam, Hamilton's quaternions (3; 26 pp.)
A.A. Tuganbaev, Semiregular, weakly regular, and π-regular rings (3; 22 pp.)
Hamiltonian algebras
A.A. Tuganbaev, Max rings and V-rings (3; 20 pp.)

C. Coalgebras

W. Michaelis, Coassociative coalgebras (3; 202 pp.)

D. Deformation Theory of Rings and Algebras (*Including Lie Algebras*)

Deformation theory of rings and algebras (general)
Yu. Khakimdzanov, Varieties of Lie algebras (2; 31 pp.)

Section 4. Other algebraic structures. Nonassociative rings and algebras. Commutative and associative algebras with extra structure

A. Lattices and Partially Ordered Sets

Lattices and partially ordered sets
A. Pultr, Frames (3; 67 pp.)
Quantales

B. Boolean Algebras

C. Universal Algebra

D. Varieties of Algebras, Groups, ...

(See also Yu. Khakimdzanov, Varieties of Lie algebras, in Section 3D)
V.A. Artamonov, Varieties of algebras (2; 29 pp.)
Varieties of groups
V.A. Artamonov, Quasivarieties (3; 23 pp.)
Varieties of semigroups

E. Lie Algebras

Yu.A. Bahturin, **M.V. Zaitsev and A.A. Mikhailov**, Infinite-dimensional Lie superalgebras (2; 34 pp.)
General structure theory
Ch. Reutenauer, Free Lie algebras (3; 17 pp.)
Classification theory of semisimple Lie algebras over **R** and **C**
The exceptional Lie algebras
M. Goze and Y. Khakimdjanov, Nilpotent and solvable Lie algebras (2; 47 pp.)
Universal enveloping algebras
Modular (ss) Lie algebras (including classification)
Infinite-dimensional Lie algebras (general)
Kac–Moody Lie algebras

F. Jordan Algebras (finite and infinite dimensional and including their cohomology theory)

G. Other Nonassociative Algebras (Malcev, alternative, Lie admissable, ...)

Mal'tsev algebras
Alternative algebras

H. Rings and Algebras with Additional Structure

Graded and super algebras (commutative, associative; for Lie superalgebras, see Section 4E)
Topological rings
Hopf algebras
Quantum groups (general)
A.I. Molev, Yangians and their applications (3; 53 pp.)
Formal groups
F. Patras, Lambda-rings (3; 26 pp.)
Ordered and lattice-ordered groups, rings and algebras
Rings and algebras with involution. C^*-algebras
Difference and differential algebra. Abstract (and p-adic) differential equations. Differential extensions
Ordered fields

I. Witt Vectors

Witt vectors and symmetric functions. Leibnitz Hopf algebra and quasi-symmetric functions

Section 5. Groups and semigroups

A. Groups

A.V. Mikhalev and A.P. Mishina, Infinite Abelian groups: methods and results (2; 36 pp.)

Simple groups, sporadic groups
Abstract (finite) groups. Structure theory. Special subgroups. Extensions and decompositions
Solvable groups, nilpotent groups, p-groups
Infinite soluble groups
Word problems
Burnside problem
Combinatorial group theory
Free groups (including actions on trees)
Formations
Infinite groups. Local properties
Algebraic groups. The classical groups. Chevalley groups
Chevalley groups over rings
The infinite dimensional classical groups
Other groups of matrices. Discrete subgroups
Reflection groups. Coxeter groups
Groups with BN-pair, Tits buildings, ...
Groups and (finite combinatorial) geometry
"Additive" group theory
Probabilistic techniques and results in group theory
Braid groups
L. Bartholdi, **R.I. Grigorchuk and Z. Šunić**, Branch groups (3; 124 pp.)

B. *Semigroups*

Semigroup theory. Ideals, radicals, structure theory
Semigroups and automata theory and linguistics

C. *Algebraic Formal Language Theory. Combinatorics of Words*

D. *Loops, Quasigroups, Heaps, ...*

E. *Combinatorial Group Theory and Topology*

Section 6. Representation and invariant theory

A. *Representation Theory. General*

Representation theory of rings, groups, algebras (general)
Modular representation theory (general)
Representations of Lie groups and Lie algebras. General

B. *Representation Theory of Finite and Discrete Groups and Algebras*

Representation theory of finite groups in characteristic zero
Modular representation theory of finite groups. Blocks
Representation theory of the symmetric groups (both in characteristic zero and modular)

Representation theory of the finite Chevalley groups (both in characteristic zero and modular
Modular representation theory of Lie algebras

C. Representation Theory of 'Continuous Groups' (*linear algebraic groups, Lie groups, loop groups, ...*) *and the Corresponding Algebras*

Representation theory of compact topological groups
Representation theory of locally compact topological groups
Representation theory of $SL_2(\mathbf{R})$, ...
Representation theory of the classical groups. Classical invariant theory
Classical and transcendental invariant theory
Reductive groups and their representation theory
Unitary representation theory of Lie groups
Finite-dimensional representation theory of the ss Lie algebras (in characteristic zero); structure theory of semi-simple Lie algebras
Infinite dimensional representation theory of ss Lie algebras. Verma modules
Representation of Lie algebras. Analytic methods
Representations of solvable and nilpotent Lie algebras. The Kirillov orbit method
Orbit method, Dixmier map, ... for ss Lie algebras
Representation theory of the exceptional Lie groups and Lie algebras
Representation theory of 'classical' quantum groups
A.U. Klimyk, Infinite dimensional representations of quantum algebras (2; 27 pp.)
Duality in representation theory
Representation theory of loop groups and higher dimensional analogues, gauge groups, and current algebras
Representation theory of Kac–Moody algebras
Invariants of nonlinear representations of Lie groups
Representation theory of infinite-dimensional groups like GL_∞
Metaplectic representation theory

D. Representation Theory of Algebras

Representations of rings and algebras by sections of sheafs
Representation theory of algebras (Quivers, Auslander–Reiten sequences, almost split sequences, ...)

E. Abstract and Functorial Representation Theory

Abstract representation theory
S. Bouc, Burnside rings (2; 64 pp.)
P. Webb, A guide to Mackey functors (2; 30 pp.)

F. Representation Theory and Combinatorics

G. Representations of Semigroups

Representation of discrete semigroups
Representations of Lie semigroups

Section 7. Machine computation. Algorithms. Tables

Some notes on this volume: Besides some general article(s) on machine computation in algebra, this volume should contain specific articles on the computational aspects of the various larger topics occurring in the main volume, as well as the basic corresponding tables. There should also be a general survey on the various available symbolic algebra computation packages.

The CoCoA computer algebra system

Section 8. Applied algebra

Section 9. History of algebra

(See also K.T. Lam, Hamilton's quaternions, in Section 3A)
History of coalgebras and Hopf algebras

This Page Intentionally Left Blank

Contents

List of Contributors

V.A. Artamonov, *Moscow State University, Moscow*
L. Bartholdi, *Université de Genève, Genève*
P.C. Eklof, *University of California at Irvine, CA*
E.E. Enochs, *University of Kentucky, Lexington, KY*
A. Facchini, *Università di Padova, Padova*
C. Faith, *Rutgers University, Piscataway, NJ*
R.I. Grigorchuk, *Steklov Mathematical Institute, Moscow*
J.A. Hermida-Alonso, *Universidad de León, León*
A. Kuku, *Abdus Salam International Centre for Theoretical Physics, Trieste*
T.Y. Lam, *University of California, Berkeley, CA*
E. Manes, *University of Massachusetts, Amherst, MA*
W. Michaelis, *University of New Orleans, New Orleans, LA*
A.I. Molev, *University of Sydney, Australia*
F. Patras, *Université de Nice, Nice*
M. Prest, *University of Manchester, Manchester*
A. Pultr, *Charles University, Prague*
C. Reutenauer, *Université du Québec à Montréal, Montréal*
L. Rodman, *College of William and Mary, Williamsburg, VA*
S.K. Sehgal, *University of Alberta, Edmonton, AB*
Z. Šunić, *Cornell University, Ithaca, NY*
A.A. Tuganbaev, *Moscow State University, Moscow*
R.H. Villarreal, *Centro de Investigación y de Estudios Avanzados del I.P.N., Mexico City*

This Page Intentionally Left Blank

Section 1A
Linear Algebra

This Page Intentionally Left Blank

On Linear Algebra over Commutative Rings

José A. Hermida-Alonso

Departamento de Matemáticas, Universidad de León, 24071 León, Spain
E-mail: demjha@unileon.es

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

The goal of this part of the Handbook of Algebra is to review some results and methods in linear algebra over commutative rings. This theory is formed by many topics which have different motivations (in some cases by their applications and in other cases by a purely algebraic interest). We have chosen two subjects, systems of linear equations and linear dynamical systems over commutative rings, as a connecting theme for this chapter. A good introduction to many topics in this area can be found in the book of B.R. McDonald [89].

An old problem in mathematics was to solve a system of linear equations over the integers and the rationals. From back in antiquity till the 18th century it was thought sufficient to give rules for linear equations to reduce a system of n linear equations with n indeterminates to a unique equation $ax = b$ with $a \neq 0$ (when the homogeneous linear equations were not linearly independent one considered that the problem was badly posed). Systems of m linear equations and n indeterminates with $m < n$ were not considered.

The problem of obtaining integer solutions for a system of linear equations over $\mathbb{Z}$ was solved by H.J. Smith; previously C. Hermite had solved the problem in a particular case. L. Kronecker and G. Frobenius introduced the rank of a matrix and provided the results for systems over $\mathbb{R}$ or $\mathbb{C}$. E. Steinitz [112] showed that a system of linear equations (S): $A\underline{x} = \underline{b}$ over $\mathbb{Z}$ has an integer solution if and only if (i) $\operatorname{rank}(A) = \operatorname{rank}(A \mid \underline{b}) = r$ where $(A \mid \underline{b})$ is the augmented matrix of (S), and (ii) the greatest common divisor of all $r \times r$ minors of A is also the greatest common divisor of all $r \times r$ minors of $(A \mid \underline{b})$.

Section 2 concerns systems of linear equations over commutative rings. The first part contains results on the characterization of some classes of rings (Prüfer domains, Prüfer rings and integrally closed rings) in terms of systems of linear equations. The goal of the second part is to relate the structure of a finite free resolution of A with criteria to solve the systems $A\underline{x} = \underline{b}$.

During the sixties the work of R.E. Kalman provided the (classical) theory of linear dynamical systems with an algebraic structure. Kalman's formulation of the central concepts did not require that the systems considered were defined over a field and consequently the theory can be generalized to linear systems over commutative rings. The interest of this generalization in applications lies when one consider special rings: polynomial rings over $\mathbb{R}$ or $\mathbb{C}$, $\mathbb{Z}$, finite rings, rings of suitable smooth real or complex functions.

The problems that appear when extending classical results (in control theory) to linear systems over commutative rings have been intensively studied from the seventies till the present time. There exists an extensive bibliography about the different problems. The book of J. Brewer et al. [8] is a good introduction for some topics in this area and the book of E. Sontag [111] is a good source for classical control theory.

The feedback classification problem (to obtain a complete set of invariants, and a canonical form, that characterizes the feedback equivalence class of a linear system) and the pole-shifting problem (to modify the characteristic polynomial of a given linear system through the use of feedback) are the topics that we treat mainly in this paper.

P.A. Brunovsky [21] solved the feedback classification problem for reachable systems over a field (in this case, the Kronecker indices are a complete set of invariants). Over an arbitrary commutative ring the feedback classification problem is wild in the sense that it is unlikely that this problem will be solved, see [16]. Classification of matrices up to equivalence and up to similarity are two particular cases of the feedback classification problem and, as is well known, both problems are open.

Section 4 contains an extended version to commutative rings of Brunovsky's theorem, the classification of some special linear systems over a principal ideal domain or over a local ring and the characterization of the pointwise feedback relation. The dynamical feedback classification problem is briefly treated in Section 5.

The stabilization problem (given a linear system $\Sigma = (A, B)$, find a feedback matrix F such that $A + BF$ has all its eigenvalues in the open left half plane) in classical control theory is the origin of the pole-shifting problem. For an n-dimensional system $\Sigma = (A, B)$ the characteristic polynomial $\chi(A, T)$ of A is not, in general, a feedback invariant associated to Σ. The pole-shifting problem consists in characterizing what monic polynomials of degree n are of the form $\chi(A + BF, T)$ where F ranges over all feedback matrices. The problem is solved completely over a field. Different classes of rings (*PA*-rings, *FC*-rings, *BCS*-rings, *PS*-rings, ...) appear in the study of pole shifting over commutative rings. The characterization of these rings in algebraic terms is the central topic of Section 5.

1. Preliminaries

Throughout this chapter R will denote a commutative ring with an identity element, $\operatorname{Spec}(R)$ the set of prime ideals of R, and $\operatorname{Max}(R)$ the set of maximal ideals of R. The varieties

$$\begin{aligned}
&V(\mathfrak{a}) = \{\mathfrak{p} \in \operatorname{Spec}(R)\colon\ \mathfrak{a} \subseteq \mathfrak{p}\},\\
&(\text{resp. } V(\mathfrak{a}) \cap \operatorname{Max}(R) = \{\mathfrak{m} \in \operatorname{Max}(R)\colon\ \mathfrak{a} \subseteq \mathfrak{m}\}),
\end{aligned}$$

where $\mathfrak{a}$ ranges over all the ideals of R, are the closed subsets for the Zariski topology on $\operatorname{Spec}(R)$ (resp. $\operatorname{Max}(R)$).

For a prime ideal $\mathfrak{p}$ of R we denote by

$$\begin{aligned}
\iota_{\mathfrak{p}} : R &\to R_{\mathfrak{p}}\\
r &\mapsto r/1
\end{aligned}$$

the canonical ring homomorphism from R to the localization $R_{\mathfrak{p}}$ of R at $\mathfrak{p}$ and by

$$\begin{aligned}
\pi_{\mathfrak{p}} : R &\to k(\mathfrak{p})\\
r &\mapsto r(\mathfrak{p})
\end{aligned}$$

the canonical homorphism from R onto the residue field $k(\mathfrak{p}) = R_{\mathfrak{p}}/\mathfrak{p}R_{\mathfrak{p}}$ of $\mathfrak{p}$. For an R-module M we denote by $M_{\mathfrak{p}}$ the $R_{\mathfrak{p}}$-module obtained from M by extension of scalars via $\iota_{\mathfrak{p}}$ (i.e. $M_{\mathfrak{p}} = M \otimes_R R_{\mathfrak{p}}$) and by $M(\mathfrak{p})$ the $k(\mathfrak{p})$-vector space obtained from M by extension of scalars via $\pi_{\mathfrak{p}}$.

1.1. *Determinantal ideals. Ranks*

DEFINITION 1. Let $A = (a_{ij})$ be an $n \times m$ matrix with entries in R and let i be a non-negative integer. The i-th determinantal ideal of A, denoted by $\mathcal{U}_i(A)$, is the ideal of R generated by all the $i \times i$ minors of A.

By construction one has

$$R = \mathcal{U}_0(A) \supseteq \mathcal{U}_1(A) \supseteq \cdots \supseteq \mathcal{U}_i(A) \supseteq \cdots$$

and $\mathcal{U}_i(A) = 0$ for $i > \min\{m, n\}$. The rank of A, denoted by $\operatorname{rank}_R(A)$, is the largest i such that $\mathcal{U}_i(A) \neq 0$.

Let $f : R \to R'$ be a ring homomorphism. The extension of A to R' is the $n \times m$ matrix $f(A) = (f(a_{ij}))$. When R' is the local ring $R_\mathfrak{p}$ (resp. the residue field $k(\mathfrak{p})$) and f is the canonical homomorphism $\iota_\mathfrak{p} : R \to R_\mathfrak{p}$ (resp. $\pi_\mathfrak{p} : R \to k(\mathfrak{p})$), we denote by $A_\mathfrak{p}$ (resp. $A(\mathfrak{p})$) the matrix $f(A)$. Then

$$\mathcal{U}_i\big(f(A)\big) = \mathcal{U}_i(A)R',$$

where $\mathcal{U}_i(A)R'$ is the extension of $\mathcal{U}_i(A)$ to R'. Consequently

$$\operatorname{rank}_R(A) \geqslant \operatorname{rank}_{R'}\big(f(A)\big)$$

and

$$\operatorname{rank}_R(A) = \max\big\{\operatorname{rank}_{R_\mathfrak{p}}(A_\mathfrak{p}) \colon \mathfrak{p} \in \operatorname{Spec}(R)\big\}.$$

DEFINITION 2. The residue rank of A, denoted by $\operatorname{res.rank}_R(A)$, is the largest i such that $\mathcal{U}_i(A) = R$ or equivalently

$$\operatorname{res.rank}_R(A) = \min\big\{\operatorname{rank}_{k(\mathfrak{p})}\big(A(\mathfrak{p})\big) \colon \mathfrak{p} \in \operatorname{Spec}(R)\big\}.$$

We denote by φ_A the homomorphism of R-modules defined by

$$\begin{aligned} \varphi_A : R^m &\to R^n \\ \underline{\alpha} &\mapsto A\underline{\alpha}. \end{aligned}$$

The submodule $\operatorname{Im}\varphi_A$ of R^n generated by the columns of A is called basic (or nondegenerate) when $\operatorname{res.rank}_R(A) \geqslant 1$. In this case

$$\dim_{k(\mathfrak{p})}\big(\big(R^n / \operatorname{Im}\varphi_A\big) \otimes_R k(\mathfrak{p})\big) < n,$$

for all prime ideal $\mathfrak{p}$ of R. We shall consider these modules in Section 4.

PROPOSITION 3. *Let A be an $n \times m$ matrix with entries in R. Then the following statements are equivalent*:

(i) *For each* $\underline{b} \in R^n$ *the system of linear equations* $(S_{\underline{b}})$: $A\underline{x} = \underline{b}$ *has a solution in* R.
(ii) *The homomorphism* φ_A *is surjective.*
(iii) $\mathcal{U}_n(A) = R$.

PROOF. $\mathcal{U}_n(A) = R$ (resp. φ_A is surjective) if and only $\mathcal{U}_n(\pi_{\mathfrak{m}}(A)) = \mathcal{U}_n(A).R/\mathfrak{m} = R/\mathfrak{m}$ (resp. $\varphi_A \otimes Id_{R/\mathfrak{m}} = \varphi_{\pi_{\mathfrak{m}}(A)}$ is surjective) for all maximal ideal $\mathfrak{m}$ of R. □

The next result is known as McCoy's theorem, see [88] and [95, p. 63].

THEOREM 4. *Let* A *be an* $n \times m$ *matrix with entries in* R. *Then the following statements are equivalent*:
(i) *The trivial solution is the unique solution of the homogeneous system* (S_h): $A\underline{x} = \underline{0}$.
(ii) *The homomorphism* φ_A *is injective.*
(iii) $\mathcal{U}_m(A)$ *is a faithful ideal* (*i.e. the annihilator ideal* $(0 :_R \mathcal{U}_m(A))$ *is zero*).

1.2. *Equivalent matrices*

DEFINITION 5. The $n \times m$ matrix A' is equivalent to the $n \times m$ matrix A if there exist invertible matrices P and Q such that $A' = PAQ$.

Let A be an $n \times m$ matrix and B an $m \times p$ matrix. Then

$$(AB)^{(i)} = A^{(i)}B^{(i)},$$

where $A^{(i)}$ and $B^{(i)}$ denote the i-th exterior power of A and B, respectively. Therefore

$$\mathcal{U}_i(AB) \subseteq \mathcal{U}_i(A) \cap \mathcal{U}_i(B).$$

THEOREM 6. *If* A *and* A' *are equivalent matrices then* $\mathcal{U}_i(A) = \mathcal{U}_i(A')$ *for all* i.

Let $\varphi : R^m \to R^n$ be an homomorphism. Suppose that A (resp. A') is the matrix of φ relative to the bases $\mathcal{B}_1$ and $\mathcal{B}_2$ (resp. $\mathcal{B}'_1$ and $\mathcal{B}'_2$) of R^m and R^n, respectively. By the above result we can put

$$\mathcal{U}_i(\varphi) = \mathcal{U}_i(A) = \mathcal{U}_i(A')$$

for all i and, in particular, we put

$$\operatorname{rank}_R(\varphi) = \operatorname{rank}_R(A).$$

Recall that an elementary divisor domain R is a domain with the property that each matrix A is equivalent to a diagonal matrix

$$\operatorname{diag}(d_1, d_2, \ldots) = \begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix},$$

where d_i divides d_{i+1} for all i. In [80] it is proved that R is an elementary divisor domain if and only if all 2×1 and 2×2 matrices over R are equivalent to a diagonal matrix.

THEOREM 7. *Let R be an elementary divisor domain. Then two $n \times m$ matrices A and A' are equivalent if and only if $\mathcal{U}_i(A) = \mathcal{U}_i(A')$ for all i.*

REMARK 8. A Bezout domain is a domain in which any finitely generated ideal is principal, see [6, p. 280]. It is clear that an elementary divisor domain is a Bezout domain. An open question is: Every Bezout domain is an elementary divisor domain? Bezout domains having only countably many maximal ideals, one-dimensional Bezout domains and the ring of real analytic functions are elementary divisor domains, see [8, Theorems 3.14 and 3.15].

PROPOSITION 9. *Let R be a local ring and A be an $n \times m$ matrix with entries in R. Then* $\operatorname{res.rank}_R(A) = r$ *if and only if A is equivalent to a matrix of the form*

$$\left(\begin{array}{c|c} Id_r & 0 \\ \hline 0 & 0 \end{array}\right)$$

where Id_r is the $r \times r$ identity matrix.

REMARK 10. In Proposition 150 we characterize the class of rings which satisfy the hypotheses of the above result.

If A is equivalent to A' then it is clear that $\operatorname{coker}\varphi_A$ is isomorphic to $\operatorname{coker}\varphi_{A'}$ but the converse is not true. When φ_A and $\varphi_{A'}$ are injective then, by the short five lemma, the converse is true. In [120] the next characterization is proved.

PROPOSITION 11. *Let A and A' be two $n \times m$ matrices with entries in R. Then* $\operatorname{coker}\varphi_A$ *is isomorphic to* $\operatorname{coker}\varphi_{A'}$ *if and only if A and A' are Fitting equivalent* (*i.e. the matrices*

$$\begin{pmatrix} A & 0 & 0 \\ 0 & Id & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} Id & 0 & 0 \\ 0 & A' & 0 \end{pmatrix}$$

are equivalent for suitable blocks of identities and zeros).

When R is a principal ideal domain or a Dedekind domain it is well know that two $n \times m$ matrices A and A' are equivalent if and only if $\operatorname{coker} \varphi_A$ is isomorphic to $\operatorname{coker} \varphi_{A'}$. In [121] the following result is proved:

PROPOSITION 12. *Suppose that R is an stable ring (i.e. if whenever the ideal generated by a and b is R, there exists a c such that $a + bc$ is a unit). Then two $n \times m$ matrices A and A' are equivalent if and only if* $\operatorname{coker} \varphi_A$ *is isomorphic to* $\operatorname{coker} \varphi_{A'}$.

REMARK 13. Stable rings were introduced in [44]. See [89, p. 53] for properties and applications of these rings.

1.3. *Fitting ideals*

Let

$$R^m \xrightarrow{\varphi_A} R^n \to M \to 0$$

and

$$R^{m'} \xrightarrow{\varphi_{A'}} R^{n'} \to M \to 0$$

be two finite free presentations of a module M. Then

$$\begin{aligned} &\mathcal{U}_{n-i}(A) = \mathcal{U}_{n'-i}(A'), \quad i \leqslant \min\{n, m\}, \\ &\mathcal{U}_i(A) = \mathcal{U}_i(A') = 0, \quad i > \min\{n, m\}, \end{aligned}$$

and, consequently, the ideals

$$\mathcal{F}_i(M) = \begin{cases} \mathcal{U}_{n-i}(A) & \text{for } 0 \leqslant i \leqslant n, \\ R & \text{for } i > n \end{cases}$$

are invariants associated to M (i.e. they only depend on M and not on the finite free presentation chosen). These ideals are know as the Fitting invariants of M, see [46].

In a similar way, see [95, p. 54], the Fitting invariants of a finitely generated R-module can be constructed. Next we collect some properties of these ideals.

PROPOSITION 14. *The Fitting invariants satisfy the following properties*:

(i) *Let M be a finitely generated R-module. Then*

$$\mathcal{F}_0(M) \subseteq \mathcal{F}_1(M) \subseteq \cdots \subseteq \mathcal{F}_i(M) \subseteq \cdots.$$

Moreover if M can be generated by m elements then $\mathcal{F}_i(M) = R$ for $i \geqslant m$.

(ii) *Let $0 \to M' \to M \to M'' \to 0$ be an exact sequence of finitely generated R-modules. Then*

$$\mathcal{F}_i(M').\mathcal{F}_j(M'') \subseteq \mathcal{F}_{i+j}(M).$$

(iii) *Suppose that the above sequence splits. Then*

$$\mathcal{F}_i(M' \oplus M'') = \sum_{j+k=i} \mathcal{F}_j(M').\mathcal{F}_k(M'').$$

(iv) *Let* $\mathfrak{a}$ *be an ideal of* R. *Then*

$$\mathcal{F}_i(R/\mathfrak{a}) = \begin{cases} \mathfrak{a} & \text{for } i = 0, \\ R & \text{for } i > 0. \end{cases}$$

(v) *Let* $f : R \to R'$ *be a homomorphism of rings. If* M *is a finitely generated* R*-module then*

$$\mathcal{F}_i(M_{(R')}) = \mathcal{F}_i(M).R',$$

where $M_{(R')}$ *is the* R'*-module obtained from* M *by extension of scalars. In particular*

$$\mathcal{F}_i(M_{\mathfrak{p}}) = \mathcal{F}_i(M)R_{\mathfrak{p}}$$

for each prime ideal $\mathfrak{p}$ *of* R.

(vi) *Suppose that* M *can be generated by* n *elements then*

$$\mathrm{Ann}_R(M)^n \subseteq \mathcal{F}_0(M) \subseteq \mathrm{Ann}_R(M).$$

In particular

$$\mathrm{Supp}(M) = V\big(\mathcal{F}_0(M)\big) = \big\{\mathfrak{p} \in \mathrm{Spec}(R)/\mathcal{F}_0(M) \subseteq \mathfrak{p}\big\},$$

where $\mathrm{Supp}(M) = \{\mathfrak{p} \in \mathrm{Spec}(R)\colon\ M_{\mathfrak{p}} \neq 0\}$ *is the support of* M.

(vii) $V(\mathcal{F}_i(M)) = \{\mathfrak{p} \in \mathrm{Spec}(R)\colon\ \dim_{k(\mathfrak{p})} M(\mathfrak{p}) \geqslant i\}$.

PROOF. See [95] and [59]. □

Let R be an elementary divisor domain and

$$R^m \xrightarrow{\varphi_A} R^n \to M \to 0$$

a finite presentation of a module M. If A is equivalent to the diagonal matrix $D = \mathrm{diag}(d_1, d_2, \ldots, d_r)$ where $d_1|d_2|\ldots|d_r$ then

$$\mathcal{F}_i(M) = \mathcal{U}_{n-i}(A) = \mathcal{U}_{n-i}(D) = \begin{cases} 0 & \text{for } i < n-r, \\ (d_1 d_2 \ldots d_{n-i}) & \text{for } n-r \leqslant i < n, \\ R & \text{for } n < i. \end{cases}$$

Conversely, if the sequence $\{\mathcal{F}_i(M)\}_{i\geqslant 0}$ is known then

$$M \cong R^r \oplus R/\big(\mathcal{F}_r(M):\mathcal{F}_{r+1}(M)\big) \oplus R/\big(\mathcal{F}_{r+1}(M):\mathcal{F}_{r+2}(M)\big) \oplus \cdots,$$

where r is the smallest integer such that $\mathcal{F}_r(M) \neq 0$.

REMARK 15. In [84] it is proved that R is an elementary divisor ring if and only if every finitely presented R-module is a direct sum of cyclic modules. See [123] for the characterization of rings R whose finitely generated modules are direct sums of cyclic modules.

Let R be a Prüfer domain (i.e. $R_{\mathfrak{p}}$ is a valuation domain for all prime ideal $\mathfrak{p}$ of R) and M a finitely presented R-module. Put

$$\overline{M} = R^r \oplus R/\big(\mathcal{F}_r(M):\mathcal{F}_{r+1}(M)\big) \oplus R/\big(\mathcal{F}_{r+1}(M):\mathcal{F}_{r+2}(M)\big) \oplus \cdots,$$

where r is the smallest integer such that $\mathcal{F}_r(M) \neq 0$. Since $R_{\mathfrak{p}}$ is an elementary divisor domain then, by property (v) of Fitting ideals, $M_{\mathfrak{p}}$ is isomorphic to $\overline{M}_{\mathfrak{p}}$ for each prime ideal $\mathfrak{p}$ of R. By the quasi-compactness of $\mathrm{Spec}(R)$ one has the next result, see [69].

PROPOSITION 16. *Let R be a Prüfer domain and M a finitely presented R-module. Then M is a direct summand of a finite sum of copies of $\overline{M}$.*

Now suppose that R is a Noetherian Prüfer domain (i.e. R is a Dedekind domain) and M is a finitely generated torsion R-module. Put $S = \bigcap_{1\leqslant i\leqslant t}(R - \mathfrak{p}_i)$ where $V(\mathcal{F}_0(M)) = \{\mathfrak{p}_1, \mathfrak{p}_2, \ldots, \mathfrak{p}_t\}$. Since $S^{-1}R$ is a principal ideal domain it follows that $S^{-1}M$ is isomorphic to $S^{-1}\overline{M}$.

PROPOSITION 17. *Let R be a Dedekind domain and M a finitely generated torsion R-module. Then M is isomorphic to $\overline{M}$.*

1.4. *Projective modules. Flat modules*

DEFINITION 18. Let R be a commutative ring and M an R-module.
- (i) M is projective if M is a direct summand of a free R-module or equivalently the functor $\mathrm{Hom}(M, \_)$ is exact.
- (ii) M is flat if the functor $M \otimes_R \_$ is exact.

See [5, Chapters I and II] and [82] for the main properties of these modules. When M is finitely generated one has the following characterization.

THEOREM 19. *Let M be a finitely generated R-module. Then*:
- (i) *M is flat if and only if M_p is a free R_p-module for all prime ideals $\mathfrak{p}$ of R.*

(ii) *M is projective if and only if M is flat and the rank function*

$$\begin{aligned}\mathrm{rk}: \mathrm{Spec}(R) &\to \mathbb{Z}\\ \mathfrak{p} &\mapsto \mathrm{rank}_{R_\mathfrak{p}} M_\mathfrak{p}\end{aligned}$$

is continuous.

Next we shall characterize, in terms of Fitting ideals, when a finitely generated R-module is flat or projective. By Propositions 9 and 14 one has:

LEMMA 20. *Let R be a local ring and M a finitely generated R-module. Then the following statements are equivalent*:

(i) *M is free of* rank *r.*

(ii) *$\mathcal{F}_{r-1}(M) = 0$ and $\mathcal{F}_r(M) = R$.*

The following result is a consequence of statement (v) of Proposition 14.

PROPOSITION 21. *Let R be a commutative ring and M a finitely generated R-module. Then the following statements are equivalent*:

(i) *M is flat* (*resp. projective*).

(ii) *The finitely generated R-module*

$$R/\mathcal{F}_0(M) \oplus R/\mathcal{F}_1(M) \oplus \cdots \oplus R/\mathcal{F}_i(M) \oplus \cdots$$

is flat (*resp. projective*).

(iii) *The cyclic R-module $R/\mathcal{F}_i(M)$ is flat* (*resp. projective*) *for each i.*

REMARK 22. For a maximal ideal $\mathfrak{m}$ of R we denote by $S_m(0)$ the kernel of the canonical ring homomorphism $\iota_\mathfrak{m} : R \to R_\mathfrak{m}$. If $\mathfrak{a}$ is an ideal of R then

(i) $R/\mathfrak{a}$ is projective if and only if $\mathfrak{a} = \bigcap_{\mathfrak{m} \in V(\mathfrak{a})} S_m(0)$ and $V(\mathfrak{a})$ is an open-closed subset of $\mathrm{Spec}(R)$.

(ii) $R/\mathfrak{a}$ is flat if and only if $\mathfrak{a} = \bigcap_{\mathfrak{m} \in V(\mathfrak{a})} S_m(0)$ and for each $\mathfrak{p} \notin V(\mathfrak{a})$ there exist Zariski open subsets U and U' such that $\mathfrak{p} \in U$, $V(\mathfrak{a}) \subseteq U'$ and $U \cap U' = \emptyset$ (see, Axiom O_{III} in [7, Ch. I, p. 57]).

Since a cyclic module $R/\mathfrak{a}$ is projective if and only if $\mathfrak{a}$ is generated by an idempotent element there follows the next result, see [95, p. 123] and [103].

THEOREM 23. *Let R be a commutative ring and M a finitely generated R-module. Then the following statements are equivalent*:

(i) *M is projective.*

(ii) *For each $i \geqslant 0$ the ideal $\mathcal{F}_i(M)$ is generated by an idempotent.*

In particular, M is projective of constant rank *r if and only if $\mathcal{F}_r(M) = R$ and $\mathcal{F}_{r-1}(M) = 0$.*

REMARK 24. Let M be a finitely generated projective R-module. If $\mathcal{F}_i(M) = (e'_i)$, where e'_i is idempotent, then the set

$$X_i = \{\mathfrak{p} \in \operatorname{Spec}(R)\colon \operatorname{rank}_{R_\mathfrak{p}} M_\mathfrak{p} = i\} = V(e'_{i-1}) \cap D(e'_i),$$

is an open-closed subset of $\operatorname{Spec}(R)$. Hence there exist idempotent elements $e_0, e_1, \ldots, e_s$ such that $X_i = D(e_i)$ for each i, $e_i e_j = 0$ for $i \neq j$ and $1 = e_0 + e_1 + \cdots + e_s$.

Suppose that the column matrix $A = (a_1, a_2, \ldots, a_n)^t$ is unimodular (i.e. $\mathcal{U}_1(A) = R$) and put $M = \operatorname{Coker}(\varphi_A)$. By Theorem 23, M is projective and hence $M \oplus R \cong R^n$ because the exact sequence

$$0 \to R \xrightarrow{\varphi_A} R^n \to M \to 0$$

splits. However, M is free if and only if $(a_1, a_2, \ldots, a_n)^t$ is the first column of an invertible matrix.

DEFINITION 25. A projective module M is called stably free or supplementable if there exists a positive integer t such that $M \oplus R^t$ is free.

THEOREM 26. *The following statements are equivalent*:

(i) *Finitely generated stably free R-modules are free.*

(ii) *Every unimodular vector in R^n is part of a basis of R^n, for each positive integer n.*

(iii) *If $\underline{b} \in R^n$ is unimodular then there exists an invertible matrix P such that $P\underline{b} = (1, 0, \ldots, 0)^t$ (i.e. $GL_n(R)$ acts transitively on unimodular vectors of R^n).*

If R is the ring of continuous real-valued functions on the sphere or the quotient ring $R = \mathbb{Z}[X_1, X_2, \ldots, X_n]/(X_1^2 + X_2^2 + \cdots + X_n^2 - 1)$ then R possesses a non free finitely generated stably free module, see [95, p. 235]. If k is a field then every finitely generated projective $k[X_1, X_2, \ldots, X_n]$-module is free, see [98,113,83].

1.5. *Flat modules and linear equations*

Let $f: R \to R'$ be a ring homomorphism and (S): $A\underline{x} = \underline{b}$ a system of linear equations over R. The extension of (S): $A\underline{x} = \underline{b}$ to R' is the system

$$\big(f(S)\big)\colon\ f(A)\underline{x} = f(\underline{b})$$

where $f(A) = (f(a_{ij}))$ and $f(\underline{b}) = (f(b_1), f(b_2), \ldots, f(b_n))^t$. When R' is a local ring $R_\mathfrak{p}$ (resp. a residue field $k(\mathfrak{p})$) and f is the canonical homomorphism $\iota_\mathfrak{p}: R \to R_\mathfrak{p}$ (resp. $\pi_\mathfrak{p}: R \to k(\mathfrak{p})$), we denote by $(S_\mathfrak{p})$ (resp. $(S(\mathfrak{p}))$ the system $(f(S))$.

THEOREM 27. *Let $f: R \to R'$ be a ring homomorphism. The following statements are equivalent*:

(i) *The R-module (via f) R' is faithfully flat (i.e. R' is a flat R-module and $\mathfrak{m}R' \neq R'$ for every maximal ideal $\mathfrak{m}$ of R).*
(ii) *If (S): $A\underline{x} = \underline{b}$ is a system of linear equations over R and $\underline{\alpha}$ is a solution of $(f(S))$ then there exist solutions $\underline{\beta}_1, \underline{\beta}_2, \ldots, \underline{\beta}_t$ of (S) such that*

$$\underline{\alpha} = \sum_{i=1}^{t} \lambda_i f(\underline{\beta}_i)$$

where $\lambda_i \in R'$ for $i = 1, \ldots, t$.

REMARK 28. The characterization of flat extensions is the same but uses homogeneous systems of linear equations instead, see [81, p. 243].

PROPOSITION 29. *Let (S): $A\underline{x} = \underline{b}$ be a system of linear equations over R. The following statements are equivalent*:
(i) *(S) has a solution in R.*
(ii) *$(S_{\mathfrak{p}})$ has a solution in $R_{\mathfrak{p}}$ for every prime ideal $\mathfrak{p}$ of R.*
(iii) *$(S_{\mathfrak{m}})$ has a solution in $R_{\mathfrak{m}}$ for every maximal ideal $\mathfrak{m}$ of R.*

PROOF. The R-modules $\prod_{\mathfrak{p}\in\mathrm{Spec}(R)} R_{\mathfrak{p}}$ and $\prod_{\mathfrak{m}\in\mathrm{Max}(R)} R_{\mathfrak{m}}$ are faithfully flat because the extension $R \to R_{\mathfrak{p}}$ is flat for every prime ideal $\mathfrak{p}$. $\square$

2. Systems of linear equations

Let (S): $A\underline{x} = \underline{b}$ be a system of n linear equations in m unknowns (i.e. $A = (a_{ij})$ is an $n \times m$ matrix with entries in R and $\underline{b} = (b_1, b_2, \ldots, b_n)^t$ is an element of R^n). If (S) has a solution in R then

$$\mathcal{U}_i(A) = \mathcal{U}_i(A \mid \underline{b}) \quad \text{for all } i \geqslant 0,$$

where $(A \mid \underline{b})$ is the augmented matrix of the system (i.e. the $n \times (m+1)$ matrix obtained from A by adding the column $\underline{b}$) and in particular

$$\mathrm{rank}_R(A) = \mathrm{rank}_R(A \mid \underline{b}).$$

The question is: When these conditions are sufficient to assure that (S) has a solution in R?

2.1. *Characterizing rings via linear equations*

LEMMA 30. *Suppose that* $\mathrm{res.rank}_R(A) = \mathrm{rank}_R(A)$. *Then (S): $A\underline{x} = \underline{b}$ has a solution in R if and only if* $\mathrm{rank}_R(A) = \mathrm{rank}_R(A \mid \underline{b})$.

PROOF. By Proposition 29 we can suppose that R is local. By Proposition 9 there exist invertible matrices P and Q such that

$$PAQ = \left(\begin{array}{c|c} Id_r & 0 \\ \hline 0 & 0 \end{array}\right).$$

The result follows by considering the system (S'): $PAQ\underline{x} = P\underline{b}$. □

The first characterization result is well known.

COROLLARY 31. *Let R be a commutative ring. The following statements are equivalent*:

(i) *R is a field.*

(ii) *A system of linear equations (S): $A\underline{x} = \underline{b}$ over R has a solution in R if and only if* $\operatorname{rank}_R(A) = \operatorname{rank}_R(A \mid \underline{b})$.

2.1.1. *Integrally closed rings*

DEFINITION 32. We say that (S): $A\underline{x} = \underline{b}$ is an overdetermined system if $\operatorname{rank}_R(A) = m$. When $m = 1$ the system is called a proportionality.

Consider the proportionality (P): $\underline{a}x = \underline{b}$ given by

$$(P)\colon \begin{cases} a_1 x = b_1, \\ a_2 x = b_2, \\ \quad\vdots \\ a_n x = b_n \end{cases}$$

and suppose that $\mathcal{U}_i(\underline{a}) = \mathcal{U}_i(\underline{a} \mid \underline{b})$ for $i = 1, 2$ and that $\mathcal{U}_1(\underline{a})$ contains a nonzero divisor. Then the extension of (P) to the total quotient ring $\mathcal{T}(R)$ of R verifies

$$\mathcal{U}_1(\underline{a})\mathcal{T}(R) = \mathcal{T}(R),$$

and $\operatorname{rank}_{\mathcal{T}(R)}(\underline{a}) = \operatorname{rank}_{\mathcal{T}(R)}(\underline{a} \mid \underline{b})$. By Lemma 30, (P) has a unique solution α in $\mathcal{T}(R)$. The condition $\mathcal{U}_1(\underline{a}) = \mathcal{U}_1(\underline{a} \mid \underline{b})$ implies that there exist elements $c_{ij} \in R$ such that

$$\begin{aligned} a_1\alpha &= a_1c_{11} + a_2c_{12} + \cdots + a_nc_{1n}, \\ a_2\alpha &= a_1c_{21} + a_2c_{22} + \cdots + a_nc_{2n}, \\ &\vdots \\ a_n\alpha &= a_1c_{n1} + a_2c_{n2} + \cdots + a_nc_{nn}. \end{aligned}$$

Therefore

$$\alpha^n + c_{n-1}\alpha^{n-1} + \cdots + c_1\alpha + c_0 = 0,$$

for some elements c_i of R (i.e. α is integral over R).

Recall that R is an integrally closed ring if every $\alpha \in \mathcal{T}(R)$ which is integral over R belongs to R. For domains see [6] and for rings with zero divisors see [73].

Considering the relation between proportionalities and overdetermined systems we prove, see [48], the following characterization.

THEOREM 33. *Let R be a commutative ring with an identity element. The following statements are equivalent*:

(i) *R is an integrally closed ring.*

(ii) *A proportionality (P): $\underline{a}x = \underline{b}$ such that $\mathcal{U}_1(\underline{a})$ contains a nonzero divisor has a solution in R if and only if $\mathcal{U}_i(\underline{a}) = \mathcal{U}_i(\underline{a} \mid \underline{b})$ for $i = 1, 2$.*

(iii) *An overdetermined system (S): $A\underline{x} = \underline{b}$, where A is an $n \times m$ matrix such that $\mathcal{U}_m(A)$ contains a nonzero divisor, has a solution in R if and only if $\mathcal{U}_i(A) = \mathcal{U}_i(A \mid \underline{b})$ for $i = m, m+1$.*

Removing the hypothesis $\mathcal{U}_1(\underline{a})$ that contains a nonzero divisor we obtain the following characterization.

THEOREM 34. *Let R be a commutative ring with an identity element. The following statements are equivalent*:

(i) *The local ring $R_{\mathfrak{m}}$ is an integrally closed domain for every maximal ideal $\mathfrak{m}$ of R.*

(ii) *A proportionality (P): $\underline{a}x = \underline{b}$ has a solution in R if and only if $\mathcal{U}_i(\underline{a}) = \mathcal{U}_i(\underline{a} \mid \underline{b})$ for $i = 1, 2$.*

Let A be an $n \times m$ matrix such that $\mathcal{U}_m(A)$ is faithful (or equivalently, the homomorphism φ_A is injective). If $a_1, a_2, \dots, a_t$ generate $\mathcal{U}_m(A)$ then

$$p(T) = a_1 + a_2 T + \cdots + a_t T^{t-1}$$

is a nonzero divisor in the polynomial ring $R[T]$, because $p(T)$ is a zero divisor in $R[T]$ if and only if there exists $r \in R$ such that $r.p(T) = 0$. From Theorems 27 and 33 we obtain the following characterization result, see [48].

THEOREM 35. *Let R be a commutative ring with an identity element. The following statements are equivalent*:

(i) *$R[T]$ is an integrally closed ring.*

(ii) *R satisfies the following properties*:
- *R is a reduced ring.*
- *A proportionality (P): $\underline{a}x = \underline{b}$ such that $\mathcal{U}_1(\underline{a})$ is a faithful ideal has a solution in R if and only if $\mathcal{U}_i(\underline{a}) = \mathcal{U}_i(\underline{a} \mid \underline{b})$ for $i = 1, 2$.*

(iii) *R satisfies the following properties*:
- *R is a reduced ring.*
- *An overdetermined system (S): $A\underline{x} = \underline{b}$, where A is an $n \times m$ matrix such that $\mathcal{U}_m(A)$ is faithful, has a solution in R if and only if $\mathcal{U}_i(A) = \mathcal{U}_i(A \mid \underline{b})$ for $i = m, m+1$.*

REMARK 36. Let R be a reduced ring and consider the following statements:

(I) $R_{\mathfrak{m}}$ is an integrally closed domain for every maximal ideal $\mathfrak{m}$ of R.

(II) $R[T]$ is an integrally closed ring.

(III) R is an integrally closed ring.

As a consequence of the above results one has the chain of implications

$$(\mathrm{I}) \Rightarrow (\mathrm{II}) \Rightarrow (\mathrm{III}).$$

It is clear that if R satisfies *property A* (i.e. each finitely generated faithful ideal contains a nonzero divisor) then (II) $\Leftrightarrow$ (III). This fact and (I) $\Rightarrow$ (II) are proved in [1]. When $\mathrm{Min}(R)$ with the Zariski topology is quasi-compact then (I) $\Leftrightarrow$ (II). Other characterizations for the integrally-closeness of $R[T]$ are given in [85,86].

2.1.2. *Prüfer rings* By Proposition 29 the existence of a solution for a system of linear equations is a local property. Using this fact one can prove, see [67], the following result which is an extension of Theorem 34 to arbitrary systems.

THEOREM 37. *Let R be a commutative ring. The following statements are equivalent*:

(i) *R is a Prüfer ring* (*i.e. every finitely generated ideal of R is flat*).

(ii) *A system of linear equations* (S): $A\underline{x} = \underline{b}$ *over R has a solution in R if and only if* $\mathcal{U}_i(A) = \mathcal{U}_i(A \mid \underline{b})$ *for* $i \geqslant 0$.

REMARK 38. When R is a domain the above result is proved in [27]. Steinitz, in [112], proves that statement (ii) holds when R is the domain of algebraic integers.

REMARK 39. See [6, p. 287] for characterizations of Prüfer domains. Note that R is a Prüfer ring if and only if $R_{\mathfrak{m}}$ is a valuation domain for all maximal ideal $\mathfrak{m}$ of R.

THEOREM 40. *Let R be a commutative ring. The following statements are equivalent*:

(i) *R is a Prüfer ring.*

(ii) *For every finitely generated ideal $\mathfrak{a}$ of R and for every system* (S): $A\underline{x} = \underline{b}$ *over R the system* $(\pi(S))$: $\pi(A)\underline{x} = \pi(\underline{b})$, *where $\pi : R \to R/\mathfrak{a}$ is the canonical homomorphism, has a solution in $R/\mathfrak{a}$ if and only if*

$$\mathcal{U}_i(A \mid \underline{b}) \subseteq \mathfrak{a}^i + \mathfrak{a}^{i-1}\mathcal{U}_1(A) + \cdots + \mathcal{U}_i(A)$$

for $i \geqslant 0$.

PROOF. See [69]. $\square$

2.2. *Systems of linear equations and finite free resolutions*

2.2.1. *Unique solution case* Let M be an R-module. Recall that a sequence $\{a_1, a_2, \ldots, a_n\}$ of elements of R is called an R-sequence on M if and only if

$$\big((a_1, a_2, \ldots, a_{i-1})M :_M a_i\big) = (a_1, a_2, \ldots, a_{i-1})M,$$

for $i = 1, 2, \ldots, n$ where

$$\big((a_1, a_2, \ldots, a_{i-1})M :_M a_i\big) = \big\{v \in M / a_i v \in (a_1, a_2, \ldots, a_{i-1})M\big\}.$$

The classical grade of an ideal $\mathfrak{a}$ on M, denoted by $gr_R\{\mathfrak{a}, M\}$, is the upper bound of the lengths of all R-sequences on M of elements of $\mathfrak{a}$. The (classical) grade of $\mathfrak{a}$ is $gr_R\{\mathfrak{a}, R\}$.

Since $\{a_1, a_2, \ldots, a_n\}$ is an R-sequence on M if and only if $\{a_1, a_2, \ldots, a_n\}$ is an $R[T]$-sequence on $M[T]$ it follows that

$$\begin{aligned} &gr_{R[T_1,T_2,\ldots,T_r]}\big\{\mathfrak{a}R[T_1, T_2, \ldots, T_r], M[T_1, T_2, \ldots, T_r]\big\} \\ &\quad \leqslant gr_{R[T_1,T_2,\ldots,T_{r+1}]}\big\{\mathfrak{a}R[T_1, T_2, \ldots, T_{r+1}], M[T_1, T_2, \ldots, T_{r+1}]\big\}. \end{aligned}$$

Now we can introduce the next concept due to Hochster, see [95, p. 148].

DEFINITION 41. Let $\mathfrak{a}$ be an ideal of R and M an R-module. The true grade or polynomial grade of $\mathfrak{a}$ on M is defined by

$$Gr_R\{\mathfrak{a}, M\} = \lim_{r\to\infty} gr_{R[T_1,T_2,\ldots,T_r]}\big\{\mathfrak{a}R[T_1, T_2, \ldots, T_r], M[T_1, T_2, \ldots, T_r]\big\}.$$

When $M = R$ we write $gr_R\{\mathfrak{a}\}$ and $Gr_R\{\mathfrak{a}\}$ instead of $gr_R\{\mathfrak{a}, R\}$ and $Gr_R\{\mathfrak{a}, R\}$, respectively.

PROPOSITION 42. *Let $\mathfrak{a}$ be an ideal of R and M an R-module.*

(i) *Suppose that $\mathfrak{a}$ is finitely generated. Then*:
- $gr_R\{\mathfrak{a}\} \geqslant 1$ *if and only if $\mathfrak{a}$ contains a nonzero divisor and*
- $Gr_R\{\mathfrak{a}\} \geqslant 1$ *if and only if $\mathfrak{a}$ is faithful.*

(ii) *When R is a Noetherian ring $gr_R\{\mathfrak{a}\} = Gr_R\{\mathfrak{a}\}$.*

(iii) *Suppose that $a_1, a_2, \ldots, a_n$ generate $\mathfrak{a}$. Then $Gr_R\{\mathfrak{a}\} \geqslant s$ if and only if*

$$\big\{f_i = a_1 + a_2T_i + \cdots + a_nT_i^{n-1}\big\}_{i=1,2,\ldots,s},$$

is an $R[T_1, T_2, \ldots, T_s]$-sequence.

(iv) *$Gr_R\{\mathfrak{a}\} \sup_{\mathfrak{b}}\{Gr_R\{\mathfrak{b}\}\}$, where $\mathfrak{b}$ ranges over all finitely generated ideals contained in $\mathfrak{a}$.*

(v) *There exists a prime ideal $\mathfrak{p}$ of R such that $\mathfrak{a} \subseteq \mathfrak{p}$ and $Gr_R\{\mathfrak{a}, M\} = Gr_R\{\mathfrak{p}, M\}$.*

(vi) *If $f : R \to R'$ is a ring homomorphism and M' is an R'-module. Then $Gr_R\{\mathfrak{a}, M'\} = Gr_{R'}\{\mathfrak{a}R', M'\}$. In particular, $Gr_R\{\mathfrak{a}, R'\} = Gr_{R'}\{\mathfrak{a}R'\}$.*

PROOF. See [95, Chapter 5] for the main properties of true grade. □

REMARK 43. Note that Theorems 33 and 35 can be formulated in terms of classical grade and true grade, respectively.

Suppose that the proportionality

$$(P)\colon \begin{cases} a_1 x = b_1, \\ a_2 x = b_2, \\ \quad \vdots \\ a_n x = b_n \end{cases}$$

satisfies $\operatorname{rank}(\underline{a}) = \operatorname{rank}(\underline{a} \mid \underline{b})$ and $Gr_R\{\mathcal{U}_1(\underline{a})\} \geqslant 2$. By statement (ii),

$$\left\{ a = \sum_{i=1}^{n} T_1^{i-1} a_i,\ a' = \sum_{i=1}^{n} T_2^{i-1} a_i \right\}$$

is an $R[T_1, T_2]$-sequence contained in $\mathcal{U}_1(\underline{a})R[T_1, T_2]$. Put $b = \sum_{i=1}^{n} T_1^{i-1} b_i$ and $b' = \sum_{i=1}^{n} T_2^{i-1} b_i$. Then (P) is equivalent to the proportionality

$$(P')\colon \begin{cases} a_1 x = b_1, \\ \quad \vdots \\ a_n x = b_n, \\ ax = b, \\ a'x = b'. \end{cases}$$

The equality $ab' = a'b$ implies that there exists $\alpha \in R[T_1, T_2]$ such that $a\alpha = b$ and $a'\alpha = b'$. Since $ab_i = a_i b$ it follows that $a_i \alpha = b_i$ for all i because a is a nonzero divisor. Therefore α is the unique solution of (P) in $R[T_1, T_2]$ and, by Theorem 27, α is the unique solution of (P) in R.

Now considering the relation between proportionalities and overdetermined systems we can prove the next result, see [107].

THEOREM 44. *Let A be an $n \times m$ matrix such that $Gr_R\{\mathcal{U}_m(A)\} \geqslant 2$. Then for $\underline{b} \in R^n$ the system (S): $A\underline{x} = \underline{b}$ has a solution in R if and only if* $\operatorname{rank}_R(A) = \operatorname{rank}_R(A \mid \underline{b})$.

Let A be an $(m+1) \times m$ matrix with $Gr_R\{\mathcal{U}_m(A)\} \geqslant 2$ and let $\underline{b} = (b_1, \ldots, b_{m+1})^t \in R^{m+1}$. Then, by the above theorem, $\underline{b} \in \operatorname{Im}(\varphi_A)$ if and only if $\det(A \mid \underline{b}) = 0$ or equivalently

$$b_1 \Delta_1 + b_2 \Delta_2 + \cdots + b_m \Delta_m = 0,$$

where $\Delta_i = (-1)^i \det A_i$ and A_i is the $m \times m$ matrix obtained from A by deleting the i-th row. So one has the next result.

THEOREM 45. *Let A be an $(m+1) \times m$ matrix. Suppose that $Gr_R\{\mathcal{U}_m(A)\} \geqslant 2$ and let g be a nonzero divisor in R. Then the sequence*

$$0 \to R^{m+1} \xrightarrow{\varphi_A} R^m \xrightarrow{\varphi_\Delta} R \to R/(g)\mathcal{U}_m(A) \to 0,$$

is exact where $\Delta = (g\Delta_1, \ldots, g\Delta_m)$.

2.2.2. *General case*

DEFINITION 46. The R-module M has a finite free resolution of length t if there exists an exact sequence

$$0 \to F_t \xrightarrow{\varphi_t} F_{t-1} \xrightarrow{\varphi_{t-1}} \cdots \to F_1 \xrightarrow{\varphi_1} F_0 \to M \to 0$$

where each F_i is a finite free R-module.

EXAMPLE 47.

(i) Let R be a principal ideal domain and M a finitely generated R-module. Then M has a finite free resolution of length $\leqslant 1$.

(ii) Let $\mathfrak{a}$ be the ideal of R generated by a regular sequence $\{a_1, a_2, \ldots, a_n\}$. Then the Koszul complex associated to $\{a_1, a_2, \ldots, a_n\}$ gives a finite free resolution of $R/\mathfrak{a}$.

(iii) If M is a finitely generated projective module then M has a finite free resolution if and only if M is stably free.

(iv) An R-module M has a finite free resolution of length $t \geqslant 1$ if and only if M has a stably free projective resolution of length t.

(v) Let R be a regular local ring and M be a finitely generated R-module. Then M has a finite free resolution.

(vi) Let k be a field, R the polynomial ring $k[T_1, T_2, \ldots, T_n]$ and M a finitely generated R-module. Then M has a finite free resolution of length $\leqslant n$.

DEFINITION 48. Let

$$0 \to F_t \xrightarrow{\varphi_t} F_{t-1} \xrightarrow{\varphi_{t-1}} \cdots \to F_1 \xrightarrow{\varphi_1} F_0 \to M \to 0$$

be a finite free resolution of M. The non negative integer

$$\mathrm{Char}_R(M) = \sum_{i=1}^{t} (-1)^i \operatorname{rank}(F_i) = \operatorname{rank}(F_0) - \operatorname{rank}(\varphi_1),$$

is an invariant associated to M called the Euler characteristic of M.

If $\mathrm{Char}_R(M) = r$, then $\mathcal{F}_i(M) = 0$ for $i < r$ and $\mathcal{F}_r(M)$ is a nonzero faithful ideal. Other properties of the Euler characteristic can be found in [95, Chapters 3 and 4].

Let

$$F'' \xrightarrow{\varphi_B} F \xrightarrow{\varphi_A} F'$$

be a complex of finite free modules (i.e. $\varphi_A \varphi_B = 0$). A classical open problem is to give necessary and sufficient conditions for the exactness of this complex. McCoy's theorem solves the question when $F'' = 0$. In [49], using Theorem 35, the problem is solved when $R[T]$ is integrally closed and $\operatorname{rank}_R A = \operatorname{rank}_R F - 1$. In [23], for the Noetherian case, and [95, p. 193], for the general case, the following result is proved:

THEOREM 49. *Suppose that*

$$(\mathbb{C}): 0 \to F_t \xrightarrow{\varphi_t} F_{t-1} \xrightarrow{\varphi_{t-1}} \cdots \to F_1 \xrightarrow{\varphi_1} F_0,$$

where $t \geqslant 1$, is a complex of finite free R-modules. Put $\operatorname{rank}(F_i) = m_i$ *and* $\operatorname{rank}(\varphi_i) = r_i$. *Then the following statements are equivalent*:

(i) $(\mathbb{C})$ *is exact.*

(ii) *The following three conditions are satisfied*:

- $\operatorname{rank}(\varphi_t) = m_t$.
- $r_{i+1} + r_i = m_i$ *for* $1 \leqslant i < t$.
- $Gr_R\{\mathcal{U}_{r_i}(\varphi_i)\} \geqslant i$ *for* $i = 1, 2, \ldots, t$.

In [24], for the Noetherian case, and [95, p. 218], for the general case, there is proved the next result on the multiplicative structure of a finite free resolutions.

THEOREM 50. *Let*

$$(\mathbb{C}): 0 \to F_t \xrightarrow{\varphi_t} F_{t-1} \xrightarrow{\varphi_{t-1}} \cdots \to F_1 \xrightarrow{\varphi_1} F_0 \to M \to 0$$

be a finite free resolution of M. Put $\operatorname{rank}(F_i) = m_i$ *and* $\operatorname{rank}(\varphi_i) = r_i$. *Then there exist ideals $\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_n$, called the factorization ideals of* $(\mathbb{C})$, *such that*

(i) $\mathcal{B}_n = R$ *and* $\mathcal{B}_i \mathcal{B}_{i-1} = \mathcal{U}_{r_i}(\varphi_i)$ *for* $i = 1, 2, \ldots, t$.

(ii) $\mathcal{B}_0$ *can be generated by* $\binom{m_0}{r_1}$ *elements. In particular if* $\operatorname{Char}_R(M) = 0$ *then* $\mathcal{B}_0$ *is a principal ideal.*

(iii) $Gr_R\{\mathcal{B}_i\} \geqslant i + 1$ *for* $i = 0, 1, \ldots, t$.

(iv) $\mathcal{B}_i$ *and* $\mathcal{U}_{r_i}(\varphi_i)$ *have the same radical ideal for* $i = 2, \ldots, t$.

(v) *If* $\mathcal{B}_0$ *is a principal ideal then it is the* (*unique*) *smallest principal ideal containing* $\mathcal{U}_{r_1}(\varphi_1) = \mathcal{F}_{\operatorname{Char}_R(M)}(M)$.

The next result is the Hilbert–Burch theorem, see [26], and it is the converse of Theorem 45.

THEOREM 51. *Let $\mathfrak{a} \neq 0$ be an ideal of R having a finite free resolution of length one. Then there exists an $(m+1) \times m$ matrix A and a nonzero divisor g such that $a = (g)\mathcal{U}_m(A)$ and $Gr_R\{\mathcal{U}_m(A)\} \geqslant 2$. In this case $\mathcal{B}_0 = (g)$ and $\mathcal{B}_1 = \mathcal{U}_m(A)$.*

We return to the main discussion. The next result, see [61], solves the initial questions in terms of finite free resolutions.

THEOREM 52. *Let A be an $n \times m$ matrix and assume that there exists an exact sequence*

$$(\mathbb{C}): 0 \to F_t \xrightarrow{\varphi_t} F_{t-1} \xrightarrow{\varphi_{t-1}} \cdots \to F_2 \xrightarrow{\varphi_2} R^m \xrightarrow{\varphi_A} R^n,$$

where each F_i is a finitely generated free R-module. Suppose that $\mathcal{U}_r(A) = (g)\mathfrak{a}$ where $r = \operatorname{rank}(A)$, *$g$ is a nonzero divisor on R and $\mathfrak{a}$ is an ideal of R such that $Gr_R\{\mathfrak{a}\} \geqslant 2$. Then the following statements are equivalent*:

(i) $Gr_R\{\mathcal{U}_{r_i}(\varphi_i)\} \geqslant i+1$ *for* $i = 2, \ldots, t$.
(ii) *For* $\underline{b} \in R^n$ *the system of linear equations* (S): $A\underline{x} = \underline{b}$ *has a solution in* R *if and only if* $\operatorname{rank}(A \mid \underline{b}) = r$ *and* $\mathcal{U}_r(A \mid \underline{b}) \subseteq (g)$.

In particular, when $Gr_R\{\mathcal{U}_r(A)\} \geqslant 2$ *the following statements are equivalent*:
(i) $Gr_R\{\mathcal{U}_{r_i}(\varphi_i)\} \geqslant i+1$ *for* $i = 2, \ldots, t$.
(ii) *For* $\underline{b} \in R^n$ *the system of linear equations* (S): $A\underline{x} = \underline{b}$ *has a solution in* R *if and only if* $\operatorname{rank}(A \mid \underline{b}) = \operatorname{rank}(A)$.

Note that by Theorem 50 the factorization $\mathcal{U}_r(A) = (g)\mathfrak{a}$ holds when $\operatorname{rank}(A) = n$. It is not difficult to prove that if R is a unique factorization domain (UFD) and $\mathfrak{b}$ is a finitely generated ideal of R then $\mathfrak{b} = (g)\mathfrak{a}$ where $Gr_R\{\mathfrak{a}\} \geqslant 2$.

3. Characteristic polynomial of an endomorphism

Let M be an R-module and let $\varphi : M \to M$ be an endomorphism of M. The operation

$$T.m = \varphi(m),$$

converts M into an $R[T]$-module. We denote by M_φ this structure of M as $R[T]$-module via φ. If $\overline{M_\varphi}$ denotes the R-module obtained from M_φ by restriction of scalars, then $(\overline{M_\varphi})_{t_T} = M_\varphi$ where t_T is multiplication by T.

The characteristic sequence of φ is the exact sequence of $R[T]$-modules given by

$$0 \to M[T] \xrightarrow{t_T - \tilde{\varphi}} M[T] \xrightarrow{\phi_\varphi} M_\varphi \to 0$$

where

$$\phi_\varphi\bigl(\Sigma m_i T^i\bigr) = \Sigma \varphi^i(m_i),$$

and $\tilde{\varphi}$ denotes the $R[T]$-endomorphism of $M[T]$ obtained from φ by extension of scalars (i.e. $\tilde{\varphi}(\Sigma m_i T^i) = \Sigma \varphi(m_i) T^i$).

Let $f : R \to R'$ be a ring homomorphism. Comparing the characteristic sequence of $\varphi_{(R')}$ with the extension to $R'[T]$ of the characteristic sequence of φ it follows that $(M_{(R')})_{\varphi_{(R')}}$ is isomorphic to $(M_\varphi)_{(R'[T])}$.

We denote by $R[\varphi]$ the commutative ring

$$R[\varphi] = \left\{ \sum_{\text{finite}} a_i \varphi^i :\ a_i \in R \right\}.$$

The kernel of the substitution homomorphism

$$\begin{aligned} R[X] &\to R[\varphi] \\ X &\mapsto \varphi \end{aligned}$$

is called the ideal of relations satisfied by φ, and it is denoted by $I(\varphi)$ (i.e. $I(\varphi) = \{p(T)\colon\ p(\varphi)=0\}$).

Recall that two endomorphisms φ and ψ of M are similar if there exists an isomorphism τ such that $\psi = \tau\varphi\tau^{-1}$. It is clear that if φ and ψ are similar then M_φ is isomorphic to M_ψ (as $R[T]$-modules) and hence $\mathcal{F}_i(M_\varphi) = \mathcal{F}_i(M_\psi)$ for all i. Consequently $\{\mathcal{F}_i(F_\varphi)\}_{i\geqslant 0}$ is a set of invariants for the similarity class of φ.

3.1. *Case of a free module*

Let F be a finitely generated free R-module and $\varphi\colon F \to F$ an endomorphism of F. We denote by $\chi(\varphi, T)$ the characteristic polynomial of φ (i.e. $\chi(\varphi, T) = \det(t_T - \tilde{\varphi})$). Considering the characteristic sequence

$$0 \to F[T] \xrightarrow{t_T - \tilde{\varphi}} F[T] \xrightarrow{\phi_\varphi} F_\varphi \to 0$$

it follows that the $R[T]$-module F_φ has a finite free resolution of length one and

$$\mathcal{F}_0(F_\varphi) = \big(\chi(\varphi, T)\big).$$

Moreover $\chi(\varphi, T)$ is the unique monic generator of $\mathcal{F}_0(F_\varphi)$. In general, $\mathcal{F}_i(F_\varphi)$ is not principal for $i \geqslant 1$.

The next result contains the basic properties of the characteristic polynomial of an endomorphism of a finitely generated free module.

PROPOSITION 53. *Let F be a finitely generated free R-module of rank n and $\varphi\colon F \to F$ an endomorphism of F.*

(i) *If $f\colon R \to R'$ is a ring homomorphism then $\chi(\varphi_{(R')}, T) = f(\chi(\varphi, T))$.*

(ii) *If F' and F'' are free R-modules and*

$$\begin{tikzcd} 0 \arrow[r] & F' \arrow[r] \arrow[d, "\varphi'"] & F \arrow[r] \arrow[d, "\varphi"] & F'' \arrow[r] \arrow[d, "\varphi''"] & 0 \\ 0 \arrow[r] & F' \arrow[r] & F \arrow[r] & F'' \arrow[r] & 0 \end{tikzcd}$$

is a commutative diagram with exact rows, then

$$\chi(\varphi, T) = \chi(\varphi', T)\chi(\varphi'', T).$$

(iii) *If ψ is another endomorphism of F then $\chi(\varphi\psi, T) = \chi(\psi\varphi, T)$.*

(iv) *If $\varphi^*\colon F^* \to F^*$ is the dual of φ, then $\chi(\varphi, T) = \chi(\varphi^*, T)$.*

(v) $\chi(\varphi, T) = T^n - Tr(\wedge^1(\varphi))T^{n-1} + \cdots + (-1)^n Tr(\wedge^n(\varphi))$, *where $\wedge^i(\varphi)$ is the i-th exterior power of φ and $Tr(\wedge^i(\varphi))$ is the trace of $\wedge^i(\varphi)$.*

(vi) (Cayley–Hamilton theorem) *The endomorphism φ satisfies its characteristic polynomial, i.e. $\chi(\varphi, \varphi) = 0$.*

(vii) (McCoy's theorem) $I(\varphi) = (\chi(\varphi, T) :_{R[T]} \mathcal{F}_1(F_\varphi))$.

Suppose that R is a field. Then $R[T]$ is a principal ideal domain and hence F_φ is completely determined by the set $\{\mathcal{F}_i(F_\varphi)\}_{i\geqslant 0}$. The i-th-invariant factor of φ, denoted by $d_i(\varphi, T)$, is the unique monic generator of the ideal $(\mathcal{F}_{n-i}(F_\varphi):\mathcal{F}_{n-i+1}(F_\varphi))$ where $n = \operatorname{rk} F$, see Section 1.3. Moreover if $\chi_i(\varphi, T)$ is the characteristic polynomial of

$$t_T : \overline{\wedge^i F_\varphi} \to \overline{\wedge^i F_\varphi}$$

then $\{\chi_i(\varphi, T)\}_{i\geqslant 0}$ is also a complete set of invariants for the similarity class of φ.

When R is an arbitrary commutative ring the set $\{\mathcal{F}_i(F_\varphi)\}_{i\geqslant 0}$ is not, in general, a complete set of invariants for the similarity class of φ. In [89, V.D.14] can be found a brief exposition, based on [94,53,54], on similarity of matrices when R is a local ring or R is the ring of integers.

3.2. *Case of a projective module*

3.2.1. *Trace and determinant* Let M be an R-module and M^* the dual module of M. The natural bilinear mapping

$$\begin{aligned} M^* \times M &\to \operatorname{Hom}_R(M, M) = \operatorname{End}_R(M) \\ (u, m) &\mapsto \varphi_{(u,m)} \end{aligned}$$

where $\varphi_{(u,m)}(m') = u(m')m$, induces a canonical homomorphism

$$\theta_M : M^* \otimes_R M \to \operatorname{End}_R(M).$$

PROPOSITION 54. *Let M be a finitely generated projective R-module. Then θ_M is an isomorphism.*

REMARK 55. In fact, see [3, II, p. 77], one has that M is projective if and only if $Id_M \in \operatorname{Im}\theta_M$ where Id_M denotes the identity endomorphism of M. In this case there exist elements $m_1, m_2, \ldots, m_t$ in M and $u_1, u_2, \ldots, u_t$ in M^* such that

$$m = \sum_{i=1}^{t} u_i(m) m_i,$$

for each $m \in M$. Hence

$$\theta_M\left(\sum_{i=1}^{t} u_i\varphi \otimes m_i\right) = \varphi.$$

The sets $\{m_i\}_{1\leqslant i\leqslant t}$ and $\{u_i\}_{1\leqslant i\leqslant t}$ are called dual bases for M. Note that M is projective if and only if dual bases can be constructed from every system of generators of M.

The bilinear mapping

$$\begin{aligned} M^* \times M &\to R \\ (u, m) &\mapsto u(m) \end{aligned}$$

induces a canonical homomorphism

$$\tau_M : M^* \otimes_R M \to R.$$

DEFINITION 56. Let M be a finitely generated projective R-module. The composition homomorphism

$$\mathrm{End}_R(M) \xrightarrow{\theta_M^{-1}} M^* \otimes_R M \xrightarrow{\tau_M} R,$$

is called the trace map and it is denoted by Tr.

PROPOSITION 57. *Let M be a finitely generated projective R-module and φ an endomorphism of M.*

(i) *If $f : R \to R'$ is a ring homomorphism then $Tr(\varphi_{(R')}) = f(Tr(\varphi))$.*

(ii) *If M' and M'' are finitely generated projective R-modules and*

$$\begin{tikzcd} 0 \arrow[r] & M' \arrow[r] \arrow[d, "\varphi'"] & M \arrow[r] \arrow[d, "\varphi"] & M'' \arrow[r] \arrow[d, "\varphi''"] & 0 \\ 0 \arrow[r] & M' \arrow[r] & M \arrow[r] & M'' \arrow[r] & 0 \end{tikzcd}$$

is a commutative diagram with exact rows, then $Tr(\varphi) = Tr(\varphi') + Tr(\varphi'')$.

(iii) *If $\varphi^* : M^* \to M^*$ is the dual of φ then $Tr(\wedge^i(\varphi)) = Tr(\wedge^i(\varphi^*))$ for $i \geqslant 0$.*

(iv) *If ψ is another endomorphism of M then $Tr(\varphi\psi) = Tr(\psi\varphi)$.*

(v) *If $\{m_i\}_{1 \leqslant i \leqslant t}$, $\{u_i\}_{1 \leqslant i \leqslant t}$ constitute dual bases for M then*

$$Tr(\varphi) = \sum_{i=1}^{t} u_i\big(\varphi(m_i)\big).$$

(vi) *$Tr(\varphi) = 0$ if and only if there exist endomorphisms $\{\alpha_i\}_{1 \leqslant i \leqslant s}$ and $\{\beta_i\}_{1 \leqslant i \leqslant s}$ such that $\varphi = \sum(\alpha_i\beta_i - \beta_i\alpha_i)$.*

(vii) *If $\{e_i\}_{1 \leqslant i \leqslant s}$ is the sequence of idempotent elements constructed in Remark* 24 *then $Tr(Id_M) = e_1 + 2e_2 + \cdots + se_s$.*

Let M be a finitely generated projective R-module and φ an endomorphism of M. Suppose that Q (resp. Q') is a finitely generated projective R-module such that $F = M \oplus Q$ (resp. $F = M \oplus Q'$) is free and let ψ (resp. ψ') be the endomorphism of F defined by

$$\psi = \left(\begin{array}{c|c} \varphi & 0 \\ \hline 0 & Id_Q \end{array}\right) \quad \text{(resp. } \psi' = \left(\begin{array}{c|c} \varphi & 0 \\ \hline 0 & Id_{Q'} \end{array}\right)\text{)}.$$

Then $\det(\psi) = \det(\psi')$ because this equality holds in $R_\mathfrak{p}$ for all prime ideals $\mathfrak{p}$ of R. Consequently we can introduce the following definition, see [51].

DEFINITION 58. We say that $\det(\psi)$ is the determinant of φ.

REMARK 59. In [2] the determinant of φ is defined by

$$\det(\varphi) = \lambda(\varphi - Id_M, 1),$$

where $\lambda(\varphi, T)$ is the determinant-trace polynomial of φ (i.e.

$$\lambda(\varphi, T) = 1 + Tr\bigl(\wedge^1(\varphi)\bigr)T + Tr\bigl(\wedge^2(\varphi)\bigr)T^2 + \cdots + Tr\bigl(\wedge^n(\varphi)\bigr)T^n,$$

where n satisfies $\wedge^n(\varphi) \neq 0$ and $\wedge^{n+1}(\varphi) = 0$). By localization arguments both definitions are equivalent.

PROPOSITION 60. *Let M be a finitely generated projective R-module and φ an endomorphism of M.*

(i) *If $f : R \to R'$ is a ring homomorphism then* $\det(\varphi_{(R')}) = f(\det(\varphi))$.

(ii) *If M' and M'' are finitely generated projective R-modules and*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & M' & \longrightarrow & M & \longrightarrow & M'' & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \varphi'} & & \downarrow{\scriptstyle \varphi} & & \downarrow{\scriptstyle \varphi''} & & \\ 0 & \longrightarrow & M' & \longrightarrow & M & \longrightarrow & M'' & \longrightarrow & 0 \end{array}$$

is a commutative diagram with exact rows, then $\det(\varphi) = \det(\varphi')\det(\varphi'')$.

(iii) *If $\varphi^* : F^* \to F^*$ is the dual of φ then* $\det(\varphi) = \det(\varphi^*)$.

(iv) *If $\{m_i\}_{1\leqslant i\leqslant t}$, $\{u_i\}_{1\leqslant i\leqslant t}$ constitute dual bases for M then* $\det(\varphi) = \det(A)$ *where A is the $t \times t$ matrix $A = (u_i(\varphi(m_j)))$.*

(v) *φ is surjective if and only if* $\det(\varphi)$ *is a unit.*

(vi) *φ is injective if and only if* $\det(\varphi)$ *is a nonzero divisor.*

3.2.2. *Characteristic polynomial*

DEFINITION 61. Let M be a finitely generated projective R-module and φ an endomorphism of M. The (Goldman) characteristic polynomial $\chi(\varphi, T)$ of φ is defined by

$$\chi(\varphi, T) = \det(t_T - \tilde{\varphi}).$$

PROPOSITION 62. *Let M be a finitely generated projective R-module and φ an endomorphism of M.*

(i) *If $f : R \to R'$ is a ring homomorphism then $\chi(\varphi_{(R')}, T) = f(\chi(\varphi, T))$.*

(ii) *If M' and M'' are finitely generated projective R-modules and*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & M' & \longrightarrow & M & \longrightarrow & M'' & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \varphi'} & & \downarrow{\scriptstyle \varphi} & & \downarrow{\scriptstyle \varphi''} & & \\ 0 & \longrightarrow & M' & \longrightarrow & M & \longrightarrow & M'' & \longrightarrow & 0 \end{array}$$

is a commutative diagram with exact rows, then $\chi(\varphi, T) = \chi(\varphi', T)\chi(\varphi'', T)$.

(iii) *If* $\varphi^* : F^* \to F^*$ *is the dual of* φ, *then* $\chi(\varphi, T) = \chi(\varphi^*, T)$.

(iv) (Cayley–Hamilton theorem) *The endomorphism* φ *satisfies its characteristic polynomial, i.e.* $\chi(\varphi, \varphi) = 0$.

(v) *The ideals* $I(\varphi)$ *and* $(\chi(\varphi, T))$ *have the same radical ideal.*

(vi) *If* $\{m_i\}_{1\leqslant i\leqslant t}$, $\{u_i\}_{1\leqslant i\leqslant t}$ *constitute dual bases for* M *then* $\chi(\varphi, T) = \chi(\varphi_A, T)$ *where* A *is the* $t \times t$ *matrix* $A = (u_i(\varphi(m_j)))$.

(vii) *If* $\{e_i\}_{1\leqslant i\leqslant t}$ *is the sequence of idempotent elements constructed in Remark* 24 *then* $\chi(0, T) = e_0 + e_1 T + \cdots + e_s T^s$ *where* 0 *is the zero homorphism.*

3.3. *Case of a module that has a finite free resolution*

3.3.1. *MacRae's invariant* We say that an R-module E is an elementary module if there exists an exact sequence of the form

$$0 \to F \to F \to E \to 0,$$

where F is a finitely generated free R-module (i.e. E possesses a finite free resolution of length one and its Euler characteristic is zero). In this case $\mathcal{F}_0(E)$ is a principal ideal generated by a nonzero divisor.

A $\mathbb{Z}$-module E is elementary if and only if E is a finite group, in this case the invariant $\mathcal{F}_0(E)$ is equivalent to the order of E. If F is a finitely generated free R-module and φ is an endomorphism of F then F_φ is an elementary $R[T]$-module, in this case $\mathcal{F}_0(E) = (\chi(\varphi, T))$.

DEFINITION 63. An R-module M possesses an elementary resolution of length n if there exists an exact sequence

$$0 \to E_n \to E_{n-1} \to \cdots \to E_1 \to E_0 \to M \to 0$$

in which each E_i is an elementary module.

REMARK 64. In [95, p. 81] it is proved that M admits an elementary resolution of finite length if and only if M has a finite free resolution and $\text{Ann}(M)$ contains a nonzero divisor.

Let

$$0 \to E_n \to E_{n-1} \to \cdots \to E_1 \to E_0 \to M \to 0$$

be an elementary resolution of M. In the total quotient ring $\mathcal{T}(R)$ of R we consider the fractional ideal

$$\mathcal{G}(M) = \mathcal{F}_0(E_0)\mathcal{F}_0(E_1)^{-1}\mathcal{F}_0(E_2)\mathcal{F}_0(E_3)^{-1}\cdots.$$

Since $\mathcal{F}_0(E_i)$ is a principal ideal generated by a nonzero divisor then

$$\mathcal{G}(M) = \frac{\alpha}{\beta}R,$$

where α and β are nonzero divisors in R.

THEOREM 65. *Suppose that M possesses an elementary resolution. Then*:

(i) *$\mathcal{G}(M)$ is an invariant associated to M (i.e. it is independent of the finite elementary resolution of M chosen).*

(ii) *$\mathcal{G}(M)$ is a principal ideal of R generated by a nonzero divisor and $\mathcal{F}_0(M) \subseteq \mathcal{G}(M)$. Further $\mathcal{G}(M)$ is the smallest principal ideal containing $\mathcal{F}_0(M)$.*

PROOF. See [87]. □

DEFINITION 66. The ideal $\mathcal{G}(M)$ is called the MacRae invariant of M.

PROPOSITION 67. *Suppose that M possesses an elementary resolution. Then the following statements are satisfied*:

(i) *If $0 \to M' \to M \to M'' \to 0$ is an exact sequence of R-modules that have finite length elementary resolutions, then $\mathcal{G}(M) = \mathcal{G}(M')\mathcal{G}(M'')$.*

(ii) *Let $f : R \to R'$ be a ring homomorphism. If $M_{(R')}$ possesses an elementary resolution, then $\mathcal{G}(M_{(R')}) \subseteq \mathcal{G}(M)R'$.*

(iii) *If $\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_t$ are the factorization ideals of a finite free resolution of M then $\mathcal{B}_0 = \mathcal{G}(M)$.*

PROOF. See [87] and [95, p. 225]. □

3.3.2. *MacRae's invariant and the characteristic polynomial* Let

$$0 \to F_t \xrightarrow{u_t} F_{t-1} \xrightarrow{u_{t-1}} \cdots \to F_1 \xrightarrow{u_1} F_0 \xrightarrow{u_0} M \to 0$$

be a finite free resolution of M and φ an endomorphism of M. Then there exist endomorphisms $\varphi_i : F_i \to F_i$ that make the diagram

$$\begin{tikzcd}
0 \arrow[r] & F_t \arrow[r, "u_t"] \arrow[d, "\varphi_t"] & F_{t-1} \arrow[r, "u_{t-1}"] \arrow[d, "\varphi_{t-1}"] & \cdots \arrow[r] & F_1 \arrow[r, "u_1"] \arrow[d, "\varphi_1"] & F_0 \arrow[r, "u_0"] \arrow[d, "\varphi_0"] & M \arrow[r] \arrow[d, "\varphi"] & 0 \\
0 \arrow[r] & F_t \arrow[r, "u_t"] & F_{t-1} \arrow[r, "u_{t-1}"] & \cdots \arrow[r] & F_1 \arrow[r, "u_1"] & F_0 \arrow[r, "u_0"] & M \arrow[r] & 0
\end{tikzcd}$$

commutative. Considering the associated characteristic sequences it follows that the diagram of $R[T]$-modules

$$\begin{tikzcd}
 & 0 \arrow[d] & & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & \\
0 \arrow[r] & F_t[T] \arrow[r, "\tilde{u}_t"] \arrow[d, "\omega_{\varphi_t}"] & \cdots \arrow[r] & F_1[T] \arrow[r, "\tilde{u}_1"] \arrow[d, "\omega_{\varphi_1}"] & F_0[T] \arrow[r, "\tilde{u}_0"] \arrow[d, "\omega_{\varphi_0}"] & M[T] \arrow[r] \arrow[d, "\omega_{\varphi}"] & 0 \\
0 \arrow[r] & F_t[T] \arrow[r, "\tilde{u}_t"] \arrow[d, "\phi_{\varphi_t}"] & \cdots \arrow[r] & F_1[T] \arrow[r, "\tilde{u}_1"] \arrow[d, "\phi_{\varphi_1}"] & F_0[T] \arrow[r, "\tilde{u}_0"] \arrow[d, "\phi_{\varphi_0}"] & M[T] \arrow[r] \arrow[d, "\phi_{\varphi}"] & 0 \\
0 \arrow[r] & (F_t)_{\varphi_t} \arrow[r, "u_t"] \arrow[d] & \cdots \arrow[r] & (F_1)_{\varphi_1} \arrow[r, "u_1"] \arrow[d] & (F_0)_{\varphi_0} \arrow[r, "u_0"] \arrow[d] & M_{\varphi} \arrow[r] \arrow[d] & 0 \\
 & 0 & & 0 & 0 & 0 &
\end{tikzcd}$$

is commutative with exact rows and columns, where $\omega_{\varphi} = t_T - \tilde{\varphi}$ and $\omega_{\varphi_i} = t_T - \tilde{\varphi}_i$.

THEOREM 68. *Let*

$$0 \to F_t \xrightarrow{u_t} F_{t-1} \xrightarrow{u_{t-1}} \cdots \to F_1 \xrightarrow{u_1} F_0 \xrightarrow{u_0} M \to 0$$

be a finite free resolution of M and φ an endomorphism of M. Then:

(i) *The exact sequence of $R[T]$-modules*

$$0 \to (F_t)_{\varphi_t} \xrightarrow{u_t} (F_{t-1})_{\varphi_{t-1}} \to \cdots \to (F_1)_{\varphi_1} \xrightarrow{u_1} (F_0)_{\varphi_0} \xrightarrow{u_0} M_{\varphi} \to 0$$

is an elementary resolution of M_{φ}.

(ii) *The MacRae invariant $\mathcal{G}(M_{\varphi})$ is generated by a monic polynomial of degree* $\mathrm{Char}_R(M)$.

Now we can introduce the following definition, see [64].

DEFINITION 69. The unique monic generator of $\mathcal{G}(M_{\varphi})$ is called the characteristic polynomial of φ and it is denoted by $\chi(\varphi, T)$.

REMARK 70. By construction $\chi(\varphi, T)$ is an invariant associated to φ and M. When M is free the above definition gives the classic characteristic polynomial.

Next we collect the main properties of $\chi(\varphi, T)$, see [64,66].

THEOREM 71. *Let M be an R-module that has a finite free resolution with Euler characteristic n, and let φ be an endomorphism of M.*

(i) *If M' and M'' are R-modules that have finite free resolutions and*

$$\begin{tikzcd} 0 \arrow[r] & M' \arrow[r] \arrow[d, "\varphi'"] & M \arrow[r] \arrow[d, "\varphi"] & M'' \arrow[r] \arrow[d, "\varphi''"] & 0 \\ 0 \arrow[r] & M' \arrow[r] & M \arrow[r] & M'' \arrow[r] & 0 \end{tikzcd}$$

is a commutative diagram with exact rows, then

$$\chi(\varphi, T) = \chi(\varphi', T)\chi(\varphi'', T).$$

(ii) *If ψ is another endomorphism of M then $\chi(\varphi\psi, T) = \chi(\psi\varphi, T)$.*

(iii) *If $f : R \to R'$ is a homomorphism of rings with unit and $M_{(R')}$ has a finite free resolution then $f(\chi(\varphi, T))$ divides $\chi(\varphi_{(R')}, T)$ and the equality holds if and only if* $\mathrm{Char}_{R'}(M_{(R')}) = n$.

(iv) (Generalized Cayley–Hamilton theorem) *For every $\Delta \in \mathcal{F}_n(M)$ one has*

$$\Delta\chi(\varphi, \varphi) = 0.$$

In particular the classical Cayley–Hamilton theorem holds when

$$\big(0 :_M \mathcal{F}_n(M)\big) = 0.$$

(v) (Generalized McCoy theorem)

$$\mathcal{F}_n(M)\big(\chi(\varphi, T) :_{R[T]} \mathcal{F}_1(M_\varphi)\big) \subseteq I(\varphi) \subseteq \big(\chi(\varphi, T) :_{R[T]} \mathcal{F}_1(M_\varphi)\big).$$

Moreover if M is torsion free (*i.e. for $m \in M$ and for a finitely generated faithful ideal $\mathfrak{a}$ of R the equality $\mathfrak{a}m = 0$ implies $m = 0$*) *then*

$$I(\varphi) = \big(\chi(\varphi, T) :_{R[T]} \mathcal{F}_1(M_\varphi)\big)$$

(*i.e. the classical McCoy theorem holds*).

REMARK 72. Let R be a reduced ring, M an R-module that has a finite free resolution with Euler characteristic n, and φ an endomorphism of M. In [64], under the assumption that $\overline{\wedge^i F_\varphi}$ has a finite free resolution for all i, there is constructed a sequence of polynomials $\{d_i(\varphi, T)\}_{1 \leqslant i \leqslant s}$, called invariant factors of φ, verifying:

(i) $d_1(\varphi, T)d_2(\varphi, T)\dots d_s(\varphi, T) = \chi(\varphi, T)$.

(ii) $d_i(\varphi, T)$ divides $d_{i+1}(\varphi, T)$ for $1 \leqslant i \leqslant s-1$.

(iii) $(\prod_{j=1}^{s-i} d_j(\varphi, T))$ is the smallest principal ideal of $R[T]$ generated by a monic polynomial that contains $\mathcal{F}_i(M_\varphi)$ for $0 \leqslant i \leqslant s-1$.

(iv) When R is a domain there exists a free submodule F of M, invariant with respect to φ, such that the restriction of φ to F is similar to the endomorphism of F defined by the diagonal block matrix

$$\begin{pmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_s \end{pmatrix}$$

where D_i is the companion matrix of $d_i(\varphi, T)$.

3.3.3. *Trace and determinant* Let M be an R-module that has a finite free resolution and φ an endomorphism of M. Consider the commutative diagram

$$\begin{tikzcd} 0 \arrow[r] & F_t \arrow[r, "u_t"] \arrow[d, "\varphi_t"] & F_{t-1} \arrow[r, "u_{t-1}"] \arrow[d, "\varphi_{t-1}"] & \cdots \arrow[r] & F_1 \arrow[r, "u_1"] \arrow[d, "\varphi_1"] & F_0 \arrow[r, "u_0"] \arrow[d, "\varphi_0"] & M \arrow[r] \arrow[d, "\varphi"] & 0 \\ 0 \arrow[r] & F_t \arrow[r, "u_t"] & F_{t-1} \arrow[r, "u_{t-1}"] & \cdots \arrow[r] & F_1 \arrow[r, "u_1"] & F_0 \arrow[r, "u_0"] & M \arrow[r] & 0 \end{tikzcd}$$

In [2] the trace of φ is defined as

$$Tr(\varphi) = \sum_{i=0}^{t} (-1)^i Tr(\varphi_i)$$

(in fact, the definition is given under the assumption that M has a finite projective resolution). Moreover it is proved that this definition possesses most of the basic properties of the trace map for the free case.

To define determinant of φ as

$$\det(\varphi) = \prod_{i=0}^{t} \det(\varphi_i)^{(-1)^i}$$

causes problems because either $\det(\varphi)$ cannot be defined (for example, if $\det(\varphi_i)$ is a zero divisor for some odd i) or it can happen that $\det(\varphi)$ cannot be an element of R. In [96] the following definition of determinant is introduced.

DEFINITION 73. Let M be an R-module that has a finite free resolution and φ an endomorphism of M. The determinant of φ is the element of R given by

$$\det(\varphi) = (-1)^n \chi(\varphi, 0),$$

where $\mathrm{Char}_R(M) = n$.

By Theorem 71 one has the classical basic properties of det. The next result, see [96], relates the character (surjective or injective) of φ and the character (unit or nonzero divisor) of $\det(\varphi)$.

THEOREM 74. *Let M be an R-module that has a finite free resolution, n the Euler characteristic of M and φ an endomorphism of M.*

(i) *If φ is surjective then* $\det(\varphi)$ *is a unit.*

(ii) *If* $\det(\varphi)$ *is a unit then the induced homomorphisms $\varphi : \Delta M \to \Delta M$ and $\overline{\varphi} : M/t(M) \to M/t(M)$ are surjective, where $\Delta \in \mathcal{F}_r(M)$ and $t(M)$ is the submodule of M formed by all the torsion elements of M.*

(iii) *If $n > 0$ and φ is injective then* $\det(\varphi)$ *is a nonzero divisor on R. The converse is true when $t(M) = 0$.*

(iv) *Suppose that $\chi(\varphi, \varphi) = 0$ (i.e. φ verifies the classical Cayley–Hamilton theorem). Then:*

- *φ is surjective if and only if* $\det(\varphi)$ *is a unit of R.*
- $\det(\varphi)$ *is a nonzero divisor on M if and only if*

$$\operatorname{Ker}(\varphi) = \operatorname{Ker}\bigl(\operatorname{Adj}(\varphi)\bigr) = 0,$$

where $\operatorname{Adj}(\varphi)$ *is the adjoint endomorphism of φ (i.e.* $\operatorname{Adj}(\varphi) = (-1)^{n+1}(\varphi^{n-1} + a_{n-1}\varphi^{n-2} + \cdots + a_1 Id_M)$ *when $\chi(\varphi, T) = T^n + a_{n-1}T^{n-1} + \cdots + a_1 T + a_0$).*

4. Feedback classification problem

DEFINITION 75. An m-input n-dimensional linear dynamical system Σ over R is a pair of matrices (A, B), where $A = (a_{ij})$ is an $n \times n$ matrix and $B = (b_{ij})$ is an $n \times m$ matrix with entries in R. When B consists of a single column vector $\underline{b} = (b_1,\ b_2, \ldots, b_n)^t$ the system $\Sigma = (A, \underline{b})$ is called a single-input system.

REMARK 76. In Section 5 we need to work with linear systems over projective modules. An m-input linear dynamical system Σ, over a finitely generated projective R-module. M, is a pair (φ, ψ) where φ is an endomorphism of M and $\psi : R^m \to M$ is an homomorphism.

Let $f : R \to R'$ be a ring homomorphism and $\Sigma = (A, B)$ a linear system over R (resp. $\Sigma = (\varphi, \psi)$ a linear system over a projective R-module M). The extension of Σ to R' is the linear system $f(\Sigma) = (f(A), f(B))$ (resp. $f(\Sigma) = (\varphi_{(R')}, \psi_{(R')})$). The natural extensions to $R_{\mathfrak{p}}$ and $k(\mathfrak{p}) = R_{\mathfrak{p}}/\mathfrak{p}R_{\mathfrak{p}}$ are denoted by $\Sigma_{\mathfrak{p}}$ and $\Sigma(\mathfrak{p})$, respectively.

DEFINITION 77. The feedback group $\mathbf{F}_{n,m}$ is the group, acting on m-input, n-dimensional linear dynamical systems $\Sigma = (A, B)$ generated by the following three types of transformations:

(i) $A \mapsto A' = PAP^{-1}$; $B \mapsto B' = PB$ for some invertible matrix P. This transformation is consequence of a change of basis in R^n.

(ii) $A \mapsto A' = A$; $B \mapsto B' = BQ$ for some invertible matrix Q. This transformation is consequence of a change of basis in R^m.

(iii) $A \mapsto A' = A + BF$; $B \mapsto B' = B$ for some $m \times n$ matrix F, which is called a feedback matrix.

The system Σ' is feedback equivalent to Σ if it is obtained from Σ by means of an element of $\mathbf{F}_{n,m}$.

From the definition it follows that $\Sigma' = (A', B')$ is feedback equivalent to $\Sigma = (A, B)$ if and only if there exist an invertible $n \times n$ matrix P, an invertible $m \times m$ matrix Q and an $m \times n$ matrix K such that

$$\begin{cases} A' = P(A + BF)P^{-1}, \\ B' = PBQ. \end{cases}$$

In particular, $\Sigma = (A, 0)$ is feedback equivalent to $\Sigma' = (A', 0)$ if and only if A and A' are similar matrices and $\Sigma = (0, B)$ is feedback equivalent to $\Sigma' = (0, B')$ if and only if B and B' are equivalent matrices.

REMARK 78. Let F and G be two $n \times l$ matrices over R. The matrix $TF + G$ over the polynomial ring $R[T]$ is called the pencil of F and G. Two pencils $TF + G$ and $TF' + G'$ are Kronecker equivalent if there exist invertible constant matrices P and Q such that

$$TF' + G' = P(TF + G)Q.$$

Given a linear system $\Sigma = (A, B)$, consider the pencil

$$T(Id, 0) + (-A, B) = (TId - A, B)$$

of the block matrices $(Id, 0)$ and $(-A, B)$. Then Σ is feedback equivalent to Σ' if and only if the corresponding pencils are Kronecker equivalent.

REMARK 79. Two linear systems $\Sigma = (\varphi, \psi)$ and $\Sigma' = (\varphi', \psi')$ over a finitely generated projective R-module M are feedback equivalent if and only if there exist an automorphism τ of P, an automorphism ϖ of R^m and an homomorphism $\phi : P \to R^m$ such that

$$\begin{cases} \varphi' = \tau(\varphi + \psi\phi)\tau^{-1}, \\ \psi' = \tau\psi\varpi. \end{cases}$$

The feedback classification problem is: Given a linear system Σ over R obtain a complete set of invariants, and a canonical form, that characterize the feedback equivalence class of Σ (i.e. Σ' and Σ yield the same invariants if and only if they are feedback equivalent).

The feedback classification problem is what is known as a wild problem and is open in the general case. However in some cases it is possible to obtain a solution.

4.1. *Brunovsky systems. Classical case*

DEFINITION 80. An n-dimensional linear system $\Sigma = (A, B)$ over R is reachable if the columns of the block matrix (called the reachability matrix of Σ)

$$A * B = (B, AB, \ldots, A^{n-1}B)$$

generate R^n.

By Proposition 3, Σ is reachable if and only if $\mathcal{U}_n(A * B) = R$ or equivalently if and only if $\Sigma(\mathfrak{m})$ is reachable for all maximal ideals $\mathfrak{m}$ of R. Moreover in this case $\mathcal{U}_1(B) = R$ (i.e. $\operatorname{Im}\varphi_B$ is a basic submodule of R^n, see Section 1.1).

REMARK 81. A linear system $\Sigma = (\varphi, \psi)$ over a finitely generated projective R-module M is reachable if and only if $\Sigma(\mathfrak{m})$ is reachable for all maximal ideals $\mathfrak{m}$ of R. Moreover in this case $\operatorname{Im}\psi$ is a basic submodule of M (i.e. $\dim_{k(\mathfrak{p})}(M(\mathfrak{p})) > \dim_{k(\mathfrak{p})}((M/\operatorname{Im}\varphi_A) \otimes_R k(\mathfrak{p}))$).

Let $\Sigma = (A, \underline{b})$ be a reachable single-input n-dimensional system. The set $\{\underline{b}, A\underline{b}, \ldots, A^{n-1}\underline{b}\}$ is a basis of R^n and

$$(\underline{b}, A\underline{b}, \ldots, A^{n-1}\underline{b})^{-1} A (\underline{b}, A\underline{b}, \ldots, A^{n-1}\underline{b}) = C(A),$$

where $C(A)$ is the companion matrix of A. Put $P = (\hat{\underline{b}}, c(A)\hat{\underline{b}}, \ldots, c(A)^{n-1}\hat{\underline{b}})$ where $c(A)$ is the transpose matrix of $C(A)$ and $\hat{\underline{b}} = (0, 0, \ldots, 1)^t$. Then

$$P(\underline{b}, A\underline{b}, \ldots, A^{n-1}\underline{b})^{-1} A (\underline{b}, A\underline{b}, \ldots, A^{n-1}\underline{b}) P^{-1} = c(A).$$

Taking an adequate feedback matrix one has the following result.

PROPOSITION 82. *Let $\Sigma = (A, \underline{b})$ be a reachable single-input n-dimensional system. Then Σ is feedback equivalent to the system*

$$\widehat{\Sigma} = \left(\widehat{A} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \ \hat{\underline{b}} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \right).$$

In particular, there exists a unique class of reachable single-input n-dimensional systems.

The next objective is to study the multi-input case ($m > 1$). We introduce the following class of reachable linear systems.

DEFINITION 83. Let $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_s)$ be a partition of the integer n (i.e. $\kappa_1, \kappa_2, \ldots, \kappa_s$ are positive integers such that $\kappa_1 \geqslant \kappa_2 \geqslant \cdots \geqslant \kappa_s$ and $\kappa_1 + \kappa_2 + \cdots + \kappa_s = n$) and let m be an integer with $m \geqslant s$. The Brunovsky linear form associated to κ and m is the reachable m-input, n-dimensional linear dynamical system $\Sigma_\kappa = (A_\kappa, B_\kappa)$ given by

$$A_\kappa = \begin{pmatrix} A_{\kappa_1} & 0 & \cdots & 0 \\ 0 & A_{\kappa_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & & A_{\kappa_s} \end{pmatrix},$$

$$B_\kappa = \begin{pmatrix} \underline{b}_{\kappa_1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \underline{b}_{\kappa_2} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \underline{b}_{\kappa_s} & 0 & \cdots & 0 \end{pmatrix},$$

where each $\Sigma_{\kappa_i} = (A_{\kappa_i}, \underline{b}_{\kappa_i})$ is the single input κ_i-dimensional canonical form. The integers $\{\kappa_i\}_{1 \leqslant i \leqslant s}$ are called the Kronecker indices of Σ_κ.

DEFINITION 84. We say that $\Sigma = (A, B)$ is a Brunovsky linear system if and only if Σ is feedback equivalent to a Brunovsky linear form.

Let $\Sigma = (A, B)$ be an m-input n-dimensional linear dynamical system over R. Denote by N_i^Σ the submodule of R^n generated by the columns of the $n \times im$ matrix

$$(A * B)_i^\Sigma = (B, AB, \ldots, A^{i-1}B)$$

and put $M_i^\Sigma = R^n / N_i^\Sigma$ for $1 \leqslant i \leqslant n$. The following result, see [63], contains the main properties of these modules.

PROPOSITION 85. *Let $\Sigma = (A, B)$ be an m-input, n-dimensional linear dynamical system over R. Then*:

(i) $(0) = N_0^\Sigma \subseteq N_1^\Sigma \subseteq \cdots \subseteq N_n^\Sigma$.

(ii) *The canonical homomorphism*

$$\begin{aligned} \varphi_i : N_i^\Sigma / N_{i-1}^\Sigma &\to N_{i+1}^\Sigma / N_i^\Sigma \\ \underline{x} + N_{i-1}^\Sigma &\to A\underline{x} + N_i^\Sigma \end{aligned}$$

is surjective for $1 \leqslant i \leqslant n-1$.

(iii) *If Σ is feedback equivalent to Σ' then N_i^Σ and M_i^Σ are isomorphic to $N_i^{\Sigma'}$ and $M_i^{\Sigma'}$ respectively, for* $1 \leqslant i \leqslant n$.

(iv) *If Σ is a reachable single-input n-dimensional system then the modules $\{N_i^\Sigma\}_{1 \leqslant i \leqslant n}$ and $\{M_i^\Sigma\}_{1 \leqslant i \leqslant n}$ are free.*

(v) *If Σ is a Brunovsky linear system then the modules $\{N_i^\Sigma\}_{1 \leqslant i \leqslant n}$ and $\{M_i^\Sigma\}_{1 \leqslant i \leqslant n}$ are free.*

The following result proves that the converse of statement (v) is true.

THEOREM 86. *Let $\Sigma = (A, B)$ be an m-input n-dimensional reachable linear system over R. Suppose that finitely generated projective R-modules are free. Then the following statements are equivalent*:

(i) *$\Sigma = (A, B)$ is a Brunovsky linear system.*
(ii) *M_i^Σ is a free R-module for $1 \leqslant i \leqslant n$.*

PROOF. Sketch from [63]. Assume (ii). Let p be the positive integer such that $N_p^\Sigma = R^n$ and $N_{p-1}^\Sigma \neq R^n$. The exact sequences

$$0 \to N_i^\Sigma / N_{i-1}^\Sigma \to M_{i-1}^\Sigma \to M_i^\Sigma \to 0, \quad 1 \leqslant i \leqslant p$$

split because M_i^Σ is free for $1 \leqslant i \leqslant p$. Therefore $N_i^\Sigma / N_{i-1}^\Sigma$ is a free R-module for $1 \leqslant i \leqslant p$. Put $s_i = \operatorname{rank}_R N_i^\Sigma / N_{i-1}^\Sigma$. By statement (ii) of Proposition 85 it follows that

$$s_1 \geqslant s_2 \geqslant \cdots \geqslant s_p.$$

There exists a decomposition

$$R^n \simeq R^n / N_{p-1}^\Sigma \oplus N_{p-1}^\Sigma / N_{p-2}^\Sigma \oplus \cdots \oplus N_2^\Sigma / N_1^\Sigma \oplus N_1^\Sigma$$

because the exact sequence

$$0 \to N_{i-1}^\Sigma \to N_i^\Sigma \to N_i^\Sigma / N_{i-1}^\Sigma \to 0$$

splits for $1 \leqslant i \leqslant p$.

Let $\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_{s_p}$ be elements of $N_1^\Sigma = \operatorname{Im} \varphi_B$ such that

$$\left\{ A^{p-1}\hat{g}_1 + N_{p-1}^\Sigma, A^{p-1}\hat{g}_2 + N_{p-1}^\Sigma, \ldots, A^{p-1}\hat{g}_{s_p} + N_{p-1}^\Sigma, \right\}$$

is a basis of $R^n / N_{p-1}^\Sigma = N_p^\Sigma / N_{p-1}^\Sigma$. $\square$

By statement (ii) of Proposition 85, $N_p^\Sigma / N_{p-1}^\Sigma$ is a direct summand of $N_{p-1}^\Sigma / N_{p-2}^\Sigma$ via the matrix A. Then there exist elements $\hat{g}_{s_p+1}, \hat{g}_{s_p+2}, \ldots, \hat{g}_{s_{p-1}}$ of $N_1^\Sigma = \operatorname{Im} \varphi_B$ such that

$$\begin{aligned} &\left\{ A^{p-2}\hat{g}_1 + N_{p-2}^\Sigma, A^{p-2}\hat{g}_2 + N_{p-2}^\Sigma, \ldots, A^{p-2}\hat{g}_{s_p} + N_{p-2}^\Sigma, \right. \\ &\left. A^{p-2}\hat{g}_{s_p+1} + N_{p-2}^\Sigma, A^{p-2}\hat{g}_{s_p+2} + N_{p-2}^\Sigma, \ldots, A^{p-2}\hat{g}_{s_{p-1}} + N_{p-2}^\Sigma, \right\} \end{aligned}$$

is a basis of $N_{p-1}^\Sigma / N_{p-2}^\Sigma$. Consequently

$$\begin{aligned} &\left\{ A^{p-1}\hat{g}_1, A^{p-1}\hat{g}_2, \ldots, A^{p-1}\hat{g}_{s_p}, \right. \\ &\left. A^{p-2}\hat{g}_1, A^{p-2}\hat{g}_2, \ldots, A^{p-2}\hat{g}_{s_p}, \quad A^{p-2}\hat{g}_{s_p+1}, A^{p-2}\hat{g}_{s_p+2}, \ldots, A^{p-2}\hat{g}_{s_{p-1}} \right\} \end{aligned}$$

is a basis of $R^n / N_{p-1}^\Sigma \oplus N_{p-1}^\Sigma / N_{p-2}^\Sigma$.

Iterating the process there exist elements $\hat{g}_{s_{p-1}+1}, \dots, \hat{g}_{s_2}, \hat{g}_{s_2+1}, \dots, \hat{g}_{s_1}$ of $N_1^{\Sigma} = \operatorname{Im} \varphi_B$ such that

$$\begin{array}{lllllllllll} \widehat{\mathfrak{B}} = \{ & A^{p-1}\hat{g}_1, & \dots, & A^{p-1}\hat{g}_{s_p}, & & & & & & & \\ & A^{p-2}\hat{g}_1, & \dots, & A^{p-2}\hat{g}_{s_p}, & \dots, & A^{p-2}\hat{g}_{s_{p-1}}, & & & & & \\ & \vdots & \dots, & \vdots & \dots, & \vdots & & & & & \\ & A\hat{g}_1, & \dots, & A\hat{g}_{s_p}, & \dots, & A\hat{g}_{s_{p-1}}, & \dots, & A\hat{g}_{s_2} & & & \\ & \hat{g}_1, & \dots, & \hat{g}_{s_p}, & \dots, & \hat{g}_{s_{p-1}}, & \dots, & \hat{g}_{s_2} & \dots, & \hat{g}_{s_1} \} & \end{array}$$

is a basis of R^n.

Put $\kappa_i = \max\{j \in \mathbb{Z}_+ / A^{j-1}\hat{g}_i \in \widehat{\mathfrak{B}}\}$ for $1 \leqslant i \leqslant s_1$. Then

$$\begin{aligned} \kappa_1 = \kappa_2 = \cdots = \kappa_{s_p} &= p, \\ \kappa_{s_p+1} = \kappa_{s_p+2} = \cdots &= \kappa_{s_{p-1}}, \\ &\vdots \\ \kappa_{s_2+1} = \kappa_{s_2+2} = \cdots &= \kappa_{s_1}. \end{aligned}$$

Moreover $\kappa_1 \geqslant \kappa_2 \geqslant \cdots \geqslant \kappa_{s_1}$ and $\kappa_1 + \kappa_2 + \cdots + \kappa_{s_1} = n$. The result follows, analogously to the single-input case, modifying $\widehat{\mathfrak{B}}$ and taking an adequate feedback matrix.

COROLLARY 87. *Suppose that $\Sigma = (A, B)$ is a Brunovsky linear system. Then the class of feedback equivalence of Σ is characterized by each one of the following sets*:

(i) *The Kronecker indices* $\{\kappa_i\}_{1 \leqslant i \leqslant s}$.

(ii) $\{n_i = \operatorname{rank}_R M_i^{\Sigma}\}_{1 \leqslant i \leqslant n}$ *or equivalently* $\{n - n_i = \operatorname{rank}_R N_i^{\Sigma}\}_{1 \leqslant i \leqslant n}$.

(iii) $\{s_i = n_{i-1} - n_i = \operatorname{rank}_R N_i^{\Sigma} / N_{i-1}^{\Sigma}\}_{1 \leqslant i \leqslant n}$.

Let $\Sigma = (A, B)$ be an m-input n-dimensional linear system and consider the finite presentation

$$R^{im} \xrightarrow{\varphi_{(A*B)_i^{\Sigma}}} R^n \to M_i^{\Sigma} \to 0$$

of M_i^{Σ}. By Theorem 23 M_i^{Σ} is projective of rank n_i if and only if

$$\mathcal{F}_j\big(M_i^{\Sigma}\big) = \mathcal{U}_{n-j}\big((A * B)_i^{\Sigma}\big) = \begin{cases} 0 & \text{for } 0 \leqslant j < n_i, \\ R & \text{for } j \geqslant n_i. \end{cases}$$

Consequently we can determine if Σ is a Brunovsky linear system and, if the answer is positive, we can obtain a complete set of invariants for the feedback class of Σ.

COROLLARY 88 (Classical case). *A reachable linear system over a field is feedback equivalent to a Brunovsky linear system.*

The above result is due to P. Brunovsky, see [21]. In [111, Section 4.2], E.D. Sontag gives a complete discussion, based on [76], of this case.

Next we expose a decomposition result for nonreachable linear systems. When R is a field this result is known as the Kalman controllability decomposition.

PROPOSITION 89. *Let $\Sigma = (A, B)$ be an n-dimensional system over R such that*

$$\mathcal{U}_i(A * B) = \begin{cases} R & \text{for } 0 \leqslant i < r, \\ 0 & \text{for } i > r. \end{cases}$$

Suppose that finitely generated projective R-modules are free. Then Σ is feedback equivalent to a system of the form

$$\widetilde{\Sigma} = \left(\widetilde{A} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix},\ \widetilde{B} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}\right),$$

where $\Sigma_1 = (A_1, B_1)$ is a reachable m-input r-dimensional system.

PROOF. By Theorem 23 the exact sequence

$$0 \to N_n^{\Sigma} \to R^n \to M_n^{\Sigma} \to 0$$

splits. The result follows considering the decomposition

$$R^n \cong M_n^{\Sigma} \oplus N_n^{\Sigma}. \qquad \square$$

4.2. *Pointwise feedback relation*

If two linear systems Σ and Σ' over R are feedback equivalent then the extensions $\Sigma(\mathfrak{p})$ and $\Sigma'(\mathfrak{p})$ are feedback equivalent for all prime ideal $\mathfrak{p}$ of R. The converse is not true in general.

DEFINITION 90. Two reachable linear systems Σ and Σ' are pointwise (resp. closed pointwise) feedback equivalent if and only if $\Sigma(\mathfrak{p})$ (resp. $\Sigma(\mathfrak{m})$) is feedback equivalent to $\Sigma'(\mathfrak{p})$ (resp. $\Sigma'(\mathfrak{m})$) for all prime (resp. maximal) ideal $\mathfrak{p}$ (resp. $\mathfrak{m}$) of R.

By Corollary 87 and Proposition 14, one has the following characterization result, see [30].

THEOREM 91. *Let Σ and Σ' be two reachable m-input, n-dimensional linear systems over R. Then the following statements are equivalent*:

(i) *Σ and Σ' are pointwise feedback equivalent.*

(ii) *The ideals $\mathcal{F}_j(M_i^{\Sigma})$ and $\mathcal{F}_j(M_i^{\Sigma'})$ have the same radical ideal* (*i.e. the closed sets $V(\mathcal{F}_j(M_i^{\Sigma}))$ and $V(\mathcal{F}_j(M_i^{\Sigma'}))$ of* $\operatorname{Spec}(R)$ *are equal*) *for $i \in \{1, \ldots, n\}$ and $j \in \{0, \ldots, n\}$.*

As a consequence of the above result one has that Σ is pointwise feedback equivalent to a Brunovsky system if and only if the function

$$\begin{aligned} \operatorname{Spec}(R) &\to \mathbb{Z} \\ \mathfrak{p} &\mapsto \dim_{k(\mathfrak{p})} M_i^{\Sigma}(\mathfrak{p}) \end{aligned}$$

is constant for each $i \in \{1, \ldots, n\}$.

REMARK 92. In the above result one obtains a complete set of invariants for the pointwise feedback equivalence class of Σ. An open question is to obtain a canonical form for the class of Σ.

COROLLARY 93. *Let Σ and Σ' be two reachable m-input, n-dimensional linear systems over R. Then the following statements are equivalent*:

(i) *Σ and Σ' are closed pointwise feedback equivalent.*

(ii) $V(\mathcal{F}_j(M_i^{\Sigma})) \cap \operatorname{Max}(R) = V(\mathcal{F}_j(M_i^{\Sigma'})) \cap \operatorname{Max}(R)$ *for all positive integers i and j.*

THEOREM 94. *The following statements are equivalent*:

(i) *R is an absolutely flat ring* (*i.e. every R-module is flat*).

(ii) *Two reachable linear systems are feedback equivalent if and only if they are pointwise feedback equivalents.*

PROOF. See [30]. □

Next we shall consider the closed pointwise feedback relation when R is a ring of real-valued functions.

Let $R = \mathfrak{C}(X, \mathbb{R})$ be the ring of real-valued functions over a Haussdorf topological space X. If $\mathfrak{a}$ is an ideal of R we denote by $Z(\mathfrak{a})$ the set of zeros of $\mathfrak{a}$, that is

$$Z(\mathfrak{a}) = \bigl\{x \in X\colon\ f(x) = 0, \forall f \in \mathfrak{a}\bigr\}.$$

If $\mathfrak{a}$ is generated by $f_1, f_2, \ldots, f_k$ then $Z(\mathfrak{a})$ is the set of zeros of the function $f_1^2 + f_2^2 + \cdots + f_k^2$.

The map

$$\begin{aligned} \xi : X &\to \operatorname{Max} R \\ x &\mapsto \mathfrak{m}_x = \{f \in R / f(x) = 0\} \end{aligned}$$

is injective and continuous. Moreover if X is compact (i.e. quasi-compact and Haussdorf) then ξ is an homeomorphism and

$$\xi\bigl(Z(\mathfrak{a})\bigr) = V(\mathfrak{a}) \cap \operatorname{Max}(R).$$

Let $\Sigma = (A, B)$ be a linear system over $\mathfrak{C}(X, \mathbb{R})$. For $x \in X$ we denote by $\Sigma(x)$ the linear system over $\mathbb{R}$ given by $\Sigma(x) = (A(x), B(x))$ where $A(x) = (a_{ij}(x))$ and $B(x) = (b_{ij}(x))$. As a consequence of Corollary 93, we have:

THEOREM 95. *Let $R = \mathfrak{C}(X, \mathbb{R})$ be the ring of real-valued functions over a compact topological space X. If Σ and Σ' are two reachable m-input, n-dimensional linear systems over R then the following statements are equivalent*:

(i) *$\Sigma(x)$ is feedback equivalent to $\Sigma'(x)$ for all $x \in X$.*

(ii) *$Z(\mathcal{F}_j(M_i^{\Sigma})) = Z(\mathcal{F}_j(M_i^{\Sigma'}))$ for all positive integers i and j.*

By Theorem 94 and [50, p. 62] one has

THEOREM 96. *Let $R = \mathfrak{C}(X, \mathbb{R})$ be the ring of real-valued functions over a compact topological space X. The following statements are equivalent*:

(i) *X is a P-space.*

(ii) *Σ is feedback equivalent to Σ' if and only if $\Sigma(x)$ is feedback equivalent to $\Sigma'(x)$ for all $x \in X$.*

4.3. *Classifying systems over a principal ideal domain*

The goal of this section is to study the feedback classification problem when R is a principal ideal domain (PID) and specially when R is the ring of integers $\mathbb{Z}$ or the polynomial ring $k[T]$ over a field k. In [16] it is proved that the category of reachable systems over $\mathbb{Z}$ is wild in the sense of classical representation theory and hence it is unlikely that canonical forms will be found. However in some cases it is possible to solve the classification problem.

4.3.1. *2-dimensional systems* Let R be a PID and $\Sigma = (A, B)$ a reachable m-input 2-dimensional linear system over R. Since $\mathcal{U}_1(B) = R$ there exist invertible matrices P and Q such that

$$PBQ = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & d & \dots & 0 \end{pmatrix}$$

where d is a generator of $\mathcal{U}_2(B)$. Considering the relevant actions of the feedback group we obtain the following result.

PROPOSITION 97. *Let $\Sigma = (A, B)$ be a 2-dimensional reachable linear system over a principal ideal domain R. Then Σ is feedback equivalent to a system of the form*

$$\Sigma_{\{a,d\}} = \left(\widehat{A} = \begin{pmatrix} 0 & 0 \\ a & 0 \end{pmatrix},\ \widehat{B} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & d & \dots & 0 \end{pmatrix} \right)$$

where a and d are coprime.

We say that the system $\Sigma_{\{a,d\}}$ is a normal form associated to Σ. Since d is a generator of $\mathcal{U}_2(B)$ it follows that d is a feedback invariant of Σ. But a is not an invariant associated to Σ.

The following result, [62], characterizes when two normal forms are feedback equivalent.

THEOREM 98. *Two normal forms $\Sigma_{\{a,d\}}$ and $\Sigma_{\{a',d'\}}$ are feedback equivalent if and only if the following two conditions are satisfied*:

(i) *The principal ideals (d) and (d') are equal.*

(ii) *There exist a unit u of R and an element h of R such that $a' \equiv uh^2 a \pmod{d}$.*

For two elements a and a' of R such that $(a,d) = (a',d) = R$ we put $a \sim a'$ if and only there exist a unit u of R and an element h of R such that $a' \equiv uh^2 a \pmod{d}$. So one has an equivalence relation, denoted by $\sim$, on the group of units $(R/(d))^*$ of $R/(d)$. The class of a by this relation is denoted by $[\tilde{a}]$.

COROLLARY 99. *Let R be a PID and Σ be a 2-dimensional reachable linear system over R. If $\Sigma_{\{a,d\}}$ is a normal form associated to Σ then the pair $\{[\tilde{a}], d\}$ is a complete set of invariants for the feedback class of Σ.*

REMARK 100. Consider the normal formal $\Sigma_{\{a,d\}}$. Then

$$M_1^{\Sigma_{\{a,d\}}} = R^2/N_1^{\Sigma_{\{a,d\}}} \cong R/(d),$$

and

$$M_2^{\Sigma_{\{a,d\}}} = R^2/N_2^{\Sigma_{\{a,d\}}} = 0.$$

Consequently, for reachable systems over a PID the set $\{M_i^{\Sigma}\}_{1 \leqslant i \leqslant n}$ is not, in general, a complete system of invariants for the feedback class of Σ.

REMARK 101. In Section 5.3.1 we shall study the systems which are characterized by the pair $\{[\tilde{1}], d\}$.

4.3.2. *Single input* Let R be a PID and $\Sigma = (A, \underline{b})$ a single input n-dimensional system over R. We say that $\Sigma = (A, \underline{b})$ is weakly reachable if

$$\operatorname{rank}_R(A * \underline{b}) = n.$$

Note that Σ is weakly reachable if and only if the extension of Σ to $\mathcal{T}(R)$ (the total field of fractions of R) is reachable.

Let $d_1 \neq 0$ be a generator of $\mathcal{U}_1(\underline{b})$. Then there exists an $n \times n$ invertible matrix P_1 such that $P_1 \underline{b} = (d_1, 0, \ldots, 0)^t$. Therefore Σ is feedback equivalent to a system of the form

$$\left(\left(\begin{array}{c|c} a_{11} & \underline{a}^t \\ \hline \underline{b}_1 & A_1 \end{array} \right), \begin{pmatrix} d_1 \\ \underline{0} \end{pmatrix} \right),$$

where A_1 is an $(n-1) \times (n-1)$ matrix, $\underline{b}_1$ is a column matrix, $\underline{a}^t$ is a row matrix and $\underline{0}$ is the zero column. Moreover the ideal (d_1) is a feedback invariant associated to Σ.

Consider the single input $(n-1)$-dimensional system $\Sigma_1 = (A_1, \underline{b}_1)$ and let $d_2 \neq 0$ be a generator of $\mathcal{U}_1(\underline{b}_1)$. Then there exists an $(n-1) \times (n-1)$ invertible matrix P_2 such that $P_2\underline{b} = (d_2, 0, \dots, 0)^t$. Hence Σ is feedback equivalent to a linear system of the form

$$\left(\begin{pmatrix} * & * & * \\ d_2 & * & * \\ \underline{0} & \underline{b}_2 & A_2 \end{pmatrix}, \begin{pmatrix} d_1 \\ 0 \\ \underline{0} \end{pmatrix}\right).$$

Iterating the process one has the following result, see [29].

THEOREM 102. *Let $\Sigma = (A, \underline{b})$ be a single input n-dimensional linear system that is weakly reachable. Then there exist nonzero elements $d_1, \dots, d_n$ of R such that Σ is feedback equivalent to a system of the form*

$$\Sigma^{\Delta} = \left(A^{\Delta}, \underline{b}^{\Delta}\right) = \left(\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,n-1} & a_{1n} \\ d_2 & a_{22} & \cdots & a_{2,n-1} & a_{2n} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & \cdots & d_n & a_{nn} \end{pmatrix}, \begin{pmatrix} d_1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}\right).$$

Moreover the ordered sequence of ideals $\{(d_1), (d_2), \dots, (d_n)\}$ is a feedback invariant associated to Σ.

The system $\Sigma^{\Delta} = (A^{\Delta}, \underline{b}^{\Delta})$ is called a normal form associated to $\Sigma = (A, \underline{b})$. It is easy to prove that $\{(d_1), (d_2), \dots, (d_n)\}$ is not a complete system of invariants for the feedback class of Σ.

Next we study when two normal forms, with the same set $\{(d_1), (d_2), \dots, (d_n)\}$, are feedback equivalent.

First note that if $\varepsilon_1, \dots, \varepsilon_n$ are units of R and P is the diagonal matrix

$$P = \operatorname{diag}(\varepsilon_1, \varepsilon_1\varepsilon_2, \dots, \varepsilon_1 \dots \varepsilon_{n-1}, \varepsilon_1 \dots \varepsilon_n),$$

then

$$\left(PA^{\Delta}P^{-1}, P\underline{b}^{\Delta}\right) = \left(\begin{pmatrix} * & * & \cdots & * & * \\ \varepsilon_2 d_2 & * & \cdots & * & * \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & * & * \\ 0 & 0 & \cdots & \varepsilon_n d_n & * \end{pmatrix}, \begin{pmatrix} \varepsilon_1 d_1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}\right).$$

Consequently we can suppose that both normal forms have the same ordered sequence, $d_1, d_2, \dots, d_n$. In [29] there is proved the following result.

THEOREM 103. *Consider two normal forms*

$$\Sigma = (A, \underline{b}) = \left(\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,n-1} & a_{1n} \\ d_2 & a_{22} & \cdots & a_{2,n-1} & a_{2n} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & \cdots & d_n & a_{nn} \end{pmatrix}, \begin{pmatrix} d_1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \right),$$

and

$$\Sigma' = (A', \underline{b}) = \left(\begin{pmatrix} a'_{11} & a'_{12} & \cdots & a'_{1,n-1} & a'_{1n} \\ d_2 & a'_{22} & \cdots & a'_{2,n-1} & a'_{2n} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & a'_{n-1,n-1} & a'_{n-1,n} \\ 0 & 0 & \cdots & d_n & a'_{nn} \end{pmatrix}, \begin{pmatrix} d_1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \right)$$

with the same ordered sequence $d_1, d_2, \ldots, d_n$. Then the following statements are equivalent:

(i) *Σ and Σ' are feedback equivalent.*

(ii) *There exists one and only one feedback action (P, F) such that*

$$\begin{cases} A' = P(A + \underline{b}F)P^{-1}, \\ P\underline{b} = \underline{b}. \end{cases}$$

Moreover this feedback action is of the form

$$P = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1,n-1} & x_{1n} \\ 0 & 1 & \ldots & x_{2,n-1} & x_{2n} \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \ldots & 1 & x_{n-1,n} \\ 0 & 0 & & 0 & 1 \end{pmatrix}, \qquad F = (x_{01}, x_{02}, \ldots, x_{0n}),$$

for some elements $x_{ij} \in R$ with $0 \leqslant i < j \leqslant n$.

When $R = \mathbb{Z}$ or $R = k[T]$ it is possible to construct a canonical form associated to a single input system that is weakly reachable.

PROPOSITION 104. *Let $\Sigma = (A, \underline{b})$ be an n-dimensional normal form over $\mathbb{Z}$ (resp. $k[T]$). Then, Σ is feedback equivalent to a unique normal form*

$$\Sigma_c = (A_c, B_c) = \left(\begin{pmatrix} \hat{a}_{11} & \hat{a}_{12} & \cdots & \hat{a}_{1,n-1} & \hat{a}_{1n} \\ \hat{d}_2 & \hat{a}_{22} & \cdots & \hat{a}_{2,n-1} & \hat{a}_{2n} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & \hat{a}_{n-1,n-1} & \hat{a}_{n-1,n} \\ 0 & 0 & \cdots & \hat{d}_n & \hat{a}_{nn} \end{pmatrix}, \begin{pmatrix} \hat{d}_1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \right),$$

that satisfies the following properties:

(i) $\hat{d}_i > 0$ (*resp.* $\hat{d}_i$ *is a monic polynomial*) *for each* $i = 1, \ldots, n$.

(ii) $0 \leqslant \hat{a}_{ij} < \hat{d}_i$ (*resp.* $\deg(\hat{a}_{ij}) < \deg(\hat{d}_i)$) *for each* $i = 1, \ldots, n$ *and each* $j = i, i+1, \ldots, n$.

The system $\Sigma_c = (A_c, B_c)$ is the canonical form associated to Σ and it is characterized by $n(n+3)/2$ invariants.

REMARK 105. The normal form over $\mathbb{Z}$ given by

$$\Sigma = \left(\begin{pmatrix} 6 & 6 \\ 3 & 4 \end{pmatrix}, \begin{pmatrix} 5 \\ 0 \end{pmatrix} \right)$$

is not feedback equivalent to the form

$$\Sigma' = \left(\begin{pmatrix} 1 & 1 \\ 3 & 1 \end{pmatrix}, \begin{pmatrix} 5 \\ 0 \end{pmatrix} \right)$$

however; $6 \equiv 1 \pmod 5$ and $4 \equiv 1 \pmod 3$.

4.3.3. *The local case* Let R be a local ring with maximal ideal $\mathfrak{m}$ and $\Sigma = (A, B)$ a reachable m-input n-dimensional linear dynamical system over R. Put

$$\xi_1 = \text{res.rank}_R(B) = \max\bigl\{i \colon \mathcal{U}_i(B) = R\bigr\} \geqslant 1.$$

Then Σ is feedback equivalent to a linear system of the form

$$\left(\left(\begin{array}{c|c} 0 & 0 \\ \hline B^{(2)} & A^{(2)} \end{array} \right), \left(\begin{array}{c|c} Id_{\xi_1} & 0 \\ \hline 0 & Y_1 \end{array} \right) \right),$$

where Y_1 is a matrix with entries in $\mathfrak{m}$.

The ξ_1-input, $(n - \xi_1)$-dimensional linear dynamical system

$$\Sigma_2 = \bigl(A^{(2)}, B^{(2)}\bigr)$$

is reachable. Put $\xi_2 = \text{res.rank}_R(B)$. Then Σ is feedback equivalent to a linear system of the form

$$\left(\left(\begin{array}{c|c} 0 & 0 \\ \hline \begin{array}{c|c} Id_{\xi_2} & 0 \\ \hline 0 & Z_2 \end{array} & \begin{array}{c|c} 0 & 0 \\ \hline B^{(3)} & A^{(3)} \end{array} \end{array}\right), \left(\begin{array}{c|c} Id_{\xi_1} & 0 \\ \hline 0 & Y_2 \end{array}\right)\right),$$

where Y_2 and Z_2 are matrices with entries in $\mathfrak{m}$.

Iterating the process, see [32], the following result can be proved.

PROPOSITION 106. *There exist positive integers $\xi_1, \xi_2, \dots, \xi_s$ such that Σ is feedback equivalent to a system of the form*

$$\widetilde{\Sigma} = (\widetilde{A}, \widetilde{B}) = \left(\widetilde{A} = \left(\begin{array}{c|c} 0 & 0 \\ \hline \widetilde{B}^{(2)} & \widetilde{A}^{(2)} \end{array}\right), \widetilde{B} = \left(\begin{array}{c|c} Id_{\xi_1} & 0 \\ \hline 0 & X_1 \end{array}\right)\right),$$

with

$$\widetilde{A}^{(j)} = \left(\begin{array}{c|c} 0 & 0 \\ \hline \widetilde{B}^{(j+1)} & \widetilde{A}^{(j+1)} \end{array}\right), \qquad \widetilde{B}^{(j)} = \left(\begin{array}{c|c} Id_{\xi_j} & 0 \\ \hline 0 & X_j \end{array}\right),$$

for $j = 2, \dots, s-1$ and

$$\widetilde{A}^{(s)} = (0), \qquad \widetilde{B}^{(s)} = (Id_{\xi_s} \mid 0),$$

where all the entries of the matrices $X_1, \dots, X_{s-1}$ are in the maximal ideal $\mathfrak{m}$ of R.

The positive integers $\{\xi_i\}_{1 \leqslant i \leqslant s}$ are related to the feedback invariants $\{M_i^\Sigma\}_{1 \leqslant i \leqslant n}$ by the equalities

$$\dim_{R/\mathfrak{m}}\left(M_i^\Sigma \otimes_R R/\mathfrak{m}\right) = n - \left(\sum_{j=1}^{i} \xi_j\right) \quad \text{for } i < s$$

and

$$M_i^\Sigma = (0) \quad \text{for } s \leqslant i \leqslant n.$$

Using the above normal form, see [32], the following result can be proved.

THEOREM 107. *Let R be a discrete valuation ring and Σ a reachable n-dimensional linear dynamical system over R. Suppose that M_i^Σ is free for $i = 2, \dots, n$. Then $\{M_1^\Sigma, \dots, M_n^\Sigma\}$ is a complete set of invariants for the feedback class of Σ.*

5. Pole shifting

The characteristic polynomial $\chi(A, T)$ of A is not, in general, a feedback invariant associated to linear system $\Sigma = (A, B)$. In this section we study the changes in the poles of Σ by the action of the feedback group. First we introduce some terminology.

DEFINITION 108. Let $\Sigma = (A, B)$ be an m-input n-dimensional linear dynamical system over R.

(i) A monic polynomial $f(T)$ of degree n is assignable to Σ if there exists an $n \times m$ matrix F such that $\chi(A + BF, T) = f(T)$.

(ii) The system Σ is coefficient assignable when all monic polynomials of degree n are assignable to Σ.

(iii) We say that r poles can be assigned to Σ if for every r elements $\lambda_1, \dots, \lambda_r$ of R there exists a monic polynomial $g(T)$ of degree $n - r$ such that the polynomial

$$f(T) = (T - \lambda_1) \cdots (T - \lambda_r) g(T),$$

is assignable to Σ.

(iv) The system Σ is pole assignable when n poles can be assigned to Σ.

The set of all assignable polynomials to Σ is a feedback invariant associated to Σ because $\chi(PAP^{-1}, T) = \chi(A, T)$ for each invertible matrix P.

REMARK 109. Let M be a projective R-module of constant rank n and $\Sigma = (\varphi, \psi)$ an m-input linear system over M, see Remark 76. A monic polynomial $f(T) \in R[T]$ of degree n is assignable to Σ if there exists an homomorphism $\phi : P \to R^m$ such that $f(T) = \chi(\varphi + \psi\phi, T)$, see Section 3.2. Similarly to the free case, we say that r poles can be assigned to Σ if for every elements $\lambda_1, \dots, \lambda_r$ of R there exists a monic polynomial $g(T)$ of degree $n - r$ such that the polynomial

$$f(T) = (T - \lambda_1) \cdots (T - \lambda_r) g(T),$$

is assignable to Σ. The system Σ is pole assignable when n poles can be assigned to Σ.

5.1. *Pole shifting for Brunovsky systems*

Consider the single input n-dimensional canonical form

$$\widehat{\Sigma} = \left(\widehat{A} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \ \underline{\hat{b}} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \right),$$

and $f(T) = T^n - a_n T^{n-1} - \cdots - a_2 T - a_1 \in R[T]$. Then the row matrix

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

verifies $\chi(\widehat{A} + \hat{\underline{b}} F, T) = f(T)$. By Proposition 82 one has

PROPOSITION 110. *Let Σ be a reachable single input system. Then Σ is coefficient assignable.*

The following result reduces the multi-input case to the single-input case.

PROPOSITION 111. *Let $\Sigma_\kappa = (A_\kappa, B_\kappa)$ be the m-input n-dimensional Brunovsky linear form associated to the partition $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_s)$ of n. Then there exist an $m \times n$ matrix F and a vector $\underline{v} \in R^m$ such that the single-input linear system $\Sigma = (A_\kappa + B_\kappa F,\ B_\kappa \underline{v})$ is reachable.*

PROOF. With respect to the standard bases $\{\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_m\}$ and $\{\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_n\}$ of R^m and R^n, respectively one has

$$B_\kappa \underline{u}_i = \underline{e}_{\kappa_1 + \kappa_2 + \cdots + \kappa_i},$$

for $i = 1, 2, \ldots, s$. The result follows considering $\underline{v} = \underline{u}_s$ and F the matrix defined by

$$F \underline{e}_{\kappa_1 + \kappa_2 + \ldots + \kappa_i + 1} = \underline{u}_i,$$

for $i = 1, 2, \ldots, s - 1$ and

$$F \underline{e}_j = 0$$

for $j \in \{1, 2, \ldots, n\} \setminus \{\kappa_1 + \kappa_2 + \cdots + \kappa_i + 1\}_{i=1,2,\ldots,s-1}$. $\square$

COROLLARY 112. *Let Σ be a Brunovsky linear system. Then Σ is coefficient assignable and hence pole assignable.*

As a consequence of this result one has that if R is a field then a reachable system is coefficient assignable. The following result, called the pole-shifting theorem, characterizes what polynomials are assignable to a non reachable system over a field.

THEOREM 113. *Let K be a field and $\Sigma = (A, B)$ an n-dimensional system over K. Suppose that* $\operatorname{rank}_K(A * B) = r$. *Then there exists a monic polynomial $\chi_u(T)$ of degree $n - r$ such that the assignable polynomials to Σ are precisely those of the form*

$$f(T) = g(T)\chi_u(T),$$

where $g(T)$ is an arbitrary monic polynomial of degree r. In particular r poles can be assigned to Σ.

PROOF. By Proposition 89, Σ is feedback equivalent to a linear system of the form

$$\widetilde{\Sigma} = \left(\widetilde{A} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \widetilde{B} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}\right),$$

where $\Sigma_1 = (A_1, B_1)$ is a reachable m-input r-dimensional system. If $F = (F_1, F_2)$ is a feedback matrix then

$$\chi(\widetilde{A} + \widetilde{B}F, T) = \chi(A_1 + B_1F_1, T)\chi(A_3).$$

The result follows applying Corollary 112 to the linear system Σ_1. □

5.2. *How many poles can be assigned to a system?*

Let $\mathbb{F}_q$ be the finite field of q elements. If D is a diagonal $q \times q$ matrix with all the elements of $\mathbb{F}_q$ down the main diagonal then the system $\Sigma = (D, 0)$ satisfies $\operatorname{rank}_K(A * B) = 0$; however, 1 pole can be assigned to Σ. Consequently the converse of the pole-shifting theorem is not, in general, true.

PROPOSITION 114. *Let K be a field and $\Sigma = (A, B)$ a linear system over K. Then*:

(i) *Σ is reachable if and only if Σ is pole assignable.*

(ii) *If K is an infinite field then r poles can be assigned to Σ if and only if* $\operatorname{rank}_K(A * B) = r$.

Next we shall bound the number of poles that we can assigned to Σ when R is an arbitrary commutative ring.

DEFINITION 115. Let $\Sigma = (A, B)$ be a linear system over R. The residue rank of Σ is defined by

$$\text{res.rk}(\Sigma) = \text{res.rk}(A * B) = \min\bigl\{\operatorname{rank}\bigl(A(\mathfrak{m}) * B(\mathfrak{m})\bigr)\colon\ \mathfrak{m} \in \operatorname{Max}(R)\bigr\}.$$

By construction, the residue rank is a feedback invariant associated to Σ. Moreover $\text{res.rk}(\Sigma) = n$, where n is the dimension of Σ, if and only if Σ is reachable.

PROPOSITION 116. *Let R be a commutative ring and $\Sigma = (A, B)$ be a linear system over R.*

(i) *If Σ is pole assignable then Σ is reachable.*

(ii) *Suppose that R is a residually infinite ring (i.e. the residue field $R/\mathfrak{m}$ is infinite for every maximal ideal $\mathfrak{m}$ of R). If r poles can be assigned to Σ then* $\text{res.rk}(\Sigma) \geqslant r$.

PROOF. If r poles can be assigned to Σ (in particular if Σ is pole assignable) then r poles can be assigned to $\Sigma(\mathfrak{m})$ for all maximal ideals $\mathfrak{m}$ of R. The result is consequence of Propositions 114 and 81. □

In [31] it is proved that in some cases the converse of statement (ii) of the above result is true.

PROPOSITION 117. *Let R be a Noetherian ring. Then the following statements are equivalent*:

(i) *If r poles can be assigned to Σ then* $\text{res.rk}(\Sigma) \geqslant r$.

(ii) *R is a residually infinite ring.*

5.3. *Pole assignability*

Let R be a commutative ring and Σ a linear system over R. By Proposition 116, if Σ is pole assignable then Σ is reachable. Later on we shall give examples to show that the converse is not true in general (note that, by Proposition 114, if R is a field the converse is true). On the other hand there exists a ring R and a system Σ over R such that $\text{res.rk}(\Sigma)$ poles cannot be assigned to Σ.

DEFINITION 118. Let R be a commutative ring. We say that

(i) R has the coefficient assignability property (or R is a *CA*-ring) if every reachable system over R is coefficient assignable.

(ii) R has the pole assignability property (or R is a *PA*-ring) if every reachable system over R is pole assignable.

(iii) R has the pole-shifting property (or R is a *PS*-ring) if $\text{res.rk}(\Sigma)$ poles can be assigned to Σ for every system Σ.

Clearly *CA*-ring implies *PA*-ring and *PS*-ring implies *PA*-ring. By Proposition 114 a field K is a *CA*-ring and, by Theorem 113, a *PS*-ring.

By Propositions 110 and 116, the concepts Σ is reachable, Σ is pole assignable and Σ is coefficient assignable are equivalent for a single input linear system Σ over a commutative ring R. On the other hand, we have proved (see Proposition 111) that a Brunovsky linear system is coefficient assignable by reduction of the multi-input case to the single-input case. These facts give a reason to introduce the following notion.

DEFINITION 119. Let $\Sigma = (A, B)$ be an n-dimensional system over R.

(i) Σ has the feedback cyclization property if there exist a feedback matrix F and a vector $\underline{u}$ such that the single input system $\Sigma' = (A + BF, B\underline{u})$ is reachable.

(ii) The ring R has the feedback cyclization property (or R is an *FC*-ring) if all reachable linear system has the feedback cyclization property.

It is clear that an *FC*-ring is a *CA*-ring and hence a *PA*-ring.

REMARK 120. Let $\Sigma = (A, B)$ be an n-dimensional linear system over R. For a nonnegative integer r we denote by $\Sigma(r)$ the $(n + r)$-dimensional system

$$\Sigma(r) = \left(\begin{pmatrix} 0 & 0 \\ 0 & A \end{pmatrix}, \begin{pmatrix} Id_r & 0 \\ 0 & B \end{pmatrix} \right).$$

In [41] the equivalence between the following statements is proved: (i) Σ is reachable, (ii) $\Sigma(r)$ has the feedback cyclization property for some non negative integer $r \leqslant n^2$, and (iii) $\Sigma(r)$ is coefficient assignable for some non negative integer $r \leqslant n^2$. When R is a Dedekind domain then $r \leqslant n-1$, see [13], and if R is a principal ideal domain then $r \leqslant 1$, see [17].

We say that two systems Σ and Σ' are dynamically feedback equivalent if and only if $\Sigma(r)$ and $\Sigma'(r)$ are feedback equivalent for some non negative integer r. The dynamical feedback classification problem, in particular over principal ideal domains, is open.

5.3.1. *Feedback cyclization property*

THEOREM 121.

(i) *A field K is an FC-ring.*
(ii) *If R is an FC-ring and $\mathfrak{a}$ is an ideal of R then $R/\mathfrak{a}$ is an FC-ring.*
(iii) *R is an FC-ring if and only if $R/J(R)$ is an FC-ring where $J(R)$ is the Jacobson radical of R.*
(iv) *A semilocal ring is an FC-ring.*
(v) *The power series ring $R[[T]]$ is an FC-ring if and only if R is an FC-ring.*
(vi) *The direct product $\prod_{i \in I} R_i$ is an FC-ring if and only if R_i is an FC-ring for each i.*

With notations of Section 4.3.1 in [62] the following result is proved which characterizes when a 2-dimensional system over a PID has the feedback cyclization property.

THEOREM 122. *Let Σ be a 2-dimensional reachable linear system over a principal ideal domain R. Then the pair $\{[\tilde{1}], d\}$ is a system of invariants for Σ if and only if Σ has the feedback cyclization property.*

For an element $d \in R$ we denote by $\omega(d)$ the cardinal of the quotient set $(R/(d))^*/\sim$. Put

$$\widehat{B} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & d & \dots & 0 \end{pmatrix}.$$

By Corollary 99, $\omega(d)$ is the cardinal of the set of all feedback classes of reachable 2-dimensional systems $\Sigma = (A, B)$ where B is equivalent to $\widehat{B}$.

As a consequence of the above theorem one has:

COROLLARY 123. *Let R be a PID. The following statements are equivalent*:

(i) *All reachable 2-dimensional systems over R have the feedback cyclization property.*
(ii) *If a and d are two coprime elements of R then there exist a unit u of R and an element h of R such that $a \equiv uh^2 \pmod d$.*
(iii) *Two reachable 2-dimensional systems $\Sigma = (A, B)$ and $\Sigma' = (A', B')$ are feedback equivalent if and only if B is equivalent to B'.*
(iv) *$\omega(d) = 1$ for all $d \in R$.*

REMARK 124. The above result it is proved in [25] when $R/(p)$ has characteristic different from 2 for all irreducible element p of R.

COROLLARY 125. *All reachable 2-dimensional systems over $\mathbb{C}[T]$ have the feedback cyclization property.*

REMARK 126. In [18] it is proved that if K is an algebraically closed field then the polynomial ring $K[T]$ is a *CA*-ring. Moreover if the characteristic of K is different from zero then $K[T]$ is not an *FC*-ring. It is an open problem to determine whether or not $\mathbb{C}[T]$ has the feedback cyclization property.

In [62] the invariant $\omega(d)$ when R is $\mathbb{Z}$ or $\mathbb{R}[T]$ is constructed. Consequently, by Corollary 123, we can determine when a reachable 2-dimensional linear system over these rings has the feedback cyclization property.

PROPOSITION 127. *Let $d(T) \in \mathbb{R}[T]$. Suppose that $d(T)$ has t different roots on $\mathbb{R}$. Then*

$$\omega(d(T)) = \begin{cases} 1 & \text{if } t = 0, \\ 2^{t-1} & \text{if } t > 0. \end{cases}$$

In particular $\omega(d(T)) = 1$ if and only if $d(T)$ has at most one real root.

LEMMA 128. *Let $d \in \mathbb{Z}$ and suppose that*

$$d = 2^r p_1^{r_1} p_2^{r_2} \dots p_t^{r_t},$$

where p_i is an odd prime for each i.

(i) *The number $\delta(d)$ of different solutions of the equation $T^2 - 1 = 0$ in $\mathbb{Z}/(d)$ is given by*

$$\delta(d) = \begin{cases} 2^t & \text{if } r = 0, 1, \\ 2^{t+1} & \text{if } r = 2, \\ 2^{t+2} & \text{if } r \geqslant 3. \end{cases}$$

(ii) *Let $\partial(d)$ be the number of solutions of the equation $T^2 + 1 = 0$ in $\mathbb{Z}/(d)$. Then $\partial(d) \geqslant 1$ if and only if $r \leqslant 1$ and $p_i \equiv 1 \pmod 4$ for $i = 1, 2, \dots, t$.*

PROPOSITION 129. *Let $d = 2^r p_1^{r_1} p_2^{r_2} \dots p_t^{r_t}$ where p_i is an odd prime for each i. Then*

$$\omega(d) = \begin{cases} \delta(d) & \text{if } \partial(d) \geqslant 1, \\ \delta(d)/2 & \text{if } \partial(d) = 0. \end{cases}$$

In particular $\omega(d) = 1$ if and only if either $d = 2$ or $d = 4$ or $d = 2^r p^s$ where $r \in \{0, 1\}$ and p is a prime integer such that $p \equiv 3 \pmod 4$.

COROLLARY 130. *$\mathbb{Z}$ and $\mathbb{R}[T]$ are not FC-rings.*

Let $\Sigma = (A, B)$ be a reachable m-input n-dimensional system over R. If R is an FC-ring then there exists $\underline{u} \in R^m$ such that $B\underline{u}$ is part of a basis of R^n and hence $B\underline{u}$ is unimodular (i.e. $\mathcal{U}_1(Bu) = R$).

DEFINITION 131. We say that R has the GCU-property if and only if whenever $\Sigma = (A, B)$ is a reachable linear system over R, there exists $\underline{u} \in R^m$ such that $B\underline{u}$ is unimodular.

REMARK 132. The ring R has the UCU-property if whenever B is a matrix with unit content there exists $\underline{u} \in R^m$ such that $B\underline{u}$ is unimodular. The UCU-property is a strong form of the GCU-property.

THEOREM 133. *Let R be a commutative ring with one in its stable range* (*i.e. if whenever the ideal generated by a and b is R, there exists c in R such that $a + bc$ is a unit*). *Then R is an FC-ring if and only if R has the GCU-property.*

PROOF. Let $\Sigma = (A, B)$ be a reachable linear system over R. There exists $\underline{u} \in R^m$ such that $B\underline{u}$ is unimodular because R has the GCU-property. Since R has one in its stable range then $GL(n)$ is transitive on unimodular columns of R^n and hence Σ is feedback equivalent to a system of the form

$$\left(\left(\begin{array}{cc} 0 & 0\ldots 0 \\ \hline & \\ & * \\ & \end{array}\right), \left(\left(\begin{array}{c|c} 1 & 0\ldots 0 \\ \hline 0 & \\ \vdots & * \\ 0 & \end{array}\right)\right)\right).$$

The result follows by induction, see [12]. □

COROLLARY 134. *A Bezout domain with one in its stable range is an FC-ring.*

Recall that a commutative ring R has many units if for every primitive polynomial $f(T) \in R[T]$ there exists $\alpha \in R$ such that $f(\alpha)$ is a unit.

COROLLARY 135. *A commutative ring R with many units is an FC-ring.*

PROOF. By [89, p. 337] R has one in its stable range. Next we prove that R has the GCU-property. Let $\Sigma = (A, B)$ be a reachable n-dimensional linear system over R, then B has a unimodular vector in its image if and only if the polynomial

$$f(T_1, T_2, \ldots, T_n, U_1, U_2, \ldots, U_m) = (T_1, T_2, \ldots, T_n)B\begin{pmatrix} U_1 \\ \vdots \\ U_m \end{pmatrix},$$

represents a unit in R. Since the coefficients of f are all elements of B it follows that f is primitive because $\mathcal{U}_1(B) = R$. By [89, p. 337] there are units $\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_n$ of R such that $f(\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_n)$ is a unit. Therefore R has the GCU-property and hence, by Theorem 133, R is an FC-ring. □

REMARK 136. The following statements hold, see [89, p. 336]:

(i) Let R be a ring and let $S = \{f(T) \in R[T]\colon\ f(T) \text{ is primitive}\}$. Then the ring of fractions $S^{-1}R[T]$ of $R[T]$ at S has many units.

(ii) $R[[T]]$ is a ring with many units if and only if R has many units. However, for no commutative ring R does $R[T]$ have many units.

(iii) If each R_i for $i \in I$ is a ring with many units then the direct product $\prod_{i\in I} R_i$ and the direct limit $\varinjlim R_i$ (if it exists) have many units.

(iv) A zero-dimensional ring R has many units if and only if $R/\mathfrak{m}$ is infinite for all maximal ideal $\mathfrak{m}$ of R.

(v) The ring $\overline{\mathbb{Z}}$ of all algebraic integers in $\mathbb{C}$ has many units.

The ring R is a local-global ring (or R has the local-global principle), see [43], if whenever $f(T_1, T_2, \ldots, T_n) \in R[T_1, T_2, \ldots, T_n]$ represents a unit in R_p for each prime ideal $\mathfrak{p}$ of R, then $f(T_1, T_2, \ldots, T_n)$ represents a unit in R. In [12] it is proved the following result:

PROPOSITION 137. *A local-global ring is an FC-ring.*

5.3.2. *PA-rings. PS-rings* In [8, Theorem 3.9] the following result is proved:

THEOREM 138. *Let R be a PA-ring and $\Sigma = (A, B)$ a reachable n-dimensional linear system over R. Then $\mathcal{B} = \operatorname{Im} \varphi_B$ contains a rank one summand of R^n. Moreover, if rank one projective R-modules are free then R has the GCU-property.*

REMARK 139. Note that, by Theorem 23, $\operatorname{Im} \varphi_B$ contains a rank one summand of R^n if and only if there exists a matrix V such that

$$\mathcal{U}_i(BV) = \begin{cases} R & \text{for } i = 1, \\ 0 & \text{for } i \geqslant 2. \end{cases}$$

Suppose that $\Sigma = (A, B)$ is a reachable m-input n-dimensional system over a PA-ring R. Let P_1 be a rank one summand of R^n and put $R^n = P_1 \oplus P_2$. Since the composition homomorphism

$$R^m \xrightarrow{\varphi_B} R^n = P_1 \oplus P_2 \xrightarrow{\pi_1} P_1,$$

is surjective there exists a rank one summand P_1' of R^m isomorphic, via φ_B, to P_1. Put $R^m = P_1' \oplus P_2'$. With respect to these decompositions of R^n and R^m one has that B is defined by the matrix of homomorphisms

$$B = \begin{pmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{pmatrix} : P_1' \oplus P_2' \to P_1 \oplus P_2,$$

where b_{11} is an isomorphism. Consider the automorphism of R^n given by the matrix of homomorphisms

$$U = \begin{pmatrix} Id_{P_1} & 0 \\ -b_{21}b_{11}^{-1} & Id_{P_2} \end{pmatrix} : P_1 \oplus P_2 \to P_1 \oplus P_2.$$

Put

$$UAU^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ q_{21} & q_{22} \end{pmatrix} : P_1 \oplus P_2 \to P_1 \oplus P_2,$$

and

$$F = \begin{pmatrix} -b_{11}^{-1}a_{11} & -b_{11}^{-1}a_{12} \\ 0 & 0 \end{pmatrix}.$$

Then $\Sigma = (A, B)$ is feedback equivalent, via the feedback action given by (U, F), to the system (defined by matrices of homomorphisms)

$$\Sigma' = \left(\begin{pmatrix} 0 & 0 \\ a_{21} & a_{22} \end{pmatrix}, \begin{pmatrix} b_{11} & 0 \\ 0 & b_{22} \end{pmatrix} \right).$$

Now, in some cases, it is possible to apply induction considering the system

$$\Gamma = \big(a_{22}, (b_{22}|a_{21}b_{11})\big),$$

which is defined over P_2, see Remark 79.

Note that if P_1 and P_2 are free then Σ' is of the form

$$\left(\left(\begin{array}{cc} 0 & 0 \ldots 0 \\ \hline & \\ & * \\ & \end{array} \right), \left(\begin{array}{c|c} 1 & 0 \ldots 0 \\ \hline 0 & \\ \vdots & * \\ 0 & \end{array} \right) \right).$$

In this way one has the following converse of Theorem 138:

THEOREM 140. *Suppose that rank one projective R-modules are free. Then the following statements are equivalent*:

(i) *R is a PA-ring.*

(ii) *R has the GCU-property.*

(iii) *Each reachable system is feedback equivalent to one of the form*

$$\left(\left(\begin{array}{cc} 0 & 0 \ldots 0 \\ \hline & \\ & * \\ & \end{array} \right), \left(\begin{array}{c|c} 1 & 0 \ldots 0 \\ \hline 0 & \\ \vdots & * \\ 0 & \end{array} \right) \right).$$

PROOF. See [8, Theorem 3.10]. □

COROLLARY 141. *An elementary divisor ring is a PA-ring.*

To study of systems over projective modules suggests the introduction of the following terminology.

DEFINITION 142. Let R be a commutative ring. We say that

(i) R has the coefficient assignability property for projective modules (or R is a *CAP*-ring) if every reachable system over a finitely generated projective R-module is coefficient assignable.

(ii) R has the pole assignability property for projective modules (or R is a *PAP*-ring) if every reachable system over a finitely generated projective R-module is pole assignable.

(iii) R has the pole-shifting property for projective modules (or R is a *PSP*-ring) if res.rk(Σ) poles can be assigned to Σ for every system defined over a finitely generated projective R-module.

REMARK 143. In [58] the *PAP*-rings are called *PA*-rings and our *PA*-rings are called *PAF*-rings.

The following result, see [58, Corollary 12], generalizes Theorem 138.

PROPOSITION 144. *Let M be a finitely generated projective R-module and let $\Sigma = (\varphi, \psi)$ be a pole assignable system over M. Then $\mathcal{B} = \operatorname{Im}\psi$ contains a rank one summand of M. In particular, if R is a PAP-ring then all projective R-modules split into rank one summands.*

Let $\Sigma = (\varphi, \psi)$ be a reachable system over a finitely generated projective R-module M. In Remark 81 it is proved that $\mathcal{B} = \operatorname{Im}\psi$ is a basic submodule of M. Moreover when Σ is pole assignable, by Proposition 144, $\mathcal{B} = \operatorname{Im}\psi$ contains a rank one summand of M. This gives a reason to introduce the next concept.

DEFINITION 145. A commutative ring R has the *BCS*-property (or R is a *BCS*-ring) if every basic submodule of a finitely generated projective R-module M contains a rank one summand of M.

REMARK 146. These rings were introduced in [58] and were called rings that satisfy the property (†). In [12] the *BCS* property is called *UCS* property. Finally the above nomenclature is introduced in [119].

The following result is essential in the pole assignability theory.

THEOREM 147. *If R is a BCS-ring then R is a PAP-ring.*

PROOF. See [58]. □

Note that, by Proposition 144, a *PAP*-ring R is a *BCS*-ring if for each basic submodule $\mathcal{B}_0$ there exists a reachable system of the form $\Sigma = (\varphi, \psi_0)$ with $\mathcal{B}_0 = \operatorname{Im}\psi_0$.

THEOREM 148. *A BCS-ring is a PSP-ring.*

PROOF. See [31]. □

Let $\mathcal{B}_0 = \operatorname{Im}\psi_0$ be a basic submodule of a finitely generated projective R-module M. It is clear that for each A the system $\Sigma = (A, \psi_0)$ satisfies $\operatorname{res.rk}(\Sigma) \geqslant 1$. The conjecture is: A *PSP*-ring is a *BCS*-ring. In [31] it is proved that the conjecture is true when R is a Prüfer domain.

PROPOSITION 149. *The following rings are BCS-rings*:

(i) *Elementary divisor rings.*
(ii) *Dedekind domains.*
(iii) *Both the rings of continuous and* $\mathfrak{C}^\infty$ *real valued functions on* X*, where* X *is a connected* 1*-dimensional manifold.*
(iv) *The ring* $\mathfrak{C}^0(X;\mathbb{C})$ *of continuous complex valued functions over a connected manifold* X*, with* $\dim(X) \leqslant 2$.
(v) 0*-dimensional rings.*
(vi) *Semilocal rings.*
(vii) *The polynomial ring* $V[x]$ *where* V *is a semilocal principal ideal domain.*
(viii) 1*-dimensional domains.*
(ix) 1*-dimensional Noetherian rings.*
(x) *Rings that have the UCU-property.*

PROOF. Statements (i) and (x) are clear. For (ii) and (iii) see [58]. For (iv) see [122]. For (v), (vi), (vii), (viii) and (ix) see [119]. □

The next result is the natural generalization of Theorem 140.

PROPOSITION 150. *Let* R *be a ring. The following statements are equivalents*:

(i) R *is a BCS-ring and projective* R*-modules of finite rank are free.*
(ii) *For every matrix* B *with* $\mathcal{U}_1(B) = R$ *there exist invertible matrices* P *and* Q *such that*

$$PBQ = \begin{pmatrix} 1 & \underline{0}^t \\ \underline{0} & B_1 \end{pmatrix}.$$

(iii) *Each linear system* Σ *with* $\operatorname{res.rk}(\Sigma) \geqslant 1$ *is feedback equivalent to one on the form*

$$\left(\left(\begin{array}{ccc} 0 & \cdots & 0 \\ \hline & & \\ & * & \\ & & \end{array} \right), \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & * & \\ 0 & & & \end{array} \right) \right).$$

As a consequence one has:

THEOREM 151. *Let R be a Bezout domain, then the following statements are equivalent*:

(i) *R is a PSP-ring.*

(ii) *R is a BCS-ring.*

(iii) *R is an elementary divisor domain.*

References

[1] T. Akiba, *Integrally-closedness of polynomial rings*, J. Math. Soc. Japan **6** (1980), 67–75.

[2] G. Almkvist, *Endomorphisms of finitely generated modules over a commutative ring*, Ark. Math. **2** (1973), 263–301.

[3] N. Bourbaki, *Éléments de Mathématique. Algèbre (Chap. 1–3),* Hermann, Paris (1970).

[4] N. Bourbaki, *Éléments de Mathématique. Algèbre Commutative (Chap. 8–9),* Masson, Paris (1983).

[5] N. Bourbaki, *Éléments de Mathématique. Algèbre Commutative (Chap. 1–4),* Masson, Paris (1985).

[6] N. Bourbaki, *Éléments de Mathématique. Algèbre Commutative (Chap. 5–7)*, Masson, Paris (1985).

[7] N. Bourbaki, *Éléments de Mathématique. Topologie Genérale*, Hermann, Paris (1971).

[8] J.W. Brewer, J.W. Bunce and F.S. Van Vleck, *Linear Systems over Commutative Rings*, Marcel Dekker, New York (1986).

[9] J.W. Brewer, T. Ford, L. Klinger and W. Schmale, *When does the ring $K[y]$ have the coefficient assignment property?*, J. Pure Appl. Algebra **112** (1996), 239–246.

[10] J.W. Brewer, W. Heinzer and D. Lantz, *The pole assignability property in polynomial rings over GCD-domains*, J. Algebra **97** (1985), 461–466.

[11] J.W. Brewer, D. Katz and W. Ullery, *On the pole assignability property over commutative rings*, J. Pure Appl. Algebra **48** (1987), 1–7.

[12] J.W. Brewer, D. Katz and W. Ullery, *Pole assignability in polynomial rings, power series rings, and Prüfer domains*, J. Algebra **106** (1987), 265–286.

[13] J.W. Brewer and L. Klingler, *Dynamic feedback over commutative rings*, Linear Algebra Appl. **98** (1988), 137–168.

[14] J.W. Brewer and L. Klingler, *Pole shifting for families of systems: The dimension one case*, Math. Control Signals Systems **1** (3) (1988), 285–292.

[15] J.W. Brewer and L. Klingler, *Dynamic feedback over rings whose homomorphic images are local-global*, Comm. Algebra **17** (7) (1989), 1759–1773.

[16] J.W. Brewer and L. Klinger, *On feedback invariants for linear dynamical systems*, Linear Algebra Appl. **325** (2001), 209–220.

[17] J.W. Brewer, L. Klinger and W. Schmale, *The dynamic feedback cyclization problem for principal ideal domains*, J. Pure Appl. Algebra **92** (1994), 213–225.

[18] J.W. Brewer, L. Klinger and W. Schmale, *$C[y]$ is a CA-ring and coefficient assignment is properly weaker that feedback cyclization over a PID*, J. Pure Appl. Algebra **97** (1995), 265–273.

[19] J.W. Brewer, C. Naudé and G. Naudé, *On Bezout domains, elementary divisor rings, and pole assignability*, Comm. Algebra **12** (1984), 2987–3003.

[20] J.W. Brewer and W. Schmale, *(A, B)-cyclic submodules*, Linear Algebra Appl. **301** (1999), 63–80.

[21] P.A. Brunovsky, *A classification of linear controllable systems*, Kybernetika **3** (1970), 173–187.

[22] D.A. Buchsbaum and D. Eisenbud, *What annilates a module?*, J. Algebra **47** (2) (1968), 231–243.

[23] D.A. Buchsbaum and D. Eisenbud, *What makes a complex exact?*, J. Algebra **25** (1973), 259–268.

[24] D.A. Buchsbaum and D. Eisenbud, *Some structure theorems for finite free resolutions*, Adv. Math. **12** (1974), 84–139.

[25] R. Bumby, E.D. Sontag, H.J. Sussmann and W. Vasconcelos, *Remarks on the pole-shifting problem over rings*, J. Pure Appl. Algebra **20** (1981), 113–127.

[26] L. Burch, *On ideals of finite homological dimension in local rings*, Proc. Cambridge Philos. Soc. **64** (1968), 941–946.

[27] P. Camion, L.S. Levi and H.B. Mann, *Linear equations over a commutative ring*, J. Algebra **19** (1972), 432–446.

[28] A. Campillo-López, T. Sánchez-Giralda, *Finitely generated projective modules and Fitting ideals*, Collect. Math. **30** (2) (1979), 3–8.

[29] M. Carriegos and J.A. Hermida-Alonso, *Canonical forms for single input linear systems*, Systems Control Lett., to appear.

[30] M. Carriegos, J.A. Hermida-Alonso and T. Sánchez-Giralda, *The pointwise feedback relation for linear dynamical systems*, Linear Algebra Appl. **279** (1998), 119–134.

[31] M. Carriegos, J.A. Hermida-Alonso and T. Sánchez-Giralda, *Pole-shifting for linear systems over commutative rings*, Linear Algebra Appl. **346** (2002), 97–107.

[32] M. Carriegos and T. Sánchez-Giralda, *Canonical forms for linear dynamical systems over commutative rings: The local case*, Lecture Notes in Pure Appl. Math., Vol. 221, Springer-Verlag, Berlin (2001), 113–133.

[33] J.L. Casti, *Linear Dynamical Systems*, Academic Press, New York (1987).

[34] W.S. Ching and B.F. Wyman, *Duality and the regulator problem for linear systems over commutative rings*, J. Comput. System Sci. **14** (3) (1977), 360–368.

[35] S. Cox and R. Pendleton, *Rings for which certain flat modules are projective*, Trans. Amer. Math. Soc. **104** (1970), 52–61.

[36] J. Dieudonné, *Abrege d'histoire des Mathematiques 1700–1900*, Hermann, Paris (1978).

[37] D. Eisenbud, *A survey of some results on finite free resolutions*, Proceedings of International Congress of Math. (1974).

[38] R. Eising, *Pole assignment for systems over rings,* Systems Control Lett. **2** (1982), 225–229.

[39] V.P. Elizarov, *Systems of equations over a commutative ring*, Russian Math. Surveys (1992), 175–177.

[40] V.P. Elizarov, *General solution of a system of homogeneous linear equations over a commutative ring*, Russian Math. Surveys (1993), 153–154.

[41] E. Emre and P. Kargonekar, *Regulation of splits linear systems over rings: Coefficient assignment and observers*, IEEE Trans. Autom. Control **27** (1982), 104–113.

[42] S. Endo, *On semi-hereditary rings*, J. Math. Soc. Japan **13** (1961), 109–119.

[43] D. Estes and R. Guralnick, *Module equivalence: Local to global when primitive polynomials represent units*, J. Algebra **77** (1982), 138–157.

[44] D. Estes and J. Ohm, *Stable range in commutative rings*, J. Algebra **7** (3) (1967), 343–362.

[45] E.G. Evans and P. Griffith, *Syzygies*, London Math. Soc. Lecture Note Series, Vol. 105, Springer-Verlag, Berlin (1985).

[46] H. Fitting, *Die Determinantenideale Moduls*, Sahresberichtd Deutschen Mathem. Vereinigung **46** (1936), 195–228.

[47] J. Ferrer, M.I. Garcia and F. Puerta, *Brunovsky local form of an holomorphic family of pairs of matrices*, Linear Algebra Appl. **253** (1997), 175–198.

[48] A. deFrancisco and J.A. Hermida-Alonso, *Characterizing when the rings R and $R[T]$ are integrally closed via linear equations over R*, J. Algebra **217** (1999), 628–649.

[49] A. deFrancisco and J.A. Hermida-Alonso, *A criterion for exactness of a complex of length two*, J. Algebra, to appear.

[50] L. Gillman and M. Jerison, *Ring of Continuous Functions,* Springer-Verlag, Berlin (1976).

[51] O. Goldman, *Determinants in projective modules*, Nagoya Math. J. **18** (1961), 27–36.

[52] G. Golub, B. Kagström and P. Van Dooren, *Direct block tridiagonalization of single-input single-output systems,* Systems Control Lett. **18** (1992), 109–120.

[53] R.M. Guralnick, *Similarity of matrices over local rings*, Linear Algebra Appl. **41** (1981), 161–172.

[54] R.M. Guralnick, *A note on the local principle for similarity of matrices*, Linear Algebra Appl. **30** (1980), 241–245.

[55] M. Hazewinkel, *A partial survey of the uses of algebraic geometry in systems and control theory*, Proc. Severi Centennial Conference, Rome (1979).

[56] M. Hazewinkel and A.-M. Perdon, *On families of systems: pointwise-local-global isomorphism problem*, Int. J. Control **33** (1981), 713–726.

[57] M. Hazewinkel, *Lectures on Invariants, Representations and Lie Algebras in Systems and Control Theory,* Lecture Notes in Math., Vol. 1029, Springer-Verlag, Berlin (1983).

[58] M.L.J. Hautus and E.D. Sontag, *New results on pole-shifting for parametrized families of systems*, J. Pure Appl. Algebra **40** (1986), 229–244.

[59] J.A. Hermida-Alonso, *Teoría de Ideales de Fitting y Clases de Chern*, Mem. y Monogr. Math. Inst. Jorge Juan C.S.I.C., Vol. XXIX, Madrid (1983).

[60] J.A. Hermida-Alonso, *Linear equations over commutative rings and grade theory*, J. Algebra **148** (2) (1992), 497–503.

[61] J.A. Hermida-Alonso, *Linear equations over commutative rings and finite free resolutions*, J. Algebra **164** (2) (1994), 452–446.

[62] J.A. Hermida-Alonso, M.P. Pérez and T. Sánchez-Giralda, *Feedback invariants for linear dynamical systems over a principal ideal domain*, Linear Algebra Appl. **218** (1995), 29–45.

[63] J.A. Hermida-Alonso, M.P. Pérez and T. Sánchez-Giralda, *Brunovsky's canonical form for linear dynamical systems over commutative rings*, Linear Algebra Appl. **233** (1996), 131–147.

[64] J.A. Hermida-Alonso and M. Pisonero, *Invariant factors of an endomorphism and finite free resolutions*, Linear Algebra Appl. **187** (1993), 201–206.

[65] J.A. Hermida-Alonso and M. Pisonero, *Cyclic endomorphisms of a module*, Bull. Belg. Math. Soc., Ser. A **45** (1993), 173–181.

[66] J.A. Hermida-Alonso and M. Pisonero, *What polynomials satisfies a given endomorphism?*, Linear Algebra Appl. **216** (1995), 125–138.

[67] J.A. Hermida-Alonso and T. Sánchez-Giralda, *Linear equations over commutative rings and determinantal ideals,* J. Algebra **99** (1) (1986), 72–79.

[68] J.A. Hermida-Alonso and T. Sánchez-Giralda, *Sur les anneaux de Prüfer*, Travaux en Cours **22** (1987), 301–307.

[69] J.A. Hermida-Alonso and T. Sánchez-Giralda, *Some criteria for solvability of systems of linear equations over modules,* Lecture Notes in Math., Vol. 1328, Springer-Verlag, Berlin (1988), 122–135.

[70] J.A. Hermida-Alonso and T. Sánchez-Giralda, *Finite free resolutions of length two and Koszul generalized complex*, Comm. Algebra **22** (2) (1994), 579–594.

[71] M. Heymann, *Comments on pole assignment in multi-input controllable systems*, IEEE Trans. Autom. Contr. **13** (1969), 748–749.

[72] D. Hilbert, *Über die Theory der algebraischen Formen*, Math. Ann. **36** (1890), 473–534.

[73] J.A. Huckaba, *Commutative Rings with Zero Divisors,* Marcel Dekker, New York (1988).

[74] N. Ito and H. Inaba, *Dynamic feedback (A, B)-invariant submodules for linear systems over commutative Noetherian domains,* Linear Algebra Appl. **282** (1998), 123–129.

[75] R.E. Kalman, *Lectures on Controllability and Observability,* C.I.M.E. Summer Course (1968).

[76] R.E. Kalman, *Kronecker invariants and feedback*, Ordinary Differential Equations, Academic Press, New York (1972), 459–471.

[77] R.E. Kalman, P. Falb and M. Arbib, *Topics in Mathematical Systems Theory*, McGraw-Hill, New York (1969).

[78] E.W. Kamen, *Lectures on algebraic system theory: linear systems over rings*, NASA Contractor Report 3016 (1978).

[79] E.W. Kamen, *Block-form control of linear time-invariant discrete-time systems defined over commutative rings*, Linear Algebra Appl. **205** (1994), 805–829.

[80] I. Kaplansky, *Elementary divisors and modules*, Trans. Amer. Math. Soc. **66** (1949), 464–491.

[81] J.P. Lafon, *Les Formalismes Fondamentaux de l'Algebre Commutative*, Hermann, Paris (1974).

[82] J.P. Lafon, *Ideals and modules*, Handbook of Algebra, Vol. 1, M. Hazewinkel, ed., Elsevier, Amsterdam (1996), 706–727.

[83] T.Y. Lam, *Serre's Conjecture*, Lecture Notes in Math., Vol. 635, Springer-Verlag, Berlin (1978).

[84] M.D. Larsen, W.J. Lewis and T.S. Shores, *Elementary divisor rings and finitely presented R-modules*, Trans. Amer. Math. Soc. **187** (1974), 231–248.

[85] T. Lucas, *Characterizing when $R[X]$ is integrally closed*, Proc. Amer. Math. Soc. **105** (1989), 861–867.

[86] T. Lucas, *Characterizing when $R[X]$ is integrally closed. II,* J. Pure Appl. Algebra **61** (1989), 49–52.

[87] R.E. MacRae, *An application of the Fitting invariants*, J. Algebra **2** (1965), 153–169.

[88] N.H. McCoy, *Rings and Ideals*, Carus Math. Monogr., Vol. 8. Math. Assoc. America (1948).

[89] B.R. McDonald, *Linear Algebra over Commutative Rings*, Marcel Dekker, New York (1984).

[90] B.R. McDonald, *Similarity of matrices over Artinian principal ideal rings*, Linear Algebra Appl. **21** (1978), 153–162.
[91] F. Minnaar and G. Naudé, *On pole assignability and feedback cyclisation for systems over rings of finite dimension*, Quaestiones Math. **12** (1988), 13–24.
[92] A.S. Morse, *Ring models for delay differential systems*, Automatica **12** (1976), 529–531.
[93] C.G. Naudé and G. Naudé, *Comments on pole assignability over rings*, Systems Control Lett. **6** (1985), 113–115.
[94] M. Newman, *Integral Matrices*, Pure and Applied Math., Academic Press, New York (1972).
[95] D.G. Northcott, *Finite Free Resolutions*, Cambridge University Press, Cambridge, MA (1976).
[96] M. Pisonero, *Determinants on modules that have a finite free resolution*, Linear and Multilinear Algebra **38** (1994), 97–107.
[97] Y. Quentel, *Sur la compacité du spectre minimal d'un anneau*, Bull. Soc. Math. France **99** (1971), 265–272.
[98] D. Quillen, *Projective modules over polynomial rings*, Invent. Math. **36** (1976), 167–171.
[99] Y. Rouchaleau, *Linear, discrete time, finite dimensional systems over some classes of commutative rings*, PhD thesis, Stanford (1972).
[100] Y. Rouchaleau and B.F. Wyman, *Linear dynamical systems over integral domain*, J. Comput. System Sci. **9** (2) (1974), 129–142.
[101] A. Sáez-Schwedt and T. Sánchez-Giralda, *Canonical forms for* 2*-dimensional linear systems over commutative rings*, Comm. Algebra, to appear.
[102] T. Sánchez-Giralda, *Exactitud de complejos y sistemas de ecuaciones*, Rev. Real Acad. Cienc. Exact., Fis. Nat. **82** (1988), 165–167.
[103] T. Sánchez-Giralda and A. Campillo, *Finitely generated projective modules and Fitting ideals*, Collect. Math. **30** (2) (1979).
[104] A. Seidenberg, *Constructions in algebra*, Trans. Amer. Math. Soc. **197** (1974), 273–313.
[105] P. Sharma, *On pole assignment problem in polynomial rings*, Systems Control Lett. **5** (1984), 49–54.
[106] P. Sharma, *Some results on pole-placement and reachability*, Systems Control Lett. **6** (1986), 325–328.
[107] D.W. Sharpe, *Grade and the theory of linear equations*, Linear Algebra Appl. **18** (1977), 25–32.
[108] H.J. Smith, *Collected Mathematical Papers*, Vol. 2, Oxford (1894).
[109] E. Sontag, *Linear systems over commutative rings: A survey*, Ric. di Automatica **7** (1976), 1–34.
[110] E. Sontag, *An introduction to the stabilization problem for parametrized families of linear systems*, Contemp. Math. **47** (1985), 369–400.
[111] E.D. Sontag, *Mathematical Control Theory*, Springer-Verlag, Berlin (1990), 2nd edn (1998).
[112] E. Steinitz, *Rechteckige Systeme und Moduln in algebraischen Zahlköper,* Math. Ann. **71** (1912), 328–354.
[113] A.A. Suslin, *Projective modules over a polynomial ring are free,* Dokl. Akad. Nauk SSSR **229** (1976), 1160–1164.
[114] A. Tannenbaum, *On pole assignability over polynomial rings*, Systems Control Lett. **2** (1982), 13–16.
[115] A. Tannenbaum, *Polynomial rings over arbitrary fields in two or more variables are not pole assignable*, Systems Control Lett. **2** (1982), 222–224.
[116] K. Tchon, *On structural instability of normal forms of affine control systems subject to static state feedback*, Linear Algebra Appl. **121** (1989), 95–104.
[117] W.V. Vasconcelos, *On finitely generated flat modules*, Trans. Amer. Math. Soc. (1969), 505–512.
[118] W.V. Vasconcelos, *Annihilators of modules with a finite free resolution*, Proc. Amer. Math. Soc. **9** (1971), 440–442.
[119] W.V. Vasconcelos and C.A. Weibel, *BCS rings*, J. Pure Appl. Algebra **52** (1988) 173–185.
[120] R.B. Warfield, *Stable equivalence of matrices and resolutions*, Comm. Algebra **6** (17) (1978), 1811–1828.
[121] R.B. Warfield, *Cancellation of modules and groups and stable range of endomorphism rings*, Pacific J. Math. **91** (1980), 457–485.
[122] C.A. Weibel, *Complex continuous functions on surfaces are BCS rings,* Systems Control Lett. (1988).
[123] R. Wiegand and S. Wiegand, *Commutative rings whose finitely generated modules are direct sums of cyclics*, J. Algebra **32** (1974), 152–172.
[124] W.A. Wonham, A.S. Morse, *Feedback invariants of linear multivariable systems,* Automatica **8** (1972), 33–100.

This Page Intentionally Left Blank

Correction and Addition

Leiba Rodman
Department of Mathematics, College of William and Mary, Williamsburg, VA 23187-8795, USA
E-mail: lxrodm@math.wm.edu

It has been pointed out to me by several colleagues that the statement that appears in [3], on p. 136, line 6 down, is incorrect. The ring $\mathbb{Z}[\lambda]$ of polynomials in one variable λ with integer coefficients is not a Bezout domain (in other words, not every finitely generated ideal of $\mathbb{Z}[\lambda]$ is principal), and therefore a fortiori cannot be a Smith domain. For example, the ideal of $\mathbb{Z}[\lambda]$ generated by the constant 3 and by λ is not principal (see, e.g., Section 25.3 in [1]).

An example of a Smith domain which is not a PID is provided by the ring $\mathcal{A}(\Omega)$ of complex valued analytic functions defined on a fixed open set Ω in the complex plane, with pointwise algebraic operations. Let $\Omega_0 \subset \Omega$ be a *discrete* set, i.e. a countable set of points in Ω with no accumulation points in Ω itself. Consider the set $S(\Omega_0)$ of functions $f \in \mathcal{A}(\Omega)$ with the property that $f(z_0) = 0$ for every $z_0 \in \Omega_0 \setminus \Omega(f)$, where $\Omega(f)$ is a finite set (which depends on f). It is easy to see, using the Weierstrass theorem on existence of analytic functions on Ω having a prescribed discrete set of zeros, that $S(\Omega_0)$ is an ideal, but not a principal ideal, of $\mathcal{A}(\Omega)$. It follows from Theorem 3.3 in [2] that $\mathcal{A}(\Omega)$ is a Smith domain.

I thank all who brought the incorrect statement to my attention.

References

[1] N. Bloch, *Abstract Algebra with Applications*, Prentice-Hall, Englewood Cliffs, NJ (1987).
[2] I. Gohberg and L. Rodman, Analytic matrix functions with prescribed local data, *J. d'Analyse Math.* **40** (1981), 90–128.
[3] L. Rodman, *Matrix functions*, Handbook of Algebra, Vol. 1, M. Hazewinkel, ed., Elsevier, Amsterdam (1996), 117–154.

This Page Intentionally Left Blank

Section 2A
Category Theory

This Page Intentionally Left Blank

Monads of Sets

Ernie Manes

Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, USA

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1. Introduction

Monads have been studied for about forty years. They first surfaced to codify resolutions for sheaf cohomology. Today they are established as a standard concept of category theory. Recently, monad research has been shaped by applications to functional programming languages.

The focus of this chapter is on monads of sets – monads in the category $\mathcal{S}$ of sets and (total) functions.

The larger study of monads in a general category puts the theory of monads of sets in proper perspective. The goal of this chapter, then, is a brief, but balanced account of monads of sets in its current incipient stage, with the hope of stimulating further research. I apologize in advance to those whose important contributions have been omitted or inadequately treated owing to my ignorance or to lack of space.

2. Origins

Eilenberg and MacLane wrote the founding paper of category theory [19] in 1945. When Cartan and Eilenberg's book [12] on homological algebra was published eleven years later, a tentative category-theoretic foundation appeared only as a short appendix by D. Buchsbaum called "Exact Categories", the forerunner of Abelian categories. In the introduction to MacLane's book [72] on homology published in 1963, he refers to the fact that algebraic systems and their homomorphisms constitute a "category", with the quotes being his, and he does not mention Abelian categories until p. 254. Such caution, even on the part of the founders, is a clear indication that categorical foundations were not widely accepted by the mid 1960s the way they would be today.

2.1. *Godement*

Simplicial objects in a category were introduced by [21]. For textbook accounts relating simplicial objects to homotopy, homology and monads see [29], [72] and, especially, [73, Chapter VII]. In a 1964 book [32] on homological algebra and sheaf theory, Godement provided (but only, once again, in a brief appendix) a category-theoretic foundation based on monads for the cohomology of an object in a category. In a third section on "the fundamental construction" of this appendix, he proposed the striking idea that underlying many of the constructions of the body of the text were simplicial objects in the category $[\mathcal{K} \to \mathcal{K}]$ of endofunctors of the category $\mathcal{K}$. The objects of this category are the functors $\mathcal{K} \to \mathcal{K}$ and the morphisms between such functors are, not surprisingly, the natural transformations between them. The identity morphism of F is the identity natural transformation $id_F : F \to F$ whose Xth component is id_{FX} and composition of natural transformations is pointwise, $(\sigma \circ \tau)_X = \sigma_X \tau_X$.

In modern parlance, $[\mathcal{K} \to \mathcal{K}]$ is more than a category, it is a 2-*category*. The main idea is that there are two kinds of composition. The first one is the category composition which we have written $\circ$, the pointwise composition of natural transformations as has already been described. This is called *horizontal composition*. The second composition (which will be written as juxtaposition) is called *vertical composition* and is defined as follows for $\sigma : F \to F'$, $\tau : G \to G'$.

$$(GF \xrightarrow{\tau\sigma} G'F')_X = (\tau_{F'X})(G\sigma_X) = (G'\sigma_X)(\tau_{FX}) \tag{1}$$

The two versions are equivalent because τ is natural.

The relationship between horizontal and vertical composition is summed up by what Godement called "the rules of the functorial calculus". To state the rules we use upper case Latin letters for functors $\mathcal{K} \to \mathcal{K}$ and Greek letters for natural transformations between such functors. For $\sigma : F \to G$, $H\sigma : HF \to HG$ is the natural transformation $(H\sigma)_X = H(\sigma_X)$ and $\sigma K : FK \to GK$ is the natural transformation $(\sigma K)_X = \sigma_{KX}$. For $\sigma : F \to G$, $\tau : G \to M$ the rules are

$$\begin{aligned} &(\sigma L)K = \sigma(LK) \\ &(ED)\sigma = E(D\sigma) \\ &(D\sigma)H = D(\sigma H) \\ &H(\tau \circ \sigma)D = (H\tau D) \circ (H\sigma D) \end{aligned} \tag{2}$$

Godement's idea was to start with a functor $T : \mathcal{K} \to \mathcal{K}$ equipped with two natural transformations $\eta : id \to T$ and $\mu : TT \to T$ and then to define the data for an (unaugmented) simplicial object in $[\mathcal{K} \to \mathcal{K}]$ for the sequence $T, T^2 = TT, T^3, \ldots$ of iterates of T with faces and degeneracies as follows:

$$\begin{aligned} &d_n^i = T^i \mu T^{n-i} : T^{n+1} \to T^n \quad (n > 0,\ 1 \leqslant i \leqslant n) \\ &s_n^i = T^i \eta T^{n-i} : T^n \to T^{n+1} \quad (n \geqslant 0,\ 0 \leqslant i \leqslant n) \end{aligned} \tag{3}$$

Godement then showed that this data satisfies the simplicial identities if and only if the following three diagrams commute:

$$\begin{tikzcd} T \arrow[r, "T\eta"] \arrow[dr, "id"'] & TT \arrow[d, "\mu"] & T \arrow[l, "\eta T"'] \arrow[dl, "id"] \\ & T & \end{tikzcd} \qquad \begin{tikzcd} TTT \arrow[r, "T\mu"] \arrow[d, "\mu T"'] & TT \arrow[d, "\mu"] \\ TT \arrow[r, "\mu"'] & T \end{tikzcd} \tag{4}$$

This leads us to the main definition of this chapter.

DEFINITION 2.1. A *monad* in a category $\mathcal{K}$ is $\mathbf{T} = (T, \eta, \mu)$ where $T : \mathcal{K} \to \mathcal{K}$ is a functor and $\eta : id \to T$, $\mu : TT \to T$ are natural transformations subject to the three commutative diagrams of (4) above.

We will generally use boldface unitalicized letters such as **S** for the monad and then S for the corresponding functor. As a rule the generic symbols η, μ will be used for all monads. Another monad would be, say, $\mathbf{Group} = (Group, \eta, \mu)$.

The term "monad" surfaced almost immediately in 1967 in [6, Definition 5.4.1, p. 39]. MacLane's influential [73] popularized it.

2.2. *Monads as monoids*

To most mathematicians, a monoid is a set X with an associative binary operation m equipped with a two-sided unit u. Category theorists [73, Chapter VII] see this as the special case of a monoid in the category $\mathcal{S}$ of sets and functions with Cartesian product as the tensor and the terminal object (1-element set) as the tensor unit. From this viewpoint, X is an object of $\mathcal{S}$, and m and u are morphisms

$$m : X \times X \to X, \qquad u : 1 \to X$$

The universal mapping property of the product induces a natural isomorphism

$$\alpha_{XYZ} : (X \times Y) \times Z \to X \times (Y \times Z)$$

thus a monoid (X, m, u) is defined by three commutative diagrams

$$\begin{tikzcd}
X \cong X \times 1 \arrow[r, "id \times m"] \arrow[dr, "id"'] & X \times X \arrow[d, "m"] & 1 \times X \cong X \arrow[l, "m \times id"'] \arrow[dl, "id"] \\
& X &
\end{tikzcd}$$

$$\begin{tikzcd}
(X \times X) \times X \cong X \times (X \times X) \arrow[r, "id \times m"] \arrow[d, "m \times id"'] & X \arrow[d, "m"] \\
X \arrow[r, "m"'] & X
\end{tikzcd}$$

To make the analogy with the diagrams of (4) precise, make the following observations. $(X, Y) \mapsto X \times Y$ is a functor $\mathcal{S} \times \mathcal{S} \to \mathcal{S}$ (where, on morphisms, $(f \times g)(x, y) = (fx, gy)$). Also, $F,\ G \mapsto GF$ is a functor $[\mathcal{K} \to \mathcal{K}] \times [\mathcal{K} \to \mathcal{K}] \to [\mathcal{K} \to \mathcal{K}]$ for any category $\mathcal{K}$ which, on morphisms, maps σ, τ to its vertical composition (1) $\sigma\tau$. Functoriality is verified using the Godement rules (2).

Now 1 is a unit for the bifunctor $X \times Y$ in $\mathcal{S}$ in that there are natural isomorphisms $1 \times X \cong X \cong X \times 1$. Even better, the identity functor $id : \mathcal{K} \to \mathcal{K}$ is a unit for composition on the nose, $id\ F = F = F\ id$. Where $X \times Y$ has an associativity isomorphism, functor composition is associative on the nose: $H(GF) = (HG)F$. To wrap it up, if Cartesian product in $\mathcal{S}$ is replaced with composition of functors, a monoid in $[\mathcal{K} \to \mathcal{K}]$ is just a monad in $\mathcal{K}$. Monads are monoids! For much more detail and a larger framework see [73, Chapter VII].

This observation makes the definition of a monad seem more natural. It is the higher level of structure of $[\mathcal{K} \to \mathcal{K}]$, being a 2-category not just a category, which allows the relative complexity of the simplicial identities from monoid axioms. Nothing like this is possible for ordinary monoids.

An additional advantage of looking at monads as monoids is that it offers a natural definition of morphism between monads, since "monoid homomorphism" is a clear concept. So here is that definition.

DEFINITION 2.2. If $\mathbf{S} = (S, \eta, \mu)$, $\mathbf{T} = (T, \eta, \mu)$ are monads in $\mathcal{K}$, a *morphism of monads* or *monad map* is a natural transformation $\lambda : S \to T$ such that the following diagram commutes

$$\begin{tikzcd} id \arrow[r, "\eta"] \arrow[dr, "\eta"'] & S \arrow[d, "\lambda"] & SS \arrow[l, "\mu"'] \arrow[d, "\lambda\lambda"] \\ & T & TT \arrow[l, "\mu"] \end{tikzcd}$$

Note that the corresponding diagram for ordinary monoids produces ordinary monoid homomorphisms. This definition was first given by [82].

EXAMPLE 2.3. There is at least one monad in a given category $\mathcal{K}$, namely the *identity monad* $id = (id, \eta, \mu)$ where $id : \mathcal{K} \to \mathcal{K}$ is the identity functor, $\eta_X = id_X$, $\mu_X = id_X$. For any monad $\mathbf{T}$, $\eta : \boldsymbol{id} \to \mathbf{T}$ is a monad map.

2.3. *Huber*

In the introduction to [38], Huber states

> "Our main tool will be the semisimplicial standard construction, originally devised by R. Godement to generate flabby resolutions in the category of sheaves. Since also the Hochschild homology theory of associative algebras and, moreover, the whole theory of derived functors in the categories of modules may be obtained with the aid of standard constructions, these turn out to be one of the most powerful tools of homological and homotopical algebra".

Here the term "standard construction" refers to what we would today call a *comonad*.

DEFINITION 2.4. A *comonad* in a category $\mathcal{K}$ is a monad $\mathbf{G} = (G, \varepsilon, \delta)$ in the opposite category $\mathcal{K}^{op}$. Thus, in terms of $\mathcal{K}$, $G : \mathcal{K} \to \mathcal{K}$ is a functor and $\varepsilon : G \to id$ and $\delta : G \to GG$ are natural transformations subject to

$$\begin{tikzcd}
G & GG \arrow[l, "G\varepsilon"'] \arrow[r, "\varepsilon G"] & G \\
& G \arrow[ul, "id"] \arrow[u, "\delta"] \arrow[ur, "id"'] &
\end{tikzcd}
\qquad
\begin{tikzcd}
GGG & GG \arrow[l, "G\delta"'] \\
GG \arrow[u, "\delta G"] & G \arrow[l, "\delta"] \arrow[u, "\delta"']
\end{tikzcd}$$

We leave it to the reader to draw the diagrams for comonad maps.

It is time to ask where monads (or, equivalently, comonads) come from. Huber's paper is best remembered today because of its fourth section which shows how monads and comonads arise from pairs of adjoint functors; curiously, this work was not mentioned in the introduction to the paper.

Adjoint functors were introduced by Kan [47] in 1958. This concept is central to monad theory so it is best that we review the definition and establish notation. Algebraists are familiar with a idea of an algebra of a certain type (such as group, ring or lattice) being freely generated by a set, so we will use this concept to characterize adjointness.

DEFINITION 2.5. Let $U : \mathcal{A} \to \mathcal{K}$ be a functor. A *free $\mathcal{A}$-object generated by* a $\mathcal{K}$-object K is a pair (A, η) with A an $\mathcal{A}$-object and $\eta : K \to UA$ all with the following universal mapping property: $\forall\, f : K \to UB\ \exists!\, \psi : A \to B$ such that $(U\psi)\eta = f$. Here, $\exists!$ means "there exists a unique…".

$$\begin{tikzcd}
K \arrow[rr, "\eta"] \arrow[dr, "f"'] & & UA \arrow[dl, "U\psi"] \\
& UB &
\end{tikzcd}$$

It is evident that if (B, f) also has this universal mapping property then ψ is an isomorphism. Thus free objects are unique when they exist.

DEFINITION 2.6. An *adjointness* $(\mathcal{A}, \mathcal{K}, U, F, \eta, \varepsilon)$ is given by a pair of functors $U : \mathcal{A} \to \mathcal{K}$, $F : \mathcal{K} \to \mathcal{A}$ together with natural transformations $\eta : id \to UF$, $\varepsilon : FU \to id$ such that the following two equations of natural transformations hold:

$$\begin{tikzcd}
U \arrow[r, "\eta U"] \arrow[dr, "id"'] & UFU \arrow[d, "U\varepsilon"] \\
& U
\end{tikzcd}
\qquad
\begin{tikzcd}
F \arrow[r, "F\eta"] \arrow[dr, "id"'] & FUF \arrow[d, "\varepsilon F"] \\
& F
\end{tikzcd}$$

We say F is *left adjoint to* U and that U is *right adjoint to* F.

There is an immediate duality principle since U, F are also functors $U : \mathcal{A}^{op} \to \mathcal{K}^{op}$, $F : \mathcal{K}^{op} \to \mathcal{A}^{op}$ with the same η, ε but now considered as morphisms in the opposite categories so that now U has F as a *right* adjoint. Hence if $(\mathcal{A}, \mathcal{K}, U, F, \eta, \varepsilon)$ is an adjointness, so is $(\mathcal{K}^{op}, \mathcal{A}^{op}, F, U, \varepsilon, \eta)$

The following is standard category theory, but we sketch the construction because of its importance.

PROPOSITION 2.7. *For a functor $U : \mathcal{A} \to \mathcal{K}$, the following two statements are equivalent.*

1. *Every $\mathcal{K}$-object generates a free $\mathcal{A}$-object.*
2. *There exists an adjointness $(\mathcal{A}, \mathcal{K}, U, F, \eta, \varepsilon)$.*

PROOF. $(1 \Rightarrow 2)$ Let (FK, η_X) be free over K. The following commutative square defines Ff by the universal property:

$$\begin{tikzcd} K \arrow[r, "\eta_K"] \arrow[d, "f"'] & UFK \arrow[d, "UFf"] \\ L \arrow[r, "\eta_L"'] & UFL \end{tikzcd} \tag{5}$$

It is routine to check using the uniqueness of Ff that F is a functor and then η is natural. The universal property also defines ε:

$$\begin{tikzcd} UA \arrow[r, "\eta_{UA}"] \arrow[dr, "id"'] & UFUA \arrow[d, "U\varepsilon_A"] \\ & UA \end{tikzcd} \tag{6}$$

The remaining details that an adjointness results are left to the reader (or, see [76, pp. 123–124]).

$(2 \Rightarrow 1)$ The following formula will be important later. Define

$$\psi = FK \xrightarrow{Ff} FUA \xrightarrow{\varepsilon_A} A \tag{7}$$

It is routine to check that such ψ establishes the universal mapping property. $\square$

We are now ready for the theorem of [38].

THEOREM 2.8. *Every adjointness $(\mathcal{A}, \mathcal{K}, U, F, \eta, \varepsilon)$ induces a monad (T, η, μ) defined by*

$$T = \mathcal{K} \xrightarrow{F} \mathcal{A} \xrightarrow{U} \mathcal{K}, \qquad \eta = \eta, \qquad TT \xrightarrow{\mu} T = UFUF \xrightarrow{U\varepsilon F} UF$$

PROOF. The monad laws (4) follow from the adjointness identities as follows:

$$\mu \circ (T\eta) = (U\varepsilon F) \circ (F\eta) = U\big((\varepsilon F) \circ (F\eta)\big) = U id_F = id_T$$
$$\mu \circ (\eta T) = (U\varepsilon F) \circ (\eta U F) = \big((U\varepsilon) \circ (\eta U)\big) F = id_U F = id_T$$

For the remaining law, apply U to the ε-naturality square induced by ε_{FK}. □

It follows dually that $G = FU$, $\varepsilon = \varepsilon$, $\delta = F\eta U$ is a comonad on $\mathcal{A}$.

In [71, Section 3 "La Méthode Universelle pour les Constructions Standard"], MacLane anticipated monads and Theorem 2.8. What was missing was Proposition 2.7 – Kan's paper on adjoint functors [47] had not yet appeared.

Adjoint functors are everywhere (see [73, Chapter IV] for a discussion). Thus Huber's theorem makes it easy to get our hands on numerous examples of monads and comonads. Let's pause to give two examples of monads. Many others will be given as the article unfolds.

EXAMPLE 2.9. Let $\mathcal{A}$ be the category of semigroups and semigroup homomorphisms with underlying set functor $\mathcal{A} \to \mathcal{S}$. The free semigroup generated by X may be constructed as $(X^+, \cdot)$ where X^+ is the set of non-empty lists $x_1, \ldots, x_n$ $(n > 0)$ with each $x_i \in X$ and the semigroup operation is list concatenation. The resulting adjointness gives rise to the monad $TX = X^+$, $\eta_X x = x$, $\mu_X(w_1 \cdots w_n) = w_1 \cdots w_n$.

The notations of conventional mathematics are not adequate for the description of μ here. We intend, for example, that μ_X should map the length-2 list $(abc)(de)$ in $\{a, b, c, d, e\}^{++}$ to the single list $abcde$. Functional programmers use more precise notation such as $\mu_X[w_1, \ldots, w_n] = \mathit{Flatten}([w_1, \ldots, w_n])$.

EXAMPLE 2.10. This construction works in great generality [76, 5.11, p. 251]. Let $\mathcal{A}$ be the category of topological semigroups and continuous semigroup homomorphisms and let $U : \mathcal{A} \to \mathbf{Top}$ be the forgetful functor to topological spaces and continuous maps. Then U has a left adjoint. The free topological semigroup generated by the space $(X, \mathcal{T})$ is the semigroup X^+ with the smallest topology $\widehat{\mathcal{T}}$ rendering continuous all $g : X^+ \to (Y, \mathcal{S})$ where $(Y, \mathcal{S})$ is a topological space for which there exists a compatible topological semigroup structure making g a semigroup homomorphism. The induced monad is $T(X, \mathcal{T}) = (X^+, \widehat{\mathcal{T}})$ with η and μ as in the previous example, all necessary continuities being easy to establish.

2.4. *An alternate definition*

Let $P^{-1} : \mathcal{S} \to \mathcal{S}$ be defined as $P^{-1}X = 2^X$, the set of all subsets of X, and mapping $X \xleftarrow{f} Y$ to its inverse image map $P^{-1}f : 2^X \to 2^Y$, $A \mapsto f^{-1}A$. P^{-1} is a representable functor (identifying subsets with 2-valued maps) and representable functors tend to have a left adjoint. This is so in this case. The reader may verify that given a set K, $(P^{-2}K, prin_K)$

with $P^{-2} = P^{-1}P^{-1}$, $prin_K x = \{A \subset X\colon\ x \in A\}$, is freely generated by K. To establish the universal mapping property:

$$\begin{tikzcd}
K \arrow[r, "prin_K"] \arrow[dr, "f"'] & P^{-2}K \arrow[d, "P^{-1}\psi"] & P^{-1}K \\
& P^{-2}L & P^{-1}L \arrow[u, "\psi"]
\end{tikzcd}$$

define $\psi(b) = \{x \in K\colon\ B \in f(x)\}$. By Huber's theorem we have the following example of a monad of sets.

EXAMPLE 2.11. The *double contravariant power set monad* is $(P^{-2}, prin, \mu)$ where

$$\mu_X(\overline{\mathcal{A}}) = \left\{A \subset X\colon\ \left\{\mathcal{B} \in P^{-2}X\colon\ A \in \mathcal{B}\right\} \in \overline{\mathcal{A}}\right\}$$

A problem with Definition 2.1 is that there are seven axioms to verify (two to show that T is a functor, two to show that η and μ are natural transformations and the three laws (4). Of particular difficulty is the μ-law which requires chasing two morphisms from $TTTX$ to TX. For instance, trying to directly verify that the data of the previous example constitute a monad would require looking at an element of $2^{2^{2^{2^{2^{2^X}}}}}$.

The argument using Huber's theorem seems to bypass the this type of complexity, but it requires that we already know an adjointness giving rise to the monad. To underscore this point, we look at another example for which there is no obvious pair of adjoint functors.

EXAMPLE 2.12. Let $P\colon \mathcal{S} \to \mathcal{S}$ be the *covariant* power set functor mapping X to 2^X but for $f\colon X \to Y$, $Pf(A)$ is the direct image $\{f(x)\colon\ x \in A\} \subset Y$. That this is a functor is routine. The *families monad* is (P^2, η, μ) where $\eta_X(x) = \{\{x\}\}$ and

$$\mu_X(\overline{\mathcal{A}}) = \left\{\bigcup_{\mathcal{A}\in\Gamma} \Theta_{\mathcal{A}}\colon\ \Gamma \in \overline{\mathcal{A}}, (\Theta_{\mathcal{A}}) \in \prod_{\mathcal{A}\in\Gamma} \mathcal{A}\right\}$$

The reader is *not* invited to verify the axioms at this stage. We will first offer an equivalent definition of monad which has only three axioms and in which T is never iterated. This version of monad was first written down in [76, Exercise 1.3.12, p. 32].

DEFINITION 2.13. A *monad in extension form* in a category $\mathcal{K}$ is $\mathbf{T} = (T, \eta, (-)^{\#})$ where T maps each $\mathcal{K}$-object K to a $\mathcal{K}$-object TK, η assigns to each $\mathcal{K}$-object K a $\mathcal{K}$ morphism $\eta_X\colon X \to TX$, and $(-)^{\#}$ assigns to each $\mathcal{K}$-morphism $\alpha\colon X \to TY$ another $\mathcal{K}$-morphism $\alpha^{\#}\colon TX \to TY$ subject to the following three axioms:

$$\alpha^{\#}\eta_X = \alpha$$
$$\eta^{\#} = id_X$$
$$\left(\beta^{\#}\alpha\right)^{\#} = \beta^{\#}\alpha^{\#}$$

Huber's theorem suggests the following heuristics for these axioms: There is a free object construction underlying the monad for which η is the inclusion of the generators and for which $\alpha^\#$ is the unique extension of α given by the universal mapping property. That $\alpha^\# \eta_X = \alpha$ expresses that $\alpha^\#$ indeed extends α. As $(\eta_X)^\# \eta_X = \eta_X = id_{TX}\eta_X$, and as surely id_{TX} is a valid ψ, uniqueness gives $(\eta_X)^\# = id_{TX}$. Similarly, because the set of valid ψ is surely closed under composition, since $(\beta^\#\alpha)^\# \eta_X = \beta^\#\alpha = \beta^\#\alpha^\#\eta_X$, $(\beta^\#\alpha)^\# = \beta^\#\alpha^\#$.

These heuristics are a bit hazy, but the definition itself is not. We will prove later in this section that it is indeed true that every monad arises by Huber's theorem, in general in more than one way. The next result shows that the two definitions of monad agree.

PROPOSITION 2.14. *If $(T, \eta, (-)^\#)$ is a monad in extension form, then T is a functor via*

$$Tf = (X \xrightarrow{f} Y \xrightarrow{\eta_Y} TY)^\# \tag{8}$$

and then (T, η, μ) is a monad if

$$\mu_X = (id_{TX})^\# \tag{9}$$

Conversely, if (T, η, μ) is a monad then $(T, \eta, (-)^\#)$ is a monad in extension form if

$$\alpha^\# = TX \xrightarrow{T\alpha} TTY \xrightarrow{\mu_Y} TY \tag{10}$$

Under these passages, the two concepts are coextensive.

PROOF. Let $(T, \eta, (-)^\#)$ be a monad in extension form. T is a functor:

$$\begin{aligned} T id_X &= (\eta_X id_X)^\# = id_{TX} \\ T(gf) &= (\eta_Z gf)^\# = \big((\eta_Z g)^\# \eta_Y f\big)^\# = (\eta_Z g)^\# (\eta_Y f)^\# = (Tg)(Tf) \end{aligned}$$

η is natural:

$$(Tf)\eta_X = (\eta_Y f)^\# \eta_X = \eta_Y f$$

μ is natural:

$$\begin{aligned} \mu_Y(TTf) &= (id_{TY})^\# \big(\eta_Y(Tf)\big)^\# = \big((id_{TY})^\# \eta_{TY}(Tf)\big)^\# = \big(id_{TY}(Tf)\big)^\# \\ &= \big((Tf)id_{TX}\big)^\# = \big((\eta_Y f)^\# id_{TX}\big)^\# = (\eta_Y f)^\# (id_{TX})^\# = (Tf)\mu_X \end{aligned}$$

The three monad laws:

$$\begin{aligned} \mu_X(T\eta_X) &= (id_{TX})^\# (\eta_{TX}\eta_X)^\# = \big((id_{TX})^\# \eta_{TX}\eta_X\big)^\# = (\eta_X)^\# = id_{TX} \\ \mu_X \eta_{TX} &= (id_{TX})^\# \eta_{TX} = id_{TX} \\ \mu_X \mu_{TX} &= (id_{TX})^\# (id_{TTX})^\# = \big((id_{TX})^\# id_{TTX}\big)^\# = (\mu_X)^\# \\ &= \big((id_{TX})^\# \eta_{TX}\mu_X\big)^\# = (id_{TX})^\# (\eta_{TX}\mu_X)^\# = \mu_X(T\mu_X) \end{aligned}$$

Now the converse direction. The two η laws:

$$(\eta_X)^\# = \mu_X(T\eta_X) = id_{TX}$$
$$\alpha^\#\eta_X = \mu_Y(T\alpha)\eta_X = \mu_Y\eta_{TY}\alpha(\eta \text{ n.t.}) = \alpha$$

The remaining extension law:

$$\begin{aligned}(\beta^\#\alpha)^\# &= \mu_Z T\big(\mu_Z(T\beta)\alpha\big) = \mu_Z(T\mu_Z)(TT\beta)(T\alpha)\\ &= \mu_Z\mu_{TZ}(TT\beta)(T\alpha) \quad (\mu\text{-monad law})\\ &= \mu_Z(T\beta)\mu_Y(T\alpha) \quad (\mu \text{ is natural})\\ &= \beta^\#\alpha^\#\end{aligned}$$

To complete the proof we must check that these passages are mutually inverse. If $(T,\eta,\mu) \mapsto (T,\eta,(-)^\#) \mapsto (\widehat{T},\eta,\hat{\mu})$ then

$$\widehat{T}f = (\eta_Y f)^\# = \mu_Y T(\eta_Y f) = \mu_Y(T\eta_Y)Tf = Tf$$
$$\hat{\mu}_X = (id_{TX})^\# = \mu_X(T id_X) = \mu_X$$

If $(T,\eta,(-)^\#) \mapsto (T,\eta,\mu) \mapsto (T,\eta,(-)^\star)$ then

$$\alpha^\star = \mu_Y(T\alpha) = (id_{TY})^\#(\eta_{TY}\alpha)^\# = \big((id_{TY})^\#\eta_{TY}\alpha\big)^\# = (id_{TY}\alpha)^\# = \alpha^\# \qquad \square$$

Armed with this result we are now ready to establish that the families monad of Example 2.12 really is a monad. Consider the extension formula given by

$$\alpha^\#(\mathcal{A}) = \left\{\bigcup_{x\in A} B_x\colon\ A \in \mathcal{A}, \forall x \in A\ B_x \in \alpha x\right\}$$

The formula originally given for μ_X is evidently $(id_{TX})^\#$. Thus we have only to check the three extension axioms of Definition 2.13. The two η laws are routine. The details of the composition law are as follows.

$$\begin{aligned}C \in \beta^\#\alpha^\#(\mathcal{A}) \quad &\Leftrightarrow \quad \exists B \in \alpha^\#(\mathcal{A})\ \forall y \in B\ \exists C_y \in \beta y \quad C = \bigcup_{y\in B} C_y\\ &\Leftrightarrow \quad \exists A \in \mathcal{A}\ \forall x \in A\ \exists B_x \in \alpha x\ \forall y \in \bigcup_{x\in A} B_x\ \exists C_y \in \beta y\\ &\qquad\quad C = \bigcup_{y\in B} C_y\\ &\Leftrightarrow \quad \exists A \in \mathcal{A}\ \forall x \in A\ \exists B_x \in \alpha x\ \forall y \in B_x\ \exists C_{xy} \in \beta y\\ &\qquad\quad C = \bigcup_{x\in A}\bigcup_{y\in B_x} C_{xy}\end{aligned}$$

On the other hand,

$$
\begin{aligned}
C \in \left(\beta^{\#}\alpha\right)^{\#}(\mathcal{A}) \quad &\Leftrightarrow \quad \exists A \in \mathcal{A}\ \forall x \in A\ \exists C_x \in \beta^{\#}\alpha x \quad C = \bigcup_{x \in A} C_x \\
&\Leftrightarrow \quad \exists A \in \mathcal{A}\ \forall x \in A\ \exists B_x \in \alpha x\ \forall y \in B_x\ \exists C_{xy} \in \beta y \\
&\qquad\quad C = \bigcup_{x \in A} \bigcup_{y \in B_x} C_{xy}
\end{aligned}
$$

which is exactly the same.

We next describe monad maps in extension form.

PROPOSITION 2.15. *Let* **S**, **T** *be monads in* $\mathcal{K}$ *and for each object* X *of* $\mathcal{K}$ *let* $\lambda_X : SX \to TX$ *be a morphism in* $\mathcal{K}$*. Then* λ *is a monad map if and only if* λ *preserves* η *and for every* $\alpha : X \to SY$ *the following square commutes*

$$
\begin{tikzcd}
SX \arrow[r, "\lambda_X"] \arrow[d, "\alpha^{\#}"'] & TX \arrow[d, "(\lambda_Y\alpha)^{\#}"] \\
SY \arrow[r, "\lambda_Y"'] & TY
\end{tikzcd}
$$

PROOF. First let λ be a monad map as in Definition 2.2. Then

$$
\begin{aligned}
(\lambda_Y\alpha)^{\#} &= \mu_Y T(\lambda_Y\alpha)\lambda_X = \mu_Y(T\lambda_Y)(T\alpha)\lambda_X \\
&= \mu_Y(T\lambda_Y)\lambda_{SY}(S\alpha) \quad (\lambda \text{ is natural}) \\
&= \mu_Y(\lambda\lambda)_Y(S\alpha) \quad (\text{definition of } \lambda\lambda) \\
&= \lambda_X\mu_Y(S\alpha) \quad (\text{monad map}) \\
&= \lambda_X\alpha^{\#}
\end{aligned}
$$

Conversely, let λ preserve η and $(-)^{\#}$ and show that λ is natural and preserves μ.

$$
\begin{aligned}
&\lambda_Y(Sf) = \lambda_Y(\eta_Y f)^{\#} = (\lambda_Y\eta_Y f)^{\#}\lambda_X = (\eta_Y f)^{\#}\lambda_X = (Tf)\lambda_X \\
&\lambda_Y\mu_Y = \lambda_Y(id_{SY})^{\#} = (\lambda_Y id_{SY})^{\#}\lambda_{SY} = \mu_Y T(\lambda_Y)\lambda_{SY} = \mu_Y(\lambda\lambda)_Y
\end{aligned}
$$

□

EXAMPLE 2.16. The *power set monad* **P** is defined by

$$
\begin{aligned}
&PX = 2^X \\
&\eta_X x = \{x\} \\
&\alpha^{\#}(A) = \bigcup_{x \in A} \alpha x
\end{aligned}
$$

P is the same covariant power set functor as before, $Pf(A) = \bigcup_{x\in A} fx$. $prin : \mathbf{P} \to \mathbf{P}^{-2}$ is a monad map to the double contravariant power set monad if $prin_X(A) = \{B\colon\ A \subset B\}$ is the principal filter generated by A. $sing : \mathbf{P} \to \mathbf{P}^2$ is a monad map to the families monad if $sing_X(A) = \{A\}$.

2.5. *Eilenberg and Moore*

In [20], Eilenberg and Moore showed that, given a monad $\mathbf{T}$ in $\mathcal{K}$, there exists an adjointness $(K^{\mathbf{T}}, \mathcal{K}, U^{\mathbf{T}}, F^{\mathbf{T}}, \eta, \varepsilon)$ whose induced monad is precisely $\mathbf{T}$ – that is, every monad arises from Huber's Theorem 2.8. In that paper, monads were called "triples", not a very colorful term. The objects of the category $K^{\mathbf{T}}$, the main object of study of the next section, were not named at all.

We have already thought of TX as a "free object" with inclusion of the generators η and unique extensions $\alpha^{\#}$. By carrying this idea to its logical conclusion, we can motivate the category $K^{\mathbf{T}}$.

PROPOSITION 2.17. *Let $\mathbf{T}$ be a monad in $\mathcal{K}$. Then a $\mathcal{K}$-morphism $\psi : TX \to TY$ has form $\alpha^{\#}$ for some α if and only if the following square commutes.*

$$\begin{tikzcd} TTX \arrow[r, "T\psi"] \arrow[d, "\mu_X"'] & TTY \arrow[d, "\mu_Y"] \\ TX \arrow[r, "\psi"'] & TY \end{tikzcd}$$

PROOF. If $\psi = \alpha^{\#}$ then $\psi = \mu_Y(T\alpha)$. Thus

$$\begin{aligned} \psi\mu_X &= \mu_Y(T\alpha)\mu_X = \mu_Y\mu_{TY}(TT\alpha) \quad (\mu \text{ is natural}) \\ &= \mu_Y(T\mu_Y)(TT\alpha) \quad (\text{monad } \mu\text{-law}) \end{aligned}$$

so 2.17 holds. Conversely, define $\alpha = \psi\eta_X$. Then

$$\alpha^{\#} = \mu_Y(T\alpha) = \mu_Y(T\psi)(T\eta_X) = \psi\mu_X(T\eta_X) = \psi \qquad \square$$

Of course, α is unique: $\alpha^{\#} = \beta^{\#} \Rightarrow \alpha = \alpha^{\#}\eta_X = \beta^{\#}\eta_X = \beta$.

In light of this, think of 2.17 as saying that ψ is a morphism from (TX, μ_X) to (TY, μ_Y). To make this precise, we have the following key definition.

DEFINITION 2.18. Let $\mathbf{T}$ be a monad in $\mathcal{K}$. A $\mathbf{T}$-*algebra* is a pair (X, ξ) with $\xi : TX \to X$ a $\mathcal{K}$-morphism such that the two laws on the left below hold. A $\mathbf{T}$-*homomorphism* $f : (X, \xi) \to (Y, \theta)$ is a $\mathcal{K}$-morphism $f : X \to Y$ such that the square on the right commutes.

$$\begin{tikzcd} X \arrow[r, "\eta_X"] \arrow[dr, "id"'] & TX \arrow[d, "\xi"] & TTX \arrow[l, "T\xi"'] \arrow[d, "\mu_X"] \\ & X & TX \arrow[l, "\xi"] \end{tikzcd} \qquad \begin{tikzcd} TX \arrow[r, "Tf"] \arrow[d, "\xi"'] & TY \arrow[d, "\theta"] \\ X \arrow[r, "f"'] & Y \end{tikzcd}$$

ξ is called the *structure map* of (X, ξ).

It is obvious from the monad laws (4) that (TX, μ_X) is a $\mathbf{T}$-algebra. Thus the definition above expresses, after the fact, that ξ is a $\mathbf{T}$-homomorphism extending id_X.

By the functoriality of T, $\mathbf{T}$-algebras and their homomorphisms form a category $K^{\mathbf{T}}$ with underlying $\mathcal{K}$-object functor $U^{\mathbf{T}} : K^{\mathbf{T}} \to \mathcal{K}$ defined by forgetting the structure maps.

The following proposition due to [20, Theorem 2.2, p. 384] puts it together:

PROPOSITION 2.19. *(TX, μ_X) is the free $\mathbf{T}$-algebra generated by X. The resulting adjointness $(K^{\mathbf{T}}, \mathcal{K}, U^{\mathbf{T}}, F^{\mathbf{T}}, \eta, \varepsilon)$ induces $\mathbf{T}$.*

PROOF. Given a $\mathbf{T}$-algebra (Y, θ) and a $\mathcal{K}$-morphism $f : X \to Y$ we must show that there exists a unique $\mathbf{T}$-homomorphism $f^{\#} : (TX, \mu_X) \to (Y, \theta)$ with $f^{\#}\eta_X = f$. To that end, $Tf : (TX, \mu_X) \to (TY, \mu_Y)$ is a $\mathbf{T}$-homomorphism because $\mu : TT \to T$ is natural. We have already observed that $\theta : (TX, \mu_X) \to (X, \xi)$ is a $\mathbf{T}$-homomorphism and that the composition of two $\mathbf{T}$-homomorphisms is again one. Thus $f^{\#}$ defined by

$$f^{\#} = TX \xrightarrow{Tf} TY \xrightarrow{\theta} Y \tag{11}$$

is a $\mathbf{T}$-homomorphism. (The earlier formula (10) is consistent with the above.) Such $f^{\#}$ extends f because

$$f^{\#}\eta_X = \theta(Tf)\eta_X = \theta\eta_Y f = id_X f = f$$

Now suppose $\psi : (TX, \mu_X) \to (X, \xi)$ is a $\mathbf{T}$-homomorphism extending f. Then

$$\psi = \psi\mu_X(T\eta_X) = \theta(T\psi)(T\eta_X) = \theta T(\psi\eta_X) = \theta(Tf) = f^{\#}$$

The proof of Proposition 2.7 constructs the adjointness $(K^{\mathbf{T}}, \mathcal{K}, U^{\mathbf{T}}, F^{\mathbf{T}}, \eta, \varepsilon)$ where $F^{\mathbf{T}}X = (TX, \mu_X)$, $F^{\mathbf{T}}f = Tf : (TX, \mu_X) \to (TY, \mu_Y)$ and $\varepsilon_{(X,\xi)} : (TX, \mu_X) \to (X, \xi)$ is ξ. Applying Theorem 2.8 recovers $\mathbf{T}$. $\square$

A problem with the $\mathbf{T}$-algebra laws as in Definition 2.18 is that T is iterated. The following result gives two extension-form definitions which do not iterate T. One of them is a conditional statement which is frequently very easy to use. The other is an equation, and it turns out to be possible to generalize it to relational $\mathbf{T}$-models (see [81]).

PROPOSITION 2.20. *Let $\mathbf{T}$ be a monad in $\mathcal{K}$ and let $\xi : TX \to X$ satisfy $\xi\eta_X = id_X$. Equivalent are*

1. *(X, ξ) is a $\mathbf{T}$-algebra, that is, $\xi(T\xi) = \xi\mu_X$.*
2. *$\forall \alpha, \beta : A \to TX$, $\xi\alpha = \xi\beta \Rightarrow \xi\alpha^{\#} = \xi\beta^{\#}$.*
3. *$\forall \alpha : A \to TX$, the following square commutes.*

$$\begin{tikzcd}
TA \arrow[r, "T(\xi\alpha)"] \arrow[d, "\alpha^{\#}"'] & TX \arrow[d, "\xi"] \\
TX \arrow[r, "\xi"'] & X
\end{tikzcd}$$

PROOF. ($1 \Rightarrow 2$) $\xi\alpha^{\#}$ is the unique homomorphism extending $\xi\alpha$.

($2 \Rightarrow 3$) $T(\xi\alpha) = (\eta_X \xi\alpha)^{\#}$ so let $\beta = \eta_X \xi\alpha$. As $\xi\beta = \xi\eta_X\xi\alpha = \xi\alpha$, $\xi T(\xi\alpha) = \xi\beta^{\#} = \xi\alpha^{\#}$.

($3 \Rightarrow 1$) Set $\alpha = id_{TX}$. □

In the next example, it is natural to refer to both forms of the definition of a **T**-algebra.

EXAMPLE 2.21. Let **Semigroup** be the category of semigroups and semigroup homomorphisms with underlying set functor $U : \mathbf{Semigroup} \to \mathcal{S}$ and let **T** be the monad of sets of Example 2.9. Then $\mathcal{S}^{\mathbf{T}} = \mathbf{Semigroup}$.

More precisely, there exists an isomorphism of categories $\Psi : \mathbf{Semigroup} \to \mathcal{S}^{\mathbf{T}}$ which is *over* $\mathcal{S}$ in that the following triangle of functors commutes.

$$\begin{tikzcd}
\mathbf{Semigroups} \arrow[rr, "\Psi"] \arrow[dr, "U"'] & & \mathcal{S}^{\mathbf{T}} \arrow[dl, "U^{\mathbf{T}}"] \\
& \mathcal{S} &
\end{tikzcd}$$

To this end, given a semigroup $(X, \star)$ define $\Psi(X, \star) = (X, \xi)$ where $\xi(x_1 \cdots x_n) = x_1 \star \cdots \star x_n$. As $\eta_X x$ is the length-1 list x, it is clear that $\xi\eta_X = id_X$. If $\alpha : A \to TX$, $\xi\alpha^{\#}(a_1 \cdots a_n) = \xi\alpha(a_1) \star \cdots \star \xi\alpha(a_n)$ so $\xi\alpha = \xi\beta \Rightarrow \xi\alpha^{\#} = \xi\beta^{\#}$. Thus Ψ is well-defined on objects. Define $\Psi^{-1}(X, \xi) = (X, \star)$ where $x \star y = \xi(xy)$. To check associativity, we'll use functional programming notation for lists, $[x, y]$ instead of xy. Then

$$\begin{aligned} x \star (y \star z) &= \xi\big[\xi[x], \xi[y, z]\big] = \xi(T\xi)\big[[x], [y, z]\big] = \xi\mu_X\big[[x], [y, z]\big] \\ &= \xi[x, y, z] = \xi\mu_X\big[[x, y], [z]\big] = \cdots = (x \star y) \star z \end{aligned}$$

checking that these passages are mutually inverse and that the homomorphisms are the same is routine.

2.6. *Kleisli*

In [51], Kleisli gave an adjointness $(K_{\mathbf{T}}, \mathcal{K}, U_{\mathbf{T}}, F_{\mathbf{T}}, \eta, \varepsilon)$ which induces **T**. The paper was published at about the same time as that of Eilenberg and Moore [20] and neither was aware of the other's work. The two constructions are almost always different. In a later section, $\mathcal{S}_{\mathbf{T}}$ will provide a semantic category for programming languages.

DEFINITION 2.22. Let **T** be a monad in $\mathcal{K}$. The *Kleisli category* $K_{\mathbf{T}}$ of **T** has the same objects as $\mathcal{K}$ and a $K_{\mathbf{T}}$-morphism from X to Y is a $\mathcal{K}$-morphism $X \to TY$. The identity morphism of X is η_X. For $\alpha : X \to TY$, $\beta : Y \to TZ$, the composition $\beta\alpha$ in $K_{\mathbf{T}}$ is $\beta^{\#}\alpha$.

That $K_{\mathbf{T}}$ is a category is essentially the axioms of Definition 2.13.

Kleisli proved the following.

PROPOSITION 2.23. *Let* $\mathbf{T}$ *be a monad in* $\mathcal{K}$. *Define* $F_{\mathbf{T}}:\mathcal{K}\to K_{\mathbf{T}}$ *by* $F_{\mathbf{T}}X = X$, $F_{\mathbf{T}}f = \eta_Y f$. *Define* $U_{\mathbf{T}}:K_{\mathbf{T}}\to\mathcal{K}$ *by* $U_{\mathbf{T}}X = TX$, $U_{\mathbf{T}}\alpha = \alpha^{\#}$. *Define* $\eta = \eta$ (*noting that* $U_{\mathbf{T}}F_{\mathbf{T}} = T$), $\varepsilon_X = id_{TX}$. *Then* $(K_{\mathbf{T}}, \mathcal{K}, U_{\mathbf{T}}, F_{\mathbf{T}}, \eta, \varepsilon)$ *is an adjointness which induces* $\mathbf{T}$.

In most situations, $X \mapsto TX$ is injective. In such cases, $K_{\mathbf{T}}$ is isomorphic to the full subcategory $\mathcal{F}$ of $K^{\mathbf{T}}$ of all objects of form (TX, μ_X). The Kleisli construction is then just the restriction of the Eilenberg–Moore construction since $F^{\mathbf{T}}$ factors through $\mathcal{F}$.

EXAMPLE 2.24. For $\mathbf{P}$ the power set monad of Example 2.16, $\mathcal{S}_{\mathbf{P}}$ is the category of sets and relations. Here a relation $R:X\to Y$ is thought of as a function $X\to PY$, so that $xRy \Leftrightarrow y\in Rx$. If $S:Y\to PZ$ is another relation, the Kleisli composition is the usual one, $x\ (\beta^{\#}\alpha)\ z \Leftrightarrow \exists y\ xRy \wedge ySz$. η_X corresponds to the usual diagonal relation.

PROPOSITION 2.25. *For any monad* $\mathbf{T} = (T, \eta, (-)^{\#})$ *in* $\mathcal{K}$, *equivalent are*

(1) η *is pointwise monic,*
(2) T *is faithful,*
(3) $F_{\mathbf{T}}:\mathcal{K}\to\mathcal{K}$ *is a subcategory.*

A monad satisfying any of these equivalent properties is said to be *nontrivial*.

PROOF. $(1\Rightarrow 2)$ As the functor $F_{\mathbf{T}}$ is bijective on objects, it is a subcategory if and only if it is faithful and this is obvious if η is pointwise monic.

$(2\Rightarrow 3)$ For $f, g: X\to Y$, if $Tf = Tg$ then $F_{\mathbf{T}}f = \eta_Y f = (Tf)\eta_Y = (Tg)\eta_Y = \eta_Y g = F_{\mathbf{T}}g$. Applying the hypothesis, $f = g$.

$(3\Rightarrow 1)$ If $\eta_Y f = \eta_Y g$ then $(Tf)\eta_Y = (Tg)\eta_Y$. As Tf, Tg are $\mathbf{T}$-homomorphisms agreeing on generators, $Tf = Tg$. As T is faithful, $f = g$ and η_Y is monic. □

PROPOSITION 2.26. *Let* $\mathbf{T}$ *be a monad in* $\mathcal{S}$. *Then* $\mathbf{T}$ *is nontrivial if and only if some* $\mathbf{T}$-*algebra has at least two elements.*

PROOF. Let (Y, θ) be a $\mathbf{T}$-algebra with at least two elements. Given a set X choose a sufficiently large set n such that there exists an injective function $f: X\to Y^n$. Let $pr_i: Y^n\to Y$ be the ith coordinate projection, and let $f_i: X\to Y$ be $pr_i f$. Define $f^{\#}: TX\to Y^n$ by $pr_i f^{\#} = f_i^{\#}$. As $f^{\#}\eta_X = f$ is monic, so is η_X. This shows that $\mathbf{T}$ is nontrivial. The converse is obvious since TX has at least two elements whenever X does, if η is pointwise monic. □

It is immediate from the preceding proposition that there are exactly two trivial monads of sets. For one, TX has one element for all sets X. For the other, TX has one element if X is nonempty but $T\emptyset = \emptyset$.

3. Universal algebra

Universal algebra is the study of equationally definable classes. The 1935 founding paper by Birkhoff [9] was limited to finitary operations, which seemed natural in that most algebraic structures such as groups, rings, lattices, Jordan algebras and so forth can be axiomatized using operations that require at most three variables. Later, in 1959, Słominski [102] published a treatise on universal algebra with infinitary operations. Examples of these are associated with analysis more than algebra by most mathematicians (e.g., the Boolean σ-algebras of measure theory) but some of the theory of varieties is the same without the finitary restriction. As we shall discuss in detail in this section, the work of Lawvere, Linton and Beck (and others) led to a new classification of algebraic structure which did not mention operations and equations at all. Among the "categories of $\mathcal{K}$-objects with structure" with underlying $\mathcal{K}$-object functor $U : \mathcal{A} \to \mathcal{K}$, are the *monadic* ones which are isomorphic to $U^{\mathbf{T}} : K^{\mathbf{T}} \to \mathcal{K}$ for some monad $\mathbf{T}$ in $\mathcal{K}$. The categories monadic over $\mathcal{S}$ are precisely those equationally definable classes which have free algebras. As we shall see in Example 3.19, compact Hausdorff spaces are monadic over $\mathcal{S}$ (although there is no cardinality bound on the number of variables required by operations) whereas complete lattices are not. The category of Abelian groups is a finitary variety, but its dual – the category of compact Abelian groups – is also monadic over $\mathcal{S}$.

There is a general theory of monadic completion (Definitions 3.18, 3.20 and Theorem 3.36) which provides a best monadic approximation for a large class of set-valued functors. There is also an existence theorem (Theorem 3.35) to the effect that, under reasonable conditions, every forgetful functor between monadic functors (e.g., rings are groups, algebras over a field are vector spaces) has a left adjoint.

3.1. *Varieties*

DEFINITION 3.1. A *signature* is a set Σ of pairs (n, ω) where n is a cardinal. We denote $\Sigma_n = \{\omega\colon\ (n, \omega) \in \Sigma\}$ the set of *n-ary operation symbols* of Σ. Σ is *finitary* if for each $(n, \omega) \in \Sigma$, n is a finite cardinal.

A *Σ-algebra* is a pair (X, δ) where for all $\omega \in \Sigma_n$, δ assigns an n-ary operation $\delta_\omega : X^n \to X$. The *$\Sigma$-homomorphisms* $f : (X, \delta) \to (Y, \varepsilon)$ preserve the operations, $f(\delta_\omega(x_i)) = \varepsilon_\omega(f x_i)$. Evidentially, Σ-algebras and their homomorphisms form a category with identity morphisms and composition at the level $\mathcal{S}$. We call this category Σ-**Alg** and write the underlying set functor as $U_\Sigma : \Sigma\text{-}\mathbf{Alg} \to \mathcal{S}$.

Category theory must come to grips with "sets that are too large". All standard texts treat this issue, so we shall not belabor it here. The objects of $\mathcal{S}$ are small sets. The Cartesian product of all the sets in $\mathcal{S}$ is not.

With the operations defined in the usual coordinatewise fashion, a product of Σ-algebras is again one and is the categorical product in Σ-**Alg**. A subset of a Σ-algebra (X, δ) which is closed under all the operations δ_ω is a *subalgebra* of (X, δ) and is a Σ-algebra in its own right. It is clear that the image of any Σ-homomorphism is a subalgebra.

THEOREM 3.2. *Let Σ be a* bounded *signature, that is, Σ is a small set. Then U_Σ: Σ-**Alg** $\to \mathcal{S}$ has a left adjoint.*

PROOF. Given Σ-homomorphisms $f, g : (X, \delta) \to (Y, \varepsilon)$, the inclusion of $\{x\colon\ fx = gx\}$ is a subalgebra and so provides the equalizer of f, g in Σ-**Alg**. Thus Σ-**Alg** has and U_Σ

preserves all small limits. By the Freyd general adjoint functor theorem, it suffices to prove that U_Σ satisfies the solution set condition. Specifically, in the diagram below:

$$\begin{tikzcd} X \arrow[r, "f"] \arrow[dr, "g"'] & (Y,\delta) \arrow[d, "\psi"] \\ & (Z,\varepsilon) \end{tikzcd}$$

given X there exists a small set $\mathcal{F}$ of algebra-valued f such for each algebra-valued g there is at least one $f \in \mathcal{F}$ and Σ-algebra homomorphism ψ making the triangle commute. We sketch the idea. Every subset A of an algebra generates a subalgebra $\langle A\rangle$ by intersecting all the subalgebras containing A. As a solution set choose all $f : X \to (Y,\delta)$ with Y a cardinal and $(Y,\delta) = \langle f(X)\rangle$. For fixed Y, f ranges over a small set and, because Σ is bounded, δ ranges over a small set. The hard part is to show that Y is bounded once X is fixed. Such is not a general principle for reasonable categories of structured sets; for example, the category of topological spaces has arbitrarily large spaces with a 1-element dense subset. It is shown in [93, Proposition 1.3, p. 99] that it suffices to choose $Y \leqslant \alpha^\beta$ with α any infinite cardinal exceeding the cardinal of X and of each Σ_n and with β any infinite cardinal for which $\Sigma_n \neq \emptyset \Rightarrow n < \beta$. Such α and β exist because Σ is bounded. □

DEFINITION 3.3. Let Σ be a bounded signature. A *Σ-equation* is (n,ω,ν) with $\omega,\ \nu \in F_\Sigma n$ where $F_\Sigma n : \mathcal{S} \to \Sigma$-**Alg** is a left adjoint of U_Σ whose existence is guaranteed by Theorem 3.2. (Strictly speaking, this functor is unique only up to natural equivalence, so one should choose a specific one). A Σ-algebra (X,δ) *satisfies* the equation (n,ω,ν) if for all $f : n \to X$, the unique Σ-homomorphism $f^\# : F_\Sigma n \to (X,\delta)$ extending f maps ω and ν to the same value. If E is any class of equations, a (Σ, E)*-algebra* is a Σ-algebra which satisfies all the equations in E.

We denote the category of all (Σ, E)-algebras and its underlying set functor as $U_{\Sigma,E} : (\Sigma, E)$-**Alg** $\to \mathcal{S}$. An *equationally definable class* is a category of form (Σ, E)-**Alg**, and (Σ, E) is an *equational presentation* for it.

EXAMPLE 3.4. An equational presentation for rings with unit is as follows. Take $\Sigma_0 = \{0,1\}$, $\Sigma_1 = \{-\}$, $\Sigma_2 = \{+,\star\}$, $\Sigma_n = \emptyset$ if $n > 2$. One may construct the free Σ-algebra generated by X by applying (only) the following rules:

atomic expressions: $X \uplus \Sigma_0 \subset FX$
derived expressions: $\omega \in \Sigma_n \wedge e_1,\ldots,e_n \in FX \;\Rightarrow\; \omega e_1 \cdots e_n \in FX$

Here $\uplus$ means disjoint union. Let $3 = \{a,b,c\}$. If (X,δ) is a Σ-algebra and $f : 3 \to X$, $f^\#(\star + ab + 10) = \delta_\star(\delta_+(a,b), \delta_+(\delta_1,\delta_0))$. In normal infix notation, this would be written $(a+b)\star(1+0)$. Rings are defined by the usual equations. The associative law of multiplication is expressed by the Σ-equation $(3,\ \star a \star bc,\ \star\star abc)$. In more conventional notation, one writes $a \star (b \star c) = (a \star b) \star c$ and we will usually write things this way.

DEFINITION 3.5. A *variety* of Σ-algebras is a class of Σ-algebras which is closed under the formation of products, subalgebras and homomorphic images.

It is routine to check that if $f : (X, \delta) \to (Y, \varepsilon)$ is a bijective Σ-homomorphism, then f^{-1} also is, so that f is an isomorphism in Σ-**Alg**. In that case, any variety containing (X, δ) also contains (X, ε) because an isomorphism is a unary product and a variety is closed under products.

It is obvious that (Σ, E)-**Alg** is a variety of Σ-algebras for any equational presentation (Σ, E). It is a well-known theorem of Birkhoff [9] that the converse is true. We turn to a proof based on some category-theoretic principles which will be useful in their own right.

DEFINITION 3.6. Let $\mathcal{A}$ be a complete category with image factorization system $(\mathcal{E}, \mathcal{M})$ and let $\mathcal{B}$ be a full subcategory of $\mathcal{A}$. $\mathcal{B}$ is a *quasivariety* in $\mathcal{A}$ if $\mathcal{B}$ is closed under products and $\mathcal{M}$-subobjects. The latter means that for all $m : X \to Y$ in $\mathcal{M}$, if Y is a $\mathcal{B}$-object, so too is X.

$\mathcal{B}$ is an *$\mathcal{E}$-reflective category* of $\mathcal{A}$ if the inclusion functor $U : \mathcal{B} \to \mathcal{A}$ has a left adjoint F such that for all $\mathcal{A}$-objects A the "inclusion of the generators" $\eta_A : A \to FB$ is in $\mathcal{E}$. (We are abusing notation here; purists would write UFB for FB.) Such η_A is called the *reflection map* of A.

For emphasis, we draw the diagram of the universal property of the reflection map.

$$\begin{tikzcd} A \arrow[r, "\eta_A"] \arrow[dr, "f"'] & FB \arrow[d, "\psi"] \\ & C \end{tikzcd}$$

$\forall C$ in $\mathcal{B}$ $\forall f : A \to C$ $\exists!\ \psi : FB \to C$ with $\psi\eta_A = f$.

The reader may be surprised to learn that the "correct" definition of "variety" at this level is *not* a quasivariety closed under $\mathcal{E}$-quotients. See Theorem 3.41.

DEFINITION 3.7. Let $\mathcal{E}$ be a class of morphisms in a category $\mathcal{A}$. $\mathcal{A}$ is *$\mathcal{E}$-cowellpowered* if for each object X there exists a small set $\mathcal{E}_0$ of morphisms in $\mathcal{E}$ with domain X such that, as shown in the diagram below,

$$\begin{tikzcd} & X \arrow[dl, "e"'] \arrow[dr, "e_0"] & \\ Y \arrow[rr, "\psi"'] & & Y_0 \end{tikzcd}$$

if $e \in \mathcal{E}$ there exists $e_0 \in \mathcal{E}_0$ and an isomorphism ψ such that the diagram commutes.

Cowellpowered means that there is a small set of "$\mathcal{E}$-quotients". For $\mathcal{E}$ the class of surjective functions, $\mathcal{S}$ is $\mathcal{E}$-cowellpowered via the set of all surjections from X to X/R as R

ranges over the set of all equivalence relations on X. On the other hand, if $\mathcal{E}$ is all functions, $\mathcal{S}$ is not cowellpowered.

PROPOSITION 3.8. *If $\mathcal{A}$ is a locally small complete category with image factorization system $(\mathcal{E}, \mathcal{M})$ such that $\mathcal{A}$ is $\mathcal{E}$-cowellpowered then every quasivariety is $\mathcal{E}$-reflective and conversely.*

PROOF. Let $\mathcal{B}$ be a quasivariety. As every equalizer is in $\mathcal{M}$, $\mathcal{B}$ is closed under limits, that is, $\mathcal{B}$ has and $U : \mathcal{B} \to \mathcal{A}$ preserves small limits. For each object A choose a small set $\mathcal{E}_A$ of $\mathcal{E}$-morphisms as per the definition of $\mathcal{E}$-cowellpowered. Consider $f : A \to B$ with B in $\mathcal{B}$ and choose image factorization $f = A \xrightarrow{e_f} I_f \xrightarrow{m_f} B$ in $\mathcal{A}$ with e_f in $\mathcal{E}_A$. As I_f is in $\mathcal{B}$ and as $\mathcal{A}$ is locally small, e_f ranges over a small set. This provides the solution set condition, so U has a left adjoint. By the universal property of the reflection, the image factorization of the reflection has a monic part which is an isomorphism, so $\mathcal{B}$ is an $\mathcal{E}$-reflective subcategory. The converse (which is less important for us) is established by the general technique of showing for a product, subalgebra or homomorphic image, that the desired object has a reflection map which must be an isomorphism. $\square$

We are now ready to prove Birkhoff's remarkable theorem.

THEOREM 3.9. *Let Σ be a bounded signature. Let $\mathcal{A}$ be any class of Σ-algebras and let E be the class of all equations satisfied by every algebra in $\mathcal{A}$. Then each (Σ, E)-algebra is a homomorphic image of a subalgebra of a product of algebras in $\mathcal{A}$.*

An immediate corollary is that every variety of Σ-algebras is equationally definable.

PROOF. Let X be a set and let $(F_\Sigma X, \eta_X : X \to F_\Sigma X)$ be the free Σ-algebra generated by X. Let $\theta_X : F_\Sigma X \to F_{\Sigma,E} X$ be the surjective reflection map that results from Proposition 3.8 applied to Σ-**Alg** with $\mathcal{E}$ = surjective homomorphisms, $\mathcal{M}$ = injective homomorphisms. Then $(F_{\Sigma,E} X, \hat{\eta_X})$ is the free (Σ, E)-algebra generated by X, where $\hat{\eta_X} = \theta_X \eta_X$. This shows in passing that

$$\text{for bounded } \Sigma,\ U_{\Sigma,E} : (\Sigma, E)\text{-}\mathbf{Alg} \to \mathcal{S} \text{ has a left adjoint.} \tag{12}$$

Now let n be a cardinal, and suppose $\sigma \neq \tau \in F_{\Sigma,E} n$. Choose $\bar{\sigma}, \bar{\tau} \in F_\Sigma n$ with $\theta_n \bar{\sigma} = \sigma$, $\theta_n \bar{\tau} = \tau$. As θ_n is a Σ-homomorphism and $\theta_n \bar{\sigma} \neq \theta_n \bar{\tau}$, whereas $F_{\Sigma,E} n$ satisfies all equations in E, the Σ-equation $(n, \bar{\sigma}, \bar{\tau})$ is not in E. By the definition of E, there exists an algebra $(A_{\sigma\tau}, \delta_{\sigma\tau})$ in $\mathcal{A}$ and a function $f_{\sigma\tau} : n \to A_{\sigma\tau}$ such that $(f_{\sigma\tau})^\#(\bar{\sigma}) \neq (f_{\sigma\tau})^\#(\bar{\tau})$. Let $\psi_{\sigma\tau} : F_{\Sigma,E} n \to (A_{\sigma\tau}, \delta_{\sigma\tau})$ be the unique Σ-homomorphism with $\psi_{\sigma\tau}\, \theta_n = (f_{\sigma\tau})^\#$. This exists because all algebras in $\mathcal{A}$ satisfy E. Let (A, δ) be the Σ-algebra product of all $(A_{\sigma\tau}, \delta_{\sigma\tau})$. Then $\psi : F_{\Sigma,E} n \to (A, \delta)$ defined by $\psi\omega = (\psi_{\sigma\tau}(\omega)\colon\ \sigma \neq \tau \in F_{\Sigma,E} n)$ is an injective Σ-homomorphism. Since each $F_{\Sigma,E} X$ is isomorphic to a $F_{\Sigma,E} n$ with n a cardinal, we have shown that every free (Σ, E)-algebra is isomorphic to a subalgebra of a product of algebras in $\mathcal{A}$. If now (X, ε) is any (Σ, E)-algebra, the unique Σ-homomorphism $\xi : F_{\Sigma,E} X \to (X, \varepsilon)$ with $\xi\, \hat{\eta_X} = id_X$ is a surjection. $\square$

3.2. *Lawvere*

In his thesis [55], its published condensed summary [56] and two other publications [57, 58], Lawvere invented a category-theoretic model of varieties and explored various adjointness results. Lawvere was embarked on a program of making adjointness fundamental in the foundations of mathematics, and results of this type were already in his thesis. Meanwhile, even a conventional textbook on varieties [13] published at that time included the theorem, much stressed by Lawvere, that forgetful functors between varieties have left adjoints.

Different equational presentations can give rise to the same variety (up to an isomorphism that commutes with the underlying set functors). There are many well known examples, for example Boolean algebras $(X, \vee, \wedge, 0, 1)$ and Boolean rings $(X, +, \star, 0, 1)$. Lawvere's model was intrinsically presentation-invariant. "Boolean algebras" had only one "algebraic theory", a certain category. The models of that theory, the Boolean algebras themselves, are certain set-valued functors from (the opposite category of) the theory. The presentation invariance is achieved by putting into the theory category *all* derived operations (modulo the equations); this idea was already well known in universal algebra. What is entirely new is that simple axioms suffice to construct an algebraic theory without any reference to a presentation. The effects are far-reaching in directions no one could have anticipated. We shall see later in this article how a school of thought initiated later by Elgot led to the use of algebraic theories to study the semantics of programming languages.

We shall now briefly describe Lawvere's algebraic theories and their models.

DEFINITION 3.10. Let $\mathcal{S}_0$ be the full subcategory of $\mathcal{S}$ of the finite cardinals. An *algebraic theory* is a pair $(\mathcal{A}, i)$ where $\mathcal{A}$ is a category and $i : \mathcal{S}_0 \to \mathcal{A}$ is a coproduct-preserving functor which is bijective on objects.

The next result shows that algebraic theories are closely related to Kleisli categories of a monad of sets, and also shows that it is very easy to get examples of both.

PROPOSITION 3.11. *Let $\mathcal{C}$ be any category and A be an object of $\mathcal{C}$ which has all finite copowers. Let $\mathcal{K}$ be the category whose objects are sets and whose morphisms $m \to n$ are $\mathcal{C}$-morphisms $m \cdot A \to n \cdot A$ (where $m \cdot A$ is a chosen m-fold copower of A). Let $\mathcal{A}$ be the full subcategory of $\mathcal{K}$ of all finite cardinals. Then the following statements hold.*

(1) *$\mathcal{A}$ is an algebraic theory and every algebraic theory has that form.*
(2) *$\mathcal{K}$ is isomorphic to the Kleisli category of a monad of sets and every such Kleisli category has that form.*

PROOF. To see that $\mathcal{A}$ is an algebraic theory, use the triangle

$$\begin{tikzcd} A \arrow[r, "in_j"] \arrow[dr, "in_{f(j)}"'] & m \cdot A \arrow[d, "i(f)"] \\ & n \cdot A \end{tikzcd}$$

to show how the functor i is defined on morphisms. By the coproduct property, a morphism $m \to n$ amounts to an function $m \to Tn$ where Tn is the set of all $\mathcal{C}$-morphisms $A \to n \cdot A$. η_n maps an element of n to its coproduct injection. An m-tuple $\alpha : m \to Tn$ induces a $\mathcal{C}$-morphism $\hat{\alpha} : m \cdot A \to n \cdot A$ by $\hat{\alpha}\, in_i = \alpha(i)$. Define $\alpha^{\#} f = \hat{\alpha} f$. □

For a variety of (Σ, E)-algebras, the free algebra functor $F_{\Sigma,E} : \mathcal{S} \to (\Sigma, E)$-**Alg** has a right adjoint, so is coproduct-preserving. In all but trivial cases, this functor is injective on objects. It follows that if $\mathcal{A}$ is the full subcategory of all (Σ, E)-algebras of form $F_{\Sigma,E}\, n$ with n a finite cardinal and if $i : \mathcal{S}_0 \to \mathcal{A}$ is the restriction of $F_{\Sigma,E}$ then $(\mathcal{A}, i)$ is an algebraic theory.

Universal algebraists view $F_{\Sigma,E}\, n$ as equivalence classes (under the equations) of all derived Σ-operations on the "variables" in n. A morphism $F_{\Sigma,E}\, m \to F_{\Sigma,E}\, n$ in the algebraic theory, by freeness, is tantamount to a function $m \to F_{\Sigma,E}\, n$, that is, a morphism is but an m-tuple of equivalence classes of derived n-ary operations. An algebra of this theory, then, should be a set X with an interpretation $X^n \to X^m$ (note the reversal of m, n) of each such morphism as an m-tuple of actual n-ary operations. This led Lawvere to define algebras as follows.

DEFINITION 3.12. If $(\mathcal{A}, i)$ is an algebraic theory, an *$(\mathcal{A}, i)$-algebra* is a functor $X : \mathcal{A}^{op} \to \mathcal{S}$ which preserves finite products. If X, Y are algebras, a *morphism* $\tau : X \to Y$ is a natural transformation.

Here, the "carrier set" is $X(1)$ and the function underlying the morphism τ is τ_1. For $\omega : i(1) \to i(n)$ in $\mathcal{A}$, the naturality square

$$\begin{tikzcd} X(1)^n \arrow[r, "\tau_1^n"] \arrow[d, "X\omega"'] & Y(1)^n \arrow[d, "Y\omega"] \\ X(1) \arrow[r, "\tau_1"'] & Y(1) \end{tikzcd}$$

shows that τ amounts to a function which "commutes with the operations". We know that $\tau_n = \tau_1^n$ because a special case of ω is the jth element of n, $j : 1 \to n$ for which $Y\omega$ is the jth product projection (since Y preserves finite products).

Hence Lawvere's presentation-invariant version of $U_{\Sigma,E} : (\Sigma, E)$-**Alg** $\to \mathcal{S}$ is the functor from $(\mathcal{A}, i)$-algebras to $\mathcal{S}$ which maps $X \mapsto X(1)$, $\tau \mapsto \tau_1$.

There is a monad-theoretic way to express an algebra as a representation of equivalence classes of derived operations. The earliest published reference of such a construction is [52, Section 3] and it is done at the level of closed categories.

DEFINITION 3.13. Let X be a set. Generalizing Example 2.11 by replacing 2 with X, define the *double dualization monad* $\mathbf{T}_X = (T, \eta, (-)^{\#})$ by

$$T_X\, m = X^{X^m}$$

$$\eta_m\, i = pr_i \quad (x_j \colon j \in m) \mapsto x_i \tag{13}$$

$$\alpha^{\#}(X^m \xrightarrow{f} X)(n \xrightarrow{g} X) = f(m \xrightarrow{\alpha} X^{X^n} \xrightarrow{pr_g} X)$$

The monad axioms can be directly verified. Alternatively, check that this monad arises from the left adjoint to the contravariant representable functor $\mathcal{S}(-, X) : \mathcal{S}^{op} \to \mathcal{S}$.

PROPOSITION 3.14. *Let X be a set with double dualization monad $\mathbf{T}_X$. Then the following two statements hold*:

(1) *For any functor $T : \mathcal{S} \to \mathcal{S}$, the passage from functions $\xi : TX \to X$ to natural transformations $\gamma : T \to T_X$ defined by*

$$\gamma_n : Tn \to X^{X^n}, \quad (\gamma_n\omega)(n \xrightarrow{f} X) = (Tn \xrightarrow{Tf} TX \xrightarrow{\xi} X)(\omega)$$

is well-defined and bijective with inverse

$$\gamma \mapsto TX \xrightarrow{\gamma_X} X^{(X^X)} \xrightarrow{pr_{id}} X$$

(2) *For any monad $\mathbf{T}$ in $\mathcal{S}$, the inverse passages above restrict to a bijection between $\mathbf{T}$-algebra structure maps and monad maps $\mathbf{T} \to \mathbf{T}_X$.*

PROOF. The first statement follows from the well-known Yoneda lemma first proved in [112]. Natural transformations $\mathcal{S}^{op}(-, X) \to F$ correspond bijectively to the elements of FX for any functor $F : \mathcal{S}^{op} \to \mathcal{S}$. Applying this to $F = X^{T(-)}$ gives $X^{TX} \cong n.t.(X^{(-)}, X^{T(-)})$ which can be recurried to the desired natural transformation $T \to T_X$. We turn now to the second statement. For $f \in X^n$, $pr_f\gamma_n\eta_n = \xi(Tf)\eta_n = \xi\eta_X f$, so γ satisfies the η-law for a monad map if and only if $\xi\eta_X f = f$ $\forall f \in X^n$ and this is clearly equivalent to the algebra law $\xi\eta_X = id_X$. Thus γ is a monad map if and only if for all $\alpha : m \to Tn$, the following square commutes:

$$\begin{tikzcd}
Tm \arrow[r, "\gamma_m"] \arrow[d, "\alpha^{\#}"'] & X^{X^m} \arrow[d, "(m \xrightarrow{\alpha} Tn \xrightarrow{\gamma_n} X^{X^n})^{\#}"] \\
Tn \arrow[r, "\gamma_n"'] & X^{X^n}
\end{tikzcd}$$

Chasing this diagram for $g : n \to X$ shows that the square above amounts to the perimeter of the following diagram:

$$\begin{tikzcd}
Tm \arrow[r, "\alpha^{\#}"] \arrow[d, "T\alpha"'] \arrow[dr, phantom, "\text{(A)}"] & Tn \arrow[r, "Tg"] \arrow[d, phantom, "\text{(B)}"] & TX \arrow[dd, "\xi"] \\
TTn \arrow[ur, "\mu_n"] \arrow[d, "TTg"'] & & \\
TTX \arrow[uurr, "\mu_X"] \arrow[r, "T\xi"'] & TX \arrow[r, "\xi"'] \arrow[ur, phantom, "\text{(C)}"] & X
\end{tikzcd}$$

In the diagram, (A) is (10), (B) is the naturality of μ and (C) is the algebra law in 2.18. Thus if (X, ξ) is a **T**-algebra, the perimeter commutes and γ is a monad map. Conversely, if the perimeter commutes it becomes the needed algebra law Proposition 2.20(3) if $g = id_X$. □

In short, the algebra (X, ξ) interprets the abstract n-ary operations $\omega \in Tn$ as an actual operation

$$X^n \xrightarrow{\xi_\omega} X, \quad \xi_\omega(f) = f^{\#}(\omega) \tag{14}$$

3.3. *Linton*

Beginning with [62] and followed up with a series of papers [63–69] Linton extended Lawvere's presentation-invariant definitions to allow for infinitary operations and for algebras over categories other than $\mathcal{S}$. Indeed, Linton was the first to go beyond bounded signatures. His thesis [61] explored the junction between category theory and analysis. Examples of "unbounded algebra" include compact groups (with Abelian ones as a subvariety) and many other examples relevant in analysis.

The equivalence of these extended theories with monads is stated variously among the papers, a formal announcement having been made quite early in [63]. Some details are given in [70].

In this section we treat the "semantic comparison functor" which provides the "best algebraic approximation" to a functor. The main ideas had already appeared in [20,55,62]. The version we develop here is tailored toward monads.

Let $U : \mathcal{A} \to \mathcal{K}$ be any functor. Using an example of the "comma category" notation introduced in Lawvere's thesis [55, pp. 16, 17], for K an object of $\mathcal{K}$ let (K, U) be the category with objects all (A, a) with $a : K \to UA$ in $\mathcal{K}$ and morphisms $f : (A, a) \to (B, b)$ such that the following triangle commutes.

$$\begin{tikzcd} K \arrow[r, "a"] \arrow[dr, "b"'] & UA \arrow[d, "Uf"] \\ & UB \end{tikzcd}$$

Thus, a free $\mathcal{A}$-object generated by K is simply an initial object of (K, U).

DEFINITION 3.15. A functor $U : \mathcal{A} \to \mathcal{K}$ is *tractable at* the $\mathcal{K}$-object K if the functor $\Delta_K : (K, U) \to \mathcal{K}$ mapping (A, a) to UA and $f : (A, a) \to (B, b)$ to $Uf : UA \to UB$ has a limit. U is *tractable* if it is tractable at each $\mathcal{K}$-object.

Because of its importance, let's review the definition of limit in this context. A *limit* of Δ_K is a pair (TK, ψ^K) where ψ^K assigns a $\mathcal{K}$-morphism $\psi^K_{A,a} : TK \to UA$ to each (A, a) in (K, U) so that the following triangle commutes for all $f : (A, a) \to (B, b)$:

$$\begin{tikzcd} TK \arrow[r, "\psi^K_{A,a}"] \arrow[dr, "\psi^K_{B,b}"'] & UA \arrow[d, "Uf"] \\ & UB \end{tikzcd} \tag{15}$$

The universal mapping property of the limit is that for every other "cone" (L, φ) as shown in the diagram on the left below

$$\begin{tikzcd} L \arrow[r, "\varphi_{A,a}"] \arrow[dr, "\varphi_{B,b}"'] & UA \arrow[d, "Uf"] \\ & UB \end{tikzcd} \qquad \begin{tikzcd} TK \arrow[r, "\psi^K_{A,a}"] & UA \\ L \arrow[u, "g"] \arrow[ur, "\varphi_{A,a}"'] & \end{tikzcd} \tag{16}$$

there exists a unique $\mathcal{K}$-morphism $g : L \to TK$ such that for all (A, a) in (K, U) the triangle on the right above commutes.

PROPOSITION 3.16. *If $U : \mathcal{A} \to \mathcal{K}$ has a left adjoint $F : \mathcal{K} \to \mathcal{A}$ then U is tractable.*

PROOF. Define $TK = UFK$, $\psi^K_{A,a} = Ug$ for the unique $g : FK \to A$ with $(Ug)\eta_A = a$. □

If $\mathcal{K}$ is small-complete and $\mathcal{A}$ is small, Δ_K is a small diagram so U is tractable. In this case U is very unlikely to have a left adjoint, however. For example, if for every cardinal n there exists a $\mathcal{K}$-object K_n with at least n endomorphisms, then a free $\mathcal{A}$-object generated by K_n would also have at least n endomorphisms. This cannot happen for all n if $\mathcal{A}$ is small category.

What we have learned so far is that if U is tractable at K with limit (TK, ψ^K) of Δ_K, TK will be (U of) the free object generated by K if that exists. So we may regard TK as the "best approximation to U of the free object". With this in mind, we characterize TK in terms of operations when $\mathcal{K} = \mathcal{S}$. This result was essentially in Lawvere's thesis [55] and plays a prominent role in [62].

PROPOSITION 3.17. *If $U : \mathcal{A} \to \mathcal{S}$ is tractable at n with $\lim \Delta_n = (Tn, \psi^n)$ then $Tn \cong n.t.(U^n, U)$.*

PROOF. For (A, a) in (n, U) define

$$\begin{aligned} &\psi^n_{A,a} : n.t.\big(U^n, U\big) \to UA \\ &U^n \xrightarrow{\gamma} U \mapsto \gamma_A(a) \end{aligned}$$

Given a cone (L, φ) as in (16), define $g : L \to n.t.(U^n, U)$ by mapping $x \in L$ to the natural transformation γ by defining $\gamma_A : UA^n \to UA$, $a \mapsto \varphi_{A,a}(x)$. □

While tractability is not directly invoked in the proof above, without it $n.t.(U^n, U)$ would fail to be a small set.

For $U = U_{\Sigma,E} : (\Sigma, E)$-**Alg** $\to \mathcal{S}$, each derived operation is defined on each algebra and commutes with all homomorphisms so is a natural transformation $U^n \to U$. Lawvere and Linton stressed that there are no other such natural transformations. Perhaps the sharpest statement of this appears in [42].

The next construction is due to [1] who called the monadic completion a "codensity triple".

DEFINITION 3.18. Let $U : \mathcal{A} \to \mathcal{K}$ be a tractable functor, with $\lim \Delta_K = (TK, \psi^K)$. The *monadic completion* of U is the monad $(T, \eta, (-)^\#)$ defined by the universal properties of (T, ψ) as follows. The (A, a)-indexed family $a : K \to UA$ is itself a cone, and so gives rise to η by

$$\begin{tikzcd} TK \arrow[r, "\psi^K_{A,a}"] & UA \\ K \arrow[u, "\eta_K"] \arrow[ur, "a"'] & \end{tikzcd}$$

For $\alpha : K \to TL$, (TK, ψ_{A,a_α}) is a cone if for $a : L \to UA$, $a_\alpha = K \xrightarrow{\alpha} TL \xrightarrow{\psi^L_{A,a}} UA$, so there exists unique $\alpha^\#$ with

$$\begin{tikzcd} TL \arrow[r, "\psi^L_{A,a}"] & UA \\ TK \arrow[u, "\alpha^\#"] \arrow[ur, "\psi^K_{A,a_\alpha}"'] & \end{tikzcd}$$

Let's check the monad axioms. As $\psi^L_{A,a}(\alpha^\# \eta_K) = \psi^K_{A,a_\alpha}\, \eta_K = \psi^L_{A,a}\alpha$, it follows from the uniqueness property that $\alpha^\# \eta_K = \alpha$. As $\psi^K_{A,a}(\eta_K)^\# = \psi^K_{A,\psi^K_{A,a}}\, \eta_K = \psi^K_{A,a}$, it follows from the universal property that $(\eta_K)^\# = id_{TK}$. Now let $\alpha K \to TL$, $\beta : L \to TM$, $a : M \to UA$. We have

$$\begin{aligned} \psi^M_{A,a}\beta^\#\alpha^\# = \psi^L_{A,(\psi^M_{A,a}\beta)}\alpha^\# &= \psi^K_{A,(\psi^L_{A,\psi^M_{A,a}\beta}\alpha)} \\ &= \psi^K_{A,(\psi^M_{A,a}\beta^\#\alpha)} = \psi^M_{A,a}\big(\beta^\#\alpha\big)^\# \end{aligned}$$

so $\beta^\#\alpha^\# = (\beta^\#\alpha)^\#$.

EXAMPLE 3.19 (*cf.* [48, Theorem 3.1]). Let $U : \mathcal{A} \to \mathcal{S}$ be the inclusion of the 301-morphism full subcategory with objects $1, 2$ and 2×2. Let $\mathbf{T} = (T, \eta, (-)^\#)$ be the monadic completion of U.

We must figure out what $\mathbf{T}$ is. To this end, we know $Tn = n.t.(U^n, U)$. This is a small set because, as $\mathcal{A}$ is small, U is tractable. If $\omega : U^n \to U$ is natural, $\omega_2 : 2^n \to 2$ is an n-ary operation which we claim is a Boolean algebra homomorphism, that is, an ultrafilter on the set n. Verification is as follows.

If $x : 1 \to 2$ is a constant, the naturality square

$$\begin{tikzcd} 1^n \arrow[r, "\omega_1"] \arrow[d, "x^n"'] & 1 \arrow[d, "x"] \\ 2^n \arrow[r, "\omega_2"'] & 2 \end{tikzcd}$$

shows that $\omega_2(0) = 0$, $\omega_2(1) = 1$. Similarly, the naturality square induced by the negation map $\neg : 2 \to 2$ shows that ω_2 preserves negation. Now consider the diagram

$$\begin{tikzcd} 2^n \times 2^n \arrow[r, "\psi"] \arrow[dr, "\wedge"'] & (2 \times 2)^n \arrow[r, "\omega_{2\times 2}"] \arrow[d, "\wedge^n"] & 2 \times 2 \arrow[d, "\wedge"] \\ & T \arrow[r, "\omega_2"'] & 2 \end{tikzcd}$$

Here ψ is the obvious bijection. Since $\omega_{2\times 2}\,\psi = \omega_2 \times \omega_2$ (consider the naturality squares induced by the $pr_i : 2 \times 2 \to 2$), it follows from the diagram that ω_2 preserves disjunction. Writing βn for the set of ultrafilters on n (which is a natural notation since this space is the beta-compactification of the discrete space n), we have so far described a function $\Gamma : n.t.(U^n, U) \to \beta n$, $\omega \mapsto \{A \subset n \colon \omega_2(A) = \mathit{True}\}$. Here, $2 = \{\mathit{False}, \mathit{True}\}$ and we identify A with its characteristic function $n \to 2$.

Now define a function in the reverse direction $\Lambda : \beta n \to n.t.(U^n, U)$ as follows. For A one of the three objects of $\mathcal{A}$, think of A^n as the compact Hausdorff space resulting from the product topology induced by the discrete topology on the finite set A. For $\mathcal{U} \in \beta n$, let $\Lambda(\mathcal{U})_A : A^n \to A$ map $f : n \to A$ to the unique point in A to which the ultrafilter $f(\mathcal{U}) = \{B \subset A \colon f^{-1}(B) \in \mathcal{U}\}$ on A converges.

We leave it as an exercise for the reader to show that Γ, Λ are mutually inverse so that the monadic completion of U has form $(\beta, \eta, (-)^\#)$. We also leave it to the reader to unwind the definitions to show that for $\alpha : X \to \beta Y$,

$$\begin{aligned} \eta_X(x) &= prin(x) = \{A \subset X \colon x \in A\} \\ \alpha^\#(\mathcal{U}) &= \big\{B \subset Y \colon \{x \in X \colon B \in \alpha x\} \in \mathcal{U}\big\} \end{aligned} \tag{17}$$

This defines the *ultrafilter monad* in $\mathcal{S}$, which we denote henceforth as $\boldsymbol{\beta}$. A different construction of this monad will be given in Example 4.6. The β-algebras are precisely the category of compact Hausdorff spaces and continuous maps. Originally proved in the author's thesis [74], various proofs have appeared in textbooks [73, p. 153], [76, pp. 60–63], [45, Section 5.2] and journal articles [90,100]. Also, see [2,18,33]. It is well-known that a topological space is compact T2 if and only if each ultrafilter converges uniquely and, indeed, the structure map $\xi : \beta X \to X$ is ultrafilter convergence. A β-homomorphism f satisfies "if $\mathcal{U}$ is an ultrafilter converging to x then $f(\mathcal{U}) = \{B \subset Y\colon\ f^{-1}B \in \mathcal{U}\}$ converges to fx" and this is a well-known characterization of continuity.

DEFINITION 3.20. Let $U : \mathcal{A} \to \mathcal{S}$ be a tractable functor with monadic completion $\mathbf{T}$. The *semantic comparison functor* is the functor over $\mathcal{K}$, $\Phi : \mathcal{A} \to \mathcal{K}^{\mathbf{T}}$, defined by $\Phi(A) = (UA, \rho_A)$ where $\rho_A = \psi^{UA}_{A,id_{UA}}$.

Here "over $\mathcal{K}$" means, recall, that the following diagram of functors commutes:

$$\begin{tikzcd} \mathcal{A} \arrow[rr, "\Phi"] \arrow[dr, "U"'] & & \mathcal{K}^{\mathbf{T}} \arrow[dl, "U^{\mathbf{T}}"] \\ & \mathcal{K} & \end{tikzcd}$$

PROPOSITION 3.21. *For a tractable functor $U : \mathcal{A} \to \mathcal{K}$, the following hold.*

(1) *The semantic comparison functor of the definition immediately preceding is well-defined and $\rho : TU \to U$ is a natural transformation.*
(2) *If U has a left adjoint then the monadic completion of U is isomorphic to the monad induced by Huber's theorem and the semantic comparison functor preserves the free functors as well, that is, the following triangle commutes.*

$$\begin{tikzcd} \mathcal{A} \arrow[rr, "\Phi"] & & \mathcal{K}^{\mathbf{T}} \\ & \mathcal{K} \arrow[ul, "F"] \arrow[ur, "F^{\mathbf{T}}"'] & \end{tikzcd}$$

(3) *For $a : K \to UA$, $\psi^K_{A,a} = \alpha^{\#} : (TK, \mu_K) \to \Phi A$.*

PROOF. (1) We first show that $\Phi A = (UA, \rho_A)$, $\rho_A = \psi^{UA}_{A,id_{UA}}$ is a $\mathbf{T}$-algebra. The η-law is immediate from the definition of η in Definition 3.18. Now let $\alpha, \beta : K \to TUA$ with $\rho_A \alpha = \rho_A \beta$. By the definition of $(-)^{\#}$ in 3.18 we have

$$\rho_A \alpha^{\#} = \psi^{UA}_{A,id_{UA}} \alpha^{\#} = \psi^K_{A,\rho_A \alpha} = \psi^K_{A,\rho_A \beta} = \rho_A \beta^{\#}$$

Thus Φ is well-defined on objects. Let $f : A \to B$ be a morphism in $\mathcal{A}$. We must show that $Uf : (UA, \rho_A) \to (UB, \rho_B)$ is a $\mathbf{T}$-homomorphism. This amounts to the commutativity of the square

$$\begin{tikzcd}
TUA \arrow[r, "TUf"] \arrow[d, "\rho_A"'] & TUB \arrow[d, "\rho_B"] \\
UA \arrow[r, "Uf"'] & UB
\end{tikzcd}$$

which is equivalent to the assertion that $\rho : TU \to U$ is natural. We have

$$\rho_B(TUf) = \psi^{UB}_{B,id_{UB}}\big(\eta_{UB}(Uf)\big)^{\#} = \psi^{UA}_{B,a}$$

where

$$a = \psi^{UB}_{B,id_{UB}}\big(\eta_{UB}(Uf)\big) = \big(\psi^{UB}_{B,id_{UB}}\eta_{UB}\big)(Uf) = id_{UB}(Uf) = Uf$$

so that $\rho_B(TUf) = \psi^{UB}_{B,Uf}$. As ψ is a limit cone we have the commutative triangle

$$\begin{tikzcd}
TUA \arrow[r, "\psi^{UA}_{A,id_{UA}}"] \arrow[dr, "\psi^{B}_{B,Uf}"'] & UA \arrow[d, "Uf"] \\
 & UB
\end{tikzcd}$$

so $\rho_B(TUf) = (Uf)\rho_A$ as desired.

(2) Now suppose $(\mathcal{A}, \mathcal{K}, U, F, \eta, \varepsilon)$ is an adjointness. Define the monadic completion $T = UF$ and ψ as in the proof of Proposition 3.16 so that $\psi^K_{A,a}$ is Ug if $g : FK \to A$ satisfies $(Ug)\eta_K = a$. Clearly the η of Definition 3.18 is the original one from the adjointness. Given $a : L \to UA$, $\alpha : K \to TL$, $\psi^L_{A,a} = Ug$ if $(Ug)\eta_L = a$ and $\alpha^{\#}$ (as induced by the adjointness) $= Uh$ if $(Uh)\eta_K = \alpha$ so $U(gh)\eta_K = (Ug)((Uh)\eta_K) = (Ug)\alpha = \psi^A_{A,a}\alpha = a_\alpha$ and $\psi^L_{A,a}\alpha^{\#} = (Ug)(Uh) = U(gh) = \psi^K_{A,a_\alpha}$. This shows that the adjointness-induced $\alpha^{\#}$ also satisfies the definition of $\alpha^{\#}$ in Definition 3.18. Hence the two monads are the same. Moreover, $TK \xrightarrow{\eta_{TK}} UFUFK \xrightarrow{U\varepsilon_{FK}} TK = id_{TK}$ shows that $\rho_{FK} = \psi^{TK}_{FK,id_{TK}} = (id_{TK})^{\#} = \mu_K$ so that $\Phi FK = (TK, \rho_{FK}) = (TK, \mu_K) = F^{\mathbf{T}}K$ as desired. We'll leave it to you the reader to check that $\Phi F = F^{\mathbf{T}}$ on morphisms as well.

(3) We know that $\psi^K_{A,a}\eta_K = a$, so it remains to show that for $a : K \to UA$, the following square commutes:

$$\begin{tikzcd}
TTK \arrow[r, "T\psi^K_{A,a}"] \arrow[d, "\mu_K"'] & TUA \arrow[d, "\psi^{UA}_{A,id_{UA}}"] \\
TK \arrow[r, "\psi^K_{A,a}"'] & UA
\end{tikzcd}$$

We compute both paths. On the one hand,

$$\psi^K_{A,a}\mu_K = \psi^K_{A,a}(id_{TK})^\# = \psi^{TK}_{A,\psi^K_{A,a}}$$

On the other hand,

$$\psi^{UA}_{A,id_{UA}}\big(T\psi^K_{A,a}\big) = \psi^{UA}_{A,id_{UA}}\big(\eta_{UA}\psi^K_{A,a}\big)^\# = \psi^{TK}_{A,\psi^K_{A,a}}$$

the same result. □

The universal property of semantic comparison must wait for Theorem 3.36.

3.4. *Monadic functors*

All mathematicians know that categories of algebras have special properties. For example a bijective homomorphism is an isomorphism. In topology, however, the student is taught to beware that a bijective continuous map is not always a homeomorphism unless the spaces are compact Hausdorff; but then the spaces *are* algebras, the algebras of the ultrafilter monad.

The goal of this subsection is lofty: to characterize algebraic structure in categorical terms.

The first theorem of this type (for finitary algebras) is in Lawvere's thesis [55, Theorem 1, p. 79]. Upgraded versions of this theorem which apply to infinitary algebras are [40, Theorem 5.2, p. 18] and the theorem of Linton below. We will state Lawvere's theorem after giving the necessary definitions.

Say that an object G of a category $\mathcal{A}$ is a *generator* if whenever $f \neq g : A \to B$ there exists $x \in A$ with $fx \neq fy$. Here "$x \in A$" means that x is a morphism $x : G \to A$. Say that an object G is *abstractly finite* if the n-fold copower $n \cdot G$ exists for every cardinal n and every $x \in n \cdot G$ factors through $F \cdot G$ for some finite subset F of n.

Let $p, q : A \to B$ in a category with a chosen object G. Say that (p, q) is *G-reflexive* if $\forall y \in B\ \exists x \in A$ with $px = y = qx$. Say that (p, q) is *G-symmetric* if $\forall x \in A\ \exists x' \in A$ with $px = qx'$, $qx' = px$. Say that (p, q) is *G-transitive* if given $x, y \in A$ with $qx = py$ there exists $z \in A$ with $pz = px$ and $qz = qy$. Finally, say that (p, q) is a *G-equivalence relation on A* if it is jointly monic, reflexive, symmetric and transitive.

Recall that a pair $p, q : A \to B$ of morphisms in a category is a *kernel pair* if there is a pullback square

$$\begin{tikzcd} A \arrow[r, "p"] \arrow[d, "q"'] & B \arrow[d, "f"] \\ B \arrow[r, "f"'] & Q \end{tikzcd}$$

for some f and that a morphism *is a coequalizer* if it is the coequalizer of some pair of maps.

Lawvere's theorem was stated for algebras (functors) over his theories. We state it in more conventional universal-algebraic terms.

THEOREM 3.22. *A category $\mathcal{A}$ is equivalent to a variety of finitary algebras if and only if it has finite limits and it has an abstractly finite generator G which is projective with respect to coequalizers, is such that $\{x\colon\ x \in A\}$ is a small set for each object A, and is such that for all cardinals n, every G-equivalence relation on $n \cdot G$ is a kernel pair.*

Of course the underlying set functor is $UA = \{x\colon\ x \in X\}$ which is just the functor represented by G. In the language of universal algebra, G-equivalence relations and kernel pairs are both congruences. In the category of topological spaces with G the one-point space, a G-equivalence relation is an equivalence relation whose inclusion into the product space is continuous, but a kernel pair must have the subspace topology.

We next state Linton's characterization theorem [62, Proposition 3, p. 88].

THEOREM 3.23. *Let $U : \mathcal{A} \to \mathcal{S}$ be tractable. Then U is equivalent to a variety of* (*not necessarily finitary*) *algebras if and only if U has a left adjoint, $\mathcal{A}$ has kernel pairs and coequalizers of pairs, a morphism h is a coequalizer if and only if Uh is surjective and $p, q : A \to B$ is a kernel pair if and only if $[Up, Uq] : UA \to UB \times UB$ is monic and its image is an equivalence relation on UB.*

The theorems of Lawvere and Linton are limited to algebraic structure over $\mathcal{S}$. In his thesis [4], Beck introduced a theory of *monadic functors* $\mathcal{A} \to \mathcal{K}$ (at that time, he called them "tripleable") . We shall see later in this section that when $\mathcal{K}$ is $\mathcal{S}$, varieties of algebras as in Linton's theorem result.

DEFINITION 3.24. A functor $U : \mathcal{A} \to \mathcal{K}$ is *weakly monadic* if there exists a monad $\mathbf{T}$ in $\mathcal{K}$ and an equivalence of categories Ψ over $\mathcal{K}$ as shown in the triangle below:

$$\begin{tikzcd} \mathcal{A} \arrow[rr, "\Psi"] \arrow[dr, "U"'] & & \mathcal{K}^{\mathbf{T}} \arrow[dl, "U^{\mathbf{T}}"] \\ & \mathcal{K} & \end{tikzcd}$$

If Ψ is even an isomorphism of categories, U is *monadic*.

Beck proposed [4, p. 73] that monadicity

> "... be regarded as a new type of mathematical structure, parallel to but not necessarily definable in terms of other known types of structure, such as algebraic, equational, topological, ordered, ..."

Monadic over $\mathcal{K}$ means "algebraic over $\mathcal{K}$"; Beck's new idea is that this is a relative term.

It is clear that a monadic functor must have a left adjoint. The theorems of Lawvere and Linton already suggest that additional conditions hinge on the construction of quotients – not surprising, since an arbitrary algebra is a quotient of a free one. For algebraic categories over $\mathcal{S}$, quotients are constructed as coequalizers of kernel pairs and the theory surrounding this is generally known as the "first isomorphism theorem" in algebra.

In thinking about algebras of a monad, Beck isolated a new type of quotient construction. To begin, observe that any $\mathbf{T}$-algebra (X,ξ) in $\mathcal{K}^{\mathbf{T}}$ gives rise to the following commutative diagram in $\mathcal{K}$.

$$\begin{tikzcd} TX \arrow[r, "\eta_{TX}"] \arrow[d, "\xi"'] \arrow[dr, phantom, "(\eta\text{ n.t.})"] & TTX \arrow[r, "\mu_X"] \arrow[d, "T\xi"] \arrow[dr, phantom, "(\mu\text{-law})"] & TX \arrow[d, "\xi"] & = id_{TX} \\ X \arrow[r, "\eta_X"'] & TX \arrow[r, "\xi"'] & X & = id_X \end{tikzcd} \tag{18}$$

(the notation "$=id_{TX}$" etc. means that the row composes to an identity morphism). A natural abstraction is the following commutative diagram:

$$\begin{tikzcd} L \arrow[r, "d"] \arrow[d, "q"'] & K \arrow[r, "f"] \arrow[d, "g"] & L \arrow[d, "q"] & = id_L \\ Q \arrow[r, "e"'] & L \arrow[r, "q"'] & Q & = id_Q \end{tikzcd} \tag{19}$$

An easy way to remember the diagram above is that $(f,q)\colon g \to q$ is split epic in the diagram category $\mathcal{K}^{\to}$.

PROPOSITION 3.25. *Given the diagram* (19) *in any category* $\mathcal{K}$, $q = coeq(f,g)$ *is the coequalizer of* f *and* g.

PROOF. That $qf = qg$ is given. If $rf = rg$ we must construct unique $\psi\colon Q \to R$ with $\psi q = r$. As q is given (split) epic, uniqueness is clear. For existence, define $\psi = re$. Then $\psi q = req = rgd = rfd = rid_L = r$. □

In (19) we call (f,g,q) a *contractible coequalizer* (with *contraction* (d,e)). If $G\colon\mathcal{K} \to \mathcal{L}$ is any functor, (Gf, Gg, Gq) is again a contractible coequalizer with contraction (Gd, Ge) because all functors preserve commutative diagrams and identity morphisms. A universal construction preserved by all functors is said to be *absolute*. In [90], Paré proved that all absolute colimits have that property by virtue of a commutative diagram.

The relationship between contractible coequalizers and coequalizers of kernel pairs is very direct in $\mathcal{S}$. Let $E \subset X \times X$ be an equivalence relation on the set X with projections $a, b : E \to X$ and let $q : X \to X/E$ be the canonical projection to the quotient set. Then (a, b) is the kernel pair of q because $xEy \Leftrightarrow q(x) = q(y)$; moreover, $q = coeq(a, b)$ is the coequalizer of a, b in $\mathcal{S}$. By the axiom of choice, there exists a section $e : X/E \to X$ with $qe = id_{X/E}$. As $q(eq) = (qe)q = id_{X/E}q = q\,id_X$, it follows from the pullback property of the kernel pair that there exists $d : X \to E$ with $fd = id_X$, $eq = gd$ – that is, $q = coeq(a, b)$ is a contractible coequalizer in $\mathcal{S}$ with contraction (d, e).

Beck's characterization theorem for monadic functors to be presented shortly may be roughly paraphrased "a functor $\mathcal{A} \to \mathcal{K}$ is monadic if and only if every object in $\mathcal{A}$ is a contractible-coequalizer-quotient of a free object". From this perspective, the classical theory of quotient sets in algebra is based on the special property that epics split in $\mathcal{S}$. The use of contractible coequalizers not only avoids the axiom of choice but works in arbitrary categories.

DEFINITION 3.26. Let $U : \mathcal{A} \to \mathcal{K}$ be a functor. A pair $f, g : a \to B$ of $\mathcal{A}$-morphisms is *U-contractible* if there exists a contractible coequalizer

$$\begin{tikzcd} UB \arrow[r, "d"] \arrow[d, "q"'] & UA \arrow[r, "Uf"] \arrow[d, "Ug"] & UB \arrow[d, "q"] & = id_{UB} \\ Q \arrow[r, "e"'] & UB \arrow[r, "q"'] & Q & = id_Q \end{tikzcd}$$

in $\mathcal{K}$. *U creates coequalizers of U-contractible pairs* if given a U-contractible pair f, g with contractible coequalizer in $\mathcal{K}$ as shown above, there exists a unique morphism $\bar{q} : B \to Q$ in $\mathcal{A}$ with $U\bar{q} = q$; and, moreover, $\bar{q} = coeq(f, g)$ in $\mathcal{A}$.

For the underlying set functor U from topological spaces to $\mathcal{S}$, if f, g is U-contractible with contractible coequalizer in $\mathcal{S}$ as above, the quotient topology on Q is the unique lift $\overline{q}$ of q which is the coequalizer of f, g but it is not the unique lift of q since any topology coarser than the quotient topology is also a lift. Thus U does not create coequalizers of U-contractible pairs.

Before stating the Beck theorem, we isolate a useful Lemma.

LEMMA 3.27. *Let $\mathbf{T}$ be a monad in $\mathcal{K}$, let (X, ξ) be a $\mathbf{T}$-algebra and let $q : X \to Q$ be split epic in $\mathcal{K}$. Suppose that the following square commutes.*

$$\begin{tikzcd} TX \arrow[r, "Tq"] \arrow[d, "\xi"'] & TQ \arrow[d, "\theta"] \\ X \arrow[r, "q"'] & Q \end{tikzcd}$$

Then (Q, θ) *is a* $\mathbf{T}$*-algebra and* $q : (X, \xi) \to (Y, \theta)$ *is a final map.*

"Final" means that if (S, σ) is a $\mathbf{T}$-algebra and $g : Q \to S$ in $\mathcal{K}$ is such that $gq : (X, \xi) \to (S, \sigma)$ is a $\mathbf{T}$-homomorphism then necessarily $q : (Y, \theta) \to (S, \sigma)$ is a $\mathbf{T}$-homomorphism. For example, quotient maps are final in the category of topological spaces.

PROOF. As η is natural, $\theta\eta_Q q = \theta(Tq)\eta_X = q\xi\eta_X = q = id_Q q$ so, as q is epic, $\theta\eta_Q = id_Q$. For the μ-law, consult the diagram below. The four inner parallelograms not (A) and the perimeter all

$$
\begin{tikzcd}
TTX \arrow[rrr, "T\xi"] \arrow[dr, "TTq"] \arrow[ddd, "\mu_X"'] & & & TX \arrow[dl, "Tq"'] \arrow[ddd, "\xi"] \\
& TTQ \arrow[r, "T\theta"] \arrow[d, "\mu_Q"'] \arrow[dr, phantom, "(A)"] & TQ \arrow[d, "\theta"] & \\
& TQ \arrow[r, "\theta"'] & Q & \\
TX \arrow[ur, "Tq"] \arrow[rrr, "\xi"'] & & & X \arrow[ul, "q"']
\end{tikzcd}
$$

commute. But then (A) also commutes because TTq is epic. Thus $q : (X, \xi) \to (Q, \theta)$ is a $\mathbf{T}$-homomorphism. To prove that it is a final map, let (Y, ν) be a $\mathbf{T}$-algebra, and let $h : Q \to Y$ be such that $hq : (X, \xi) \to (Y, \nu)$ is a $\mathbf{T}$-homomorphism. Then

$$\nu(Th)(Tq) = \nu T(hq) = hq\xi = h\theta(Tq)$$

As Tq is epic, $\nu(Th) = h\theta$, so $h : (Q, \theta) \to (Y, \nu)$ is a $\mathbf{T}$-homomorphism. $\square$

We are now ready to state and prove the *Beck tripleableness theorem*. The essence of the theorem appears in Beck's thesis [4, Theorem 1, p. 10].

THEOREM 3.28. *Let* $U : \mathcal{A} \to \mathcal{K}$ *be a functor. The* U *is monadic if and only if* U *has a left adjoint and creates coequalizers of* U*-contractible pairs.*

PROOF. Suppose first that U is monadic. There is no loss of generality in assuming $U = U^{\mathbf{T}}$ for a monad $\mathbf{T}$ in $\mathcal{K}$. U has a left adjoint by Proposition 2.19. Now let $f, g : (A, \nu) \to$

(X, ξ) be $\mathbf{T}$-homomorphisms with contractible coequalizer $q = coeq(f, g) : X \to Q$ in $\mathcal{K}$. As contractible coequalizers are absolute, $Tq = coeq(Tf, Tg)$ in $\mathcal{K}$. As f, g are homomorphisms,

$$q\xi(Tf) = qf\nu = qg\nu = q\xi(Tg)$$

which induces a unique morphism $\tau : TQ \to Q$ in $\mathcal{K}$ with $\tau(Tq) = q\xi$. By Lemma 3.27, (Q, τ) is a $\mathbf{T}$-algebra and $q = coeq(f, g)$ in $K^{\mathbf{T}}$. Also, $q : (X, \xi) \to (Q, \tau)$ is the only lift because Tq is epic.

Conversely, Let $(\mathcal{A}, \mathcal{K}, U, F, \eta, \varepsilon)$ be an adjointness and let U create coequalizers of U-contractible pairs. Let $\Phi : \mathcal{A} \to K^{\mathbf{T}}$ be the semantic comparison functor. We shall show that Φ is an isomorphism of categories. Let (K, ξ) be a $\mathbf{T}$-algebra. Then $\varepsilon_{FK}, F\xi : FTK \to FK$ is a U-contractible pair of $\mathcal{A}$-morphisms with contractible coequalizer

$$\begin{tikzcd}
TK \arrow[r, "\eta_{TK}"] \arrow[d, "\xi"'] & TTK \arrow[r, "\mu_K"] \arrow[d, "T\xi"] & TK \arrow[d, "\xi"] & = id_{TK} \\
K \arrow[r, "\eta_K"'] & TK \arrow[r, "\xi"'] & K & = id_K
\end{tikzcd}$$

noting that $T\xi = UF\xi$, $\mu_K = U\varepsilon_K$. As U creates coequalizers of U-contractible pairs, there exists unique $\bar{\xi} : FK \to A$ in $\mathcal{A}$ with $U\bar{\xi} = \xi$ (and $UA = K$ in particular). As $(U\bar{\xi})\eta_K = \xi\eta_K = id_K$, $\bar{\xi} = (id_K)^{\#} = \varepsilon_A$. Thus $\Phi A = (UA, U\varepsilon_A) = (K, \xi)$, and Φ is surjective on objects. Applying similar constructions to the algebra $\Phi A = (UA, U\varepsilon_A)$ shows that $U\varepsilon_A = coeq(U\varepsilon_{FUA}, UFU\varepsilon_A)$ is a contractible coequalizer so that ε_A is the unique morphism ψ with $U\psi = U\varepsilon_A$. Additionally, ε_A is the coequalizer of $\varepsilon_{FUA}, FU\varepsilon_A$, a fact we shall need shortly. Thus $\Phi A = \Phi B \Rightarrow \varepsilon_A = \varepsilon_B \Rightarrow A = B$, so now Φ is bijective on objects. Let $f, g : A \to B$ with $Uf = Ug$. As ε is natural, $f\varepsilon_A = \varepsilon_B(FUf) = \varepsilon_B(FUg) = g\varepsilon_A$. But we just saw that ε_A is epic, so it follows that $f = g$. This shows that U is a faithful functor. As $U^{\mathbf{T}}\Phi = U$, Φ is faithful as well. It remains only to show that Φ is full. Let $f : \Phi A \to \Phi B$ be a $\mathbf{T}$-homomorphism. By definition, the following square commutes.

$$\begin{tikzcd}
TUA \arrow[r, "UFf"] \arrow[d, "U\varepsilon_A"'] & TUB \arrow[d, "U\varepsilon_B"] \\
UA \arrow[r, "f"'] & UB
\end{tikzcd}$$

As discussed above, $\varepsilon_A = coeq(\varepsilon_{FUA}, FU\varepsilon_A)$ in $\mathcal{A}$, and U preserves this coequalizer. As is routinely checked, there exists a unique $\bar{f} : A \to B$ with $\bar{f}\varepsilon_A = \varepsilon_B(Ff)$ by the coequalizer property. As $(U\bar{f})(U\varepsilon_A) = f(U\varepsilon_A)$ we must have $U\bar{f} = f$. Thus $\Phi\bar{f} = f$ and Φ is full. $\square$

[17,90] prove variants of Beck's theorem. In this regard, also see the survey in [3, p. 139].

As a prelude to the next proposition, let us call an equational presentation (Σ, E) *tractable* if $U_{\Sigma,E}$ is. When U_Σ is not tractable, the Σ-equations of Definition 3.3 involve elements of large sets, but there is no reason why the definitions are invalid (a "higher universe" for $F_\Sigma n$ is needed). The point is that (Σ, E) can be tractable even if Σ isn't. As a trivial example, take any tractable (Σ, E) and add a proper class of new unary operations u each satisfying the equation $ux = x$. It is also possible for a class of (Σ, E)-algebras to be of interest without being tractable. It was independently proved by [30,35] (which both appeared in the same journal volume) that complete Boolean algebras are not tractable at $\aleph_0$. Also see [105]. Thus we shall try to develop theory in the non-tractable case when possible.

PROPOSITION 3.29. *Let (Σ, E) be a (not necessarily tractable) equational specification. Then $U_{\Sigma,E} : (\Sigma, E)$-**Alg** $\to \mathcal{S}$ creates coequalizers or $U_{\Sigma,E}$-contractible pairs. If (Σ, E) is tractable, $U_{\Sigma,E}$ is monadic.*

As a result of this proposition, we now have a great many examples of set-valued monadic functors. The proof of the coequalizer condition is simple and adapts well to many other situations.

PROOF. Write $U = U_{\Sigma,E}$. Let $f, g : A \to B$ in $\mathcal{A}$ and consider a contractible coequalizer in $\mathcal{K}$ as follows:

$$\begin{tikzcd}
UB \arrow[r, "d"] \arrow[d, "h"'] & UA \arrow[r, "Uf"] \arrow[d, "Ug"] & UB \arrow[d, "h"] & = id_{UB} \\
C \arrow[r, "e"'] & UB \arrow[r, "h"'] & C & = id_C
\end{tikzcd}$$

By the standard theory of quotient algebras, it suffices to prove that the equivalence relation of h is a subalgebra of $B \times B$ since it is then a congruence whose quotient algebra structure is the unique lift of h which is moreover the coequalizer of f, g in $\mathcal{A}$. To this end, let E be the equivalence relation of g which is a congruence because g is a homomorphism. Let R be the image of $E \xrightarrow{i} A \times A \xrightarrow{f\times f} B \times B$ which is a subalgebra of $B \times B$ because f is a homomorphism. To show: $xRy \Leftrightarrow hx = hy$. First let $x\ R\ y$. We will write f, g instead of the fussy Uf, Ug. There exist $a_x, a_y \in A$ with $fa_x = x$, $fa_y = y$, $ga_x = ga_y$. Thus $hx = hfa_x = ga_x = hga_y = hfa_y = hy$. Conversely, let $hx = hy$. Define $a_x = dx$, $a_y = dy$. Then $fa_x = fdx = x$, $fa_y = fdy = y$, $ga_x = gdx = ehx = ehy = gdy = ga_y$, so $x\ R\ y$. This proves the first statement. The second statement is immediate from the Beck Theorem 3.28 because of the result (12) which asserts that tractable equational presentations induce free algebras. □

We note that the converse of the theorem just proved is also true. We will show this in Theorem 3.44.

EXAMPLE 3.30. Let $\mathcal{A}$ be the category of real separated topological linear spaces with continuous linear maps, let $\mathcal{K}$ be the category of Hausdorff topological spaces and continuous maps and let $U : \mathcal{A} \to \mathcal{K}$ be the usual underlying functor. Then U is monadic.

To see that U has frees, use the special adjoint functor theorem – the Hahn–Banach theorem asserts that $\mathbb{R}$ is a cogenerator for $\mathcal{A}$. It remains to show that U creates coequalizers of U-contractible pairs, so let $f, g : A \to B$ in $\mathcal{A}$ and let $UB \xrightarrow{q} Q = coeq(Uf, Ug)$ be a contractible coequalizer in $\mathcal{K}$. Without topology, real vector spaces is equationally definable (describe $(\lambda, x) \mapsto \lambda x$ as an $\mathbb{R}$-indexed family of unary operations). By Proposition 3.29 and the Beck theorem, there exists a unique vector space structure on Q making q linear which is, moreover, the coequalizer of f, g at the level of vector spaces. As such, the kernel $q^{-1}(0)$ is a vector subspace of B. Additionally, any split epic of topological spaces is a quotient map, so the original continuous map q is a quotient map and the topology on Q is Hausdorff by the definition of $\mathcal{K}$. Thus the kernel of q is a closed linear subspace. The resulting separated topological linear quotient structure on Q must coincide with the vector space and quotient structures already introduced since the vector space structure is unique and since there is only one quotient topology. One detail remains: that $q = coeq(f, g)$ in $\mathcal{A}$. But we already know this at the level of vector spaces. Given $t : B \to C$ in $\mathcal{A}$ with $tf = tg$, the unique linear $\psi : Q \to C$ with $\psi q = t$ must be continuous because q is a quotient map and t is continuous.

3.5. *Monadic comparison*

In this section, our last at the level of general categories, we study functors over $\mathcal{K}$ with emphasis on monadicity.

DEFINITION 3.31. Let $\mathcal{D}$ be a small category. A functor $U : \mathcal{A} \to \mathcal{K}$ *creates limits of type* $\mathcal{D}$ if for every functor $\Delta : \mathcal{D} \to \mathcal{A}$ and every limit $\psi_i : K \to U\Delta i$ of $U\Delta$ there exists unique $\overline{\psi_i} : \overline{K} \to \Delta_i$ in $\mathcal{A}$ with $U\overline{\psi_i} = \psi_i$ for all i; and, moreover, $\overline{\psi} : \overline{K} \to \Delta_i$ is a limit of Δ. U *creates limits* if U creates limits of type $\mathcal{D}$ for every small category $\mathcal{D}$.

It is obvious that if $\mathcal{K}$ is complete and U creates limits then $\mathcal{A}$ is complete and U preserves limits. In this case every limit in $\mathcal{A}$ is created.

A couple of other observations are in order. First of all, if U creates limits then, in particular, U must be faithful. For if $f, g : A \to B$ with $Uf = Ug$, id_{UA} is the equalizer of Uf, Ug so there must exist unique $i : A \to A$ with $Ui = id_{UA}$. Obviously $i = id_A$ so as id_A is, moreover, the equalizer of f, g, $f = g$. It is then immediate from this that the limit projections of a created limit form an initial family in $\mathcal{A}$.

PROPOSITION 3.32. *Every monadic functor creates limits.*

PROOF. We assume without loss that $U = U^{\mathbf{T}}$ for some monad $\mathbf{T}$ in $\mathcal{K}$. In the notation of the previous definition, let $\Delta i = (X_i, \xi_i)$ in $K^{\mathbf{T}}$. Consider the following diagram, where

$\tau : i \to j$ in $\mathcal{D}$.

$$\begin{tikzcd}
TK \arrow[r, "T\psi_i"] \arrow[d, "\xi"'] \arrow[dr, phantom, "\text{(A)}"] & TX_i \arrow[r, "TU\Delta\tau"] \arrow[d, "\xi_i"] \arrow[dr, phantom, "\text{(B)}"] & TX_j \arrow[d, "\xi_j"] \\
K \arrow[r, "\psi_i"'] & X_i \arrow[r, "U\Delta\tau"'] & X_j \quad = \psi_j
\end{tikzcd}$$

Square (B) commutes because $U\Delta\tau$ is $\mathbf{T}$-homomorphism. As

$$\begin{aligned}
U(\Delta\tau)\xi_i(T\psi_i) &= \xi_j(TU\Delta\tau)(T\psi_i) = \xi_j T(U\Delta\tau\psi_i) \\
&= \xi_j(T\psi_j) \quad ((K, \psi) \text{ is a cone})
\end{aligned}$$

There exists a unique ξ making (A) commute by the limit property. The algebra axioms on (K, ξ) and the fact that $\psi_i : (K, \xi) \to \Delta i$ is a limit in $K^{\mathbf{T}}$ all hinge on proving, in each case, that two K-valued morphisms are equal. This is a routine verification since the ψ_i are jointly monic. □

The above theorem is well known in algebra. For example, if X_i is a family of groups, there is exactly one group structure on the Cartesian product set making the projections $X \to X_i$ group homomorphisms and, moreover, this is the categorical product in the category of groups.

The preceding proposition does *not* hold for weakly monadic functors. For example, the category of all groups whose underlying set is a cardinal is weakly monadic over $\mathcal{S}$, but if $\psi_i : K \to UX_i$ is a limit in $\mathcal{S}$ with K not a cardinal, the ψ_i cannot be lifted. On the other hand, this category is monadic over the full subcategory of $\mathcal{S}$ of all cardinals.

DEFINITION 3.33. $U : \mathcal{A} \to \mathcal{K}$ is a *Beck functor* if U creates limits, U preserves limits and U creates coequalizers of U-contractible pairs.

PROPOSITION 3.34. *Every monadic functor is Beck. Every $U_{\Sigma,E} : (\Sigma, E)$-$\mathbf{Alg}$ is Beck.*

PROOF. Since every functor with a left adjoint preserves limits, we know that every monadic functor is Beck from Theorem 3.28. That $U_{\Sigma,E}$ creates limits (it suffices to check products and equalizers since $\mathcal{S}$ is complete) is standard algebra and then $U_{\Sigma,E}$ preserves limits again because $\mathcal{S}$ is complete. By Proposition 3.29 we are done. □

The proof of the following is safely left to the reader.

PROPOSITION 3.35. *Given the commutative diagram of functors*

$$\begin{tikzcd} \mathcal{A} \arrow[rr, "\Gamma"] \arrow[dr, "U"'] & & \mathcal{B} \arrow[dl, "V"] \\ & \mathcal{K} & \end{tikzcd}$$

The following statements hold.

(1) *If U creates coequalizers of U-contractible pairs and if V creates coequalizers of V-contractible pairs then Γ creates coequalizers of Γ-contractible pairs.*

(2) *If U, V are Beck, so too is Γ.*

We are finally ready to establish the universal property of the semantics comparison functor.

THEOREM 3.36. *Let $U : \mathcal{A} \to \mathcal{K}$ be tractable with semantic comparison functor $\Phi : \mathcal{A} \to K^{\mathbf{T}}$ as in Definition* 3.20. *Let $V : \mathcal{B} \to \mathcal{K}$ be any* (*not necessarily tractable*) *Beck functor, and let $\Psi : \mathcal{A} \to \mathcal{B}$ be any functor over $\mathcal{K}$. Then there exists a unique functor Γ over $\mathcal{K}$ making the following triangle commute*:

$$\begin{tikzcd} \mathcal{A} \arrow[rr, "\Phi"] \arrow[dr, "\Psi"'] & & K^{\mathbf{T}} \arrow[dl, "\Gamma"] \\ & \mathcal{B} & \end{tikzcd}$$

PROOF. Consider the functor $\Xi_K : (K, U) \to K^{\mathbf{T}}$ mapping $f : (A, a) \to (B, b)$ to $f : \Phi A \to \Phi B$. As $U^{\mathbf{T}} \Xi_K = \Delta_K : (K, U) \to \mathcal{K}$ with limit $\psi^K_{A,a} : TK \to UA$ and as $\psi^K_{A,a} = U^{\mathbf{T}} a^\#$ by Proposition 3.21(3),

$$(TK, \mu_K) \xrightarrow{a^\#} \Phi A \quad (\text{for } K \xrightarrow{a} UA)$$

is a (created) limit in $K^{\mathbf{T}}$. As Γ must preserve limits, the only possible definition of $\Gamma(TK, \mu_K)$ is as the unique created limit

$$\Gamma(TK, \mu_K) \xrightarrow{a^\#} \Phi A \quad (\text{for } K \xrightarrow{a} UA)$$

in $\mathcal{B}$, with underlying $\mathcal{K}$-object TK. The typical $\mathbf{T}$-homomorphism $(TK, \mu_K) \to (TL, \mu_L)$ is $\alpha^\#$ for $\alpha : K \to TL$. For $b : L \to UA$, the $\mathbf{T}$-homomorphisms $b^\# \alpha^\#, (b^\# \alpha)^\# : (TK, \mu_K) \to \Phi A$ are equal, because they agree on generators. By the initiality of the family of limit projections of a created limit, $\alpha^\# : \Gamma(TK, \mu_K) \to \Gamma(TL, \mu_L)$ is a $\mathcal{B}$-morphism. So far we have shown that if $\mathcal{F}$ is the full subcategory of all (TK, μ_K)

in $K^{\mathbf{T}}$, that $\Gamma : \mathcal{F} \to \mathcal{B}$ is a well-defined functor over $\mathcal{K}$ and is the only such functor on $\mathcal{F}$ that could restrict a functor extending Ψ.

Now let (K, ξ) be a $\mathbf{T}$-algebra so that $\xi = coeq(\mu_K, T\xi)$ is a contractible coequalizer with contraction (η_{TK}, η_K) in $\mathcal{K}$. There is thus a unique $\mathcal{B}$-structure $\Gamma(K, \xi)$ admitting $\xi : \Gamma(TK, \mu_K) \to \Gamma(K, \xi)$, and so this is the only possible definition of $\Gamma(K, \xi)$.

At this point there is a potential worry that $\Gamma(TK, \mu_K)$ is now defined in two different ways. To be careful, let's temporarily write $\Gamma_0 : \mathcal{F} \to \mathcal{B}$ for the first functor so that the admissibility of $\xi : \Gamma_0(TK, \mu_K) \to \Gamma(K, \xi)$ defines $\Gamma(K, \xi)$. As all $\mathbf{T}$-homomorphisms $(TTK, \mu_{TK}) \to (TK, \mu_K)$ are admissible $\Gamma_0(TTK, \mu_{TK}) \to \Gamma_0(TK, \mu_K)$ including μ_K, we see that $\Gamma_0(TK, \mu_K) = \Gamma(TK, \mu_K)$ after all.

As $\xi : \Gamma(TK, \mu_K) \to \Gamma(K, \xi)$ is a created coequalizer in $\mathcal{B}$, it is a final map by Proposition 3.27. If $f : K \to L$ in $\mathcal{K}$ and (L, θ) is a $\mathbf{T}$-algebra, $Tf : \Gamma(K, \xi) \to \Gamma(TL, \mu_L)$ and $\theta : \Gamma(TL, \mu_L) \to \Gamma(L, \theta)$ are $\mathcal{B}$-admissible. If f is a $\mathbf{T}$-homomorphism $(K, \xi) \to (L, \theta)$, $f\xi : \Gamma(TK, \mu_K) \to \Gamma(L, \theta)$ is $\mathcal{B}$-admissible because $f\xi = \theta(Tf)$. It follows from the finality of ξ that $f : \Gamma(K, \xi) \to \Gamma(L, \theta)$ is $\mathcal{B}$-admissible. Thus Γ is a well-defined functor over $\mathcal{K}$.

On the one hand, $\psi^K_{A, id_{UA}} : \Gamma(TUA, \mu_{UA}) \to \Psi A$ is $\mathcal{B}$-admissible, being a limit projection. On the other hand, this same map is the structure map of ΦA by the definition of Φ. Thus $\Psi A = \Gamma \Phi A$. The proof is complete. □

COROLLARY 3.37. *Let* $\mathbf{T}$ *be a monad in* $\mathcal{K}$ *and let* $\mathcal{F}$ *be a full subcategory of* $\mathcal{K}^{\mathbf{T}}$ *which contains* (TK, μ_K) *for every* $\mathcal{K}$*-object* K. *Let* $V : \mathcal{B} \to \mathcal{K}$ *be a Beck functor. Then each functor* $\Gamma_0 : \mathcal{F} \to \mathcal{B}$ *over* $\mathcal{K}$ *extends uniquely over* $\mathcal{K}$ *to* $\mathcal{K}^{\mathbf{T}}$.

PROOF. The inclusion $\mathcal{F} \to K^{\mathbf{T}}$ is the semantic comparison functor of $\mathcal{F} \to K^{\mathbf{T}} \xrightarrow{U^{\mathbf{T}}} \mathcal{K}$ by Proposition 3.16. □

EXAMPLE 3.38. Let (Σ, E) be any (not necessarily tractable) equational presentation. Let A_1 denote the unique 1-element (Σ, E)-algebra and assume that there exist (Σ, E)-algebras A_2, A_4 with, respectively, 2 and 4 elements such that any of the 301 functions of form $A_i \to A_j$ are Σ-homormorphisms. Then any compact Hausdorff space has canonical (Σ, E)-algebra structure.

For let $\mathcal{G}$ be the full subcategory $\{A_1, A_2, A_4\}$ of (Σ, E)-**Alg**. By Example 3.19, the inclusion $\Phi : \mathcal{G} \to \mathcal{S}^{\boldsymbol{\beta}} =$ compact T2 spaces is the semantics comparison functor, so that the inclusion $\Psi : \mathcal{G} \to (\Sigma, E)$-**Alg** extends uniquely to the desired functor $\Gamma : \mathcal{S}^{\boldsymbol{\beta}} \to (\Sigma, E)$-**Alg** by Theorem 3.36. In this case there are (Σ, E)-algebras of every cardinality (if X is an infinite set, remove a point, assign the discrete topology, and replace the point with the topology of the one-point compactification).

The universal property of semantic comparison constructs forgetful functors between categories of algebras. We now seek to learn more about such functors. We'll start with a familiar example from algebra. Ring homomorphisms $\varphi : Q \to R$ correspond to functors $\Gamma : R$-**Mod** $\to Q$-**Mod** where Γ assigns the same Abelian group with Q-action $qx = \varphi(q)x$. Here, of course, R-**Mod** is the equationally definable class of modules over the ring R with rx an R-indexed family of unary operations. Thus R-**Mod** is monadic with monad $\mathbf{T}_R$. Ring homomorphisms are also coextensive with monad maps $\lambda : \mathbf{T}_Q \to \mathbf{T}_R$. Here, $T_R X$ is the underlying set of the free module $\bigoplus_X R$ generated by X, where $\lambda_X(r_x) = (\varphi r_x)$.

The idea that forgetful functors between general algebra categories correspond to maps of their theories in the opposite direction dates to Lawvere's thesis [55]. The statement for monads is as follows.

THEOREM 3.39. *Let* $\mathbf{S}, \mathbf{T}$ *be monads in* $\mathcal{K}$, *then there exists a bijective functorial correspondence between monad maps* $\lambda : \mathbf{S} \to \mathbf{T}$ *and functors* $\Lambda : K^{\mathbf{T}} \to K^{\mathbf{S}}$ *over* $\mathcal{K}$. *The mutually inverse passages are*

$$\Lambda(K, \xi) = (K, SK \xrightarrow{\lambda_K} TK \xrightarrow{\xi} K)$$

and

$$\lambda_K = SK \xrightarrow{S\eta_K} STK \xrightarrow{\nu_K} TK \quad (TK, \nu_K) = \Lambda(TK, \mu_K)$$

PROOF. Let's first check $(K, \xi\lambda_K)$ is an $\mathbf{S}$-algebra when (K, ξ) is a $\mathbf{T}$-algebra and λ is a monad map. We have $\xi\lambda_K\eta_K = \xi\eta_K = id_K$, which is the η-law. Now let $\alpha, \beta : W \to SK$. Writing $\alpha^{\#\#}$ for $\mathbf{S}$-extension and $\gamma^{\#}$ for $\mathbf{T}$-extension, we have the monad map law

$$\begin{tikzcd} SW \arrow[r, "\lambda_W"] \arrow[d, "\alpha^{\#\#}"'] & TW \arrow[d, "(\lambda_K\alpha)^{\#}"] \\ SK \arrow[r, "\lambda_K"'] & TK \end{tikzcd}$$

and a similar square for β. Thus if $(\xi\lambda_K)\alpha = (\xi\lambda_K)\beta$, also $\xi(\lambda_K\alpha) = \xi(\lambda_K\beta)$ so

$$(\xi\lambda_K)\alpha^{\#\#} = \xi(\lambda_K\alpha)^{\#}\lambda_W = \xi(\beta_K\alpha)^{\#}\lambda_W = (\xi\lambda_K)\beta^{\#\#}$$

That a $\mathbf{T}$-homomorphism $f : (K, \xi) \to (L, \theta)$ is an $\mathbf{S}$-homomorphism $f : (K, \lambda_K\xi) \to (L, \lambda_L\theta)$ is obvious because λ is natural. So far: $\Lambda : K^{\mathbf{T}} \to K^{\mathbf{S}}$ is a well-defined functor over $\mathcal{K}$.

Next, given Λ over $\mathcal{K}$, we check that λ is a monad map.

$$\begin{aligned} \lambda_K\eta_K &= \nu_K\big(S\eta_K^T\big)\eta_K^S \\ &= \nu_K\eta_{TK}^S\eta_K^T \quad (\eta^T \text{ is natural}) \\ &= \eta_K \quad \big((TK, \nu_K) \text{ is an } \mathbf{S}\text{-algebra}\big) \end{aligned}$$

To establish the square above for $\alpha : W \to SK$,

$$(\lambda_K\alpha)^{\#}\lambda_W = \nu_K S(\lambda_K\alpha)^{\#}\big(S\eta_W^T\big) = \nu_K S\big((\lambda_K\alpha)^{\#}\eta_W^T\big) = \nu_K S(\lambda_K\alpha)$$

which is the formula for $(\lambda_K \alpha)^{\#\#}$. But this is the same as $\lambda_K \alpha^{\#\#}$ because λ_K is an **S**-homomorphism. Thus $\Lambda \mapsto \lambda$ is well-defined.

We must show that the two passages are mutually inverse. If $\lambda \mapsto \Lambda \mapsto \hat{\lambda}$, then $\hat{\lambda}_K = SK \xrightarrow{S\eta_K} STK \xrightarrow{\nu_K} TK$ where $\nu_K = STK \xrightarrow{\lambda_{TK}} TTK \xrightarrow{T\mu} TK$. Thus

$$\begin{aligned}\hat{\lambda}_K &= \mu_k \lambda_{TK}(S\eta_K)\\ &= \mu_K(T\eta_K)\lambda_K \quad (\lambda \text{ n.t.})\\ &= \lambda_K \quad (\text{monad law})\end{aligned}$$

Now let $\Lambda \mapsto \lambda \mapsto \widehat{\Lambda}$. Then $\widehat{\Lambda}(K,\xi) = (K, \xi\lambda_K)$ so $\xi\lambda_K : (SK, \mu_K^S) \to \widehat{\Lambda}(K,\xi)$ is an **S**-homomorphism and, being a structure map, $\widehat{\Lambda}(K,\xi)$ is the unique **S**-algebra admitting $\xi\lambda_K$. On the other hand, $\xi\lambda_K$ is an **S**-homomorphism $(SK, \mu_K) \to \Lambda(K,\xi)$ so $\Lambda = \widehat{\Lambda}$. Finally, the passage is functorial because $\nu_K(\lambda_K \xi) = (\nu_k \lambda_k)\xi$. $\square$

By Proposition 3.35 one might suspect that forgetful functors between categories of algebras are often monadic themselves. We'll show in Corollary 3.50 that this is always true over $\mathcal{S}$. For example, the forgetful functors from rings to monoids and Abelian groups are both monadic. In general, all that lacks is a left adjoint. Often the adjoint functor theorems provide it since such theorems assert that under reasonable conditions a functor which preserves limits will have a left adjoint. But there is another technique due to Linton which constructs free objects as a coequalizer as we now describe.

THEOREM 3.40. *Let $U : \mathcal{A} \to \mathcal{K}$ have a left adjoint with adjointness* $(\mathcal{A}, \mathcal{K}, U, F, \eta, \varepsilon)$. *Assume that each pair $t, a : A \to B$ of $\mathcal{A}$-morphisms has a coequalizer. Let* **T** *be a monad in $\mathcal{K}$ and let $\Gamma : \mathcal{A} \to K^{\mathbf{T}}$ be any functor over $\mathcal{K}$. Then Γ has a left adjoint.*

PROOF. We give only an outline of the construction. For full details, see [76, pp. 182–183]. Let (K, ξ) be a **T**-algebra. Write $\Gamma FK = (UFK, \theta_K)$. Define $\mathcal{A}$-morphisms $a, b : FTK \to FK$ by

$$\begin{aligned}a &= FTK \xrightarrow{FT\eta_K} FTUFK \xrightarrow{F\theta_K} FUFK \xrightarrow{\varepsilon_{FK}} FK\\ b &= FTK \xrightarrow{F\xi} FK\end{aligned}$$

Let $p : FK \to A$ be *coeq*(a, b) in $\mathcal{A}$. Define $\delta = K \xrightarrow{\eta_K} UFK \xrightarrow{Up} UA$ in $\mathcal{K}$. Then $\delta : (K, \xi) \to \Gamma A$ is a **T**-homomorphism and (A, δ) is the free $\mathcal{A}$-object generated by (K, ξ). $\square$

We turn next to a version of the Birkhoff variety theorem valid in $K^{\mathbf{T}}$. To this end we must be sure to clarify what "subalgebras" and "quotient algebras" of a **T**-algebra (X, ξ) would be.

If $i : A \to X$ is monic, there is at most one ξ_0 such that the following square commutes.

$$\begin{tikzcd} TA \arrow[r, "Ti"] \arrow[d, "\xi_0"'] & TX \arrow[d, "\xi"] \\ A \arrow[r, "i"'] & X \end{tikzcd} \tag{20}$$

In that case (A, ξ_0) is a $\mathbf{T}$-algebra. To check this, $i\xi_0\eta_A = \xi(Ti)\eta_A = \xi\eta_X i = id_X i = i\,id_A$ and i is monic so $\xi_0\eta_A = id_A$. For the second axiom, let $\alpha, \beta : W \to TA$ with $\xi_0\alpha = \xi_0\beta$. Then $i\xi_0\alpha^\# = \xi(Ti)\alpha^\# = \xi((Ti)\alpha)^\# = ((Ti)\beta)^\# = i\xi_0\beta^\#$ so, again as i is monic, $\xi_0\alpha^\# = \xi_0\beta^\#$.

We call either (A, ξ_0) or $i : A \to X$ a *subalgebra of* (X, ξ). If $\mathcal{M}$ is a class of monics, an $\mathcal{M}$-*subalgebra* is a subalgebra $i : A \to X$ with $i \in \mathcal{M}$.

Surprisingly, Birkhoff's theorem in the context of an image factorization system $(\mathcal{E}, \mathcal{M})$ is with respect to $\mathcal{M}$-subalgebras but *not* $\mathcal{E}$-quotient algebras. Rather, a Birkhoff subcategory $\mathcal{B}$ of $K^{\mathbf{T}}$ is *closed under* $U^{\mathbf{T}}$*-split epics* which means that if $q : (K, \xi) \to (L, \theta)$ is a $\mathbf{T}$-homomorphism with (K, ξ) in $\mathcal{B}$ and $q : K \to L$ a split epic in $\mathcal{K}$, then (L, θ) is also in $\mathcal{B}$. Where $\mathcal{E}$ comes in is in that the monad map corresponding to the inclusion $\mathcal{B} \to K^{\mathbf{T}}$ is pointwise in $\mathcal{E}$.

We turn now to the formal statement of the theorem.

THEOREM 3.41. *Let $\mathcal{K}$ a locally small category with small limits. Let $(\mathcal{E}, \mathcal{M})$ be an image factorization system for $\mathcal{K}$ such that $\mathcal{K}$ is $\mathcal{E}$-cowellpowered. Let $\mathcal{A}$ be a full subcategory of $K^{\mathbf{S}}$ where $\mathbf{S}$ is a monad in $\mathcal{K}$ such that S preserves $\mathcal{E}$. Then the following statements are equivalent.*

(1) *$\mathcal{A}$ is closed under products, $\mathcal{M}$-subobjects and $U^{\mathbf{S}}$-split epics.*

(2) *If $U : \mathcal{A} \to \mathcal{K}$ is the restriction of $U^{\mathbf{S}}$ to $\mathcal{A}$, then U is monadic and the monad map λ corresponding to the inclusion $\mathcal{A} \to U^{\mathbf{S}}$ is pointwise in $\mathcal{E}$.*

When these conditions hold, the functor part of the monad corresponding to U also preserves $\mathcal{E}$. Further, $\mathcal{A}$ is a full reflective subcategory of $K^{\mathbf{S}}$ whose reflection maps are all in $\mathcal{E}$ and λ_K is the reflection of (SK, μ_K).

We say that such $\mathcal{A}$ is a *Birkhoff subcategory of* $K^{\mathbf{S}}$ (with respect to $(\mathcal{E}, \mathcal{M})$).

PROOF. $(1 \Rightarrow 2)$. If $f : (K, \xi) \to (L, \theta)$ is an $\mathbf{S}$-homomorphism and if f has $(\mathcal{E}, \mathcal{M})$-factorization $K \xrightarrow{e} I \xrightarrow{m} L$ in $\mathcal{K}$, consider the diagram

$$\begin{tikzcd} SK \arrow[r, "Se"] \arrow[d, "\xi"'] \arrow[dr, phantom, "(A)"] & SI \arrow[r, "Sm"] \arrow[d, "\theta_0"] \arrow[dr, phantom, "(B)"] & SL \arrow[d, "\theta"] \\ K \arrow[r, "e"'] & I \arrow[r, "m"'] & L \end{tikzcd} \quad = f$$

Here the perimeter commutes because f is an **S**-homomorphism. Since S preserves $\mathcal{E}$, there exists unique θ_0 making (A) and (B) each commute, so that (I, θ_0) is an $\mathcal{M}$-subalgebra of (L, θ) with e, m now both homomorphisms. What this shows is that $(\mathcal{E}, \mathcal{M})$ provides an image factorization system for $K^{\mathbf{S}}$ whose epics are those homomorphisms, $U^{\mathbf{S}}$ of which are in $\mathcal{E}$, and $\mathcal{M}$ similarly.

It is then immediate that $\mathcal{A}$ is a quasivariety in $K^{\mathbf{S}}$ so that by Proposition 3.8 each **S**-algebra X has a reflection $r : X \to X_{\mathcal{A}}$ in $\mathcal{A}$. If r has image factorization me the homomorphism $e : X \to I$ also has the universal property of the reflection (because morphisms in $\mathcal{E}$ are epic and because $\mathcal{A}$ is closed under $\mathcal{M}$-subalgebras) so m is an isomorphism and $r \in \mathcal{E}$.

U has a left adjoint F with FK the reflection of (SK, μ_K). If $a, b : A \to B$ in $\mathcal{A}$ is U-contractible with coequalizer $q : UB \to Q$ in $\mathcal{K}$, there exists unique **S**-homomorphic lift $q : B \to \overline{Q}$ and then $q = coeq(a, b)$ in $K^{\mathbf{S}}$. As $\mathcal{A}$ is closed under U-split epics, $\overline{A} \in \mathcal{A}$ so, a fortiori, $q = coeq(a, b)$ in $\mathcal{A}$. By Theorem 3.28, U is monadic.

The monad $\mathbf{T}$ for U is induced by the known adjointness $(\mathcal{A}, \mathcal{K}, U, F, \eta, \varepsilon)$. Thus the reflection $\lambda_K : (SK, \mu_K) \to (TK, \mu_K)$ defines T and $\mu : TT \to T$. Also, $id \xrightarrow{\eta} S \xrightarrow{\lambda} T$ defines η. What we need to do is show that the monad map corresponding to the inclusion of $\mathcal{A}$ in $K^{\mathbf{S}}$ is the reflection maps λ since we already know each λ_K is an epic in $\mathcal{E}$. Hence we must prove $\lambda_K = \nu_K(S\eta^T_K)$. To that end, let the **S**-structure map of (TK, μ_K) be $\nu_K : STK \to TK$. We have

$$\big(\nu_K\big(S\eta^T_K\big)\big)\eta^S_K = \nu_K \eta^S_{TK} \eta^T_K \quad (\eta^S \text{ natural}) = id_{TK}\eta^T_K = \eta^T_K$$

As λ_K, $\nu_K(S\eta^T_K)$ are two **S**-homomorphisms agreeing on generators, they are equal, as we hoped to prove. To see that T preserves $\mathcal{E}$, let $e : K \to L$ in $\mathcal{E}$ and consider the naturality square

$$\begin{tikzcd} SK \arrow[r, "\lambda_K"] \arrow[d, "Se"'] & TK \arrow[d, "Te"] \\ SL \arrow[r, "\lambda_L"'] & TL \end{tikzcd}$$

Any image factorization system satisfies $gf \in \mathcal{E} \Rightarrow g \in \mathcal{E}$ so $Te \in \mathcal{E}$ because Se, λ_K, λ_L are all in $\mathcal{E}$.

$(2 \Rightarrow 1)$. The functor $\Lambda : K^{\mathbf{T}} \to K^{\mathbf{S}}$ is a full subcategory $\mathcal{A}$ of $K^{\mathbf{S}}$ as follows. If $(K, \xi\lambda_K) = (K, \theta\lambda_K)$, $\xi = \theta$ because λ_K is epic, so Λ is injective on objects. Similarly, in the diagram

$$\begin{tikzcd} SK \arrow[r, "\lambda_K"] \arrow[d, "Sf"'] \arrow[dr, phantom, "\text{(A)}"] & TK \arrow[r, "\xi"] \arrow[d, "Tf"] \arrow[dr, phantom, "\text{(B)}"] & K \arrow[d, "f"] \\ SL \arrow[r, "\lambda_L"'] & TL \arrow[r, "\theta"'] & L \end{tikzcd}$$

(A) always commutes because λ is natural, so that if $f:(K,\xi\lambda_K)\to(L,\theta\lambda_L)$ is an **S**-homomorphism (i.e., (A, B) commutes) then (B) commutes preceded by the epic λ_K and so in fact commutes. This shows that Λ is a full functor.

By Proposition 3.35, the inclusion $\mathcal{A}=K^{\mathbf{T}}\to K^{\mathbf{S}}$ creates and preserves limits, so it is clear that $\mathcal{A}$ is closed under products. Let $m:B\to A$ in $\mathcal{M}$ with A in $\mathcal{A}$. If $r:B\to B_{\mathcal{A}}$ is the reflection of B in $\mathcal{A}$ there exists a commutative triangle

$$\begin{tikzcd} B \arrow[r, "e"] \arrow[dr, "m"'] & B_{\mathcal{A}} \arrow[d] \\ & A \end{tikzcd}$$

so that, by standard properties of image factorization systems, $r\in\mathcal{E}\cap\mathcal{M}$ and hence is an isomorphism. As isomorphisms are unary products, it then follows that B is in $\mathcal{A}$.

Finally, let $q:(K,\xi)\to(L,\theta)$ in $K^{\mathbf{S}}$ with (K,ξ) in $\mathcal{A}$, and with $d:L\to K$ in $\mathcal{K}$ such that $qd=id_L$. We must show (L,θ) is in $\mathcal{A}$ as well. There exists a **T**-algebra $(K,\hat{\xi})$ with $\hat{\xi}\lambda_K=\xi$. Consider the diagram

$$\begin{tikzcd} SK \arrow[r, "\lambda_K"] \arrow[d, "Sq"'] \arrow[dr, phantom, "\text{(A)}"] & TK \arrow[r, "\hat{\xi}"] \arrow[d, "Tq"] \arrow[dr, phantom, "\text{(B)}"] & K \arrow[d, "q"] & =\xi \\ SL \arrow[r, "\lambda_L"'] & TL \arrow[r, "\hat{\theta}"'] & L & =\theta \end{tikzcd}$$

where $\hat{\theta}$ is yet to be constructed. As λ is natural, (A) commutes. It suffices to find $\hat{\theta}$ in $\mathcal{K}$ with $\hat{\theta}\lambda_L=\theta$ since then (A, B) commutes and λ_K epic $\Rightarrow$ (B) commutes $\Rightarrow$ (by Lemma 3.27) that $(L,\hat{\theta})$ is a **T**-algebra and so $(L,\theta)=(L,\hat{\theta}\lambda_L)$ is in $\mathcal{A}$. To that end define

$$\hat{\theta}=TL\xrightarrow{Td}TK\xrightarrow{\hat{\xi}}K\xrightarrow{q}L$$

As every **T**-homomorphism is an **S**-homomorphism, $\hat{\theta}\lambda_L:(SL,\mu_L)\to(L,\theta)$ is an **S**-homomorphism. We have

$$\begin{aligned}\hat{\theta}\lambda_L\eta_L &= q\hat{\xi}(Td)\lambda_L\eta_L = q\hat{\xi}\lambda_K(Sd)\eta_L \quad (\lambda \text{ natural})\\ &= q\xi(Sd)\eta_L=\theta(Sq)(Sd)\eta_L=\theta\eta_L \quad (\text{as } qd=id_L)\\ &= id_L\end{aligned}$$

But $\theta:(SL,\mu_L)\to(L,\theta)$ is the unique **S**-homomorphic extension of id_L, so $\hat{\theta}\lambda_L=\theta$ as desired. □

We conclude this subsection with the example of topological semigroups. For a much more general treatment at the level of topological categories (there called "fibre-complete categories") see [76, Section 3.5].

EXAMPLE 3.42. The category of topological semigroups and continuous semigroup homomorphisms is monadic over topological spaces.

Let $\mathcal{A}$ be the category of all $(X, \star, \mathcal{T})$ with $(X, \star)$ a semigroup and $(X, \mathcal{T})$ a topological space (but no relation between the two structures) and with morphisms all continuous semigroup homomorphisms. Let $U : \mathcal{A} \to \mathbf{Top}$ be the functor which forgets the semigroup structure. Semigroups is monadic over $\mathcal{S}$, being equationally definable. U has a left adjoint $(X, \mathcal{T}) \mapsto (FX, \widehat{\mathcal{T}})$ where FX is the free semigroup generated by X and $\widehat{\mathcal{T}}$ is the smallest topology $\mathcal{W}$ such that $\alpha^{\#} : (FX, \mathcal{W}) \to (Y, *, \mathcal{S})$ is continuous where α ranges over all continuous $(X, \mathcal{T}) \to (Y, *, \mathcal{W})$ and $\alpha^{\#}$ denotes the semigroup-homomorphic extension. If $a, b : (X, \star, \mathcal{T}) \to (Y, *, \mathcal{S}$ is U-contractible, with contractible coequalizer $q : (Y, \mathcal{T}) \to (Z, \mathcal{W})$, there exists a unique semigroup lift $q : (X, *) \to (Z, \bullet)$ which is, moreover, the semigroup equalizer of a, b. But $q : (Y, *, \mathcal{S}) \to (Z, \bullet, \mathcal{W})$ is the coequalizer in $\mathcal{A}$ since a split epic in **Top** is a quotient map.

So far, $U : \mathcal{A} \to \mathbf{Top}$ is monadic. The true topological semigroups are the full subcategory $\mathcal{B}$ of $\mathcal{A}$ of all $(X, \star, \mathcal{T})$ for which $\star : (X, \mathcal{T})^2 \to (X, \mathcal{T})$ is jointly continuous. It is well known that $\mathcal{B}$ is closed under products and subsemigroups with the subspace topology. This suggests that we attempt to apply Theorem 3.41 with $\mathcal{M} =$ topological subspaces (that is, all initial monics). This forms an image factorization system in **Top** with $\mathcal{E} =$ all continuous surjections. That F preserves $\mathcal{E}$ is clear. We have only to show that $\mathcal{B}$ is closed under U-split quotients. Let $(X, \star, \mathcal{T})$ in $\mathcal{B}$ and $(Q, \bullet, \mathcal{W})$ in $\mathcal{A}$ and consider the commutative square

$$\begin{tikzcd}
X \times X \arrow[r, "\star"] \arrow[d, "q \times q"'] & X \arrow[d, "q"] \\
Q \times Q \arrow[r, "\bullet"'] & Q
\end{tikzcd}$$

where q is split epic in **Top**. Since $q \times q$ is also split epic it must be a quotient map, hence final in **Top**. Thus $\bullet$ is continuous because $\bullet(q \times q) = q \star$ is, and $(Q, \bullet, \mathcal{W})$ is in $\mathcal{B}$. Notice that $\mathcal{B}$ is not closed under $\mathcal{E}$. If $(X, \star, \mathcal{T})$ is not in $\mathcal{B}$ and if $\mathcal{T}_d$ is the discrete topology on X, $(X, \star, \mathcal{T}_d)$ is in $\mathcal{B}$ and the identity function from $(X, \star, \mathcal{T}_d)$ to $(X, \star, \mathcal{T})$ in $\mathcal{E}$.

3.6. *Monadic functors of sets*

Our goals in this subsection are to develop further abstract theory, to provide a number of examples and to explore some duality questions.

We begin by clarifying the relationship between **T**-algebras of sets and equationally definable classes.

DEFINITION 3.43. A monad **T** in $\mathcal{S}$ is *bounded by the cardinal* α if $\forall x\ \forall \omega \in TX\ \exists \beta < \alpha\ \exists f : \beta \to X$ with ω in the image of Tf. If such α exists, **T** is *bounded*. Otherwise, **T** is *unbounded*. **T** is *finitary* if **T** is bounded by $\aleph_0$.

THEOREM 3.44. *A functor $U : \mathcal{A} \to \mathcal{S}$ is monadic if and only if there exists a tractable equational presentation (Σ, E) with $\mathcal{A}$ isomorphic over $\mathcal{S}$ to (Σ, E)-**Alg**.*

PROOF. That tractable equational classes are monadic was seen in Proposition 3.29. For the converse, we may assume that $U = U^{\mathbf{T}}$ for $\mathbf{T}$ a monad in $\mathcal{S}$ and construct an isomorphism over $\mathcal{S}$, $\mathcal{S}^{\mathbf{T}} \cong (\Sigma, E)$-**Alg** for some (Σ, E) (which must be tractable because $U^{\mathbf{T}}$ has a left adjoint). We know that the elements of a free (Σ, E)-algebra generated by n can be thought of as all the derived n-ary operations. It is harmless to include all such operations as primitive operations if the appropriate equations are imposed (i.e., if f, g are "primitive" operations and $h(x, y) = f(x, g(y, x))$ is a derived operation then h can be introduced as a primitive operation so long as the equation defining h is added to E). This suggests the following construction. If $\mathbf{T}$ is unbounded, define $\Sigma_n = Tn$ for all cardinals n. If $\mathbf{T}$ is bounded by α use this definition for $n < \alpha$ but set $\Sigma_n = \emptyset$ otherwise. In either case, $\forall \nu \in TX\ \exists n\ \exists \omega \in \Sigma_n\ \exists f : n \to X\ (Tf)\omega = \nu$ since, if $\mathbf{T}$ is unbounded, we can choose n as the cardinality of X.

It was earlier explained that Tn may be thought of as in $n.t.(U^n, U)$. For the sake of clarity, we'll review the constructions in more detail. First of all, observe that each $\omega \in \Sigma_n$ may be regarded as a natural transformation $id^n \to T$ via $\omega_X : X^n \to TX$, $n \xrightarrow{a} X \mapsto (Ta)\omega$. Verification that the naturality square (A) in the diagram below commutes for any function f is trivial.

$$\begin{tikzcd}
X^n \arrow[r, "\omega_X"] \arrow[d, "f^n"'] \arrow[dr, phantom, "\text{(A)}"] & TX \arrow[r, "\xi"] \arrow[d, "Tf"] \arrow[dr, phantom, "\text{(B)}"] & X \arrow[d, "f"] \\
Y^n \arrow[r, "\omega_Y"'] & TY \arrow[r, "\theta"'] & Y
\end{tikzcd}$$

Thus if f is a $\mathbf{T}$-homomorphism (B) we see that the functor Γ over $\mathcal{S}$ defined by $\Gamma(X, \xi) = (X, \delta)$, $\delta_\omega = \xi\omega_X$ is well-defined.

$$\begin{tikzcd}
\mathcal{S}^{\mathbf{T}} \arrow[r, "\Gamma"] \arrow[dr, "U^{\mathbf{T}}"'] & \Sigma\text{-}\mathbf{Alg} \arrow[d, "U_\Sigma"] \\
& \mathcal{S}
\end{tikzcd}$$

If (X, ξ), (X, θ) are distinct $\mathbf{T}$-algebras, $\exists \nu \in TX$ with $\xi\nu \neq \theta\nu$. By hypothesis, $\exists n\ \exists \omega \in \Sigma_n\ \exists t : n \to X$ with $(Tt)\omega = \nu$, that is, $\omega_X t = \nu$. If $\Gamma(X, \xi) = (X, \delta)$, $\Gamma(X, \theta) = (X, \varepsilon)$, $\delta_\omega t = \xi\nu \neq \theta\nu = \varepsilon_\omega t$. This shows that Γ is injective on objects.

We next argue that Γ is a full functor. Let f be a Σ-homomorphism $\Gamma(X, \xi) \to \Gamma(Y, \theta)$, that is, $\forall n\ \forall \omega \in \Sigma_n$ the perimeter of (A, B) above commutes. We must show that (B) commutes. This is immediate from the fact that each $\nu \in TX$ is in the image of ω_X for some $n, \omega \in \Sigma_n$.

At this point, Γ is a full subcategory $\mathcal{B}$ of Σ-**Alg**. By Theorem 3.9 it remains to show that $\mathcal{B}$ is closed under products, subalgebras and quotient algebras. Even without tractability, it is clear that U_Σ creates small limits, so it is clear as well that $\mathcal{B}$ is closed under limits, products in particular.

Let A be a Σ-subalgebra of $\Gamma(X,\xi)$ with inclusion $i: A \to X$. For $\omega \in \Sigma_n$ with induced natural transformation $\omega : id^n \to T$ consider the diagram

$$\begin{tikzcd} A^n \arrow[r, "\omega_A"] \arrow[d, "i^n"'] \arrow[dr, phantom, "\text{(A)}"] & TA \arrow[r, "\xi_0"] \arrow[d, "Ti"] \arrow[dr, phantom, "\text{(B)}"] & A \arrow[d, "i"] \\ X^n \arrow[r, "\omega_X"'] & TX \arrow[r, "\xi"'] & X \end{tikzcd}$$

As $\xi\omega_X$ is the typical Σ-operation on $\Gamma(X,\xi)$, $\xi\omega_X i^n$ factors through A by hypothesis, and the unique $d: A^n \to A$ with $id = \xi\omega_X i^n$ is then the Σ-interpretation of ω in the subalgebra A. Our goal is to find ξ_0 such that (B) commutes as then (A,ξ_0) is a **T**-algebra (by (20.3)) with $\Gamma(A,\xi_0) = A$ as a Σ-subalgebra of $\Gamma(X,\xi)$. (In more detail: $\xi_0\omega_A = d$ because (A) commutes by naturality). Let $\nu \in TA$. By the definition of Σ, $\exists n\ \exists\omega \in \Sigma_n\ \exists t \in A^n\ \omega_A(t) = \nu$. Then $\xi(Ti)\nu = (\xi(Ti)\omega_A)t = (\xi\omega_X i^n)t \in A$ so $\xi(Ti)$ factors through A as desired.

Finally, we show that $\mathcal{B}$ is closed under quotients. The situation is summed up in the diagram

$$\begin{tikzcd} X^n \arrow[r, "\omega_X"] \arrow[d, "h^n"'] \arrow[dr, phantom, "\text{(A)}"] & TX \arrow[r, "\xi"] \arrow[d, "Th"] \arrow[dr, phantom, "\text{(B)}"] & X \arrow[d, "h"] \\ Q^n \arrow[r, "\omega_Q"'] & TQ \arrow[r, "\theta"'] & Q \end{tikzcd}$$

where (X,ξ) is a **T**-algebra, h is surjective and $\omega \in Tn$. We must find θ such that (B) commutes – (Q,θ) is then a **T**-algebra by Lemma 3.27. By the axiom of choice there exists $d: Q \to X$ with $hd = id_Q$. Define

$$\theta = TQ \xrightarrow{Td} TX \xrightarrow{\xi} X \xrightarrow{h} Q$$

Let $\sigma \in TX$. Then $\exists n\ \exists\omega \in Tn\ \exists t \in X^n \omega_X t = \sigma$. We have

$$(h\xi)\sigma = (h\xi\omega_X)t$$

and

$$\begin{aligned}\big(\theta(Th)\big)\sigma &= \big(h\xi(Td)(Th)\big)\sigma = \big(h\xi\big(T(dh)\big)\omega_X\big)t \\ &= (h\xi\omega_X)(dht) \quad (\omega \text{ n.t.})\end{aligned}$$

But as Q is a Σ-algebra quotient of $\Gamma(X,\xi)$, if $ht = ht' \in Q^n$, $(h\xi\omega_X)t = (h\xi\omega_X)t'$. Setting $t' = dht$, indeed $ht' = hdhd = ht$, so (B) commutes and we are done. $\square$

It might seem that Theorem 3.44 would be the last word on monadic functors of sets. In practice, however, there are many natural set-theoretic constructions that produce monads and, in such cases, it may be difficult to identify what the algebras are. We'll explore some examples.

EXAMPLE 3.45. Let **P** be the power set monad of Example 2.16. Then $\mathcal{S}^{\mathbf{P}}$ may be identified with the category of complete sup-semilattices and morphisms which preserve all suprema. The structure map $\xi : PX \to X$ is $\xi(A) = Sup\, A$.

One guesses this result by realizing that $A \in PX$ can be written as $A = \bigcup(\{x\}\colon\ x \in A) = \bigcup(\eta_X\, x\colon\ x \in A)$ which expresses A as the supremum operation on its variables $\eta_X\, x$. But PX should be the X-ary operations of **P**. Let's check the details. Define Σ by $\Sigma_n = \{sup_n\}$ for each cardinal n. Define the class E of equations to be the following two schemes, where we write *sup* for whichever sup_n applies.

Partition associativity: If $(I_j\colon\ j \in J)$ partitions I,

$$sup\big(sup(x_j\colon\ j \in I_j)\colon\ j \in J\big) = sup(x_i\colon\ i \in I).$$

Idempotency: If $x_i = x$, $sup(x_i) = x$.

It is routine to show that (Σ, E)-**Alg** is complete semilattices. If (Y, sup) is a complete semilattice and if $f : X \to Y$ there exists unique sup-preserving $f^{\#}$ such that the following triangle commutes

$$\begin{tikzcd} X \arrow[r, "\eta_X"] \arrow[dr, "f"'] & (PX, \bigcup) \arrow[d, "f^{\#}"] \\ & (Y, sup) \end{tikzcd}$$

namely $f^{\#}(A) = sup(fx\colon\ x \in A)$. Thus (Σ, E) is tractable and the monad induced by the adjointness is readily seen to be the power set monad.

EXAMPLE 3.46. Let $\mathbf{P}^{-2}$ be the double contravariant power set monad. Then $\mathcal{S}^{\mathbf{P}^{-2}}$ is the category **CABA** of complete atomic Boolean algebras and complete Boolean algebra homomorphisms. The structure map $\xi : P^{-2}X \to X$ is

$$\xi(\mathcal{A}) = Sup(x\colon\ x \text{ is an atom, } {\uparrow}x \in \mathcal{A})$$

where ${\uparrow}x = \{y\colon\ y \geqslant x\}$.

To seek an operational representation of $\mathcal{A} \in P^{-2}X$, we can start with $\mathcal{A} = \bigcup(\{A\}\colon\ A \in \mathcal{A})$. The problem is then to figure out how to represent $\{A\}$ in terms of the "variables"

$\eta_X\ x = prin(x) = \{B\colon\ x \in B\}$. We have

$$\bigcap\big\{prin(x)\colon\ x \in A\big\} = prin(A) = \{B\colon\ B \supset A\}$$
$$\bigcap\big\{prin(x)'\colon\ x \notin A\big\} = \{B\colon\ B \subset A\}$$

so that

$$\mathcal{A} = \bigcup\Big(\bigcap\big(prin(x)\colon\ x \in A\big) \cap \bigcap\big(prin(x)'\colon\ x \notin A\big)\colon\ A \in \mathcal{A}\Big) \tag{21}$$

At this stage, complete Boolean algebras come to mind. This cannot be right, because free such algebras do not exist by the theorems of [30,35] alluded to earlier. On the other hand, 2^{2^X} is an atomic Boolean algebra, so is in **CABA**. Moreover, it is well-known that every complete atomic Boolean algebra is isomorphic to the algebra of all subsets of some set and that the homomorphisms $\psi : 2^Y \to 2^Z$ are given by $\psi = g^{-1}$ for unique function $g : Z \to Y$. Moreover, $(2^{2^X}, prin_X)$ is the free complete atomic Boolean algebra generated by X, as shown in the triangle below.

$$\begin{tikzcd} X \arrow[r, "prin_X"] \arrow[dr, "f"'] & 2^{2^X} \arrow[d, "f^\#"] \\ & 2^Y \end{tikzcd}$$

There is at most one such morphism $f^\#$ in **CABA** by (21). To see that $f^\#$ exists, define $f^\# = (\hat{f})^{-1}$ where $\hat{f} : Y \to 2^X$ is defined by $x \in \hat{f}\,y \Leftrightarrow y \in f\,x$. Then

$$f^\#\big(prin(x)\big) = \big\{y \in Y : \hat{f}\,y \in prin(x)\big\} = \big\{y \in Y\colon\ y \in f\,x\big\} = f\,x$$

We leave it to the reader to verify that the induced monad is $\mathbf{P}^{-2}$.

To see that the structure map $\xi : 2^{2^{2^X}} \to 2^X$ is as claimed, observe that for $i = id_{2^X} : 2^X \to 2^X$, $\hat{i} = prin_X : X \to 2^{2^X}$, so $\xi = (prin_X)^{-1}$. But for $\overline{\mathcal{A}} \in 2^{2^{2^X}}$,

$$\begin{aligned} prin^{-1}(\overline{\mathcal{A}}) &= \big\{x \in X\colon\ prin(x) \in \overline{\mathcal{A}}\big\} = Sup\big(\{x\}\colon\ prin(x) \in \overline{\mathcal{A}}\big) \\ &= Sup(A\colon\ A \text{ is an atom and } \uparrow A \in \overline{\mathcal{A}}) \end{aligned}$$

There are a couple of further points to explore about **CABA**. Observe that $\lambda : \mathbf{P}^{-2} \to \mathbf{P}^{-2}$ defined by

$$\lambda_X(\mathcal{A}) = \{A \subset X\colon\ X - A \notin \mathcal{A}\} \tag{22}$$

is a monad involution, $\lambda \circ \lambda = id_{\mathbf{P}^{-2}}$. λ corresponds to the more familiar duality $\Gamma : \mathbf{CABA} \to \mathbf{CABA}$ over $\mathcal{S}$ mapping $(X, \leqslant)$ to $(X, \geqslant)$.

We remarked in Example 2.16 that the principal filter map $prin : \mathbf{P} \to \mathbf{P}^{-2}$ is a monad map. This tells us that each complete atomic Boolean algebra (X, ξ) with supremum operation *Sup* is also a complete semilattice under formula of Theorem 3.39 with supremum operation Sup_0 given by

$$\begin{aligned} Sup_0(A) &= \xi\big(prin(A)\big) = Sup(x\colon\ x \text{ is an atom}, A \subset \uparrow x) \\ &= Sup(x\colon\ x \text{ is an atom}, x \in A) \end{aligned}$$

so, by atomicity, the two supremum operations coincide.

The next example is due to [75, Proposition 7.1].

EXAMPLE 3.47. Let $\mathbf{T}$ be any monad in $\mathcal{S}$. Let $\mathcal{A}$ be the category of *compact* $\mathbf{T}$*-algebras*. The objects are all $(X, \mathcal{T}, \xi)$ with $(X, \mathcal{T})$ a compact Hausdorff topological space and with (X, ξ) a $\mathbf{T}$-algebra such that $\forall n\ \forall \omega \in Tn\delta_\omega : X^n \to X$ is jointly continuous. Morphisms are continuous $\mathbf{T}$-homomorphisms. Then the forgetful functor $U : \mathcal{A} \to \mathcal{S}$ is monadic.

Here, recall, $\delta_\omega(f) = f^\#(\omega)$ for $f \in X^n$, where $f^\# : (Tn, \mu_n) \to (X, \xi)$ is the unique $\mathbf{T}$-homomorphic extension. Of course X^n has the product topology induced by $(X, \mathcal{T})$. If $\mathbf{T}$ corresponds to groups, semigroups, rings, ... the usual category of topological algebras results. Let's check the details. Compact Hausdorff spaces are monadic over $\mathcal{S}$ by Example 3.19, so have equational presentation (Σ_1, E_1). Similarly, $\mathbf{T}$-algebras have equational presentation (Σ_2, E_2). Then compact $\mathbf{T}$-algebras have equational presentation $(\Sigma_1 + \Sigma_2, E_1 + E_2 + E_3)$ where $+$ denotes disjoint union and E_3 records the equations that the Σ_2-operations are Σ_1-homomorphisms. (Note: Draw the diagram and you will see that this is symmetric – the Σ_1-operations are Σ_2-homomorphisms.) This operation on monadic functors is often written as tensor product. The result is always equational as just noted, so we have only to show in the case at hand that $U : \mathcal{A} \to \mathcal{S}$ satisfies the solution set condition. Let $(X, \mathcal{T}, \xi)$ in $\mathcal{A}$ and let D be a $\mathbf{T}$-subalgebra of (X, ξ). Let $E = \overline{D}$ be the topological closure of D. For any $\mathbf{T}$-operation $f : X^n \to X$ of (X, ξ), we have

$$\begin{aligned} f\big(E^n\big) &= f\big(\big(\overline{D}\big)^n\big) = f\big(\overline{D^n}\big) \subset \overline{(f(D^n))} \quad \text{(continuity)} \\ &\subset \overline{D} \quad (A \text{ is a } \mathbf{T}\text{-algebra}) \\ &= E \end{aligned}$$

so E is again a $\mathbf{T}$-subalgebra and hence is a compact $\mathbf{T}$-algebra. But then any $a : K \to UA$ factors through UE if E is closure of the $\mathbf{T}$-subalgebra generated by the image of a. As the cardinality of E cannot exceed that of βTK, the solution set condition is assured.

The tractability of the tensor product of monadic functors has been considered more generally by [58,28]. For a much deeper analysis see [41].

See [67, Corollary 2, p. 81], [3, Section 9.3], [76, Section 3.7] for results on the question of when $K^{\mathbf{T}}$ has colimits. The following basic result is adequate for our needs.

THEOREM 3.48. *Let* $\mathbf{T}$ *be a monad in* $\mathcal{K}$ *such that each pair* $f, g : (X, \xi) \to (Y, \theta)$ *has a coequalizer in* $K^{\mathbf{T}}$*. Let* Δ *be a small category. Then if* $\mathcal{K}$ *has colimits of type* Δ*, so does* $K^{\mathbf{T}}$*.*

PROOF. For K in $\mathcal{K}$ let $\widehat{K}$ be the functor $\Delta \to \mathcal{K}$ which is constantly K, i.e., for $t: i \to j$ in Δ, $\widehat{K}i = K$ and $\widehat{K}t = id_K$. Then $\mathcal{K}$ has limits of type Δ if and only of $\Lambda: \mathcal{K} \to \mathcal{K}^\Delta$, $K \mapsto \widehat{K}$ has a left adjoint. Let $(U^{\mathbf{T}})^\Delta$ be the functor $(K^{\mathbf{T}})^\Delta \to \mathcal{K}^\Delta$ which forgets the $\mathbf{T}$-algebra structure. Then the following square commutes.

$$\begin{tikzcd} K^{\mathbf{T}} \arrow[r, "\Lambda"] \arrow[d, "U^{\mathbf{T}}"'] & (K^{\mathbf{T}})^\Delta \arrow[d, "(U^{\mathbf{T}})^\Delta"] \\ \mathcal{K} \arrow[r, "\Lambda"'] & \mathcal{K}^\Delta \end{tikzcd}$$

Now apply Theorem 3.40. $\square$

COROLLARY 3.49. *For any monad* $\mathbf{T}$ *in* $\mathcal{S}$, $\mathcal{S}^{\mathbf{T}}$ *is small cocomplete.*

PROOF. Apply the preceding theorem. We shall show that $f, g: (X, \xi) \to (Y, \theta)$ has a coequalizer in $\mathcal{S}^{\mathbf{T}}$. By Theorem 3.44 we need only cite the well-known construction for (Σ, E)-algebras – divide out by the smallest congruence containing $\{(fx, gx)\colon x \in X\} \subset Y \times Y$. $\square$

COROLLARY 3.50. *Every functor over* $\mathcal{S}$ *between monadic functors is itself monadic.*

PROOF. This is immediate from Proposition 3.35 and Theorem 3.40 in view of the preceding theorem. $\square$

An intriguing question is to characterize, in some familiar way, the opposite of a known category. Well known nontrivial examples are Stone duality between Boolean algebras and Stone spaces and Pontryagin duality for locally compact Abelian groups. This is a large topic, admirably treated in [45] (see especially Chapter VI; also see the extensive bibliography there). A certain fragment of this theory is captured in the "contravariant representation theorem" due to Linton in [69], and we turn now in the direction of a statement of that result. Some background will be needed.

We seek to characterize $\mathcal{A}^{op}$ particularly when $\mathcal{A} = \mathcal{S}^{\mathbf{T}}$. Our approach will be to seek a functor $H: \mathcal{A}^{op} \to \mathcal{S}$ with a left adjoint in such a way that the semantic comparison functor is a full subcategory, thereby representing $\mathcal{A}^{op}$ as being some of the $\mathbf{S}$-algebras for appropriate $\mathbf{S}$. Up to natural equivalence, $H = \mathcal{A}(-, J)$ where J is free on one generator with respect to H (i.e., element of HA = function $1 \to HA = \mathcal{A}^{op}$-morphism $J \to A = \mathcal{A}$-morphism $A \to J$).

Based on these observations, we start with a functor $U: \mathcal{A} \to \mathcal{S}$ such that $\mathcal{A}$ has and U preserves all small limits (certainly true if $\mathcal{A} = \mathcal{S}^{\mathbf{T}}$), and a chosen object J of $\mathcal{A}$. The functor $\mathcal{A}(-, J): \mathcal{A}^{op} \to \mathcal{S}$ has left adjoint $n \mapsto J^n$, the product being taken in $\mathcal{A}$ not $\mathcal{A}^{op}$. The resulting monadic completion is a generalization of the double dualization monads of Definition 3.13 which is the special case $\mathcal{A} = \mathcal{S}$. Let's record the general definition.

DEFINITION 3.51. Let $\mathcal{A}$ be a locally small category and let J be an object of $\mathcal{A}$ for which the nth power J^n exists in $\mathcal{A}$ for all sets n in $\mathcal{S}$. The *double dualization monad* $\mathbf{D}_J$ *in* $\mathcal{S}$ *induced by* J is given by

$$
\begin{aligned}
&D_J X = \mathcal{A}\big(J^X, J\big)\\
&\eta_X x = J^X \xrightarrow{pr_x} J\\
&\alpha^\#(J^X \xrightarrow{f} J) = J^Y \xrightarrow{\hat{\alpha}} J^X \xrightarrow{f} J
\end{aligned}
$$

where $\alpha : X \to \mathcal{A}(J^Y, J)$, $\hat{\alpha}(g)(x) = (\alpha x)(g)$. The *double dualization monad* $\mathbf{D}^J$ *in* $\mathcal{A}$ *induced by* J is given by

$$
\begin{aligned}
&D^J A = J^{\mathcal{A}(A,J)}\\
&pr_{A \xrightarrow{f} J}\, \eta_A = f\\
&pr_{B \xrightarrow{f} J}\, \alpha^\# = pr_{f\alpha}
\end{aligned}
$$

$\mathbf{D}_J$ is the monad induced by the adjointness $(\mathcal{A}^{op}, \mathcal{K}, \mathcal{A}(-, J), n \mapsto J^n, \eta, \varepsilon)$. The same adjointness induces a comonad in $\mathcal{A}^{op}$, that is, a monad in $\mathcal{A}$, namely $\mathbf{D}^J$. We leave these routine verifications to the reader.

A further background result of general interest is the following one about quasivarieties.

PROPOSITION 3.52. *Let $\mathcal{A}$ be a locally small category with small limits and with image factorization system $(\mathcal{E}, \mathcal{M})$. Let J be any object of $\mathcal{A}$ and let $\mathcal{B}$ be the quasivariety generated by J (i.e., the intersection of all quasivarieties containing J). Then the following statements hold.*

(1) *X is in $\mathcal{B}$ if and only if $ev_X : X \to J^{\mathcal{A}(A,J)}$ is in $\mathcal{M}$.*

(2) *For X in $\mathcal{A}$, if ev_X has $\mathcal{E}$-$\mathcal{M}$ factorization me, e is the reflection of X in $\mathcal{B}$.*

PROOF. If ev_X is in $\mathcal{M}$ then X is an $\mathcal{M}$-subobject of a product of copies of J, so is in $\mathcal{B}$. Conversely let X be an object of $\mathcal{B}$. For any image factorization system, if $m_i : A_i \to B_i$ is in $\mathcal{M}$, so is $\prod m_i : \prod A_i \to \prod B_i$, so the class of $\mathcal{M}$-subobjects of powers of J is closed under products, and hence coincides with $\mathcal{B}$. As such, there exists I and $m : X \to J^I$ with $m \in \mathcal{M}$. Since the square below commutes,

$$
\begin{tikzcd}
X \arrow[r, "ev_X"] \arrow[d, "m"'] \arrow[dr, phantom, "\text{(A)}", near start] & J^{\mathcal{A}(X,J)} \arrow[dl, "\psi"] \arrow[d, "pr_{pr_i m}"] \\
J^I \arrow[r, "pr_i"'] & J \arrow[ul, phantom, "\text{(B)}", near start]
\end{tikzcd}
$$

there exists unique ψ such that (B) commutes and then (A) also commutes because it does followed by each pr_i. Thus $ev_X \in \mathcal{M}$. We turn to the proof of the second statement. For any X in $\mathcal{A}$, write $ev_X = X \xrightarrow{e} B \xrightarrow{m} J^{\mathcal{A}(X,J)}$ with $e \in \mathcal{E},\ m \in \mathcal{M}$. Clearly $B \in \mathcal{B}$. If also $C \in \mathcal{B}$ and $f : X \to C$, consider the diagram

$$\begin{tikzcd}
X \arrow[r, "e"] \arrow[dd, "f"'] & B \arrow[d, "m"] \arrow[ddl, "g"'] \\
 & J^{\mathcal{A}(X,J)} \arrow[d, "D^J f"] \\
C \arrow[r, "ev_i"'] & J^{\mathcal{A}(C,J)}
\end{tikzcd}$$

The perimeter of the rectangle commutes. By the first part of the proof, $ev_C \in \mathcal{M}$ so the desired reflection-induced g is defined by diagonal fill-in. $\square$

We are now ready to state the *contravariant representation theorem* of [69, Section 2]. For a proof see that reference and use the preceding proposition.

THEOREM 3.53. *Let* $\mathbf{T}$ *be a monad in* $\mathcal{S}$ *such that every injective* $\mathbf{T}$*-homomorphism is the equalizer in* $\mathcal{S}^{\mathbf{T}}$ *of some pair of maps and let* J *be an injective cogenerator of* $\mathcal{S}^{\mathbf{T}}$ *inducing the double dualization monad* $\mathbf{D}_J$ *in* $\mathcal{S}$. *Then the semantic comparison functor* Φ *of the functor* $\mathcal{S}^{\mathbf{T}}(-,J) : (\mathcal{S}^{\mathbf{T}})^{op} \to \mathcal{S}$ *is a full reflective subcategory. Moreover,* (J, pr_{id_J}) *is a* $\mathbf{D}_J$*-algebra and the quasivariety it generates consists of all objects isomorphic to an object in the image of* Φ.

EXAMPLE 3.54. Let $\mathbf{T}$ be the identity monad in $\mathcal{S}$, $J = 2$. Then $\mathbf{D}_2$ is the double contravariant power set monad whose algebras are complete atomic Boolean algebras (Example 3.46). Monics are equalizers in $\mathcal{S}$ and every set with at least two elements is an injective cogenerator. The semantics comparison functor

$$\begin{tikzcd}
\mathcal{S}^{op} \arrow[rr, "\Phi"] \arrow[dr, "\mathcal{S}(-,2)"'] & & \mathbf{CABA} \arrow[dl] \\
 & \mathcal{S} &
\end{tikzcd}$$

with $\Phi X = 2^X$ is an equivalence of categories.

EXAMPLE 3.55. Let $\mathcal{S}^{\mathbf{T}}$ be the category $\mathbf{BA}$ of Boolean algebras, and let $J = 2$. Since an ultrafilter on X amounts to a Boolean algebra homomorphism $2^X \to 2$, it is no surprise that $\mathbf{D}_2 = \boldsymbol{\beta}$ as is easily checked. Thus the semantics comparison functor of $(-,2) : (\mathbf{BA})^{op} \to \mathcal{S}$ maps $(\mathbf{BA})^{op}$ to the category of compact Hausdorff spaces which are subspaces of a power of the two-element discrete space. But these are just the compact Hausdorff totally disconnected spaces.

Realistically, it is just as much work to prove that 2 is an injective cogenerator in **BA** and that monics are equalizers in **BA** as it is to prove Stone duality directly, but it is still nice to see that this is one of the examples of contravariant representation.

With regard to the next example see [17,43,89].

EXAMPLE 3.56. Complex commutative $C^\star$-algebras are monadic over $\mathcal{S}$ and its opposite is equivalent to the category of compact Hausdorff spaces. This is generally known as Gelfand duality. Negrepontis proved this using contravariant representation. Isbell provided a concrete equational description of commutative $C^\star$-algebras.

4. Semantics of programming languages

Central issues in computer programming language design include maintenance of code, safety, perspicuous data structures and formal methods to prove that programs are correct to specification. An *operational semantics* of a programming language is a formal description of what the computer will do when a program is run. The manuals that come with commercial programming languages provide (usually informal) operational semantics. A *denotational semantics* gives a mathematical description (called a *denotation*) of a program in some formal system more friendly to mathematical analysis than the programming language itself. For example, functional programs have a denotation as a closed term in an appropriate version of the λ-calculus. Such denotations can usually be reduced to a normal form.

There are many approaches to denotational semantics, each providing certain mathematical insights into the analysis of programs. A general discussion of this subject is far beyond the scope of this paper. Here, we shall focus on certain denotational models involving monads.

4.1. *Elgot*

Elgot is the first to use a fragment of category theory, namely the first order theory of categories with finite coproducts $X_1 + \cdots + X_n$, as a framework for denotational semantics. (The first-order nature of category theory was strongly emphasized in Lawvere's thesis [55].) Here coproducts are thought of as a "disjoint union" which is precisely the case in $\mathcal{S}$. In another familiar category of structures, the modules over a ring, the binary coproduct is quite different from a union being the direct sum $X \oplus Y$ which is the Cartesian product module with coproduct injections $in_X(x) = (x, 0)$, $in_Y(y) = (0, y)$. Hence axioms are needed for "coproducts to be like disjoint unions", a point to which we shall return shortly.

Elgot considered a morphism of form $X_1 + \cdots + X_m \to Y_1 + \cdots + Y_n$ to be the semantics of a network element with m input lines and n output lines where the ith input line receives inputs from the value object X_i and the jth output line takes values in the object Y_j. For example, for $f, g : X \to Y$ and P a test on X, the construct

$$\textit{if } P \textit{ then } f \textit{ else } g : X \to Y$$

is synthesized as follows. Represent P (as well as not-P) by a coproduct

$$P \xrightarrow{i} X \xleftarrow{i'} P'$$

Given $t : P \to Z$, $u : P' \to Z$, the coproduct property asserts that there exists a unique morphism $\psi : X \to Z$ with $\psi i = t$ and $\psi i' = u$; write such ψ as $\langle t, u\rangle$. Then

$$\textit{if } P \textit{ then } f \textit{ else } g = \langle f i, g i'\rangle$$

In general, all loop-free flowschemes find semantics in a category with finite coproducts.

Iteration is an additional construct which postulates the existence of an iteration operation of form

$$X \xrightarrow{f} X + Y \mapsto X \xrightarrow{f^\dagger} Y$$

Here, in the simplest case (with the disjoint union interpretation of $X + Y$ of paramount importance), $f(x)$ is in exactly one of X, Y so that f can be iterated until (if ever) the result is in Y and that value in Y is $f^\dagger(x)$. One of Elgot's main observations is that the fundamental fixed point equation for $f^\dagger$ is easily expressed in the denotational language, amounting to the assertion that the following diagram commutes:

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[dr, "f^\dagger"'] & X + Y \arrow[d, "\langle f^\dagger, id_Y\rangle"] \\
& Y
\end{tikzcd}$$

Many workers developed Elgot's ideas. We refer the reader to [11] for a textbook treatment and an extensive bibliography. For Elgot's contributions, see [10].

We turn now to discussing Elgot's choice of category and what that has to do with monads. We begin with a proposition.

PROPOSITION 4.1. *Let* $\mathbf{T}$ *be a monad in* $\mathcal{K}$. *Then the canonical functor* $F_{\mathbf{T}} : \mathcal{K} \to K_{\mathbf{T}}$ *preserves coproducts.*

PROOF. $F_{\mathbf{T}}$ preserves all colimits, in fact, since it has $U_{\mathbf{T}}$ as a right adjoint. It is instructive, however, to see the construction details for coproducts. Consider the diagram

$$\begin{tikzcd}
X_i \arrow[r, "in_i"] \arrow[dr, "\alpha_i"'] & X \arrow[r, "\eta_X"] \arrow[d, "\alpha"] & TX \arrow[dl, "\alpha^\#"] \\
& TY &
\end{tikzcd}$$

Here, $in_i : X_i \to X$ is a coproduct in $\mathcal{K}$. We must show that $\eta_X in_i : X_i \to TX$ is a coproduct in $K_{\mathbf{T}}$. Given α_i as shown, there exists unique α in $\mathcal{K}$ with $\alpha in_i = \alpha_i$. Then $\alpha^{\#}(\eta_X in_i) = \alpha_i$ so $\alpha \circ (\eta_X in_i) = \alpha_i$ in $K_{\mathbf{T}}$. If $\beta : X \to TY$ also satisfied $\beta \circ (\eta_X in_i) = \alpha_i$ in $K_{\mathbf{T}}$ then, in $\mathcal{K}$, $\beta in_i = \beta^{\#} \eta_X in_i = \alpha_i$ so $\beta = \alpha$. □

Elgot chose Lawvere's algebraic theories as a denotational framework. This was a surprising development since these categories were invented as a means to describe their models, not as useful categories in their own right. The theory could have been developed with few changes in the Kleisli category of a monad of sets (recall Proposition 3.11), it being a minor quibble that these have arbitrary rather than just finite coproducts. Both Lawvere theories and Kleisli categories for monads of sets have disjoint union as coproduct (in the latter case, use the preceding proposition).

There is little difference in the denotational semantics between Lawvere theories and Kleisli categories so long as one sticks to the first order theory of categories with finite coproducts. The difference becomes significant in the section after next where the denotational syntax includes T, η and $(-)^{\#}$ as well.

The next example appears in [22].

EXAMPLE 4.2. Let D be a fixed set of "external memory states". As in Proposition 3.11, let $\mathcal{L}$ be the Lawvere theory whose objects (as always) are the finite cardinals and for which a morphism $m \to n$ is a function $m \times D \to n \times D$ in $\mathcal{S}$, noting that the m-fold copower $m \cdot D$ is the binary Cartesian product $m \times D$ in $\mathcal{S}$.

In this category, $f : m \times D \to n \times D$ has the interpretation that if the ith instruction starts in external state d and $f(i, d) = (j, e)$ then the new external state is e and the next instruction to be executed is the jth.

4.2. *Continuous lattices*

A formal operational semantics for a programming language defines its syntax recursively. Consider the following (oversimplified) mutually recursive equations for tests and expressions (involving variables and given constants):

$$Test = Expr > Expr \mid Expr == Expr \mid Test \vee Term \mid Test \wedge Test \mid \neg Test \tag{23}$$

$$Expr = Term \mid if\ Test\ then\ Expr\ else\ Expr \tag{24}$$

An example of an expression with a test is

$$if\ \big(x + y^2 > 20\big) \wedge \neg(y == 3)\ then\ x + y\ else\ x^2 * y$$

To explain the syntax, one would read Equation (23) as "a test is either $e_1 > e_2$ where e_1, e_2 are expressions or $e_1 == e_2$ where e_1, e_2 are expressions ($==$ means equals, syntactically different from the equals sign in the equation) or $t_1 \vee t_2$ where t_1, t_2 are tests, …", the equation for expressions similarly. Here the set *Term* of terms such as $x + y^2$ is defined elsewhere. The sets *Test*, *Expr* are defined by these "equations" by the *Kleene fixed point theorem* [50] which is stated as follows.

PROPOSITION 4.3. *Let $(X, \leqslant)$ be a partially ordered set with a least element $\perp$ and with suprema of ascending chains. Let $f : (X, \leqslant) \to (X, \leqslant)$ be* continuous, *i.e., f preserves suprema of ascending chains. Then $\bigvee_n f^n(\perp)$ exists and is the least fixed point of f.*

To apply this theorem to the current example, let A be the program alphabet and let A^+ denote the set of all non-empty strings on A (including all elements of *Term*, *Test* and *Expr*). Let $(X, \leqslant) = 2^{A^+} \times 2^{A^+}$ with the product order induced by subset inclusion. Thus $\perp = (\emptyset, \emptyset)$. Define f by

$$\begin{aligned} f(W, V) = (&V > V \cup V == V \cup W \vee W \cup W \wedge W \cup \neg W, \\ &\mathit{Term} \cup \ \mathit{if}\ W\ \mathit{then}\ V\ \mathit{else}\ V) \end{aligned}$$

That f is continuous is routine. Set (*Test*, *Expr*) to be the least fixed point of f. Thus

$$\begin{aligned} f(\emptyset, \emptyset) &= (\emptyset, \mathit{Term}) \\ f^2(\emptyset, \emptyset) &= (\mathit{Term} > \mathit{Term} \cup \mathit{Term} == \mathit{Term}, \mathit{Term}) \\ &\ \ \cdots \end{aligned}$$

is the inclusion chain whose union is the desired pair of sets. This operational syntax sets up structural induction for a denotational semantics relative to a specified Σ-algebra X that interprets the terms as elements of X. The semantics of a test is then a subset of X^m for some m and the semantics of an expression is a function $X^n \to X$ for some n.

New problems arise trying to carry out this style of operational syntax and denotational semantics for a functional programming language. Consider the following.

$$\begin{aligned} &\mathit{plus} :: \mathit{Integer} \to (\mathit{Integer} \to \mathit{Integer}) \\ &\mathit{map}(\mathit{plus}\ 5)\ [1, 2, 3] \end{aligned}$$

The first line describes the type of the addition function *plus*. This could have been implemented in uncurried form

$$\mathit{plus} :: (\mathit{Integer},\ \mathit{Integer}) \to \mathit{Integer}$$

so that *plus*$(5, 7)$ returns 12, but *plus* 5 would be undefined. With the original definition, *plus* 5 is the function $\lambda x(x + 5)$, λ-calculus notation for the function $x \mapsto x + 5$, and this is very useful to have. For example, in the second line above we can map the function *plus* 5 over the list $[1, 2, 3]$ to return $[6, 7, 8]$.

More shall be said about functional programming in the next section. Here we notice a new problem, the complexity of the set of values. In the scenario of (23), (24), an unspecified fixed set can serve as the set of values and its denotational semantics is the carrier set of the algebra X above. Here, however, values can be either basic values X or functions such as *plus*5 since, as we just saw, *plus*5 is an argument to the map operator and arguments to operators are values. Many other functions would, in fact, need to be values also, but at the very least we would need to deal with the following set V of values defined recursively by the fixed point equation

$$V = X \mid [V \to V] \tag{25}$$

where $[A \to B]$ is the set of functions from A to B. At the level of operational syntax, there is a least fixed point solution in the same style as for (23), (24). At the denotational level, where one must implement what the symbols mean, it is much less clear what to do. We expect that (25) is to be true as an equation, at least to the extent that the left- and right hand sides are isomorphic in a natural way. If $[V \to V]$ means all functions from V to itself, the cardinality of the right hand side is always greater than that of the left. The problems are more subtle if a more restricted class of functions is used, but in the end one can show that there is no meaningful denotational semantics for (25) if this equation is in the category $\mathcal{S}$ of sets.

D.S. Scott proposed a revolutionary approach to denotational semantics based on partially ordered sets. Early papers are [97,99]. Scott's idea was that, for denotational purposes, the domain and codomain of a computable function should be a partially ordered set, where the order relation $x \leqslant y$ is interpreted as "y has at least as much information as x". There should be a least element $\perp$ to express "no information" and each ascending chain $x_n \leqslant x_{n+1}$ should have a supremum x (with x_n "approximating x", with bigger n giving better approximations). Many additional properties have been suggested over the years and there is yet no complete consensus on any one set of axioms. One of the earliest was the concept of a *continuous lattice* announced in [98]. There are many equivalent definitions of continuous lattice. The one we give here is that of Proposition 2.4 in the paper of Scott just cited.

Toward that definition, let L be a complete lattice. Recall that a subset D of a partially ordered set is *directed* if it is nonempty and if $\forall x,\ y \in D\ \exists z \in D$ with $x \leqslant z$, $y \leqslant z$. What was subsequently called the *Scott topology* on L has as open sets those $U \subset L$ satisfying (a) U is an upper set, that is, $u \in U$ and $x \geqslant u \Rightarrow x \in U$, and (b) for all directed (hence nonempty, by definition) $D \subset L$, $\bigvee D \in U \Rightarrow U \cap D \neq \emptyset$. Following a suggestion of [24], a Scott open set is an observable, representing the computational manifestation of a point. $x \in U$ means "observing U is consistent with x", so the meaning of axiom (a) is clear. Axiom (b) states that if x is the limit of the finite approximations x_n, then any observation of x must be consistent with one of the approximations.

That said, a *continuous lattice* is a complete lattice L with the property that the following equation holds for each $x \in L$:

$$x = \bigvee \Bigl(\bigwedge U \colon\ x \in U, U \text{ Scott open} \Bigr) \tag{26}$$

Note that each $\bigwedge U \leqslant x$. Thinking of the directed set ($\bigwedge U \colon\ x \in U, U$ Scott open) as the set of approximations of x, axiom (26) says that each x is the limit of its approximations. We refer the reader to [31] for a wealth of information about continuous lattices.

Scott was able to show that continuous lattices could be used for a satisfactory denotational semantics. He showed that for $[D \to D]$ the set of continuous endomorphisms of D, there exists a nontrival continuous lattice D and a suitable isomorphism for (25).

The *Lawson topology* on a continuous lattice is the smallest topology whose open sets include the Scott open ones as well as all sets of form $\{y\colon\ y \not\geqslant x\}$ (for all x). The Lawson topology is compact Hausdorff and binary infimum is Lawson continuous. See [31, Section III.1].

A. Day and O. Wyler, working independently, discovered the filter monad (which we shall define shortly) and proved the theorem immediately below. See [14,111].

PROPOSITION 4.4 [14]. *A complete lattice L is continuous if and only if for all families $(D_i\colon\ i \in I)$ of directed subsets of L,*

$$\bigwedge_{i\in I} \bigvee D_i = \bigvee_{(d_i)\in\prod D_i} \bigwedge_{i\in I} d_i$$

THEOREM 4.5. *The category of continuous lattices and morphisms which preserve all infima and directed suprema is monadic over $\mathcal{S}$ being, in fact, the algebras over the filter monad.*

Day and Wyler used different proofs and we will outline each. But first we shall describe the filter monad in $\mathcal{S}$. Later in this section we will construct a similar monad in the category of $T0$ topological spaces and study a special class of monads called *Kock–Zöberlein* monads whose algebras are always a subcategory of the base category.

If X is a set, a *filter* on X is a non-empty collection $\mathcal{F}$ of subsets of X satisfying $\forall F, G \subset X,\ F, G \in \mathcal{F} \Leftrightarrow F \cap G \in \mathcal{F}$. For $A \subset X$, $prin(A) = \{B \subset X\colon\ A \subset B\}$ is a filter on X, the *principal filter* generated by A. $prin(\emptyset) = 2^X$ is the *improper filter*; all other filters are *proper*. If $\mathcal{F}$ is a proper filter, $\emptyset \notin \mathcal{F}$.

EXAMPLE 4.6. The *filter monad* $\mathbf{F} = (F, \eta, (-)^\#)$ in $\mathcal{S}$ is defined by

$$\begin{aligned}
&FX = \text{set of all filters on } X\\
&\eta_X x = prin(x) \quad (\text{i.e., } prin(\{x\}))\\
&\text{For } \alpha : X \to FY, \alpha^\#(\mathcal{F}) = \big\{B \subset Y\colon\ \{x \in X\colon\ B \in \alpha x\} \in \mathcal{F}\big\}
\end{aligned}$$

The monad laws are checked as follows.

$$\begin{aligned}
\eta_X^\#(\mathcal{F}) &= \big\{B \subset X\colon\ \{x \in X\colon\ B \in prin(x)\} \in \mathcal{F}\big\}\\
&= \big\{B \subset X\colon\ \{x \in X\colon\ x \in B\} \in \mathcal{F}\big\}\\
&= \{B \subset X\colon\ B \in \mathcal{F}\} = \mathcal{F}\\
\alpha^\# \eta_X x &= \alpha^\#\big(prin(x)\big)\\
&= \big\{B \subset X\colon\ \{y \in X\colon\ B \in \alpha y\} \in prin(x)\big\}\\
&= \{B \subset X\colon\ B \in \alpha x\} = \alpha x
\end{aligned}$$

For $\alpha : X \to FY$, $\beta : Y \to FZ$,

$$\begin{aligned}(\beta^{\#}\alpha)^{\#}(\mathcal{F}) &= \big\{C \subset Z\colon \big\{x \in X\colon C \in \beta^{\#}\alpha x\big\} \in \mathcal{F}\big\}\\ &= \big\{C \subset Z\colon \big\{x \in X\colon \{y \in Y\colon C \in \beta y\} \in \alpha x\big\} \in \mathcal{F}\big\}\\ &= \big\{C \subset Z\colon \{y \in Y\colon C \in \beta y\} \in \alpha^{\#}(\mathcal{F})\big\}\\ &= \beta^{\#}\alpha^{\#}(\mathcal{F})\end{aligned}$$

Moreover, for $f : X \to Y$, it is easily checked that

$$(Ff)\mathcal{F} = \big\{B \subset Y\colon f^{-1}(B) \in \mathcal{F}\big\} \tag{27}$$

We note that the ultrafilter monad (17) is a submonad of **F**.

We first outline Day's proof that $\mathcal{S}^{\mathbf{F}}$ = continuous lattices (with morphisms that preserve directed suprema and all infima). First consider the complete lattice $(FX, \subset)$ and note that infima are intersections (the empty one being the improper filter) and that *directed* suprema are unions. For $A \subset X$ define

$$\Box A = \{\mathcal{F} \in FX\colon A \in \mathcal{F}\} \tag{28}$$

It is obvious that $\Box A$ is Scott open. Moreover, if $\mathcal{F} \in \overline{\mathcal{A}}$ with $\overline{\mathcal{A}}$ Scott open, $\mathcal{F} = \bigcup(prin(A)\colon A \in \mathcal{F}) \in \overline{\mathcal{A}} \Rightarrow \exists A \in \mathcal{F}$ with $prin(A) \in \overline{\mathcal{A}}$ so that for $\mathcal{G} \in \Box A$, $A \in \mathcal{G} \Rightarrow prin(A) \subset \mathcal{G} \Rightarrow \mathcal{G} \in \overline{\mathcal{A}}$. This shows that $\{\Box A\colon A \subset X\}$ is a base for the Scott topology (noting that $\Box(A \cap B) = (\Box A) \cap (\Box B)$). It is now a short calculation to show that $(FX, \subset)$ is a continuous lattice. For $\mathcal{F} \in FX$,

$$\begin{aligned}\bigcup\Big(\bigcap \overline{\mathcal{A}}\colon \mathcal{F} \in \overline{\mathcal{A}}, \overline{\mathcal{A}}\text{ Scott open}\Big) &= \bigcup\Big(\bigcap \overline{\mathcal{A}}\colon A \in \mathcal{F}, \Box A \subset \overline{\mathcal{A}}\Big)\\ &= \bigcup\Big(\bigcap \Box A\colon A \in \mathcal{F}\Big)\\ &= \bigcup\big(prin(A\colon A \in \mathcal{F})\big) = \mathcal{F}\end{aligned}$$

We next see that (FX, η_X) is the free continuous lattice generated by X. For let $(Y, \leqslant)$ be a continuous lattice and let $f : X \to Y$ be any function. We must show that there exists unique $f^{\#} : (FX, \subset) \to (Y, \leqslant)$ preserving directed suprema and all infima and which extends f. Uniqueness is clear since $f^{\#}$ must satisfy

$$\begin{aligned}f^{\#}(\mathcal{F}) &= f^{\#}\Big(\bigcup_{F \in \mathcal{F}} \bigcap_{x \in F} prin(x)\Big)\\ &= \bigvee_{F \in \mathcal{F}} \bigwedge_{x \in F} fx\end{aligned}$$

Using this formula to define $f^{\#}$, it is routine to check that $f^{\#}(prin(x)) = fx$ and that $f^{\#}$ preserves directed suprema. Use the equation of Proposition 4.4 to show that $f^{\#}$ preserves infima as follows. For the empty infimum, $f^{\#}(2^X) = \bigwedge(\bigvee f(F)\colon\ F \subset X) = \bigwedge \emptyset$ is the greatest element of Y. For $(\mathcal{F}_i)$ a non-empty family of filters, define

$$D_i = \left\{ \bigwedge f(F)\colon\ F \in \mathcal{F}_i \right\}$$

Then each D_i is directed. We have

$$\begin{aligned}
\bigwedge_i f^{\#}(\mathcal{F}_i) = \bigwedge_i \bigvee D_i &= \bigvee_{(F_i)\in\prod\mathcal{F}_i} \bigwedge_i \bigwedge_{x\in F_i} fx \\
&= \bigvee_{(F_i)\in\prod\mathcal{F}_i} \Bigl(\bigwedge_{x\in\bigcup F_i} fx \Bigr) \\
&\leqslant \bigvee_{F\in\bigcap\mathcal{F}_i} \bigwedge_{x\in F} fx \quad \Bigl(\text{because } \bigcup F_i \in \bigcap \mathcal{F}_i\Bigr) \\
&= f^{\#}\Bigl(\bigcap \mathcal{F}_i\Bigr)
\end{aligned}$$

whereas $f^{\#}(\bigcap \mathcal{F}_i) \leqslant \bigvee_i f^{\#}(\mathcal{F}_i)$ is trivial.

One can complete the proof that $\mathcal{S}^{\mathbf{F}}$ is continuous lattices. It is *not* enough, by Theorem 3.44, to cite the equational presentation of [31, Theorem 2.3 (DD^*), p. 58], since these morphisms preserve all suprema, not just the directed ones.

Wyler's proof begins with the observation, easily established, that the ultrafilter monad β is a submonad of $\mathbf{F}$ as is the power set monad $\mathbf{P}$ via the monad map *prin*: $\mathbf{P} \to \mathbf{F}$ which maps A to $prin(A)$. It follows that every $\mathbf{F}$-algebra is both a compact Hausdorff space and a complete inf-semilattice. It is isomorphic to consider the $\mathbf{P}$-algebras to be complete sup-semilattices, but the former choice is more natural here by the following reasoning. By Theorem 3.44 and Day's proof, the $\mathbf{P}$-structure map of the $\mathbf{F}$-algebra $\xi : FX \to X$ maps $A \subset X$ to $\xi(prin(A)) = \xi(\bigcap_{x\in A} prin(x)) = \bigwedge_{x\in A} \xi(prin(x)) = \bigwedge A$. Thus the forgetful functor from the continuous lattices established by Day's proof to $\mathbf{P}$-algebras remembers the infimum structure.

Now back to Wyler's proof, taking the point of view that the $\mathbf{F}$-algebras are yet unknown but are, at least, compact Hausdorff spaces and complete inf-semilattices with morphisms that are continuous and preserve all infima. By the ultrafilter theorem, which states that every filter is the intersection of its containing ultrafilters, $\xi : FX \to X$ must satisfy $\xi(\mathcal{F}) = \xi(\bigcap \mathcal{U}\colon\ \mathcal{F} \subset \mathcal{U} \in \beta X) = \bigwedge_{\mathcal{F}\subset\mathcal{U}\in\beta X} \xi(\mathcal{U})$, where ξ preserves infima by virtue of being a $\mathbf{P}$-homomorphism (Theorem 3.44 again). But $\xi(\mathcal{U})$ is the unique point to which the ultrafilter $\mathcal{U}$ converges. This shows that the $\mathbf{F}$-algebra structure is determined by the compact Hausdorff and complete inf-semilattice structures.

Many details are necessary to complete the proof. At the end, the compact Hausdorff topology assigned to a continuous lattice is precisely the Lawson topology. Much more detail is given in Wyler's paper and in [45, Section VII.3]

We turn next to the idea of a Kock–Zöberlein monad. To begin, an *order-enriched category* is a pair $(\mathcal{K}, \leqslant)$ where $\mathcal{K}$ is a category and $\leqslant$ assigns a partial order $\leqslant_{BC}$ to the set $\mathcal{K}(B, C)$ of $\mathcal{K}$-morphisms from B to C in such a way that for all $f : A \to B$,

$g_1, g_2 : B \to C$, $h : C \to D$, $g_1 \leqslant_{BC} g_2 \;\Rightarrow\; hg_1 f \leqslant_{AD} hg_1 f$. In most cases, we drop the subscripts and simply write $\leqslant$.

Recall that a topological space is $T0$ if $x \neq y \;\Rightarrow\; \mathcal{N}_x \neq \mathcal{N}_y$, where $\mathcal{N}_x$ is the filter of all neighborhoods of x.

EXAMPLE 4.7. Let **T0** be the category of $T0$ topological spaces and continuous maps. It follows that $x \leqslant y$ if $\mathcal{N}_x \subset \mathcal{N}_y$ is a partial order on a given $T0$ space, the so-called *specialization order*. (**T0**, $\leqslant$) is an order-enriched category under the pointwise specialization order since $f : X \to Y$ is continuous if and only if $\forall N \in \mathcal{N}_{fx} \exists M \in \mathcal{N}_x$ with $f(M) \subset N$.

DEFINITION 4.8. Let $(\mathcal{K}, \leqslant)$ be an order-enriched category and let $\mathbf{T} = (T, \eta, (-)^{\#})$ be a monad in $\mathcal{K}$. Then $\mathbf{T}$ is *Kock–Zöberlein* if the following two conditions hold.

$$f \leqslant g \;\Rightarrow\; Tf \leqslant Tg \tag{29}$$

$$\eta_{TX} \leqslant T\eta_X \tag{30}$$

The fundamental facts about Kock–Zöberlein monads, at a more general level than order-enriched categories, were established by [53]. The ideas presented here were adapted from Kock's paper and carried further by [23].

EXAMPLE 4.9. The *open filter monad* $\mathbf{F}^\circ = (F^\circ, \eta, (-)^{\#})$ is the monad in **T0** described as follows.

(Note the use of the interior symbol $(-)^\circ$ to distinguish this monad from the filter monad $\mathbf{F}$ of sets.) For X a $T0$-space, let $\Omega(X)$ denote the topology of open subsets of X. A *filter* on $\Omega(X)$ is a subset $\mathcal{F}$ of $\Omega(X)$ satisfying $\forall A, B \in \Omega(X)$, $A, B \in \mathcal{F} \Leftrightarrow A \cap B \in \mathcal{F}$. Then

$$F^\circ X = \big\{\mathcal{F} \subset \Omega(X) \colon \ \mathcal{F} \text{ is a filter on } \Omega(X)\big\}$$
$$\eta_X x = prin(x) = \big\{U \in \Omega(X) \colon \ x \in U\big\}$$
$$\text{For } \alpha : X \to F^\circ Y, \quad \alpha^{\#}\mathcal{F} = \big\{V \in \Omega(Y) \colon \ \{x \in X \colon \ V \in \alpha x\} \in \mathcal{F}\big\}$$

For the discrete topology $\Omega(X) = 2^X$, $F^\circ X = FX$.

This is an important example so, with an eye toward future generalizations, we shall verify the asserted properties in some detail. The proof of the monad laws is too similar to Example 4.6 to bear repeating, but there is more to do. We must define $F^\circ X$ as a $T0$-space and then prove that η_X is continuous and that $\alpha^{\#}$ is continuous when α is. After that, the Kock–Zöberlein axioms must be verified.

For $U \in \Omega(X)$ define

$$\Box U = \{\mathcal{F} \in F^\circ X \colon \ U \in \mathcal{F}\}$$

We have $\Box(U \cap V) = (\Box U) \cap (\Box V)$, so we define $F^\circ X$ as a topological space by choosing $\{\Box U \colon \ U \in \Omega(X)\}$ as a base for the open sets. To see that this is $T0$, let $\mathcal{F} \neq \mathcal{G} \in F^\circ X$ so that there exists (say) $U \in \mathcal{F}$ with $U \notin \mathcal{G}$. Then $\mathcal{F} \in \Box U$ so that $\Box U$ is a neighborhood of $\mathcal{F}$. Suppose that $\Box U$ were also a neighborhood of $\mathcal{G}$. Then some $V \in \mathcal{G}$ exists with $\Box V \subset \Box U$. But then $U \in \mathcal{G}$, the desired contradiction.

The continuity of η_X is clear because it is immediate that

$$\eta_X^{-1}(\Box V) = \{x \in X\colon\ V \in prin(x)\} = V$$

Now let $\alpha : X \to F^\circ Y$ be continuous. Notice that the definition of $\alpha^\#$ is equivalently written

$$\alpha^\#(\mathcal{F}) = \{B \in \Omega(Y)\colon\ \mathcal{F} \in \Box(\alpha^{-1}(\Box B))\}$$

Thus, for $V \in \Omega(Y)$,

$$\begin{aligned}(\alpha^\#)^{-1}(\Box V) &= \{\mathcal{F} \in F^\circ X\colon\ \alpha^\#(\mathcal{F}) \in \Box V\} = \{\mathcal{F} \in F^\circ X\colon\ V \in \alpha^\#(\mathcal{F})\}\\ &= \{\mathcal{F} \in F^\circ X\colon\ \mathcal{F} \in \Box(\alpha^{-1}(\Box V))\}\\ &= \Box\{\alpha^{-1}(\Box V)\}\end{aligned}$$

so $\alpha^\#$ is also continuous.

We must now show $\mathbf{F}^\circ$ is Kock–Zöberlein. Toward that end, we first observe that the specialization order on $F^\circ X$ is subset inclusion:

$$\begin{aligned}\mathcal{N}_{\mathcal{F}} \subset \mathcal{N}_{\mathcal{G}} \quad &\Leftrightarrow \quad \forall_{U \in \Omega(X)} \mathcal{F} \in \Box U \ \Rightarrow\ \mathcal{G} \in \Box U\\ &\Leftrightarrow \quad \forall U \in \Omega(X)\ U \in \mathcal{F} \ \Rightarrow\ U \in \mathcal{G}\\ &\Leftrightarrow \quad \mathcal{F} \subset \mathcal{G}\end{aligned}$$

That done, we show that $f \leqslant g \Leftrightarrow \forall V \in \Omega Y f^{-1}V \subset g^{-1}V$, as follows. If $f \leqslant g$ and $V \in \mathcal{N}_{fx}$ then $V \in \mathcal{N}_{gx}$ so $gx \in V$. Conversely, let $N \in \mathcal{N}_{fx}$. There exists $x \in U \in \Omega(X)$ with $f(U) \subset N^\circ$, the interior of N. By hypothesis, $g(U) \subset N^\circ$ so $N \in \mathcal{N}_{gx}$ and $f \leqslant g$.

Given $f, g : X \to Y$ continuous with $f \leqslant g$ and $\mathcal{F} \in F^\circ X$, we must show $(F^\circ f)\mathcal{F} \subset (F^\circ g)\mathcal{F}$. Routinely, $(F^\circ h)\mathcal{L} = \{V \in \Omega(Y)\colon\ h^{-1}V \in \mathcal{L}\}$. Thus if $V \in (F^\circ f)\mathcal{F}$, $f^{-1}(V) \in \mathcal{F}$. As $f \leqslant g$, $g^{-1}(V) \supset f^{-1}(V)$ so $g^{-1}(V) \in \mathcal{F}$ and $V \in (F^\circ g)\mathcal{F}$.

Finally, it must be shown that $\eta_{F^\circ X}(\mathcal{F}) \subset F^\circ \eta_X(\mathcal{F})$. We have

$$\begin{aligned}\eta_{F^\circ X}(\mathcal{F}) &= prin(\mathcal{F}) = \{\overline{U} \in \Omega(F^\circ X)\colon\ \mathcal{F} \in \overline{U}\}\\ F^\circ \eta_X(\mathcal{F}) &= \{\overline{U} \in \Omega(F^\circ X)\colon\ \eta_X^{-1}(\overline{U}) \in \mathcal{F}\}\end{aligned}$$

Let $\overline{U} \in \eta_{F^\circ X}(\mathcal{F})$. As $\overline{U} \in \Omega F^\circ(X)$, there exists $U \in \mathcal{F}$ with $\Box U \subset \overline{U}$. As $U = \eta_X^{-1}(\Box U)$, $\Box U \in F^\circ \eta_X(\mathcal{F})$. As $F^\circ \eta_X(\mathcal{F})$ is a filter, $U \in F^\circ \eta_X(\mathcal{F})$. All details in the verification that $\boldsymbol{F^\circ}$ is a Kock–Zöberlein monad in **T0** have now been completed.

We return now to a general Kock–Zöberlein monad $\mathbf{T}$ in an order-enriched category $(\mathcal{K}, \leqslant)$. The next proposition shows that $K^{\mathbf{T}}$ is a subcategory of $\mathcal{K}$.

PROPOSITION 4.10. *For* $\mathbf{T}$ *Kock–Zöberlein in an order-enriched* $(\mathcal{K}, \leqslant)$ *and* X *an object of* $\mathcal{K}$*, there is at most one* $\xi : TX \to X$ *with* $\xi\eta_X = id_X$ *and such that* (X, ξ) *is a* $\mathbf{T}$*-algebra.*

PROOF. Let $\xi\eta_X = id_X$. Using the naturality of η and (30),

$$\eta_X\xi = (T\xi)\eta_{TX} \leqslant (T\xi)(T\eta_X) = T(\xi\eta_X) = id_{TX} \tag{31}$$

If also $\theta\eta_X = id_{TX}$,

$$\xi = \xi id_{TX} \geqslant \xi\eta_X\theta = \theta$$

and $\theta \geqslant \xi$ similarly, so $\xi = \theta$. To see that (X, ξ) is a **T**-algebra, use (31) three times and (30):

$$\begin{aligned}
\xi(T\xi) &= \xi(T\xi)id_{TTX} \leqslant \xi(T\xi)\eta_{TX}\eta_X\xi\mu_X \\
&\leqslant \xi(T\xi)(T\eta_X)\eta_X\xi\mu_X = \xi\mu_X \\
&= \xi\mu_X id_{TXX} \leqslant \xi\mu_X(T\eta_X)(T\xi) \\
&= \xi(T\xi)
\end{aligned}$$

so that $\xi(T\xi) = \xi\mu_X$. □

[14, Theorem 4.4] (see also [23, 6.2]) proved that the category of $\boldsymbol{F}^\circ$-algebras is, once again, the category of continuous lattices and morphisms which preserve infima and directed suprema. The objects are considered as $T0$ spaces via the Scott topology and the morphisms are provably Scott-continuous.

We put the cap on the previous proposition by identifying the subcategory $K^{\mathbf{T}}$ of $\mathcal{K}$ in intrinsic terms as the injective objects. Recall that if $\mathcal{K}$ is a category and $\mathcal{M}$ is any class of $\mathcal{K}$-morphisms, an object I is *$\mathcal{F}$-injective* if given

$$\begin{tikzcd}
A \arrow[r, "m"] \arrow[dr, "f"'] & X \arrow[d, "g"] \\
& I
\end{tikzcd}$$

with m in $\mathcal{M}$ and f as shown then there exists at least one g with $gm = f$. For example, if $\mathcal{M}$ is the class of closed subspaces in the category of topological spaces and continuous maps, the Tietze extension theorem states that the unit interval is injective. Again, in the category of Abelian groups, the injectives with respect to subgroups are the divisible groups.

THEOREM 4.11. *Let $(\mathcal{K}, \leqslant)$ be an order-enriched category with Kock–Zöberlein monad* **T**. *Let $\mathcal{F}$ be the class of T-*embeddings*, which are those $m : A \to X$ for which there exists $\hat{m} : TX \to TA$ such that $\hat{m}(Tm) = id_{TA}$, $id_{TX} \leqslant (Tm)\hat{m}$. Then the* **T**-*algebras are precisely the $\mathcal{F}$-injectives.*

PROOF. For any X, $\mu_X(T\eta_X) = id_{TX}$ and $(T\eta_X)\mu_X \geqslant id_{TX}$ using (30) and the proof of the previous proposition. Thus η_X is a T-embedding. Hence for injective X there exists ξ with $\xi\eta_X = id_X$ so (X, ξ) is a **T**-algebra. Conversely, let X be a **T**-algebra and consider

$$X \xleftarrow{f} A \xrightarrow{m} B$$

with m a T-embedding. Consider the diagram

$$\begin{tikzcd} A \arrow[rr, "f"] \arrow[d, "m"'] & & X \\ B \arrow[r, "\eta_B"'] & TB \arrow[r, "\hat{m}"'] & TA \arrow[u, "f^{\#}"'] \end{tikzcd}$$

Then

$$f^{\#}\hat{m}\eta_B m = f^{\#}\hat{m}(Tm)\eta_A = f^{\#}\eta_A = f$$

so X is $\mathcal{M}$-injective. □

In the founding paper [98, Theorem 2.12], Scott proved that the continuous lattices are precisely the subspace-injective $T0$ spaces. We know from Day's work that continuous lattices are also the $\boldsymbol{F}^{\circ}$-algebras so, following the proof in [23], we recover Scott's theorem from Theorem 4.11 by showing that the F°-embeddings coincide with subspaces. The advantage of this approach is that it suggests a method to develop further examples, as is done in the paper of Escardó just cited. For example, he proves that "continuous Scott domains" are the dense-subspace-injectives in **T0**.

Notice, by the way, that the $\mathbf{F}^{\circ}$-algebras are not a full subcategory of **T0**. Morphisms that preserve infima and directed suprema are Scott continuous and morphisms that are Scott continuous preserve directed suprema, but do not generally preserve infima. As an example, consider the Sierpinski space $\{True, False\}$ which arises as the continuous lattice $False < True$ with $\{True\}$ as the only nontrivial Scott open set. If X is a continuous lattice, the Scott open sets correspond to the S-valued continuous maps. But only those Scott open sets which are closed under infima have infimum-preserving characteristic functions.

Let $f : X \to Y$ in **T0**. We wish to show that f is a subspace (i.e., a monic such that $\Omega(X) = \{f^{-1}(V)\colon\ V \in \Omega(Y)\}$) if and only if f is an F°-embedding.

LEMMA 4.12. *For any $f : X \to Y$ in an order-enriched category, there exists at most one $g : Y \to X$ with $gf \leqslant id_X$, $id_Y \leqslant fg$*

PROOF. If also $hf \leqslant id_X$, $id_Y \leqslant fh$ then

$$g = g\, id_Y \leqslant g(fh) = (gf)h \leqslant id_X h = h$$

and $h \leqslant g$ similarly. □

Consider an arbitrary continuous map $f : X \to Y$ of $T0$ spaces. For $U \in \Omega(X)$ define

$$\widehat{U} = \{y \in Y\colon\ \exists y \in V \in \Omega(Y) \text{ with } f^{-1}(V) \subset U\}$$

and then define $g : F^\circ Y \to F^\circ X$ by

$$g(\mathcal{G}) = \{U \in \Omega(X)\colon\ \widehat{U} \in \mathcal{G}\}$$

As $U \mapsto \widehat{U}$ preserves intersection, $g(\mathcal{G})$ is a filter. As $g^{-1}(\Box U) = \Box\widehat{U}$, g is continuous. We have

$$\begin{aligned} g(F^\circ f)\mathcal{F} &= g\{B \subset Y\colon\ f^{-1}(B) \in \mathcal{F}\} \\ &= \{U \in \Omega(X)\colon\ f^{-1}(\widehat{U}) \in \mathcal{F}\} \\ &\subset \mathcal{F} \quad (\text{as } f^{-1}(\widehat{U}) \subset U) \end{aligned}$$

and also

$$\begin{aligned} (F^\circ f)g\mathcal{G} &= (F^\circ f)\{U \in \Omega(X)\colon\ \widehat{U} \in \mathcal{G}\} \\ &= \{W \in \Omega(Y)\colon\ \{y \in Y\colon\ \exists y \in V \in \Omega(Y) \\ &\qquad\qquad \text{with } f^{-1}(V) \subset f^{-1}(W)\} \in \mathcal{G}\} \\ &\subset \mathcal{G} \end{aligned}$$

By the previous lemma, g is unique with these properties. To be an F°-embedding we must also have $\mathcal{F} \subset g(F^\circ f)\mathcal{F}$. Hence

$$f \text{ is an } F^\circ\text{-embedding} \quad \Leftrightarrow \quad U \in \mathcal{F} \Rightarrow f^{-1}(\widehat{U}) \in \mathcal{F}$$

If f is a subspace and $U \in \mathcal{F} \in F^\circ X$, $\exists V \in \Omega(Y)$ with $f^{-1}(V)$ (which is essentially $X \cap V) = U$. It follows that $f^{-1}(\widehat{U}) = U$, so f is an F°-embedding. Conversely, let f be an F°-embedding. Let $x_1 \neq x_2 \in X$ so that there exists (say) $U \in \Omega(X)$ with $x_1 \in U$, $x_2 \notin U$. If $\mathcal{F} = \mathcal{N}_x$, $f^{-1}(U) \in \mathcal{F}$ so $\exists V \in \Omega(Y)$ with $x_1 \in f^{-1}(V) \subset U$. As $fx_1 \in V$, $fx_2 \notin V$, f is a monic function. Finally, let $U \in \Omega(X)$ and let $\mathcal{F} = \{W \in \Omega(X)\colon\ U \subset W\}$. Then $f^{-1}(\widehat{U}) \in \mathcal{F}$ so $U \subset f^{-1}(\widehat{U}) \subset U$ and there exists $V \in \Omega(Y)$ with $f^{-1}(V) = U$. Thus X has the subspace topology.

4.3. *Functional programming*

In [84–86], Moggi took an arbitrary category (of "types") as a framework for denotational semantics and introduced monads in that category for "notions of computation" with the paradigm

Ta: the type of computations of type a

η: values are computations

$\alpha^\#$: evaluate the computation and apply α to the resulting value

(The idea that values and computations of values are different types originates with [96].) Though anticipated by Elgot's work, what is new here is the explicit inclusion of T, η, $(-)^{\#}$ in the denotational syntax. Moggi emphasized theoretical issues concerned with reasoning about denotations of programs. Almost immediately in [109], however, Wadler proposed that monads could be implemented as datatypes in a practical way and this led to the monad class in the lazy functional language Haskell. (This language is named after Haskell Curry because of its close connections to the λ-calculus.) The monad class declaration in the Haskell prelude begins

$$\begin{aligned}&\textit{Class Monad } T\\ &\textit{return} :: a \to Ta\\ &\textit{bind} :: (a \to Tb) \to Ta \to Tb\end{aligned}$$

Here, *return* x is the computation that simply returns the value x whereas $\textit{bind}\,\alpha\omega$ is so named because it is implemented as "let $x \Leftarrow \omega$ *in* αx". In more traditional mathematical notation, think of a, b as arbitrary sets. Then *return* is a function $\eta_a : a \to Ta$ and $\textit{bind}\,\alpha\omega = \alpha^{\#}(\omega)$. Here $a \to b \to c$ is parsed $a \to (b \to c)$ which corresponds, by uncurrying, to $(a \times b) \to c$.

Though the denotational syntax – the first order theory of categories with T, η, $(-)^{\#}$ – is adequate to state the monad axioms, it is up to the programmer to check that these and other needed laws hold for a candidate monad. The type inference engine cannot do this, in essence, because testing for equality of two computable functions is undecidable in general.

In pure functional programming, every computation is a function acting on input arguments. Forbidden are "side effects" such as global variables with assigned values. (In large programs, difficult-to-find errors often result from the presence of different global variables with the same name – this cannot happen if such variables are not allowed!). Monads of sets are used to interpret the functions in $\mathcal{S}_{\mathbf{T}}$ for various $\mathbf{T}$ throughout the program. The idea is to update code by changing only the monad rather than restructuring the code. For example, if a numerical algorithm uses division it is a good idea to handle an attempt to divide by 0. A set E of error handlers defines a monad by

$$\begin{aligned}&TX = X + E\\ &\eta_X = in_X, \text{ the coproduct injection}\\ &\alpha^{\#}(\omega) = \begin{cases}\alpha(\omega) & \text{if } \omega \in X,\\ \omega & \text{if } \omega \in E\end{cases}\end{aligned} \tag{32}$$

Declaring this monad instead of no monad (i.e., the identity monad) allows error handling without needing to alter the innards of the original code.

[108] used the monad of Example 4.2 to design a simpler algorithm for array update than had been proposed up to that time. While Wadler admitted that his algorithm could be written without monads, it was the monad style of thinking that led to its discovery.

[94] introduced "CPS transforms" to modify the λ-calculus to allow a "continuation passing style" of programming. Moggi argued that it was more natural to leave the λ-calculus alone and use the double-dualization monad of 3.13 to do CPS instead.

These remarks only scratch the surface. Much has been written about the use of monads in functional programming. We refer the reader to the books [8,39,91] and to [25,27,46,59, 49,60,83,87,88,92,104,106,107,110].

Space does not permit coverage of such wealth of material. As such, we devote the balance of this section to the use of monads for collection data types, following the monad paradigm

$$\begin{aligned} &TX\colon \text{ collections of elements of } X \\ &\eta_X x\colon \text{ singletons are collections} \\ &\alpha^{\#}\colon \text{ collection union} \end{aligned} \tag{33}$$

The foremost collection data type of functional programming (important in programming generally) is the list. A *list* with members in X is an n-tuple (usually written $x_1 \cdots x_n$ rather than $(x_1, \ldots, x_n)$) with each x_i in X. Here, n is the *length* of the list. The unique list with $n = 0$ is not excluded and is called the *empty list*, written Λ. The set of all lists on X is denoted $X^\star$.

EXAMPLE 4.13. The *list monad*, $\mathbf{list} = (list, \eta, (-)^{\#})$, is defined by

$$\begin{aligned} &list\ X = X^\star \\ &\eta_X x = x \quad \text{(the length-1 list)} \\ &\alpha^{\#}(x_1 \cdots x_n) = \alpha(x_1) \cdots \alpha(x_n) \end{aligned}$$

Here, the juxtaposition in the definition of $\alpha^{\#}$ is called *list concatenation*. For example, the concatenation of cat, dog, cowbird (considered as lists of characters) is the list catdogcowbird. An explicit symbol for binary concatenation is $++$ (two plus signs). It is clear that $(X^\star, ++, \Lambda)$ is a monoid, indeed the free monoid generated by X. Verification of the monad axioms is routine. Notice that $SX = X^+$ is a submonad of $\mathbf{T}$ and is exactly the monad of the adjointness discussed in Example 2.9. $\mathbf{T}$-algebras are monoids, $\mathbf{S}$-algebras are semigroups. In both cases, the structure map $X^\star \to X$, $X^+ \to X$ realizes formal multiplication as the monoid or semigroup multiplication.

Using Haskell names but standard mathematical notations we now indicate a few of the functional programming list-processing tools. We shall start with some operators which can be defined for any monad $(T, \eta, (-)^{\#})$ of sets.

EXAMPLE 4.14. List concatenation is just $\mu_X : TTX \to TX$ when $\mathbf{T}$ is the list monad. Thus $+ + X = \mu_X$.

EXAMPLE 4.15. Mapping a function over a list is a standard operator, e.g., map $f x_1 \cdots x_n$ is $(f x_1) \cdots (f x_n)$. The map operator $map : [X \to Y] \to [X^\star \to Y^\star]$ is just $f \mapsto Tf$ when $\mathbf{T}$ is the list monad.

EXAMPLE 4.16. $unzip : (X \times Y)^\star \to X^\star \times Y^\star$ defined by

$$unzip(x_1, y_1) \cdots (x_n, y_n) = (x_1 \cdots x_n, y_1 \cdots y_n)$$

is definable in terms of map by mapping the two projection functions $X \leftarrow X \times Y \rightarrow Y$ over the input list. Thus, for any monad,

$$unzip\colon T(X \times Y) \longrightarrow TX \times TY, \quad \omega \mapsto \big(Tpr_X(\omega), Tpr_Y(\omega)\big)$$

We know that the list monad has monoids as algebras which, in turn, has standard equational presentation in terms of a binary operation xy and unit e with equations

$$\begin{aligned} x(yz) &= (xy)z \\ xe &= x \\ ex &= x \end{aligned}$$

By the Birkhoff variety theorem 3.9, adding further equations will produce a quotient monad $\theta\colon \mathbf{list} \to \mathbf{S}$, so that the elements of SX are equivalence classes of lists. If we are careful to impose only *balanced* equations which have the same set of variables in both terms, then (noting that the original three monoid equations are balanced) any derivation showing that two lists are in the same equivalence class by repeated equation substitution will never change the set of variables. Thus the set *mem*(ω) of *members* of $\omega \in SX$ is well defined by

$$mem(\omega) = \{x_1, \ldots, x_n\} \quad \text{if } \theta(x_1 \cdots x_n) = \omega \tag{34}$$

In short, such an **S** is a collection monad. (A formal definition of this concept appears below.)

By adding the balanced equation $xy = yx$, SX is the set of *finite bags* of X, that is, "finite sets in which repetition counts". If two fruits of the same type are considered indistinguishable, a physical bag containing apples, pears and oranges is a good example of a bag. Two lists $v, w \in X^\star$ represent the same bag if and only if there exists a bijection $\sigma : X \to X$ with $\sigma^\star v = w$. Notice that SX is the free Abelian monoid generated by X and that **S**-algebras are Abelian monoids.

Starting with bags and adding the further balanced equation $xx = x$ yields the finite subset submonad $\mathbf{P}_{\mathit{fin}}$ of the power set monad of Example 2.16. A collection in $\mathbf{P}_{\mathit{fin}}X$ is a finite subset of X, that is, a list in which neither order nor repetition counts. The algebras are sup-semilattices (or inf-semilattices, which is isomorphic).

Let us summarize these two new monads.

EXAMPLE 4.17. The bags monad, $\mathbf{bag} = (bag, \eta, (-)^\#)$, is defined as follows. Write a bag on X as the formal sum $\sum_{x \in X} n_x x$ where $n_x \in \mathbb{N}$ is the number of occurrences of x.

$$\begin{aligned} bagX &= \left\{ \sum_{x \in X} n_x x\colon \{x\colon n_x > 0\} \text{ is finite} \right\} \\ \eta_X x &= \sum_{y \in X} \delta_x^y y \quad \text{(Kronecker delta)} \end{aligned}$$

$$\alpha^{\#}\left(\sum_{x\in X} n_x x\right) = \sum_{y\in Y} m_y y, \quad m_y = \sum_{x\in X} n_x m_{xy} \quad \text{if } \alpha(x) = \sum_{y\in Y} m_{xy} y$$

$$(bag\ f)\left(\sum_{x\in X} n_x x\right) = \sum_{y\in Y} m_y y, \quad m_y = \sum_{fx=y} n_x$$

EXAMPLE 4.18. The finite subsets monad, $\mathbf{P}_{fin} = (P_{fin}, \eta, (-)^{\#})$, is defined by

$$\begin{aligned}
&P_{fin} X = \{F \subset X\colon\ F \text{ is finite}\}\\
&\eta_X x = \{x\}\\
&\alpha^{\#}(F) = \bigcup_{x\in F} \alpha(x)\\
&(P_{fin} f)(F) = f(F) = \bigl\{f(x)\colon\ x \in F\bigr\}
\end{aligned}$$

We next set out to define the "shape" of a collection. The first paper to study shape in the current context (but for functors, not monads) is [44]. Jay's basic idea is that a term such as $\tau\omega x\nu y x z$ for $\tau \in \Sigma_3,\ \omega \in \Sigma_1,\ \nu \in \Sigma_2$ has shape $\tau\omega \bullet \nu \bullet\bullet\bullet$ and list of data $xyxz$. Reconstructing a collection from its shape and data cannot always be done (see [79, Section 7]). The concept of shape, however, can be defined for any monad of sets.

DEFINITION 4.19. Let 1 be the one-element set $\{\bullet\}$. A $\mathbf{T}$-*shape* is an element of $T1$. Let $!\colon X \to 1$ be the unique function. The *shape map* is $shape_X = T!\colon TX \to T1$.

The definition is not surprising. We obtain the shape of a collection by substituting $\bullet$ for each collection member thereby obliterating any variation of the data and leaving only the shape.

For the list monad, *list* $1 \cong \mathbb{N}$ and shape is just length. For **bag**, *bag* $1 = \mathbb{N}$ once again and $shape_X(\sum_{x\in X} n_x x)$ is the arithmetic sum $\sum_{x\in X} n_x$. For $\mathbf{P}_{fin}$, there are only two shapes, "empty" and "nonempty".

EXAMPLE 4.20. For Σ a finitary signature as in Definition 3.1, let $\mathbf{T}_\Sigma$ be the free Σ-algebra monad induced by the adjointness of Theorem 3.2 whose algebras are the Σ-algebras. The inductive construction of $T_\Sigma X$, generalizing Example 3.4, is well-known in universal algebra (see [13] for example) and is as follows:

$$\begin{aligned}
&x \in T_\Sigma X \quad \text{for } x \in X\\
&\text{if } \tau_1, \ldots, \tau_n \in T_\Sigma X, \omega \in \Sigma_n \quad \text{then } \omega\tau_1 \cdots \tau_n \in T_\Sigma X
\end{aligned}$$

The monad structure is inductively defined by $\eta_X x = x$, $\alpha^{\#}(x) = \alpha(x)$, $\alpha^{\#}(\omega\tau_1 \cdots \tau_n) = \omega(\alpha^{\#}\tau_1)\cdots(\alpha^{\#}\tau_n)$.

We regard this as a collection monad with the leaves of the trees as collection members. "Tree concatenation" $++X$ works by regarding a tree whose leaves are trees on Y to be a tree on Y by replacing leaf x with subtree $\alpha(x)$. The *map* f operator replaces each leaf x with fx. An example of *unzip* is $unzip(\omega(x, y)(a, b)) = (\omega x a, \omega y b)$. The shape map produces the usual tree shapes, e.g., $shape_X(\omega x y) = \omega \bullet \bullet$.

An important operator on lists is $\mathit{filter}: 2^X \times X \to X^\star$ which maps (A, ω) to the list that results by deleting from ω all $x \notin A$. For example, $\mathit{filter}(\mathit{even}?, 12 \cdots 12 \cdots 9)$ would return 2468. There is a natural way to define filter for a pair $(\mathbf{T}, \perp)$ where $\mathbf{T}$ is a monad of sets and $\perp \in T\emptyset$. Define $2^X \to [X \to TX]$ mapping $A \subset X$ to $\alpha_A : X \to TX$ with $\alpha_A x = \eta_X x$ if $x \in A$, $\alpha_A x = \perp$ (via $T\emptyset \to TX$) if $x \notin A$. Then define $\mathit{filter}(A, \omega) = \alpha_A^\#(\omega)$.

For lists, $\perp$ must be chosen as the unique element Λ of $\emptyset^\star$. For $A = \{2, 4, 6, 8, 10\}$, $\omega = 12 \cdots 10$, $\alpha_A^\#(\omega)$ is the list concatenation $\Lambda 2 \Lambda 4 \cdots 8 \Lambda 10 = 246810$. For $\mathbf{T}_\Sigma$ with $\perp \in \Sigma_0$, $\mathit{filter}(\{a, b\}, \omega a \omega b c)$ is $\omega a \omega b \perp$.

The list operators

$$\mathit{reverse}: X^\star \to X^\star, \quad \mathit{reverse}(x_1 \cdots x_n) = x_n \cdots x_1$$

and

$$\begin{aligned} &\mathit{zip}: X^\star \times Y^\star \to (X \times Y)^\star, \\ &\mathit{zip}(x_1 \cdots x_m, y_1 \cdots y_n) = (x_1, y_1) \cdots (x_t, y_t), \quad t = \mathit{min}(m, n) \end{aligned}$$

seem to depend on the order of the elements in the list. It is hard to imagine how one could zip two sets, for example, given that we expect $\mathit{zip} \circ \mathit{unzip} = \mathit{id}$. Reverse and zip make sense for trees however, that is, for the monad $\mathbf{T}_\Sigma$. For example, $\mathit{reverse}(\omega ab) = \omega ba$ and $\mathit{zip}(\omega ab, \omega cd) = \omega(a, c)(b, d)$. Here, for zip, we require that both trees have the same shape. For lists, where shapes are natural numbers, one shape is always a subshape of another and this allowed a definition of zip for any two lists. For trees, there would be no way to zip, say, ωab and τcd. We shall have more to say about reverse and zip later.

Other list operators which one would hope to eventually understand for more general monads include

$$\begin{aligned} &\mathit{first}: X^\star \to X + \{\perp\}, && x_1 \cdots x_n \mapsto x_1 \quad \text{if } n > 0, \perp \text{ if } n = 0 \\ &\mathit{tail}: X^\star \to X^\star, && x_1 \cdots x_n \mapsto x_2 \cdots x_n \quad (= \Lambda \text{ if } n = 0) \\ &\mathit{drop}: \mathbb{N} \times X^\star \to X^\star, && (k, x_1 \cdots x_n) \mapsto x_{k+1} \cdots x_n \quad (= \Lambda \text{ if } k \geqslant n) \\ &\mathit{take}: \mathbb{N} \times X^\star \to X^\star, && (k, x_1 \cdots x_n) \mapsto x_1 \cdots x_t, \quad t = \mathit{min}(k, n) \end{aligned}$$

(In first, $\perp$ is any abstract new element, not necessarily one in $T\emptyset$.) An especially important list operator is *right fold*:

$$\begin{aligned} &\mathit{foldr}: [X \times Y \to Y] \times Y \times X^\star \to Y \\ &\mathit{foldr}(f, y, x_1 \cdots x_n) = y \oplus \big(x_1 \oplus (x_2 \oplus \cdots \oplus (x_{n-1} \oplus x_n) \cdots\big) \end{aligned}$$

where $x \oplus y = f(x, y)$.

We proceed now in search of an axiomatic theory of collection monads and subclasses of such. Collection properties should have an operational semantics that relates directly to the issues facing the functional programmer as well as a denotational semantics which provides a tool to establish mathematical relationships between collection subclasses. To begin, let $2 = \{True, False\}$ be the Boolean type. For $A \subset X$, consider the function

$$memTest_A : X^\star \to 2$$
$$memTest_A(x_1 \cdots x_n) = True \quad \Leftrightarrow \quad \forall_{1\leqslant i\leqslant n} x_i \in A$$

Let $\chi_A : X \to 2$ be the characteristic function of A and let $\wedge : 2 \times 2 \to 2$ be Boolean "and". Then, routinely,

$$memTest_A(w) = foldr\big(\wedge, True, (map\chi_A)\big) \tag{35}$$

For example, $memTest_A(abc) = True \wedge (\chi_A(a) \wedge (\chi_A(b) \wedge \chi_A(c)))$. This shows how to write a program to test collection membership in A given a program to test element membership in A.

What axioms are required to do the same for an arbitrary monad of sets? This question feels, at first, like one of the least important of a great many such questions. In fact, however, answering only this question has far-reaching consequences and will lead to surprising connections.

The key observation is that $(2, \wedge)$ is a monoid (in the usual way) and hence is an algebra of the list monad. Its structure map ψ is exactly

$$\psi : 2^\star \to 2, \quad \psi(w) = foldr(\wedge, True, w)$$

Thus the map of (35) can be written in monadic terms as

$$memTest_A = \psi(T\chi_A) \tag{36}$$

This construction was anticipated by [7, Lemma 1, p. 14] and categorical generalizations of this paper were studied by [103], but neither author considered algebras over a monad.

DEFINITION 4.21. Let $\mathbf{T}$ be a monad of sets. For $A \subset X$, $\omega \in TX$, A is a *support* of ω if "$\omega \in TA$", that is, ω is in the image of $TA \to TX$. For example, Definition 3.43 may be expressed by saying that $\mathbf{T}$ is bounded if some cardinal κ exists such that every $\omega \in TX$ has a support of cardinal at most κ. We write $supp(\omega) = \{A \subset X\colon\ A \text{ is a support of } \omega\}$. A *support classifier* of $\mathbf{T}$ is a $\mathbf{T}$-algebra $(2, \psi)$ such that for every set X and for every subset A of X with inclusion map $i : A \to X$, the following square is a pullback

$$\begin{tikzcd}
TA \arrow[r, "!"] \arrow[d, "Ti"'] & 1 \arrow[d, "True"] \\
TX \arrow[r, "\chi_A^\#"'] & (2, \psi)
\end{tikzcd}$$

Note that $\chi_A^\# = memTest_A$ as in (36) by (11). Thus the pullback property amounts to the statement "A is a support of ω if and only if $memTest_A(\omega) = True$".

A support classifier $(2, \psi)$ is unique when it exists. To see this, use the formula $memTest_A = \psi(T\chi_A)$ for $A = \{True\} \subset 2$. For then $\chi_A = id_2$ and the following square is a pullback:

$$\begin{tikzcd} T1 \arrow[r] \arrow[d, "T(True)"'] & 1 \arrow[d, "True"] \\ T2 \arrow[r, "\psi"'] & (2, \psi) \end{tikzcd}$$

This says that $\psi = \chi_{\mathcal{Z}}$ if $\mathcal{Z}$ is the image of $T(True) : T1 \to T2$.

For the list monad, χ_Z is indeed $x_1 \cdots x_n \mapsto x_1 \wedge \cdots \wedge x_n$ because $x_1 \wedge \cdots \wedge x_n = True \Leftrightarrow x_1 = x_2 = \cdots = x_n = True$. Thus $(2, \wedge)$ is the support classifier for the list monad.

DEFINITION 4.22. A pullback of $A \xrightarrow{f} B \xleftarrow{g} C$ with g monic is an *inverse image square* (because $f^{-1}(C)$ gives the pullback in $\mathcal{S}$). Say that a functor is *taut* if it preserves inverse image squares. Say that a natural transformation is *taut* if each naturality square induced by a monic is a pullback.

THEOREM 4.23 [81, Theorem 3.3]. *For a monad* $\mathbf{T}$ *of sets, equivalent are*:

(1) $\mathbf{T}$ *has a support classifier*;
(2) $\mathbf{T} = (T, \eta, \mu)$ *is nontrivial and* T, η *and* μ *are taut*;
(3) *There exists a taut monad map* $\mathbf{T} \to \mathbf{F}$ *to the filter monad*;
(4) $supp : \mathbf{T} \to \mathbf{F}$ *is a well-defined taut monad map.*

If any, hence all, of these conditions hold, say that $\mathbf{T}$ is a *taut* monad. It is known [81, Proposition 3.12] that every submonad of the filter monad is taut and has taut inclusion functor.

COROLLARY 4.24. *If* L *is a continuous lattice and* $\mathbf{T}$ *is a taut monad then* (L, ξ) *is a* $\mathbf{T}$-*algebra if*

$$\xi(\omega) = \bigvee_{A \in supp(\omega)} \bigwedge A$$

PROOF. Use Theorem 3.39 and (26). □

It it not hard to see that if $L = 2$ with $False < True$, then the corresponding $\mathbf{T}$-algebra is the support classifier of $\mathbf{T}$.

We can now define

DEFINITION 4.25. A monad $\mathbf{T}$ of sets is a *collection monad* if it is taut and if for every $\omega \in TX$, $supp(\omega)$ is a principal filter. The generator of this filter is called the set of *members* of ω, written $mem(\omega)$. Thus $A \in supp(\omega) \Leftrightarrow mem(\omega) \subset A$.

The following characterization, essentially that of [79, Definition 4.1] without the finitary restriction, provides an operational definition for collection monads, in contrast to the denotational definitions so far. It also explains why we called $\alpha^\#$ "collection union" in the collection paradigm.

THEOREM 4.26. *Let* **T** *be a monad of sets. Then* **T** *is a collection monad if and only if*

(1) **T** *is nontrivial*;
(2) Members exist, *that is, if* $\omega \in TX$ *there exists a minimum support* $mem(\omega)$ *of* ω;
(3) Members are collected, *that is, for* $\omega \in TX$ *and* $\alpha : X \to TY$, $mem(\alpha^\#(\omega)) = \bigcup_{x \in mem(\omega)} mem(\alpha x)$.

PROOF. First, let **T** be a collection monad. Any taut monad satisfies (1) and any collection monad satisfies (2) by definition. Regard the power set monad **P** as a submonad of the filter monad **F** via $prin : \mathbf{P} \to \mathbf{F}$ so that

$$supp = \mathbf{T} \xrightarrow{mem} \mathbf{P} \xrightarrow{prin} \mathbf{F}$$

Thus $mem : \mathbf{T} \to \mathbf{P}$ is a monad map because *prin* is a submonad and *supp* is a monad map. Hence, for $\alpha : X \to TY$ we have the commutative square

$$\begin{tikzcd} TX \arrow[r, "mem_X"] \arrow[d, "\alpha^\#"'] \arrow[dr, phantom, "\text{(A)}"] & PX \arrow[d, "(mem_Y \alpha)^\#"] \\ TY \arrow[r, "mem_Y"'] & PY \end{tikzcd}$$

which immediately gives (3). Conversely, assume (1, 2, 3). (3) implies that (A) commutes, as just discussed. Let $\hat{x} = \eta_X(x)$. Clearly $mem(\hat{x}) \subset \{x\}$. To see that $mem(\hat{x}) = \{x\}$ (so that $mem : \mathbf{T} \to \mathbf{P}$ is a monad map) we must show $\hat{x} \notin T\emptyset$. Suppose otherwise, so that there exists $\omega_0 \in T\emptyset$ with $(Te)\omega_0 = \hat{x}$ for $e : \emptyset \to X$ the empty inclusion. For any function $f : X \to X$, $fe = e$. By naturality of η,

$$\begin{aligned} \eta_X(fx) &= (Tf)\hat{x} = (Tf)\big((Te)\omega_0\big) \\ &= \big(T(fe)\big)\omega_0 = (Te)\omega_0 \\ &= \hat{x} = \eta_X(x) \end{aligned}$$

By (1) and Proposition 2.25, $fx = x$. But the arguments above hold replacing X with any superset of $\{x\}$ so it may be assumed without loss of generality that X has two or more elements at the same time that every endomorphism of X is the identity. This is contradictory, so $\hat{x} \in T\emptyset$ is impossible and $mem : \mathbf{T} \to \mathbf{P}$ is a monad map. It is in fact a taut monad map as follows. If $A \subset X$ with inclusion map i, diagram (A) above becomes

$$\begin{array}{ccc} TA & \xrightarrow{mem_A} & PA \\ {\scriptstyle Ti}\downarrow & & \downarrow{\scriptstyle Pi} \\ TX & \xrightarrow[mem_X]{} & PX \end{array}$$

This diagram is a pullback because $mem_X(\omega) \in A \Rightarrow \omega \in TA$. Thus *mem* is a taut monad map. It is directly seen (without invoking the fact, mentioned above, that this is true for all submonads of $\mathbf{F}$) that $prin : \mathbf{P} \to \mathbf{F}$ is taut. By Theorem 4.23, we are done. □

COROLLARY 4.27. *For any complete inf-semilattice L and collection monad $\mathbf{T}$, if ξ is defined by*

$$\xi(\omega) = \bigwedge mem(\omega)$$

then (L, ξ) is a $\mathbf{T}$-algebra.

Equational presentations are a useful tool to refine collection notions. For instance, if we start with the signature $\Sigma_0 = \{\Lambda\}$, $\Sigma_2 = \{\star\}$ then the induced monad $\mathbf{T}_\Sigma$ is a collection monad for which the members of each Σ-tree are its leaves. For example,

$$mem(\star x \star \Lambda z) = \{x, z\}$$

We can refine this collection notion by imposing the equations

$$\begin{aligned} x \star (y \star z) &= (x \star y) \star z \\ x \star \Lambda &= x \\ \Lambda \star x &= x \end{aligned}$$

giving rise to the quotient monad $TX = X^*$ of lists, including the empty list $[\Lambda]$ (using $[-]$ for equivalence class under the equations). Adding the further equation

$$x \star y = y \star x$$

removes the order from the lists, resulting in bags. Adding yet another equation

$$x \star x = x$$

also removes repetitions, giving sets (that is, the power set monad $\mathbf{P}$). Of key importance is that all five equations above are *balanced*, that is, we equate only pairs of trees with the same set of leaves. By applying such equations successively, the set of members never changes so, for the resulting quotient monad $\theta : \mathbf{T}_\Sigma \to \mathbf{T}$, the following diagram commutes:

$$\begin{tikzcd} \mathbf{T}_\Sigma \arrow[rr, "\theta"] \arrow[dr, "mem^\Sigma"'] & & \mathbf{T} \arrow[dl, "mem"] \\ & \mathbf{P} & \end{tikzcd}$$

The following theorem provides a formal characterization, showing that all collection monads arise by this equational process.

THEOREM 4.28. *Let* $\mathbf{T}$ *be a monad of sets. The following three conditions are equivalent.*

(1) $\mathbf{T}$ *is a collection monad.*
(2) *There exists a signature* Σ *and a taut monad quotient* $\theta : \mathbf{T}_\Sigma \to \mathbf{T}$.
(3) *There exists a tractable equational presentation* (Σ, E) *with all equations in* E *balanced such that* $\mathbf{T}$ *is isomorphic to the monad* $\mathbf{T}_{\Sigma,E}$ *for* (Σ, E)*-algebras.*

PROOF. $(1 \Rightarrow 2)$ For n a cardinal, define $\Sigma_n = \{\omega \in Tn\colon\ mem(\omega) = n\}$, noting that this may be empty. The proof of [79, Theorem 4.14] works without assuming that $\mathbf{T}$ is finitary.

$(2 \Rightarrow 3)$ By the construction of 3.41 specialized to $\mathcal{K} = \mathcal{S}$, $\mathcal{E} =$ surjective, $\mathcal{M} =$ monic, $\mathcal{S}^{\mathbf{T}}$ is a Birkhoff subcategory of $\mathcal{S}^{\mathbf{T}}$, so has form (Σ, E)-*Alg*. By the proof of Birkhoff's Theorem 3.9, $(\omega, \nu) \in E \Leftrightarrow \exists X \theta_X \omega = \theta_X \nu$. (The same proof works for unbounded signatures.) If $\theta_X \omega = \theta_X \nu$, $A \subset X$ and $\theta_X \nu \in TA$, then $\theta_X \omega \in TA$ as well because θ is taut. This shows that E is balanced.

$(3 \Rightarrow 1)$ $\mathbf{T}_\Sigma$ is a collection monad. As E is balanced, there exist functions $mem_X : T_{\Sigma,E} X \to PX$ such that $mem_X \theta_X = mem_X^\Sigma$. It follows from elementary category theory (such as [79, Lemma 1.2(3)]) that *mem* is a taut monad map because mem^Σ is a taut monad map and θ is a monad map. The proof is complete. □

A collection monad is *ordered* if "changing the order of the members always changes the collection". For example, the list monad is ordered because under the permutation $x \mapsto y$, $y \mapsto x$, $z \mapsto z$, $xyxyyzyz$ is different from $yxyxxzxz$. The sets monad $\mathbf{P}$ is not ordered because $\{x, y, z\} = \{y, x, z\}$. A collection monad is *zipped* if unzip has an inverse. Appropriate precise definitions are given below. To set them up, we make some observations.

Firstly, if $\sigma \in X!$ (the set of bijections $X \to X$) and Σ is a signature, $T\sigma : T_\Sigma X \to T_\Sigma X$ indeed behaves like *map* sending, e.g., $\rho x \rho y \tau x$ (where $\rho \in \Sigma_2$, $\tau \in \Sigma_1$) to $\rho(\sigma x)\rho(\sigma y)\tau(\sigma x)$. This works as well for equivalence classes under a balanced presentation, e.g., $[\rho x \rho y \tau x] \mapsto [\rho(\sigma x)\rho(\sigma y)\tau(\sigma x)]$. Hence, to "change the order of the members of collection $\omega \in TX$", choose $\sigma \in X!$ and apply $T\sigma$ to ω.

Secondly, if we form the pullback $TY \leftarrow ss(X, Y) \to TX$ of the shape maps $TX \xrightarrow{T!} T1 \xleftarrow{T!} TY$ ("*ss*" is for "same shape"), then

$$ss(X, Y) = \big\{(\omega, \nu) \in TX \times TY\colon\ \omega, \nu \text{ have the same shape}\big\}$$

Since the diagram

$$\begin{tikzcd} X \times Y \arrow[r, "pr_y"] \arrow[d, "pr_X"'] & Y \arrow[d, "!"] \\ X \arrow[r] & 1 \end{tikzcd}$$

commutes (indeed, is a pullback), it follows from Definition 4.16 of $unzip : T(X \times Y) \to TX \times TY$, that *unzip* factors through $ss(X, Y)$.

Here are the formal definitions.

DEFINITION 4.29. A collection monad is *ordered* if given $\sigma \in X!$, $\omega \in TX$, if $(T\sigma)\omega = \omega$ then $\forall x \in mem(\omega)\ \sigma x = x$.

DEFINITION 4.30. A collection monad $\mathbf{T}$ is *zipped* if, given sets X, Y, the *unzip* map considered as a function $unzip : T(X \times Y) \to ss(X, Y)$ as discussed above, is bijective.

Of course, the inverse to *unzip* is then the *zip map* $zip : ss(X, Y) \to T(X \times Y)$. It is clear that both definitions are properties solely of the functor T.

We next give denotational equivalents.

THEOREM 4.31 [79, Theorem 6.10]. *For a collection monad* $\mathbf{T}$*, the following three conditions are equivalent.*

(1) $\mathbf{T}$ *is ordered.*

(2) *There exists a family* $\gamma_X : TX \to X^*$ *such that*

$$\begin{tikzcd} TA \arrow[r, "\gamma_A"] \arrow[d, "Ti"'] & A^* \arrow[d, "i^*"] \\ TX \arrow[r, "\gamma_X"'] & X^* \end{tikzcd}$$

is a pullback whenever $i : A \to X$ *is monic.*

(3) T *preserves equalizers of pairs of monics.*

PROOF. The cited proof carries over mutatis mutandis, provided finite cardinals are replaced by arbitrary ones. The proof that the third condition implies the first depends heavily on the axiom of choice, even in the finitary case. □

THEOREM 4.32 [79, Proposition 7.10]. *For a collection monad* $\mathbf{T}$, $\mathbf{T}$ *is zipped if and only if* T *preserves pullbacks.*

By elementary category theory, any functor defined on a category with binary powers which preserves pullbacks, also preserves equalizers. ($i : E \to X$ is the equalizer of $f, g : X \to Y$ if and only if

$$\begin{tikzcd} E \arrow[r, "fi=gi"] \arrow[d, "i"'] & Y \arrow[d, "{[id_Y, id_Y]}"] \\ X \arrow[r, "{[f,g]}"'] & Y \times Y \end{tikzcd}$$

is a pullback). This proves

COROLLARY 4.33. *A zipped collection monad is ordered.*

The previous corollary, although intuitive, does not appear to be easy to prove directly from the operational definitions.

Is it possible to reverse the order of the members of an ordered collection monad? For this notion to make sense, a collection should have a unique list of members in a canonical way. Such happens for the *shapely monads* of [79, Section 8] but not for ordered monads in general since the construction of τ in Theorem 4.31 involves an arbitrary well-ordering, so is not in any way canonical.

What happens for the list monad? Here, $reverse_X : X^* \to X^*$ is a monad involution. In general, the pointwise equalizer of two monad maps is a submonad. Applying this construction to $reverse_X, id_X : X^* \to X^*$ gives the submonad **PAL** of palindromes. Intersecting this with the submonad **OL** of lists of odd length gives the submonad **OP** of palindromes of odd length. Call the resulting algebras, respectively, *palindrome algebras* and *odd-palindrome algebras*.

OL-algebras are "ternary semigroups", that is, sets with a ternary operation $[xyz]$ satisfying the associative law

$$\big[xy[abc]\big] = \big[x[yab]z\big] = \big[[xya]bc\big]$$

Of course, all semigroups are **OL**-algebras by Theorem 3.39. If xy is a semigroup, $[xyz] = xyz$ is the desired forgetful functor. A ternary semigroup which does not come from a semigroup is the set of odd integers with $[xyz] = x + y + z$.

Odd-palindrome algebras are equationally presented with a single binary operation $x \star y$ satisfying the single equations

$$(x \star y) \star z = x \star \big(y \star (x \star z)\big)$$

The forgetful functor from semigroups is defined by $x \star y = xyx$. In general, if (X, ξ) is an odd-palindrome algebra,

$$x \star y = \xi(xyx)$$
$$\xi(x_1 \cdots x_{n-1} x_n x_{n-1} \cdots x_1) = x_1 \star \big(x_2 \star (\cdots x_{n-2} \star (x_{n-1} \star x_n) \cdots)\big)$$

There are two odd-palindrome algebra structures on any Abelian group with at least two elements and neither comes from a semigroup:

$$x \star y = -y$$
$$x \star y = 2x - y$$

It is possible to characterize palindrome algebras as $(X, \star, d)$ with $(X, \star)$ an odd-palindrome algebra and d a unary operation, all subject to two additional equations:

$$d(x \star y) = x \star \big(y \star (dx)\big)$$
$$x \star (dx) = d(dx)$$
$$(dx) \star y = x \star (x \star y)$$

An example is any group with $x \star y = xyx$, $dx = x^{-1}$. There is also the following trivial example: If $\perp \in X$, $x \star y = x$, $dx = \perp$.

We noted in Example 2.21 that the submonad $TX = X^+$ of nonempty lists has semigroups as algebras. The maps *first* and *last* can be defined (without $\perp$) as usual by

$$\mathit{first}: X^+ \to X, \quad x_1 \cdots x_n \mapsto x_1$$
$$\mathit{last}: X^+ \to X, \quad x_1 \cdots x_n \mapsto x_n$$

These maps can be thought of as the structure maps, respectively, of the semigroups $xy = x$, $xy = y$, but there is another way to think of them, namely as monad maps to the identity monad. Algebras of the identity monad are just sets and the two corresponding forgetful functors from sets to semigroups are exactly the two constructions just given. The equalizer of the monad maps *first* and *last* gives rise to the submonad **FL** of non-empty lists consisting of all words with the same first and last symbol. An equational characterization of the **FL**-algebras is not known at this time.

Any pointwise Cartesian product of monads is a monad and, in particular, $id \times id$ is a monad in $\mathcal{S}$. Here, $\eta_X x = (x, x)$ and, for $\alpha = [\alpha_1, \alpha_2]: X \to Y \times Y$, $\alpha^{\#}(x_1, x_2) = (a, d)$ if $\alpha_1(x_1) = (a, b)$ and $\alpha_2(x_2) = (c, d)$. The algebras of this monad are a well-known variety of semigroups, namely rectangular bands, which are characterized by the "anti-palindrome identity" $xyx = x$. (The free algebra $X \times X$ has semigroup operation $(x, y)(a, b) = (x, b)$.) The monad map $X^+ \to X \times X$ corresponding to the forgetful functor from rectangular bands to semigroups is precisely $w \mapsto (\mathit{first}(w), \mathit{last}(w))$.

5. By way of conclusion

This chapter must conclude, but the proper development of monads of sets is just beginning.

A recent further characterization of taut monads is announced in [80]. (The proof will appear elsewhere.) We briefly discuss this result here. The school of the theory of program verification known as *dynamic logic* (with semantics via *dynamic algebras*) and its precursors is partially represented by the works [15,16,26,36,37,54,95] and the references there.

There are two principal operations in dynamic logic. The first is $[\alpha]Q$, whose standard semantics is the set of all beginning states such that, after program α executes, all terminating states (if any) belong to Q. The second is $\langle\alpha\rangle Q$ whose standard semantics is the set of all beginning states such that, after program α executes, there exists at least one terminating state in Q. The book [78] introduced "Boolean categories" to capture loop-free dynamic logic with four axioms in the first-order theory of categories. Such categories interface well with [11] so that a full theory with iteration can also be given in the first-order theory of categories.

The axioms on a Boolean category are easily stated and we shall now turn in that direction. Say that $i : A \to X$ is a *summand* (*of* X) if there exists $i' : A' \to X$ such that $A \xrightarrow{i} X \xleftarrow{i'} A'$ is a coproduct. If $i : A \to X$, $j : B \to X$ are summands, say that $i \leqslant j$ if there exists $k : A \to B$ with $jk = i$. Such $\leqslant$ is reflexive and transitive. We write $Summ(X)$ for the poset of $\leqslant$-antisymmetry classes of summands of X.

DEFINITION 5.1 [78, Definition 4.4]. A category is *Boolean* if it has finite coproducts (including an initial object 0), if every diagram $X \xrightarrow{\alpha} Y \xleftarrow{j} Q$ with j a summand has a pullback $X \xleftarrow{i} [\alpha]Q \xrightarrow{g} Q$ with i a summand, if every summand pulls back coproducts to coproducts and if whenever $X \xrightarrow{id} X \xleftarrow{id} X$ is a coproduct, $X \cong 0$.

In the book cited above, it is shown that, in a Boolean category, each poset $Summ(X)$ is a Boolean algebra. If $(-)'$ denotes Boolean complement, we then define $\langle\alpha\rangle Q = ([\alpha]Q')'$.

DEFINITION 5.2. A monad $\mathbf{T}$ of sets is a *standard Boolean monad* if $\mathbf{T}$ is nontrivial, if the Kleisli category $\mathcal{S}_{\mathbf{T}}$ is a Boolean category and if summands are standard, that is, if $i : A \to TX$ (qua morphism $A \to X$ in $\mathcal{S}_{\mathbf{T}}$) is a summand then there exists $B \subset X$ and an $\mathcal{S}_{\mathbf{T}}$-isomorphism $\alpha : A \to TB$ such that $(\eta_X j) \circ \alpha = i$ in $\mathcal{S}_{\mathbf{T}}$, that is, $(Tj)\alpha = i$ in $\mathcal{S}$.

Here, then, is the promised new result.

THEOREM 5.3. *A monad of sets is taut if and only if it is a standard Boolean monad*

As a consequence of this result, dynamic algebra is inherent in collection monads, a surprising connection between data structures and program verification. Ramifications await development.

We have not previously mentioned the *distributional paradigm* for monads of sets:

TX: the set of $\mathbf{T}$-distributions on X
$\eta_X x$: point-mass distribution
$\alpha^{\#}$: distributional extension

An important example is probability distributions. For this, define the sum $\sum x_i$, for (x_i) any family of real numbers, to be the limit (if it exists) of the net of finite subsums.

EXAMPLE 5.4. **Prob** $= (Prob, \eta, (-)^{\#})$ is the monad of sets defined by

$$prob(X) = \text{all formal sums } \sum_{x\in X} \lambda_x x, \quad \lambda_x \in [0,\infty), \ \sum_{x\in X} \lambda_x = 1$$

$$\eta_X x = \sum_{y\in X} \delta_x^y \quad \text{(Kronecker delta)}$$

$$\alpha^{\#}\Big(\sum_{x\in X} \lambda_x x\Big) = \sum_{y\in Y} \rho_y y, \quad \alpha x = \sum_{y\in Y} \rho_{xy} y, \quad \rho_y = \sum_{x\in X} \lambda_x \rho_{xy}$$

This monad is a collection monad, with $mem(\sum \lambda_x x) = \{x \in X\colon\ \lambda_x > 0\}$.

The distributional paradigm was introduced in [77] (see [34] for a textbook treatment). This paper also introduced the "monad comprehensions" reinvented in [108]. Most of the distributional examples are collection monads. The time is ripe to develop the interactions between these two approaches.

The use of monads in functional programming, which subsumes the study of collection monads, makes it clear that many situations arise where the effects of two or more monads are required together. If **S**, **T** are monads of sets, TS and ST are two (usually different) candidates for the "combined" monad. The papers [49,60,106] consider the problem of when TS is again a monad. The original paper on this problem is [5] where it is shown that ways of inducing a monad structure on ST subject to natural conditions are classified by certain natural transformations $\lambda : TS \to ST$ called *distributive laws*. Beck showed that the **ST**-algebras relative to a given distributive law λ are tantamount to (X, ξ, θ) where (X, ξ) is an **S**-algebra, (X, θ) is a **T**-algebra, and the following diagram commutes:

$$\begin{tikzcd} TSX \arrow[rr, "\lambda_X"] \arrow[dr, "\theta(T\xi)"'] & & STX \arrow[dl, "\xi(S\theta)"] \\ & X & \end{tikzcd}$$

Beck's centerpiece example was how rings are built from **T**-algebra = monoid, **S**-algebra = Abelian group and λ the distributive law of multiplication over addition, viz.

$$\lambda_X\big((x+y)(a+b)\big) = xa + xb + ya + yb$$

An intriguing possible example of a distributive law is the Dempster–Shafer belief function on X which amounts to an element of *Prob* P_0X [101, pp. 38–39], where P_0X is the submonad of the power set monad **P** of "finite non-empty subsets". We do not know if this is a monad at this time, but it is natural conjecture that it is, using Dempters rule of combination as a distributive law.

In conclusion, much remains to be done with monads of sets. I hope that many will contribute to their development.

References

[1] H. Appelgate and M. Tierney, *Categories with models*, Lecture Notes in Math., Vol. 80, Springer-Verlag, Berlin (1969), 156–243.

[2] M. Barr, *Relational algebras*, Lecture Notes in Math., Vol. 137, Springer-Verlag, Berlin (1970), 39–55.

[3] M. Barr and C. Wells, *Toposes, Triples and Theories*, Springer-Verlag, Berlin (1984).

[4] J. Beck, *Triples, Algebras and Cohomology*, Dissertation, Columbia University (1967).

[5] J. Beck, *Distributive laws*, Lecture Notes in Math., Vol. 80, Springer-Verlag, Berlin (1969), 119–140.

[6] J. Bénabou, *Introduction to bicategories*, Lecture Notes in Math., Vol. 47, Springer-Verlag, Berlin (1967), 1–77.

[7] R. Bird, *An introduction to the theory of lists*, Logic of Programming and Calculi of Discrete Design, M. Broy, ed., NATO ASI Series F., Vol. 36, Springer-Verlag, Berlin (1987), 3–42.

[8] R. Bird, *Introduction to Functional Programming Using Haskell*, 2nd edn, Prentice Hall Europe, London (1998).

[9] G. Birkhoff, *On the structure of abstract algebras*, Proc. Cambridge Philos. Soc. **31** (1935), 433–454.

[10] S.L. Bloom, ed., *Calvin C. Elgot, Selected Papers*, Springer-Verlag, Berlin (1982).

[11] S. Bloom and Z. Ésik, *Iteration Theories: The Equational Logic of Iterative Processes*, Springer-Verlag, Berlin (1993).

[12] H. Cartan and S. Eilenberg, *Homological Algebra*, Princeton Univ. Press, Princeton (1956).

[13] P.M. Cohn, *Universal Algebra*, Harper and Row, New York (1965).

[14] A. Day, *Filter monads, continuous lattices and closure systems*, Canadian J. Math. **XXVII** (1975), 50–59.

[15] E.W. Dijkstra, *A Discipline of Programming*, Prentice Hall, Englewood Cliffs, NJ (1976).

[16] E.W. Dijkstra and C.S. Scholten, *Predicate Calculus and Program Semantics*, Springer-Verlag, Berlin (1990).

[17] J. Duskin, *Variations on Beck's tripleability criterion*, Lecture Notes in Math., Vol. 106, Springer-Verlag, Berlin (1969), 74–129.

[18] G.A. Edgar, *The class of topological spaces is equationally definable*, Algebra Universalis **3** (1973), 139–146.

[19] S. Eilenberg and S. MacLane, *General theory of natural equivalences*, Trans. Amer. Math. Soc. **58** (1945), 231–294.

[20] S. Eilenberg and J.C. Moore, *Adjoint functors and triples*, Illinois J. Math. **9** (1965), 381–398.

[21] S. Eilenberg and J.A. Zilber, *Semi-simplicial complexes and singular homology*, Ann. of Math. **51** (1950), 499–513.

[22] C.C. Elgot, *Algebraic theory of program schemes*, Lecture Notes in Math., Vol. 188, Springer-Verlag, Berlin (1971), 71–88.

[23] M. Escardó, *Injective spaces via the filter monad*, Topology Proc. **22** (1997), 97–110.

[24] M. Escardó, *The patch topology of a stably compact space*, Northeastern Conference in Topological Methods in Computer Science, April 19, 1999, University of Connecticut, Stamford, Connecticut, USA, unpublished.

[25] A. Filinski, *Representing monads*, Proceedings of POPL'94 Symposium Principles of Programming Languages, ACM Press (1994), 446–457.

[26] M.J. Fischer and R.E. Ladner, *Propositional dynamic logic of regular programs*, J. Comput. System Sci. 18 (1979), 286–294.

[27] M. Fokkinga, *Monadic maps and folds for arbitrary datatypes*, Memoranda Informatica 94-28, University of Twente (1994).

[28] P. Freyd, *Algebra-valued functors in general and tensor products in particular*, Colloquium Math. **14** (1966), 89–106.

[29] P. Gabriel and M. Zisman, *Calculus of Fractions and Homotopy Theory*, Springer-Verlag, Berlin (1967).

[30] H. Gaifmann, *Infinite Boolean polynomials I*, Fund. Math. **54** (1964), 229–250.

[31] G. Giertz et al., *A Compendium of Continuous Lattices*, Springer-Verlag, Berlin (1980).

[32] R. Godement, *Topologie Algébrique et Théorie des Faisceaux*, Hermann, Paris (1964).

[33] H. Gonshor, *An application of nonstandard analysis to category theory*, Fund. Math. **104** (1979), 75–83.

[34] I.R. Goodman and H.T. Nguyen, *Uncertainty Models for Knowledge-Based Systems*, North-Holland, Amsterdam (1985).

[35] A.W. Hales, *On the non-existence of free complete Boolean algebras*, Fund. Math. **54**, 45–66.

[36] D. Harel, A.R. Meyer and V.R. Pratt, *Computability and completeness in logics of programs*, Proceedings of 9th ACM Symposium on the Theory of Computing (1977), 261–268.

[37] C.A.R. Hoare, *An axiomatic basis for computer programming*, Comm. ACM **12** (1969), 576–580, 583.

[38] P.J. Huber, *Homotopy theory in general categories*, Math. Ann. **144** (1961), 361–385.

[39] P. Hudak, *The Haskell School of Expression: Learning Functional Programming through Multimedia*, Cambridge University Press, New York (2000).

[40] J.R. Isbell, *Subobjects, adequacy, completeness and categories of algebras*, Rozprawy Matematyczne XXXVI, Warsaw (1964), 32 pp.

[41] J.R. Isbell, *General functorial semantics I*, Amer. J. Math. **XCIV** (1972), 535–596.

[42] J.R. Isbell, *Functorial implicit operations*, Israel J. Math. **15** (1973), 185–188.

[43] J.R. Isbell, *Generating the algebraic theory of $C(X)$*, Algebra Universalis **15** (1982), 153–155.

[44] C.B. Jay, *A semantics for shape*, Sci. Comput. Programming **25** (1995), 251–283.

[45] P.T. Johnstone, *Stone Spaces*, Cambridge University Press (1982).

[46] M.P. Jones, *Functional programming with overloading and higher-order polymorphism*, Advanced Functional Programming, J. Jeuring and E. Meijer, eds, Bästad, Sweden, Springer-Verlag, Berlin (1995), 97–136.

[47] D.M. Kan, *Adjoint functors*, Trans. Amer. Math. Soc. **87** (1958), 294–329.

[48] J.F. Kennison and D. Gildenhuys, *Equational completion, model-induced triples and pro-objects*, J. Pure Appl. Algebra **1** (1971), 317–346.

[49] D.J. King and P. Wadler, *Combining monads*, Proceedings of the Fifth Annual Glasgow Workshop on Functional Programming, Ayr, Scotland, J. Launchbury and P.M. Sansom, eds, Springer-Verlag, Berlin (1992), 134–143.

[50] S.C. Kleene, *Introduction to Metamathematics*, Van Nostrand, New York (1972).

[51] H. Kleisli, *Every standard construction is induced by a pair of adjoint functors*, Proc. Amer. Math. Soc. **16** (1965), 544–546.

[52] A. Kock, *On double-dualization monads*, Mathematisk Institute, Aarhus Universitet, Lecture Notes, Vol. 38 (1969).

[53] A. Kock, *Monads for which structures are adjoint to units*, J. Pure Appl. Algebra **104** (1995), 41–59.

[54] D. Kozen, *A representation theorem for models of $*$-free PDL*, Lecture Notes in Computer Science, Vol. 85, Springer-Verlag, Berlin (1980), 351–362.

[55] F.W. Lawvere, *Functorial Semantics of Algebraic Theories*, Dissertation, Columbia University (1963).

[56] F.W. Lawvere, *Functorial semantics of algebraic theories*, Proc. Nat. Acad. Sci. USA **50** (1963), 869–872.

[57] F.W. Lawvere, *Algebraic theories, algebraic categories, and algebraic functors*, The Theory of Models, Proceedings of the 1963 International Symposium at Berkeley, J.W. Addison, L. Henkin and A. Tarski, eds, North-Holland, Amsterdam (1965), 413–418.

[58] F.W. Lawvere, *Some algebraic problems in the context of functorial semantics of algebraic theories*, Lecture Notes in Math., Vol. 61, Springer-Verlag, Berlin (1968), 41–61.

[59] K. Lellahi and V. Tannen, *A calculus for collections and aggregates*, Category Theory and Computer Science, E. Moggi and G. Rosolini, eds, Springer-Verlag, Berlin (1997), 261–280.

[60] S. Liang, P. Hudak and M.P. Jones, *Monad transformers and modular interpreters*, Proceedings of POPL'95 Symposium Principles of Programming Languages, ACM Press (1995), 333–343.

[61] F.E.J. Linton, *Functorial measure theory*, Dissertation, Columbia University (1963).

[62] F.E.J. Linton, *Some aspects of equational categories*, Proceedings of the Conference on Categorical Algebra, La Jolla, 1965, S. Eilenberg, D.K. Harrison, S. MacLane and H. Röhrl, eds, Springer-Verlag, Berlin (1966), 84–94.

[63] F.E.J. Linton, *Triples vs. theories*, Notices Amer. Math. Soc. **13** (1966), 227.

[64] F.E.J. Linton, *Autonomous equational categories*, J. Math. Mech. **15** (1966), 637–642.

[65] F.E.J. Linton, *An outline of functorial semantics*, Lecture Notes in Math., Vol. 80, Springer-Verlag, Berlin (1969), 7-52.

[66] F.E.J. Linton, *Applied functorial semantics*, Lecture Notes in Math., Vol. 80, Springer-Verlag, Berlin (1969), 53–74.

[67] F.E.J. Linton, *Coequalizers in categories of algebras*, Lecture Notes in Math., Vol. 80, Springer-Verlag, Berlin (1969), 75–90.
[68] F.E.J. Linton, *Relative functorial semantics: adjointness results*, Lecture Notes in Math., Vol. 99, Springer-Verlag, Berlin (1969), 384–418.
[69] F.E.J. Linton, *Applied functorial semantics I*, Annali di Matematica Pura ed Applicata **86** (1970), 1–14.
[70] F.E.J. Linton, *Relative functorial semantics III, triples vs. theories*, Unpublished Manuscript (1974), 3 pages.
[71] S. MacLane, *Homologie des anneaux et des modules*, Colloque de topologie Algèbrique, Louvain (1956), 55–80.
[72] S. MacLane, *Homology*, Academic Press, New York (1963).
[73] S. MacLane, *Categories for the Working Mathematician*, Springer-Verlag, Berlin (1971).
[74] E.G. Manes, *A Triple Miscelleny: Some Aspects of the Theory of Algebras over a Triple*, Dissertation, Wesleyan University (1967).
[75] E.G. Manes, *A triple-theoretic construction of compact algebras*, Lecture Notes in Math., Vol. 80, Springer-Verlag, Berlin (1969), 91–118.
[76] E.G. Manes, *Algebraic Theories*, Springer-Verlag, Berlin (1976).
[77] E.G. Manes, *A class of fuzzy theories*, J. Math. Anal. Appl. **85** (1982), 409–451.
[78] E.G. Manes, *Predicate Transformer Semantics*, Cambridge University Press, New York (1992).
[79] E.G. Manes, *Implementing collection classes with monads*, Math. Structures Comput. Sci. **8** (1998), 231–276.
[80] E.G. Manes, *Boolean monads for functional programming*, Proceedings of World Multiconference on Systemics, Cybernetics and Informatics ISAS/SCI 2001, Orlando, Florida, USA, XVII, N. Callaos, S. Long and M. Loutfi, eds (2001), 189–193.
[81] E.G. Manes, *Taut monads and T0 spaces*, Theoret. Comput. Sci. **275** (2002), 79–109.
[82] J.-M. Maranda, *On fundamental constructions and adjoint functors*, Canadian Math. Bull. **9** (1966), 581–591.
[83] E. Meijer and J. Jeuring, *Merging monads and folds for functional programming*, Proceedings of the Fifth Annual Glasgow Workshop on Functional Programming, Ayr, Scotland, J. Launchbury and P.M. Sansom, eds, Springer-Verlag, Berlin (1992), 228–266.
[84] E. Moggi, *Computational lambda-calculus and monads*, Proceedings of Symposium on Logic and Computer Science, IEEE, Computer Society Press, Washington, DC (1989), 14–23.
[85] E. Moggi, *Notions of computation and monads*, Inform. and Comput. **93** (1991), 55–92.
[86] E. Moggi, *A semantics for evaluation logic*, Fund. Inform. **22** (1995), 117–152.
[87] P. Mulry, *Monads and algebras in the semantics of partial data types*, Theor. Comput. Sci. **99** (1992), 141–155.
[88] P. Mulry, *Monads in semantics*, Electronic Notes in Theoretical Computer Science **12** (1998), www.elsevier.nl/locate/entcs/volume12.html, 12 pp.
[89] J.W. Negrepontis, *Duality in algebra from the point of view of analysis*, J. Algebra **19** (1971), 228–253.
[90] R. Paré, *On absolute colimits*, J. Algebra **19** (1971), 80–95.
[91] S.L. Peyton Jones, *The Implementation of Functional Programming Languages*, Prentice Hall International, New York (1987).
[92] S.L. Peyton Jones and P. Wadler, *Imperative functional programming*, Proceedings of POPL'93 Symposium Principles of Programming Languages, ACM Press (1993), 71–84.
[93] R.S. Pierce, *Introduction to the Theory of Abstract Algebras*, Holt, Rinehart and Winston, New York (1968).
[94] G.D. Plotkin, *Call-by-name, call-by-value and the λ-calculus*, Theoret. Comput. Sci. **1** (1975), 125–159.
[95] V.R. Pratt, *Semantical considerations on Floyd–Hoare Logic*, Proceedings of 17th IEEE Symposium on the Foundations of Computer Science (1976), 109–121.
[96] J. Reynolds, *The essence of Algol*, Algorithmic Languages, de Bakker and Van Vleet, eds, North-Holland, Amsterdam (1981), 345–372.
[97] D.S. Scott, *Outline of a mathematical theory of computation*, Technical Monograph PRG-2, Oxford University Computing Laboratory (1970).
[98] D.S. Scott, *Continuous lattices*, Lecture Notes in Math., Vol. 234, Springer-Verlag, Berlin (1972), 97–136.

[99] D.S. Scott and C. Strachey, *Towards a mathematical semantics for computer languages*, Proceedings of the Symposium on Computers and Automata, J. Fox, ed., Polytechnic Institute of Brooklyn Press (1971), 19–46.

[100] Z. Semadeni, *A simple topological proof that the underlying set functor from compact spaces is monadic*, Lecture Notes in Math., Vol. 378, Springer-Verlag, Berlin (1974), 429–435.

[101] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton (1976).

[102] J. Słominski, *The theory of abstract algebras with infinitary operations*, Dissertationes Math. **18** (1959).

[103] M. Spivey, *A categorical approach to the theory of lists*, Lecture Notes in Computer Science, Vol. 375, Springer-Verlag, Berlin (1989), 399–408.

[104] M. Spivey, *A functional theory of exceptions*, Sci. Comput. Programming **14** (1990), 25–42.

[105] R.M. Solovay, *New proof of a theorem of Gaifmann and Hales*, Bull. Amer. Math. Soc. **72** (1966), 282–284.

[106] G.L. Steele, *Building interpreters by composing monads*, Proceedings of POPL'94 Symposium Principles of Programming Languages, ACM Press (1994), 474–492.

[107] P. Trinder, *Comprehensions, a query language for DBPLs*, Database Programming Langues: Bulk Types and Persistent Data, Third International Workshop, August 27–30, 1991, Nafplion, Greece, P. Kanellakis and J.W. Schmidt, eds, Morgan Kaufmann, San Mateo, CA, USA (1991), 55–68.

[108] P. Wadler, *Comprehending monads*, Math. Structures Comput. Sci. **2** (1992), 461–493.

[109] P. Wadler, *The essence of functional programming*, Proceedings of POPL'92 Symposium Principles of Programming Languages, ACM Press (1992), 1–14.

[110] P. Wadler, *Monads for functional programming*, Advanced Functional Programming, Bästad, Sweden, J. Jeuring and E. Meijer, eds, Springer-Verlag, Berlin (1995), 24–52.

[111] O. Wyler, *Algebraic theories of continuous lattices*, Lecture Notes in Math., Vol. 871, Springer-Verlag, Berlin (1981), 390–413.

[112] N. Yoneda, *On the homology of modules*, J. Fac. Sci. Tokyo, Sec. I. **7** (1954), 193–227.

This Page Intentionally Left Blank

Section 2C
Algebraic K-theory

This Page Intentionally Left Blank

Classical Algebraic K-Theory: The Functors K_0, K_1, K_2

Aderemi Kuku

The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
E-mail: kuku@ictp.trieste.it

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

Introduction

Algebraic K-theory has grown phenomenally in various directions in the last three decades as a multidisciplinary subject whose methods and contents span many areas of mathematics – notably algebra, number theory, topology, geometry (algebraic, differential, noncommutative) and functional analysis. As such, it has grown to become one of the most unifying forces in mathematical research. The subject has also recorded outstanding success in the investigations and solutions of many famous problems (see [6]).

It is generally accepted that algebraic K-theory actually started with Grothendieck's construction of an Abelian group $K(\mathcal{A})$ (now denoted $K_0(\mathcal{A})$) associated to a suitable subcategory of an Abelian category (e.g., for a scheme X, $\mathcal{A} = \mathcal{P}(X)$ the category of locally free sheaves of O_X-modules or $\mathcal{A} = \mathcal{M}(X)$, the category of coherent sheaves of O_X-modules). This construction was done by A. Grothendieck during his reformulation and proof of his generalised Riemann–Roch theorem, see [15] or [31]. However, there were earlier works which were later recognized as proper constituents of the subject, e.g., J.H.C. Whitehead's construction of $\mathrm{Wh}(\pi_1(X))$ (X a topological space) [119] or even much earlier work of Dedekind and Weber [19] on ideal class groups.

Meanwhile, M.F. Atiyah and F. Hirzebruch [4,5], studied for any finite CW-complex X, the Abelian group $K_0(\mathcal{A})$ where for $k = \mathbb{R}$ or $\mathbb{C}$, $\mathcal{A} = \mathrm{Vect}_k(X)$ the category of finite dimensional k-vector bundles on X in what became known as topological K-theory. Now, the realisation by R.G. Swan [95] that when X is a compact space, the category $\mathrm{Vect}_{\mathbb{C}}(X)$ is equivalent to the category $\mathcal{P}(\mathbb{C}(X))$ of finitely generated projective modules over the ring $\mathbb{C}(X)$ of complex-valued continuous functions on X, provided the initial connection between topological K-theory and algebraic K-theory. Moreover, the fact that when X is affine (i.e. $X = \mathrm{Spec}(A)$, A a commutative ring), the category $\mathcal{P}(X)$ is equivalent to $\mathcal{P}(A)$, the category of finitely generated projective A-modules, also confirms the appropriateness of $K_0(\mathcal{P}(A))$ (A any ring with identity) as a good definition of K_0 of a ring A, usually written $K_0(A)$.

The groups $K_0(A)$ for various types of rings A (e.g., Dedekind domains, number fields, group-rings, orders, C^*-algebras, etc.) have been subjected to intense studies over the years especially because the groups $K_0(A)$ for relevant A's are replete with applications at first in several areas of (pure) mathematics and later in some areas of applied mathematics and physics. For example, C.T.C. Wall [114] showed that if X is a connected space dominated by a finite CW-complex, then there is a well defined obstruction ω in $K_0(\mathbb{Z}\pi_1(X))$ such that X has the homotopy type of a finite complex if and only if $\omega = 0$ (see 2.2.12). Moreover, K_0 of C^*-algebras is connected with noncommutative geometry (see [17] or 1.4.3(iv)); K_0 of Dedekind domains with class groups of number theory (see Section 2); K_0 of orders and group-rings with representation theory (see [18] or Section 3), etc. Furthermore, $K_0(\mathcal{C})$ is also well defined for other types of category $\mathcal{C}$ (e.g., symmetric monoidal categories, see Section 1.4).

The definition of K_1, due to H. Bass, was inspired by Atiyah–Hirzebruch topological K-theory $K^{-n}(X) := \widetilde{K}(S^n(X))$ where $S(X)$ is the suspension of X and $\widetilde{K}(Y) := \mathrm{Ker}(K^0(Y) \to K(*))$ for any connected paracompact space Y and $*$ a point of Y (see [3] or [38]). H. Bass defined K_1 of a ring A, $K_1(A)$, modelled on the description of bundles on SX by clutching. Because for any finite group G, $\mathrm{Wh}(G)$ (the Whitehead group of G), defined as a quotient of $K_1(\mathbb{Z}G)$ (see Section 6.10) houses some topological invariants known as "Whitehead torsion" when $G=\pi_1(X)$ (X a finite CW-complex), computations of $K_1(\mathbb{Z}G)$ and also of $SK_1(\mathbb{Z}G) := \mathrm{Ker}(K_1(\mathbb{Z}G) \to K_1(QG))$, became of interest in topology since $\mathrm{rank}\, K_1(\mathbb{Z}G) = \mathrm{rank}\,\mathrm{Wh}(G)$ and $SK_1(\mathbb{Z}G)$ is the full torsion subgroup of $\mathrm{Wh}(G)$ (see Section 6.10 or [64]). Also of interest are computations of $SK_1(\widehat{\mathbb{Z}}_pG)$ (see, e.g., [64] or [42]). Moreover, stability considerations of $K_1(A)$ yielded results of finite generation of $K_1(\mathbb{Z}G)$ and finiteness of $SK_1(\mathbb{Z}G)$ [7,8] as well as a key step to the solution of the congruence subgroup problem for $SL_n(A)$ where A is the ring of integers of a number field F [13].

The definition of $K_2^M(A)$, A any ring, is due to J. Milnor [63]. As will be seen in a forthcoming chapter on higher K-theory, $K_2^M(A)$ coincides with the Quillen K-groups $K_2(\mathcal{P}(A)) = K_2(A)$ but for $n \neq 2$, $K_n^M(A)$, defined only for commutative rings A, are in general different from $K_n(A)$ even though there are maps between them, as well as connections with other theories, e.g., Galois and étale cohomology theories, Brauer groups, etc. yielding famous conjectures – e.g., Milnor, Bloch–Kato conjectures, etc.

We now briefly review the contents of this chapter. Section 1 introduces the Grothendieck group associated to a semigroup A and the ring associated to a semi-ring (1.1) leading to discussions on K_0 of rings and K_0 of symmetric monoidal categories with copious examples – topological K-groups $K_0(X)$, $KU(X)$, $K_G^0(X)$, Burnside rings, representation rings, Witt rings, Picard groups – $\mathrm{Pic}(R)$ (R, a commutative ring with identity), $\mathrm{Pic}(X)$ (X a locally ringed space); K_0 of Azamaya algebras, etc. We also briefly indicate how to realise K_0 as "Mackey" functors yielding induction theory for K_0 of group-rings (see [48]) – a topic that will be a subject of another chapter in more generality (see [24,25,46]).

Section 2 deals with class groups of Dedekind domains, orders and group rings, and we also briefly discuss Wall's finiteness obstruction as an application. In Section 3, we discuss K_0 of an exact category with copious examples while Section 4 exposes some fundamental properties and examples of K_0 of exact categories, e.g., devissage, resolution and localisation theorems – that will be seen in more generality in a forthcoming chapter on higher K-theory. We also discuss K_0 of the category of nilpotent endomorphisms with consequent fundamental theorems for K_0, G_0 of rings and schemes.

In Section 5 we discuss $K_1(A)$ with the observation that the definition due to Bass coincides with Quillen's $K_1(\mathcal{P}(A))$ or $\pi_1(BGL(A)^+)$. We also briefly discuss K_1 of local rings and skew fields; Mennicke symbols and some stability results for K_1. In Section 6, which deals with SK_1 of orders and group rings, we, among other things, call attention to the fact that when R is the ring of integers in a number field, Λ an R-order, then $K_1(\Lambda)$ is a finitely generated Abelian group, $SK_1(\Lambda), SK_1(\widehat{\Lambda}_p)$ are finite groups, see [64], with the observation that these results have since been generalised, i.e. for all $n \geqslant 1$, $K_n(\Lambda)$ is finitely generated and $SK_n(\Lambda)$, $SK_n(\widehat{\Lambda}_p)$ are finite for all prime ideals p of R (see [49, 50]). Similarly, for a maximal order Γ in a p-adic semi-simple algebra Σ, the result that $SK_1(\Gamma)=0$ iff Σ is unramified over its centre [41] has been extended for all $n \geqslant 1$ (i.e. $SK_{2n}(\Gamma)=0$ and $SK_{2n-1}(\Gamma)=0$ iff Σ is unramified over its centre) (see [44]). We refer to copious computations of SK_1 of orders and group rings in [64] and discuss Whitehead torsion in Section 6.10.

Section 7 is devoted to discussing some $K_1 - K_0$ exact sequences – Mayer–Vietoris, localisation sequences and the exact sequence associated to an ideal. The localisation sequence leads to the introduction of the fundamental theorem for K_1.

The last section (Section 8) deals with a rather brief review of the functor K_2 due to J. Milnor [63]. We observe that $K_2(A) = H_2(E(A), \mathbb{Z})$ for any ring A with identity and that when A is a field, division ring, local or semi-local ring, $K_2(A)$ is generated by symbols. We then briefly discuss the connections between K_2, Brauer group of fields and Galois cohomology leading to the Merkurjev–Suslin theorem (8.2.4) which we discuss in the context of the Bloch–Kato conjecture for higher-dimensional K-theory of fields with a brief review of the current status of the conjecture. We end Section 8 with applications of K_2 to pseudo-isotopy of manifolds and Bloch's formula for Chow groups.

In anticipation of the forthcoming chapter on higher K-theory we have included references to some results on higher K-theory that apply to lower K-groups as well as constitute generalisations of such known results for K_0, K_1 or K_2.

NOTES ON NOTATION. If R is a Dedekind domain with quotient field F, $\mathbf{p}$ any prime ideal of R, we write $R_{\mathbf{p}}$ for the localisation of R at $\mathbf{p}$, $\widehat{R}_{\mathbf{p}}(\widehat{F}_{\mathbf{p}})$ for the completion of R (resp. F) at $\mathbf{p}$. If Λ is an R-order in a semi-simple F-algebra Σ, we write $\Lambda_{\mathbf{p}}$ for $R_{\mathbf{p}} \otimes_R \Lambda$, $\widehat{\Lambda}_{\mathbf{p}}$ for $\widehat{R}_{\mathbf{p}} \otimes_R \Lambda$, and $\widehat{\Sigma}_{\mathbf{p}} = \widehat{F}_{\mathbf{p}} \otimes_F \Sigma$. For all $n \geqslant 0$, we write $SK_n(\Lambda) = \mathrm{Ker}(K_n(\Lambda) \to K_n(\Sigma))$ and $SG_n(\Lambda) = \mathrm{Ker}(G_n(\Lambda) \to G_n(\Sigma))$.

We shall write $\mathcal{G}Set$ for the category of G-sets (G a group), $\mathcal{R}ings$ for the category of rings with identity and homomorphisms preserving identity, $\mathcal{CR}ings$ for the category of commutative rings and ring homomorphisms preserving identity.

Many other notations used are defined in the text.

1. Some basic Grothendieck group constructions and examples

1.1. *Grothendieck group associated with a semi-group*

1.1.1. Let $(A, +)$ be an Abelian semi-group. Define a relation '$\sim$' on $A \times A$ by $(a, b) \sim (c, d)$ if there exists $u \in A$ such that $a + d + u = b + c + u$. One can easily check that '$\sim$' is an equivalence relation. Let $\overline{A}$ denote the set of equivalence classes of '$\sim$', and write $[a, b]$ for the class of (a, b) under '$\sim$'. We define addition ($\dot{+}$) on $\overline{A}$ by $[a, b] \dot{+} [c, d] = [a + c, b + d]$. Then $(\overline{A}, \dot{+})$ is an Abelian group in which the identity element is $[a, a]$ and the inverse of $[a, b]$ is $[b, a]$.

Moreover, there is a well-defined additive map $f : A \to \overline{A} : a \to [a + a, a]$ which is, in general, neither injective nor surjective. However, f is injective iff A is a cancellation semi-group, i.e. iff $a + c = b + c$ implies that $a = b$ for all $a, b, c \in A$, see [48] or [38].

1.1.2. It can be easily checked that $\overline{A}$ possesses the following universal property with respect to the map $f : A \to \overline{A}$. Given any additive map $h : A \to B$ from A to an Abelian group B, then there exists a unique map $g : \overline{A} \to B$ such that $h = gf$.

1.1.3. DEFINITION. $\overline{A}$ is usually called the Grothendieck group of A or the group completion of A and denoted by $K(A)$.

1.1.4. REMARKS.

(i) The construction of $K(A) = \overline{A}$ above can be shown to be equivalent to the following:
Let $(F(A), \dot{+})$ be the free Abelian group freely generated by the elements of A, and $R(A)$ the subgroup of $F(A)$ generated by all elements of the form $a \dot{+} b - (a+b)$, $a, b \in A$. Then $K(A) \simeq F(A)/R(A)$.

(ii) If A, B, C are Abelian semi-groups together with bi-additive map $f : A \times B \to C$, then f extends to a unique bi-additive map $\bar{f} : \overline{A} \times \overline{B} \to \overline{C}$ of the associated Grothendieck groups. If A is a semi-ring, i.e. an additive Abelian group together with a bi-additive multiplication $A \times A \to A$ $(a, b) \to ab$, then the multiplication extends uniquely to a multiplication $\overline{A} \times \overline{A} \to \overline{A}$ which makes $\overline{A}$ into a ring (commutative if A is commutative) with identity $1 = [1+1, 1]$ in $\overline{A}$ if $1 \in A$.

(iii) If B is a semi-module over a semi-ring A, i.e. if B is an Abelian semi-group together with a bi-additive map $A \times B \to B : (a, b) \to a \cdot b$ satisfying $a'(ab) = (a'a)b$ for $a, a' \in A$, $b \in B$, then the associated Grothendieck group $\overline{B}$ is an $\overline{A}$-module.

(iv) If $A = \{1, 2, 3 \ldots\}$, $\overline{A} = K(A) = \mathbb{Z}$. Hence the construction in 1.1.1 is just a generalisation of the standard procedure of constructing integers from the natural numbers.

(v) A sub-semi-group A of an Abelian semi-group B is said to be cofinal in B if for any $b \in B$, there exists $b' \in B$ such that $b + b' \in A$. It can be easily checked that $K(A)$ is a subgroup of $K(B)$ if A is cofinal in B.

1.2. K_0 *of a ring*

1.2.1. For any ring Λ with identity, let $\mathcal{P}(\Lambda)$ be the category of finitely generated projective Λ-modules. Then the isomorphism classes $I\mathcal{P}(\Lambda)$ of objects of $\mathcal{P}(\Lambda)$ form an Abelian semi-group under direct sum '$\oplus$'. We write $K_0(\Lambda)$ for $K(I\mathcal{P}(\Lambda))$ and call $K_0(\Lambda)$ the Grothendieck group of Λ. For any $P \in \mathcal{P}(\Lambda)$, we write (P) for the isomorphism class of P (i.e. an element of $I\mathcal{P}(\Lambda)$) and $[P]$ for the class of (P) in $K_0(\Lambda)$.

If Λ is commutative, then $I\mathcal{P}(\Lambda)$ is a semi-ring with tensor product $\otimes_\Lambda$ as multiplication which distributes over '$\oplus$'. Hence $K_0(\Lambda)$ is a ring by 1.1.4(ii).

1.2.2. REMARKS.

(i) $K_0 : \mathcal{R}ings \to \mathcal{A}b : \Lambda \to K_0(\Lambda)$ is a functor – since any ring homomorphism $f : \Lambda \to \Lambda'$ induces a semi-group homomorphism $I\mathcal{P}(\Lambda) \to I\mathcal{P}(\Lambda') : P \to P \otimes \Lambda'$ and hence a group homomorphism $K_0(\Lambda) \to K_0(\Lambda')$.

(ii) K_0 is also a functor: $\mathcal{CR}ings \to \mathcal{CR}ings$.

(iii) $[P] = [Q]$ in $K_0(\Lambda)$ iff P is stably isomorphic to Q in $\mathcal{P}(\Lambda)$, i.e. iff $P \oplus \Lambda^n \simeq Q \oplus \Lambda^n$ for some integer n. In particular $[P] = [\Lambda^n]$ for some n iff P is stably free, see [8] or [7].

1.2.3. EXAMPLES.

(i) If Λ is a field or a division ring or a local ring or a principal ideal domain, then $K_0(\Lambda) \simeq \mathbb{Z}$.

NOTE. The proof in each case is based on the fact that any finitely generated Λ-module is free and Λ satisfies the invariant basis property (i.e. $\Lambda^r \simeq \Lambda^s \Rightarrow r = s$). So $I\mathcal{P}(\Lambda) \simeq \{1, 2, 3, \ldots\}$ and so, $K_0(\Lambda) \simeq \mathbb{Z}$ by 1.1.4(iv), see [8] or [79].

(ii) Any element of $K_0(\Lambda)$ can be written as $[P] - r[\Lambda]$ for some integer $r > 0$, $P \in \mathcal{P}(\Lambda)$ or as $s[\Lambda] - [Q]$ for some $s > 0$, $Q \in \mathcal{P}(\Lambda)$ (see [8] or [105]). If we write $\widetilde{K}_0(\Lambda)$ for the quotient of $K_0(\Lambda)$ by the subgroup generated by $[\Lambda]$, then every element of $\widetilde{K}_0(\Lambda)$ can be written as $[P]$ for some $P \in \mathcal{P}(\Lambda)$, see [105] or [8].

(iii) If $\Lambda \simeq \Lambda_1 \times \Lambda_2$ is a direct product of two rings Λ_1, Λ_2 then $K_0(\Lambda) \simeq K_0(\Lambda_1) \times K_0(\Lambda_2)$ (see [105] for a proof).

(iv) Let G be a semi-simple simply connected affine algebraic group over an algebraically closed field. Let A be the coordinate ring of G. Then $K_0(A) \simeq \mathbb{Z}$.

REMARKS. See [55] for a proof of this result which says that all algebraic vector bundles on G are stably trivial. The result is due to A. Grothendieck.

(v) $K_0(k[x_0, x_1, \ldots, x_n]) \simeq \mathbb{Z}$. This result is due to J.P. Serre, see [83].

1.3. K_0 *of a ring via idempotents*

1.3.1. For any ring Λ with identity, let $M_n(\Lambda)$ be the set of $n \times n$ matrices over Λ, and write $M(\Lambda) = \bigcup_{n=1}^{\infty} M_n(\Lambda)$. Also let $GL_n(\Lambda)$ be the group of invertible $n \times n$ matrices over Λ and write $GL(\Lambda) = \bigcup_{n=1}^{\infty} GL_n(\Lambda)$. For $P \in \mathcal{P}(\Lambda)$ there exists $Q \in \mathcal{P}(\Lambda)$ such that $P \oplus Q \simeq \Lambda^n$ for some n. So, we can identify with each $P \in \mathcal{P}(\Lambda)$ an idempotent matrix $p \in M_n(\Lambda)$ (i.e. $p : \Lambda^n \to \Lambda^n$) which is the identity on P and '0' on Q.

Note that if p, q are idempotent matrices in $M(\Lambda)$, say $p \in M_r(\Lambda)$, $q \in M_s(\Lambda)$, corresponding to $P, Q \in \mathcal{P}(\Lambda)$, then $P \simeq Q$ iff it is possible to enlarge the sizes of p, q (by possibly adding zeros in the lower right-hand corners) such that p, q have the same size ($t \times t$, say) and are conjugate under the action of $GL_t(\Lambda)$, see [79].

Let $\mathrm{Idem}(\Lambda)$ be set of idempotent matrices in $M(\Lambda)$. It follows from the last paragraph that $GL(\Lambda)$ acts by conjugation on $\mathrm{Idem}(\Lambda)$, and so, we can identify the semi-group $I\mathcal{P}(\Lambda)$ with the semi-group of conjugation orbits $(\mathrm{Idem}(\Lambda))\hat{}$ of the action of $GL(\Lambda)$ on $\mathrm{Idem}(\Lambda)$ where the semi-group operation is induced by $(p, q) \to \left(\begin{smallmatrix} p & 0 \\ 0 & q \end{smallmatrix}\right)$. $K_0(\Lambda)$ is the Grothendieck group of this semi-group $(\mathrm{Idem}(\Lambda))\hat{}$.

1.3.2. REMARKS.

(i) Computing K_0-groups via idempotents is particularly useful when Λ is an involutive Banach algebra or C^*-algebra (see [17,21] for example).

(ii) Also the methods of computing K_0-groups via idempotents are used to prove the following results 1.3.2 and 1.3.3 below.

1.3.3. THEOREM [79]. *If* $\{\Lambda_i\}_{i \in I}$ *is a direct system of rings* (*with identity*), *then*

$$K_0(\Lambda) = \varinjlim_{i \in I} K_0(\Lambda_i).$$

For proof see [79].

1.3.4. THEOREM (Morita equivalence for K_0 of rings). *For any ring Λ and any natural number $n > 0$, $K_0(\Lambda) \simeq K_0(M_n(\Lambda))$.*

PROOF. Follows from 1.3.3 since $\mathrm{Idem}(M_n(\Lambda) = \mathrm{Idem}(\Lambda)$ and $GL(M_n(\Lambda)) \simeq GL(\Lambda)$. □

1.3.5. COROLLARY. *If Λ is a semi-simple ring, then $K_0(\Lambda) \simeq \mathbb{Z}^r$ for some positive integer r.*

PROOF (*Sketch*). Let $V_1, \dots, V_r$ be simple Λ-modules. By the Wedderburn theorem, $\Lambda \simeq \prod_{i=1}^r M_{n_i}(D_i)$ where $D_i = \mathrm{Hom}_\Lambda(V_i, V_i)$ and $\dim_{D_i}(V_i) = n_i$. Hence $K_0(\Lambda) \simeq \prod_{i=1}^r K_0(M_{n_i}(D_i)) \simeq \prod_{i=1}^r K_0(D_i) \simeq \mathbb{Z}^r$ by 1.2.3(i) and (iii) as well as 1.3.4. □

1.4. K_0 *of symmetric monoidal categories*

1.4.1. DEFINITION. A symmetric monoidal category is a category $\mathcal{C}$ equipped with a functor $\perp : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$ and a distinguished object "0" such that $\perp$ is "coherently associative and commutative" in the sense of MacLane, that is,

(i) $A \perp 0 \simeq A \simeq 0 \perp A$.
(ii) $A \perp (B \perp C) \simeq (A \perp B) \perp C$.
(iii) $A \perp B \simeq B \perp A$ for all $A, B, C \in \mathcal{C}$.

Moreover, the following diagrams commute.

(i)
$$\begin{tikzcd}
(A \perp (0 \perp B)) \arrow[r, "\sim"] \arrow[d, "\wr"] & (A \perp 0) \perp B \arrow[d, "\wr"] \\
A \perp B \arrow[r, "\sim"] & B \perp A
\end{tikzcd}$$

(ii)
$$\begin{tikzcd}
A \perp 0 \arrow[rr, "\sim"] \arrow[dr, "\sim"'] & & 0 \perp A \arrow[dl, "\sim"] \\
& A &
\end{tikzcd}$$

(iii)
$$\begin{tikzcd}
A \perp (B \perp (C \perp D)) \arrow[r, "\sim"] \arrow[d, "\wr"] & (A \perp B) \perp (C \perp D) \arrow[d, "\wr"] \\
A \perp ((B \perp C) \perp D) \arrow[d, "\wr"] & ((A \perp B) \perp C) \perp D \\
(A \perp (B \perp C)) \perp D \arrow[ur, "\sim"] &
\end{tikzcd}$$

Let $I\mathcal{C}$ be the set of isomorphism classes of objects of $\mathcal{C}$. Clearly, if $\mathcal{C}$ is small, then $(I\mathcal{C}, \perp)$ is an Abelian semi-group (in fact a monoid), and we write $K_0^\perp(\mathcal{C})$ for $K(I\mathcal{C}, \perp)$ or simply $K_0(\mathcal{C})$ when the context is clear.

In other words, $K_0^\perp(\mathcal{C}) = F(\mathcal{C})/R(\mathcal{C})$ where $F(\mathcal{C})$ is the free Abelian group on the isomorphism classes (C) of $\mathcal{C}$-objects, and $R(\mathcal{C})$ the subgroup of $F(\mathcal{C})$ generated by $(C' \perp C'') - (C') - (C'')$ for all C', C'' in $\mathrm{ob}(\mathcal{C})$.

1.4.2. REMARKS.

(i) $K_0^\perp(\mathcal{C})$ satisfies a universal property as in 1.1.
(ii) If $\mathcal{C}$ has another composition '0' that is associative and distributive with respect to $\perp$, then $K_0^\perp(\mathcal{C})$ can be given a ring structure through '0' as multiplication and we

shall sometimes denote this ring by $K_0^{\perp}(\mathcal{C}, \perp, 0)$ or $K_0(\mathcal{C}, \perp, 0)$ or just $K_0(\mathcal{C})$ if the context is clear.

1.4.3. EXAMPLES.

(i) If Λ is any ring with identity, then $(\mathcal{P}(\Lambda), \oplus)$ is a symmetric monoidal category (s.m.c.) and $K_0^{\oplus}(\Lambda) = K_0(\Lambda)$ as in 1.2.1.

(ii) If Λ is commutative, then $K_0^{\oplus}(\Lambda)$ is a ring where $(\mathcal{P}(\Lambda), \oplus)$ has the further composition '$\otimes$'.

(iii) Let X be a compact topological space and for $F = \mathbb{R}$ or $\mathbb{C}$, let $\mathbf{VB}_F(X)$ be the (symmetric monoidal) category of (finite-dimensional) vector bundles on X. Then $I\mathbf{VB}_F(X)$ is an Abelian monoid under Whitney sum '$\oplus$'. It is usual to write $KO(X)$ for $K_0^{\oplus}(\mathbf{VB}_{\mathbb{R}}(X))$ and $KU(X)$ for $K_0^{\oplus}(VB_{\mathbb{C}}(X))$. Note that if X, Y are homotopy equivalent, then $KO(X) = KO(Y)$ and $KU(X) = KU(Y)$. Moreover, if X is contractible, we have $KO(X) = KU(X) = \mathbb{Z}$ (see [3] or [38]).

(iv) Let X be a compact space, $\mathbb{C}(X)$ the ring of $\mathbb{C}$-valued functions on X. By a theorem of R.G. Swan [95], there exists an equivalence of categories $\Gamma : VB_{\mathbb{C}}(X) \to \mathcal{P}(\mathbb{C}(X))$ taking a vector bundle: $E \xrightarrow{p} X$ to $\Gamma(E)$, where $\Gamma(E) = \{\text{sections } s : X \to E \mid ps = 1\}$. This equivalence induces a group isomorphism $I : KU(X) \simeq K_0(\mathbb{C}(X))$.

This isomorphism (I) provides the basic initial connection between algebraic K-theory (right-hand side of I) and topological K-theory (left-hand side of I) since the K-theory of $\mathcal{P}(\Lambda)$ for an arbitrary ring Λ could be studied instead of the K-theory of $\mathcal{P}(\mathbb{C}(X))$.

Now, $\mathbb{C}(X)$ is a commutative C^*-algebra and the Gelfand–Naimark theorem [17] says that any commutative C^*-algebra Λ has the form $\Lambda = \mathbb{C}(X)$ for some locally compact space X. Indeed, for any commutative C^*-algebra Λ, we could take X as the spectrum of Λ, i.e. the set of all nonzero homomorphisms from Λ to $\mathbb{C}$ with the topology of pointwise convergence. Noncommutative geometry is concerned with the study of noncommutative C^*-algebras associated with "noncommutative" spaces and K-theory (algebraic and topological) of such C^*-algebras has been extensively studied and connected to some (co)homology theories (e.g., Hochschild and cyclic (co)homology theories) of such algebras through Chern characters (see, e.g., [21,54,17,22]).

(v) Let G be a group acting continuously on a topological space X. The category $VB_G(X)$ of complex G-vector bundles on X is symmetric monoidal under Whitney sum '$\oplus$' and we write $K_G^0(X)$ for the Grothendieck group $K_0(VB_G(X))$. If X is a point, $VB_G(X)$ is the category of representations of G in $\mathcal{P}(\mathbb{C})$ and $K_G^0(X) = R(G)$, the representation ring of G.

If G acts trivially on X, then $K_G^0(X) \simeq KU(X) \otimes_{\mathbb{Z}} R(G)$ (see [81] or [82]).

(vi) Let *FSet* be the category of finite sets, $\dot{\cup}$ the disjoint union. Then (*FSet*, $\dot{\cup}$) is a symmetric monoidal category and $K_0^{\dot{\cup}}(\mathit{FSet}) \simeq \mathbb{Z}$ (see [48]).

(vii) Let R be a commutative ring with identity. Then $\mathrm{Pic}(R)$, the category of finitely generated projective R-modules of rank one (or equivalently the category of algebraic line bundles L over R) is a symmetric monoidal category and $K_0^{\otimes}(\mathrm{Pic}(R)) = \mathrm{Pic}(R)$, the Picard group of R.

(viii) The category $\operatorname{Pic}(X)$ of line bundles on a locally ringed space is a symmetric monoidal category under '$\otimes$' and $K_0^{\otimes}(\operatorname{Pic}(X)) := \operatorname{Pic}(X)$ is called the Picard group of X. Observe that when $X = \operatorname{Spec}(R)$, we recover $\operatorname{Pic}(R)$ in (vii). It is well known that $\operatorname{Pic}(X) \simeq H^1(X, O_X^*)$, see [33] or [67].

(ix) Let R be a commutative ring with identity. An R-algebra Λ is called an Azumaya algebra if there exists another R-algebra Λ' such that $\Lambda \otimes_R \Lambda' \simeq M_n(R)$ for some positive integer n. Let $\operatorname{Az}(R)$ be the category of Azumaya algebras. Then $(\operatorname{Az}(R), \otimes_R)$ is a symmetric monoidal category. Moreover, the category $F\mathcal{P}(R)$ of faithfully projective R-modules is symmetric monoidal with respect to $\perp = \otimes_R$ if the morphisms in $F\mathcal{P}(R)$ are restricted to isomorphisms. There is a monoidal functor $F\mathcal{P}(R) \to \operatorname{Az}(R) : P \to \operatorname{End}_R(P)$ inducing a group homomorphism $K_0(F\mathcal{P}(R)) \xrightarrow{\varphi} K_0(\operatorname{Az}(R))$. The cokernel of φ is called the Brauer group of R and denoted by $\operatorname{Br}(R)$. Hence $\operatorname{Br}(R)$ is the Abelian group generated by isomorphism classes $[\Lambda]$ with relations $[\Lambda \otimes_R \Lambda'] = [\Lambda] + [\Lambda']$ and $[\operatorname{End}_R(P)] = 0$.

If R is a field F, then $\operatorname{End}_R(P) \simeq M_n(F)$ for some n and $\operatorname{Br}(F)$ is the Abelian group generated by isomorphism classes of central simple F-algebras with relations $[\Lambda \otimes \Lambda'] = [\Lambda] + [\Lambda']$ and $[M_n(F)] = 0$ (see [79]).

(x) Let G be a finite group, $\mathcal{C}$ any small category. Let $\mathcal{C}_G$ be the category of G-objects in $\mathcal{C}$ or equivalently, the category of G-representations in $\mathcal{C}$, i.e. objects of $\mathcal{C}_G$ are pairs $(X, U : G \to \operatorname{Aut}(X))$ where $X \in \operatorname{ob}(\mathcal{C})$ and U is a group homomorphism from G to the group of $\mathcal{C}$-automorphisms of X. If $(\mathcal{C}, \perp)$ is a symmetric monoidal category, so is $(\mathcal{C}_G, \dot{\perp})$ where for

$$\bigl(X, U : G \to \operatorname{Aut}(X)\bigr), \quad \bigl(X', U' : G \to \operatorname{Aut}(X')\bigr)$$

in $\mathcal{C}_G$, we define

$$(X, U) \mathbin{\dot{\perp}} (X', U') := \bigl(X \perp X', U \perp U' : G \to \operatorname{Aut}(X \perp X')\bigr),$$

where $U \perp U'$ is defined by the composition

$$G \xrightarrow{U \times U'} \operatorname{Aut}(X) \times \operatorname{Aut}(X') \to \operatorname{Aut}(X \perp X').$$

So we obtain the Grothendieck group $K_0^{\dot{\perp}}(\mathcal{C}_G)$.

If $\mathcal{C}$ possesses a further associative composition '0' such that $\mathcal{C}$ is distributive with respect to $\perp$ and '0', then so is $\mathcal{C}_G$, and hence $K_0^{\dot{\perp}}(\mathcal{C}_G)$ is a ring.

EXAMPLE. (a) If $\mathcal{C} = \mathcal{P}(R), \perp = \oplus$, '0' $= \otimes_R$ where R is a commutative ring with identity, then $\mathcal{P}(R)_G$ is the category of RG-lattices (see [48,18,47] and $K_0(\mathcal{P}(R)_G)$ is a ring usually denoted by $G_0(R, G)$. Observe that when $R = \mathbb{C}$, $G_0(\mathbb{C}, G)$ is the usual representation ring of G denoted in the literature by $R(G)$. Also see 3.1.4(iv).

(b) If $\mathcal{C} = \mathit{FSets}$, '$\perp$' = disjoint union, '0' – Cartesian product. Then $K_0(\mathcal{C}_G)$ is the Burnside ring of G usually denoted by $\Omega(G)$. See [48].

(xi) Let G be a finite group, S a G-set. We can associate with S a category $\mathbf{S}$ as follows: $\mathrm{ob}(\mathbf{S}) = \{s \mid s \in S\}$. For $s, t \in S$, $\mathrm{Hom}_{\mathbf{S}}(s,t) = \{(g,s) \mid s \in G,\ gs = t\}$, where composition is defined for $t = gs$ by $(h,t) \cdot (g,s) = (hg,s)$ and the identity morphism $s \to s$ is given by (e,s) where e is the identity element of G. Now let $(\mathcal{C}, \perp)$ be a symmetric monoidal category and let $[\mathbf{S}, \mathcal{C}]$ be the category of covariant functors $\zeta : \mathbf{S} \to \mathcal{C}$. The $([\mathbf{S}, \mathcal{C}], \dot{\perp})$ is also a symmetric monoidal category where $(\zeta \mathbin{\dot{\perp}} \eta)_{(g,s)} = \zeta_s \perp \eta_s \to \zeta_{gs} \perp \eta_{gs}$. We write $K_0^G(S, \mathcal{C})$ for the Grothendieck group of $[\mathbf{S}, \mathcal{C}]$.

If $(\mathcal{C}, \perp)$ possesses an additional composition '0' that is associative and distributive with respect to '$\perp$', then $K_0^G(S, \mathcal{C})$ can be given a ring structure (see [48]).

Note that for any symmetric monoidal category $(\mathcal{C}, \perp)$, $K_0^G(-, \mathcal{C}) : GSet \to \mathcal{A}b$ is a 'Mackey functor' (see [48] and the chapter on Mackey functors in volume 2 of this Handbook), and that when $\mathcal{C}$ possesses an additional composition '0' discussed above, then $K_0^G(-, \mathcal{C}) : GSet \to \mathcal{A}b$ is a 'Green functor' (see [48]). We shall discuss these matters in further details under abstract representation theory – a forthcoming chapter.

(xii) Let A be an involutive Banach algebra and $\mathrm{Witt}(A)$ the group generated by isomorphism classes $[Q]$ of invertible Hermitian forms Q on $P \in \mathcal{P}(A)$ with relations $[Q_1 \oplus Q_2] = [Q_1] + [Q_2]$ and $[Q] + [-Q] = 0$. Define a map $\varphi : K_0(A) \to \mathrm{Witt}(A)$ by $[P] \mapsto$ class of (P, Q) with Q positive. If A is a C^*-algebra with 1, then there exists on any $P \in \mathcal{P}(A)$ an invertible form Q satisfying $Q(x,x) \geqslant 0$ for all $x \in P$ and in this case $\varphi : K_0(A) \to \mathrm{Witt}(A)$ is an isomorphism. However, φ is not an isomorphism in general for arbitrary involutive Banach algebras. See [17].

(xiii) Let F be a field and $\mathrm{Sym}\,B(F)$ the category of symmetric inner product spaces (V, β) – V a finite-dimensional vector space over F and $\beta : V \otimes V \to F$ a symmetric bilinear form. Then $(\mathrm{Sym}\,B(F), \perp)$ is a symmetric monoidal category where $(V, \beta) \perp (V', \beta')$ is the orthogonal sum of (V, β) and (V', β') defined as the vector space $V \oplus V'$ together with a bilinear form $\beta^* : (V \oplus V', V \oplus V') \to F$ given by $\beta^*(v \oplus v', v_1 \oplus v_1') = \beta(v, v_1) + \beta'(v', v_1')$.

If we define composition $(V, \beta) \odot (V', \beta')$ as the tensor product $V \otimes V'$ together with a bilinear form $\beta^*(v \otimes v', v_1 \otimes v_1') = \beta(v, v_1)\beta'(v', v_1')$, then $K_0(\mathrm{Sym}\,B(F), \perp, \odot)$ is a commutative ring with identity.

The Witt ring $W(F)$ is defined as the quotient of $K_0(\mathrm{Sym}\,B(F))$ by the subgroup $\{nH\}$ generated by the hyperbolic plane $H = \left(F^2, \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\right)$.

For more details about $W(F)$ see [80].

2. K_0 and class groups of Dedekind domains, orders and group-rings

2.1. *K_0 and class groups of Dedekind domains*

2.1.1. An integral domain R with quotient field F is called a Dedekind domain if it satisfies any of the following equivalent conditions

(i) Every ideal in R is projective (i.e. R is hereditary).

(ii) Every nonzero ideal $\mathbf{a}$ of R is invertible (that is $\mathbf{a}\mathbf{a}^{-1} = R$ where $\mathbf{a}^{-1} = \{x \in F \mid x\mathbf{a} \subset R\}$.

(iii) R is Noetherian, integrally closed and every nonzero prime ideal is maximal.
(iv) R is Noetherian, and $R_{\mathbf{m}}$ is a discrete valuation ring for all maximal ideals $\mathbf{m}$ of R.
(v) Every nonzero ideal is uniquely a product of prime ideals.

2.1.2. EXAMPLE. $\mathbb{Z}$, $F[x]$, are Dedekind domains. So is the ring of integers in a number field.

2.1.3. DEFINITION. A fractional ideal of a Dedekind domain R (with quotient field F) is an R-submodule $\mathbf{a}$ of F such that $s\mathbf{a} \subset R$ for some $s \neq 0$ in F. Then $\mathbf{a}^{-1} = \{x \in F \mid x\mathbf{a} \subset R\}$ is also a fractional ideal. Say that $\mathbf{a}$ is invertible if $\mathbf{a}\mathbf{a}^{-1} = R$. The invertible fractional ideals form a group which we denote by I_R. Also each element $u \in F^*$ determines a principal fractional ideal Ru. Let P_R be the subgroup of I_R consisting of all principal fractional ideals. The (ideal) class group of R is defined as I_R/P_R and denoted by $C\ell(R)$.

It is well known that if R is the ring of integers in a number field, then $C\ell(R)$ is finite see [18].

2.1.4. DEFINITION. Let R be a Dedekind domain with quotient field F. An R-lattice is a finitely generated torsion free R-module. Note that any R-lattice M is embeddable in a finite-dimensional F-vector space V such that $F \otimes_R M = V$. Moreover, every R-lattice M is R-projective (since R is hereditary and M can be written as a direct sum of ideals) (see 2.1.5 below – Steinitz's theorem). For $P \in \mathcal{P}(R)$ define the R-rank of P as the dimension of the vector space $F \otimes_R P$ and denote this number by $\mathrm{rk}(P)$.

2.1.5. THEOREM ([18], Steinitz theorem). *Let R be a Dedekind domain. Then*

(i) *If $M \in \mathcal{P}(R)$, then $M = \mathbf{a}_1 \oplus \mathbf{a}_2 \oplus \cdots \oplus \mathbf{a}_n$ where n is the R-rank of M and each $\mathbf{a}_i$ is an ideal of R.*
(ii) *Two direct sums $\mathbf{a}_1 \oplus \mathbf{a}_2 \oplus \cdots \oplus \mathbf{a}_n$ and $\mathbf{b}_1 \oplus \mathbf{b}_2 \oplus \cdots \oplus \mathbf{b}_n$ of nonzero ideals of R are R-isomorphic if and only if $n = m$ and the ideal class of $\mathbf{a}_1\mathbf{a}_2 \cdots \mathbf{a}_n$ = ideal class of $\mathbf{b}_1\mathbf{b}_2 \cdots \mathbf{b}_n$.*

2.1.6. DEFINITION. The ideal class associated to M as in 2.1.5, is called the Steinitz class and is denoted by $\mathrm{St}(M)$.

2.1.7. THEOREM. *Let R be a Dedekind domain. Then*

$$K_0(R) \simeq \mathbb{Z} \oplus C\ell(R).$$

SKETCH OF PROOF. Define a map

$$Q = (\mathrm{rk}, \mathrm{st}) : K_0(R) \to \mathbb{Z} \times C\ell(R)$$

by

$$(\mathrm{rk}, \mathrm{st})[P] = \big(\mathrm{rk}\, P, \mathrm{st}(P)\big),$$

where $\mathrm{rk}\, P$ is the R-rank of P (2.1.4) and $\mathrm{st}(P)$ is the Steinitz class of P. We have $\mathrm{rk}(P \oplus P^1) = \mathrm{rk}(P) + \mathrm{rk}(P^1)$ and $\mathrm{st}(P \oplus P^1) = \mathrm{st}(P) \cdot \mathrm{st}(P^1)$. So φ is a homomorphism that can easily be checked to be an isomorphism, the inverse being given by $\eta : \mathbb{Z} \times C\ell(R) \to K_0(R)$, $(n, (\mathbf{a})) \to n[R] + [\mathbf{a}]$.

2.1.8. REMARKS.

(i) It follows easily from Steinitz's theorem that $\mathrm{Pic}(R) \simeq C\ell(R)$ for any Dedekind domain R.

(ii) Let R be a commutative ring with identity, $\mathrm{Spec}(R)$ the set of prime ideals of R. For $P \in \mathcal{P}(R)$ define $r_P : \mathrm{Spec}(R) \to \mathbb{Z}$ by $r_P(\mathbf{p}) =$ rank of $P_{\mathbf{p}}$ over $R_{\mathbf{p}} =$ dimension of $P_{\mathbf{p}}/\mathbf{p}_{\mathbf{p}} P_{\mathbf{p}}$. Then r_P is continuous where $\mathbb{Z}$ is given the discrete topology (see [8] or [101]). Let $H_0(R) :=$ group of continuous functions $\mathrm{Spec}(R) \to \mathbb{Z}$. Then we have a homomorphism $r : K_0(R) \to H_0(R) : r([P]) = r_P$ (see [8]). One can show that if R is a one-dimensional commutative Noetherian ring then $(\mathrm{rk}, \det) : K_0(R) \to H_0(R) \oplus \mathrm{Pic}(R)$ is an isomorphism – a generalisation of 2.1.7 which we recover by seeing that for Dedekind domains R, $H_0(R) \simeq \mathbb{Z}$. Note that $\det : K_0(R) \to \mathrm{Pic}(R)$ is defined by $\det(P) = \Lambda^n P$ if the R-rank of P is n. (See [8].)

(iii) Since a Dedekind domain is a regular ring, $K_0(R) \simeq G_0(R)$.

2.2. *Class groups of orders and group rings*

2.2.1. DEFINITION. Let R be a Dedekind domain with quotient field F. An R-order Λ in a finite-dimensional semi-simple F-algebra Σ is a subring of Σ such that (i) R is contained in the centre of Λ, (ii) Λ is a finitely generated R-module, and (iii) $F \otimes_R \Lambda = \Sigma$.

EXAMPLE. For a finite group G, the group ring RG is an R-order in FG when $\mathrm{char}(F)$ does not divide $|G|$.

2.2.2. DEFINITION. Let R, F, Σ be as in 2.2.1. A maximal R-order Γ in Σ is an order that is not properly contained in any other R-order in Σ.

EXAMPLES. (i) R is a maximal R-order in F.

(ii) $M_n(R)$ is a maximal R-order in $M_n(F)$.

2.2.3. REMARKS. Let R, F, Σ be as in 3.2.1. Then

(i) Any R-order Λ is contained in at least one maximal R-order in Σ (see [18]).

(ii) Every semi-simple F-algebra Σ contains at least one maximal order. However, if Σ is commutative, then Σ contains a unique maximal order, namely, the integral closure of R in Σ (see [18] or [74]).

(iii) If Λ is an R-order in Σ, then $\Lambda_{\mathbf{p}}$ is an $R_{\mathbf{p}}$-order in Σ for any prime = maximal ideal $\mathbf{p}$ of R. Moreover, $\Lambda = \bigcap_{\mathbf{p}} \Lambda_{\mathbf{p}}$ (intersection within Σ).

(iv) In any R-order Λ, every element is integral over R (see [18] or [75]).

2.2.4. DEFINITION. Let R, F, Σ, Λ be as in 2.2.1. A left Λ-lattice is a left Λ-module which is also an R-lattice (i.e. finitely generated and projective as an R-module).

A Λ-ideal in Σ is a left Λ-lattice $M \subset \Sigma$ such that $FM \subset \Sigma$.

Two left Λ-lattices M, N are said to be in the same genus if $M_{\mathbf{p}} \simeq N_{\mathbf{p}}$ for each prime ideal $\mathbf{p}$ of R. A left Λ-ideal is said to be locally free if $M_{\mathbf{p}} \simeq \Lambda_{\mathbf{p}}$ for all $\mathbf{p} \in \mathrm{Spec}(R)$. We write $M \vee N$ if M and N are in the same genus.

2.2.5. DEFINITION. Let R, F, Σ be as in 2.2.1, Λ an R-order in Σ. Let $S(\Lambda) = \{\mathbf{p} \in \mathrm{Spec}(R) \mid \widehat{\Lambda}_{\mathbf{p}}$ is not a maximal $\widehat{R}_{\mathbf{p}}$-order in $\widehat{\Sigma}\}$. Then $S(\Lambda)$ is a finite set and $S(\Lambda) = \emptyset$ iff Λ is a maximal R-order. Note that the genus of a Λ-lattice M is determined by the isomorphism classes of modules $\{M_{\mathbf{p}} \mid \mathbf{p} \in S(\Lambda)\}$ (see [18] or [74]).

2.2.6. THEOREM. *Let L, M, N be lattices in the same genus. Then $M \oplus N \simeq L \oplus L'$ for some lattice L' in the same genus. Hence, if M, M' are locally free Λ-ideals in Σ, then $M \oplus M' = \Lambda \oplus M''$ for some locally free ideal M''.*

2.2.7. DEFINITION. Let R, F, Σ be as in 2.2.1. The idèle group of Σ, denoted $J(\Sigma)$ is defined by $J(\Sigma) := \{(\alpha_{\mathbf{p}}) \in \prod(\widehat{\Sigma}_{\mathbf{p}})^* \mid \alpha_{\mathbf{p}} \in \widehat{\Lambda}^*_{\mathbf{p}}$ almost everywhere$\}$. For $\alpha = (\alpha_{\mathbf{p}}) \in J(\Sigma)$, define

$$\Lambda\alpha := \Sigma \cap \left\{\bigcap_{\mathbf{p}} \widehat{\Lambda}_{\mathbf{p}}\alpha_{\mathbf{p}}\right\} = \bigcap_{\mathbf{p}}\{\Sigma \cap \widehat{\Lambda}_{\mathbf{p}}\alpha_{\mathbf{p}}\}.$$

The group of principal idèles, denoted $u(\Sigma)$ is defined by $u(\Sigma) = \{\alpha = (\alpha_{\mathbf{p}}) \mid \alpha_{\mathbf{p}} = x \in \Sigma^*$ for all $\mathbf{p} \in \mathrm{Spec}(R)\}$. The group of unit idèles is defined by

$$U(\Lambda) = \prod_{\mathbf{p}}(\Lambda_{\mathbf{p}})^* \subseteq J(\Sigma).$$

REMARKS.

(i) $J(\Sigma)$ is independent of the choice of the R-order Λ in Σ since if Λ' is another R-order, then $\Lambda_{\mathbf{p}} = \Lambda'_{\mathbf{p}}$ a.e.

(ii) $\Lambda\alpha$ is isomorphic to a left ideal of Λ and $\Lambda\alpha$ is in the same genus as Λ. Call $\Lambda\alpha$ a locally free (rank 1) Λ-lattice or a locally free fractional Λ-ideal in Σ. Note that any $M \in g(\Lambda)$ can be written in the form $M = \Lambda\alpha$ for some $\alpha \in J(\Sigma)$ (see [18]).

(iii) If $\Sigma = F$ and $\Lambda = R$, we also have $J(F), u(F)$ and $U(R)$ as defined above.

(iv) For $\alpha, \beta \in J(\Sigma)$, $\Lambda\alpha \oplus \Lambda\beta \cong \Lambda \oplus \Lambda\alpha\beta$ (see [18]).

2.2.8. DEFINITION. Let F, Σ, R, Λ be as in 2.2.1. Two left Λ-modules M, N are said to be stably isomorphic if $M \oplus \Lambda^{(k)} \simeq N \oplus \Lambda^{(k)}$ for some positive integer k. If F is a number field, then $M \oplus \Lambda^{(k)} \simeq N \oplus \Lambda^{(k)}$ iff $M \oplus \Lambda \simeq N \oplus \Lambda$. We write $[M]$ for the stable isomorphism class of M.

2.2.9. THEOREM [18]. *The stable isomorphism classes of locally free ideals form an Abelian group $C\ell(\Lambda)$ called the locally free class group of Λ where addition is given by $[M] + [M'] = [M'']$ whenever $M \oplus M' \simeq \Lambda \oplus M''$. The zero element is (Λ) and inverses exist since $(\Lambda\alpha) \oplus (\Lambda\alpha^{-1}) \simeq \Lambda \oplus \Lambda$ for any $\alpha \in J(\Sigma)$.*

2.2.10. THEOREM. *Let R, F, Λ, Σ be as in* 2.2.1. *If F is an algebraic number field, then $C\ell(\Lambda)$ is a finite group.*

PROOF (*Sketch*). If L is a left Λ-lattice, then there exists only a finite number of isomorphism classes of left Λ-lattices M such that $FM \simeq FL$ as Σ-modules. In particular, there exists only a finite number of isomorphism classes of left Λ ideals in Σ (see [18] or [75]). □

2.2.11. REMARKS. Let R, F, Λ, Σ be as in 2.2.1.

(i) If $\Lambda = R$, then $C\ell(\Lambda)$ is the ideal class group of R.

(ii) If Γ is a maximal R-order in Σ, then every left Λ-ideal in Σ is locally free. So, $C\ell(\Gamma)$ is the group of stable isomorphism classes of all left Γ-ideals in Σ.

(iii) Define a map $J(\Sigma) \to C\ell(\Lambda)$; $\alpha \to [\Lambda\alpha]$. Then one can show that this map is surjective and that the kernel is $J_0(\Sigma)\Sigma^* U(\Lambda)$ where $J_0(\Sigma)$ is the kernel of the reduced norm acting on $J(\Sigma)$. So $J(\Sigma)/(J_0(\Sigma)\Sigma^* U(\Lambda)) \simeq C\ell(\Lambda)$ (see [18]).

(iv) If G is a finite group such that no proper divisor of $|G|$ is a unit in R, then $C\ell(RG) \simeq SK_0(RG)$. Hence $C\ell(\mathbb{Z}G) \simeq SK_0(\mathbb{Z}G)$ for every finite group G (see [18] or [99]).

For computations of $C\ell(RG)$ for various R and G see [18].

2.2.12. *An application – the Wall finiteness obstruction theorem.* Let R be a ring. A bounded chain complex $C = (C_*, d)$ of R-modules is said to be of finite type if all the C_j's are finitely generated. The Euler characteristic of $C = (C_*, d)$ is given by: $\chi(C) = \sum_{i=-\infty}^{\infty}(-1)^i[C_i]$, and we write $\bar{\chi}(C)$ for the image of $\chi(C)$ in $\widetilde{K}_0(R)$.

The initial motivation for Wall's finiteness obstruction theorem stated below was the desire to find out when a connected space has the homotopy type of a CW-complex. If X is homotopically equivalent to a CW-complex, the singular chain complex $S_*(X)$ with local coefficients is said to be finitely dominated if it is chain homotopic to a complex of finite type. Let $R = \mathbb{Z}\pi_1(X)$, the integral group-ring of the fundamental group of X. Wall's finite obstruction theorem stated below implies that a finitely dominated complex has a finiteness obstruction in $\widetilde{K}_0(R)$ and is chain homotopic to a complex of finite type of free R-modules if and only if the finiteness obstruction vanishes. More precisely we have the following

THEOREM [114]. *Let (C_*, d) be a chain complex of projective R-modules which is homotopic to a chain complex of finite type of projective R-modules. Then (C_*, d) is chain homotopic to a chain complex of finite type of free R-modules if and only if $\bar{\chi}(C) = 0$ in $\widetilde{K}_0(R)$.*

NOTE. For further applications in this direction see [115,105,85].

3. K_0 of exact and Abelian categories – definitions and examples

3.1. *K_0 of exact categories and examples*

3.1.1. DEFINITION. An exact category is an additive category $\mathcal{C}$ embeddable as a full subcategory of an Abelian category $\mathcal{A}$ such that $\mathcal{C}$ is equipped with a class $\mathcal{E}$ of short exact sequences $0 \to M' \to M \to M'' \to 0$ (I) satisfying

(i) $\mathcal{E}$ is a class of all sequences (I) in $\mathcal{C}$ that are exact in $\mathcal{A}$.
(ii) $\mathcal{E}$ is closed under extensions in $\mathcal{A}$, i.e. if (I) is an exact sequence in $\mathcal{A}$ and $M', M'' \in \mathcal{C}$, then $M \in \mathcal{C}$.

3.1.2. DEFINITION. For a small exact category $\mathcal{C}$, define the Grothendieck group $K_0(\mathcal{C})$ of $\mathcal{C}$ as the Abelian group generated by isomorphism classes (C) of $\mathcal{C}$-objects subject to the relation $(C') + (C'') = (C)$ whenever $0 \to C' \to C \to C'' \to 0$ is an exact sequence in $\mathcal{C}$.

3.1.3. REMARKS.
(i) $K_0(\mathcal{C}) \simeq \mathcal{F}/\mathcal{R}$ where $\mathcal{F}$ is the free Abelian group on the isomorphism classes (C) of $\mathcal{C}$-objects and $\mathcal{R}$ the subgroup of $\mathcal{F}$ generated by all $(C) - (C') - (C'')$ for each exact sequence $0 \to C' \to C \to C'' \to 0$ in $\mathcal{C}$. Denote by $[C]$ the class of (C) in $K_0(\mathcal{C}) = \mathcal{F}/\mathcal{R}$.
(ii) The construction satisfies the following property: If $\chi : \mathcal{C} \to A$ is a map from $\mathcal{C}$ to an Abelian group A given that $\chi(C)$ depends only on the isomorphism class of $\mathcal{C}$ and $\chi(C) = \chi(C') + \chi(C'')$ for any exact sequence $0 \to C' \to C \to C'' \to 0$, then there exists a unique $\chi' : K_0(\mathcal{C}) \to A$ such that $\chi(C) = \chi'([C])$ for any $\mathcal{C}$-object C.
(iii) Let $F : \mathcal{C} \to \mathcal{D}$ be an exact functor between two exact categories $\mathcal{C}, \mathcal{D}$ (i.e. F is additive and takes short exact sequences in $\mathcal{C}$ to such sequences in $\mathcal{D}$). Then F induces a group homomorphism $K_0(\mathcal{C}) \to K_0(\mathcal{D})$.
(iv) Note that an Abelian category $\mathcal{A}$ is also an exact category and the definition of $K_0(\mathcal{A})$ is the same as in 2.1.2.

3.1.4. EXAMPLES.
(i) Any additive category is an exact category as well as a symmetric monoidal category under '$\oplus$', and $K_0(\mathcal{C})$ is a quotient of the group $K_0^{\oplus}(\mathcal{C})$ defined in 1.4.1.

If every short exact sequence in $\mathcal{C}$ splits, then $K_0(\mathcal{C}) = K_0^{\oplus}(\mathcal{C})$. For example, $K_0(\Lambda) = K_0(\mathcal{P}(\Lambda)) = K_0^{\oplus}(\mathcal{P}(\Lambda))$ for any ring Λ with identity.
(ii) Let Λ be a (left) Noetherian ring. Then the category $\mathcal{M}(\Lambda)$ of finitely generated (left)-Λ-modules is an exact category and we denote $K_0(\mathcal{M}(\Lambda))$ by $G_0(\Lambda)$. The inclusion functor $\mathcal{P}(\Lambda) \to \mathcal{M}(\Lambda)$ induces a map $K_0(\Lambda) \to G_0(\Lambda)$ called the Cartan map. For example, $\Lambda = RG$, R a Dedekind domain, G a finite group, yields a Cartan map $K_0(RG) \to G_0(RG)$.

If Λ is left Artinian, then $G_0(\Lambda)$ is free Abelian on $[S_1], \dots, [S_r]$ where the S_i are distinct classes of simple Λ-modules while $K_0(\Lambda)$ is free Abelian on $[I_1], \dots, [I_\ell]$ and the I_i are distinct classes of indecomposable projective Λ-modules (see [18]). So, the map $K_0(\Lambda) \to G_0(\Lambda)$ gives a matrix a_{ij} where $a_{ij} =$ the number of times S_j occurs in a composition series for I_i. This matrix is known as the Cartan matrix.

If Λ is left regular (i.e. every finitely generated left Λ-module has finite resolution by finitely generated projective left Λ-modules), then it is well known that the Cartan map is an isomorphism (see [18]).

For example, if R is a Dedekind domain with quotient field F and Λ is a maximal R-order in a semi-simple F-algebra, Σ, then $K_0(\Lambda) \simeq G_0(\Lambda)$ since Λ is regular. (See [18] or [25] for further information on Cartan maps.)

(iii) Let R be a commutative ring with identity, Λ an R-algebra. Let $\mathcal{P}_R(\Lambda)$ be the category of left Λ-lattices, i.e. Λ-modules which are finitely generated and projective as R-modules. Then $\mathcal{P}_R(\Lambda)$ is an exact category and we write $G_0(R,\Lambda)$ for $K_0(\mathcal{P}_R(\Lambda))$. If $\Lambda = RG$, G a finite group, we write $\mathcal{P}_R(G)$ for $\mathcal{P}_R(RG)$ and also write $G_0(R,G)$ for $G_0(R,RG)$. If $M, N \in \mathcal{P}_R(\Lambda)$, then, so is $M \otimes_R N$ and hence the multiplication given in $G_0(R,G)$ by $[M][N] = (M \otimes_R N)$ makes $G_0(R,G)$ a commutative ring with identity.

(iv) If R is a commutative regular ring and Λ is an R-algebra that is finitely generated and projective as an R-module (e.g., $\Lambda = RG$, G a finite group or R is a Dedekind domain with quotient field F and Λ is an R-order in a semi-simple F-algebra) then $G_0(R,\Lambda) \simeq G_0(\Lambda)$.

SKETCH OF PROOF. Define a map $\varphi : G_0(R,\Lambda) \to G_0(\Lambda)$ by $\varphi[M] = [M]$. Then φ is a well defined homomorphism. Now for $M \in \mathcal{M}(\Lambda)$, there exists an exact sequence $0 \to L \to P_{n-1} \xrightarrow{\varphi_{n-1}} P_{n-2} \to \cdots \to P_0 \to M \to 0$ where $P_i \in \mathcal{P}(\Lambda)$, $L \in \mathcal{M}(\Lambda)$. Now, since $\Lambda \in \mathcal{P}(R)$, each $P_i \in \mathcal{P}(R)$ and hence $L \in \mathcal{P}(R)$. So $L \in \mathcal{P}_R(\Lambda)$. Now define $\delta[M] = [P_0] - [P_1] + \cdots + (-1)^{n-1}[P_{n-1}] + (-1)^n[L] \in G_0(R,\Lambda)$. One easily checks that $\delta f = 1 = f\delta$. $\square$

(v) Let X be a scheme (see [33]), $\mathcal{P}(X)$ the category of locally free sheaves of O_X-modules of finite rank (or equivalently the category of finite-dimensional (algebraic) vector bundles on X). Then $\mathcal{P}(X)$ is an exact category and we write $K_0(X)$ for $K_0(\mathcal{P}(X))$ (see [70]).

If $X = \operatorname{Spec}(A)$ for some commutative ring A, then we have an equivalence of categories $\mathcal{P}(X) \to \mathcal{P}(A) : E \to \Gamma(X,E) = \{A\text{-module of global sections}\}$, with the inverse equivalence $\mathcal{P}(A) \to \mathcal{P}(X)$ given by $P \to \widetilde{P} : U \to O_X(U) \otimes_A P$. Hence $K_0(X) \simeq K_0(A)$.

(vi) Let X be a Noetherian scheme (i.e. X can be covered by affine open sets $\operatorname{Spec}(A_i)$ where each A_i is Noetherian), then the category $\mathcal{M}(X)$ of coherent sheaves of O_X-modules is exact. We write $G_0(X)$ for $K_0(\mathcal{M}(X))$. If $X = \operatorname{Spec}(A)$ then we have an equivalence of categories $\mathcal{M}(X) \simeq \mathcal{M}(A)$ and $G_0(X) \simeq G_0(A)$.

(vii) Let G be a finite group, S a G-set, $\mathbf{S}$ the category associated to S (see 1.4.3(xi)), $\mathcal{C}$ an exact category, and $[\mathbf{S},\mathcal{C}]$ the category of covariant functors $\zeta : \mathbf{S} \to \mathcal{C}$. We write ζ_s for $\zeta(s)$, $s \in S$. Then $[\mathbf{S},\mathcal{C}]$ is an exact category where a sequence $0 \to \zeta' \to \zeta \to \zeta'' \to 0$ in $[\mathbf{S},\mathcal{C}]$ is defined to be exact if $0 \to \zeta'_s \to \zeta_s \to \zeta''_s \to 0$ is exact in $\mathcal{C}$ for all $s \in S$. Denote by $K_0^G(S,\mathcal{C})$ the K_0 of $[\mathbf{S},\mathcal{C}]$. Then $K_0^G(-,\mathcal{C}) : \mathbf{GSet} \to \mathbf{Ab}$ is a functor that can be proved to be a 'Mackey functor' (see [24] or [48]).

It can also be shown (see [48] or [47]) that if $S = G/G$, the $[G/G,\mathcal{C}] \simeq \mathcal{C}_G$ in the notation of 1.4.3(x). Also, constructions analogous to the one above can be done for G a profinite group (see [46]) or compact Lie groups [52].

Now if R is a commutative Noetherian ring with identity, we have $[G/G, \mathbf{P}(R)] \simeq \mathcal{P}(R)_G \simeq \mathcal{P}_R(RG)$ (see [48] or [47]), and so, $K_0^G[G/G, \mathcal{P}(R)) \simeq K_0(\mathcal{P}(R)_G) \simeq G_0(R, G)$ and that if R is regular $K_0(\mathcal{P}(R)_G) \simeq G_0(R, G) \simeq G_0(RG)$. This provides an initial connection between K-theory of representations of G in $\mathcal{P}(R)$ and K-theory of the group ring RG.

In particular, when $R = \mathbb{C}, \mathcal{P}(\mathbb{C}) = \mathcal{M}(\mathbb{C})$ and $K_0(\mathcal{P}(\mathbb{C})_G) \simeq G_0(\mathbb{C}, G) = G_0(\mathbb{C}G) =$ the Abelian group of characters $\chi : G \to \mathbb{C}$ (see [18]), as already observed in §1.

(viii) Let X be a compact topological space and $F = \mathbb{R}$ or $\mathbb{C}$. Then the category $VB_F(X)$ of vector bundles over X is an exact category. We had earlier observed (see §1) that $VB_F(X)$ is also a symmetric monoidal category. Since every short exact sequence in $VB_F(X)$ splits, we have $K_0(VB_F(X)) \simeq K_0^{\oplus}(VB_F(X))$.

4. Some fundamental results on K_0 of exact and Abelian categories

In this section, we discuss some of the results that will be seen in more generality when higher K-groups are treated in a forthcoming chapter.

4.1. *Devissage theorem*

4.1.1. DEFINITION. Let $\mathcal{C}_0 \subset \mathcal{C}$ be exact categories. The inclusion functor $\mathcal{C}_0 \to \mathcal{C}$ is exact and hence induces a homomorphism $K_0(\mathcal{C}_0) \to K_0(\mathcal{C})$. A $\mathcal{C}_0$-filtration of an object A in $\mathcal{C}$ is a finite sequence of the form: $0 = A_0 \subset A_1 \subset \cdots \subset A_n = A$ where each $A_i/A_{i-1} \in \mathcal{C}_0$.

4.1.2. LEMMA. *If $0 \subset A_0 \subset A_1 \subset \cdots \subset A_n = A$ is a $\mathcal{C}_0$-filtration, then $[A] = \Sigma[A_i/A_{i-1}]$, $1 \leqslant i \leqslant n$, in $K_0(\mathcal{C})$.*

4.1.3. THEOREM (Devissage theorem). *Let $\mathcal{C}_0 \subset \mathcal{C}$ be exact categories such that $\mathcal{C}_0$ is Abelian. If every $A \in \mathcal{C}$ has a $\mathcal{C}_0$-filtration, then $K_0(\mathcal{C}_0) \to K_0(\mathcal{C})$ is an isomorphism.*

PROOF. Since $\mathcal{C}_0$ is Abelian, any refinement of a $\mathcal{C}_0$-filtration is also a $\mathcal{C}_0$-filtration. So, by the Zassenhaus lemma, any two finite filtrations have equivalent refinements, that is, refinements such that the successive factors of the first refinement are, up to a permutation of the order of their occurrences, isomorphic to those of the second.

So, if $0 \subset A_0 \subset A_1 \subset \cdots \subset A_n = A$ is any $\mathcal{C}_0$-filtration of A in $\mathcal{C}$, then

$$J(A) = \Sigma[A_i/A_{i-1}] \quad (1 \leqslant i \leqslant n)$$

is well defined, since $J(A)$ is unaltered by replacing the given filtration with a refinement.

Now let $0 \to A' \xrightarrow{\alpha} A \xrightarrow{\beta} A'' \to 0$ be an exact sequence in $\mathcal{C}$. Obtain a filtration for A by $0 = A_0 \subset A_1 \subset \cdots \subset A_n = A'$ for A' and $\beta^{-1}(A^0) \subset \beta^{-1}(A^1) \subset \cdots \subset \beta^{-1}(A'')$ if $A^0 \subset A^1 \subset \cdots \subset A''$ is a $\mathcal{C}_0$-filtration of A''. Then $0 = A_0 \subset A_1 \subset \cdots \subset A_k \subset \beta^{-1}(A^0) \subset \beta^{-1}(A^1) \subset \cdots \subset \beta^{-1}(A^n)$ is a filtration of A.

So, $J(A) = J(A') + J(A'')$. Hence J induces a homomorphism $K_0(\mathcal{C}) \to K_0(\mathcal{C}_0)$. We also have a homomorphism $i : K_0(\mathcal{C}_0) \to K_0(\mathcal{C})$ induced by the inclusion functor $i : \mathcal{C}_0 \to \mathcal{C}$. Moreover, $i \circ J = 1_{K_0(\mathcal{C})}$ and $J \circ i = 1_{K_0(\mathcal{C})}$. Hence $K_0(\mathcal{C}_0) \simeq K_0(\mathcal{C})$. □

4.1.4. COROLLARY. *Let* **a** *be a nilpotent two-sided ideal of a Noetherian ring* R. *Then* $G_0(R/\mathbf{a}) \simeq G_0(R)$.

PROOF. If $M \in \mathcal{M}(R)$, then $M \supset \mathbf{a}M \supset \cdots \supset \mathbf{a}^k M = 0$ is an $\mathcal{M}(R/\mathbf{a})$ filtration of $\mathcal{M}$. Result follows from 4.1.3. □

4.1.5. EXAMPLES.

(i) Let R be an Artinian ring with maximal ideal $\mathbf{m}$ such that $\mathbf{m}^r = 0$ for some r. Let $k = R/\mathbf{m}$ (e.g., $R = Z/p^r$, $k = F_p$).

In 4.1.3, put $\mathcal{C}_0 =$ category of finite-dimensional k-vector spaces and $\mathcal{C}$, the category of finitely generated R-modules. Then, we have a filtration

$$0 = \mathbf{m}^r M \subset \mathbf{m}^{r-1} M \subset \cdots \subset \mathbf{m}M \subset M \quad \text{of } M,$$

where $M \in \operatorname{ob}\mathcal{C}$. Hence by 4.1.3, $K_0(\mathcal{C}_0) \simeq K_0(\mathcal{C})$.

(ii) Let X be a Noetherian scheme, $\mathcal{M}(X)$ the category of coherent sheaves of O_X-modules, $i : Z \subset X$ the inclusion of a closed subscheme. Then $\mathcal{M}(Z)$ becomes an Abelian subcategory of $\mathcal{M}(X)$ via the direct image $i : \mathcal{M}(Z) \subset \mathcal{M}(X)$. Let $\mathcal{M}_Z(X)$ be the Abelian category of coherent sheaves of O_X-modules supported on Z, **a** an ideal sheaf in O_X such that $O_X/\mathbf{a} \simeq O_Z$. Then every $M \in \mathcal{M}_Z(X)$ has a finite filtration $M \supset M\mathbf{a} \supset M\mathbf{a}^2 \supset \cdots$ and so, by devissage $K_0(\mathcal{M}_Z(X) \simeq K_0(\mathcal{M}(Z) \simeq G_0(Z))$. See §4.3.4 for more examples of applications of devissage.

4.2. *Resolution theorem and examples*

4.2.1. RESOLUTION THEOREM [8,67]. *Let* $\mathcal{A}_0 \subset \mathcal{A}$ *be an inclusion of exact categories. Suppose that every object of* $\mathcal{A}$ *has a finite resolution by objects of* $\mathcal{A}_0$ *and that if* $0 \to M' \to M \to M'' \to 0$ *is an exact sequence in* $\mathcal{A}$, *then* $M \in \mathcal{A}_0$ *implies that* $M', M'' \in \mathcal{A}_0$. *Then* $K_0(\mathcal{A}_0) \simeq K_0(\mathcal{A})$.

4.2.2. EXAMPLES. (i) Let R be a regular ring. Then, for any $M \in \operatorname{ob}\mathcal{M}(R)$, there exists $P_i \in \mathcal{P}(R)$, $i = 0, 1, \ldots, n$, such that the sequence $0 \to P_n \to P_{n-1} \to \cdots \to M \to 0$ is exact. Put $\mathcal{A}_0 = \mathcal{P}(R)$, $\mathcal{A} = \mathcal{M}(R)$ in 4.6. Then we have $K_0(R) \simeq G_0(R)$ (see [8]).

(ii) Let $\mathbf{H}(R)$ be the category of all R-modules having finite homological dimension, i.e. having a finite resolution by finitely generated projective R-modules: $\mathbf{H}_n(R)$ the subcategory of modules having resolutions of length $\leqslant n$. Then by the resolution theorem 4.2.1, applied to $\mathbf{P}(R) \subseteq \mathbf{H}(R)$ we have $K_0(R) \simeq K_0\mathbf{H}(R) \simeq K_0\mathbf{H}_n(R)$ for all $n \geqslant 1$ (see [8] or [101]).

(iii) Let $\mathcal{C}$ be an exact category and $\operatorname{Nil}(\mathcal{C})$ the category whose objects are pairs (M, ν) where $M \in \mathcal{C}$ and ν is a nilpotent endomorphism of M, i.e. $\nu \in \operatorname{End}_{\mathcal{C}}(M)$. Let $\mathcal{C}_0 \subset \mathcal{C}$ be an exact subcategory of $\mathcal{C}$ such that every object of $\mathcal{C}$ has a finite $\mathcal{C}_0$-resolution. Then every object of $\mathrm{Nil}(\mathcal{C})$ has a finite $\mathrm{Nil}(\mathcal{C}_0)$-resolution and so, by 4.2.1, $K_0(\mathrm{Nil}(\mathcal{C}_0)) \simeq K_0(\mathrm{Nil}(\mathcal{C}))$.

(iv) In the notation of (iii), we have two functors $Z : \mathcal{C} \to \mathrm{Nil}(\mathcal{C}) : Z(M) = (M, 0)$ (where '0' denotes zero endomorphism) and $F : \mathrm{Nil}(\mathcal{C}) \to \mathcal{C} : F(M, \nu) = M$ satisfying $FZ = 1_{\mathcal{C}}$ and hence a split exact sequence $0 \to K_0(\mathcal{C}) \xrightarrow{Z} K_0(\mathrm{Nil}(\mathcal{C})) \to \mathrm{Nil}_0(\mathcal{C}) \to 0$ which defines $\mathrm{Nil}_0(\mathcal{C})$ as cokernel of Z.

If Λ is a ring, and $\mathbf{H}(\Lambda)$ is the category defined in (ii) above, then we denote $\mathrm{Nil}_0(\mathcal{P}(\Lambda))$ by $\mathrm{Nil}_0(\Lambda)$. If S is a central multiplicative system in Λ, $\mathbf{H}_S(\Lambda)$ the category of S-torsion objects of $\mathbf{H}(\Lambda)$ and $\mathcal{M}_S(\Lambda)$ the category of finitely generated S-torsion Λ-modules, one can show that if $S = T_+ = \{t^i\}$ – a free Abelian monoid on one generator t, then there exists isomorphisms of categories $\mathcal{M}_{T_+}(\Lambda[t]) \simeq \mathrm{Nil}(\mathcal{M}(\Lambda))$ and $\mathbf{H}_{T_+}(\Lambda[t]) \simeq \mathrm{Nil}(\mathbf{H}(\Lambda))$ and an isomorphism of groups: $K_0(\mathbf{H}_{T_+}(\Lambda[t])) \simeq K_0(\Lambda) \oplus \mathrm{Nil}_0(\Lambda)$. Hence $K_0(\mathrm{Nil}\,\mathbf{H}(\Lambda)) \simeq K_0(\Lambda) \oplus \mathrm{Nil}_0(\Lambda)$. See [8,101,67,32] for further information.

(v) The fundamental theorem for K_0 says that:

$$K_0\big(\Lambda[t, t^{-1}]\big) \simeq K_0(\Lambda) \oplus K_{-1}(\Lambda) \oplus NK_0(\Lambda) \oplus NK_0(\Lambda),$$

where $NK_0(\Lambda) := \mathrm{Ker}(K_0(\Lambda[t] \xrightarrow{\tau_+} K_0(\Lambda))$ where τ_+ is induced by augmentation $t = 1$, and K_{-1} is the negative K-functor $K_{-1} : \mathbf{Rings} \to$ Abelian groups defined by H. Bass in [8]. For generalisation of this fundamental theorem to higher K-theory, see [67].

4.3. K_0 *and localisation in Abelian categories*

We close this section with a discussion leading to a localisation short exact sequence 4.3.2 and then give copious examples to illustrate the use of the sequence.

4.3.1. A full subcategory $\mathcal{B}$ of an Abelian category $\mathcal{A}$ is called a Serre subcategory if whenever $0 \to M' \to M \to M'' \to 0$ is an exact sequence in $\mathcal{A}$, then $M \in \mathcal{B}$ if and only if $M', M'' \in \mathcal{B}$. We now construct a quotient Abelian category $\mathcal{A}/\mathcal{B}$ whose objects are just objects of $\mathcal{A}$. $\mathrm{Hom}_{\mathcal{A}/\mathcal{B}}(M, N)$ is defined as follows: If $M' \subseteq M$, $N' \subseteq N$ are subobjects such that $M/M' \in \mathrm{ob}(\mathcal{B})$, $N' \in \mathrm{ob}(\mathcal{B})$, then there exists a natural homomorphism $\mathrm{Hom}_{\mathcal{A}}(M, N) \to \mathrm{Hom}_{\mathcal{A}}(M', N/N')$. As M', N' range over such pairs of objects, the groups $\mathrm{Hom}_{\mathcal{A}}(M', N/N')$ form a direct system of Abelian groups and we define $\mathcal{A}/\mathcal{B}(M, N) = \varinjlim_{(M',N')} \mathcal{A}(M', N/N')$.

The quotient functor $T : \mathcal{A} \to \mathcal{A}/\mathcal{B}$ defined by $M \to T(M) = M$ is such that

(i) $T : \mathcal{A} \to \mathcal{A}/\mathcal{B}$ is an additive functor.

(ii) If $\mu \in \mathrm{Hom}_{\mathcal{A}}(M, N)$, then $T(\mu)$ is null if and only if $\mathrm{Im}(\mu) \in \mathrm{ob}(\mathcal{B})$. Also $T(\mu)$ is epimorphism if and only if $\mathrm{coker}\,\mu \in \mathrm{ob}(\mathcal{B})$ and it is a monomorphism iff $\mathrm{Ker}(\mu) \in \mathrm{ob}(\mathcal{B})$. Hence $T(\mu)$ is an isomorphism if and only if μ is a $\mathcal{B}$-isomorphism.

4.3.2. REMARKS. Note that $\mathcal{A}/\mathcal{B}$ satisfies the following universal property: If $T' : \mathcal{A} \to \mathcal{D}$ is an exact functor such that $T'(M) \simeq 0$ for all $M \in \mathcal{B}$, then there exists a unique exact functor $U : \mathcal{A}/\mathcal{B} \to \mathcal{D}$ such that $T' = U \circ T$.

4.3.3. THEOREM [8,35]. *Let $\mathcal{B}$ be a Serre subcategory of an Abelian category $\mathcal{A}$. Then there exists an exact sequence*

$$K_0(\mathcal{B}) \to K_0(\mathcal{A}) \to K_0(\mathcal{A}/\mathcal{B}) \to 0.$$

4.3.4. EXAMPLES.

(i) Let Λ be a Noetherian ring, $S \subset \Lambda$ a central multiplicative subset of Λ, $\mathcal{M}_S(\Lambda)$ the category of finitely generated S-torsion Λ-modules. Then $\mathcal{M}(\Lambda)/\mathcal{M}_S(\Lambda) \simeq \mathcal{M}(\Lambda_S)$ see [8,35,101] and so the exact sequence in 4.3.3 becomes

$$K_0\bigl(\mathcal{M}_S(\Lambda)\bigr) \to G_0(\Lambda) \to G_0(\Lambda_S) \to 0. \tag{I}$$

(ii) If Λ in (i) is a Dedekind domain R with quotient field F, and $S = R - 0$, then $K_0(\mathcal{M}_S(R)) \simeq \bigoplus_{\mathbf{m}} G_0(R/\mathbf{m}) = \bigoplus_m K_0(R/\mathbf{m})$ where $\mathbf{m}$ runs through the maximal ideals of R. Now, since $K_0(R/\mathbf{m}) \simeq \mathbb{Z}$ and $K_0(R) \simeq \mathbb{Z} \oplus C\ell(R)$ the sequence (I) yields the exactness of

$$\bigoplus \mathbb{Z} \to \mathbb{Z} \oplus C\ell(R) \to \mathbb{Z} \to 0.$$

(iii) Let Λ be a Noetherian ring, $S = \{s^i\}$ for some $s \in S$. Then $K_0(\mathcal{M}_S(R)) \simeq G_0(R/sR)$ (by devissage) yielding the exact sequence

$$G_0(\Lambda/s\Lambda) \to G_0(\Lambda) \to G_0\left(\Lambda\left(\frac{1}{s}\right)\right) \to 0.$$

(iv) Let R be the ring of integers in a p-adic field F, Γ a maximal R-order in a semi-simple F-algebra Σ, $S = R - 0$, then $K_0(\mathcal{M}_S(\Gamma)) \simeq G_0(\Gamma/\pi\Gamma) \simeq K_0(\Gamma/\operatorname{rad}\Gamma)$ (see [18] or [42]) where πR is the unique maximal ideal of R.

(v) If R is the ring of integers in a number field F, Λ an R-order in a semi-simple F-algebra Σ, let $S = R - 0$. Then $K_0(\mathcal{M}_S(\Lambda)) \simeq \bigoplus G_0(\Lambda/p\Lambda)$ (see [18] or [45]) where p runs through all the prime ideals of R.

(vi) Let X be a Noetherian scheme, U an open subscheme of X, $Z = X - U$, let $\mathcal{A} = \mathcal{M}(X)$ the category of coherent (sheaves of) O_X-modules, $\mathcal{B}$ the category of O_X-modules whose restriction to U is zero (i.e. the category of coherent modules supported on Z). Then $\mathcal{A}/\mathcal{B}$ is the category of coherent O_U-modules and so, (I) becomes $G_0(Z) \to G_0(X) \to G_0(U) \to 0$ (see 4.1.5(ii) or [67]).

(vii) Let Λ be a (left) Noetherian ring, $\Lambda[t]$ the polynomial ring in the variable t, $\Lambda[t, t^{-1}]$ the Laurent polynomial ring. Then $\Lambda[t, t^{-1}] = \Lambda[t]_S$ where $S = \{t^i\}$. Now, the map $\varepsilon : \Lambda[t] \to \Lambda,\ t \to 0$ induces an inclusion $\mathcal{M}(\Lambda) \subset \mathcal{M}(\Lambda[t])$ and the canonical map $i : \Lambda[t] \to \Lambda[t]_S = \Lambda[t, t^{-1}],\ t \to t/1$, yields an exact functor $\mathcal{M}(\Lambda[t]) \to \mathcal{M}(\Lambda[t, t^{-1}])$. So from 4.3.3, we have the localisation sequence

$$G_0(\Lambda) \xrightarrow{\varepsilon_*} G_0\bigl(\Lambda[t]\bigr) \to G_0\bigl(\Lambda[t, t^{-1}]\bigr) \to 0. \tag{II}$$

Now $\varepsilon_* = 0$ since for any Λ, the exact sequence of $\Lambda[t]$-modules $0 \to N[t] \xrightarrow{t} N[t] \to N \to 0$ yields

$$\varepsilon_*[N] = \big[N[t]\big] - \big[N[t]\big] = 0.$$

So, $G_0(\Lambda[t]) \simeq G_0(\Lambda[t, t^{-1}])$ from (II) above. This proves the first part of the fundamental theorem for G_0 of rings 4.3.5 below.

4.3.5. THEOREM (Fundamental theorem for G_0 of rings). *If Λ is a left Noetherian ring, then the inclusions $\Lambda \overset{i}{\hookrightarrow} \Lambda[t] \overset{j}{\hookrightarrow} \Lambda[t, t^{-1}]$ induce isomorphisms*

$$G_0(\Lambda) \cong G_0\big(\Lambda[t]\big) \cong G_0\big(\Lambda[t, t^{-1}]\big).$$

PROOF. See [8] or [101] for the proof of the second part. □

4.3.6. REMARKS. (i) The fundamental Theorem 4.3.5 above can be generalised to schemes (see [67]). If X is a scheme, write $X[s]$ for $X \times \mathrm{Spec}(\mathbb{Z}[s])$ and $X[s, s^{-1}]$ for $X \times \mathrm{Spec}(\mathbb{Z}[s, s^{-1}])$. When X is Noetherian, the map $\varepsilon : X \to X[s]$ defined by $s = 0$ induces an inclusion $\mathcal{M}(X) \subset \mathcal{M}(X[s])$ and hence a transfer map $\varepsilon_* : G_0(X) \to G_0(X[s])$. So we have a localisation exact sequence

$$G_0(X) \xrightarrow{\varepsilon_*} G_0\big(X[s]\big) \to G_0\big(X[s, s^{-1}]\big) \to 0.$$

We also have a fundamental theorem similar to 4.3.5 as follows

4.3.7. THEOREM (Fundamental theorem for G_0 of schemes). *If X is a Noetherian scheme, then the flat maps $X[s, s^{-1}] \overset{j}{\hookrightarrow} X[s] \overset{i}{\hookrightarrow} X$ induce isomorphisms $G_0(X) \cong G_0(X[s]) \cong G_0(X[s, s^{-1}])$.*

4.3.8. REMARKS.

(i) If we put $X = \mathrm{Spec}(\Lambda)$ in 4.3.7, Λ is Noetherian ring, we recover 4.3.5.

(ii) For all $n \geqslant 0$, there are fundamental theorems for G_n of rings and schemes (see [67] or [86]) and these will be discussed in a forthcoming chapter on higher K-theory.

(iii) There is a generalisation of 4.3.5 due to A. Grothendieck as follows: Let R be a commutative Noetherian ring, Λ a finite R-algebra, T a free Abelian group or monoid with a finite basis. Then $G_0(\Lambda) \to G_0(\Lambda[T])$ is an isomorphism, see [8].

(iv) If Λ is a (left) Noetherian regular ring, so are $\Lambda[t]$ and $\Lambda[t, t^{-1}]$. Since $K_0(R) \cong G_0(R)$ for any Noetherian regular ring R, we have from 4.3.5 that $K_0(\Lambda) \simeq K_0(\Lambda[t]) \simeq K_0\Lambda[t, t^{-1}]$. Furthermore, if T is a free Abelian group or monoid with a finite basis, then $K_0(A) \to K_0(\Lambda[T])$ is an isomorphism (see [8]).

5. K_1 of rings

5.1. *Definitions and basic properties*

5.1.1. Let R be a ring with identity, $GL_n(R)$ the group of invertible $n \times n$ matrices over R. Note that $GL_n(R) \subset GL_{n+1}(R)$: $A \mapsto \left(\begin{smallmatrix} A & 0 \\ 0 & 1 \end{smallmatrix}\right)$. Put $GL(R) = \varinjlim GL_n(R) = \bigcup_{n=1}^{\infty} GL_n(R)$. Let $E_n(R)$ be the subgroup of $GL_n(R)$ generated by elementary matrices $e_{ij}(a)$ where $e_{ij}(a)$ is the $n \times n$ matrix with 1's along the diagonal, a in the (i, j)-position and zeros elsewhere. Put $E(R) = \varinjlim E_n(R)$.

NOTE. The $e_{ij}(a)$ satisfy the following.

(i) $e_{ij}(a)e_{ij}(b) = e_{ij}(a + b)$ for all $a, b \in R$.

(ii) $[e_{ij}(a), e_{jk}(b)] = e_{ik}(ab)$ for all $i \neq k,\ a, b \in R$.

(iii) $[e_{ij}(a), e_{k\ell}(b)] = 1$ for $j \neq i \neq \ell \neq k$.

5.1.2. LEMMA. *If $A \in GL_n(R)$, then $\left(\begin{smallmatrix} A & 0 \\ 0 & A^{-1} \end{smallmatrix}\right) \in E_{2n}(R)$.*

PROOF. First observe that for any $C \in M_n(R)$, $\left(\begin{smallmatrix} I_n & 0 \\ C & I_n \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} I_n & C \\ 0 & I_n \end{smallmatrix}\right)$ are in $E_{2n}(R)$, where I_n is the identity $n \times n$ matrix. Hence

$$\begin{pmatrix} A & 0 \\ 0 & A^{-1} \end{pmatrix} = \begin{pmatrix} I_n & A \\ 0 & I_n \end{pmatrix} \begin{pmatrix} I_n & 0 \\ -A^{-1} & I_n \end{pmatrix} \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \begin{pmatrix} I_n & 0 \\ -A & I_n \end{pmatrix} \in E_{2n}(R)$$

since

$$\begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} = \begin{pmatrix} I_n & -I_n \\ 0 & I_n \end{pmatrix} \begin{pmatrix} I_n & 0 \\ I_n & I_n \end{pmatrix} \begin{pmatrix} I_n & -I_n \\ 0 & I_n \end{pmatrix}.$$

□

5.1.3. THEOREM (Whitehead lemma).

(i) *$E(R) = [E(R), E(R)]$, i.e. $E(R)$ is perfect.*

(ii) *$E(R) = [GL(R), GL(R)]$.*

PROOF (*Sketch*). (i) It follows from properties (ii) of elementary matrices that $[E(R), E(R)] \subset E(R)$. Also, $E_n(R)$ is generated by elements of the form $e_{ij}(a) = [e_{ik}(a), e_{kj}(1)]$ and so $E(R) \subset [E(R), E(R)]$. So, $E(R) = [E(R), E(R)]$.

(ii) For $A, B \in GL_n(R)$,

$$\begin{pmatrix} ABA^{-1}B^{-1} & 0 \\ 0 & I_n \end{pmatrix} = \begin{pmatrix} AB & 0 \\ 0 & (AB)^{-1} \end{pmatrix} \begin{pmatrix} A^{-1} & 0 \\ 0 & A \end{pmatrix} \begin{pmatrix} B^{-1} & 0 \\ 0 & B \end{pmatrix} \in E_{2n}(R).$$

Hence $[GL(R), GL(R)] \subset E(R)$ by 5.1.2. Also, from (i) above, $E(R) \subset [E(R), E(R)] \subset [GL(R), GL(R)]$. Hence $E(R) = [GL(R), GL(R)]$. □

5.1.4. DEFINITION.

$$\begin{aligned} K_1(R) &:= GL(R)/E(R) = GL(R)/\big[GL(R), GL(R)\big] \\ &= H_1\big(GL(R), \mathbb{Z}\big). \end{aligned}$$

5.1.5. REMARKS.

(i) For an exact category $\mathcal{C}$, the Quillen definition of $K_n(\mathcal{C})$, $n \geqslant 0$ coincides with the above definition of $K_1(R)$ when $\mathcal{C} = \mathcal{P}(R)$ (see [69] or [67]). We hope to discuss the Quillen construction in a forthcoming chapter.

(ii) The above Definition 4.1.4 is functorial, i.e. any ring homomorphism $R \to R'$ induces an Abelian group homomorphism $K_1(R) \to K_1(R')$.

(iii) $K_1(R) = K_1(M_n(R))$ for any positive integer n and any ring R.

(iv) $K_1(R)$, as defined above, coincides with $K_{\mathrm{det}}(\mathcal{P}(R))$ where $K_{\mathrm{det}}(\mathcal{P}(R))$ is a quotient of the additive group generated by all isomorphism classes $[P, \mu]$, $P \in \mathcal{P}(R)$, $\mu \in \mathrm{Aut}(P)$ (see [18] or [8]).

5.1.6. If R is a commutative, the determinant map $\det: GL_n(R) \to R^*$ commutes with $GL_n(R) \to GL_{n+1}(R)$ and hence defines a map $\det: GL(R) \to R^*$ which is surjective since given $a \in R^*$, there exists $A = \begin{pmatrix} a & 0 \\ 0 & I \end{pmatrix}$ such that $\det A = a$. Now det induces a map $\det: GL(R)/[GL(R), GL(R)] \to R^*$, i.e. $\det: K_1(R) \to R^*$. Moreover, $\alpha(a) = \begin{pmatrix} a & 0 \\ 0 & I \end{pmatrix}$ for all $a \in R^*$ defines a map $\alpha : R^* \to K_1(R)$ and $\det \alpha = 1_R$. Hence $K_1(R) \simeq R^* \oplus SK_1(R)$ where $SK_1(R) := \mathrm{Ker}(\det: K_1(R) \to R^*)$. Note that $SK_1(R) = SL(R)/E(R)$ where $SL(R) = \lim_{n\to\infty} SL_n(R)$ and $SL_n(R) = \{A \in GL_n(R) \mid \det A = 1\}$. Hence $SK_1(R) = 0$ if and only if $K_1(R) \simeq R^*$.

5.1.7. EXAMPLES.

(i) If F is a field, then $K_1(F) \simeq F^*$, $K_1(F[x]) \simeq F^*$.

(ii) If R is a Euclidean domain (for example $\mathbb{Z}$, $\mathbb{Z}[i] = \{a + bi;\ a, b \in \mathbb{Z}\}$, polynomial ring $F[x]$, F a field) then $SK_1(R) = 0$, i.e. $K_1(R) \simeq R^*$ (see [63,79]).

(iii) If R is the ring of integers in a number field F, then $SK_1(R) = 0$ (see [13] or [79]).

(iv) If R is a Noetherian ring of Krull dimension $\leqslant 1$ with finite residue fields and all maximal ideals, then $SK_1(R)$ is torsion [8].

5.2. *K_1 of local rings and skew fields*

5.2.1. THEOREM [18,79]. *Let R be a noncommutative local ring. Then there exists a homomorphism* $\det: GL_n(R) \to R^*/[R^*, R^*]$ *for each positive integer n such that*

(i) $E_n(R) \subset \mathrm{Ker}(\det)$

(ii)
$$\det\begin{pmatrix} \alpha_1 & & & \\ & \alpha_2 & & 0 \\ & & \ddots & \\ & 0 & & \ddots & \\ & & & & \alpha_n \end{pmatrix} = \bar{\alpha}_1\bar{\alpha}_2\cdots\bar{\alpha}_n \quad \textit{where } \alpha_i \in R^*$$

for all i and $\alpha \to \bar{\alpha}$ is the natural map

$$R^* \to (R^*)^{ab} = R^*/[R^*, R^*].$$

(iii)
$$\begin{tikzcd} GL_n(R) \arrow[rr] \arrow[dr] & & GL_{n+1}(R) \arrow[dl] \\ & (R^*)^{ab} & \end{tikzcd}$$
commutes

NOTE. The homomorphism 'det' above is usually called Dieudonné determinant because it was J. Dieudonné who first introduced the ideas in 5.2.1 for skew fields (see [23]).

5.2.2. THEOREM [79]. *Let R be a noncommutative local ring. Then the natural map $GL_1(R) = R^* \hookrightarrow GL(R)$ induces a surjection $R^*/[R^*, R^*] \to K_1(R)$ whose kernel is the subgroup generated by the images of all elements $(1 - xy)/(1 - yx)^{-1} \in R^*$ for all x, y in the unique maximal ideal $\mathbf{m}$ of R.*

5.2.3. THEOREM [79]. *If R is a skew field then $K_1(R) \cong R^*/[R^*, R^*]$.*

5.3. *Mennicke symbols*

5.3.1. Let R be a commutative ring with identity, $a, b \in R$. Choose $c, d \in R$ such that $ad - bc = 1$, i.e. such that $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL_2(R)$. Define Mennicke symbols $[a, b] \in SK_1(R)$ as the class of $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SK_1(R)$. Then

(i) $[a, b]$ is well defined.
(ii) $[a, b] = [b, a]$ if $a \in R^*$.
(iii) $[a_1 a_2, b] = [a_1, b][a_2, b]$ if $a_1 a_2 R + bR = R$.
(iv) $[a, b] = [a + rb, b]$ for all $r \in R$.

We have the following result

5.3.2. THEOREM [8]. *If R is a commutative ring of Krull dimension $\leqslant 1$, then the Mennicke symbols generate $SK_1(R)$.*

REMARKS. See [8] or [13] for further details on Mennicke symbols.

5.4. *Stability for K_1*

5.4.1. Stability results are very useful for reducing computations of $K_1(R)$ to computations of matrices over R of manageable size.

Let A be any ring with identity. An integer n is said to satisfy stable range condition (SR_n) for $GL(A)$ if whenever $r > n$, and $(a_1, a_2, \ldots, a_r)$ is a unimodular row, then there exists $b_1, b_2, \ldots, b_{r-1} \in A$ such that $(a_1 + a_r b_1, a_2 + a_r b_2, \ldots, a_{r-1} + a_r b_{r-1})$ is unimodular. Note that $(a_1, a_2, \ldots, a_r) \in A^r$ unimodular says that $(a_1, a_2, \ldots, a_r)$ generates the unit ideal, i.e. $\Sigma A a_i = A$ (see [8]). For example, any semi-local ring satisfies SR_2 (see [101] or [8]).

5.4.2. THEOREM [8,101]. *If SR_n is satisfied, then*

(i) $GL_m(A)/E_m(A) \to GL(A)/E(A)$ *is onto for all* $m \geqslant n$.

(ii) $E_m(A) \lhd GL_m(A)$ *if* $m \geqslant n+1$.

(iii) $GL_m(A)/E_m(A)$ *is Abelian for* $m \geqslant 2n$.

For further information on K_1-stability, see [8,101,106].

6. K_1, SK_1 of orders and group-rings; Whitehead torsion

6.1. Let R be the ring of integers in a number field F, Λ an R-order in a semi-simple F-algebra Σ. First we have the following result (see [8]).

6.1.1. THEOREM. *$K_1(\Lambda)$ is a finitely generated Abelian group.*

PROOF. The proof relies on the fact that $GL_n(\Lambda)$ is finitely generated and also that $GL_n(\Lambda) \to K_1(\Lambda)$ is surjective (see [8]). □

6.1.2. REMARKS. Let R be a Dedekind domain with quotient field F, Λ an R-order in a semi-simple F-algebra Σ. The inclusion $\Lambda \hookrightarrow \Sigma$ induces a map $K_1(\Lambda) \to K_1(\Sigma)$. Putting $SK_1(\Lambda) = \operatorname{Ker}(K_1(\Lambda) \to K_i(\Sigma))$, it means that understanding $K_1(\Lambda)$ reduces to understanding $K_1(\Sigma)$ and $SK_1(\Lambda)$. Since Σ is semi-simple, $\Sigma = \bigoplus \Sigma_i$ where $\Sigma_i = M_{n_i}(D_i)$, D_i a skew field. So $K_1(\Sigma) = \bigoplus K_1(D_i)$.

One way of studying $K_1(\Lambda)$ and $SK_1(\Lambda)$, $K_1(\Sigma)$ is via reduced norms. We consider the case where R is the ring of integers in a number field or p-adic field F.

Let R be the ring of integers in a number field or p-adic field F. Then there exists a finite extension E of F such that $E \otimes \Sigma$ is a direct sum of full matrix algebras over E, i.e. E is a splitting field of Σ. If $a \in \Sigma$, the element $1 \otimes a \in E \otimes \Sigma$ may be represented by a direct sum of matrices and the reduced norm of a, written $\operatorname{nr}(a)$ is defined as the product of their determinants. We then have $\operatorname{nr}: GL(\Sigma) \to C^*$ where $C =$ centre of Σ (if $\Sigma = \bigoplus_{i=1}^m \Sigma_i$ and $C = \bigoplus_{i=1}^m C_i$ we could compute $\operatorname{nr}(a)$ component-wise via $GL(\Sigma_i) \to C_i^*$). Since C^* is Abelian we have $\operatorname{nr}: K_1(\Sigma) \to C^*$. Composing this with $K_1(\Lambda) \to K_1(\Sigma)$ we have a reduced norm map $\operatorname{nr}: K_1(\Lambda) \to K_1(\Sigma) \to C^*$.

From the discussion below, it will be clear that an alternative definition of $SK_1(\Lambda) = \{x \in K_1(\Lambda) \mid \operatorname{nr}(x) = 1\}$.

6.1.3. THEOREM. *Let R be the ring of integers in a number field F, Λ an R-order in a semi-simple F-algebra Σ. In the notation of* 6.1.2, *let U_i be the group of all nonzero elements $a \in C_i$ such that $\beta(a) > 0$ for each embedding $\beta: C_i \to \mathbb{R}$ at which $\mathbb{R} \otimes_{C_i} \Sigma_i$ is not a full matrix algebra over $\mathbb{R}$. Then*

(i) *the reduced norm map yields an isomorphism* $\operatorname{nr}: K_1(\Sigma) \cong \prod_{i=1}^m U_i$,

(ii) $\operatorname{nr}: K_1(\Lambda) \subset \prod_{i=1}^m (U_i \cap R_i^*)$ *where R_i is the ring of integers in C_i.*

PROOF. See [18]. □

6.1.4. REMARKS.

(i) If Γ is a maximal R-order in Σ, then we have equality in (ii) of 6.1.3, i.e. $\mathrm{nr}(K_1(\Gamma)) = \prod_{i=1}^{m}(U_i \cap R_i^*)$. (See [18].) Hence $\operatorname{rank} K_1(\Gamma) = \operatorname{rank} \prod_{i=1}^{m}(U_i \cap R_i^*)$.

(ii) If Λ is any R-order in Σ, then $\mathrm{nr}(K_1(\Lambda))$ is of finite index in S^* (see [18]).

(iii) For all $n \geqslant 1$, $K_n(\Lambda)$ is finitely generated and $SK_n(\Lambda)$ is finite (see [49] or [50]).

6.1.5. THEOREM. *Let R be the ring of integers in a number field F, Λ any R-order in a semi-simple F-algebra Σ. Then $SK_1(\Lambda)$ is a finite group.*

PROOF. See [8]. The proof involves showing that $SK_1(\Lambda)$ is torsion and observing that $SK_1(\Lambda)$ is also finitely generated as a subgroup of $K_1(\Lambda)$ see 6.1.1. □

The next results are local versions of 6.1.3 and 6.1.5.

6.1.6. THEOREM. *Let R be the ring of integers in a p-adic field F, Γ a maximal R-order in a semi-simple F-algebra Σ. In the notation of* 6.1.2, *we have*

(i) $\mathrm{nr}: K_1(\Sigma) \simeq C^*$;

(ii) $\mathrm{nr}: K_1(\Gamma) \cong S^*$ *where $S = \oplus R_i$ and R_i is the ring of integers in C_i.*

6.1.7. THEOREM.

(i) *Let F be a p-adic field (i.e. any finite extension of $\widehat{Q}_P$), R the ring of integers of F, Λ any R-order in a semi-simple F-algebra Σ. Then $SK_1(\Lambda)$ is finite.*

(ii) *Let R be the ring of integers in a p-adic field F, $\mathbf{m}$ the maximal ideal of $R, q = |R/\mathbf{m}|$. Suppose that Γ is a maximal order in central division algebra over F. Then $SK_1(\Gamma)$ is a cyclic group of order $(q^n - 1)/q - 1$. $SK_1(\Gamma) = 0$ iff $D = F$.*

6.1.8. REMARKS.

(i) For the proof of 6.1.7, see [41] and [64].

(ii) It follows from 6.1.6 that $\operatorname{rank} K_1(\Gamma) = \operatorname{rank}(S^*)$ for any maximal order Γ in a p-adic semi-simple F-algebra.

(iii) If in 6.1.3 and 6.1.5 $R = Z, F = Q, G$ a finite group, we have that rank of $K_1(ZG) = s - t$ where s is the number of real representations of G, and t is the corresponding number of rational representations of G. (See [64].)

(iv) Computation of $SK_1(ZG)$ for various groups has attracted extensive attention because of its applicability in topology. For details of such computations, see [64].

(v) That for all $n \geqslant 1, SK_n(\mathbb{Z}G), SK_n(\widehat{\mathbb{Z}}_p G)$ are finite groups are proved in [49,50].

(vi) It also is known that if Γ is a maximal order in a semi-simple F-algebra Σ, then $SK_{2n}(\Gamma) = 0$ and $SK_{2n-1}(\Gamma) = 0$ for all $n \geqslant 1$ iff Σ is unramified over its centre, see [44]. These generalisations will be discussed in a forthcoming chapter on higher K-theory.

6.2. *Whitehead torsion*

J.H.C. Whitehead (see [119]) observed that if X is a topological space with fundamental group G, and $R = \mathbb{Z}G$, then the elementary row and column transformations of matrices over R have some natural topological meaning. To enable him to study homotopy between spaces, he introduced the group $\mathrm{Wh}(G) = K_1(ZG)/\omega(\pm G)$ where ω is the map $G \to GL_1(ZG) \to GL(ZG) \to K_1(ZG)$, such that if $f : X \to Y$ is a homotopy equivalence, then there exists an invariant $\tau(f)$ in $\mathrm{Wh}(G)$ such that $\tau(f) = 0$ if and only if f is a simple homotopy equivalence, i.e. $\tau(f) = 0$ iff f is induced by elementary deformations transforming X to Y. The invariant $\tau(f)$ is known as Whitehead torsion. (See [61].)

Now, it follows from 6.1.1 that $\mathrm{Wh}(G)$ is finitely generated when G is a finite group. Moreover, it is also well known that $\mathrm{Tor}(K_1(\mathbb{Z}G)) = (\pm 1) \times G^{ab} \times SK_1(\mathbb{Z}G)$ where $SK_1(\mathbb{Z}G) = \mathrm{Ker}(K_1(\mathbb{Z}G) \to K_1(QG))$, see [64]. So rank $K_1(\mathbb{Z}G) = \operatorname{rank} \mathrm{Wh}(G)$ and it is well known that $SK_1(\mathbb{Z}G)$ is the full torsion subgroup of $\mathrm{Wh}(G)$ (see [64]). So, computations of $\mathrm{Tor}(K_1(\mathbb{Z}G))$ reduce essentially to computations of $SK_1(\mathbb{Z}G)$. The last two decades have witnessed extensive research on computations of $SK_1(\mathbb{Z}G)$ for various groups G (see [64]). More generally, if R is the ring of integers in a number field or a p-adic field F, there has been extensive effort in understanding the groups $SK_n(RG) = \mathrm{Ker}(K_n(RG) \to K_n(FG))$ for all $n \geqslant 1$. (See [49–51].) More generally still, if Λ is an R-order in a semi-simple F-algebra Σ (i.e. Λ is a subring of Σ, finitely generated as an R-module and $\Lambda \otimes_R F = \Sigma$), there has been extensive effort to compute $SK_n(\Lambda) = \mathrm{Ker}(K_n(\Lambda) \to K_n(\Sigma))$ (see [49–51]) the results of which apply to $\Lambda = RG$. We shall discuss these computations further in the forthcoming chapter on higher K-theory.

Note also that Whitehead torsion is useful in the classifications of manifolds (see [64] or [61]).

7. Some $K_1 - K_0$ exact sequences

7.1. *Mayer–Vietoris sequence*

7.1.1. Let

$$\begin{tikzcd} A \arrow[r, "f_1"] \arrow[d, "f_2"'] & A_1 \arrow[d, "g_1"] \\ A_2 \arrow[r, "g_1"'] & A' \end{tikzcd} \tag{I}$$

be a commutative square of ring homomorphisms satisfying

(i) $A = A_1 \times_A A_2 = \{(a_1, a_2) \in A_1 \times A_2 \mid g_1(a_1) = g_2(a_2)\}$, i.e. given $a_1 \in A_1, a_2 \in A_2$ such that $g_1 a_1 = g_2 a_2$, then there exists one and only one element $a \in A$ such that $f_1(a) = a_1$, $f_2(a) = a_2$.

(ii) At least one of the two homomorphisms g_1, g_2 is surjective. The square (I) is then called a Cartesian square of rings.

7.1.2. THEOREM. *Given a Cartesian square of rings as in* 7.1.1, *then there exists an exact sequence* $K_1(A) \xrightarrow{\alpha_1} K_1(A_1) \oplus K_1(A_2) \xrightarrow{\beta_1} K_1(A') \xrightarrow{\delta} K_0(A) \xrightarrow{\alpha_0} K_0(A_1) \oplus K_0(A_2) \xrightarrow{\beta_0} K_0(A')$.

NOTE. Call this sequence the Mayer–Vietoris sequence associated to the Cartesian square (I). For details of the proof of 7.1.2, see [63].

SKETCH OF PROOF. The maps α_i, β_i $(i = 0, 1)$ are defined as follows: For $x \in K_i(A)$, $\alpha_i(x) = (f_{1*}(x), f_{2*}(x))$ and for $(y, z) \in K_i(A_1) \oplus K_i(A_2)$ $i = 0, 1$, $\beta_i(y, z) = g_{1*}y - g_{2*}z$. The boundary map $\delta : K_1(A') \to K_0(A)$ is defined as follows: Represent $x \in K_1(A')$ by a matrix $\gamma = (a_{ij})$ in $GL_r(A')$. This matrix determines an automorphism $\gamma : A'^n \to A'^n$. Let $\gamma(z_j) = \Sigma a_{ij} z_j$ where $\{z_j\}$ is a standard basis for A'^n. Let $P(\gamma)$ be the subgroup of $A_1^n \times A_2^n$ consisting of $\{(x, y) \mid \gamma g_1^n(x) = g_2^n(y)\}$ where $g_1^n : A_1^n \to A'^n$, $g_2^n : A_2^n \to A'^n$ are induced by g_1, g_2, respectively. We need the following

7.1.3. LEMMA.
(i) *If there exists $(b_{ij}) \in GL_n(A_2)$ which maps to $\gamma = (a_{ij})$, then $P(\gamma) \simeq A^n$.*
(ii) *If g_2 is surjective, then $P(\gamma)$ is a finitely generated projective A-module.*

For the proof of 7.1.3 see [63].
Conclusion of definition of δ: Now define

$$\delta[\gamma] = \big[P(\gamma)\big] - [A^n] \in K_0(A)$$

and verify exactness of the sequence 7.1.2 as an exercise. □

7.1.4. COROLLARY. *If A is a ring and $\mathbf{a}_1, \mathbf{a}_2$ ideals of A such that $\mathbf{a}_1 \cap \mathbf{a}_2 = 0$, then there exists an exact sequence*

$$\begin{aligned} K_1(A) \;&\to\; K_1(A/\mathbf{a}_1) \oplus K_1(A/\mathbf{a}_2) \to K_1\big(A/(\mathbf{a}_1 + \mathbf{a}_2)\big) \\ &\xrightarrow{\delta} K_0(A) \to K_0(A/\mathbf{a}_1) \oplus K_0(A/\mathbf{a}_2) \to K_0\big(A/(\mathbf{a}_1 \oplus \mathbf{a}_2)\big). \end{aligned}$$

PROOF. Follows by applying 7.1.2 to the Cartesian square:

$$\begin{tikzcd} A \arrow[r, "f_1"] \arrow[d, "f_2"'] & A/\mathbf{a}_1 \arrow[d, "g_1"] \\ A/\mathbf{a}_2 \arrow[r, "g_2"'] & A/(\mathbf{a}_1 + \mathbf{a}_2) \end{tikzcd} \tag{II}$$

□

7.1.5. EXAMPLE. Let G be a finite group of order n, $A = ZG$. Let $\mathbf{a}_1$ be the principal ideal of A generated by $b = \sum_{g \in G} g$, $\mathbf{a}_2$ the augmentation ideal $= \{\sum r_g g \mid \sum r_g = 0\}$. Then $\mathbf{a}_1 \cap \mathbf{a}_2 = 0$. So, $A_2 = A/\mathbf{a}_2 \simeq \mathbb{Z}$, $A' = A/(\mathbf{a}_1 + \mathbf{a}_2) \simeq \mathbb{Z}/n\mathbb{Z}$ from the Cartesian squares (I) and (II) above.

Now suppose that $|G| = p$, a prime. Let $G = \langle x \rangle$. Put $t = f_1(x)$. Then, A_1 has the form $\mathbb{Z}[t]$ with a single relation $\sum_{i=0}^{p-1} t^i = 0$. So, A_1 may be identified with $Z[\xi]$ where ξ is a primitive p-th root of unity.

We now have the following:

7.1.6. THEOREM. *If $|G| = p$, then $f_1 : K_0(ZG) \cong K_0(Z[\xi])$ is an isomorphism. Hence $K_0(ZG) \simeq Z \oplus C\ell(\mathbb{Z}[\xi])$.*

PROOF. From 7.1.2, we have an exact sequence

$$\begin{aligned} K_1\big(Z[\xi]\big) \oplus K_1(Z) \to K_1(Z/pZ) \xrightarrow{\delta} K_0(\mathbb{Z}G) \\ \to K_0\big(Z[\xi]\big) \oplus K_0(Z) \to K_0(Z/pZ). \end{aligned}$$

Now since $g_{2*} : K_0(\mathbb{Z}) \simeq K_0(Z/pZ)$ is an isomorphism, the result will follow once we show that $\delta = 0$. To show that $\delta = 0$, it suffices to show that $K_1(Z[\xi]) \to K_1(Z/pZ)$ is onto. Let r be a positive integer prime to p. Put $u = 1 + \xi + \cdots + \xi^{r-1} \in Z[\xi]$. Let $\xi^r = \eta$, $\eta^s = \xi$, for some $s > 0$. Then $\nu = 1 + \eta + \cdots + \eta^{s-1} \in Z[\xi]$. In $Q(\xi)$, we have

$$\nu = (\nu^s - 1)/(\eta - 1) = (\xi - 1)/(\xi^r - 1) = 1/u.$$

So, $u \in (Z[\xi])^*$, i.e. given $r \in (Z/pZ)^* \simeq K_1(Z/pZ)$, there exists $u \in (Z[\xi])^*$ such that $g_{1*}(u) = r$. That $K_0(\mathbb{Z}G) \simeq Z \oplus C\ell(Z[\xi])$ follows from 2.1.7. □

7.1.7. REMARKS. (i) The Mayer–Vietoris sequence 7.1.2 can be extended to the right to negative K-groups defined by H. Bass in [8]. More precisely, there exists functors K_{-n}, $n \geqslant 1$, from rings to Abelian groups such that the sequence

$$\cdots \to K_0(A') \to K_{-1}(A) \to K_{-1}(A_1) \oplus K_{-1}(A_2) \to K_{-1}(A') \to \cdots$$

is exact.

(ii) The Mayer–Vietoris sequence 7.1.2 can be extended beyond K_2 under special circumstances that will be discussed in the forthcoming chapter on higher K-theory.

7.2. *Exact sequence associated to an ideal of a ring*

7.2.1. Let A be a ring, $\mathbf{a}$ any ideal of A. The canonical map $f : A \to A/\mathbf{a}$ induces $f_* : K_i(A) \to K_i(A/\mathbf{a})$, $i = 0, 1$. We write $\overline{A}$ for $A/\mathbf{a}$ and for $M \in \mathcal{P}(A)$ we put $\overline{M} = M/\mathbf{a}M \simeq \overline{A} \otimes_A M$. Let $K_0(A, \mathbf{a})$ be the Abelian group generated by expressions of the form $[M, f, N]$, $M, N \in \mathcal{P}(A)$, where $f : \overline{A} \otimes_A M \simeq \overline{A} \otimes_A N$ with relations defined as follows:

For $L, M, N \in \mathcal{P}(A)$ and $\overline{A}$-isomorphisms $f : \overline{L} \simeq \overline{M}$, $g : \overline{M} \simeq \overline{N}$, we have

$$[L, gf, N] = [L, d, M] + [M, g, N].$$

(ii) Given exact sequences

$$0 \to M_1 \to M_2 \to M_3 \to 0; \quad 0 \to N_1 \to N_2 \to N_3 \to 0,$$

where $M_i, N_i, N_i \in \mathcal{P}(A)$, and given $\overline{A}$-isomorphisms $f_i : \overline{M_i} \simeq \overline{N_i}$ $(i = 1, 2, 3)$ which commute with the maps associated with the given sequences, we have $[M_2, f_2, N_2] = [M_1, f_1, N_1] + [M_3, f_3, N_3]$.

7.2.2. THEOREM. *There exists an exact sequence*

$$K_1(A) \to K_1(\overline{A}) \xrightarrow{\delta} K_0(A, \mathbf{a}) \xrightarrow{\eta} K_0(A) \to K_0(\overline{A}).$$

7.2.3. REMARKS.

(i) We shall not prove the above result in detail but indicate how the maps δ, η are defined leaving the rest as an exercise. It is clear how the maps $K_i(A) \to K_i(\overline{A})$, $i = 0, 1$ are defined. The map δ assigns to each $f \in GL_n(\overline{A})$ the triple $[A^n, f, A^n] \in K_0(A, \mathbf{a})$ while the map η takes $[M, f, N]$ onto $[M] - [N]$ for $M, N \in \mathcal{P}(A)$ such that $f : \overline{M} \simeq \overline{N}$.

(ii) The exact sequence 7.2.2 could be extended to K_2 and beyond with appropriate definitions of $K_i(A, \mathbf{a})$, $i \geqslant 1$. We shall discuss this in the context of higher K-theory in a forthcoming chapter, see [67].

7.3. *Localisation sequences*

7.3.1. Let S be a central multiplicative system in a ring A, $\mathbf{H}_S(A)$ the category of finitely generated S-torsion A-modules of finite projective dimension. Note that an A-module M is S-torsion if there exists $s \in S$ such that $sM = 0$, and that an A-module has finite projective dimension if there exists a finite $\mathbf{P}(A)$-resolution, i.e. there exists an exact sequence (I) $0 \to P_n \to P_{n-1} \to \cdots \to P_0 \to M \to 0$ where $P_i \in \mathbf{P}(A)$. Then we have the following theorem.

7.3.2. THEOREM. *With notation as in* 7.3.1, *there exist natural homomorphisms* δ, ε *such that the following sequence is exact*:

$$K_1(A) \to K_1(A_S) \xrightarrow{\delta} K_0\big(\mathbf{H}_S(A)\big) \xrightarrow{\varepsilon} K_0(A) \to K_0(A_S),$$

where A_S is the ring of fractions of A with respect to S.

PROOF. We shall not prove exactness in detail but indicate how the maps δ and ε are defined leaving details of proof of exactness at each point as an exercise.

Let $M \in \mathbf{H}_S(A)$ have a finite $\mathbf{P}(A)$-resolution as in 7.3.1 above. Define $\varepsilon([M]) = \Sigma(-1)^i[P_i] \in K_0(A)$. We define δ as follows: If $\alpha \in GL_m(A_S)$, let $s \in S$ be a common denominator for all entries of α such that $\beta = s\alpha$ has entries in A. We claim that $A^n/\beta A^n \in \mathbf{H}_S(A)$ and $A^n/sA^n \in \mathbf{H}_S(A)$. That they have finite $\mathbf{P}(A)$-resolutions follow from the exact sequences

$$0 \to A^n \xrightarrow{\beta} A^n \to A^n/\beta A^n \to 0 \quad \text{and}$$
$$0 \to A^n \xrightarrow{s} A^n \to A^n/sA^n \to 0.$$

To see that $A/\beta A^n$ is S-torsion, let $t \in S$ be such that $\alpha^{-1}t = \gamma$ has entries in A. Then $\gamma A^n \subset A^n$ implies that $tA^n \subset \alpha A^n$ and hence that $stA^n \subset s\alpha A^n = \beta A^n$. Then $st \in S$ annihilates $A^n/\beta A^n$.

We now define

$$\delta[\alpha] = [A^n/\beta A^n] - [A^n/sA^n].$$

So

$$\begin{aligned}\varepsilon\delta[\alpha] &= \varepsilon[A^n/\beta A^n] - \varepsilon[A^n/sA^n]\\ &= \big([A^n] - [A^n]\big) - \big([A^n] - [A^n]\big) = 0.\end{aligned}$$

□

7.3.3. REMARKS.

(i) Putting $A = \Lambda[t]$ and $S = \{t^i\}_{i \geqslant 0}$ in 7.3.2, we obtain an exact sequence

$$\begin{aligned}K_1\big(\Lambda[t]\big) \to K_1\big(\Lambda[t, t^{-1}]\big) \xrightarrow{\partial} K_0\big(H_{\{t^i\}}(\Lambda[t])\big)\\ \to K_0\big(\Lambda[t]\big) \to K_0\big(\Lambda[t, t^{-1}]\big)\end{aligned}$$

which is an important ingredient in the proof of the following result called the fundamental theorem for K_1 (see [8]).

(ii) FUNDAMENTAL THEOREM for K_1

$$K_1\big(\Lambda[t, t^{1-}]\big) \simeq K_1(\Lambda) \oplus K_0(\Lambda) \oplus NK_1(\Lambda) \oplus NK_1(\Lambda),$$

where $NK_1(\Lambda) = \mathrm{Ker}(K_1(\Lambda[t]) \xrightarrow{\tau} K_1(\Lambda))$ and τ is induced by the augmentation $\Lambda[t] \to \Lambda$ $(t = 1)$.

(iii) In the forthcoming section on higher K-theory, we shall discuss the extension of the localisation sequence 7.3.2 to the left for all $n \geqslant 1$ as well as some further generalisations of the sequence.

8. The functor K_2 – brief review

In this section we provide a brief review of the functor K_2 due to J. Milnor, see [63].

8.1. *K_2 of a ring – definitions and basic properties*

8.1.1. Let A be a ring. The Steinberg group of order n ($n \geqslant 3$) over A, denoted $\mathrm{St}_n(A)$ is the group generated by $x_{ij}(a)$, $i \neq j$, $1 \leqslant i, j \leqslant n$, $a \in A$, with relations

(i) $x_{ij}(a)x_{ij}(b) = x_{ij}(a+b)$.
(ii) $[x_{ij}(a), x_{k\ell}(b)] = 1$, $j \neq k$, $i \neq \ell$.
(iii) $[x_{ij}(a), x_{jk}(b)] = x_{ik}(ab)$, i, j, k distinct.
(iv) $[x_{ij}(a), x_{ki}(b)] = x_{kj}(-ba)$, $j \neq k$.

Since the generators $e_{ij}(a)$ of $E_n(A)$ satisfy relations (i)–(iv), above, we have a unique surjective homomorphism $\varphi_n : \mathrm{St}_n(A) \to E_n(A)$ given by $\varphi_n(x_{ij}(a)) = e_{ij}(a)$. Moreover, the relations for $\mathrm{St}_{n+1}(A)$ include those of $\mathrm{St}_n(A)$ and so, there are maps $\mathrm{St}_n(A) \to \mathrm{St}_{n+1}(A)$. Let $\mathrm{St}(A) = \varinjlim_n \mathrm{St}_n(A)$, $E(A) = \varinjlim_n E_n(A)$, then we have a canonical map $\varphi : \mathrm{St}(A) \to E(A)$.

8.1.2. DEFINITION. Define $K_2^M(A)$ as the kernel of the map $\varphi : \mathrm{St}(A) \to E(A)$.

8.1.3. THEOREM. *$K_2^M(A)$ is Abelian and is the centre of* $\mathrm{St}(A)$. *So* $\mathrm{St}(A)$ *is a central extension of $E(A)$.*

PROOF. See [63]. □

8.1.4. DEFINITION. An exact sequence of groups of the form $1 \to A \to E \xrightarrow{\varphi} G \to 1$ is called a central extension of G by A if A is central in E. Write this extension as (E, φ). A central extension (E, φ) of G by A is said to be universal if for any other central extension (E', φ') of G, there is a unique morphism $(E, \varphi) \to (E', \varphi)$.

8.1.5. THEOREM. $\mathrm{St}(A)$ *is the universal central extension of $E(A)$. Hence there exists a natural isomorphism $K_2^M(A) \simeq H_2(E(A), Z)$.*

PROOF. The last statement follows from the fact that for a perfect group G (in this case $E(A)$), the kernel of the universal central extension (E, φ) (in this case $(\mathrm{St}(A), \varphi)$) is naturally isomorphic to $H_2(G, Z)$ (in this case $H_2(E(A), \mathbb{Z})$).

For the proof of the first part see [63]. □

8.1.6. DEFINITION. Let A be a commutative ring, $u \in A^*$

$$w_{ij}(u) := x_{ij}(u)x_{ji}(-u^{-1})x_{ij}(u).$$

Define $h_{ij}(u) := w_{ij}(u)w_{ij}(-1)$. For $u, v \in A^*$, one can easily check that $\varphi([h_{12}(u), h_{13}(v)]) = 1$.

So $[h_{12}(u), h_{13}(v)] \in K_2(A)$. It can be shown that $[h_{12}(u), h_{13}(v)]$ is independent of the indices 1, 2, 3. We write $\{u, v\}$ for $[h_{12}(u), h_{13}(v)]$ and call this the Steinberg symbol.

8.1.7. THEOREM. *Let A be a commutative ring. The Steinberg symbol $\{,\} : A^* \times A^* \to K_2(A)$ is skew symmetric and bilinear, i.e.*

$$\{u, v\} = \{v, u\}^{-1} \quad \textit{and} \quad \{u_1u_2, v\} = \{u_1, v\}\{u_2, v\}.$$

PROOF. See [63]. □

8.1.8. THEOREM. *Let A be a field, division ring, local ring or a semi-local ring. Then $K_2^M(A)$ is generated by symbols.*

PROOF. See [20] or [109] or [26]. □

8.1.9. THEOREM (Matsumoto). *If F is a field, then $K_2^M(F)$ is generated by $\{u, v\}$, $u, v \in F^*$, with relations*

(i) $\{uu', v\} = \{u, v\}\{u', v\}$.
(ii) $\{u, vv'\} = \{u, v\}\{u, v'\}$.
(iii) $\{u, 1 - u\} = 1$,

i.e. $K_2^M(F)$ is the quotient of $F^* \otimes_{\mathbb{Z}} F^*$ by the subgroup generated by the elements $x \otimes (1 - x)$, $x \in F^*$.

8.1.10. EXAMPLES. Writing K_2 for K_2^M:

(i) $K_2(\mathbb{Z})$ is cyclic or order 2. See [63].
(ii) $K_2(\mathbb{Z}(i)) = 1$, so is $K_2(\mathbb{Z}\sqrt{-7})$, see [63].
(iii) $K_2(\mathbb{F}_q) = 1$ if F_q is a finite field with q elements. See [63].
(iv) If F is a field $K_2(F[t]) \simeq K_2(F)$, see [63].

More generally, $K_2(R[t]) \simeq K_2(R)$ if R is a commutative regular ring.

8.1.11. REMARKS.

(i) There is a definition by J. Milnor of higher K-theory of fields $K_n^M(F)$, $n \geqslant 1$, which coincides with $K_2^M(F)$ above for $n = 2$. More precisely,

$$K_n^M(F) := \underbrace{F^* \otimes F^* \otimes \cdots \otimes F^*}_{n \text{ times}} / \{a_1 \otimes \cdots \otimes a_n \mid a_i + a_j = 1 \text{ for some } i \neq j,\ a_i \in F^*\},$$

i.e. $K_n^M(F)$ is the quotient of $F^* \otimes F^* \otimes \cdots \otimes F^*$ (n times) by the subgroup generated by all $a_1 \otimes a_2 \otimes \cdots \otimes a_n$, $a_i \in F$ such that $a_i + a_j = 1$ for some $i \neq j$. Note that $K_*^M(F) = \bigoplus_{n \geqslant 0}^{\infty} K_n^M(F)$ is a ring.

(ii) The higher K-groups defined by D. Quillen [69,67], namely $K_n(\mathcal{C})$, $\mathcal{C}$ an exact category $n \geqslant 0$ and $K_n(A) = \pi_n(BGL(A)^+)$, $n \geqslant 1$, coincides with $K_2^M(A)$ above when $n = 2$ and $\mathcal{C} = \mathcal{P}(A)$.

8.2. *Connections with Brauer group of fields and Galois cohomology*

8.2.1. Let F be field and $B_r(F)$ the Brauer group of F, i.e. the group of stable isomorphism classes of central simple F-algebras with multiplication given by tensor product of algebras. See [58].

A central simple F-algebra A is said to be split by an extension E of F if $E \otimes A$ is E-isomorphic to $M_r(E)$ for some positive integer r. It is well known (see [58]) that such E can be taken as some finite Galois extension of F. Let $B_r(F, E)$ be the group of stable isomorphism classes of E split central simple algebras. Then $B_r(F) := B_r(F, F_s)$ where F_s is the separable closure of F.

8.2.2. THEOREM [58]. *Let E be a Galois extension of a field F, $G = \mathrm{Gal}(E/F)$. Then there exists an isomorphism $H^2(G, E^*) \cong B_r(F, E)$. In particular $B_r(F) \cong H^2(G, F_s^*)$ where $G = \mathrm{Fal}(F_s/F) = \varinjlim \mathrm{Gal}(E_i/F)$, where E_i runs through the finite Galois extensions of F.*

8.2.3. Now, for any $m > 0$, let μ_m be the group of m-th roots of 1, $G = \mathrm{Gal}(F_s/F)$, we have the Kummer sequence of G-modules

$$0 \to \mu_m \to F_s^* \to F_s^* \to 0$$

from which we obtain an exact sequence of Galois cohomology groups

$$F^* \xrightarrow{m} F^* \to H^1(F, \mu_m) \to H^1(F, F_s^*) \to \cdots,$$

where $H^1(F, F_s^*) = 0$ by Hilbert theorem 90. So we obtain isomorphism $\chi_m : F^*/mF^* \cong F^* \otimes \mathbb{Z}/m \to H^1(F, \mu_m)$.

Now, the composite

$$\begin{aligned} F^* \otimes_{\mathbb{Z}} F^* &\to (F^* \otimes_{\mathbb{Z}} F^*) \otimes \mathbb{Z}/m \to H^1(F, \mu_m) \otimes H^1(F, \mu_m) \\ &\to H^2\big(F, \mu_m^{\otimes 2}\big) \end{aligned}$$

is given by $a \otimes b \to \chi_m(a) \cup \chi_m(b)$ (where $\cup$ is cup product) which can be shown to be a Steinberg symbol inducing a homomorphism $g_{2,m} : K_2(F) \otimes \mathbb{Z}/m\mathbb{Z} \to H^2(F, \mu_m^{\otimes 2})$.

We then have the following result due to A.S. Merkurjev and A.A. Suslin, see [59].

8.2.4. THEOREM [59]. *Let F be a field, m an integer > 0 such that the characteristic of F is prime to m. Then the map*

$$g_{2,m} : K_2(F)/m\, K_2(F) \to H^2\big(F, \mu_m^{\otimes 2}\big)$$

is an isomorphism where $H^2(F, \mu_m^{\otimes 2})$ can be identified with the m-torsion subgroup of $B_r(F)$.

8.2.5. REMARKS. By generalising the process outlined in 8.2.3 above, we obtain a map

$$g_{n,m} : K_n^M(F)/m\ K_n^M(F) \to H^n\big(F, \mu_m^{\otimes 2}\big). \tag{I}$$

It is a conjecture of Bloch–Kato that $g_{n,m}$ is an isomorphism for all F, m, n. So, 8.2.4 is the $g_{2,m}$ case of the Bloch–Kato conjecture when m is prime to the characteristic of F. Furthermore, A. Merkurjev proved in [58], that 8.2.4 holds without any restriction on F with respect to m.

It is also a conjecture of Milnor that $g_{n,2}$ is an isomorphism. In 1996, V. Voevodsky proved that $g_{n,2}r$ is an isomorphism for any r. See [110].

8.3. *Some applications in algebraic topology and algebraic geometry*

8.3.1. *K_2 and pseudo-isotopy.* Let $R = \mathbb{Z}G$, G a group. For $u \in R^*$, put $w_{ij}(u) := x_{ij}(u)x_{ji}(-u^{-1})x_{ij}(u)$. Let W_G be the subgroup of $\mathrm{St}(R)$ generated by all $w_{ij}(g)$, $g \in G$. Define $\mathrm{Wh}_2(G) = K_2(R)/(K_2(R) \cap W_G)$.

Now let M be a smooth n-dimensional compact connected manifold without boundary. Two diffeomorphisms h_0, h_1 of M are said to be isotopic if they lie in the same path component of the diffeomorphism group. h_0, h_1 are said to be pseudo-isotopic if there is a diffeomorphism of the cylinder $M \times [0,1]$ restricted to h_0 on $M \times (0)$ and to h_1 on $M \times (1)$. Let $P(M)$ be the pseudo-isotopy space of M, i.e. the group of diffeomorphism h of $M \times [0,1]$ restricting to the identity on $M \times (0)$. Computation of $\pi_0(P(M^n))$ helps to understand the differences between isotopies and we have the following result due to Hatcher and Wagoner.

THEOREM [34]. *Let M be an n-dimensional ($n \geqslant 5$) smooth compact manifold with boundary. Then there exists a surjective map*

$$\pi_0\big(P(M)\big) \to \mathrm{Wh}_2\big(\pi_1(X)\big),$$

where $\pi_1(X)$ is the fundamental group of X.

8.3.2. *Bloch's formula for Chow groups.* Let X be a regular scheme of finite type over a field F, $CH^r(X)$ the Chow group of codimension r cycles on X modulo rational equivalence (see [33]). The functors K_n, $n \geqslant 0$, are contravariant functors from the category of schemes to the category of graded commutative rings, see [67]. Now we can sheafify the presheaf $U \to K_r(U)$ for $r \geqslant 0$ to obtain a sheaf $\mathcal{K}_{r,X}$. The stalk of $\mathcal{K}_{r,X}$ at $x \in X$ can be shown to be $K_r(O_{X,x})$. The following result, known as Bloch's formula, provides a K_2-theoretic formula for $CH^2(X)$.

THEOREM. *Let X be a regular scheme of finite type over a field F. Then there is a natural isomorphism*

$$H^2(X, \mathcal{K}_{r,X}) \simeq CH^2(X).$$

REMARK. D. Quillen proved a generalisation of the above result, i.e. $H^2(X, \mathcal{K}_{r,X}) \simeq CH^r(X)$ for all $r > 1$ in [67].

References

[1] J.F. Adams, *Vector fields on spheres*, Ann. of Math. **75** (1962), 603–632.

[2] R. Alperin, R.K. Dennis and M. Stein, *The Non-triviality of $SK_1(\mathbb{Z}[\pi])$*, Lecture Notes in Math., Vol. 353, Springer, Berlin (1973), 1–7.

[3] M.F. Atiyah, *K-Theory*, W.A. Benjamin, New York (1967).

[4] M. Atiyah and F. Hirzebruch, *Riemann–Roch theorems for differentiable manifolds*, Bull. Amer. Math. Soc. **65** (1959), 276–281.

[5] M. Atiyah and F. Hirzebruch, *Quelques théorèmes do non plongement pour les variétés différentiables*, Bull. Soc. Math. France **89** (1959), 383–396.

[6] A. Bak, *Editorial*, K-Theory **1** (1987), 1–4.

[7] H. Bass, *K-theory and stable algebra*, Publ. IHES **22** (1964), 1–60.

[8] H. Bass, *Algebraic K-Theory*, Benjamin, New York (1968).

[9] H. Bass, *Unitary algebraic K-theory*, Algebraic K-Theory III, Lecture Notes in Math., Vol. 343, Springer, Berlin (1973), 57–265.

[10] H. Bass, *Algebraic K-theory: a historical survey*, Proc. ICM, Vancouver, 1974, Vol. I (1975), 277–283.

[11] H. Bass, A.O. Kuku and C. Pedrini (eds), *Algebraic K-Theory and its Applications*, ICTP K-Theory, Proceedings, World Scientific, Singapore (1999).

[12] H. Bass and M.P. Murthy, *Grothendieck groups and Picard groups of Abelian group rings*, Ann. of Math. **86** (1967), 16–73.

[13] H. Bass, J. Milnor and J.-P. Serre, *Solution of the congruence subgroup problem for SL_n ($n \geqslant 3$) and Sp_{sn} ($n \geqslant 2$)*, Publ. IHES **33** (1967), 59–137.

[14] H. Bass and J. Tate, *The Milnor ring of a global field*, Algebraic K-Theory II, Lecture Notes in Math., Vol. 342, Springer, Berlin (1973), 349–446.

[15] A. Borel and J.-P. Serre, *Le théorème de Riemann–Roch*, Bull. Soc. Math. France **8** (1958), 97–136.

[16] A. Cartan and S. Eileniberg, *Homological Algebra*, Princeton Univ. Press (1956).

[17] A. Connes, *Non-Commutative Geometry*, Academic Press, New York (1994).

[18] C.W. Curtis and I. Reiner, *Methods of Representation Theory II*, J. Wiley, New York, Brisbane (1987).

[19] R. Dedekind and M. Weber, *Theorie der algebraischen Funktionen einer Veränderlichen* (= R. Dedekind, Gesam. Math. Werker, Vol. I, Braunschweig (1932), 238–349), Crelle's J. **92** (1882), 181–290.

[20] R.K. Dennis and M. Stein, *K_2 of radical ideals and semi-local rings revisited*, Algebraic K-Theory II, Lecture Notes in Math., Vol. 342, Springer, Berlin (1973), 281–303.

[21] D.N. Diep, A.O. Kuku and N.Q. Tho, *Non-commutative Chern characters of compact Lie group C^*-algebra*, K-Theory **17** (1999), 195–208.

[22] D.N. Diep, A.O. Kuku and N.Q. Tho, *Non-commutative Chern characters of compact quantum groups*, J. Algebra **226** (2000), 311–331.

[23] J. Dieudonné, *Les déterminants sur un corps non commutatif*, Bull. Soc. Math. France **71** (1943), 27–45.

[24] A.W.M. Dress, *Induction and structure theorems for orthogonal representations of finite groups*, Ann. of Math. **102** (1975), 291–325.

[25] A.W.M. Dress and A.O. Kuku, *The Cartan map for equivariant higher algebraic K-groups*, Comm. Algebra **7** (9) (1981), 727–746.

[26] S. Gersten, *On the functor K_2*, J. Algebra **17** (1971), 212–237.

[27] S. Gersten, *On Mayer–Vietoris functors and algebraic K-theory*, J. Algebra **18** (1971), 51–88.

[28] D. Grayson, *Higher Algebraic K-Theory II (after Daniel Quillen)*, Lecture Notes in Math., Vol. 551, Springer, Berlin (1976), 217–240.

[29] D. Grayson, *Finite generation of K-groups of a curve over a finite field (after Daniel Quillen)*, Lecture Notes in Math., Vol. 966, Springer, Berlin (1982), 69–90.

[30] D. Grayson, *A survey on K-theory of fields*, Contemp. Math. **83** (1989), 31–55.

[31] A. Grothendieck, *Classes de faisceaux et théréme de Riemann–Roch*, Mimeographed notes, Princeton, 1957; Théorie des Intersections et Théoréme de Riemann–Roch (SGA6), Lecture Notes in Math., Vol. 225, Springer, Berlin (1971), 20–77.

[32] M. Hazewinkel, *Operations in the K-theory of endomorphisms*, J. Algebra **84** (1983), 285–304.

[33] R. Hartshorne, *Algebraic Geometry*, Springer, New York (1977).

[34] A. Hatcher and J. Wagoner, *Pseudo-isotopies of compact manifolds*, Astérisque **6** (1973).

[35] A. Heller, *Some exact sequences in algebraic K-theory*, Topology **4** (1965), 389–408.

[36] A. Heller and I. Reiner, *Grothendieck groups of orders in semisimple algebras*, Trans. Amer. Math. Soc. **112** (1964), 344–355.

[37] G. Higman, *The units of group rings*, Proc. London Math. Soc. **46** (1940), 231–248.

[38] M. Karoubi, *K-Theory: An Introduction*, Springer, Berlin (1978).

[39] M. Karoubi and O. Villamayor, *K-théorie algébrique et K-théorie topologique*, Math. Scand. **28** (1971), 265–307.

[40] A.O. Kuku, *Some algebraic K-theoretical applications of LF and NF functors*, Proc. Amer. Math. Soc. **37** (2) (1973), 363–365.

[41] A.O. Kuku, *Whitehead group of orders in p-adic semi-simple algebras*, J. Algebra **25** (1973), 415–418.

[42] A.O. Kuku, *Some finiteness theorems in the K-theory of orders in p-adic algebras*, J. London Math. Soc. **13** (1) (1976), 122–128.

[43] A.O. Kuku, *SK_n of orders and G_n of finite rings*, Lecture Notes in Math., Vol. 551, Springer, Berlin (1976), 60–68.

[44] A.O. Kuku, *SG_n of orders and group rings*, Math. Z. **165** (1979), 291–295.

[45] A.O. Kuku, *Higher algebraic K-theory of group rings and orders in algebras over number fields*, Comm. Algebra **10** (8) (1982), 805–816.

[46] A.O. Kuku, *Equivariant K-theory and the cohomology of profinite groups*, Lecture Notes in Math., Vol. 1046, Springer, Berlin (1984), 234–244.

[47] A.O. Kuku, *K-theory of group rings of finite groups over maximal orders in division algebras*, J. Algebra **91** (1) (1984), 18–31.

[48] A.O. Kuku, *Axiomatic theory of induced representations of finite groups*, Group Representations and its Applications, A.O. Kuku (ed.), Les Cours du CIMPA Nice, N4 (1985).

[49] A.O. Kuku, *K_n, SK_n of integral group rings*, Contemp. Math. **55** (1986), 333–338.

[50] A.O. Kuku, *Some finiteness results in the higher K-theory of orders and group rings*, Topology Appl. **25** (1987), 185–191.

[51] A.O. Kuku, *Ranks of K_n and G_n of orders and group rings of finite groups over integers in number fields*, J. Pure Appl. Algebra **138** (1999), 39–44.

[52] A.O. Kuku, *Equivariant higher K-theory for compact Lie group actions*, Beitrage zur Algebra und Geometrie **41** (1) (2000), 141–150.

[53] T.-Y. Lam, *Induction techniques for Grothendieck groups and Whitehead groups of finite groups*, Ann. Sci. École Norm. Sup. (Paris) **1** (1968), 91–148.

[54] J.L. Loday, *Cyclic Homology*, Springer, Berlin (1992).

[55] R. Martin, *Anneaux de Grothendieck de Varieties de drapeant*, Bull. Soc. Math. France **104** (1976), 337–348.

[56] H. Matsumoto, *Sur les sous-groupes arithmètiques des groupes semi-simples déployés*, University of Paris Thesis, 1968, Ann. Sci. École Norm. Sup. (Paris) **2** (1969), 1–62.

[57] A. Merkurjev, *On the norm residue symbol of degree 2*, Dokl. Akad. Nauk SSSR **261** (1981), 542–547 (in Russian); Soviet Math. Dokl. **24** (1981), 546–551.

[58] A.S. Merkurjev, *K_2 of fields and the Brauer group*, AMS Contemp. Math. **55** (1983), 529–546.

[59] A. Merkurjev and A. Suslin, *K-cohomology of Severi–Brauer varieties and the norm residue homomorphism*, Izv. Akad. Nauk SSSR **46** (1982) (in Russian); Math. USSR-Izv. **21** (1983), 307–340.

[60] A. Merkurjev and A. Suslin, *The norm residue homomorphism of degree three*, Izv. Akad. Nauk. SSSR **54** (1990), 339–356 (in Russian); Math. USSR-Izv. **36** (1991), 349–367.

[61] J. Milnor, *Whitehead torsion*, Bull. Amer. Math. Soc. **72** (1966), 358–426.

[62] J. Milnor, *Algebraic K-theory and quadratic forms*, Invent. Math. **9** (1970), 318–344.

[63] J. Milnor, *Introduction to Algebraic K-Theory*, Princeton Univ. Press (1971).

[64] K. Oliver, *Whitehead Groups of Finite Groups*, Cambridge Univ. Press (1988).

[65] D. Quillen, *Cohomology of groups*, Proc. ICM Nice 1970, Vol. 2, Gauthier-Villars, Paris (1971), 47–52.

[66] D. Quillen, *On the cohomology and K-theory of the general linear groups over a finite field*, Ann. of Math. **96** (1972), 552–586 (revised version of 1970 preprint "The K-theory associated to a finite field").

[67] D. Quillen, *Higher algebraic K-theory I*, Algebraic K-Theory I, Lecture Notes in Math., Vol. 341, Springer, Berlin (1973), 85–147.

[68] D. Quillen, *Finite generation of the groups K_i of rings of algebraic integers*, Algebraic K-Theory I, Lecture Notes in Math., Vol. 341, Springer, Berlin (1973), 195–214.

[69] D. Quillen, *Higher K-theory for categories with exact sequences*, New Developments in Topology, London Math. Soc. Lecture Notes, No. 11, Cambridge Univ. Press (1974), 95–110.

[70] D. Quillen, *Higher algebraic K-theory*, Proc. Intern. Congress Math., Vancouver, 1974, Vol. I, Canad. Math. Soc. (1975), 171–176.

[71] D. Quillen, *Projective modules over polynomial rings*, Invent. Math. **36** (1976), 167–171.

[72] D. Quillen, *On the group completion of a simplicial monoid*, Mem. Amer. Math. Soc., Vol. 529 (1994), 89–105 (1971 MIT Preprint).

[73] I. Reiner, *Integral representations of cyclic groups of prime order*, Proc. Amer. Math. Soc. **8** (1957), 142–146.

[74] I. Reiner, *Maximal Orders*, Academic Press, London (1979).

[75] I. Reiner and K.W. Roggenkamp, *Integral Representations*, Lecture Notes, Vol. 744, Springer, Berlin (1979).

[76] I. Reiner and S. Ullom, *Class groups of integral group rings*, Trans. Amer. Math. Soc. **170** (1972), 1–30.

[77] S. Rim, *Modules over finite groups*, Ann. of Math. **69** (1959), 700–712.

[78] S. Rim, *Projective class groups*, Trans. Amer. Math. Soc. **98** (1961), 459–467.

[79] J. Rosenberg, *Algebraic K-Theory and its Applications*, Springer, Berlin (1994).

[80] W. Scharlau, *Quadratic and Hermitian Forms*, Springer, Berlin (1985).

[81] G. Segal, *Equivariant K-theory*, Publ. Math. J. IHES **34** (1968).

[82] G. Segal, *Representation ring of a compact Lie group*, IHES **34** (1968), 113–128.

[83] J.-P. Serre, *Modules projectifs et espaces fibrés à fibre vectorielle*, Exposé 23, Seminaire P. Dubreil 1957/58, Secrétariat Mathématique, Paris (1958).

[84] J.-P. Serre, *Local Fields*, Springer, Berlin.

[85] L. Siebenmann, *Obstruction to finding a boundary for an open manifold of dimension greater than five*, Princeton University thesis (1965); *The structure of tame ends*, Notices Amer. Math. Soc. **13** (1966), 862.

[86] V. Srinivas, *Algebraic K-Theory*, Progress in Math., Vol. 90, Birkhauser (1991).

[87] R. Steinberg, *Génératerus, relations et revètements de groupes algébriques*, Colloq. Théorie des Groupes Algébriques, Gauthier-Villars, Paris (1962), 113–127 (in French).

[88] A. Suslin, *Projective modules over polynomial rings are free*, Dokl. Akad. Nauk. SSSR **229** (1976), 1063–1066; Soviet Math. Dokl. **17** (1976), 1160–1164.

[89] A. Suslin, *The structure of the special linear group over rings of polynomials*, Izv. Akad. Nauk SSSR Ser. Mat. **41** (1977), 235–252.

[90] A. Suslin, *Mennicke symbols and their applications in the K-theory of fields*, Lecture Notes in Math., Vol. 966, Springer, Berlin (1982), 334–356.

[91] A.A. Suslin, *On the K-theory of algebraically closed fields*, Invent. Math. **73** (1983), 241–245.

[92] A.A. Suslin, *On the K-theory of local fields*, J. Pure Appl. Alg. **34** (1984), 301–318.

[93] R. Swan, *Projective modules over finite groups*, Bull. Amer. Math. Soc. **65** (1959), 365–367.

[94] R. Swan, *Induced representations and projective modules*, Annals **71** (1960), 552–578.

[95] R. Swan, *Vector bundles and projective modules*, Trans. Amer. Math. Soc. **105** (1962), 264–277.

[96] R. Swan, *Nonabelian homological algebra and K-theory*, Proc. Sympos. Pure Math., Vol. XVII (1970), 88–123.

[97] R. Swan, *Algebraic K-theory*, Proc. ICM Nice 1970, Vol. 1, Gauthier-Villars, Paris (1971), 191–199.

[98] R. Swan, *Excision in algebraic K-theory*, J. Pure Appl. Algebra **1** (1971), 221–252.

[99] R. Swan, *K-Theory of Finite Groups and Orders*, Lecture Notes, Vol. 149, Springer, Berlin (1970).

[100] R. Swan, *Topological examples of projective modules*, Trans. Amer. Math. Soc. **230** (1977), 201–234.

[101] R. Swan, *Algebraic K-Theory*, Lecture Notes in Math., Vol. 76, Springer, Berlin (1968).

[102] J. Tate, *Symbols in arithmetic*, Proc. ICM Nice 1970, Vol. 1, Gauthier-Villars, Paris (1971), 201–211.

[103] J. Tate, *Relations between K_2 and Galois cohomology*, Invent. Math. **36** (1976), 257–274.
[104] J. Tate, *On the torsion in K_2 of fields*, Algebraic Number Theory (1976 Kyoto Internat. Sympos.), Japan Soc. Promotion Sci., Tokyo (1977).
[105] K. Vadarajan, *Finiteness Obstruction of C.T.C. Wall*, Wiley-Interscience (1989).
[106] L. Vaserstein, *On the stabilization of the general linear group over a ring*, Math. USSR-Sb. **8** (1969), 383–400.
[107] L. Vaserstein, *The stable range of rings and the dimension of topological spaces*, Funktsional. Anal. i Prilozhen. **5** (1971), 17–27 (in Russian).
[108] L. Vaserstein, *Foundations of algebraic K-theory*, Uspekhi Mat. Nauk **31** (1976), 87–149 (in Russian); Russian Math. Surveys **31** (1976), 89–156.
[109] W. van der Kallen, *Le K_2 des nombres duaux*, C. R. Acad. Sci. Paris **273** (1971), A1204–A1207.
[110] V. Voevodsky, *The Milnor conjecture*, Preprint (1996).
[111] T. Vorst, *Polynomial extensions and excision for K_1*, Math. Ann. **244** (1979), 193–204.
[112] F. Waldhausen, *Algebraic K-theory of generalized free products*, Revised version of 1973 preprint, Ann. of Math. **108** (1978), 135–256.
[113] C.T.C. Wall, *Graded Brauer groups*, J. Reine Angew. Math. **213** (1964), 187–199.
[114] C.T.C. Wall, *Finiteness conditions for CW-complexes*, Ann. of Math. **81** (1965), 56–69.
[115] C.T.C. Wall, *Surgery of non-simply-connected manifolds*, Ann. of Math. **84** (1966), 217–276.
[116] C.T.C. Wall, *Foundations of algebraic L-theory*, Algebraic K-Theory III, Lecture Notes in Math., Vol. 343, Springer, Berlin (1973), 266–300.
[117] C.T.C. Wall, *Norms of units in group rings*, Proc. London Math. Soc. **29** (1974), 593–632.
[118] J.H.C. Whitehead, *Simplicial spaces, nucleii and m-groups*, Proc. London Math. Soc. **45** (1939), 243–327.
[119] J.H.C. Whitehead, *Simple homotopy types*, Amer. J. Math. **72** (1950), 1–57.
[120] E. Witt, *Theorie der quadratischen Formen in beliebigen Körpern*, J. Reine Angew. Math. **176** (1937), 31–44.

Section 2D
Model Theoretic Algebra

This Page Intentionally Left Blank

Model Theory for Algebra

Mike Prest

Department of Mathematics, University of Manchester, Manchester M13 9PL, UK
E-mail: mprest@maths.man.ac.uk

Contents

Abstract

The purpose of this chapter is to give a general introduction to the basic ideas and techniques from model theory.

I begin with some general remarks concerning model theory and its relationship with algebra. There follows a "mini-course" on first order languages, structures and basic ideas in model theory. Then there is a series of subsections which describe briefly some topics from model theory.

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1. Model theory and algebra

There is a variety of ways in which people have described the relationship between model theory, algebra and logic. Certainly, model theory fits naturally between, and overlaps, algebra and logic. Model theory itself has a "pure" aspect, where we investigate structures and classes of structures which are delineated using notions from within model theory, and it has an "applied" aspect, where we investigate structures and classes of structures which arise from outside model theory.

The first aspect is exemplified by stability theory where we assume just that we are dealing with a class of structures (cut out by some axioms) in which there is a "notion of independence" satisfying certain reasonable conditions. The investigation of such classes and the development of structure theory within such classes was a major project of Shelah and others (see [49,50,23]). Out of it have grown other projects and directions, in particular, "geometrical stability theory" which has close links with algebraic geometry (see, e.g., [41]).

The second aspect is exemplified by the model theory of fields (or groups, or modules, or ...). Here the techniques used arise mostly from the specific area but there is some input of model-theoretic ideas, techniques and theorems. The input from model theory is typically not from the most highly developed "internal" parts of the area but one can be fairly sure that at least the compactness theorem 2.10 will figure as well as a certain perspective. The model-theoretic perspective, of course, leads one to ask questions which may not be algebraically natural but it may also lead to fresh ideas on existing algebraic questions.

For example, within the model theory of modules one may aim to classify the complete theories of modules over a given ring. Model-theoretically this is a natural project because of the central role played in model theory by elementary classes. Algebraically it does not seem very natural, even though it can be described in purely algebraic terms (by making use of the notion of ultraproduct). Nevertheless, this project did lead to unexpected discoveries and algebraic applications (see the companion chapter on model theory and modules in this volume for example).

In its development model theory has looked very much towards algebra and other areas outside logic. It has often taken ideas from these areas, extracted their content within a framework provided by logic, developed them within that context and applied the results back to various areas of algebra (as well as parts of analysis and geometry see, e.g., [8,32, 53]). An example of this process is provided by the concept of being "algebraic over" a set of elements (see the subsection on this below). The inspirational example here is the notion of an element of a field being algebraic (as opposed to transcendental) over a subfield. This leads to a general and fundamental model theoretic notion which applies in many different contexts.

That phase of the development of model theory provided most of the ideas mentioned in this article. For some snapshots of model theory as it is now, one may look at the various (especially survey) articles mentioned above, below and in the bibliography.

2. The basics

Most, but not all, model theory uses first-order finitary logic. In this article I mostly confine myself to the first-order, finitary context. By a "formal language", or just "language", I will, from now, mean a first-order, finitary language.

> "First order" means that the quantifiers range over elements of a given structure (but also see the subsection on many-sorted structures). In a second-order language we also have quantifiers which range over arbitrary subsets of structures (there are intermediate languages where there are restrictions on the subsets). A second-order language is, of course, much more expressive that a first-order one but we lose the compactness theorem. One of the points of model theory is that if a property can be expressed by a first-order formula or sentence then we know that it is preserved by certain constructions.
> "Finitary" means that the formulas of the ($L_{\omega\omega}$-)language are all finite strings of symbols. An $L_{\infty\omega}$-language is one where arbitrary conjunctions and disjunctions of formulas are formulas. In such a language one can express the property of a group being torsion by $\forall x \bigvee_{i=1}^{\infty} x^i = e$. In an $L_{\infty\infty}$-language one also allows infinite strings of quantifiers. See [19], [27, Section 2.8] for these languages and for some algebraic applications of them.
> From the great variety of languages that have been considered by logicians, it is the first-order finitary ones which have proved to be most useful for applications in algebra.

The concept of a first order, finitary formula is rather basic but is often rather quickly passed over in accounts written for non-logicians. Certainly it is possible to do a, perhaps surprising, amount of model theory without mentioning formulas and so when writing an article for algebraists for instance, one may wish to minimise mention of, or even entirely avoid, talking about formulas because one knows that this will be a stumbling block to many readers. Although this is possible and sometimes even desirable, this is not the course that is taken here.

A **formula** (by which I will always mean a formula belonging to a formal language) is a string of symbols which can be produced in accordance with certain rules of formation. In general a formula will contain occurrences of variables $(x, y, \ldots)$. Some of these will be **bound** (or **within the scope of**) a quantifier. For example, in $\forall x \exists y (x \neq y)$ this is true of the occurrence of x and that of y (the "x" in $\forall x$ is counted as part of the quantifier, not as an occurrence of x). Some may be **free**. For example, in $\forall x(\exists y(x \neq y)) \wedge (x = z)$ the unique occurrence of z is free as is the second occurrence of x. The **free variables** of a formula are those which occur free somewhere in the formula. A formula without free variables is called a **sentence** and such a formula is either true or false in a given structure for the relevant language (for instance, $\forall x \exists y (x \neq y)$ is true in a structure iff that structure has at least two distinct elements). We write $M \models \phi$ if the sentence ϕ is true in the structure M. We write $\phi(x_1, \ldots, x_n)$ to indicate that the set of free variables of the formula ϕ is contained in $\{x_1, \ldots, x_n\}$ (it is useful not to insist that each of x_1 to x_n actually occurs free in ϕ). Given a formula $\phi(x_1, \ldots, x_n)$ of a language L, given a structure M for that language and given elements $a_1, \ldots, a_n \in M$ we may replace every free occurrence of x_i in ϕ by a_i – the result we denote by $\phi(a_1, \ldots, a_n)$ – and then we obtain a **formula with parameters** which is now a statement which is either true or false in M: we write $M \models \phi(a_1, \ldots, a_n)$ if it is true in M. For example, if $\phi(x, z)$ is $\forall x(\exists y(x \neq y)) \wedge x = z$ and if $a, b \in M$ then $M \models \phi(a, b)$ iff M has at least two elements and if $a = b$.

> One cannot literally replace an occurrence of a variable by an element of a structure. Rather, one enriches the language by "adding names (new constant symbols) for elements of the structure" and then, using the same notation for an element and for its name, $\phi(a_1, \ldots, a_n)$ becomes literally a sentence of a somewhat larger language. See elsewhere (for example, [10,24]) for details.
> Also see those references for the precise definition of the satisfaction relation, $M \models \phi$ and $M \models \phi(a_1, \ldots, a_n)$, between structures and sentences/formulas with parameters. It is a natural inductive definition and one does not normally have to refer to it in order to understand the content of the relation in particular cases.

A formal language has certain basic ingredients or building blocks. Some of these, such as the symbol $\wedge$ which represents the operation of conjunction ("and"), are common to all languages: others are chosen according to the intended application. Then one has certain rules which delimit exactly the ways in which the formulas of the language may be built up from these ingredients.

The ingredients common to all languages are: an infinite stock of **variables** (or **indeterminates**); the **logical connectives**, $\wedge$ (**conjunction** "and"), $\vee$ (**disjunction** "or"), $\neg$ (**negation** "not"), $\rightarrow$ (**implication** "implies"), $\leftrightarrow$ (**bi-implication** "iff"); the **universal quantifier** $\forall$ ("for all") and the **existential quantifier** $\exists$ ("there exists"); a symbol, $=$, for equality. We also need to use parentheses, (and), to avoid ambiguity but there are conventions which reduce the number of these and hence aid readability of formulas. The language which is built up from just this collection of symbols we denote by L_0 and call the **basic language** (**with equality**). The formulas of this language are built up in a natural way, as follows.

The basic language. We use letters such as x, y, u, v and indexed letters such as $x_1, x_2, \ldots$ for variables. We also abuse notation (in the next few lines and in general) by allowing these letters to range over the set of variables, so x for instance is a "generic" variable.

The definition of the formulas of the language is inductive. First we define the atomic formulas (the most basic formulas) and then we say how the stock of formulas may be enlarged by inductively combining formulas already constructed.

If x and y are any two variables then $x = y$ is an **atomic formula** (so $x = x, u = x, \ldots$ are atomic formulas). There are no more atomic formulas (for this language).

If ϕ and ψ are formulas then the following also are formulas: $(\phi \wedge \psi)$, $(\phi \vee \psi)$, $(\neg\phi)$, $(\phi \rightarrow \psi)$, $(\phi \leftrightarrow \psi)$. Any formula constructed from the atomic formulas using only these operations is said to be **quantifier-free**.

EXAMPLES 2.1. $(x = y \wedge x = u)$ is a formula, so is $(x = y \rightarrow u = v)$, as is $((x = y \wedge x = u) \vee (x = y \rightarrow u = v))$, as is $(\neg((x = y \wedge x = u) \vee (x = y \rightarrow u = v)))$. In order to increase readability we write, for instance, $(x \neq y)$ rather than $(\neg x = y)$. We may also drop pairs of parentheses when doing so does not lead to any ambiguity in reading a formula. Some conventions allow the removal of further parentheses. Just as $\times$ has higher priority than $+$ (so $2+3\times 4$ equals 14 not 20) we assign $\neg$ higher priority than $\wedge$ and $\vee$, which have higher priority than $\rightarrow$ and $\leftrightarrow$. The assignment of priorities to $\forall x$ and $\exists x$ is rather less consistent.

If ϕ is a formula and x is any variable then $(\forall x\phi)$ and $(\exists x\phi)$ are formulas.

A **formula** is any string of symbols which is formed in accordance with these rules.

A further convention sometimes used is to write $\forall x(\phi)$ or just $\forall x\phi$ for $(\forall x\phi)$ and similarly for $\exists$ and even for a string of quantifiers.

EXAMPLES 2.2. $(\forall x\,\exists y(x=y\vee x=z))\vee(x\neq y\wedge\forall u\,u=z)$ is a formula which, with more parentheses shown, would be $((\forall x(\exists y(x=y\vee x=z)))\vee(x\neq y\wedge(\forall u\,u=z)))$.

> We remark that only $\neg$, $\wedge$ and $\exists$ (say) are strictly necessary since one has, for instance, that $\phi\vee\psi$ is logically equivalent to $\neg(\neg\phi\wedge\neg\psi)$ and that $\forall x\phi$ is logically equivalent to $\neg\exists x\neg\phi$. This allows proofs which go by induction on complexity (of formation) of formulas to be shortened somewhat since fewer cases need be considered.

The optional extras from which we may select to build up a more general language, L, are the following: function symbols; relation symbols; constant symbols. Each function symbol and each relation symbol has a fixed **arity** (number of arguments). These optional symbols are sometimes referred to as the **signature** of the particular language.

EXAMPLE 2.3. Suppose that we want a language appropriate for groups. We could take the basic language L_0 and select, in addition, just one binary function symbol with which to express the multiplication in the group. In this case it would be natural to use operation, $x*y$, rather than function, $f(x,y)$, notation and that is what we do in practice. Since inverse and identity are determined once we add the group axioms we need select no more. For instance, the group axiom which says that every element has a right inverse could be written $\forall x\exists y\forall z((x*y)*z=z\wedge z*(x*y)=z)$. But it would make for more easily readable formulas if we give ourselves a unary (= 1-ary) function symbol with which to express the function $x\mapsto x^{-1}$ and a constant symbol with which to "name" the identity element of the group. Again, we use the natural notation and so would have, among the axioms for a group written in this language, $\forall x(x*x^{-1}=e)$.

For many purposes the choice of language is not an issue so long as the collection of definable sets (see the subsection on these) remains unchanged. But change of language does change the notion of substructure and it is also crucial for the question of quantifier-elimination. For instance, we may consider the p-adic field $\mathbb{Q}_p$ as a structure for the language of ordered fields supplemented by predicates, P_n for each integer $n\geqslant 2$, by interpreting $P_n(\mathbb{Q}_p)$ to be the set of elements of $\mathbb{Q}_p$ which are n-th powers. In this language every formula is equivalent, modulo the theory of this structure, to one without quantifiers [32]: we say that $\mathbb{Q}_p$ has elimination of quantifiers in this language but this is certainly not true of $\mathbb{Q}_p$ regarded as a structure just for the language of ordered fields (note that the property of being an n-th power in this latter language requires an existential quantifier for its expression).

EXAMPLE 2.4. Suppose that we want a language appropriate for ordered rings (such as the reals). We could take a minimal set consisting just of two binary function symbols, $+$ and $\times$ (for addition and multiplication), together with a binary relation symbol, $\leqslant$, for the order on the ring. But we could also have constant symbols, 0 and 1, a unary function symbol, $-$, for negative (and/or a symbol for subtraction), and binary relation symbols, $\geqslant$, $<$ and $>$, for the relations "associated" to $\leqslant$. If we are dealing with ordered fields note that we cannot introduce a unary function symbol for multiplicative inverse because that operation is not total. We could, however, introduce the symbol as an informal abbreviation in formulas since this partial operation is certainly definable by a formula of our language.

The inductive part of the definition of formula for a general language is identical to that for L_0 and it is in the atomic formulas that the difference lies. First, we say that a **term** of the language L is any expression built up from the variables, the constant symbols (if there are any) and any already constructed terms by using the function symbols (if there are any).

> Any variable is a term. Any constant symbol is a term. If f is an n-ary function symbol and if $t_1, \dots, t_n$ are terms then $f(t_1, \dots, t_n)$ is a term. For example, if L is a language for rings, with function and constant symbols $+, \times, -, 0, 1$, and if we use the axioms for rings to replace some terms by equivalent terms, then a term may be identified with a (non-commutative) polynomial, with integer coefficients, in the variables.

An **atomic formula** is any expression of the form $t_1 = t_2$ where t_1, t_2 are terms or of the form $R(t_1, \dots, t_n)$ where R is an n-ary relation symbol and $t_1, \dots, t_n$ are terms.

EXAMPLE 2.5. Examples of atomic formulas in a language for ordered rings are: $xy + z \leqslant x^2y - 1$ and $xy - yx = 0$.

> One point which we have rather glossed over is: exactly what is a language? (that is, as a mathematical object, what is it?). The simplest answer is to regard a formal language simply as the set of all formulas of the language and that is what we shall do. In practice, however, one usually describes a language by giving the building blocks and the ways in which these may be combined.

Suppose that L is a language. An L-**structure** is a set M together with, for each optional symbol (constant, function, relation) of the language, a specific element of, function on, relation on the set (of course the functions and relations must have the correct arity). Just as we refer to "the group G" rather than "the group $(G, *)$" usually we refer to "the structure M". When we need to be more careful (for example if a certain set is being considered as the underlying set of two different structures for the same language or of structures for two different languages) then we may use appropriate notation, such as that just below.

> An L-structure is $\mathbf{M} = (M; c^{\mathbf{M}}, \dots, f^{\mathbf{M}}, \dots, R^{\mathbf{M}}, \dots)$ where: M is a set; for each constant symbol c of L, $c^{\mathbf{M}}$ is an element of M; for each function symbol f of L with arity, say, n, $f^{\mathbf{M}} : M^n \to M$ is an n-ary function on M; for each relation symbol R of L with arity, say, n, $R^{\mathbf{M}} \subseteq M^n$ is an n-ary relation on M.

Suppose that L is a language. A property of L-structures is **elementary**, or **axiomatisable** or **first-order expressible**, if there is a set, T, of sentences of L such that an L-structure M has that property iff M satisfies (every sentence in) T.

EXAMPLE 2.6. The property of being Abelian is an elementary property of groups since it is equivalent to satisfying the sentence $\forall x \forall y(x * y = y * x)$.

The property of being torsion free is an elementary property of groups since it is equivalent to satisfying the set $\{\forall x(x^n = e \to x = e)\colon\ n \geqslant 1\}$ of sentences. Here x^n is just an abbreviation for $x * x * \cdots * x$ (n "x"s). (Strictly speaking, we should say "The property of being an Abelian group is an elementary property of L-structures, where L is a language appropriate for groups." but, having noted that being a group is an elementary property, it is harmless and natural to extend the terminology in such ways.)

The property of being torsion is not an elementary property of groups. Note that the formal rendition, $\forall x \exists n(x^n = e)$, of this property is not a sentence of the language for groups since the quantifier $\exists n$ should range over the elements of the group, not over the integers. The fact that it is not an elementary property will be proved later (3.8).

EXAMPLE 2.7. Let L be a language for rings. The property of being a field of characteristic $p \neq 0$ is **finitely axiomatisable** (meaning that there is a finite number of sentences, equivalently a single sentence (their conjunction), which axiomatises it).

The property of being (a field) of characteristic zero needs infinitely many sentences of the form $1 + 1 + \cdots + 1 \neq 0$. It follows from the compactness theorem 2.10 that if σ is any sentence in L which is true in all fields of characteristic zero then there is an integer N such that all fields of characteristic greater than N satisfy σ.

Let T be a set of sentences of L (briefly, an L-**theory**). We say that an L-structure M is a **model of** T and write $M \models T$ if M satisfies every sentence of T. A class $\mathcal{C}$ of L-structures is **elementary**, or **axiomatisable**, if there is some L-theory T such that $\mathcal{C} = \mathrm{Mod}(T) = \{M\colon\ M \models T\}$. If $\mathcal{C}$ is any class of L-structures then the **theory of** $\mathcal{C}$, $\mathrm{Th}(\mathcal{C}) = \{\sigma \in L\colon\ \forall M \in \mathcal{C} M \models \sigma\}$, is the set of sentences of L satisfied by every member of $\mathcal{C}$. We write $\mathrm{Th}(M)$ for $\mathrm{Th}(\{M\})$ and call this the (**complete**) **theory of** M ("complete" since for every sentence σ either σ or $\neg\sigma$ is in $\mathrm{Th}(M)$).

EXAMPLE 2.8. We show that the class of algebraically closed fields is an elementary one (in the language for rings). A field K is algebraically closed iff every non-constant polynomial with coefficients in K in one indeterminate has a root in K. At first sight it may seem that there is a problem since we cannot refer to particular polynomials (the coefficients are elements of a field which are, apart from 0, 1 and other integers, not represented by constant symbols or even terms of the language) but then we note that it is enough to have general coefficients. So for each integer $n \geqslant 1$ let τ_n be the sentence $\forall u_0, \ldots, u_n\ (u_n \neq 0 \to \exists x(u_n x^n + \cdots + u_1 x + u_0 = 0))$. Let $S = \{\sigma_f\} \cup \{\tau_n\colon\ n \geqslant 1\}$ where σ_f is a sentence axiomatising the property of being a field. Then a structure K for the language of rings is an algebraically closed field iff $K \models S$.

THEOREM 2.9 (Completeness theorem). *Let T be a set of sentences of L. Then* $\mathrm{Th}(\mathrm{Mod}(T))$ *is the deductive closure of T. In particular, if* $\sigma \in \mathrm{Th}(\mathrm{Mod}(T))$ *then* *there are finitely many sentences $\tau_1, \ldots, \tau_k \in T$ such that every L-structure M satisfies $\tau_1 \wedge \cdots \wedge \tau_k \to \sigma$.*

The term **deductive closure** comes from logic and refers to some notion of formal proof within a precisely defined system. The "completeness" in the name of the theorem refers to the fact that this system is strong enough to capture all consequences of any set of axioms. One can set up such a system, independent of L, in various alternative ways and then the deductive closure of a set T of sentences consists of all sentences which can be generated from T using this system (see, for example, [18] for details). Since a formal proof is of finite length the second statement follows immediately and is one of many ways of expressing the compactness theorem. Here is another.

THEOREM 2.10 (Compactness theorem). *Let T be a set of sentences of L. If every finite subset of T has a model then T has a model.*

Another reformulation of the completeness theorem is that if a set of sentences is **consistent** (meaning that no contradiction can be formally derived from it) then it has a model.

The compactness theorem can be derived independently of the completeness theorem by using the ultraproduct construction and Łos' theorem (and thus this cornerstone of model theory can be obtained without recourse to logic per se).

If two L-structures, M, N satisfy exactly the same sentences of L (that is if $\mathrm{Th}(M) = \mathrm{Th}(N)$) then we say that they are **elementarily equivalent**, and write $M \equiv N$. This is a much coarser relation than that of isomorphism (isomorphic structures are, indeed, elementarily equivalent) but it allows the task of classifying structures up to isomorphism to be split as classification up to elementary equivalence (sometimes much more tractable than classification to isomorphism) and then classification up to isomorphism within elementary equivalence classes (model theory is particularly suited to working within such classes and so, within such a context, one may be able to develop some structure theory). One may also be led to work within an elementary equivalence class even if one is investigating a specific structure M. For M shares many properties with the structures in its elementary equivalence class which is, note, $\mathrm{Mod}(\mathrm{Th}(M))$ but some of these structures may have useful properties (existence of "non-standard" elements, many automorphisms, ...) that M does not. This will be pointless if M is finite since, in this case, any structure elementarily equivalent to M must be isomorphic to M. But otherwise the elementary equivalence class of M contains arbitrarily large structures.

THEOREM 2.11 (Upwards Löwenheim–Skolem theorem). *Let M be an infinite L-structure. Let κ be any cardinal greater than or equal to the cardinality of M and the cardinality of L. Then there is an L-structure, elementarily equivalent to M, of cardinality κ. Indeed M has an elementary extension of size κ.*

(The cardinality of L is equal to the larger of $\aleph_0$ and the number of optional symbols in L.) If M is a substructure of the L-structure N then we say that M is an **elementary substructure** of N (and N is an **elementary extension** of M), and we write $M \prec N$, if for every formula $\phi = \phi(x_1, \ldots, x_n)$ of L and every $a_1, \ldots, a_n \in M$ we have $M \models$

$\phi(a_1, \ldots, a_n)$ iff $N \models \phi(a_1, \ldots, a_n)$. A way to see this definition is as follows. Let $\bar{a}$ be any finite tuple of elements of M. When regarded as a tuple of elements of an arbitrary extension N the L-expressible properties of $\bar{a}$ may change (those expressible by quantifier-free formulas will not but consider, e.g., the even integers as a subgroup of the group of all integers and look at divisibility by 2). If they do not, and if this is true for all tuples $\bar{a}$, then $M \prec N$.

The upwards Löwenheim–Skolem theorem says that if M is an infinite structure then there are arbitrarily large structures with the same first-order properties as M. The next theorem goes in the other direction.

THEOREM 2.12 (Downwards Löwenheim–Skolem theorem). *Let M be an L-structure and let κ be an infinite cardinal with* $\mathrm{card}(L) \leqslant \kappa \leqslant \mathrm{card}(M)$. *Then M has an elementary substructure of cardinality κ. If $A \subseteq M$ and if* $\mathrm{card}(A) \leqslant \kappa$ *then there is an elementary substructure $N \prec M$ of cardinality κ and with $A \subseteq N$.*

EXAMPLE 2.13. Consider the reals $\mathbb{R}$ as a structure for the language, L, of ordered rings. Then $\mathbb{R}$ has a countable elementary substructure $\mathbb{R}'$. Among the properties that $\mathbb{R}'$ shares with $\mathbb{R}$ are the intermediate value theorem for polynomials with coefficients in the structure and factorisability of polynomials into linear and quadratic terms, both these being expressible (at least indirectly) by sentences of the language.

So, for example, if L is a countable language and if T is any L-theory with an infinite model then T has a model of every infinite cardinality.

The second part of the above result can be used as follows. Start with a structure M. Produce an elementary extension M' of M which contains some element or set, B, of elements with some desired properties. Use the result (with $A = M \cup B$) to cut down to a "small" elementary extension of M containing the set B.

EXAMPLE 2.14. All algebraically closed fields of a given characteristic are elementarily equivalent. For, any two algebraically closed fields of the same characteristic and the same uncountable cardinality, say $\aleph_1$ for definiteness, are isomorphic (being copies of the algebraic closure of the rational function field in $\aleph_1$ indeterminates over the prime subfield). If K, L are algebraically closed then, by the Löwenheim–Skolem theorems, there are fields K_1, L_1 elementarily equivalent to K, L respectively and of cardinality $\aleph_1$. So if K and L, and hence K_1 and L_1, have the same characteristic then $K_1 \simeq L_1$ and so $K \equiv K_1 \equiv L_1 \equiv L$ as claimed.

A composition of elementary embeddings is elementary. Furthermore, one as the following.

THEOREM 2.15 (Elementary chain theorem). *Suppose that $M_0 \prec M_1 \prec \cdots \prec M_i \prec \cdots$ are L-structures, each elementarily embedded in the next. Then the union carries an L-structure induced by the structures on the various M_i and, with this structure, it is an elementary extension of each M_i.*

More generally, the direct limit of a directed system of elementary embeddings is an elementary extension of each structure in the system.

PROPOSITION 2.16 (Criterion for elementary substructure). *Let M be a substructure of the L-structure N. Then $M \prec N$ iff for every formula $\phi(x, \bar{y}) \in L$ and for every tuple $\bar{b}$ of elements of M (of the same length as $\bar{y}$) if $N \models \exists x \phi(x, \bar{b})$ then there is $a \in M$ such that $N \models \phi(a, \bar{b})$.*

This is proved by induction on the complexity (of formation) of formulas.

3. Topics

3.1. *Applications of the compactness theorem*

The compactness theorem pervades model theory, directly and through the proofs of numerous other theorems. A common direct use has the following form. We want to produce a structure with a specified set of properties. For some reasons we know that every finite subset of this set of properties can be realised in some structure. The compactness theorem guarantees the existence of a structure which satisfies all these properties simultaneously.

EXAMPLE 3.1. Suppose that T is the theory of (algebraically closed) fields and let σ be a sentence of the language of rings. Suppose that there is, for each of infinitely many distinct prime integers p, an (algebraically closed) field K_p which satisfies σ. Then there is an (algebraically closed) field of characteristic zero which satisfies σ. For consider the set $T \cup \{\sigma\} \cup \{1 + \cdots + 1$ (n "1"s) $\neq 0$: $n \geqslant 1\}$. By assumption every finite subset of this set has a model (K_p for p large enough) and so, by compactness, this set has a model K, as required.

Note the corollary: if σ is a sentence in the language of rings then there is an integer N such that for all primes $p > N$, and for $p = 0$, every algebraically closed field of characteristic p satisfies σ or else every such field satisfies $\neg\sigma$. For otherwise there would be infinitely many primes p such that there is an algebraically closed field of characteristic p satisfying σ and the same would be true for $\neg\sigma$. So there would, by the first paragraph, be an algebraically closed field of characteristic 0 which satisfies σ and also one which satisfies $\neg\sigma$, contradicting the fact 2.14 that all algebraically closed fields of characteristic 0 are elementarily equivalent.

Another type of use has the following form (in fact it is just the kind of use already introduced but omitting explicit enrichment of the language with new constant symbols). In this case we want to produce an element, tuple or even an infinite set of elements with some specified properties. Again, we know for some reasons that these properties are finitely satisfiable. The compactness theorem says that they are simultaneously satisfiable. This type of use usually takes place within the context of the models of a complete theory.

COROLLARY 3.2. *Let M be a structure, $B \subseteq M$, and let $n \geqslant 1$. Suppose that Φ is a set of formulas* ***over*** *B (that is, with parameters in B) which is finitely satisfied in M. Then Φ is*

realised *in an elementary extension of* M. (*That is, there is a tuple* $\bar{c}$ *of elements in some elementary extension of* M *such that* $\bar{c}$ *satisfies every formula in* Φ.)

EXAMPLE 3.3. Perhaps the best known example is the construction of infinitesimals ("construction" is not an accurate term: "pulled out of a hat" is closer to the truth). Consider the reals, $\mathbb{R}$, regarded as an ordered field (in a suitable language). Let $\Phi(x)$ be the set $\{x > 0\} \cup \{x + \cdots + x \leqslant 1 \ (n \text{ "}x\text{"s}){:}\ n \geqslant 1\}$ of formulas which, taken together, describe an element which is strictly greater than zero but less than $\frac{1}{n}$ for each $n \geqslant 1$. Of course, no element of $\mathbb{R}$ satisfies all these formulas but any finite subset of Φ does have a solution in $\mathbb{R}$. So, using the compactness theorem, there is an elementary extension $\mathbb{R}^*$ of $\mathbb{R}$ which contains a realisation of Φ: and such a realisation is an infinitesimal so far as the copy of $\mathbb{R}$ sitting inside $\mathbb{R}^*$ (as an elementary substructure) is concerned.

EXAMPLE 3.4 (*Bounds in polynomial rings*). There is a host of questions concerning ideals in polynomial rings, of which the following is a basic example. Consider the polynomial ring $R = K[X_1, \dots, X_t]$ where K is a field. Let $f, g_1, \dots, g_n \in R$. If f belongs to the ideal $I = \langle g_1, \dots, g_n \rangle$ then there are polynomials $h_1, \dots, h_n \in R$ such that $f = h_1 g_1 + \cdots + h_n g_n$. There is no a priori bound on the (total) degrees of the h_i which might be needed but if f does belong to I then there are, in fact, such polynomials h_i with degree bounded above by a function which depends only on the degree of f and the degrees of the g_i. Similarly, if f belongs to the radical of I – that is, some power of f belongs to I – then the minimal such power can be bounded above by a function of the degree of f and the degrees of the g_i only. In many cases such bounds arise directly from explicit computation procedures but existence of such bounds often may be obtained, often very easily, by use of the compactness theorem. See, e.g., [15] and references therein. Such methods are used and extended in [11] to obtain Lang–Weil-type estimates on the sizes of definable subsets of finite fields $\mathbb{F}_{p^n}$ as $n \to \infty$.

EXAMPLE 3.5 (*Polynomial maps*). Suppose that V is an algebraic subvariety of $\mathbb{C}^n$ (that is, V is the set, $V_{\mathbb{C}}(g_1, \dots, g_k)$, of common zeroes, in $\mathbb{C}^n$, of some set, $g_1, \dots, g_k$, of polynomials in $\mathbb{C}[X_1, \dots, X_n]$) and let $f : V \to V$ be a polynomial map (that is $f(\bar{a}) = (f_1(\bar{a}), \dots, f_n(\bar{a}))$ for some polynomials $f_1, \dots, f_n$). Suppose that f is injective. Then f is onto [3]. This can be proved as follows.

Notice that the assertion is true if we replace $\mathbb{C}$ by a finite field (simply because V is then a finite set). It follows that the assertion is true if $\mathbb{C}$ is replaced by the algebraic closure of any finite field (for this is a union of finite fields F and, if F is large enough to contain the coefficients of the polynomials g_i, f_j, then $V_F(g_1, \dots, g_k)$ is closed under f).

Next observe that the assertion is expressible by a sentence in the language of fields. Of course the polynomials $g_1, \dots, g_k$ have to be replaced by polynomials with general coefficients as do $f_1, \dots, f_n$ (as in the argument that "algebraically closed" is axiomatisable) and the argument must be applied to each member of an infinite set of sentences (since any single sentence refers to polynomials of bounded total degree) but, having noted this, we may easily express the conditions "$(x_1, \dots, x_n) \in V(g_1, \dots, g_k)$" (meaning $(x_1, \dots, x_n) \in V_K(g_1, \dots, g_k)$, where now K can be any field), "f is surjective" and "f is injective".

The fact that I have not written down the relevant sentences is rather typical in model theory since, with some experience, it can be clear that certain conditions are expressible by sentences (often rather indigestible if actually written down) of the relevant language (it is also often clear "by compactness" that certain properties are not so expressible).

Thus we have our sentence, σ (corresponding to some bound on the degrees), true in each field $\widetilde{\mathbb{F}_p}$ which is the algebraic closure of a field of prime characteristic. By 3.1 it follows that there is an algebraically closed field of characteristic 0 which satisfies σ. Finally we use the fact, proved earlier, that all algebraically closed fields of a given characteristic are elementarily equivalent and hence we deduce that $\mathbb{C}$ satisfies σ and, therefore, satisfies the original assertion. (Indeed, this argument shows that every algebraically closed field satisfies it.)

3.2. *Morphisms and the method of diagrams*

A **morphism** $\alpha:\mathbf{M}\to\mathbf{N}$ between L-structures is, of course, just a structure-preserving map.

> Precisely, we require that: for each constant symbol c of L we have $\alpha(c^{\mathbf{M}}) = c^{\mathbf{N}}$; for each n-ary function symbol f of L and n-tuple $\bar{a}$ from M we have $\alpha(f^{\mathbf{M}}(\bar{a})) = f^{\mathbf{N}}(\alpha\bar{a})$; for each n-ary relation symbol R of L and n-tuple $\bar{a}$ from M we have $R^{\mathbf{M}}(\bar{a})$ implies $R^{\mathbf{N}}(\alpha\bar{a})$ (where $\alpha\bar{a}$ denotes $(\alpha a_1,\dots,\alpha a_n)$ if $\bar{a}=(a_1,\dots,a_n)$).
> If the language contains relation symbols then a bijective morphism need not be an isomorphism (exercise: give a counterexample in posets) and for **isomorphism** one needs the stronger condition "$R^{\mathbf{M}}(\bar{a})$ iff $R^{\mathbf{N}}(\alpha\bar{a})$".

A **substructure**, $\mathbf{N}$, of an L-structure $\mathbf{M}$ is given by a subset N of M which contains all the interpretations, $c^{\mathbf{M}}$, of constant symbols in M and which is closed under all the functions, $f^{\mathbf{M}}$, on M. Then we make it an L-structure by setting $c^{\mathbf{N}} = c^{\mathbf{M}}$, $f^{\mathbf{N}} = f^{\mathbf{M}} \upharpoonright N^n$ and $R^{\mathbf{N}} = R^{\mathbf{M}} \cap N^n$ for each constant symbol, c, n-ary function symbol, f, and n-ary relation symbol R of L. More generally we have the notion of an **embedding** of L-structures: a monic morphism which satisfies the additional condition "$R^{\mathbf{M}}(\bar{a})$ iff $R^{\mathbf{N}}(\alpha\bar{a})$" seen in the definition of isomorphism above.

The "method of diagrams" is a means of producing morphisms between L-structures. Suppose that M is an L-structure. Enrich the language L by adding a new constant symbol $[a]$ for every element $a \in M$, thus obtaining the language denoted L_M. Of course M has a natural enrichment to an L_M-structure, given by interpreting $[a]$ as a for each $a\in M$. The **atomic diagram** of M is the collection of all L_M-sentences of the form $t_1 = t_2$, $t_1 \neq t_2$, $R(t_1,\dots,t_n)$, $\neg R(t_1,\dots,t_n)$ satisfied by M, where the t_i are terms of L_M and R is any (n-ary) relation symbol of L. This is the collection of all basic positive and negative relations (in the informal sense) which hold between the elements of M. The **positive atomic diagram** of M is the subset of the atomic diagram containing just the positive atomic sentences (i.e. those of the forms $t_1 = t_2$ and $R(t_1,\dots,t_n)$). For instance, if M is a ring then the positive atomic diagram is (equivalent to) the multiplication and addition tables of M and the atomic diagram further contains a record of all polynomial combinations of elements of M which are non-zero. The **full diagram** of M is the L_M-theory of M (i.e. all sentences of L_M satisfied by M).

Let D be any of the above "diagrams" and let $T' = \mathrm{Th}(M) \cup D$ – a set of L_M-sentences (by $\mathrm{Th}(M)$ I mean the theory of the L-structure M, not the enriched L_M-structure). If N' is any model of T' then first note that the reduction of the L_M-structure N' to an L-structure N (i.e. forget which elements $[a]^{N'}$ interpret the extra constant symbols $[a]$) satisfies $\mathrm{Th}(M)$ and hence $N \equiv M$. Second, we have a natural map, α, from M to N ($= N'$ as a set) given by taking $a \in M$ to the interpretation, $[a]^{N'}$, of the corresponding constant symbol, $[a]$, in N'. Because N' satisfies at least the positive atomic diagram of M it is immediate that α is a morphism of L-structures.

THEOREM 3.6. *Let M be an L-structure and let D be any of the above diagrams. Let N' be a model of* $\mathrm{Th}(M) \cup D$ *and let N be the reduct of N' to L. Then N is elementarily equivalent to M. Furthermore, if $\alpha : M \to N$ is the map defined above then*:

(a) *if D is the positive atomic diagram of M then α is a morphism*;
(b) *if D is the atomic diagram of M then α is an embedding*;
(c) *if D is the full diagram of M then α is an elementary embedding.*

There are variants of this. For instance, if D is the atomic diagram of M and if we take N' to be a model of just this (so drop $\mathrm{Th}(M)$) then we have that α is an embedding to the L-structure N (which need not be elementarily equivalent to M). Or we can add names just for elements of some substructure A of M and then we obtain a morphism α from that substructure to N.

3.3. *Types and non-standard elements*

Suppose that M is an L-structure. Let $\bar{a} = (a_1, \ldots, a_n)$ be a tuple of elements of M. The **type** of $\bar{a}$ in M is the set of all L-formulas satisfied by $\bar{a}$ in M. Thus it is the "L-description" of this tuple (or, better, of how this tuple sits in M). If $f \in \mathrm{Aut}(M)$ is an automorphism then the type of $f\bar{a} = (fa_1, \ldots, fa_n)$ is equal to the type of $\bar{a}$ since, for any formula ϕ, we have $M \models \phi(\bar{a})$ iff $M \models \phi(f\bar{a})$ because f is an isomorphism.

More generally, we may want a description of how $\bar{a}$ sits in M with respect to some fixed set, $B \subseteq M$, of parameters. So we define the **type** of $\bar{a}$ in M **over** B to be $\mathrm{tp}^M(\bar{a}/B) = \{\phi(\bar{x}, \bar{b})\colon\ \phi \in L, \bar{b} \text{ in } B\}$ – the set of all formulas (in some fixed tuple, $\bar{x}$, of free variables matching $\bar{a}$) with parameters from B satisfied by $\bar{a}$ in M. We write $\mathrm{tp}(\bar{a})$ for $\mathrm{tp}(\bar{a}/\emptyset)$. It is immediate that if N is an elementary extension of M then $\mathrm{tp}^N(\bar{a}/B) = \mathrm{tp}^M(\bar{a}/B)$ and so we often drop the superscript.

Note that $\mathrm{tp}(\bar{a}/B)$ has the following properties: it is a set of formulas with parameters from B in a fixed sequence of n free variables; it is closed under conjunction and under implication; it is consistent (it does not contain any contradiction such as $x_1 \neq x_1$); it is maximal such (it is "complete", equivalently, for every formula $\phi(\bar{x}, \bar{b})$ with parameters from B either this formula or its negation is contained in $\mathrm{tp}(\bar{a}/B)$). Any set of formulas satisfying these conditions is called an n-**type over** B.

THEOREM 3.7. *Let M be a structure, $B \subseteq M$, and let $n \geqslant 1$. Suppose that p is an n-type over B. Then p is **realised** in an elementary extension of M. That is, there is $N \succ M$ and $c_1, \ldots, c_n \in N$ such that* $\mathrm{tp}(\bar{c}/B) = p$. (*Then $\bar{c}$ is said to be a **realisation** of p.*)

We say "every type is realised in an elementary extension". If Φ is any **consistent** set of formulas (that is, the closure of Φ under conjunction and implication contains no contradiction) in $\bar{x}$ with parameters from B then, by Zorn's lemma, Φ is contained in a maximal consistent such set, that is, in a type over B (sometimes one says that the **partial type** Φ is contained in a **complete type**). This gives 3.2 above.

All the above goes equally for infinite tuples. In this way one can use compactness to produce not just elements, but structures as in the subsection on the Method of Diagrams, for instance.

EXAMPLE 3.8. We claimed earlier that the property of being a torsion group is not elementary: now we justify that claim. Let G be a torsion group for which there is no bound on the order of its elements (for example, let G be the direct sum of the finite groups $\mathbb{Z}/n\mathbb{Z}$ for $n \geqslant 2$). Let $\Phi = \Phi(x) = \{x^n \neq e\colon n \geqslant 1\}$ where e denotes the identity of G and where x^n is an abbreviation for the term which is a product of n x's (a slightly dangerous abbreviation since the whole point is that we cannot refer to general integers n in our formulas!). Then Φ is a partial type (in G) since any given finite subset of Φ is realised by an element of G which has high enough order. Therefore Φ is realised by some element, c say, in some elementary extension, G' say, of G. In particular the group G' is elementarily equivalent to G but it is not a torsion group (since $\Phi(c)$ says that c has infinite order). Thus the property of being torsion is not an elementary one.

3.4. *Algebraic elements*

Suppose that M is a structure, that $a \in M$ and that $B \subseteq M$. We say that a is **algebraic over** B if there is a formula $\phi(x, \bar{b})$ with parameters $\bar{b}$ from B such that a satisfies this formula, we write $M \models \phi(a, \bar{b})$, and such that the solution set, $\phi(M, \bar{b}) = \{c \in M\colon\ M \models \phi(c, \bar{b})\}$, of this formula in M is finite. In this case, if M' is an elementary extension of M then the solution sets $\phi(M', \bar{b})$ and $\phi(M, \bar{b})$ are equal (exercise: use that M satisfies a sentence which gives the size of the solution set of $\phi(x, \bar{b})$) so the relation of being algebraic over a set is unchanged by moving to an elementary extension.

If $M \models \phi(a, \bar{b})$ and f is an automorphism of M which fixes B pointwise then $M \models \phi(fa, \bar{b})$. Hence if a is algebraic over B then a has only finitely many conjugates under the action of $\mathrm{Aut}_B M$, by which we denote the group of automorphisms of M which fix B pointwise. If M is sufficiently saturated (see the subsection on saturated structures), but not for general structures M, the converse is true.

A tighter relation is that of being **definable over** a set B: this is as "algebraic over" but with the stronger requirement that the element or tuple is the *unique* solution of some formula with parameters from B (equivalently is fixed by all elements of $\mathrm{Aut}_B(M)$ in a sufficiently saturated model M).

EXAMPLE 3.9. If M is a vector space over a field then a is algebraic over B iff a is in the linear span of B iff a is definable over B. If M is an algebraically closed field then a is algebraic (in the model-theoretic sense) over B iff a is algebraic (in the usual sense) over B. Any element in the subfield, $\langle B\rangle$, generated by B is definable over B but, if the characteristic of the field is $p > 0$ then one also has that any p^n-th root of any element of $\langle B\rangle$ is definable over B.

3.5. *Isolated types and omitting types*

If we fix an integer $n \geqslant 1$ and a subset B of an L-structure M then the set, $S_n(B)$, of all n-types over B carries a natural topology which has, for a basis of clopen sets, the sets of the form $\mathcal{O}_{\phi(\bar{x},\bar{b})} = \{p \in S_n(B)\colon\ \phi(\bar{x},\bar{b}) \in p\}$. We also denote this set $S_n^T(B)$ where $T = \mathrm{Th}(M)$ to emphasise that the notion of "type" makes sense only relative to a complete theory. This is in fact the Stone space (the space of all ultrafilters) of the Boolean algebra of equivalence classes of formulas with free variables $\bar{x}$ (formulas are equivalent if they define the same subset of M and the ordering is implication). This space is compact (by the compactness theorem).

A type p is **isolated** or **principal** if there is some formula ϕ in p which proves every formula in p: for every $\psi \in p$ we have $M \models \forall\bar{x}(\phi(\bar{x}) \to \psi(\bar{x}))$. In this case $\mathcal{O}_\phi = \{p\}$ and p is an isolated point of the relevant Stone space. Such a type must be realised in every model: for by consistency of p we have $M \models \exists\bar{x}\phi(\bar{x})$, say $M \models \phi(\bar{c})$, and then, since ϕ generates p, we have $M \models \psi(\bar{c})$ for every $\psi \in p$, that is $M \models p(\bar{c})$ and $\bar{c}$ realises p, as required. For countable languages there is a converse.

THEOREM 3.10 (Omitting types theorem). *Let L be a countable language, let M be an L-structure and let $p \in S_n(\emptyset) = S_n^{\mathrm{Th}(M)}(\emptyset)$. If p is a non-isolated type then there is an L-structure N elementarily equivalent to M which* ***omits*** *p (i.e. which does not realise p).*

This is extended to cover the case of types over a subset B by enriching L by adding a name for each element of B and then applying the above result (assuming, of course, that B is countable so that the enriched language is countable). There are extensions of this result which allow sets of types to be omitted simultaneously [10, 2.2.15, 2.2.19], [24, 7.2.1]. The result is not true without the countability assumption [10, after 2.2.18].

3.6. *Categoricity and the number of models*

A theory is $\aleph_0$-**categorical** if it has just one countably infinite model up to isomorphism. More generally it is κ-**categorical** where κ is an infinite cardinal if it has, up to isomorphism, just one model of cardinality κ. If an L-theory is κ-categorical for any cardinal $\kappa \geqslant \mathrm{card}(L)$ then, by the Löwenheim–Skolem theorems, it must be complete.

EXAMPLE 3.11. Any two atomless Boolean algebras (equivalently, Boolean rings with zero socle) are elementarily equivalent since there is, up to isomorphism, just one such structure of cardinality $\aleph_0$.

THEOREM 3.12 (Morley's theorem). *Suppose that L is a countable language and that T is a complete L-theory which is κ-categorical for some uncountable cardinal κ. Then T is λ-categorical for every uncountable cardinal λ.*

The situation for $\kappa = \aleph_0$ is different. For instance, the theory of algebraically closed fields of characteristic zero is uncountably categorical but not $\aleph_0$-categorical. Indeed, examples show that $\aleph_0$-categoricity, uncountable categoricity and their negations can occur in all four combinations.

THEOREM 3.13. *Suppose that L is a countable language and that T is a complete L-theory which has an infinite model. Then the following are equivalent*:

(i) *T is $\aleph_0$-categorical*;

(ii) *for each $n \geqslant 1$ the Stone space $S_n(\emptyset)$ is finite*;

(iii) *each type in each Stone space $S_n(\emptyset)$ $(n \geqslant 1)$ is isolated*;

(iv) *for each countable model M of T and for each integer $n \geqslant 1$ there are just finitely many orbits of the action of* $\mathrm{Aut}(M)$ *on n-tuples*;

(v) *for each finite tuple, $\bar{x}$, of variables there are just finitely many formulas with free variables $\bar{x}$ up to equivalence modulo T.*

For instance, if some space $S_n(\emptyset)$ is infinite then, by compactness, it contains a non-isolated type p. Then there will be a countably infinite model which realises p but also a countable infinite model which omits p and so the theory cannot be $\aleph_0$-categorical.

EXAMPLE 3.14. Any $\aleph_0$-categorical structure is **locally finite** in the sense that every finitely generated substructure is finite. For, let M be $\aleph_0$-categorical, let $\bar{a}$ be a finite sequence of elements from M and let b, c be distinct elements of the substructure, $\langle \bar{a} \rangle$, generated by $\bar{a}$. Then $\mathrm{tp}(b/\bar{a}) \neq \mathrm{tp}(c, \bar{a})$ since the type "contains the expression of b (resp. c) in terms of $\bar{a}$". Therefore $\mathrm{tp}(b, \bar{a}) \neq \mathrm{tp}(c, \bar{a})$ are distinct $(1 + n)$-types, where n is the length of the tuple $\bar{a}$. But $S_{1+n}(\emptyset)$ is finite and, therefore, $\langle \bar{a} \rangle$ is finite. Indeed the argument shows that M is **uniformly locally finite** since if $s_n = |S_n(\emptyset)|$ then any substructure of M generated by n elements has cardinality bounded above by s_{n+1}.

A more general question in model theory is: given a complete theory T and an infinite cardinal κ what is the number, $n(\kappa, T)$, up to isomorphism, of models of T of cardinality κ? Remarkably complete results on this question and on the connected question of the existence or otherwise of structure theorems for models of T have been obtained by Shelah and others (see [50]).

3.7. *Prime and atomic models*

Suppose that M is an L-structure. Set $T = \mathrm{Th}(M)$. The elementary equivalence class, $\mathrm{Mod}(T)$, of M, equipped with the elementary embeddings between members of the class forms a category (not very algebraically interesting unless, say, T has elimination of quantifiers since there are rather few morphisms). A model of T is a **prime** model if it embeds elementarily into every model of T and is **atomic** if every type realised in it is isolated (hence is a type which must be realised in every model).

THEOREM 3.15. *Let T be a complete theory in the countable language L and suppose that T has an infinite model. Then the L-structure M is a prime model of T iff M is a countable atomic model of T. Such a model of T exists iff, for every integer $n \geqslant 1$, the set of isolated points in $S_n(\emptyset)$ is dense in $S_n(\emptyset)$ (in particular this will be so if each $S_n(\emptyset)$ is countable). If T has a prime model then this model is unique to isomorphism.*

The proof of the second statement involves a back-and-forth construction: we describe this construction in the next section.

3.8. *Back-and-forth constructions*

This is a method for producing morphisms between structures. Suppose that M is a countably infinite L-structure and that N is an L-structure. Enumerate the elements of M in a sequence $a_0, a_1, \dots, a_i, \dots$ indexed by the natural numbers. If we are to embed M in N then we need to find an element, b_0, of N such that the isomorphism type of the substructure, $\langle b_0 \rangle$, of N generated by b_0 is isomorphic to the substructure, $\langle a_0 \rangle$, of M generated by a_0. Supposing that there is such an element, fix it. So now we have a "partial embedding" from M to N (the map with domain $\{a_0\}$ (or, if one prefers, $\langle a_0 \rangle$) sending a_0 to b_0). Now we need an element b_1 of N to which to map a_1. It is necessary that the substructures $\langle b_0, b_1 \rangle$ and $\langle a_0, a_1 \rangle$ be isomorphic by the map sending a_0 (resp. a_1) to b_0 (resp. b_1). If there is such an element, b_1 say, fix it. Continue in this way. At the typical stage we have images $b_0, \dots, b_n$ for $a_0, \dots, a_n$ and we need to find an element b_{n+1} of N which "looks the same over $b_0, \dots, b_n$ as a_{n+1} does over $a_0, \dots, a_n$". In the limit we obtain an embedding of M into N.

This is the shape of a "forth" construction. Of course, the key ingredient is missing: how can we be sure that the elements b_i of the sort we want exist? And, of course, that must somehow flow from the hypotheses that surround any particular application of this construction.

We may want a stronger conclusion: that the constructed embedding of M into N be an elementary embedding. In that case the requirement that $\langle a_0, \dots, a_{n+1} \rangle$ be isomorphic (via $a_i \mapsto b_i$) to $\langle b_0, \dots, b_n \rangle$ must be replaced by the stronger requirement $\mathrm{tp}^M(a_0, \dots, a_{n+1}) = \mathrm{tp}^N(b_0, \dots, b_{n+1})$ (note that, in this case we must assume $M \equiv N$).

EXAMPLE 3.16. The random graph is formed (with probability 1) from a countably infinite set of vertices by joining each pair of vertices by an edge with probability $\frac{1}{2}$. It is characterised as the unique countable graph such that, given any finite, disjoint, sets, U, V, of vertices, there is a point not in $U \cup V$ which is joined to each vertex of U and to no vertex of V. An easy "forth" argument, using this characterising property, shows that every countable graph embeds in the random graph (and the corresponding back-and-forth argument shows the uniqueness of this graph up to isomorphism).

For a back-and-forth construction, we suppose that both M and N are countably infinite and we want to produce an isomorphism from M to N. For this we interlace the forth construction with the same construction going in the other direction. That is, on, say, even-numbered steps, we work on constructing the map α and on odd-numbered steps we work on constructing its inverse α^{-1} (in order to ensure that the resulting map α is surjective).

EXAMPLE 3.17. Let $\mathbb{Q}$ denote the rationals regarded as a partially ordered set: as such it is an example of a densely linearly ordered set without endpoints. Finitely many axioms in a language which has just one binary relation symbol suffice to axiomatise this notion. Let T_{dlo} denote the theory of densely linearly ordered sets without endpoints. Then any countable (necessarily infinite) model of T_{dlo} is isomorphic to $\mathbb{Q}$. This is a straightforward back-and-forth argument. It follows that T_{dlo} is a complete theory: all densely linearly ordered sets without endpoints are elementarily equivalent. For let $M \models T_{\text{dlo}}$. By the downwards Löwenheim–Skolem theorem there is a countable M_0 elementarily equivalent to M. But then $M_0 \simeq \mathbb{Q}$ and so $M \equiv M_0 \equiv \mathbb{Q}$, which is enough.

3.9. *Saturated structures*

A saturated structure is one which realises all types of a particular sort: a "fat" structure as opposed to the "thin" atomic structures which realise only those types which must be realised in every model. For instance, an L-structure M is **weakly saturated** if it realises every type (in every Stone space $S_n(\emptyset)$) over the empty set. An L-structure M is κ-**saturated**, where $\kappa \geqslant \aleph_0$ is a cardinal, if for every subset $A \subseteq M$ of cardinality strictly less than κ and every n (it is enough to ask this for $n = 1$) every type in $S_n(A)$ is realised in M. Such structures always exist by a (possibly transfinite) process of realising types in larger and larger models and they provide a context in which "every consistent situation (of a certain "size") can be found". More precisely we have the following (which is proved by a "forth" construction).

THEOREM 3.18 (saturated implies universal). *Let M be a κ-saturated L-structure. Then every model of the theory of M of cardinality strictly less than κ elementarily embeds in M.*

An L-structure M is **saturated** if it is $\text{card}(M)$-saturated.

THEOREM 3.19 (saturated implies homogeneous). *Suppose that M is a saturated structure and that $\bar{a}, \bar{b}$ are matching, possibly infinite, sequences of elements of M of cardinality strictly less than* $\text{card}(M)$ *and with* $\text{tp}^M(\bar{a}) = \text{tp}^M(\bar{b})$. *Then there is an automorphism, α, of M with $\alpha(\bar{a}) = \bar{b}$.*

Thus, in a saturated structure, types correspond to orbits of the automorphism group of M.

COROLLARY 3.20. *Let M be a saturated structure of cardinality κ and suppose that the theory of M has complete elimination of quantifiers (for this see later). Suppose that A, B*

are substructures of M each generated by strictly fewer than κ elements and suppose that $\beta : A \to B$ is an isomorphism. Then β extends to an automorphism of M.

Often in model theory it is convenient to work inside a ("monster") model which embeds all "small" models and which is (somewhat) homogeneous in the above sense. With some assumptions on $\mathrm{Th}(M)$ one knows that there are saturated models of $\mathrm{Th}(M)$ of arbitrarily large cardinality but, for an arbitrary theory, unless some additional (to ZFC) set-theoretic assumptions are made, such models might not exist. One does have, however, without any additional set-theoretic assumptions, arbitrarily large elementary extensions which are, for most purposes, sufficiently saturated (see [24, Section 10.2]) to serve as a "universal domain" within which to work.

Related ideas can make sense in contexts other than the category of models of a complete theory, indeed, even in non-elementary classes.

EXAMPLE 3.21 (*Universal locally finite groups*). A group G is **locally finite** if every finitely generated subgroup of G is finite. A group G is a **universal locally finite** group if G is locally finite, if every finite group embeds into G and if, whenever G_1, G_2 are finite subgroups of G and $f : G_1 \to G_2$ is an isomorphism, then there is an inner automorphism of G which extends f.

For each infinite cardinal κ there exists a universal locally finite group of cardinality κ. Up to isomorphism there is just one countable universal locally finite group (an easy back-and-forth argument). This structure is not, however, $\aleph_0$-categorical since it is not uniformly locally finite nor is it weakly saturated since it has elements of unbounded finite order but no element of infinite order. If κ is uncountable then, [33], there are many non-isomorphic universal locally finite groups of cardinality κ. Each locally finite group of cardinality κ can be embedded in a universal locally finite group of cardinality λ for each $\lambda \geqslant \kappa$.

EXAMPLE 3.22 (*Saturated structures have injectivity-type properties*). Suppose that M, N are L-structures for some language L. Suppose that N_0 is a substructure of N and that $f : N_0 \to M$ is a morphism. Suppose that for every finite tuple $\bar{a}$ from N there is an extension of f to a morphism from the substructure, $\langle N_0, \bar{a} \rangle$, of N generated by N_0 together with $\bar{a}$, to M. Let M' be a sufficiently saturated elementary extension of M (precisely, M' should be $(|N| + |L|)^+$-saturated). Then there is an extension of f to a morphism from N to M'.

To see this, enumerate N as $\{c_\alpha\}_{\alpha \in I} \cup \{d_\beta\}_{\beta \in J}$ with $\{c_\alpha\}_{\alpha \in I} = N_0$. Let $\Phi = \Phi(\{x_\beta\}_{\beta \in J}) = \{\phi(f(c_{\alpha_1}), \dots, f(c_{\alpha_m}), x_{\beta_1}, \dots, x_{\beta_n})$: ϕ is atomic and $N \models \phi(c_{\alpha_1}, \dots, c_{\alpha_m}, d_{\beta_1}, \dots, d_{\beta_n})\}$. By assumption Φ is a partial type (any finite subset of Φ mentions variables corresponding to only finitely many elements outside N_0 and then a morphism extending f and with domain including these elements provides us with a realisation of this finite subset since, as is easily seen, morphisms preserve the truth of atomic formulas). Since M' is sufficiently saturated it realises Φ, say $M' \models \Phi(\{b_\beta\}_{\beta \in J})$, and we extend f by mapping d_β to b_β for $\beta \in J$.

3.10. *Ultraproducts*

The ultraproduct construction has been extensively used in applications of model theory to algebra. In many, though by no means all, cases an appeal to the existence of suitably saturated extensions would serve equally well but the ultraproduct construction does have the advantage of being a purely algebraic one (although perhaps "construction" is not quite the right word since it inevitably appeals to Zorn's lemma at the point where a filter is extended to an ultrafilter).

We start with a set, $\{M_i\colon i \in I\}$, of L-structures, M_i, indexed by a set I. The ultraproduct construction produces a kind of "average" of the M_i. Let D be an ultrafilter on I.

> A **filter on** a set I is a filter, D, in the Boolean algebra, $\mathcal{P}(I)$, of all subsets of I. That is: $\emptyset \notin D$; $I \in D$; if $X, Y \in D$ then $X \cap Y \in D$; if $X \subseteq Y \subseteq I$ and $X \in D$ then $Y \in D$. An **ultrafilter** is a maximal filter and is characterised by the further condition: if $X \subseteq I$ then either $X \in D$ or $X^c \in D$. An ultrafilter, D, on I is **principal** if there is $i_0 \in I$ such that $D = \{X \subseteq I\colon i_0 \in X\}$. Any ultrafilter which contains a finite set must be principal. An example of a filter, sometimes called the **Fréchet filter**, on an infinite set I is the set of all cofinite sets ($X \subseteq I$ is **cofinite** if X^c is finite). Any filter can be extended to an ultrafilter: but, unless the ultrafilter is principal, there is no explicit way to describe its members (the existence of non-principal ultrafilters is just slightly weaker than the axiom of choice). Any ultrafilter containing the Fréchet filter is non-principal.

The **ultraproduct**, $\prod_{i\in I} M_i/D$, of the M_i $(i \in I)$ with respect to the ultrafilter D on I is, as a set, the product $\prod_{i\in I} M_i$ factored by the equivalence relation $\sim = \sim_D$ given by $(a_i)_{i\in I} \sim (b_i)_{i\in I}$ iff $a_i = b_i$ "D-almost-everywhere", that is, if $\{i \in I\colon a_i = b_i\} \in D$. Then the set $\prod_{i\in I} M_i/D$ is made into an L-structure by defining the constants, functions and relations pointwise almost everywhere (the defining properties of a filter give that this is well-defined).

EXAMPLE 3.23. Suppose that the M_i are groups. Then $\prod_i M_i/D$ is the product group factored by the normal subgroup consisting of all tuples $(a_i)_i$ which are equal to the identity on D-almost-all coordinates: $\prod_i M_i/D \equiv (\prod_i M_i)/\{(a_i)_i\colon \{i \in I\colon a_i = e_i\} \in D\}$ (where e_i denotes the identity element of M_i).

EXAMPLE 3.24. Suppose that the M_i are fields. Let $a = (a_i)_i/\sim \in \prod_i M_i|D$ be a non-zero element of the ultraproduct. Define the element $b = (b_i)_i/\sim$ by setting $b_i = a_i^{-1}$ for each i with $a_i \neq 0$ and setting b_i to be, say, 0 on all other coordinates. Since $J = \{i \in I\colon a_i \neq 0\} \in D$ we have $\{i \in I\colon a_i b_i = 1\} \in D$ and hence $ab = 1$. That is, $\prod_i M_i/D$ is a field.

If $D = \{X \subseteq I\colon i_0 \in X\}$ is a principal ultrafilter then the ultraproduct $\prod_i M_i/D$ is isomorphic to M_{i_0} so the interesting case is when D is non-principal.

THEOREM 3.25 (Łos' theorem). *Let M_i $(i \in I)$ be a set of L-structures and let D be an ultrafilter on I. Set $M^* = \prod_{i\in I} M_i/D$ to be the ultraproduct. If σ is a sentence of L*

then $M^* \models \sigma$ *iff* $\{i \in I\colon\ M_i \models \sigma\} \in D$ (*that is, iff "D-almost all" coordinate structures satisfy* σ).

More generally, if $\phi(x_1, \ldots, x_n)$ *is a formula and if* $a^1, \ldots, a^n \in M^*$ *with* $a^j = (a_i^j)_i/\sim$ $(j = 1, \ldots, n)$ *then* $M^* \models \phi(a^1, \ldots, a^n)$ *iff* $\{i \in I\colon\ M_i \models \phi(a_i^1, \ldots, a_i^n)\} \in D$.

EXAMPLE 3.26. Let $\mathcal{P}$ be an infinite set of non-zero prime integers and let K_p be a field of characteristic p for $p \in \mathcal{P}$. Let D be a non-principal ultrafilter on $\mathcal{P}$ and let K be the corresponding ultraproduct $\prod_p K_p/D$. Then K has characteristic zero and is an infinite model of the theory of finite fields (for such **pseudofinite** fields see [2]). For example, it has, for each integer $n \geqslant 1$, just one field extension of degree n since this is true, and can be expressed (with some work) in a first-order way of finite fields.

EXAMPLE 3.27 (*Ultraproduct proof of the compactness theorem*). Let T be a set of sentences, each finite subset of which has a model. Let X be the set of all finite subsets of T. For each $S \in X$ choose a model M_S of S. Given any $\sigma \in T$ let $\langle\sigma\rangle = \{S \in X\colon\ \sigma \in S\}$. Note that the intersection of any finitely many of these sets is non-empty and so $F = \{Y \subseteq nX\colon\ \langle\sigma_1\rangle \cap \cdots \cap \langle\sigma_n\rangle \subseteq Y$ for some $\sigma_i \in T\}$ is a filter. Let D be any ultrafilter containing F. Then $\prod_{S\in X} M_S/D$ is a model of T. For let $\sigma \in T$. Then $\langle\sigma\rangle \in D$ and, since $\sigma \in S$ implies $M_S \models \sigma$, we have $\{S\colon\ M_S \models \sigma\} \in D$ and so $\prod_{S\in X} M_S/D \models \sigma$ (by Łos' theorem), as required.

If all M_i are isomorphic to some fixed L-structure M then we denote the ultraproduct by M^I/D and call it an **ultrapower** of M. In this case the point of the construction is not to produce an "average structure" but to create "nonstandard" elements of M. For instance, any ultrapower of the real field $\mathbb{R}$ by a non-principal ultrafilter will contain infinitesimals.

A variant of the construction is to allow D to be any filter in $\mathcal{P}(I)$: the result is then called a **reduced product**. For reduced products there is a (considerably) weaker version of Łos' theorem (see [10, Section 6.2]).

EXAMPLE 3.28 (*Embeddings into general linear groups*). The following example of the use of ultraproducts is from [34]. It makes use of the fact that the ultraproduct construction, when extended in the obvious way to morphisms, is functorial. A group G is **linear of degree** n if there is an embedding of G into the general linear group $\mathrm{GL}(n, K)$ for some field K.

Suppose that G is a group such that every finitely generated subgroup is linear of degree n. Then G is linear of degree n.

For the proof, let G_i $(i \in I)$ be the collection of finitely generated subgroups of G. For each i choose an embedding $f_i : G_i \to \mathrm{GL}(n, K_i)$ for some field K_i. Let D be a non-principal ultrafilter on I and let $f = \prod_i f_i/D : \prod_i G_i/D \to \prod_i \mathrm{GL}(n, K_i)/D$. It is easy to see (for example, think in terms of matrix representations of elements of $\mathrm{GL}(n, -)$) that $\prod_i \mathrm{GL}(n, K_i)/D \simeq \mathrm{GL}(n, \prod_i K_i/D)$. By Łos' theorem $\prod_i K_i/D$ is a field. It remains to see an embedding of G into $\prod_i G_i/D$: at this point we realise that the ultrafilter D should not be arbitrary non-principal. Given an element $g \in G$ let $[g]$ denote the set of all $i \in I$ such that $g \in G_i$ and let $F = \{[g]\colon\ g \in G\}$. Since $[g_1] \cap \cdots \cap [g_t] = \{i\colon\ g_1, \ldots, g_t \in G_i\}$ the set F has the finite intersection property. (The intersection of any finitely many elements of F is non-empty and hence the collection of those subsets of I which contain such an intersection forms a filter.) Let D be any ultrafilter containing F. Now we can define the morphism from G to $\prod_i G_i/D$. So, given $g \in G$ let $\bar{g}$ be the element of $\prod_i G_i$ which has i-th coordinate equal to g if $g \in G_i$ and equal to the identity element of G otherwise. Map g to $\bar{g}/D$. Our choice of D (to contain each set $[g]$) ensures that this map is an injective homomorphism, as required.

The next result is an algebraic criterion for elementary equivalence. The result after that often lends itself to algebraic applications.

THEOREM 3.29. *Two L-structures are elementarily equivalent iff they have isomorphic ultrapowers.*

THEOREM 3.30. *A class of L-structures is elementary iff it is closed under ultraproducts and elementary substructures.*

3.11. *Structure of definable sets and quantifier elimination*

Suppose that M is an L-structure. A **definable** subset of M is one of the form $\phi(M) = \{a \in M\colon M \models \phi(a)\}$ for some formula $\phi = \phi(x) \in L$. More generally if $A \subseteq M$ then an A-**definable** subset of M is one which is definable by a formula $\phi = \phi(x, \bar{a})$ with parameters from A. Yet more generally one may consider subsets of powers, M^n, of M definable by formulas with more than one free variable. The logical operations on formulas correspond to set-theoretic operations on these sets: for instance, conjunction, negation and existential quantification correspond, respectively, to intersection, complement and projection.

For many questions it is important to understand something of the structure of these sets and the relations between them. Of particular importance are quantifier elimination results. A theory T has (**complete**) **elimination of quantifiers** if every formula is equivalent modulo T to the conjunction of a sentence and a quantifier-free formula (so if T is also a complete theory then every formula will be equivalent to one without quantifiers). In order to prove quantifier elimination for a complete theory T it is enough to show that any formula of the form $\exists y\phi(\bar{x}, y)$ with ϕ quantifier-free is itself equivalent to one which is quantifier-free. In other words, it is sufficient to show that any projection of any set which can be defined without quantifiers should itself be definable without quantifiers.

EXAMPLES 3.31. The geometric content of elimination of quantifiers is illustrated by the case of the theory of algebraically closed fields. This theory does have elimination of quantifiers, a result due to Tarski and, in its geometric form (the image of a constructible set under a morphism is constructible), to Chevalley. The elimination comes down to showing that if X is a quantifier-free definable subset of some power K^n, where K is an algebraically closed field, then the projection along, say, the last coordinate is also quantifier-free definable (of course it is definable using an existential quantifier).

Tarski also proved the considerably more difficult result that the real field has elimination of quantifiers (and hence so do all real-closed fields) in the language of ordered rings. The geometric form of this statement is fundamental in the study of real algebraic geometry. See, for example, [14].

Partial elimination of quantifiers may be useful. If every formula is equivalent, modulo a complete theory T, to an existential (equivalently, every formula is equivalent to a universal) formula then T is said to be **model-complete**. This is equivalent to the condition that every embedding between models be an elementary embedding. See, for example, [31] for more on this.

For another example, the theory of modules over any ring has a partial elimination of quantifiers: every formula is equivalent to the conjunction of a sentence and a Boolean combination of "positive primitive" (certain positive existential) formulas and numerous consequences of this can be seen in [47].

A proof of model-completeness can be a stepping stone to a proof of full quantifier-elimination and has, in itself, geometric content (see, e.g., [19,52]).

3.12. *Many-sorted structures*

A single-sorted structure is one in which all elements belong to the same set (or sort). Most model theory textbooks concern themselves with these. Yet many-sorted structures are very important within model theory and its applications. Fortunately, there is almost no difference between the model theory of single- and many-sorted structures.

Some structures are naturally many-sorted. For example, in the model-theoretic study of valued fields it is natural to have one sort for the (elements of the) field and another sort for the (elements of the) value group. One would also have a function symbol, representing the valuation, taking arguments in the field sort and values in the group sort.

Other structures can be usefully enriched to many-sorted structures. In fact, it is common now in model theory to work in the context of the many-sorted enrichment described in the following subsection.

3.13. *Imaginaries and elimination of imaginaries*

All the ideas that we have discussed up to now are quite "classical". What we describe next is more recent but now pervades work in pure model theory and in many areas of application. A precursor was the practice of treating n-tuples of elements from a structure M as "generalised elements" of the structure. Shelah went much further.

Let M be any L-structure, let $n \geqslant 1$ be an integer and let E be a (ϕ-, that is, without extra parameters) definable equivalence relation on M^n. By that we mean that there is a formula, $\psi(\bar{x}, \bar{y}) \in L$, with $l(\bar{x}) = l(\bar{y}) = n$, such that for all n-tuples, $\bar{a}, \bar{b}$ of elements of M we have $M \models \psi(\bar{a}, \bar{b})$ iff $E(\bar{a}, \bar{b})$ holds. For example, the relation of conjugacy of elements (i.e. 1-tuples) in a group is definable by the formula $\exists z(y = z^{-1}xz)$. The E-equivalence classes are regarded as generalised or **imaginary** elements of M.

Formally, one extends L to a multi-sorted language, denoted L^{eq}. This means that for each **sort** (set of the form M^n/E) we have a stock of variables and quantifiers which range just over the elements of that sort. One also adds to the language certain (definable) functions between sorts, such as the canonical projection from M^n to M^n/E for each n, E. The structure M, together with all its associated imaginary sorts M^n/E and morphisms between them, is an L^{eq}-structure, denoted M^{eq}. There is a natural equivalence between the category of models of $\mathrm{Th}(M)$ and the category of models of $\mathrm{Th}(M^{\mathrm{eq}})$ (we mean the categories where the morphisms are the elementary embeddings) and most model-theoretic properties are unchanged by moving to the much richer structure M^{eq}, a notable exception being not having elimination of imaginaries. It has proved to be enormously useful in model theory to treat these imaginary elements just as one would treat ordinary elements of a structure.

3.14. *Interpretation*

This is a long-standing theme in model theory which seems to have ever-growing uses and significance. The idea is that, "within" a structure, one may find, or "interpret", other structures (of the same kind or of quite different kinds). Then, for example, if the first structure has some good properties (finiteness conditions, dimensions, ranks...), these transfer to the interpreted structure and, conversely, if the interpreted structure has "bad" properties then this has consequences for the initial structure. Let us be somewhat more precise, using the notion of imaginary sorts that we introduced above.

Suppose that M is an L-structure and that M^n/E is some sort of M^{eq}. The structure on M induces structure on M^n/E (via reference to inverse images, in M^n, of elements in M^n/E). The set M^n/E equipped with some chosen part of all this induced structure is an L'-structure for some other language L' and is said to be **interpreted** in M. In fact it is convenient here to extend the structure M^{eq} to include, as additional sorts, all definable subsets of structures M^n/E. See [24] for more on interpretation.

EXAMPLE 3.32. If K is a field, p an irreducible polynomial in $K[X]$ and L the corresponding finite extension field then L can be interpreted as K^n, where n is the degree of p, equipped with the obvious addition and with multiplication defined according to the polynomial p.

EXAMPLE 3.33. The simplest example using quotient sorts is the interpretation of the set of un-ordered pairs of elements of a structure M. This is M^2/E where E is the equivalence relation on M^2 defined by $E((x, y), (x', y'))$ iff $M \models (x = x' \wedge y = y') \vee (x = y' \wedge y = x')$.

For examples in the context of groups and fields see, for example, Chapter 3 of [6]. For very general results on finding interpretable groups and fields (the "group configuration") see [7,40,44].

3.15. *Stability: ranks and notions of independence*

There are various ranks which may be assigned to the definable subsets of a structure. These ranks give some measure of the complexity of the structure and are technically very useful since they allow one to have some measure of the extent to which one set depends on another. They also allow one to give meaning to the statement that an element a is no more dependent on a set B than on a subset $A \subseteq B$. The notion of independence that is referred to here, and which generalises linear independence in vector spaces and algebraic independence in algebraically closed fields, exists for all so-called stable theories (and beyond, see [29]) and is defined even when there is no global assignment of ranks to definable sets. To give an idea of one of these ranks we define Morley rank.

Let M be any structure. By a definable set we will mean one which is definable by a formula perhaps using as parameters some elements of M. A definable set has Morley rank 0 exactly if it is finite. Having defined what it means to have Morley rank $\geqslant n$ we say that a set X has Morley rank $\geqslant n+1$ if there is an infinite set X_i $(i \in I)$ of definable sets, each of which is a subset of X of rank $\geqslant n$ and with the X_i pairwise disjoint and of Morley rank $\geqslant n$. The definition can be continued for arbitrary ordinals and can also be extended to types (of course a definable set or type may have Morley rank undefined). Thus Morley rank is a measure of the extent to which a definable set may be chopped up into smaller definable sets. An L-structure M is said to be **totally transcendental** or, if L is countable, ω-**stable**, if every definable subset of M has Morley rank. See [6] for the rich theory of groups with finite Morley rank.

The origin of this notion of rank, and hence of model-theoretic stability theory, was Morley's Theorem 3.12, from the proof of which it follows that an $\aleph_1$-categorical structure for a countable language must be ω-stable.

EXAMPLE 3.34. Any algebraically closed field is ω-stable since, as we have seen (2.14), any such field is $\aleph_1$-categorical. It follows that any structure interpretable in an algebraically closed field must be ω-stable: in particular this applies to affine algebraic groups. Cherlin conjectured (and Zilber has a similar conjecture) that any simple ω-stable group is an algebraic group over an algebraically closed field. For more on this influential conjecture, see [6].

References

[1] J. Adámek and J. Rosický, *Locally Presentable and Accessible Categories*, London Math. Soc. Lecture Note Ser., Vol. 189, Cambridge University Press (1994).

[2] J. Ax, *The elementary theory of finite fields*, Ann. Math. **88** (1968), 239–271.

[3] J. Ax, *A metamathematical approach to some problems in number theory*, Number Theory Institute, Proc. Symp. in Pure Math., Vol. XX, 1969, J. Lewis, ed., American Mathematical Society, Providence, RI (1971), 161–190.

[4] J. Barwise, *An introduction to first-order logic*, in: [5], pp. 5–46.

[5] J. Barwise (ed.), *Handbook of Mathematical Logic*, North-Holland, Amsterdam (1977).

[6] A. Borovik and A. Nesin, *Groups of Finite Morley Rank*, Oxford Logic Guides, Vol. 26, Oxford University Press (1994).

[7] E. Bouscaren, *The group configuration – after E. Hrushovski*, in: [38], pp. 199–209.

[8] E. Bouscaren (ed.), *Model Theory and Algebraic Geometry*, Lecture Notes in Math., Vol. 1696, Springer-Verlag, Berlin (1998).

[9] E. Bouscaren, *Introduction to model theory*, in: [8], pp. 1–18.

[10] C.C. Chang and H.J. Keisler, *Model Theory*, North-Holland, Amsterdam (1973), 3rd edn (1990) (references are to the original edition).

[11] Z. Chatzidakis, L. van den Dries and A. Macintyre, *Definable sets over finite fields*, J. Reine Angew. Math. **427** (1992), 107–135.
[12] G. Cherlin, *Model-Theoretic Algebra, Selected Topics*, Lecture Notes in Math., Vol. 521, Springer-Verlag, Berlin (1976).
[13] P.M. Cohn, *Universal Algebra*, Reidel, Dordrecht (1981).
[14] L. van den Dries, *Alfred Tarski's elimination theory for real closed fields*, J. Symbolic Logic **53** (1988), 7–19.
[15] L. van den Dries and K. Schmidt, *Bounds in the theory of polynomial rings over fields*, Invent. Math. **76** (1984), 77–91.
[16] P.C. Eklof, *Ultraproducts for algebraists*, in: [5], pp. 105–137.
[17] P.C. Eklof, *Applications to algebra*, Model Theoretic Logics, J. Barwise and S. Feferman, eds, Springer-Verlag, New York (1985), 423–441.
[18] H.B. Enderton, *A Mathematical Introduction to Logic*, Academic Press, New York (1972).
[19] T.S. Gardener and H. Schoutens, *Flattening and subanalytic sets in rigid analytic geometry*, Proc. London Math. Soc.
[20] R. Göbel (ed.), *Advances in Algebra and Model Theory*, Gordon and Breach, London (1997).
[21] B. Hart, A.H. Lachlan and M.A. Valeriote (eds), *Algebraic Model Theory*, Proc. NATO Advanced Study Inst., Toronto, 1996, Riedel, Dordrecht, London (1997).
[22] W. Hodges, *Building Models by Games*, London Math. Soc. Student Texts, Vol. 2, Cambridge University Press, Cambridge (1985).
[23] W.A. Hodges, *What is a structure theory?*, Bull. London Math. Soc. **19** (1987), 209–237.
[24] W. Hodges, *Model Theory*, Cambridge University Press, Cambridge (1993).
[25] W. Hodges, *A Shorter Model Theory*, Cambridge University Press, Cambridge (1997) (shorter than [24], that is).
[26] C.U. Jensen and H. Lenzing, *Model Theoretic Algebra*, Gordon and Breach, London (1989).
[27] R. Kaye and D. Macpherson, *Models and groups*, Automorphisms of First-Order Structures, R. Kaye and D. Macpherson, eds, Oxford University Press (1994), 3–31.
[28] H.J. Keisler, *Fundamentals of model theory*, in: [5], pp. 47–103.
[29] B. Kim and A. Pillay, *From stability to simplicity*, Bull. Symbolic Logic **4** (1998), 17–36.
[30] A. Macintyre, *On definable subsets of p-adic fields*, J. Symbolic Logic **41** (1976), 605–610.
[31] A. Macintyre, *Model completeness*, in: [5], pp. 139–180.
[32] A. Macintyre, *Twenty years of p-adic model theory*, Logic Colloquium '84, North-Holland, Amsterdam (1986), 121–153.
[33] A. Macintyre and S. Shelah, *Uncountable universal locally finite groups*, J. Algebra **43** (1976), 168–175.
[34] A.I. Mal'cev, *On faithful representations of infinite groups of matrices*, Mat. Sb. **8** (1940), 405–422; transl.: Amer. Math. Soc. Transl. (2) **2** (1956), 1–21.
[35] A. Marcja and C. Toffalori, Introduzione alla Teoria dei Modelli, Pitagora, Bologna (1998); English transl. in preparation.
[36] D. Marker, *Introduction to the model theory of fields*, in: [37], pp. 1–37.
[37] D. Marker, M. Messmer and A. Pillay, *Model Theory of Fields*, Lecture Notes in Logic, Vol. 5, Springer-Verlag, Berlin (1996).
[38] A. Nesin and A. Pillay (eds), *The Model Theory of Groups*, Notre Dame Math. Lectures, Vol. 11, Notre Dame Press, Paris (1989).
[39] A. Pillay, *Model theory, differential algebra and number theory*, Proc. ICM '94, Birkhäuser (1995).
[40] A. Pillay, *Geometric Stability Theory*, Oxford Logic Guides, Vol. 32, Oxford University Press (1996).
[41] A. Pillay, *Model theory and Diophantine geometry*, Bull. Amer. Math. Soc. **34** (1997), 405–422.
[42] A. Pillay, *Model theory of algebraically closed fields*, in: [8], pp. 61–84.
[43] B. Poizat, *Cours de Théorie des Modèles*, Nur al-Mantiq wal-Ma'rifah, Villeurbanne (1985); English transl.: *A Course in Model Theory*, Springer-Verlag, Berlin (2000).
[44] B. Poizat, *Groupes Stables*, Nur al-Mantiq wal-Ma'rifah, Villeurbanne (1987).
[45] M. Prest, *Model Theory and Modules*, London Math. Soc. Lecture Notes Ser., Vol. 130, Cambridge University Press, Cambridge (1988).
[46] M. Prest, *Interpreting modules in modules*, Ann. Pure Appl. Logic **88** (1997), 193–215.

[47] M. Prest, *Model theory and modules*, Handbook of Algebra, Vol. 3, M. Hazewinkel, ed., Elsevier, Amsterdam (2003), 227–253.

[48] Ph. Rothmaler, *Introduction to Model Theory*, Gordon and Breach, London (2000).

[49] S. Shelah, *Classification Theory and the Number of Non-Isomorphic Models*, North-Holland, Amsterdam (1978), 2nd edn (1990) (with new chapters).

[50] S. Shelah, *Classification of first order theories which have a structure theorem*, Bull. Amer. Math. Soc. **12** (1985), 227–232.

[51] F. Wagner, *Stable Groups*, London Math. Soc. Lecture Note Ser., Vol. 240, Cambridge University Press (1997).

[52] A.J. Wilkie, *Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function*, J. Amer. Math. Soc. **9** (1996), 1051–1094.

[53] A.J. Wilkie, *Model theory of analytic and smooth functions*, Models and Computability, London Math. Soc. Lecture Note Ser., Vol. 259, Cambridge University Press (1999), 407–419.

Model Theory and Modules

Mike Prest

Department of Mathematics, University of Manchester, Manchester M13 9PL, UK
E-mail: mprest@maths.man.ac.uk

Contents

Abstract

The model-theoretic investigation of modules has led to ideas, techniques and results which are of algebraic interest, irrespective of their model-theoretic significance. It is these aspects that I will discuss in this chapter, although I will make some comments on the model theory of modules per se.

Our default is that the term "module" will mean (unital) right module over a ring (associative with 1) R. The category of these is denoted Mod-R, the full subcategory of finitely presented modules will be denoted mod-R, the notation R-Mod denotes the category of left R-modules. By **Ab** we mean the category of Abelian groups.

In Part 1 we introduce the general concepts and in Part 2 we discuss these in more specific contexts.

References within the text, as well as those in the bibliography, are not comprehensive but are intended to lead the reader to a variety of sources.

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1.

Purity. Purity (pure embeddings, pure-injective modules) undoubtedly plays the central role so we will start with that. The notion of a pure embedding between modules was introduced by Cohn [28]. We say that the module A is a **pure submodule** of the module B if every finite system $\sum_i x_i r_{ij} = a_j$ $(j = 1, \dots, m)$ of R-linear equations with coefficients in A (so $r_{ij} \in R$ and $a_j \in A$) and with a solution in B has a solution in A (a solution being elements $b_1, \dots, b_n$ such that $\sum_i b_i r_{ij} = a_j$ for all j). We extend this in the obvious way to define the notion of a **pure embedding** between modules and we also say that an exact sequence $0 \to A \xrightarrow{f} B \to C \to 0$ is **pure-exact** if f is a pure embedding.

Functor categories. Let $D(R) = (R\text{-mod}, \mathbf{Ab})$ denote the category of additive functors (from now on we use the term "functor" to mean additive functor) from the category of finitely presented left modules to the category of Abelian groups. This is a Grothendieck Abelian category. It has a generating set of finitely presented objects: indeed, the representable functors $(L, -)$ for $L \in R$-mod are the finitely generated projective objects and, together, are generating. This category is **locally coherent** – any finitely generated subfunctor of a finitely presented functor is itself finitely presented – and of global dimension less than or equal to 2. A functor is finitely presented iff it is the cokernel of a map between representable functors. The full subcategory $C(R) = (R\text{-mod}, \mathbf{Ab})^{\mathrm{fp}}$ of finitely presented functors is an Abelian category and the inclusion of $(R\text{-mod}, \mathbf{Ab})^{\mathrm{fp}}$ into $(R\text{-mod}, \mathbf{Ab})$ preserves exact sequences.

Notice that the category $(R\text{-mod}, \mathbf{Ab})$ is just the category of "modules" over the "ring with many objects" R-mod (better, over a small version of this). Concepts for modules over a ring generally make good sense in this context and largely can be understood in this way (that is, as having the same content that they have for modules).

There is a full embedding of Mod-R into $(R\text{-mod}, \mathbf{Ab})$ which is given on objects by sending $M \in \text{Mod-}R$ to the functor $M \otimes_R - : R\text{-mod} \to \mathbf{Ab}$ and which is given in the natural way on morphisms. The image of this embedding consists (up to isomorphism) of the absolutely pure objects of $(R\text{-mod}, \mathbf{Ab})$. A module M is said to be **absolutely pure** (= **fp-injective**) if every embedding $M \to M'$ in Mod-R is pure (equivalently if whenever $A \xrightarrow{f} B$ is an embedding of modules such that B/fA is finitely presented then each morphism $A \xrightarrow{g} M$ can be factored as $g = hf$ for some $B \xrightarrow{h} M$) and the same definition may be made for functors. (Indeed, everything that we do here for modules can be done for the objects of a locally finitely presented Grothendieck category.)

An object E of an Abelian category is **injective** if every inclusion of the form $E \to F$ in the category is split. Every object of a Grothendieck category has an injective hull (a "smallest" injective object containing it).

THEOREM 1.1. *The following are equivalent for the exact sequence* $0 \to A \xrightarrow{f} B \to C \to 0$ *of right R-modules*:
(i) *the sequence is pure-exact*;
(ii) *for every (finitely presented) left module L the sequence* $0 \to A \otimes_R L \to B \otimes_R L \to C \otimes_R L \to 0$ *of Abelian groups is exact*;

(iii) *for every (finitely presented) left module* L *the morphism* $A \otimes_R L \xrightarrow{f \otimes_R 1_L} B \otimes_R L$ *of Abelian groups is monic*;
(iv) *for every finitely presented module* M *the sequence* $0 \to (M, A) \to (M, B) \to (M, C) \to 0$ *is exact*;
(v) *the natural transformation* $f \otimes_R - : A \otimes_R - \to B \otimes_R -$ *is a monomorphism in* $(R\text{-mod}, \mathbf{Ab})$.

PROPOSITION 1.2. *Every split embedding is pure. The composition of two pure embeddings is pure. If* $A \xrightarrow{f} B$ *is a pure embedding and if the cokernel* B/A *is finitely presented then* f *is a split embedding.*

So purity is significant in the presence of "large" (non-finitely presented) modules. Here is another indication of this.

PROPOSITION 1.3. *Any direct limit of pure embeddings is pure.*

In particular, any direct limit of split embeddings, though not necessarily split, will be pure.

Pure-injectives. A module N is **pure-injective** (also called **algebraically compact**) if whenever $N \xrightarrow{f} M$ is a pure embedding then f is split. If we think of pure embeddings as embeddings which would split if the smaller module had "enough" elements then we can see pure-injectivity as a kind of completeness property.

pp-definable subgroups. Let $G = (r_{ij})_{ij}$ be a matrix with entries in R. Let M be any module. Then $\mathrm{ann}_{M^n} G = \{\bar{c} \in M^n \colon \sum c_i r_{ij} = 0 \text{ for each } j\}$ is an $\mathrm{End}(M)$-submodule of M^n (where $\mathrm{End}(M)$ acts diagonally on M^n) and the projection of $\mathrm{ann}_{M^n} G$ to the first, say, k coordinates is a subgroup of, even an $\mathrm{End}(M)$-submodule of, M^k. Such subgroups are termed variously **finitely matrisable subgroups** [185], **subgroups of finite definition** [52] or, as we shall say, **pp-definable subgroups**.

Notice that the subgroup of M^k that we just defined consists of all k-tuples $\bar{a} = (a_1, \ldots, a_k)$ from M such that there exists a tuple, $\bar{b} = (b_1, \ldots, b_{n-k})$, of elements from M such that $(\bar{a}\bar{b})G = 0$ (where $(\bar{a}\bar{b})$ denotes the row vector $(a_1, \ldots, a_k, b_1, \ldots, b_{n-k})$). That is, it is the subgroup $\{\bar{a} \in M^k \colon M \models \exists y_1, \ldots, y_{n-k} \bigwedge_j \sum_{i=1}^k a_i r_{ij} + \sum_{l=1}^{n-k} y_l r_{k+l,j} = 0\}$ of M^k. Here one should read the symbol $\models$ as "satisfies the condition that" and $\bigwedge$ denotes repeated conjunction ("and"). If we denote by $\phi(x_1, \ldots, x_k)$ the condition (on $x_1, \ldots, x_n$) $\exists y_1, \ldots, y_{n-k} \bigwedge_j \sum_{i=1}^k x_i r_{ij} + \sum_{l=1}^{n-k} y_l r_{k+l,j} = 0$ then we may regard this as a formula of the formal language which is used for the model theory of modules. A formula of this particular shape is referred to as a **positive primitive** (or **pp**) formula. More loosely, any formula which is equivalent in all R-modules to one of this form is said to be a pp formula.

The above subgroup of M^k consists of all $\bar{a}$ which satisfy the formula ϕ (we write $M \models \phi(\bar{a})$ for that) and we denote this "solution set" of ϕ in M by $\phi(M)$. That explains the terminology "pp-definable subgroup".

Observe that, having specified the matrix G and the integer k (and hence the formula ϕ above), we obtain, by the above construction, a functor, F_ϕ, from Mod-R to **Ab** (indeed, a subfunctor of the representable functor $(R,-)^k$): namely that which takes a module M to the group $\phi(M)$ (the action on morphisms is restriction/corestriction since if $A \xrightarrow{f} B$ is any morphism of modules then $f\phi(A) \subseteq \phi(B)$). Using the fact that such a functor commutes with direct limits, together with the fact that every module is a direct limit of finitely presented modules, one sees that this functor is determined by its restriction, which we also denote F_ϕ, to mod-R.

If ϕ, ψ are pp formulas we write $\psi \leqslant \phi$ if ψ implies ϕ, that is, if $\psi(M) \leqslant \phi(M)$ for every (finitely presented) module M, that is, if F_ψ is a subfunctor, $F_\psi \leqslant F_\phi$, of F_ϕ.

THEOREM 1.4 (e.g., [103, Chapter 12]). *Every functor of the form F_ϕ is finitely presented in* (mod-R, **Ab**). *Every finitely presented functor in* (mod-R, **Ab**) *is isomorphic to one the form F_ϕ/F_ψ for some pp formulas with $\psi \leqslant \phi$.*

THEOREM 1.5 (e.g., [68, 7.1]). *For a module N the following are equivalent*:

(i) *N is pure-injective*;

(ii) *any system $\sum_{i\in I} x_i r_{ij} = a_j$ $(j \in J)$ of R-linear equations in possibly infinitely many variables x_i (but with, for each j, almost all r_{ij} zero) which is finitely solvable in N (i.e. for every finite $J' \subseteq J$ the system $\sum_{i\in I} x_i r_{ij} = a_j$ $(j \in J')$ has a solution in N) has a solution in N*;

(iii) *if C_k $(k \in K)$ is any set of cosets of pp-definable subgroups of N with the **finite intersection property** (i.e. for every finite subset $K' \subseteq K$ the intersection $\bigcap_{k\in K'} C_k$ is non-empty) then it has non-empty intersection*: $\bigcap_{k\in K} C_k \neq \emptyset$;

(iv) *N is injective over pure embeddings (if $A \xrightarrow{f} B$ is a pure embedding and if $A \xrightarrow{g} N$ is any morphism then there is a factorisation of g through f)*;

(v) *the functor $N \otimes_R -$ is an injective object of the functor category* (R-mod, **Ab**);

(vi) *For every index set I the summation map $N^{(I)} \to N$, given by $(a_i)_{i\in I} \mapsto \sum_i a_i$, factors through the canonical embedding of the direct sum $N^{(I)}$ into the direct product N^I.*

EXAMPLES 1.6. Any injective module is pure-injective. Any module which is Artinian over its endomorphism ring is pure-injective (this includes finite modules and modules which are finite-dimensional over a field contained in the centre of R). If M_R is a module then the left R-module $\mathrm{Hom}_{\mathbb{Z}}(M, \mathbb{Q}/\mathbb{Z})$ is pure-injective.

Linear and algebraic compactness. The module M is said to be **linearly compact** if every set of cosets of submodules of M which has the finite intersection property has non-empty intersection. So if M is an R-module then M is a pure-injective R-module if it is linearly compact as a module over its endomorphism ring (or any subring thereof).

In general pure-injective modules are not closed under extensions (e.g., [149, p. 436]) but one has the useful result [183]: if $0 \to A \to B \to C \to 0$ is an exact sequence with A and C linearly compact then B is linearly compact.

Pure-injective hulls. Every module M has a **pure-injective hull**, variously denoted $\overline{M}$ or $PI(M)$: M purely embeds, $M \xrightarrow{j} \overline{M}$, in $\overline{M}$; $\overline{M}$ is a pure-injective module; if $M \xrightarrow{f} N$ is a pure embedding of M into a pure-injective module N then there is a morphism (necessarily a pure embedding) $\overline{M} \xrightarrow{g} N$ with $gj = f$. This module, $\overline{M}$, is unique up to isomorphism over M. It is most efficiently produced (at least, given the corresponding theorem for injective objects of Grothendieck categories) using the following result (take the injective hull of the functor $M \otimes_R -$), as is the structure theorem that follows.

THEOREM 1.7 [52]. *Every injective object of the functor category* (R-mod, **Ab**) *has the form $N \otimes_R -$ for some pure-injective module N.*

THEOREM 1.8. *Let N be a pure-injective module. Then $N \simeq PI(\bigoplus_\lambda N_\lambda) \oplus N_c$ where each N_λ is an indecomposable pure-injective and where N_c is a* ***continuous*** (= ***superdecomposable***) *pure-injective* (*that is, one without any non-zero indecomposable summands*). *The modules N_λ, together with their multiplicities, as well as the module N_c, are determined up to isomorphism by N.*

The next result has been extensively used in the model theory of modules.

THEOREM 1.9 ([42], also see [45]). *Let N be a pure-injective module and let A be a submodule of N. Then there is a direct summand of N, denoted $H(A)$ which is determined up to isomorphism over A and which is minimal in the sense that if $A \leqslant N'$ and N' is a direct summand of N then there is an A-isomorphism of $H(A)$ with a direct summand of N'.*

In fact, $H(A)$ may be identified as that module such that $H(A) \otimes_R -$ is the injective hull of the image of $A \otimes_R -$ in $N \otimes_R -$.

Σ-pure-injectives and modules of finite endolength. Any direct product of pure-injective modules is pure-injective. A module M is said to be **Σ-pure-injective** if $M^{(\aleph_0)}$ (and hence every direct sum of copies of M) is pure-injective.

THEOREM 1.10. *A module M is Σ-pure-injective iff M has the descending chain condition on pp-definable subgroups. In particular, any module which is Artinian over its endomorphism ring is Σ-pure-injective.*

THEOREM 1.11. *If M is Σ-pure-injective then so is every pure submodule of M* (*hence such a module is a direct summand*) *as is every module in the closure of $\{M\}$ under arbitrary direct summands, direct sums, direct products and direct limits.*

Every Σ-pure-injective module is a direct sum of indecomposable pure-injective modules.

A module M is **of finite endolength** if it is of finite length when considered as a module over its endomorphism ring. Such a module must be Σ-pure-injective.

THEOREM 1.12 [45]. *The module M is of finite endolength iff for any index set I, the power M^I is a direct sum of copies of direct summands of M.*

A module which is of finite endolength and is not finitely presented is said to be **generic**. At least in the context of algebras over algebraically closed fields, such modules correspond to one-parameter families of finite-dimensional modules, see [31] and, for more general contexts, [32].

The following result indicates the model-theoretic relevance of pure-injectivity and its second part points to the special role played by the indecomposable pure-injectives.

THEOREM 1.13. [41,157] *Every module is elementarily equivalent to a pure-injective module* (*in fact, is an elementary substructure of its pure-injective hull*).

[184] *Every module is elementarily equivalent to a direct sum of indecomposable pure-injectives* (*for elementary equivalence see below*).

pp-types. There are various ways in which one may try to explain the key role played by pure-injective modules in the model theory of modules. One is to say that they are the "positively saturated" modules – those which realise every pp-type. Saturation is a notion, a kind of completeness property, from model theory and the proofs of many results in model theory involve moving to a suitable saturated elementary extension. The analogous procedure here is moving to the pure-injective hull (or some other pure-injective pure extension) of a module. The notion of a pp-type has played a central role in the model-theoretic approach to modules yet its algebraic translation is not a familiar algebraic object. I now say something about this.

Let $a \in M$. The **pp-type** of a in M is just the set of pp formulas satisfied by a in M: $\mathrm{pp}^M(a) = \{\phi(x)\colon\ M \models \phi(a)\}$. If we replace each pp formula by the corresponding subfunctor, F_ϕ, of the forgetful functor, $(R,-) \in (\text{mod-}R, \mathbf{Ab})^{\mathrm{fp}}$, then $\mathrm{pp}^M(a)$ becomes a filter in the lattice of finitely generated subfunctors of $(R,-)$. If M is finitely presented then this filter will be principal, that is there will be some pp formula $\phi_0 \in \mathrm{pp}^M(a)$ which implies all the rest (in terms of functors, ϕ_0 is such that the intersection of this filter is F_{ϕ_0} and F_{ϕ_0} belongs to the filter). I remark that the formula ϕ_0 can be computed explicitly from a presentation of M by generators and relations and from the representation of a as an R-linear combination of these generators.

We can make rather more algebraic sense of this by moving to the dual functor category $(R\text{-mod}, \mathbf{Ab})^{\mathrm{fp}}$ (see the subsection on duality of functors). In that context the element $a \in M$, that is the morphism $R \xrightarrow{a} M$, becomes the morphism $({}_RR,-) \simeq (R \otimes_R -) \xrightarrow{a\otimes -} (M \otimes -)$ in $(R\text{-mod}, \mathbf{Ab})$ and the "duals", dF_ϕ, in the sense discussed below of the functors F_ϕ with $\phi \in \mathrm{pp}^M(a)$ turn out to be exactly the finitely generated subfunctors of the kernel of $a \otimes -$. That is, the pp-type of a in M is "really" the kernel of the morphism $a \otimes - : (R \otimes -) \to (M \otimes -)$. In [103] the pp-type of an element was treated heuristically as a generalised annihilator of the element. The above change of perspective makes this precise.

Ziegler spectrum. The (right) **Ziegler spectrum**, Zg_R, of R is the topological space which has, for its points, the (isomorphism types of) indecomposable pure-injectives and which has, for a basis of open sets, those given by:

- pairs of pp-formulas – let $\phi \geqslant \psi$ be pp formulas (in any number of free variables) and set $(\phi/\psi) = \{N \in \mathrm{Zg}_R\colon\ \phi(N)/\psi(N) \neq 0\}$;
- alternatively, by finitely presented functors in $(R\text{-mod}, \mathbf{Ab})$ – let $F \in (R\text{-mod}, \mathbf{Ab})^{\mathrm{fp}}$ and set $(F) = \{N \in \mathrm{Zg}_R\colon\ (F, N \otimes -) \neq 0\}$;
- alternatively by morphisms in mod-R – let $A \xrightarrow{f} B$ be a morphism in mod-R and set $(f) = \{N \in \mathrm{Zg}_R\colon\ (f, N) : (B, N) \to (A, N)$ is not onto$\}$.

These are, indeed, equivalent: any set of one of the above forms is of the other forms (and the transformation from one to the other is quite explicit and can be found variously in [34,75,117]). Since every functor in $(\text{mod-}R, \mathbf{Ab})^{\mathrm{fp}}$ is isomorphic to F_ϕ/F_ψ for some pp formulas $\phi \geqslant \psi$ a fourth way of giving the open sets is as $\{N \in \mathrm{Zg}_R\colon\ \overrightarrow{G}N \neq 0\}$ as G ranges over $(\text{mod-}R, \mathbf{Ab})^{\mathrm{fp}}$ and where $\overrightarrow{G}$ is the unique extension of such a functor to a functor on all of Mod-R which commutes with direct limits.

THEOREM 1.14 [184, 4.9]. *The space* Zg_R *is compact, that is, every open cover of* Zg_R *has a finite subcover. The compact open sets are precisely the basic open sets* (*variously*) *described above.*

Ziegler showed that the closed subsets of this space correspond exactly to those elementary classes of modules which are closed under direct sum and direct summand (Crawley-Boevey [34] has termed such classes **definable**, which is a useful term although it has potential to be confused with the more general "axiomatisable").

Support of a module. The **support** of a module M is $\mathrm{supp}(M) = \{N \in \mathrm{Zg}_R\colon\ N$ is a direct summand of M' for some $M' \equiv M\}$. (Modules M, M' are elementarily equivalent [118], $M \equiv M'$, if they have exactly the same first-order properties – a first-order property being one which can be expressed by a sentence, or set of sentences, of the formal language for R-modules. It is equivalent to require that they have isomorphic ultrapowers.) This is a closed subset of Zg_R and every closed subset is the support of some module. One has $\mathrm{supp}(M) = \mathrm{supp}(M_1)$ iff $M^{(\aleph_0)} \equiv M_1^{(\aleph_0)}$ (if, for instance, R is an algebra over an infinite field then one has $\mathrm{supp}(M) = \mathrm{supp}(M_1)$ iff $M \equiv M_1$). One can avoid the use of elementary equivalence in this definition by working with torsion theories in the functor category, as follows.

To each module M is associated a hereditary torsion theory (for torsion theory see [168]) on the functor category $(R\text{-mod}, \mathbf{Ab})$: namely that cogenerated by $E(M \otimes -) = \overline{M} \otimes -$. So the corresponding torsion class is $\mathcal{T} = \{F \in (R\text{-mod}, \mathbf{Ab})\colon\ (F, M \otimes -) = 0\}$. Denote the associated torsion theory by τ_M. This is a torsion theory of finite type since the torsion class is determined by (is the smallest hereditary torsion class containing) the finitely presented torsion functors. It is the case that τ_M is cogenerated by a set of indecomposable injective objects. The set of points $N \in \mathrm{Zg}_R$ such that $N \otimes -$ is τ_M-torsionfree is the **support** of M as defined above.

If X is any closed subset of Zg_R and if $\tau = \tau_X$ is the corresponding torsion theory of finite type on $(R\text{-mod}, \mathbf{Ab})$ (that is, the torsion theory cogenerated by $\{N \otimes -\colon\ N \in X\}$) then the modules M with $\mathrm{supp}(M) \subseteq X$ are exactly those such that $M \otimes -$ is τ-torsionfree. One also has, for any module M, that $\mathrm{supp}(\overline{M}/M) \subseteq \mathrm{supp}(M) = \mathrm{supp}(\overline{M})$

and so, if $\mathrm{supp}(M) \subseteq X$, then $(\overline{M}/M) \otimes -$ is τ-torsionfree and hence $M \otimes -$ is τ-closed (in the sense of torsion theory). Indeed, see [53,60,73], the functors $M \otimes -$ for M with $\mathrm{supp}(M) \subseteq X$ give exactly the absolutely pure objects of the category $(R\text{-mod}, \mathbf{Ab})_\tau$ where by $(R\text{-mod}, \mathbf{Ab})_\tau$ we mean the category which is obtained by localising $(R\text{-mod}, \mathbf{Ab})$ at τ.

Isolated points. A point, N, of Zg_R is isolated (that is $\{N\}$ is open) iff there is a finitely presented functor F from R-mod to **Ab** such that $(F, N \otimes -) \neq 0$ and such that N is the only such indecomposable pure-injective. For instance, if F is a finitely presented simple object of $(R\text{-mod}, \mathbf{Ab})$ then (F) contains just one point, namely the point $N \in \mathrm{Zg}_R$ such that $N \otimes -$ is the injective hull of F. We say that $N \in \mathrm{Zg}_R$ is **isolated by a minimal pair** if $\{N\} = (\phi/\psi)$ for some pair, $\phi > \psi$ of pp formulas such that no pp formula lies strictly between ϕ and ψ. A morphism $f : A \to B$ is said to be **minimal left almost split** if it is not a split monomorphism but every morphism $g : A \to C$ which is not a split monomorphism factors through f and if every endomorphism h of B with $hf = f$ is an automorphism of B (see [5]).

THEOREM 1.15.

(a) *Suppose that $N \in \mathrm{Zg}_R$. Then N is isolated by a minimal pair iff $N \otimes -$ is the injective hull of a finitely presented simple functor in* $(R\text{-mod}, \mathbf{Ab})$.

(b) *If $N \in \mathrm{Zg}_R$ is the pure-injective hull of a finitely presented module M then N is isolated by a minimal pair iff there is a minimal left almost split map in* mod-R *with source M* (*see, e.g.,* [60, 7.7]).

(c) *If R is countable and $N \in \mathrm{Zg}_R$ is isolated then N is isolated by a minimal pair* [184].

It is not known whether or not the equivalence in part (c) holds for general rings. It does hold whenever the functor category $(R\text{-mod}, \mathbf{Ab})$ has Krull–Gabriel dimension (see below).

Related to the above one has the following.

THEOREM 1.16 ([32, 2.3], also see [186]). *An indecomposable pure-injective module is the injective hull of a simple functor iff it is the source of a left almost split map in* Mod-R.

THEOREM 1.17 [103,56]. *Let $M \in$* mod-R. *Then the pure-injective hull, $\overline{M}$, of M is indecomposable iff* $\mathrm{End}(M)$ *is local.*

THEOREM 1.18 [60, Prop. 5.4]. *Let R be a ring such that every finitely presented module is a direct sum of modules with local endomorphism ring. Then the points of* Zg_R *of the form $\overline{M}$ where M is finitely presented indecomposable are dense in* Zg_R. *If* mod-R *has almost split sequences then all such points are isolated by minimal pairs* (*e.g.,* [120, Prop. 3.7]).

In particular, if R is an Artin algebra then every indecomposable finitely generated R-module is an isolated point of Zg_R and, together, these points are dense in Zg_R [103, Chapter 13].

Closed points. As for closed points, we have the following.

THEOREM 1.19. *If* $M \in \text{Mod-}R$ *is of finite endolength then* $\text{supp}(M)$ *is a finite set of closed points of* Zg_R. *In particular, if* N *is indecomposable and of finite endolength then* N *is a closed point of* Zg_R. *For countable rings* R *we have the converse*: *if* N *is a closed point of* Zg_R *then* N *is of finite endolength.*

It is not known whether or not the last part holds for arbitrary rings. It does hold if $(R\text{-mod}, \mathbf{Ab})$ has Krull–Gabriel dimension (by the analysis of [184], see, e.g., [103, Section 10.4]).

Finiteness conditions. A central aim is to understand the Ziegler spectrum: to prove general results and to obtain descriptions of Zg_R (points and topology) for particular rings R. There are "finiteness conditions" and related dimensions and ranks which aid the analysis of Zg_R. Here we discuss the Krull–Gabriel and uniserial dimensions of the functor category $(R\text{-mod}, \mathbf{Ab})$ and the Cantor–Bendixson rank of Zg_R (also see [110,117]). Both dimensions on the functor category are obtained by successive localisation.

Krull–Gabriel dimension. Let $\mathcal{C}$ be a locally coherent Grothendieck category, such as $(R\text{-mod}, \mathbf{Ab})$. If τ is a torsion theory of finite type on $\mathcal{C}$ then the quotient category $\mathcal{C}_\tau$ is again locally coherent (with finitely presented objects exactly the objects isomorphic to localisations of finitely presented objects of $\mathcal{C}$).

Denote by $\mathcal{S}_0$ the subcategory of $\mathcal{C}^{\text{fp}}$ (the full subcategory of finitely presented objects of $\mathcal{C}$), consisting of all finitely presented objects of finite length. This is a **Serre subcategory** of $\mathcal{C}^{\text{fp}}$ (if $0 \to A \to B \to C \to 0$ is exact then $B \in \mathcal{S}_0$ iff $A, C \in \mathcal{S}_0$) and hence its closure under direct limits is the torsion class for a torsion theory, $\tau = \tau(\mathcal{C})$, of finite type on $\mathcal{C}$. The corresponding quotient category $\mathcal{C}^{(1)} = \mathcal{C}_\tau$ is obtained from $\mathcal{C}$ by "making zero" all finitely presented simple objects. Since this localised category is again locally coherent Grothendieck [60,73,105] we can repeat the process.

Having defined $\mathcal{C}^{(\alpha)}$ we define $\mathcal{C}^{(\alpha+1)}$ to be $\mathcal{C}^{(\alpha)}_{\tau(\mathcal{C}^{(\alpha)})}$. The process can be continued transfinitely as follows. Let τ_α, with corresponding torsion class denoted $\mathcal{T}_\alpha$, be the torsion theory on $\mathcal{C}$ such that $\mathcal{C}^{(\alpha+1)} = \mathcal{C}_{\tau_\alpha}$. If λ is a limit ordinal define τ_λ to be the torsion theory which has torsion class $\bigcup\{\mathcal{T}_\alpha \colon \alpha < \lambda\}$ and define $\mathcal{C}^{(\lambda)}$ to be the localisation of $\mathcal{C}$ at τ_λ.

The least α such that $\mathcal{C}^{(\alpha+1)} = 0$ (that is, such that $\mathcal{T}_\alpha = \mathcal{C}$) is the **Krull–Gabriel dimension**, $\text{KGdim}(\mathcal{C})$, [46] of $\mathcal{C}$: if there is no such α (that is, if some non-zero localisation of $\mathcal{C}$ has no finitely presented simple object) then we set $\text{KGdim}(\mathcal{C}) = \infty$ (and say that the Krull–Gabriel dimension of $\mathcal{C}$ is "undefined"). The **Krull–Gabriel dimension** of an object $C \in \mathcal{C}$ is the least α such that the image of C in $\mathcal{C}^{(\alpha+1)}$ is zero (that is, $C \in \mathcal{T}_\alpha$), if this exists, and is ∞ otherwise. We set $\text{KG}(R) = \text{KGdim}(R\text{-mod}, \mathbf{Ab})$ and refer to this also as the **Krull–Gabriel dimension** of R.

THEOREM 1.20. $\text{KG}(R) < \infty$ *iff the lattice of finitely generated subfunctors of the functor* $({}_R R, -) \in (R\text{-mod}, \mathbf{Ab})$ *has Krull dimension in the sense of Gabriel–Rentschler.*

m-dimension. Recall that Krull dimension (in the sense of [143]) is defined on posets by inductively collapsing intervals which have the descending chain condition. One can define a variant of this dimension by inductively collapsing intervals of finite length. This dimension, called "**m-dimension**" in [103], does, therefore, grow more slowly than Krull dimension but the one dimension is defined iff the other is (iff the poset contains no densely ordered sub-poset). Actually, the discussion in [103] is in terms of the lattice of pp formulas but this is isomorphic to the above lattice of finitely generated functors via the map $\phi \mapsto F_\phi$. So we have $\text{m-dim}(M) = \text{KGdim}(M)$ for every module M.

Recall that, if X is a closed subset of Zg_R then the modules M with $\text{supp}(M) \subseteq X$ correspond to the absolutely pure objects of the localised category $(R\text{-mod}, \mathbf{Ab})_\tau$ where $\tau = \tau_X$ is the torsion theory of finite type on $(R\text{-mod}, \mathbf{Ab})$, which we also denote $D(R)$, corresponding to X. In particular, the points, N, of X correspond to the indecomposable injectives, $(N \otimes -)_\tau \simeq (N \otimes -)$, of $(R\text{-mod}, \mathbf{Ab})_\tau$.

THEOREM 1.21. *Suppose that X is a closed subset of Zg_R, let τ be the corresponding torsion theory on $D(R) = (R\text{-mod}, \mathbf{Ab})$ and let $(R\text{-mod}, \mathbf{Ab})_\tau$ be the localisation of $D(R)$ at this torsion theory.*

If $\text{KGdim}((R\text{-mod}, \mathbf{Ab})_\tau) = \alpha < \infty$ then every pure-injective module with support contained in X is the pure-injective hull of a direct sum of indecomposables [45]. *Furthermore, the Cantor–Bendixson rank of X equals α* [184]. *In particular, the isolated points of X are dense in X.*

In particular if $\text{KG}(R) < \infty$ then there are no continuous pure-injective R-modules.

Cantor–Bendixson rank. The Cantor–Bendixson rank of a topological space T is defined as follows. Let T' be the set of non-isolated points of T. Inductively set $T^{(0)} = T$, $T^{(\alpha+1)} = (T^{(\alpha)})'$, $T^{(\lambda)} = \bigcap\{T^{(\alpha)}\colon \alpha < \lambda\}$ for limit ordinals λ. The **Cantor–Bendixson rank**, $\text{CB}(p)$, of a point $p \in T$ is the least α such that $p \notin T^{(\alpha)}$, and is ∞ if there is no such α. The **Cantor–Bendixson rank**, $\text{CB}(T)$, of T is the least α such that $T^{(\alpha)} = \emptyset$ if such exists, and is ∞ otherwise. If T is a compact space then $\text{CB}(T) = \max\{\text{CB}(p)\colon p \in T\}$ and there are only finitely many points of maximum CB-rank if this is less than ∞.

Isolation condition. We say that R satisfies the **isolation condition** if for every closed subset X of Zg_R every point $N \in X$ which is isolated in X is isolated by an X-minimal pair, meaning that there are $\psi \leqslant \phi$ with $\{N\} = X \cap (\phi/\psi)$ and such that the localised functor $(F_\phi/F_\psi)_{\tau=\tau_X}$ is a simple object of $(R\text{-mod}, \mathbf{Ab})_\tau$ (an equivalent condition on this pair is that for any/every M with $\text{supp}(M) = X$ we have $\phi(M) > \psi(M)$ and there is no pp-definable subgroup of M strictly between $\phi(M)$ and $\psi(M)$). (In [103] this condition was given the ad hoc name "condition $(\wedge)$".) The condition is right/left symmetric (it holds for R iff it holds for R^{op}). It is not known whether or not every ring satisfies the isolation condition (in the case of von Neumann regular rings this specialises to an existing open question about such rings, see Section 6b below).

THEOREM 1.22 [184]. *Every countable ring satisfies the isolation condition. If $\text{KG}(R) < \infty$ or, more generally, if there are no continuous pure-injective R-modules then R satisfies the isolation condition.*

THEOREM 1.23. *If R satisfies the isolation condition then* $\mathrm{CB}(\mathrm{Zg}_R) = \mathrm{KG}(R)$. *In particular this is true if R is countable.*

Uniserial dimension. The other dimension that we consider is obtained by the same general process that we used to define Krull–Gabriel dimension. If $\mathcal{C}$ is a locally finitely presented Grothendieck category let $\mathcal{S}_u$ denote the Serre subcategory of $\mathcal{C}^{\mathrm{fp}}$ which is generated by the finitely presented uniserial objects (an object is **uniserial** if the lattice of its subobjects is a chain). The same process of successive localisation that we used for Krull–Gabriel dimension now yields the notion of **uniserial dimension** (∞ or "defined" – that is, an ordinal) of a locally coherent Grothendieck category. We write $\mathrm{UD}(R)$ for the uniserial dimension of the functor category $(R\text{-mod}, \mathbf{Ab})$. Clearly $\mathrm{UD}(R) \leqslant \mathrm{KGdim}(R)$.

THEOREM 1.24 [184]. *If the uniserial dimension of* $(R\text{-mod}, \mathbf{Ab})$ *is defined then there are no continuous pure-injective R-modules. If R is countable then the converse is true.*

It is not known whether or not the converse is true for all rings.

Gabriel–Zariski spectrum. Next we consider a new topology on the set of isomorphism classes of indecomposable pure-injective R-modules. Consider the collection, $\{U^c\colon U$ a compact open subset of $\mathrm{Zg}_R\}$, of complements of compact Ziegler-open sets. We use the notation $[\phi/\psi]$, $[F]$ for the complements of (ϕ/ψ), (F), respectively. This collection of sets is closed under finite intersection and so forms a basis for a new topology: the **dual-Ziegler**, or **Zariski, topology** on Zg_R. We will call the resulting topological space the **Gabriel–Zariski spectrum**, Zar_R, of R and also the **Zariski spectrum** of mod-R.

We explain the terminology. Let R be a commutative Noetherian ring. Recall that the Zariski spectrum of R is the set, $\mathrm{Spec}(R)$, of prime ideals of R equipped with the topology which has, for a basis of open sets, those of the form $\{P \in \mathrm{Spec}(R)\colon r \notin P\}$ as r varies over R. Following Gabriel and Matlis [43,91] we replace each prime P by the injective hull, $E(R/P)$, of the factor module R/P. This is an indecomposable injective R-module (denote the set of isomorphism classes of these by Inj_R) and every indecomposable injective R-module has this form. To any finitely presented module M associate $\{E \in \mathrm{Inj}_R\colon \mathrm{Hom}(M, E) = 0\}$. These sets, as M varies, form a basis of a topology on Inj_R. Indeed, the resulting space is homeomorphic to $\mathrm{Spec}(R)$ via the identification of P with $E(R/P)$. Thus $\mathrm{Spec}(R)$ may be defined purely in terms of the module category.

This definition makes sense in any locally coherent category, in particular, in the category $D(R) = (R\text{-mod}, \mathbf{Ab})$. Since the indecomposable injective objects of $D(R)$ correspond bijectively with the indecomposable pure-injective right R-modules this topology on $D(R)$ induces one on the set Zg_R. This new topology on the set Zg_R is exactly the one we defined above, the basic open sets having the form $\{N\colon (F, N \otimes -) = 0\}$ as F ranges over the subcategory $C(R) = (R\text{-mod}, \mathrm{Ab})^{\mathrm{fp}}$, hence the name.

Despite the name, however, the space Zar_R is only "algebraic–geometric" in parts. For example, it is seldom compact and it may have infinitely many clopen points. In some examples (see, e.g., [25]) it seems to be a partial amalgamation of "geometric" and "combinatorial" pieces.

Nonetheless, there is a natural sheaf of rings over it which directly generalises the structure sheaf of a commutative Noetherian ring. We need the notion of the ring of definable scalars in order to define this.

Rings of definable scalars. Let X be a closed subset of Zg_R. Let M be a module with $\mathrm{supp}(M) = X$. Then the set of pp-definable functions from M to M forms what is called the **ring of definable scalars**, R_M, of M. This ring, rather R-algebra since there is a canonical morphism $R \to R_M$, depends only on X, so we denote it also by R_X. Every module with support contained in X is naturally an R_X-module. This ring may be defined in other ways, without direct reference to model theory.

First, choose M with $\mathrm{supp}(M) = X$ to be pure-injective and also such that $M \otimes -$ is an injective cogenerator for the torsion theory of finite type, τ, on $D(R)$ which corresponds to X. Then R_X is the bi-endomorphism ring, $\mathrm{Biend}(M)$, of M^I (that is, $\mathrm{End}({}_{\mathrm{End}(M)}M^I)$) where the power I is chosen large enough so that M^I is cyclic over its endomorphism ring.

Alternatively, consider the localisation, $({}_RR, -)_\tau$, of the forgetful functor at the torsion theory $\tau = \tau_{\mathrm{supp}(M)}$. Then the endomorphism ring, in the localised category, $D(R)_\tau$, of $({}_RR, -)_\tau$ is isomorphic to R_X.

THEOREM 1.25 [24].

(a) *Let M be a Σ-pure-injective module which is finitely generated over its endomorphism ring (e.g., let M be a module of finite endolength). Then $R_M \simeq \mathrm{Biend}(M)$.*

(b) *Suppose that $R \xrightarrow{f} S$ is an epimorphism in the category of rings. Then the forgetful functor from* Mod-S *to* Mod-R *induces an embedding of* Zg_S *as a closed subset of* Zg_R. *The ring of definable scalars of this closed subset is exactly S, regarded as an R-algebra via f.*

(c) *Let E be an injective R-module which cogenerates a torsion theory, τ, of finite type on* Mod-R. *Then the ring of definable scalars of E is precisely the corresponding localisation $R \to R_\tau$.*

A finer topology. Burke [20] introduced another topology on the underlying set of Zg_R which he (re-)named, in [22], the **full support topology** (in his thesis he called it the “types-over-formulas” topology because the basic open sets are of the form (p/ψ) where p is a pp-type (a notion from model theory) and ψ a pp formula). The closed sets for this topology are exactly the sets of the form $\{N\colon\ N \otimes - \text{ is } \tau\text{-torsionfree}\}$ where now τ ranges over all hereditary torsion theories (not just those of finite type) on (R-mod, **Ab**). Associated to any closed set of this topology is the ring of type-definable scalars, which is isomorphic to the endomorphism ring of the localised forgetful functor. Corresponding to this topology one also obtains a sheaf of rings, with stalks being rings of infinitely definable scalars, analogous to that defined below.

The sheaf of locally definable scalars. For every basic closed subset X of Zg_R we have the associated R-algebra, $R \to R_X$, of definable scalars. These sets form a basis of open sets for Zar_R and so this assignment defines a presheaf on this basis of Zar_R, hence extends to a sheaf on Zar_R, called the **sheaf of locally definable scalars**, LDef_R, of R. One can check that the stalk of this sheaf at an indecomposable injective N is just R_N. Such a ring need not be local but its centre will be and so the centre of LDef_R is a sheaf of local commutative rings.

As remarked already this is the usual definition of structure sheaf for a commutative Noetherian ring extended to a more general context. One can also check that, for such a ring R, the restriction of LDef_R to the (Ziegler-closed) subset, Inj_R, of Zg_R is just the usual structure sheaf of R.

In fact, the sheaf of locally definable scalars is just a part of a richer structure. Consider the presheaf which assigns to a closed subset X of Zg_R the corresponding localisation, $(R\text{-mod}, \mathbf{Ab})_\tau^{\mathrm{fp}} = ((R\text{-mod}, \mathbf{Ab})_\tau)^{\mathrm{fp}}$ where $\tau = \tau_X$, of the subcategory of finitely presented objects of the functor category. This sheaf of skeletally small Abelian categories is denoted $\mathbf{LDef}_R$ and has a natural interpretation in model-theoretic terms, as the sheaf of categories of imaginaries associated with the category of R-modules [116].

Duality of functors. Next we turn to duality between right and left modules. The basic duality, which is valid for all rings R, is between the categories, $C(R) = (R\text{-mod}, \mathbf{Ab})^{\mathrm{fp}}$ and $C(R^{\mathrm{op}}) = (\text{mod-}R, \mathbf{Ab})^{\mathrm{fp}}$ of finitely presented functors.

THEOREM 1.26 [4,53]. *For any ring R we have $C(R^{\mathrm{op}}) \simeq C(R)^{\mathrm{op}}$ via the contravariant functor which is defined on objects by taking $F \in C(R^{\mathrm{op}})$ to the functor, DF, in $C(R)$ which is given on objects by taking $L \in R\text{-mod}$ to $(F, - \otimes L)$ (since L is finitely presented one does have $- \otimes L \in (\text{mod-}R, \mathbf{Ab})^{\mathrm{fp}}$).*

The model-theoretic version of this duality [102] is that the lattice of pp formulas for right modules is anti-isomorphic to the lattice of pp formulas for left modules. Equivalently, the lattice of finitely generated subfunctors of the forgetful functor $(R_R, -) \in D(R^{\mathrm{op}})$ is anti-isomorphic to the lattice of finitely generated subfunctors of the functor $({}_R R, -) \simeq (R \otimes -, -) \in D(R)$. The correspondence sends an inclusion $F \xrightarrow{f} (R_R, -)$ to $dF = \ker(({}_R R, -) \xrightarrow{Df} DF)$. This was extended by Herzog to give a duality between Ziegler spectra.

Duality of spectra. Let X be a closed subset of Zg_R. Consider the corresponding Serre subcategory of $C(R) = (R\text{-mod}, \mathbf{Ab})^{\mathrm{fp}}$, $\mathcal{S}_X = \{F \in C(R)\colon (F, N \otimes -) = 0 \text{ for all } N \in X\}$. The duals, DF, of these functors form a Serre subcategory, $D\mathcal{S}_X$, of $C(R^{\mathrm{op}})$. In fact, we get exactly those functors $G \in C(R^{\mathrm{op}})$ such that $\overrightarrow{G}(N) = 0$ for every $N \in X$. This follows immediately from the formula $(F, M \otimes -) \simeq \overrightarrow{F}(M)$ [104] which is valid for any $F \in C(R)$ and $M \in \text{Mod-}R$. Denote by DX the closed subset, $\{N \in_R \mathrm{Zg}\colon (G, - \otimes N) = 0 \text{ for all } G \in D\mathcal{S}_X\}$, of ${}_R\mathrm{Zg}$ corresponding to $D\mathcal{S}_X$.

Recall that the collection of open subsets of any topological space forms a locale (a complete Heyting algebra, that is, a complete lattice in which meet distributes over arbitrary joins).

THEOREM 1.27. *Let R be any ring. Then the map $X \mapsto DX$ between closed subsets of Zg_R and ${}_R\mathrm{Zg}$ is a bijection which commutes with arbitrary unions and intersections. Thus* *the locales of open subsets of* Zg_R *and* ${}_R\mathrm{Zg}$ *are isomorphic* (*that is, these spaces are* "*homeomorphic at the level of topology*").

[57, 4.9] *If* R *is countable then there is actually a homeomorphism* (*that is,* "*at the level of points*") *between* $\mathrm{Zg}_R/\approx$ *and* ${}_R\mathrm{Zg}/\approx$ *which induces* $X \mapsto DX$. *Here* $\approx$ *denotes topological equivalence*: *the equivalence relation which identifies two points if they belong to exactly the same open sets.*

If X *is a closed subset of* Zg_R *such that* $\mathrm{KG}((R\text{-mod}, \mathbf{Ab})_\tau) < \infty$ (*here* $\tau = \tau_X$ *is the localisation corresponding to* X) *then to every point* $N \in X$ *there is a uniquely defined point* $DN \in DX$ *such that this correspondence induces a homeomorphism from* X *to* DX.

These results have many corollaries including, as well as equality of various dimensions for right and left modules, those below.

COROLLARIES 1.28.

(a) [57] *For any ring there is a bijection between definable classes of right and left modules.*
(b) *If* R *is a countable ring then there is a continuous pure-injective right module iff there is a continuous pure-injective left module.*
(c) ([57, 4.10], also [32, Section 6]) *For any ring* R *there is a bijection* (*which preserves endolength*) *between right and left indecomposable modules of finite endolength.*

Let us make the duality between closed sets more concrete. If M is any R-module, S is any subring of $\mathrm{End}(M)$ and E is an injective cogenerator for S-Mod then the corresponding dual of M is $M^* = \mathrm{Hom}_S({}_SM, {}_SE)$. This has a natural structure as a left R-module and, as such, it is pure-injective.

THEOREM 1.29 ([190], also see [57]). *Let* R *be a ring,* $M \in \text{Mod-}R$, S *a subring of* $\mathrm{End}(M)$, E *an injective cogenerator for* S-Mod. *Regard* $M^* = \mathrm{Hom}_S({}_SM, {}_SE)$ *as a left* R*-module. Then* $\mathrm{supp}(M^*) = D(\mathrm{supp}(M))$.

In some cases a suitable choice of S will give the duality $N \mapsto DN$ as $N \mapsto N^*$ for at least certain points of the spectrum. For instance, if R is an Artin algebra and we take S to be a minimal injective cogenerator for the category of modules over $Z(R)/J(Z(R))$ where $Z(R)$ denotes the centre of R and J denotes the Jacobson radical, then $DN \simeq N^*$ for N finitely presented (as well as for some infinite-dimensional modules N).

2.

Now we give information on the Ziegler spectra of various types of ring.

1. *Rings of finite representation type.* A ring R is said to be of **finite representation type** if every R-module is a direct sum of indecomposable modules and if there are, up to isomorphism, just finitely many indecomposable R-modules. The condition for right modules implies that for left modules. It is equivalent that $\mathrm{KG}(R) = 0$ and so, for such rings, Zg_R is a discrete space. It is an open question whether Zg_R being discrete implies that R is of finite representation type (see Section 6b below). If R is of finite representation type then every module is Σ-pure-injective.

2. *Pure semisimple rings.* A ring R is said to be **right pure-semisimple** if every right R-module is a direct sum of indecomposable modules. In this case every right R-module is Σ-pure-injective (in turn, this condition implies right pure-semisimplicity), R must be right Artinian, every indecomposable right R-module is finitely generated and there are, up to isomorphism, only finitely many indecomposable right R-modules of length n for each natural number n [102,190]. For such a ring $\mathrm{KG}(R) = \mathrm{CB}(\mathrm{Zg}_R) < \infty$. A ring which is right and left pure-semisimple must be of finite representation type. It is a long-standing open question whether or not every right pure-semisimple ring is of finite representation type. For Artin algebras this was shown to be so by Auslander [3]. Herzog [58] showed that it holds for PI rings. Simson has shown (e.g. see [167]) that the general problem reduces to questions about extensions of division algebras.

3. *Dedekind domains.* Let R be a commutative Dedekind domain (in fact, what we say here applies equally well if R is a non-commutative Dedekind domain which satisfies a polynomial identity). Then the R-modules of finite length are pure-injective and the indecomposable ones are exactly the isolated points of Zg_R. The set of isolated points is dense in Zg_R and the other points can obtained as suitable direct limits or inverse limits of these points. The points of Cantor–Bendixson rank 1 are the "Prüfer" modules (the injective hulls of the simple R-modules) and their duals, the "adic" modules (the completions of R at non-zero primes). There remains the quotient field of R: this is the unique point of (maximal) rank 2. We have $\mathrm{KG}(R) = \mathrm{CB}(\mathrm{Zg}_R) = 2$.

4. *Artin algebras.* The ring R is an **Artin algebra** if its centre is Artinian and if R is finitely generated as a module over its centre. Finite algebras and algebras which are finite-dimensional over a field are examples. Every module of finite length over such a ring is pure-injective and the existence of Auslander–Reiten sequences for such rings yields that the indecomposable modules of finite length are all isolated points of Zg_R. Furthermore these are exactly the isolated points and, together, they are dense in Zg_R.

Further description of Zg_R is very much tied up with description of the finitely generated modules and hence with the representation type of R. For the remainder of this subsection we assume that R is a finite-dimensional K-algebra where K is a field. For the precise definition of domestic, tame and wild we refer to [29] for instance (especially since, although the situation is clear-cut in the case where K is algebraically closed, it is not clear what the definitions of these terms should be in general, although it is clear that various particular algebras belong to the one category or another).

4a. *Domestic algebras.* A K-algebra R is domestic if there are finitely many representation embeddings mod-$R' \to$ mod-R where R' is a finite localisation of the polynomial ring $K[X]$ such that, for each integer $d \geqslant 1$, all but finitely many indecomposable R-modules of (K-)dimension d lie (up to isomorphism) in the union of the images of these embeddings. For characterisations of these algebras in terms of generic modules see [31,32]. Tame hereditary finite-dimensional K-algebras are examples, as are certain string algebras.

Let R be any tame hereditary finite-dimensional K-algebra. So, if K is algebraically closed, then R is Morita equivalent to a finite product of rings, each of which is the path algebra over K of an extended Dynkin quiver (for the general case see [35]). Suppose that R is indecomposable as a ring. Then the Ziegler spectrum of R is, roughly, composed of finitely many generically overlapping copies of Ziegler spectra of Dedekind domains, together with the discretely-indexed families of indecomposable pre-projective and pre-injective modules. One has $\mathrm{KG}(R) = \mathrm{CB}(\mathrm{Zg}_R) = 2$. See [112,150].

For some time all the values of $\mathrm{KG}(R)$ computed for R an Artin algebra had been 0 (finite representation type), 2 (tame hereditary and some algebras related by tilting) and ∞ (some tame but non-domestic algebras and wild algebras). Then, following work of Schröer [163], the values of $\mathrm{KG}(R)$ and $\mathrm{CB}(\mathrm{Zg}_R)$ were computed [25,164] for a series of domestic string algebras and turn out to give all integer values $3 \leqslant n < \omega$. See the references for details, including the explicit description of the points and the topology, which relies heavily on [149] and [21]. It has been shown by Krause [77] (for finite-dimensional algebras over an algebraically closed field) and Herzog [61] (for Artin algebras in general) that there is no Artin algebra R with $\mathrm{KG}(R) = 1$. Hence every finite value, apart from 1, occurs as $\mathrm{KG}(R)$ for some Artin algebra R. The author has conjectured that, for an Artin algebra R, we have $\mathrm{KG}(R) < \infty$ iff R is of domestic representation type.

4b. *Tame algebras.* A K-algebra R is tame if, for each integer $d \geqslant 1$ there are finitely many representation embeddings $\text{mod-}R' \to \text{mod-}R$ where R' is a finite localisation of the polynomial ring $K[X]$ such that all but finitely many indecomposable R-modules of (K-)dimension d lie (up to isomorphism) in the union of the images of these embeddings (but the number of representation embeddings needed may grow as d grows). Examples include string algebras (see [26]) and tame canonical algebras [148] and, in these examples (which are essentially those which have been so far computed), both $\mathrm{KG}(R)$ and $\mathrm{CB}(\mathrm{Zg}_R)$ turn out to be ∞ when the algebra is not domestic. There is still, however, the hope of being able to describe the spectra in these cases (following the idea that if the finite-dimensional modules can be described then so can the Ziegler spectrum), see, e.g., [151]. It was also conjectured that if R is a tame algebra then the uniserial dimension of (R-mod, $\mathbf{Ab}$) is defined and, in particular, there should be no continuous pure-injective R-modules but this has been contradicted by [138].

4c. *Wild algebras.* A K-algebra R is wild if there is a representation embedding from $\text{mod-}K\langle X, Y\rangle$, where $K\langle X, Y\rangle$ is the free K-algebra in two generators, to mod-R. Roughly, this means that there are two-parameter families (and hence n-parameter families for each n) of finite-dimensional R-modules and then the classification problem for finite-dimensional R-modules is considered to be impossible. For such an algebra we have $\mathrm{KG}(R) = \infty$ and so, at least for countable rings (but conjecturally for all rings), $\mathrm{CB}(\mathrm{Zg}_R) = \infty$. Furthermore the uniserial dimension of (R-mod, $\mathbf{Ab}$) is ∞ and there exist continuous pure-injective R-modules.

5. *Infinite-dimensional algebras.* For infinite-dimensional algebras the themes of classification in the tame case and impossibility of complete classification in the wild case continue, even though the terms tame and wild are not generally defined in this context.

5a. *Hereditary orders.* These are not far removed from the tame hereditary finite-dimensional algebras that we discussed above. A ring R is a **hereditary order** (in a central simple algebra) if R is a hereditary Noetherian prime ring which has, for its simple ring of quotients, a matrix ring over a division ring which is finite-dimensional over its centre. Equivalently, R is a hereditary Noetherian prime ring which satisfies some polynomial identity. In [30] Crawley-Boevey draws a parallel between and, indeed, links the categories of finite-length modules over a tame hereditary Artin algebra and over a hereditary order, with the maximal orders corresponding to those Artin algebras in which all tubes are homogeneous. Using this, the techniques of [112] which give the description of the Ziegler spectrum of a tame hereditary Artin algebra also yield the, almost identical, description for hereditary orders.

5b. *Generalised Weyl algebras.* The first Weyl algebra, $R = K\langle X, Y\colon\ YX - XY = 1\rangle$, over a field, K, of characteristic zero is a simple Noetherian hereditary domain which is not an order. Klingler and Levy [71] showed that the category of torsion modules over this ring is "wild" and their techniques can be used to show that there is a continuous pure-injective R-module. If M is any indecomposable R-module of finite length then the pure-injective hull, $\overline{M}$, of M is indecomposable and it follows from a result of Bavula [12] that no such point is isolated (see [120]). In [120] it is shown that the set of points of this form is dense in Zg_R and hence that there are no isolated points in Zg_R. These and related results are proved in [120] for a class of rings, certain generalised Weyl algebras in the sense of [11], which includes the first Weyl algebra.

5c. *Pullback rings.* If R, R' are two commutative discrete valuation domains and if there is an isomorphism between their residue fields then one may form the pullback in the sense of Levy [85]. An example of a ring so obtained is the algebra $K[X, Y\colon\ XY = 0]_{(X,Y)}$ which is the infinite-dimensional version of the Gelfand–Ponomarev algebras $K[X, Y\colon\ XY = 0 = X^n = Y^m]$ which are, for $m, n \geqslant 2$, $m + n \geqslant 5$, tame non-domestic string algebras. For such pullback rings Toffalori [171,172] classified the indecomposable pure-injective "separated" modules and Ebrahimi-Atani [36] classified all the indecomposable pure-injectives, N, such that $N/N.J(R)$ is of finite length. Note that the complete description of the Ziegler spectrum for such rings would include that for the above tame non-domestic Gelfand–Ponomarev algebras.

5d. *Differential polynomial rings.* Let K be a field and let d be a derivation on K: that is $d: K \to K$ is an additive map which satisfies $d(ab) = a.db + da.b$. Let R be the corresponding differential polynomial ring: R is the ring of polynomials, with (non-central) coefficients from K, in an indeterminate X with relations $aX = X.da$ $(a \in K)$. By varying K and d we obtain a variety of interesting examples. For instance, suppose that (K, d) is a universal field with derivation. Then R is an example of a **V-ring** (see [40] and the chapter on Max-rings and V-rings in this volume) – a ring in which every simple module is injective. In fact, there is a unique simple R-module, S, and the Ziegler spectrum of R consists of just three points: the injective module S; the "dual" of this module (the pure-injective hull of R is the pure-injective hull of a direct sum of copies of this dual module); the quotient division ring of R. The first two points are isolated (and S is even Σ-pure-injective), and we have $\mathrm{KG}(R) = \mathrm{CB}(\mathrm{Zg}_R) = 1$.

For another class of examples, Puninski [134] investigates the Ziegler spectrum of, and the finite length modules over, the ring of differential operators $\mathcal{D} = K[[X]][\frac{\delta}{\delta X}]$ where K is algebraically closed of characteristic zero. The general shape of the spectrum turns out to be similar to that over a commutative Dedekind domain (or over a tame hereditary finite-dimensional algebra). In particular the Cantor–Bendixson rank is 2. In [135] he describes the Ziegler spectrum of rings of the form $R = K[X, \delta]$ where K is a field of characteristic 0, δ is a derivation on K whose field of constants is algebraically closed and where it is assumed that the category of finite length modules has Auslander–Reiten sequences. Again the description is similar to that seen in the tame hereditary case.

6. *Regular rings.* The ring R is (**von Neumann**) **regular** if every finitely generated right (equivalently left) ideal is generated by an idempotent element. The following are equivalent: R is regular; the theory of R-modules admits elimination of quantifiers; every pp formula is equivalent to a quantifier-free formula; every module is absolutely pure; every embedding between modules is a pure embedding; every pure-injective module is injective. So for such rings the Ziegler spectrum is the set of isomorphism classes of indecomposable injective modules and the basic open sets are those of the form $\{N\colon\ N$ is an indecomposable injective and $Ne \neq 0\}$ see [153].

6a. *Commutative regular rings.* The **Pierce spectrum** of such a ring is the space of maximal ideals equipped with the Stone topology, which has, for a basis of open (and closed) sets, those of the form $\{M\colon\ e \notin M\}$ as e ranges over elements of R. An injective module is indecomposable iff it is the injective hull of a simple module, and so we have a natural bijection between Zg_R and the Pierce spectrum, which is easily seen to be a homeomorphism (in fact, the Ziegler, Zariski and Pierce topologies coincide). Examples are the Boolean rings (that is, Boolean algebras): these are the commutative regular rings R such that each factor ring R/M, where M is a maximal ideal of R, is the field with two elements.

A Boolean ring is **atomic** if every non-zero ideal contains a simple ideal and is **superatomic** if every factor ring is atomic. The terminology may be extended to general commutative regular rings via their Boolean algebras of idempotent elements and then the condition is equivalent to R being **semi-Artinian** (that is, every non-zero module has a non-zero socle). Then [45, Theorem 4], for a commutative regular ring R, R is superatomic/semi-Artinian iff the Cantor–Bendixson rank of Zg_R is defined iff $\mathrm{KG}(R) < \infty$ (and, of course, then we have $\mathrm{CB}(\mathrm{Zg}_R) = \mathrm{KG}(R)$). Otherwise, there are continuous (pure-)injective modules, irrespective of the cardinality of the ring. The **Pierce sheaf** is the sheaf which assigns to the ideal I the factor ring R/I. This is naturally identified with the sheaf of locally definable scalars.

6b. *(Non-commutative) regular rings.* Some of the results on commutative regular rings generalise. Trlifaj [176] showed that if R is a regular ring then R is semi-Artinian iff there is no continuous pure-injective R-module iff $\mathrm{UD}(R) < \infty$ iff m-dim$(R) < \infty$ iff $\mathrm{KG}(R) < \infty$. Also see [177] for $\mathrm{KG}(R)$ and $\mathrm{CB}(\mathrm{Zg}_R)$.

The connection between $\mathrm{KG}(R)$ and $\mathrm{CB}(\mathrm{Zg}_R)$ is, however, open even in this case. Indeed [176] there exists a regular ring R which does not satisfy the isolation condition iff there exists a regular ring R' which is simple, non-Artinian and with $|\mathrm{Zg}_{R'}| = 1$. The existence of such a ring R' is an open question. One does have [176] that the isolation condition holds for regular rings R which satisfy any of the following conditions: $|R| < 2^{\aleph_0}$; all primitive factor rings of R are Artinian; R is semi-Artinian.

If R is a regular ring with all primitive factor rings Artinian then every indecomposable injective R-module is (the injective hull of) a simple module (cf. [51]) and so Zg_R is homeomorphic to the maximal ideal space of R. In particular, Zg_R is a T_1 space. Furthermore if R is regular with all primitive factor rings Artinian then the following conditions are equivalent: Zg_R is Hausdorff; Zg_R is a normal space; Zg_R is totally disconnected; R is a biregular ring (for every $x \in R$, the ideal RxR is generated by a central idempotent). The paper [177] contains further results about the relationship between maxspec(R) and Zg_R in the general regular case.

7. *Serial rings.* The model theory of modules over serial rings was independently investigated by Eklof and Herzog [37] and by Puninski [131]. In both these papers a particularly nice basis of the Ziegler topology was found and general characterisations of indecomposable pure-injectives in terms of the ideals of R were given. Subsequently Puninski [132, 133], in the commutative case, and Reynders [144] have investigated Cantor–Bendixson rank for Ziegler spectra of serial rings. One has [133] that if R is a commutative valuation domain then $\mathrm{Kdim}(R) < \infty$ iff $\mathrm{KG}(R) < \infty$ iff $\mathrm{CB}(\mathrm{Zg}_R) < \infty$ iff there is no continuous pure-injective R-module. Furthermore [144] bounds $\mathrm{KG}(R)$ in terms of the Krull dimension of R when R is serial.

Also in [128] it is shown that if R is a commutative valuation domain then there is a superdecomposable pure-injective R-module iff the value group of R contains, as a partially ordered subset, a copy of the rationals. See also [136].

In [140] information about the continuous pure-injective modules over a commutative valuation domain is obtained.

8. *Pseudo-finite-dimensional representations of $U(sl(2, K))$.* Herzog [63] considers the closure, C, in the Ziegler spectrum of the finite-dimensional representations (regarded as representations of the universal enveloping algebra $U(sl(2, K))$) of the Lie algebra $sl(2, K)$. By a pseudo-finite-dimensional representation is meant one with support contained in C. Herzog shows that the canonical morphism from $U(sl(2, K))$ to the ring of definable scalars of this set is an epimorphism to a von Neumann regular ring which has the (continuum many) points of C as its simple torsion representations. He also extends some of the theory of weights to these representations.

9. *Stable and triangulated categories.* Benson and Gnacadja [15] show that certain of the idempotent modules of Rickard [145] in the stable module category for a finite group are pure-injective. These results have been extended by Benson and Krause [18]. Krause [81] has shown how to define the Ziegler spectrum of any compactly generated triangulated category and in [17] Benson and Krause find the projective spectrum of the Tate cohomology ring of a finite group as a part of the Zariski spectrum (in the sense used in this paper) of the group ring of that group.

10. *Modules over group rings and lattices.* The model theory of modules over group rings and of lattices over orders has been investigated, in particular by Marcja and Toffalori, especially with a view to showing that the tame/wild dichotomy corresponds to the split between (a ring having) decidable/undecidable theory of modules. See [173,174] and, for example, [175,89,87]. These papers also provide a great variety of examples of interpretations of classes of additive structures in other such classes.

11. *Decidability/undecidability.* The word problem for groups is known to be undecidable: there is no algorithm which, when input with any word w and words $w_1, \ldots, w_n$ will decide whether or not w represents the identity element in the free group factored by the normal subgroup generated by the words $w_1, \ldots, w_n$.

Baur [9,10] and others (see [103, Chapter 17]) showed that this unsolvable word problem for groups can be encoded in the theory of modules over various rings. For example the theory of $K\langle X, Y\rangle$-modules, where K is any field, encodes the word problem for groups and hence is **undecidable**, meaning that there is no algorithm which, input with any sentence from the language of R-modules, will decide whether or not it is true in all R-modules.

It has been conjectured by the author that if R is a wild K-algebra then the theory of R-modules interprets that of $K\langle X, Y\rangle$-modules and hence is undecidable (this is known to be so for strictly wild algebras, see [108]). Indeed all current evidence is in favour of the implication "wild implies undecidable" even outside the context of finite-dimensional algebras, in particular for group rings [173,174]. The evidence for the implication "tame implies decidable" is rather less compelling but, still, in the light of what is currently known, it seems not unreasonable to conjecture that this is so. Ziegler [184] showed that if enough is known about the topology of the spectrum then one obtains a decision procedure for the theory of modules. In all cases where decidability of the theory of modules has been established one has an explicit description of the spectrum. So explicit description of Ziegler spectra also has this application.

References

[1] J. Adámek and J. Rosický, *Locally Presentable and Accessible Categories*, London Math. Soc. Lecture Note Ser., Vol. 189, Cambridge University Press (1994).

[2] M. Auslander, *Coherent functors*, Proceedings of the Conference on Categorical Algebra, La Jolla, 1965, Springer-Verlag, Berlin (1966), 189–231.

[3] M. Auslander, *Large modules over Artin algebras*, Algebra, Topology and Category Theory, Academic Press, New York (1976), 1–17.

[4] M. Auslander, *Isolated singularities and existence of almost split sequences*, Notes by Louise Unger, Representation Theory II, Groups and Orders, Ottawa, 1984, Lecture Notes in Math., Vol. 1178, Springer-Verlag, Berlin (1986), 194–242.

[5] M. Auslander, I. Reiten and S.O. Smalo, *Representation Theory of Artin Algebras*, Cambridge University Press (1995).

[6] J.T. Baldwin and A.H. Lachlan, *On universal Horn classes categorical in some infinite power*, Algebra Universalis **3** (1973), 98–111.
[7] S. Baratella and M. Prest, *Pure-injective modules over the dihedral algebras*, Comm. Algebra **25** (1997), 11–32.
[8] S. Baratella and C. Toffalori, *The theory of $\mathbb{Z}C(2)^2$-lattices is decidable*, Arch. Math. Logic **37** (1998), 91–104.
[9] W. Baur, *Decidability and undecidability for theories of abelian groups with predicates for subgroups*, Compositio Math. **31** (1975), 23–30.
[10] W. Baur, *Undecidability of the theory of abelian groups with a subgroup*, Proc. Amer. Math. Soc. **55** (1976), 125–128.
[11] V. Bavula, *Generalised Weyl algebras and their representations*, Algebra i Analiz **4** (1) (1992), 75–97; English transl.: St. Petersburg Math. J. **4** (1) (1993), 71–92.
[12] V. Bavula, *The extension group of the simple modules over the first Weyl algebra*, Bull. London Math. Soc. **32** (2000), 182–190.
[13] V. Bavula and F. van Oystaeyen, *The simple modules of certain generalised crossed products*, J. Algebra **194** (1997), 521–566.
[14] A. Beligiannis, *Relative homological algebra and purity in triangulated categories*, J. Algebra **227** (2000), 268–361.
[15] D. Benson and G. Gnacadja, *Phantom maps and purity in modular representation theory. I*, Fund. Math. **161** (1999), 27–91.
[16] D. Benson and G. Gnacadja, *Phantom maps and purity in modular representation theory. II*, Algebras and Representation Theory **4** (2001), 395–404.
[17] D. Benson and H. Krause, *Generic idempotent modules for a finite group*, Algebras and Representation Theory **3** (2000), 337–346.
[18] D. Benson and H. Krause, *Pure injectives and the spectrum of the cohomology ring of a finite group*, J. Reine Angew. Math. **542** (2002), 23–51.
[19] K.I. Beidar, A.V. Mikhalev and G. Puninski, *Logical aspects of ring and module theory*, Fund. i Prikl. Mat. **1** (1995), 1–62.
[20] K. Burke, *Some model-theoretic properties of functor categories for modules*, Doctoral Thesis, University of Manchester (1994).
[21] K. Burke, *Pure-injective hulls of expanding string modules*, Preprint, University of Manchester (1997).
[22] K. Burke, *Some connections between finitely generated functors and pp-types*, Preprint, University of Manchester (1997).
[23] K. Burke, *Co-existence of Krull filtrations*, J. Pure Appl. Algebra **155** (2001), 1–16.
[24] K. Burke and M. Prest, *Rings of definable scalars and biendomorphism rings*, Model Theory of Groups and Automorphism Groups, London Math. Soc. Lecture Note Ser., Vol. 244, Cambridge University Press (1997), 188–201.
[25] K. Burke and M. Prest, *The Ziegler and Zariski spectra of some domestic string algebras*, Algebras and Representation Theory **5** (2002), 211–234.
[26] M.C.R. Butler and C.M. Ringel, *Auslander–Reiten sequences with few middle terms and applications to string algebras*, Comm. Alg. **15** (1987), 145–179.
[27] C.C. Chang and H.J. Keisler, *Model Theory*, North-Holland, Amsterdam (1973), 3rd edn (1990).
[28] P.M. Cohn, *On the free product of associative rings*, Math. Z. **71** (1959), 380–398.
[29] W. Crawley-Boevey, *On tame algebras and bocses*, Proc. London Math. Soc. **56** (1988), 451–483.
[30] W. Crawley-Boevey, *Regular modules for tame hereditary algebras*, Proc. London Math. Soc. **62** (1991), 490–508.
[31] W. Crawley-Boevey, *Tame algebras and generic modules*, Proc. London Math. Soc. **63** (1991), 241–265.
[32] W. Crawley-Boevey, *Modules of finite length over their endomorphism rings*, Representations of Algebras and Related Topics, London Math. Soc. Lecture Note Ser., Vol. 168, Cambridge University Press (1992), 127–184.
[33] W. Crawley-Boevey, *Locally finitely presented additive categories*, Comm. Algebra **22** (1994), 1641–1674.
[34] W. Crawley-Boevey, *Infinite-dimensional modules in the representation theory of finite-dimensional algebras*, Proc. ICRA VIII, CMS Conf. Proc. Ser., Vol. 23 (1998), 29–54.

[35] V. Dlab and C.M. Ringel, *Indecomposable representations of graphs and algebras*, Mem. Amer. Math. Soc., No. 173, Providence, RI (1976).
[36] S. Ehrahimi-Atani, *On pure-injective modules over pullback rings*, Comm. Algebra **28** (2000), 4037–4069.
[37] P.C. Eklof and I. Herzog, *Model theory of modules over a serial ring*, Ann. Pure Appl. Math. **72** (1995), 145–176.
[38] P. Eklof and G. Sabbagh, *Model-completions and modules*, Ann. Math. Logic **2** (1971), 251–295.
[39] A. Facchini and G. Puninski, *Σ-pure-injective modules over a serial ring*, Abelian Groups and Modules, Kluwer, Dordrecht (1995), 145–162.
[40] C. Faith, *Algebra: Rings, Modules and Categories I*, Springer-Verlag, Berlin (1973).
[41] E. Fisher, *Powers of saturated modules*, J. Symbolic Logic **37** (1972), 777.
[42] E. Fisher, *Abelian Structures*, Yale University (1974/5); Partly published as *Abelian Structures I*, Abelian Group Theory, Lecture Notes in Math., Vol. 616, Springer-Verlag, Berlin (1977), 270–322.
[43] P. Gabriel, *Des catégories abéliennes*, Bull. Soc. Math. (France) **90** (1962), 323–448.
[44] S. Garavaglia, *Decomposition of totally transcendental modules*, J. Symbolic Logic **45** (1980), 155–164.
[45] S. Garavaglia, *Dimension and rank in the model theory of modules*, Preprint, University of Michigan (1979); revised (1980).
[46] W. Geigle, *The Krull–Gabriel dimension of the representation theory of a tame hereditary Artin algebra and applications to the structure of exact sequences*, Manuscripta Math. **54** (1985), 83–106.
[47] W. Geigle and H. Lenzing, *A class of weighted projective lines arising in representation theory of finite-dimensional algebras*, Singularities, Representations of Algebras and Vector Bundles, Lecture Notes in Math., Vol. 1273, Springer-Verlag, Berlin (1987), 265–297.
[48] G. Geisler, *Zur Modelltheorie von Moduln*, Doctoral dissertation, Universität Freiburg (1994).
[49] I.M. Gelfand and V.A. Ponomarev, *Indecomposable representations of the Lorenz group*, Uspekhi Mat. Nauk **23** (2) (1968); transl.: Russian Math. Surveys **23** (2) (1968), 1–58.
[50] K.R. Goodearl, *Von Neumann Regular Rings*, Pitman, London (1979).
[51] K.R. Goodearl, *Artinian and Noetherian modules over regular rings*, Comm. Algebra **8** (1980), 477–504.
[52] L. Gruson and C.U. Jensen, *Modules algébriquement compact et foncteurs* $\varprojlim^{(i)}$, C. R. Acad. Sci. Paris **276** (1973), 1651–1653.
[53] L. Gruson and C.U. Jensen, *Dimensions cohomologiques reliées aux foncteurs* $\varprojlim^{(i)}$, Lecture Notes in Math., Vol. 867, Springer-Verlag, Berlin (1981), 243–294.
[54] R. Hartshorne, *Coherent functors*, Adv. in Math. **140** (1998), 44–94.
[55] I. Herzog, *Modules with few types*, J. Algebra **149** (1992), 358–370.
[56] I. Herzog, *The Auslander–Reiten translate*, Contemp. Math. **130** (1992), 153–165.
[57] I. Herzog, *Elementary duality of modules*, Trans. Amer. Math. Soc. **340** (1993), 37–69.
[58] I. Herzog, *A test for finite representation type*, J. Pure Appl. Algebra **95** (1994), 151–182.
[59] I. Herzog, *Dualizing rings and a characterisation of finite representation type*, C. R. Acad. Sci. Paris **322** (1996), 507–510.
[60] I. Herzog, *The Ziegler spectrum of a locally coherent Grothendieck category*, Proc. London Math. Soc. **74** (1997), 503–558.
[61] I. Herzog, *The endomorphism ring of a localised coherent functor*, J. Algebra **191** (1997), 416–426.
[62] I. Herzog, *Locally simple objects*, P. Eklof and R. Göbel (eds.), Proc. Int. Conf. on Abelian Groups and Modules, Dublin, Birkhäuser (1999), 341–351.
[63] I. Herzog, *The pseudo-finite dimensional representations of $sl(2,k)$*, Selecta Math. **7** (2001), 241–290.
[64] I. Herzog, *Pure injective envelopes*, Preprint, Ohio State University (2002).
[65] I. Herzog and V. Puninskaya, *The model theory of divisible modules over a domain*, Fund. i Prikl. Mat. **2** (1996), 563–594.
[66] M. Hochster, *Prime ideal structure in commutative rings*, Trans. Amer. Math. Soc. **142** (1969), 43–60.
[67] B. Huisgen-Zimmermann, *Purity, algebraic compactness, direct sum decompositions and representation type*, in: [83], 331–367.
[68] C.U. Jensen and H. Lenzing, *Model Theoretic Algebra*, Gordon and Breach, London (1989).
[69] I. Kaplansky, *Infinite Abelian Groups*, Univ. of Michigan Press, Ann Arbor (1954).
[70] R. Kielpinski, *On Γ-pure injective modules*, Bull. Polon. Acad. Sci. **15** (1967), 127–131.

[71] L. Klingler and L. Levy, *Wild torsion modules over Weyl algebras and general torsion modules over HNPs*, J. Algebra **172** (1995), 273–300.

[72] H. Krause, *The endocategory of a module*, CMS Conf. Proc. **18** (1996), 419–432.

[73] H. Krause, *The spectrum of a locally coherent category*, J. Pure Appl. Algebra **114** (1997), 259–271.

[74] H. Krause, *Stable equivalence preserves representation type*, Comm. Math. Helv. **72** (1997), 266–284.

[75] H. Krause, *The Spectrum of a Module Category*, Habilitationsschrift, Universität Bielefeld (1997); published as Mem. Amer. Math. Soc. No. 707 (2001).

[76] H. Krause, *Exactly definable categories*, J. Algebra **201** (1998), 456–492.

[77] H. Krause, *Generic modules over Artin algebras*, Proc. London Math. Soc. **76** (1998), 276–306.

[78] H. Krause, *Functors on locally finitely presented categories*, Colloq. Math. **75** (1998), 105–132.

[79] H. Krause, *Finitistic dimension and the Ziegler spectrum*, Proc. Amer. Math. Soc. **126** (1998), 983–987.

[80] H. Krause, *Decomposing thick subcategories of the stable module category*, Math. Ann. **313** (1999), 95–108.

[81] H. Krause, *Smashing subcatgories and the telescope conjecture – an algebraic approach*, Invent. Math. **139** (2000), 99–133.

[82] H. Krause, *Coherent functors in stable homotopy theory*, Fundamenta Math. **173** (2002), 33–56.

[83] H. Krause and C.M. Ringel (eds.), *Infinite Length Modules*, Birkhäuser (2000).

[84] T.G. Kucera and M. Prest, *Imaginary modules*, J. Symbolic Logic **57** (1992), 698–723.

[85] L. Levy, *Modules over pullbacks and subdirect sums*, J. Algebra **71** (1981), 50–61.

[86] M. Makkai and R. Parè, *Accessible Categories: the Foundations of Categorical Model Theory*, Contemp. Math., Vol. 104, Amer. Math. Soc., Providence, RI (1989).

[87] A. Marcja, M. Prest and C. Toffalori, *The stability classification for Abelian-by-finite groups and modules over a group ring*, J. London Math. Soc. **47** (1993), 212–226.

[88] A. Marcja, M. Prest and C. Toffalori, *The torsion free part of the Ziegler spectrum of RG when R is a Dedekind domain and G is a finite group*, J. Symbolic Logic, to appear.

[89] A. Marcja and C. Toffalori, *Decidable representations*, J. Pure Appl. Algebra **103** (1995), 189–203.

[90] H. Marubayashi, *Modules over bounded Dedekind prime rings. II*, Osaka J. Math. **9** (1972), 427–445.

[91] E. Matlis, *Injective modules over Noetherian rings*, Pacific J. Math. **8** (1958), 511–528.

[92] B. Mitchell, *Rings with several objects*, Adv. Math. **8** (1972), 1–161.

[93] E. Monari-Martinez, *On Σ-pure-injective modules over valuation domains*, Arch. Math. **46** (1986), 26–32.

[94] R.S. Pierce, *Modules over Commutative Regular Rings*, Mem. Amer. Math. Soc., Vol. 70 (1967).

[95] F. Point and M. Prest, *Decidability for theories of modules*, J. London Math. Soc. **38** (1988), 193–206.

[96] N. Popescu, *Abelian Categories with Applications to Rings and Modules*, Academic Press, New York (1973).

[97] M. Prest, *Some model-theoretic aspects of torsion theories*, J. Pure Appl. Algebra **12** (1978), 295–310.

[98] M. Prest, *Some model-theoretic properties of torsion theories in Abelian categories*, Doctoral Thesis, University of Leeds (1979).

[99] M. Prest, *Elementary torsion theories and locally finitely presented Abelian categories*, J. Pure Appl. Algebra **18** (1980), 205–212.

[100] M. Prest, *Tame categories of modules and decidability*, University of Liverpool (1985).

[101] M. Prest, *Model theory and representation type of algebras*, Logic Colloquium '86: Proceedings of the European Meeting of the Association for Symbolic Logic, Hull, 1986, North-Holland, Amsterdam (1988), 219–260.

[102] M. Prest, *Duality and pure-semisimple rings*, J. London Math. Soc. **38** (1988), 193–206.

[103] M. Prest, *Model Theory and Modules*, London Math. Soc. Lecture Notes Ser., Vol. 130, Cambridge University Press (1988).

[104] M. Prest, *Remarks on elementary duality*, Ann. Pure Appl. Logic **62** (1993), 183–205.

[105] M. Prest, *The (pre-)sheaf of definable scalars*, Preprint, University of Manchester (1994).

[106] M. Prest, *A note concerning the existence of many indecomposable infinite-dimensional pure-injective modules over some finite-dimensional algebras*, Preprint, University of Manchester (1995).

[107] M. Prest, *Representation embeddings and the Ziegler spectrum*, J. Pure Appl. Algebra **113** (1996), 315–323.

[108] M. Prest, *Epimorphisms of rings, interpretations of modules and strictly wild algebras*, Comm. Algebra **24** (1996), 517–531.

[109] M. Prest, *Interpreting modules in modules*, Ann. Pure Appl. Logic **88** (1997), 193–215.

[110] M. Prest, *Morphisms between finitely presented modules and infinite-dimensional representations*, Canad. Math. Soc. Conf. Proc. Ser. **24** (1998), 447–455.

[111] M. Prest, *Ideals in* mod-R *and the* ω*-radical*, Preprint, University of Manchester (1998).

[112] M. Prest, *Ziegler spectra of tame hereditary algebras*, J. Algebra **207** (1998), 146–164.

[113] M. Prest, *The Zariski spectrum of the category of finitely presented modules*, Preprint, University of Manchester (1998).

[114] M. Prest, *Sheaves of definable scalars over tame hereditary algebras*, Preprint, University of Manchester (1998).

[115] M. Prest, *Tensor product and theories of modules*, J. Symbolic Logic **64** (1999), 617–628.

[116] M. Prest, *The sheaf of locally definable scalars over a ring*, Models and Computability, London Math. Soc. Lecture Note Ser., Vol. 259, Cambridge University Press (1999), 339–351.

[117] M. Prest, *Topological and geometric aspects of the Ziegler spectrum*, in: [83], 369–392.

[118] M. Prest, *Model theory for algebra*, Handbook of Algebra, Vol. 3, M. Hazewinkel, ed., Elsevier, Amsterdam (2003), 199–226.

[119] M. Prest and V. Puninskaya, *Modules with few types over some finite-dimensional algebras*, J. Symbolic Logic, to appear.

[120] M. Prest and G. Puninski, *Some model theory over hereditary Noetherian domains*, J. Algebra **211** (1999), 268–297.

[121] M. Prest and G. Puninski, *Pure-injective envelopes of finite length modules over a Generalised Weyl Algebra*, J. Algebra **251** (2002), 150–177.

[122] M. Prest, Ph. Rothmaler and M. Ziegler, *Extensions of elementary duality*, J. Pure Appl. Algebra **93** (1994), 33–56.

[123] M. Prest, Ph. Rothmaler and M. Ziegler, *Absolutely pure and flat modules and "indiscrete" rings*, J. Algebra **174** (1995), 349–372.

[124] M. Prest and J. Schröer, *Serial functors, Jacobson radical and representation type*, J. Pure Appl. Algebra **170** (2002), 295–307.

[125] M. Prest and R. Wisbaner, *Finite presentation and purity in categories* $\sigma[M]$, Preprint, Universities of Düsseldorf and Manchester (2002).

[126] V. Puninskaya, *Injective minimal modules*, Algebra i Logika **30** (1994), 211–226; transl.: Algebra and Logic **33** (1994), 120–128.

[127] V. Puninskaya, *Modules with few types over a hereditary Noetherian prime ring*, J. Symbolic Logic **66** (2001), 271–280.

[128] G. Puninski, *Superdecomposable pure-injective modules over commutative valuation rings*, Algebra i Logika **31** (1992), 655–671; transl.: Algebra and Logic **31** (1992), 377–386.

[129] G. Puninski, *RD-formulas and W-rings*, Mat. Zametki **53** (1993), 95–103 (in Russian).

[130] G. Puninski, *On Warfield rings*, Algebra i Logika **33** (1994), 264–285; transl.: Algebra and Logic **33** (1994), 147–159.

[131] G. Puninski, *Indecomposable pure-injective modules over serial rings*, Trudy Moskov. Mat. Obshch. **56** (1995), 176–191; transl.: Trans. Moscow Math. Soc. **56** (1995), 137–147.

[132] G. Puninski, *Pure-injective modules over right Noetherian serial rings*, Comm. in Algebra **23** (1995), 1579–1591.

[133] G. Puninski, *Cantor–Bendixson rank of the Ziegler spectra over a commutative valuation domain*, J. Symbolic Logic **64** (1999), 1512–1518.

[134] G. Puninski, *Pure-injective and finite length modules over a ring of differential operators*, J. Algebra **231** (2000), 546–560.

[135] G. Puninski, *Pure-injective and finite length modules over certain rings of differential polynomials*, Proc. Moscow–Taiwan Workshop. Y. Fong et al., eds, Springer-Verlag, Berlin (2000), 171–185.

[136] G. Puninski, *Serial Rings*, Klüwer, Dordrecht (2001).

[137] G. Puninski, *When a superdecomposable module over a serial ring exists*, Preprint, University of Manchester (2001).

[138] G. Puninski, *Superdecomposable pure injective modules exist over some string algebras*, Preprint, University of Manchester (2001).

[139] G. Puninski, *The isolation property and the finitely presented Krull dimension over a serial ring*, Preprint, University of Manchester (2002).
[140] G. Puninski, *Pure-injective modules over commutative valuation domains*, Algebras and Representation Theory, to appear.
[141] G. Puninski, M. Prest and Ph. Rothmaler, *Rings described by various purities*, Comm. Algebra **27** (1999), 2127–2162.
[142] G. Puninski and R. Wisbauer, *Σ-injective modules over left duo and left distributive rings*, J. Pure Appl. Algebra **113** (1996), 55–66.
[143] R. Rentschler and P. Gabriel, *Sur la dimension des anneaux et ensembles ordonnés*, C. R. Acad. Sci. Paris, Ser. A **5** (1967), 712–715.
[144] G. Reynders, *Ziegler spectra of serial rings with Krull dimension*, Comm. Algebra **27** (1999), 2583–2611.
[145] J. Rickard, *Idempotent modules in the stable category*, J. London Math. Soc. **56** (1997), 149–170.
[146] C.M. Ringel, *The indecomposable representations of the dihedral 2-groups*, Math. Ann. **214** (1975), 19–34.
[147] C.M. Ringel, *Infinite-dimensional representations of finite-dimensional hereditary algebras*, Sympos. Math. Inst. Naz. Alta. Mat. **23** (1979), 321–412.
[148] C.M. Ringel, *Tame Algebras and Integral Quadratic Forms*, Lecture Notes in Math., Vol. 1099, Springer-Verlag, Berlin (1984).
[149] C.M. Ringel, *Some algebraically compact modules I*, Abelian Groups and Modules, A. Facchini and C. Menini, eds, Kluwer, Dordrecht (1995), 419–439.
[150] C.M. Ringel, *The Ziegler spectrum of a tame hereditary algebra*, Colloq. Math. **76** (1998), 105–115.
[151] C.M. Ringel, *Infinite length modules: some examples as introduction*, in: [83], 1–73.
[152] J.-E. Roos, *Locally Noetherian categories and generalised strictly linearly compact rings: Applications*, Category Theory, Homology Theory and their Applications, Lecture Notes in Math., Vol. 92, Springer-Verlag, Berlin (1969), 197–277.
[153] Ph. Rothmaler, *Some model theory of modules. II. On stability and categoricity of flat modules*, J. Symbolic Logic **48** (1983), 970–985.
[154] Ph. Rothmaler, *Purity in model theory*, Advances in Algebra and Model Theory, Gordon and Breach, London (1997), 445–469.
[155] Ph. Rothmaler, *Mittag-Leffler modules*, Ann. Pure Appl. Logic **88** (1997), 227–239.
[156] G. Sabbagh, *Sur la pureté dans les modules*, C. R. Acad. Sci. Paris **271** (1970), 865–867.
[157] G. Sabbagh, *Aspects logiques de la pureté dans les modules*, C. R. Acad. Sci. Paris **271** (1970), 909–912.
[158] G. Sabbagh, *Sous-modules purs, existentiellement clos et élémentaires*, C. R. Acad. Sci. Paris **272** (1970), 1289–1292.
[159] G. Sabbagh and P. Eklof, *Definability problems for rings and modules*, J. Symbolic Logic **36** (1971), 623–649.
[160] L. Salce, *Valuation domains with superdecomposable pure injective modules*, Lecture Notes Pure Appl. Math., Vol. 146, Springer-Verlag, Berlin (1993), 241–245.
[161] M. Schmidmeier, *Auslander–Reiten–Köcher für Artinische Ringe mit Polynomidentität*, Diss. Universität Munchen (1996).
[162] A. Schofield, *Representations of Rings over Skew Fields*, London Math. Soc. Lecture Notes Ser., Vol. 92, Cambridge University Press (1985).
[163] J. Schroër, *Hammocks for String Algebras*, Habilitationsschrift, Universität Bielefeld (1998).
[164] J. Schroër, *On the Krull–Gabriel dimension of an algebra*, Math. Z. **233** (2000), 287–303.
[165] J. Schroër, *The Krull–Gabriel dimension of an algebra – open problems*, in: [83], 419–424.
[166] S. Shelah, *Classification Theory and the Number of Non-Isomorphic Models*, North-Holland, Amsterdam (1978).
[167] D. Simson, *A class of potential counterexamples to the pure semisimplicity conjecture*, Advances in Algebra and Model Theory, Gordon and Breach, London (1997), 345–373.
[168] B. Stenström, *Rings of Quotients*, Springer-Verlag, Berlin (1975).
[169] W. Szmielew, *Elementary properties of abelian groups*, Fund. Math. **41** (1954), 203–271.
[170] B.R. Tennison, *Sheaf Theory*, London Math. Soc. Lecture Note Ser., Vol. 20, Cambridge University Press (1975).

[171] C. Toffalori, *Decidability for $\mathbb{Z}G$-modules when G is cyclic of prime order*, Math. Logic Quart. **42** (1996), 369–378.

[172] C. Toffalori, *Some decidability results for $\mathbb{Z}G$-modules when G is cyclic of squarefree order*, Math. Logic Quart. **42** (1996), 433–445.

[173] C. Toffalori, *Decidability for modules over a group ring. III*, Comm. Algebra **24** (1996), 331–344.

[174] C. Toffalori, *Wildness implies undecidability for lattices over group rings*, J. Symbolic Logic **62** (1997), 1429–1447.

[175] C. Toffalori, *An undecidability theorem for lattices over group rings*, Ann. Pure Appl. Logic **88** (1997), 241–262.

[176] J. Trlifaj, *Two problems of Ziegler and uniform modules over regular rings*, Abelian Groups and Modules, M. Dekker, New York (1996), 373–383.

[177] J. Trlifaj, *Uniform modules, Γ-invariants, and Ziegler spectra of regular rings*, Trends in Mathematics, Birkhäuser (1999), 327–340.

[178] L.V. Tyukavkin, *Model completeness of certain theories of modules*, Algebra i Logika **21** (1) (1982), 73–83; transl.: Algebra and Logic **21** (1) (1982), 50–57.

[179] R.B. Warfield, *Purity and algebraic compactness for modules*, Pacific J. Math. **28** (1969), 699–710.

[180] R.B. Warfield, *Serial rings and finitely presented modules*, J. Algebra **37** (1975), 187–222.

[181] E. Wilkins, *An equivalence induced by Ext and Tor applied to the finitistic weak dimension of coherent rings*, Glasgow Math. J. **40** (1998), 167–176.

[182] R. Wisbauer, *Foundations of Module and Ring Theory*, Gordon and Breach, Philadelphia (1991).

[183] D. Zelinsky, *Linearly compact rings and modules*, Amer. J. Math. **75** (1953), 79–90.

[184] M. Ziegler, *Model theory and modules*, Ann. Pure Appl. Logic **26** (1984), 149–213.

[185] W. Zimmermann, *Rein injektive direkte Summen von Moduln*, Comm. Algebra **5** (1977), 1083–1117.

[186] W. Zimmermann, *Auslander–Reiten sequences over Artinian rings*, J. Algebra **119** (1988), 366–392.

[187] W. Zimmermann, *Auslander–Reiten sequences over derivation polynomial rings*, J. Pure Appl. Algebra **74** (1991), 317–332.

[188] W. Zimmermann, *Extensions of three classical theorems to modules with maximum condition for finite matrix subgroups*, Forum Math. **10** (1998), 377–392.

[189] B. Zimmermann-Huisgen and W. Zimmermann, *Algebraically compact rings and modules*, Math. Z. **161** (1978), 81–93.

[190] B. Zimmermann-Huisgen and W. Zimmermann, *On the sparsity of representations of rings of pure global dimension zero*, Trans. Amer. Math. Soc. **320** (1990), 695–713.

[191] G. Zwara, *Tame algebras and degenerations of modules*, in: [83], 311–319.

This Page Intentionally Left Blank

Section 3A
Commutative Rings and Algebras

This Page Intentionally Left Blank

Monomial Algebras and Polyhedral Geometry

Rafael H. Villarreal

Centro de Investigación y de Estudios Avanzados del I.P.N., Departamento de Matemáticas, Apartado Postal 14-740, 07000 Mexico City, D.F., Mexico

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1. Introduction

Let $R = K[\mathbf{x}] = K[x_1, \ldots, x_n]$ be a polynomial ring in the indeterminates $x_1, \ldots, x_n$, over the field K and let

$$f_i = x^{v_i} = x_1^{v_{1i}} \cdots x_n^{v_{ni}} \quad (i = 1, \ldots, q),$$

be a finite set of monomials of R. We are interested in studying the following *monomial algebras* along with their presentation ideals:

- the *monomial ring*: $K[\mathbf{x}]/(f_1, \ldots, f_q)$,
- the *face ring* or *Stanley–Reisner ring*: $K[\mathbf{x}]/(f_1, \ldots, f_q)$, if the monomials are square-free,
- the *monomial subring*: $K[f_1, \ldots, f_q] \subset K[\mathbf{x}]$,
- the *Rees algebra*: $K[\mathbf{x}, f_1 t, \ldots, f_q t] \subset K[\mathbf{x}, t]$, which is also a monomial subring, and
- the *toric ideal*: the ideal of relations of a monomial subring.

If such monomials are square-free they are indexed by a hypergraph built on the set of indeterminates, which provides a second combinatorial structure in addition to the associated Stanley–Reisner simplicial complex.

This chapter is intended as an introduction to monomial algebras and its connections to combinatorics, graph theory and polyhedral geometry.

Some important notions from commutative algebra that have played a role in the development of the theory, such as *Cohen–Macaulay ring*, *normal ring*, *Gorenstein ring*, *integral closure*, *Hilbert series*, and *local cohomology* will be introduced.

As applications we present the upper bound theorem for the number of faces of a simplicial sphere [111], a description of the integral closure of an edge subring [108], a generalized marriage theorem for a certain family of graphs [74,129], and a study of systems of binomials in the ideal of an affine toric variety [49]. The applications and topics that we have selected illustrate the interplay between several areas of mathematics and the power of combinatorial commutative algebra techniques.

Standard references for Stanley–Reisner rings and simplicial complexes are [23, Chapter 5] and [113,131]. The initial contributions of M. Hochster [78], G. Reisner [97] and R. Stanley [111,112] were crucial for the growth and interest in the area. A great deal of effort has been directed to the understanding of diverse properties of Cohen–Macaulay simplicial complexes and their corresponding face rings [6,24,52,58,59,63,84,109,110].

There is a fruitful connection between monomial rings and monomial subrings due to the fact that the initial ideal (with respect to some term order of the variables) of a toric ideal (or any polynomial ideal) is a monomial ideal [36]. This allows to compute several invariants of projective varieties using algebraic systems such as *CoCoA* [32] and *Macaulay*2 [65]. There are some excellent book that focus in computational methods in commutative algebra and geometry [1,36,44,124]. In the center of those methods lies the Buchberger algorithm for the computation of Gröbner bases [28].

For monomial subrings and toric ideals some standard references are [23,46,117]. There is special interest in studying the family of normal homogeneous monomial subrings because of its connections with other areas such as geometry and combinatorics [20,21,27, 23,37,71,107,117,129], those subrings are studied in Section 5.

An important tool to study monomial subrings is *Normaliz* [25], which is effective in practice and can be used to find normalizations, Hilbert series, Ehrhart rings and volumes of lattice polytopes. To compute invariants of monomial subrings it is preferable to use *Normaliz* instead of *CoCoA* or *Macaulay*2, roughly speaking the reason is that the first uses integer programming techniques which are faster than Gröbner bases techniques.

One of the most striking and deep results in the theory of toric rings is the connection between Gröbner bases of toric ideals and convex polytopes discovered by Bernd Sturmfels [116,117]. Another deep result due to M. Hochster [77] is that any normal monomial subring is Cohen–Macaulay.

For the reader's convenience we have included a condensed section on commutative algebra. We will make free use of the standard terminology and notation of homological algebra (including Tor and Ext) as described in [103]. Note that (with a few exceptions) the results presented here are stated without giving detailed proofs or without giving a proof at all, but often we will point out the original source or a reference. Throughout this chapter base rings are assumed to be Noetherian and modules finitely generated.

2. Commutative algebra

In this section we are going to introduce some notions and results from commutative algebra. The main references here are [5,23,87].

A *Noetherian ring* R is a commutative ring with unit with the property that every ideal of I is finitely generated. If R is a Noetherian ring, then any finitely generated R-module M is also Noetherian.

By the Hilbert basis theorem [5] a polynomial ring $R = K[x_1, \dots, x_n]$ over a field K is Noetherian. In particular the ideal $I(X)$ of polynomials that vanish on a subset X of $\mathbb{C}^n$ is finitely generated. If R is Noetherian and I is an ideal of R, then R/I and R^n are Noetherian. Thus if R is a Noetherian ring, then any submodule of R^n is finitely generated.

THEOREM 2.1. *If M is an R-module, then M is Noetherian if and only for every ascending chain of submodules of M*

$$N_0 \subset N_1 \subset \cdots \subset N_n \subset N_{n+1} \subset \cdots \subset M$$

there is an integer k so that $N_i = N_k$ for every $i \geqslant k$.

The *spectrum* $\operatorname{Spec}(R)$ of a ring R is the set of its prime ideals. The *minimal primes* of R are the minimal primes of $\operatorname{Spec}(R)$ with respect to inclusion.

From now on by a *ring* we shall mean a Noetherian ring and by a *module* we shall mean a finitely generated module.

2.1. *Primary decomposition of modules*

Let M be an R-module and $m \in M$. The *annihilator* of m (respectively M) is:

$$\operatorname{ann}(m) = \{x \in R \mid xm = 0\} \quad (\text{respectively } \operatorname{ann}(M) = \{x \in R \mid xM = 0\}).$$

The *radical* of an ideal $I \subset R$ is $\operatorname{rad}(I) = \{x \in R \mid x^n \in I \text{ for some } n > 0\}$, the radical of I is also denoted by $\sqrt{I}$.

DEFINITION 2.2. Let M be an R-module. A prime ideal $\mathfrak{p}$ is an *associated prime* ideal of M if there is a monomorphism $\phi : R/\mathfrak{p} \to M$. The set of associated prime ideals of M will be denoted by $\operatorname{Ass}_R(M)$.

If $M = R/I$ it is usual to say that an associated prime ideal of R/I is an associated prime ideal of I and to set $\operatorname{Ass}(I) = \operatorname{Ass}(R/I)$.

Let M be an R-module. An element $x \in R$ is a *zero-divisor* of M if there is $0 \neq m \in M$ such that $xm = 0$. The set of zero divisors of M is denoted by $\mathcal{Z}(M)$. If x is not a zero-divisor of M we say that x is a *regular element* of M.

PROPOSITION 2.3. *Let M be an R-module, then $\mathcal{Z}(M) = \bigcup_{\mathfrak{p} \in \operatorname{Ass}_R M} \mathfrak{p}$.*

DEFINITION 2.4. Let M be an R-module. A submodule (respectively ideal) N of M (respectively R) is said to be *primary* if $\operatorname{Ass}_R(M/N) = \{\mathfrak{p}\}$.

THEOREM 2.5. *Let M be an R-module. If N is a submodule of M, then N has an irredundant primary decomposition $N = N_1 \cap \cdots \cap N_r$ such that*:

(a) $\operatorname{Ass}_R(M/N_i) = \{\mathfrak{p}_i\}$ *for all i.*

(b) $N \neq N_1 \cap \cdots \cap N_{i-1} \cap N_{i+1} \cap \cdots \cap N_r$ *for all i.*

(c) $\mathfrak{p}_i \neq \mathfrak{p}_j$ *if $N_i \neq N_j$.*

REMARK. (a) If $N \neq M$ and $N_1 \cap \cdots \cap N_r$ is an irredundant primary decomposition of N with $\operatorname{Ass}_R(M/N_i) = \{\mathfrak{p}_i\}$, then $\operatorname{Ass}_R(M/N) = \{\mathfrak{p}_1, \ldots, \mathfrak{p}_r\}$;

(b) if R is a ring, then the minimal primes of R are precisely the minimal primes of $\operatorname{Ass}_R(R)$.

COROLLARY 2.6. *Let R be a ring and let I be a proper ideal of R. Then I has an irredundant primary decomposition $I = \mathfrak{q}_1 \cap \cdots \cap \mathfrak{q}_r$ such that $\mathfrak{q}_i$ is a $\mathfrak{p}_i$-primary ideal and* $\operatorname{Ass}_R(R/I) = \{\mathfrak{p}_1, \ldots, \mathfrak{p}_r\}$.

Finding primary decompositions of ideals in polynomial rings over fields is difficult, for the main algorithms see [45,64,79,105]. For a specially nice treatment of the principles of primary decomposition consult the book of Wolmer V. Vasconcelos [124, Chapter 3].

2.1.1. *Modules of fractions and localizations.* Let M be an R-module and let S be a multiplicatively closed subset of R so that $1 \in S$. Then the *module of fractions* of M with respect to S, or the *localization* of M with respect to S, is

$$S^{-1}(M) = \{m/s \mid m \in M,\ s \in S\},$$

where $m/s = m_1/s_1$ if and only if $t(s_1 m - s m_1) = 0$ for some $t \in S$. In particular $S^{-1}R$ has a ring structure given by the usual rules of addition and multiplication, and $S^{-1}M$ is a module over the ring $S^{-1}R$.

EXAMPLE. (a) If $f \in R$ and $S = \{f^n \mid n \geqslant 0\}$, then $S^{-1}R$ is usually denoted by R_f. For instance if $R = \mathbb{C}[x]$ is a polynomial ring in one variable, then $R_x = \mathbb{C}[x, x^{-1}]$ is the ring of Laurent polynomials.

(b) Let $\mathfrak{p} \in \operatorname{Spec}(R)$, then $S = R \setminus \mathfrak{p}$ is a multiplicatively closed set. In this case $S^{-1}R$ is denoted by $R_\mathfrak{p}$ and is called *the localization* of R at $\mathfrak{p}$.

2.1.2. *Krull dimension and height.* By a *chain* of length n of prime ideals of a ring R we mean a strictly increasing sequence of prime ideals $\mathfrak{p}_0 \subset \mathfrak{p}_1 \subset \cdots \subset \mathfrak{p}_n$. The *Krull dimension* of R, denoted by $\dim(R)$, is the supremum of the lengths of all chains of prime ideals in R. Let $\mathfrak{p}$ be a prime ideal in R, the *height* of $\mathfrak{p}$, denoted by $\operatorname{ht}(\mathfrak{p})$ is $\operatorname{ht}(\mathfrak{p}) = \dim(R_\mathfrak{p})$. If I is an ideal of R, then its height is:

$$\operatorname{ht}(I) = \min\bigl\{\operatorname{ht}(\mathfrak{p}) \mid I \subset \mathfrak{p} \text{ and } \mathfrak{p} \in \operatorname{Spec}(R)\bigr\}.$$

The *dimension* of an R-module M is defined as $\dim(M) = \dim R/\operatorname{ann}(M)$ and its *codimension* as $\operatorname{codim}(M) = \dim(R) - \dim(M)$. A ring R is called *Artinian* if $\dim(R) = 0$.

EXAMPLE. If $R = K[x_1, \ldots, x_n]$ is a polynomial ring over a field K and $I \subset R$ is an ideal of R, then $\dim(R/I) = \dim(R) - \operatorname{ht}(I)$ and $\dim(R) = n$.

2.1.3. *Special types of rings.* A *local ring* $(R, \mathfrak{m}, K)$ is a ring R with exactly one maximal ideal $\mathfrak{m}$. The field $K = R/\mathfrak{m}$ is the *residue field* of R. The prototype of a local ring is $R_\mathfrak{p}$, the *localization* of R at a prime $\mathfrak{p}$. If $(R, \mathfrak{m}, K)$ is a local ring, then

$$\dim(R) \leqslant \dim_K \mathfrak{m}/\mathfrak{m}^2,$$

see [5]. Those rings for which the equality holds are called *regular local rings*. A ring R (not necessarily local) is *regular* if $R_\mathfrak{p}$ is a regular local ring for every $\mathfrak{p} \in \operatorname{Spec}(R)$. One of the most important properties of regular local rings is that they are unique factorization domains, see [86,87].

2.2. *Graded modules*

Let $(H,+)$ be an Abelian semigroup. An *H-graded ring* is a ring R together with a decomposition

$$R=\bigoplus_{a\in H} R_a \quad \text{(as a } \mathbb{Z}\text{-module)},$$

such that $R_aR_b\subset R_{a+b}$ for all $a,b\in H$. A *graded ring* is by definition a $\mathbb{Z}$-graded ring.

If R is an H-graded ring and M is an R-module with a decomposition

$$M=\bigoplus_{a\in H} M_a,$$

such that $R_aM_b\subset M_{a+b}$ for all $a,b\in H$, we say that M is an *H-graded module*. An element $f\in M_a$ is said to be *homogeneous* of degree a, in this case we set $\deg(f)=a$. The elements in R_a are called *forms* of degree a.

A map $\varphi\colon M\to N$ between H-graded modules is *graded* if $\varphi(M_a)\subset N_a$ for all $a\in H$. If $M=\bigoplus_{a\in H} M_a$ is an H-graded module and $N=\bigoplus_{a\in H} N_a$ is a graded submodule, that is, $N_a\subset M_a$ for all a, then M/N is an H-graded R-module with $(M/N)_a=M_a/N_a$.

EXAMPLE 2.7. Let $R=K[x_1,\dots,x_n]$ be a polynomial ring over a field K and let $d_1,\dots,d_n\in\mathbb{N}\setminus\{0\}$. We grade R by $\deg(x_i):=d_i$. For $a=(a_i)$ in $\mathbb{N}^n$ we set $x^a=x_1^{a_1}\cdots x_n^{a_n}$ and $|a|=a_1d_1+\cdots+a_nd_n$. The induced $\mathbb{N}$-grading is:

$$R=\bigoplus_{i=0}^{\infty} R_i, \quad \text{where } R_i=\bigoplus_{|a|=i} Kx^a,$$

one may extend this grading to a $\mathbb{Z}$-grading by setting $R_i=0$ for $i<0$.

DEFINITION 2.8. The *standard grading* of a polynomial ring $K[x_1,\dots,x_n]$ is the $\mathbb{N}$-grading induced by setting $\deg(x_i)=1$ for all i.

DEFINITION 2.9. Let K be a field. A *standard algebra* is a finitely generated $\mathbb{N}$-graded K-algebra

$$R=\bigoplus_{i\geqslant 0} R_i=K[y_1,\dots,y_r] \quad \text{with } y_i\in R_1 \text{ for all } i.$$

The ideal $\mathfrak{m}=R_+=\bigoplus_{i\geqslant 1} R_i$ is the irrelevant maximal ideal of R.

Hilbert polynomial. Let $R=\bigoplus_{i=0}^{\infty} R_i$ be an $\mathbb{N}$-graded ring. If M is an $\mathbb{N}$-graded R-module and R_0 is Artinian define the *Hilbert function* of M as

$$H(M,i)=\ell_{R_0}(M_i) \quad (\ell_{R_0}=\text{length w.r.t. } R_0).$$

DEFINITION 2.10. An $\mathbb{N}$-graded ring $R = \bigoplus_{i=0}^{\infty} R_i$ is called a *homogeneous ring* if $R = R_0[x_1, \ldots, x_n]$, where $\deg(x_i) = 1$.

THEOREM 2.11 (Hilbert). *Let $R = \bigoplus_{i=0}^{\infty} R_i$ be a homogeneous ring and let M be an $\mathbb{N}$-graded R-module with $d = \dim(M)$. If R_0 is Artinian, then there is a unique polynomial $P_M(t) \in \mathbb{Q}[t]$ of degree $d-1$ such that $P_M(i) = H(M, i)$ for $i \gg 0$.*

PROOF. See [23, Theorem 4.1.3]. □

DEFINITION 2.12. The polynomial $P_M(t)$ is called the *Hilbert polynomial* of M. If $P_M(t) = a_{d-1}t^{d-1} + \cdots + a_0$ is the Hilbert polynomial of M, the positive integer $e(M) := (d-1)!a_{d-1}$ is called the *multiplicity* of M.

2.3. *Cohen–Macaulay rings*

Let us introduce the class of Cohen–Macaulay rings and present some of its basic properties. Some useful characterizations of those rings will be given.

DEFINITION 2.13. Let M be an R-module. A sequence $\underline{\theta} = \theta_1, \ldots, \theta_n$ in R is *regular* on M if $(\underline{\theta})M \neq M$ and $\theta_i \notin \mathcal{Z}(M/(\theta_1, \ldots, \theta_{i-1})M)$ for all i.

PROPOSITION 2.14. *If $\underline{\theta}$ is a regular sequence of a ring R, then* $\mathrm{ht}(\underline{\theta}) = r$.

Let M be a module over a local ring $(R, \mathfrak{m})$. The *depth* of M, denoted by depth(M), is the length of a maximal regular sequence on M which is contained in $\mathfrak{m}$. The module M is said to be *Cohen–Macaulay* (C–M for short) if $\mathrm{depth}(M) = \dim(M)$. If $a \in \mathfrak{m}$ is a regular element of M, then

$$\dim M/aM = \dim M - 1 \quad \text{and} \quad \operatorname{depth} M/aM = \operatorname{depth} M - 1,$$

see [82, Chapter 3]. Thus M/aM is Cohen–Macaulay.

A ring R is Cohen–Macaulay if R is C–M as an R module. If R is a nonlocal ring, we say that R is Cohen–Macaulay if $R_{\mathfrak{p}}$ is a C–M local ring for all $\mathfrak{p} \in \mathrm{Spec}(R)$. An ideal I of R is a *Cohen–Macaulay ideal* if R/I is a Cohen–Macaulay R-module.

DEFINITION 2.15. Let $(R, \mathfrak{m})$ be a local ring of dimension d. A *system of parameters* (s.o.p) of R is a set $\underline{\theta} = \theta_1, \ldots, \theta_d$ with $\mathrm{rad}(\underline{\theta}) = \mathfrak{m}$. If R is a standard algebra, then $\underline{\theta}$ is a *homogeneous system of parameters* (h.s.o.p) of R if θ_i is homogeneous for all i and $\mathrm{rad}(\underline{\theta}) = \mathfrak{m}$.

NOTATION. Let I be an ideal of a ring R, $\nu(I)$, will denote the *minimum number of generators* of I.

The next result shows that systems of parameters always exist, see [5].

THEOREM 2.16. *If* $(R, \mathfrak{m})$ *is a local ring* (*or a standard algebra*), *then*

$$\dim(R) = \min\{\nu(I) \mid I \text{ is an ideal of } R \text{ with } \operatorname{rad}(I) = \mathfrak{m}\}.$$

PROPOSITION 2.17. *Let R be a local ring and let $\underline{\theta}$ be a s.o.p of R. Then R is a Cohen–Macaulay ring if and only if $\underline{\theta}$ is a regular sequence.*

PROPOSITION 2.18. *Let $M \neq 0$ be a Cohen–Macaulay R-module over a local ring R. Then* $\dim(R/\mathfrak{p}) = \operatorname{depth}(M)$ *for all* $\mathfrak{p} \in \operatorname{Ass} M$.

PROPOSITION 2.19. *If R is a Cohen–Macaulay local ring and I is a ideal of R, then* $\dim(R/I) = \dim(R) - \operatorname{ht}(I)$.

Minimal resolutions. Let R be a polynomial ring over a field with the standard grading and I a graded ideal of R.

Let $a \in \mathbb{N}$. The graded R-module obtained by a *shift in graduation* is defined as $R(-a) = \bigoplus_{i \geqslant 0} R(-a)_i$, where $R(-a)_i = R_{-a+i}$. By the *resolution* of R/I (see [44]), we mean its *minimal graded resolution* by free R-modules:

$$\begin{aligned} 0 \to \bigoplus_{i=1}^{b_g} R(-d_{gi}) &\xrightarrow{\varphi_g} \cdots \to \bigoplus_{i=1}^{b_k} R(-d_{ki}) \xrightarrow{\varphi_k} \cdots \\ \to \bigoplus_{i=1}^{b_1} R(-d_{1i}) &\xrightarrow{\varphi_1} R \to R/I \to 0, \end{aligned}$$

where $\operatorname{Im}(\varphi_k) \subset \mathfrak{m}R^{b_{k-1}}$ and $\mathfrak{m} = R_+$, that is, all the entries of the matrices $\varphi_1, \ldots, \varphi_g$ are in $\mathfrak{m}$. This condition is equivalent to require that at each stage we use a minimal generating set. Any two minimal resolutions are isomorphic as complexes. The integers $b_1, \ldots, b_g$ are the *Betti numbers* of R/I. The module $Z_k = \ker(\varphi_k)$ is called the *k-syzygy module* of R/I.

REMARK. (a) If K is a principal ideal domain, then all finitely generated projective $K[x_1, \ldots, x_n]$-modules are free, see [103, Theorem 4.63].

(b) If R/I has a minimal free resolution as above then, $\operatorname{pd}_R(R/I)$, the projective dimension of R/I, is equal to g.

The existence of a finite minimal free resolution as above is a consequence of the famous Hilbert's syzygy theorem:

THEOREM 2.20 [75]. *Let R be a polynomial ring over a field K. Let M be a finitely generated graded R-module. Then* $\operatorname{pd}_R(M) \leqslant \dim R$.

PROOF. See [44] or [53, Theorem 3.2]. □

THEOREM 2.21 (Auslander–Buchsbaum). *Let M be a finitely generated* (*graded*) *R-module. If R is a regular local ring* (*or a standard algebra*), *then*

$$\mathrm{pd}_R M + \mathrm{depth}\, M = \mathrm{depth}\, R.$$

PROOF. See [53, Theorem 3.1]. □

COROLLARY 2.22. *Let R be a polynomial ring and I a graded ideal, then* $\mathrm{pd}_R R/I \geqslant \mathrm{ht}\, I$, *with equality iff R/I is a Cohen–Macaulay R-module.*

DEFINITION 2.23. Let R be a polynomial ring over a field K. A graded ideal I of R is *Gorenstein* if R/I is a Cohen–Macaulay ring and the last Betti number in the minimal graded resolution of R/I is equal to 1.

3. Hilbert–Poincaré series

Hilbert–Poincaré series are introduced in this section. We use two gradings of the face ring of a simplicial complex to compute its Hilbert series in terms of the f-vector. The h-vector of a simplicial complex Δ will be introduced and studied. If Δ is Cohen–Macaulay its h-vector will satisfy some numerical constraints, see Lemma 3.10.

The computation of the Hilbert series of a graded ideal can be reduced to the computation of the Hilbert series of a monomial ideal using Gröbner basis and elimination of variables. The reader is referred to [10,12].

3.1. *Face rings of simplicial complexes*

A *simplicial complex* Δ consists of a finite set V of *vertices* and a collection Δ of subsets of V called *faces* such that

(i) If $v \in V$, then $\{v\} \in \Delta$.

(ii) If $F \in \Delta$ and $G \subset F$, then $G \in \Delta$.

Let Δ be a simplicial complex and let F be a face of Δ. Define the dimensions of F and Δ by $\dim F = |F| - 1$ and $\dim \Delta = \sup\{\dim F \mid F \in \Delta\}$ respectively. A face of dimension q is called a *q-face* or a *q-simplex*.

DEFINITION 3.1. If Δ has dimension d its *f-vector* is the $(d+1)$-tuple:

$$f(\Delta) = (f_0, \ldots, f_d),$$

where f_i is the number of i-faces of Δ. Write $f_{-1} = 1$.

Next we introduce a distinguished class of rings that play a fundamental role in algebraic combinatorics [113].

DEFINITION 3.2. Let Δ be a simplicial complex with vertices $x_1, \ldots, x_n$. If K is a field, the *Stanley–Reisner ring* or *face ring* $K[\Delta]$ is R/I_Δ, where

$$I_\Delta = \left(\left\{X_{i_1} \cdots X_{i_r} \mid i_1 < \cdots < i_r,\ \{x_{i_1}, \ldots, x_{i_r}\} \notin \Delta\right\}\right),$$

and $R = K[X_1, \ldots, X_n]$ is a polynomial ring.

A fine grading. Note that R can be endowed with a *fine* $\mathbb{Z}^n$*-grading* as follows. For $a = (a_1, \ldots, a_n) \in \mathbb{Z}^n$, set

$$R_a = \begin{cases} KX^a, & \text{if } a_i \geqslant 0 \ \text{ for } \ i = 1, \ldots, n, \\ 0, & \text{if } \ a_i < 0 \ \text{ for some } i, \end{cases}$$

where $X^a = X_1^{a_1} \cdots X_n^{a_n}$.

Let $I \subset R$ be an ideal generated by monomials. Since I is $\mathbb{Z}^n$-graded, the quotient ring R/I inherits the $\mathbb{Z}^n$-grading given by $(R/I)_a = R_a/I_a$ for all $a \in \mathbb{Z}^n$. In particular Stanley–Reisner rings have a *fine grading*.

Let M be a $\mathbb{Z}^n$-graded R-module. Each homogeneous component M_a of M is an R_0-module. Define the *Hilbert function* $H(M, a) = \ell(M_a)$, provided that the length $\ell(M_a)$ of M_a is finite for all a, and call

$$F(M, \mathbf{t}) = \sum_{a \in \mathbb{Z}^n} H(M, a)\mathbf{t}^a$$

the *Hilbert–Poincaré series* of M. Here $\mathbf{t} = (t_1, \ldots, t_n)$, where the t_i are indeterminates and $\mathbf{t}^a = t_1^{a_1} \cdots t_n^{a_n}$ for $a = (a_1, \ldots, a_n) \in \mathbb{Z}^n$.

By induction on n it follows that the polynomial ring $R = K[X_1, \ldots, X_n]$ with the fine grading has Hilbert–Poincaré series:

$$F(R, \mathbf{t}) = \sum_{a \in \mathbb{N}^n} \mathbf{t}^a = \prod_{i=1}^{n} \frac{1}{1 - t_i}.$$

The standard grading. On the other hand if R has the standard grading $\deg(X_i) = 1$ for all i and $M = \bigoplus_{i=0}^{\infty} M_i$ is an $\mathbb{N}$-graded module over R, the *Hilbert function* and *Hilbert–Poincaré series* of M are defined by

$$H(M, i) = \dim_K(M_i) \quad \text{and} \quad F(M, t) = \sum_{i=0}^{\infty} H(M, i)t^i$$

respectively.

3.2. *Hilbert series of face rings*

Let Δ be a simplicial complex and let K be a field, we denote by x_i the residue class of the indeterminate X_i in $K[\Delta]$. Thus $K[\Delta] = K[x_1, \dots, x_n]$. Define the *support* of $a \in \mathbb{Z}^n$ as $\operatorname{supp}(a) = \{x_i \mid a_i > 0\}$. If x^a is a nonzero monomial in $K[\Delta]$ we set $\operatorname{supp}(x^a) = \operatorname{supp}(a)$.

Hilbert series with the fine and standard gradings. Let $x^a \in K[\Delta]$ and let $\operatorname{supp}(a) = \{x_{i_1}, \dots, x_{i_m}\}$. Since I_Δ is generated by square free monomials we have

$$x^a \neq 0 \quad \Leftrightarrow \quad x_{i_1} \cdots x_{i_m} \neq 0 \quad \Leftrightarrow \quad x_{i_1} \cdots x_{i_m} \notin I_\Delta \quad \Leftrightarrow \quad \operatorname{supp}(a) \in \Delta.$$

Hence the nonzero monomials x^a form a K-basis of $K[\Delta]$. Therefore

$$F(K[\Delta], \mathbf{t}) = \sum_{\substack{a \in \mathbb{N}^n \\ \operatorname{supp}(a) \in \Delta}} \mathbf{t}^a = \sum_{F \in \Delta} \sum_{\substack{a \in \mathbb{N}^n \\ \operatorname{supp}(a) = F}} \mathbf{t}^a.$$

Let $F \in \Delta$. If $F = \emptyset$, then $\sum_{\operatorname{supp}(a)=F} \mathbf{t}^a = 1$, and if $F \neq \emptyset$, then

$$\sum_{\substack{a \in \mathbb{N}^n \\ \operatorname{supp}(a) \subset F}} \mathbf{t}^a = \prod_{x_i \in F} \frac{1}{1 - t_i} \quad \Rightarrow \quad \sum_{\substack{a \in \mathbb{N}^n \\ \operatorname{supp}(a) = F}} \mathbf{t}^a = \prod_{x_i \in F} \frac{t_i}{1 - t_i}.$$

Altogether we obtain that the expression for $F(K[\Delta], \mathbf{t})$ simplifies to

$$F(K[\Delta], \mathbf{t}) = \sum_{F \in \Delta} \prod_{x_i \in F} \frac{t_i}{1 - t_i}, \tag{1}$$

where the product over an empty index set is equal to 1. To compute the Hilbert series of $K[\Delta]$ as an $\mathbb{N}$-graded algebra note that for $i \in \mathbb{Z}$ we have

$$K[\Delta]_i = \bigoplus_{a \in \mathbb{Z}^n,\ |a| = i} K[\Delta]_a,$$

where $|a| = a_1 + \cdots + a_n$ for $a = (a_1, \dots, a_n)$. Observe that the Hilbert series of $K[\Delta]$ with the fine grading specializes to the Hilbert series of $K[\Delta]$ with the $\mathbb{Z}$-grading, that is, if $t_i = t$ for all i, then $F(K[\Delta], \mathbf{t}) = F(K[\Delta], t)$. Thus we have shown:

THEOREM 3.3. *The Hilbert–Poincaré series of* $K[\Delta]$ *is given by*

$$F(K[\Delta], t) = \sum_{i=-1}^{d} \frac{f_i t^{i+1}}{(1 - t)^{i+1}},$$

where $f = (f_0, \dots, f_d)$ *is the* f*-vector and* $d = \dim(\Delta) = \dim(K[\Delta]) - 1$.

COROLLARY 3.4. *If $R = K[X_1, \dots, X_n]$ is a polynomial ring, then*

$$F(R,t) = \frac{1}{(1-t)^n} \quad \textit{and} \quad H(R,m) = \binom{m+n-1}{n-1}.$$

PROPOSITION 3.5. *If* $\dim(\Delta) = d$, *then the Hilbert function of* $K[\Delta]$ *is*:

$$H\big(K[\Delta], m\big) = \sum_{i=0}^{d} \binom{m-1}{i} f_i \quad (\textit{for } m \geqslant 0). \tag{2}$$

REMARK. Recall that if $n \in \mathbb{R}$ and $k \in \mathbb{N}$ then

$$\binom{n}{k} = (-1)^k \binom{-n+k-1}{k}. \tag{3}$$

From (2) we derive that $H(K[\Delta], m)$ is a polynomial function of degree d with leading coefficient $d! f_d$, that is, the multiplicity of $K[\Delta]$ is equal to f_d.

Simplicial complexes and their h-vectors. We will write the Hilbert series of $S = K[\Delta]$ in two ways. By Theorem 3.3 one has:

$$F\big(K[\Delta], t\big) = \sum_{i=-1}^{d} \frac{f_i t^{i+1}}{(1-t)^{i+1}}. \tag{4}$$

On the other hand by the Hilbert–Serre theorem (cf. [5]) there is a (unique) polynomial $h(t) = h_0 + h_1 t + \cdots + h_r t^r \in \mathbb{Z}[t]$ so that $h(1) \neq 0$ and satisfying

$$F(S,t) = \frac{h(t)}{(1-t)^{d+1}}, \tag{5}$$

comparing (4) and (5) yields $h_k = 0$ for $k > d+1$ and

$$h_k = \sum_{i=0}^{k} (-1)^{k-i} \binom{d+1-i}{k-i} f_{i-1} \quad (0 \leqslant k \leqslant d+1). \tag{6}$$

DEFINITION 3.6. The *h-vector* of Δ is defined as

$$h(\Delta) = (h_0, \dots, h_r) \quad (r \leqslant d+1).$$

By making the substitution $t = z/(1+z)$ in the equality

$$\frac{h_0 + h_1 t + \cdots + h_{d+1} t^{d+1}}{(1-t)^{d+1}} = \sum_{i=0}^{d+1} \frac{f_{i-1} t^i}{(1-t)^i},$$

we obtain $\sum_{i=0}^{d+1} h_i z^i (1+z)^{d+1-i} = \sum_{i=0}^{d+1} f_{i-1} z^i$. Hence

$$f_{k-1} = \sum_{i=0}^{d+1} \binom{d+1-i}{k-i} h_i = \sum_{i=0}^{k} \binom{d+1-i}{k-i} h_i, \quad \text{for } 1 \leqslant k \leqslant d+1. \tag{7}$$

Recall that the *Euler characteristic* $\chi(\Delta)$ of Δ is defined as

$$\chi(\Delta) := \sum_{i=0}^{d} (-1)^i f_i,$$

and the *reduced Euler characteristic* $\tilde{\chi}(\Delta)$ of Δ is $\tilde{\chi}(\Delta) = \sum_{i=-1}^{d} (-1)^i f_i$. From Equations (6) and (7) we derive

$$\begin{aligned}
f_d &= \sum_{i=0}^{d+1} h_i, \\
h_1 &= f_0 - (d+1), \\
h_{d+1} &= (-1)^d \big(\chi(\Delta) - 1\big) = (-1)^d \tilde{\chi}(\Delta).
\end{aligned}$$

3.3. *The h-vector of a Cohen–Macaulay complex*

Let S be a standard algebra and let $\underline{\theta} = \{\theta_1, \dots, \theta_d\}$ be a subset of S of homogeneous elements. Recall that the set $\underline{\theta}$ is an *homogeneous system of parameters* (h.s.o.p for short) of S if $d = \dim(S)$ and $\operatorname{rad}(\underline{\theta}) = S_+$.

LEMMA 3.7. *Let $S = \bigoplus_{i \geqslant 0} S_i$ be a standard algebra over an infinite field K. If S is Cohen–Macaulay, then there exist a h.s.o.p $\underline{\theta} = \{\theta_1, \dots, \theta_d\}$ so that $\underline{\theta}$ is a regular sequence and $\theta_i \in S_1$ for all i.*

PROOF. Let $\operatorname{Ass}(S) = \{\mathfrak{p}_1, \dots, \mathfrak{p}_s\}$. Notice that $\operatorname{Ass}(S) = \operatorname{Min}(S)$, because S is Cohen–Macaulay. We may assume $d > 0$, otherwise there is nothing to prove. If $S_1 \subset \mathcal{Z}(S) = \bigcup_{i=1}^{s} \mathfrak{p}_i$, then $S_1 = \bigcup_{i=1}^{s} (\mathfrak{p}_i)_1$. Since K is infinite we obtain that $s = 1$ and $\mathfrak{p}_1 = S_+$, that is, $\dim S = 0$, which is a contradiction. Hence there is $\theta_1 \in S_1$ which is regular on S. The result readily follow by induction, since $S/(\theta_1)$ is Cohen–Macaulay of dimension $d-1$. $\square$

THEOREM 3.8 [112]. *Let S be a standard algebra and let $\theta_1, \dots, \theta_d$ be a h.s.o.p for S with $a_i = \deg(\theta_i)$. If A is the quotient ring $S/(\theta_1, \dots, \theta_d)$ with the natural grading, then S is Cohen–Macaulay if and only if*

$$F(S,t) = F(A,t) \Big/ \prod_{i=1}^{d} \big(1 - t^{a_i}\big). \tag{8}$$

DEFINITION 3.9. A simplicial complex Δ is said to be *Cohen–Macaulay* over a field K if the Stanley–Reisner ring $K[\Delta]$ is a Cohen–Macaulay ring.

Let us mention one of the application of Theorem 3.8 that plays a central role in Stanley's proof of the upper bound theorem for spheres.

LEMMA 3.10. *If Δ is a Cohen–Macaulay simplicial complex of dimension d on n vertices over an infinite field K, then the h-vector of Δ satisfies*

$$0 \leqslant h_i(\Delta) \leqslant \binom{i+n-d-2}{i} \quad (0 \leqslant i \leqslant d+1). \tag{9}$$

PROOF. Let $S = K[\Delta]$. By Lemma 3.7 there exists a regular h.s.o.p $\underline{\theta}$ for S so that each θ_i can be chosen of degree one, since $A = S/(\underline{\theta})S$ is Artinian, in this case Theorem 3.8 says that the h-vector of Δ is given by $h_i(\Delta) = H(A,i)$. Let $S = R/I$, where $R = K[X_1,\ldots,X_n]$ and $I = I_\Delta$. Note that $S/(\underline{\theta})S \cong \overline{R}/\bar{I}\,\overline{R}$, where $\overline{R} = R/(\underline{\theta})$ is a polynomial ring in $n-d-1$ variables and $\bar{I}$ is the image of I in $\overline{R}$. Therefore we have

$$h_i(\Delta) = H(A,i) = H(\overline{R}/\bar{I}\,\overline{R}, i) \leqslant H(\overline{R}, i) = \binom{i+n-d-2}{i},$$

as required. $\square$

4. Stanley–Reisner rings

The two main objects of study of this section are face rings and simplicial complexes. We emphasize the connection between algebraic properties of face rings and the reduced simplicial homology of Stanley–Reisner complexes. Some of the main references here are [23,78,113]. Our exposition in this section has been inspired by the detailed and beautiful presentation of the upper bound theorem and the local cohomology modules of Stanley–Reisner rings given in [23, Chapter 5]. A concern while writing this section was to include mainly the essential parts required to give a proof of both the upper bounds theorems and Reisner's theorem.

Primary decomposition. Let $R = K[X_1,\ldots,X_n]$ be a polynomial ring over a field K and let I be an ideal of R generated by square free monomials. The quotient ring $S = R/I$ will be called a *face ring*. The *Stanley–Reisner simplicial complex* Δ_I associated to I, has vertex set $V = \{x_1,\ldots,x_n\}$ and its faces are:

$$\Delta_I = \bigl\{\{x_{i_1},\ldots,x_{i_k}\} \mid i_1 < \cdots < i_k,\ X_{i_1}\cdots X_{i_k} \notin I\bigr\}.$$

A prime ideal $\mathfrak{p}$ of R is called a *face ideal* if $\mathfrak{p} = (X_{i_1},\ldots,X_{i_k})$ for some indeterminates $X_{i_1},\ldots,X_{i_k}$. An ideal I of R is called a *monomial ideal* if I is generated by a finite set of monomials, see [36,79,96,131] for details on monomial ideals and its importance in Gröbner bases theory.

PROPOSITION 4.1. *If I is a monomial ideal of R, then every associated prime ideal of I is a face ideal.*

COROLLARY 4.2. *If I is an ideal of R generated by square free monomials, then $I = \bigcap_{i=1}^{s} \mathfrak{p}_i$, where $\mathfrak{p}_1, \dots, \mathfrak{p}_s$ are the associated primes of I.*

PROOF. Let $I_1 = \bigcap_{i=1}^{s} \mathfrak{p}_i$. It suffices to prove $I_1 \subset I$. Take $f = X_{i_1}^{a_1} \cdots X_{i_r}^{a_r}$ in $I_1 = \mathrm{rad}(I)$, that is, $f^k \in I$ for some $k \geqslant 1$. Since I is generated by square free monomials we obtain $X_{i_1} \cdots X_{i_r} \in I$, hence $f \in I$. $\square$

COROLLARY 4.3. *If Δ is a simplicial complex, then*

$$I_\Delta = \bigcap_F P_F,$$

where the intersection is taken over all maximal faces F of Δ, and P_F denotes the face ideal generated by all X_i such that $x_i \notin F$.

COROLLARY 4.4. *A Cohen–Macaulay simplicial complex Δ is pure, that is, all the maximal faces have the same dimension*

PROOF. Use the previous corollary and Proposition 2.18. $\square$

4.1. *Reisner's criterion of Cohen–Macaulay complexes*

The next aim is to introduce the simplicial homology modules of a simplicial complex and the local cohomology modules of a standard algebra. A useful theorem of Hochster will be presented that relates these two notions (whenever the base ring is a field) via Hilbert series. Then as a consequence one obtains Reisner's characterization of Cohen–Macaulay complexes in terms of the vanishing of certain homology modules.

Simplicial homology. Let Δ be a simplicial complex and let F be a face of Δ. Define the dimensions of F and Δ by $\dim(F) = |F| - 1$ and $\dim(\Delta) = \sup\{\dim F \mid F \in \Delta\}$ respectively. A face of dimension q is sometimes refered to as a *q-face* or as a *q-simplex*. If $\dim(F) = \dim(\Delta)$, F is called a *facet*.

Let F be a face of Δ with vertices $v_0, v_1, \dots, v_q$. Two total orderings $v_{i_0} < \cdots < v_{i_q}$ and $v_{j_0} < \cdots < v_{j_q}$ of the vertices are *equivalent* if $(i_0, \dots, i_q)$ is an even permutation of $(j_0, \dots, j_q)$. This is an equivalence relation, and for $q > 1$ it partitions the total orderings of $v_0, \dots, v_q$ into exactly two equivalence classes. An *oriented q-simplex* of Δ is a q-simplex F with a choice of one of these equivalence classes. The oriented simplex determined by the ordering $v_0 < \cdots < v_q$ will be denoted by $[v_0, \dots, v_q]$.

Let A be a ring and let $C_q(\Delta)$ be the free A-module with basis consisting of the oriented q-simplices in Δ modulo the relations

$$[v_0, v_1, v_2, \dots, v_q] + [v_1, v_0, v_2, \dots, v_q].$$

In particular $C_q(\Delta)$ is defined for any field K and $\dim_K C_q(\Delta)$ is equal to the number of q-simplices of Δ. For $q \geqslant 1$ there is a homomorphism $\partial_q : C_q(\Delta) \to C_{q-1}(\Delta)$ induced by

$$\partial_q\big([v_0, v_1, \ldots, v_q]\big) = \sum_{i=0}^{q}(-1)^i[v_0, v_1, \ldots, v_{i-1}, \hat{v}_i, v_{i+1}, \ldots, v_q],$$

where $\hat{}$ over a symbol means that the symbol is deleted. Since $\partial_q\partial_{q+1} = 0$ we obtain the chain complex $C(\Delta) = \{C_q(\Delta), \partial_q\}$, which is called the *oriented chain complex* of Δ. The *augmented oriented chain complex* of Δ is:

$$0 \to C_d(\Delta) \xrightarrow{\partial_d} C_{d-1}(\Delta) \to \cdots \to C_0(\Delta) \xrightarrow{\varepsilon} C_{-1}(\Delta) = A \to 0,$$

where $d = \dim \Delta$ and $\varepsilon(v) = 1$ for every vertex v of Δ. This chain complex will be denoted by $(C_\star(\Delta), \varepsilon)$. Set $\partial_0 = \varepsilon$ and $C_{-1}(\Delta) = A$.

Let $Z_q(\Delta, A) = \ker(\partial_q)$, $B_q = \operatorname{im}(\partial_{q+1})$, and

$$\widetilde{H}_q(\Delta, A) = Z_q(\Delta, A)/B_q(\Delta, A), \quad \text{for } q \geqslant 0.$$

The elements of $Z_q(\Delta, A)$ and $B_q(\Delta, A)$ are called *cycles* and *boundaries* respectively and $\widetilde{H}_q(\Delta, A)$ is the qth *reduced homology group* of Δ with coefficients in A. Note that if $\Delta \neq \emptyset$, then $\widetilde{H}_i(\Delta; A) = 0$ for $i < 0$.

The *reduced Euler characteristic* $\tilde{\chi}(\Delta)$ of Δ is equal to

$$\tilde{\chi}(\Delta) = \sum_{i=-1}^{d}(-1)^i \widetilde{H}_i(\Delta; K) = \sum_{i=-1}^{d}(-1)^i f_i \quad (d = \dim \Delta).$$

Let $C_\star(\Delta)$ be the augmented chain complex of Δ over the ring A. The q-reduced singular cohomology group with coefficients in Δ is defined as

$$\widetilde{H}^q(\Delta; A) = \widetilde{H}^q\big(\operatorname{Hom}_A\big(C_\star(\Delta), A\big)\big).$$

If K is a field, there are canonical isomorphisms

$$\widetilde{H}^q(\Delta; K) \cong \operatorname{Hom}_K\big(\widetilde{H}_q(\Delta, K), K\big) \quad \text{and}$$
$$\widetilde{H}_q(\Delta; K) \cong \operatorname{Hom}_K\big(\widetilde{H}^q(\Delta, K), K\big).$$

Thus in particular we have $\widetilde{H}_q(\Delta; K) \cong \widetilde{H}^q(\Delta; K)$.

PROPOSITION 4.5. *If Δ is a nonempty simplicial complex with c connected components, then $\widetilde{H}_0(\Delta; A)$ is a free A-module of rank $c - 1$.*

Let Δ and Δ_1 be simplicial complexes with disjoint vertex sets V and W respectively. The *join* $\Delta * \Delta_1$ is the simplicial complex on the vertex set $V \cup W$ with faces $F \cup G$, where $F \in \Delta$ and $G \in \Delta_1$. The *cone* $\operatorname{cn}(\Delta) = w * \Delta$ of Δ is the join of a point $\{w\}$ with Δ.

PROPOSITION 4.6. *Let Δ be a simplicial complex and let* $\operatorname{cn}\Delta = w * \Delta$ *be the cone of* Δ. *Then* $\widetilde{H}_p(\operatorname{cn}\Delta) = 0$ *for all* p.

PROOF. See [131, Proposition 5.25]. □

Injective resolutions and Gorenstein rings. Let R be a ring and let I be an R-module. We say that I is *injective* if the functor $\operatorname{Hom}_R(\cdot, I)$ is exact. Note that this functor is always left exact.

DEFINITION 4.7. Let R be a ring and let M be an R-module. A complex

$$\mathcal{I}_\star : 0 \to I^0 \xrightarrow{\partial_0} I^1 \xrightarrow{\partial_1} I^2 \xrightarrow{\partial_2} \cdots$$

of injective R-modules is an *injective resolution* of M if $H_i(\mathcal{I}_\star) = 0$ for $i > 0$ and $H_0(\mathcal{I}_\star) = \ker(\partial_0) \cong M$.

The *injective dimension* of M, denoted $\operatorname{inj}\dim M$, is the smallest integer n for which there exist an injective resolution $\mathcal{I}_\star$ of M with $I^m = 0$ for $m > n$. If there is no such n, the injective dimension of M is infinite.

For the proofs of the next three results and for additional information on Gorenstein rings and injective resolutions see [8,23,80,103].

THEOREM 4.8. *Let* $(R, \mathfrak{m})$ *be a local ring and let* M *be an* R*-module of finite injective dimension. Then* $\dim M \leqslant \operatorname{inj}\dim M = \operatorname{depth} R$.

DEFINITION 4.9. A local ring R is *Gorenstein* if $\operatorname{inj}\dim R < \infty$. A ring R is *Gorenstein* if $R_\mathfrak{m}$ is a Gorenstein ring for any maximal ideal $\mathfrak{m}$ of R.

DEFINITION 4.10. Let R be a ring and let $N \subset M$ be R-modules. M is an essential extension of N if for any nonzero R-submodule U of M one has $U \cap N \neq 0$. An essential extension M of N is called proper if $N \neq M$.

PROPOSITION 4.11. *Let R be a ring. An R-module N is injective if and only if it has no proper essential extensions.*

DEFINITION 4.12. Let R be a ring and M an R-module. An injective module E such that $M \subset E$ is an essential extension is called an *injective hull* of M and is denoted by $E = E(M)$ or $E = E_R(M)$.

Let M be an R-module. Then M admits an injective hull and M has a *minimal injective* resolution $E_\star(M)$:

$$\begin{aligned}
&0 \to M \xrightarrow{\partial_{-1}} E_0(M) \xrightarrow{\partial_0} E_1(M) \to \cdots \\
&\quad \to E_{i-1}(M) \xrightarrow{\partial_{i-1}} E_i(M) \xrightarrow{\partial_i} E_{i+1}(M) \to \cdots,
\end{aligned}$$

where $E_0(M) = E(M)$ and $E_{i+1} = E(\operatorname{coker} \partial_{i-1})$. Here ∂_{-1} denotes the embedding $M \to E(M)$, and ∂_i is defined in a natural way. Any two minimal injective resolutions of M are isomorphic. If $\mathcal{I}_\star$ is an injective resolution of M, then $E_\star(M)$ is isomorphic to a direct summand of $\mathcal{I}_\star$.

DEFINITION 4.13. Let $(R, \mathfrak{m}, K)$ be a local ring R and let $M \neq 0$ be an R-module with depth s. The *type* of M is the number

$$r(M) = \dim_K \operatorname{Ext}_R^s(K, M).$$

THEOREM 4.14. *Let R be a local ring. Then R is a Gorenstein ring if and only if R is a Cohen–Macaulay ring of type* 1.

Local cohomology. Here we present a vanishing theorem of Grothendieck (Theorem 4.18) which is essential to prove the Reisner criterion for Cohen–Macaulay complexes. Our main references for homological algebra and local cohomology are [23,67,76,103].

Let $(R, \mathfrak{m})$ be a local ring and let M be an R-module. Denote by $\Gamma_\mathfrak{m}(M)$ the submodule of M of all the elements with support in $\{\mathfrak{m}\}$, that is,

$$\Gamma_\mathfrak{m}(M) = \left\{x \in M \mid \mathfrak{m}^k x = 0 \text{ for some } k > 0\right\}.$$

Let $\underline{x} = \{x_1, \dots, x_n\}$ be a sequence of elements in R generating an $\mathfrak{m}$-primary ideal. Set $\underline{x}^k = \{x_1^k, \dots, x_n^k\}$. The family $\underline{x}^k$ gives the $\mathfrak{m}$-adic topology on R, hence

$$\Gamma_\mathfrak{m}(M) = \left\{z \in M \mid (\underline{x})^k z = 0 \text{ for some } k \geqslant 0\right\}.$$

Since $\operatorname{Hom}_R(R/I, M) = \{x \in M \mid Ix = 0\}$ for any ideal I of R, we obtain a natural isomorphism

$$\Gamma_\mathfrak{m}(M) \cong \varinjlim \operatorname{Hom}_R\left(R/\mathfrak{m}^k, M\right) \cong \varinjlim \operatorname{Hom}_R\left(R/\left(\underline{x}^k\right), M\right). \tag{10}$$

PROPOSITION 4.15. *$\Gamma_\mathfrak{m}(\cdot)$ is a left exact additive functor.*

DEFINITION 4.16. The *local cohomology functors*, denoted by $H_\mathfrak{m}^i(\cdot)$ are the right derived functors of $\Gamma_\mathfrak{m}(\cdot)$.

REMARK. Let M and I be R-modules. (a) If $\mathcal{I}_\star$ is an injective resolution of M, then $H_\mathfrak{m}^i(M) = H^i(\Gamma_\mathfrak{m}(\mathcal{I}_\star))$ for $i \geqslant 0$, (b) $H_\mathfrak{m}^0(M) = \Gamma_\mathfrak{m}(M)$ and $H_\mathfrak{m}^i(M) = 0$ for $i < 0$. If I is injective, then $H_\mathfrak{m}^i(I) = 0$ for $i > 0$.

PROPOSITION 4.17. *If $(R, \mathfrak{m})$ is a local ring and M is an R-module, then*

$$H_\mathfrak{m}^i(M) \cong \varinjlim \operatorname{Ext}_R^i\left(R/\mathfrak{m}^k, M\right) \cong \varinjlim \operatorname{Ext}_R^i\left(R/\left(\underline{x}^k\right), M\right),$$

for $i \geqslant 0$, where $\underline{x}$ is a sequence in R generating an $\mathfrak{m}$-primary ideal.

PROOF. Recall that if $\mathcal{P}_\star$ is a projective resolution of L and $\mathcal{I}_\star$ is an injective resolution of M, then $\operatorname{Ext}_R^i(L, M)$ can be computed as follows:

$$\operatorname{Ext}_R^i(L, M) \cong H^i\big(\operatorname{Hom}_R(\mathcal{P}_\star, M)\big) \cong H^i\big(\operatorname{Hom}_R(L, \mathcal{I}_\star)\big),$$

see [76, Proposition 8.1]. Assume $\mathcal{I}_\star$ is an injective resolution of M, then

$$H_{\mathfrak{m}}^i(M) \cong H^i\big(\Gamma_{\mathfrak{m}}(\mathcal{I}_\star)\big) \quad \text{and} \quad \Gamma_{\mathfrak{m}}(\mathcal{I}_\star) \cong \varinjlim \operatorname{Hom}_R\big(R/\mathfrak{m}^k, \mathcal{I}_\star\big).$$

Therefore

$$\begin{aligned} H_{\mathfrak{m}}^i(M) &\cong H^i\big(\varinjlim \operatorname{Hom}_R\big(R/\mathfrak{m}^k, \mathcal{I}_\star\big)\big) \cong \varinjlim H^i\big(\operatorname{Hom}_R\big(R/\mathfrak{m}^k, \mathcal{I}_\star\big)\big) \\ &\cong \varinjlim \operatorname{Ext}_R^i\big(R/\mathfrak{m}^k, M\big). \end{aligned}$$

Since

$$\Gamma_{\mathfrak{m}}(M) \cong \varinjlim \operatorname{Hom}_R\big(R/\mathfrak{m}^k, M\big) \cong \varinjlim \operatorname{Hom}_R\big(R/\big(\underline{x}^k\big), M\big)$$

the second isomorphism follows using the same arguments. □

Next we recall the following vanishing theorem [67].

THEOREM 4.18 (Grothendieck). *If $(R, \mathfrak{m})$ is a local ring and M is an R-module of depth t and dimension d, then*

(a) $H_{\mathfrak{m}}^i(M) = 0$ *for $i < t$ and $i > d$.*
(b) $H_{\mathfrak{m}}^t(M) \neq 0$ *and* $H_{\mathfrak{m}}^d(M) \neq 0$.

Local cohomology of face rings. Our exposition below regarding local cohomology and the Hochster theorem follows closely that of [23] and the reader should consult this excellent reference for further details and results.

Let Δ be a simplicial complex and let

$$R = K[\Delta] = K[X_1, \ldots, X_n]/I_\Delta$$

be the Stanley–Reisner ring of Δ. Let $\mathfrak{m}$ be the maximal ideal generated by the residue classes x_i of the indeterminates X_i and let $H_{\mathfrak{m}}^i(R)$ be the local cohomology modules of R.

Consider the complex $C^\star$

$$\begin{aligned} &C^\star : 0 \to C^0 \to C^1 \to \cdots \to C^n \to 0, \\ &C^t = \bigoplus_{1 \leqslant i_1 < \cdots < i_t \leqslant n} R_{x_{i_1} \cdots x_{i_t}}, \end{aligned}$$

where R_y denotes R localized at $S = \{y^i\}_{i \geqslant 0}$ and the differentiation map $d_t : C^t \to C^{t+1}$ is given on the component

$$R_{x_{i_1} \cdots x_{i_t}} \xrightarrow{d_t} R_{x_{j_1} \cdots x_{j_{t+1}}}$$

to be the natural homomorphism

$$(-1)^{s-1} \cdot \eta : R_{x_{i_1} \cdots x_{i_t}} \to R_{(x_{i_1} \cdots x_{i_t}) x_{j_s}}$$

if $\{i_1, \ldots, i_t\} = \{j_1, \ldots, \hat{j}_s, \ldots, j_{t+1}\}$ and 0 otherwise. If $x = x_{i_1} \cdots x_{i_r}$ is in $K[\Delta]$, then $R_x \neq 0$ iff $\operatorname{supp} x \in \Delta$. Hence $H^i(C^\star) = 0$ for $i > \dim K[\Delta]$ (cf. Theorem 4.18). Recall that there is an isomorphism

$$H^i_{\mathfrak{m}}(R) \cong H^i(C^\star).$$

Note that $C^\star$ is a $\mathbb{Z}^n$-graded complex. Recall that R is itself $\mathbb{Z}^n$-graded. Let $a \in \mathbb{Z}^n$. Then $R_a \cong K$ if $a \in \mathbb{N}^n$ and $\operatorname{supp}(a) = \{x_i \mid a_i > 0\}$ is in Δ, and $R_a = 0$ otherwise. The components of C^t are of the form R_x for some $x \in R$ homogeneous in the fine grading. One defines a $\mathbb{Z}^n$-grading in R_x by:

$$(R_x)_a = \left\{ r/x^m \mid r \text{ is a monomial and } \deg(r) - m\deg(x) = a \right\}.$$

Here the term deg refers to the fine grading. Extending this grading to C^t it is clear that $C^\star$ becomes a $\mathbb{Z}^n$-graded complex. We may then equip $H^i(C^\star) \cong H^i_{\mathfrak{m}}(R)$ with the induced $\mathbb{Z}^n$-graded structure. Hence the *local cohomology modules $H^i_{\mathfrak{m}}(R)$ are $\mathbb{Z}^n$-graded modules*.

Consider $R = K[\Delta]$ as a graded K-algebra and give the module C^t the structure of $\mathbb{Z}$-graded R-module by setting

$$\left(C^t\right)_j = \bigoplus_{a \in \mathbb{Z}^n,\ |a| = j} \left(C^t\right)_a.$$

In this way $C^\star$ becomes a complex of graded $\mathbb{Z}$-modules and $H^i_{\mathfrak{m}}(R)$ inherits a $\mathbb{Z}$-graded structure. Note the isomorphisms $H^i_{\mathfrak{m}}(R) \cong H^i(C^\star)$ and

$$H^i_{\mathfrak{m}}(R)_j \cong \bigoplus_{a \in \mathbb{Z}^n,\ |a| = j} H^i(C^\star)_a.$$

Let $a \in \mathbb{Z}^n$. To determine when $(R_x)_a \neq 0$ set

$$G_a = \{x_i \mid a_i < 0\} \quad \text{and} \quad H_a = \{x_i \mid a_i > 0\}.$$

Notation. Set $F = \operatorname{supp}(x) = \{x_{i_1}, \ldots, x_{i_r}\}$, if $x = x_{i_1} \cdots x_{i_r}$ is square-free.

LEMMA 4.19. *$(R_x)_a \cong K$ if and only if $G_a \subset F$ and $F \cup H_a \in \Delta$.*

PROOF. See [23]. $\square$

Let Δ be a simplicial complex. For $F \in \Delta$ define the *link* of F as

$$\mathrm{lk}_\Delta(F) = \{H \in \Delta \mid H \cap F = \emptyset \text{ and } H \cup F \in \Delta\}.$$

The *star* of F is defined as $\mathrm{st}_\Delta(F) = \{H \in \Delta \mid H \cup F \in \Delta\}$. For simplicity sometimes we omit the index Δ in st_Δ or lk_Δ.

LEMMA 4.20. *If $a \in \mathbb{Z}^n$, then $H^p_{\mathfrak{m}}(K[\Delta])_a \cong \widetilde{H}^{p-|G_a|-1}(\mathrm{lk}_{\mathrm{st}\, H_a}\, G_a; K)$.*

PROOF. See [23]. $\square$

REMARK 4.21. If $a = 0$, then $H^p_{\mathfrak{m}}(K[\Delta])_0 = \widetilde{H}^{p-1}(\Delta; K)$.

EXAMPLE 4.22. Let $I = (x_1x_4, x_2x_4, x_3x_4)$ and $R = K[x_1, \ldots, x_4]/I$. In this case:

$$\begin{aligned} 0 \to R &\xrightarrow{d_0} R_{x_1} \oplus R_{x_2} \oplus R_{x_3} \oplus R_{x_4} \\ &\xrightarrow{d_1} R_{x_1x_2} \oplus R_{x_1x_3} \oplus R_{x_2x_3} \xrightarrow{d_2} R_{x_1x_2x_3} \to 0, \end{aligned}$$

where

$$\begin{aligned} &d_0(z) = (z, z, z, z), \\ &d_1(z_1, z_2, z_3, z_4) = (z_2 - z_1, z_3 - z_1, z_3 - z_2), \\ &d_2(w_1, w_2, w_3) = w_1 - w_2 + w_3. \end{aligned}$$

THEOREM 4.23 (Hochster). *Let Δ be a simplicial complex and let K be a field. Then the Hilbert series of the local cohomology modules of $K[\Delta]$ with respect to the fine grading is given by*

$$F\big(H^p_{\mathfrak{m}}(K[\Delta]), \mathbf{t}\big) = \sum_{F \in \Delta} \dim_K \widetilde{H}_{p-|F|-1}(\mathrm{lk}\, F; K) \prod_{x_j \in F} \frac{t_j^{-1}}{1 - t_j^{-1}}.$$

PROOF. By Lemma 4.20 and the comment before Proposition 4.5 we have

$$\begin{aligned} \dim_K H^p_{\mathfrak{m}}(K[\Delta]) &= \dim_K \widetilde{H}^{p-|G_a|-1}(\mathrm{lk}_{\mathrm{st}\, H_a}\, G_a; K) \\ &= \dim_K \widetilde{H}_{p-|G_a|-1}(\mathrm{lk}_{\mathrm{st}\, H_a}\, G_a; K). \end{aligned}$$

Let $\mathbb{Z}^n_- = \{a \in \mathbb{Z}^n \mid a_i \leqslant 0 \text{ for all } i\}$. If $H_a \neq \emptyset$, then $\mathrm{lk}_{\mathrm{st}\, H_a}\, G_a$ is acyclic by [23, Lemma 5.3.5] . If $H_a = \emptyset$, then $\mathrm{lk}_{\mathrm{st}\, H_a}\, G_a = \mathrm{lk}\, G_a$. Therefore:

$$F\big(H^p_{\mathfrak{m}}(K[\Delta]), \mathbf{t}\big) = \sum_{a \in \mathbb{Z}^n} \dim_K \widetilde{H}_{p-|G_a|-1}(\mathrm{lk}_{\mathrm{st}\, H_a}\, G_a; K)\mathbf{t}^a$$

$$= \sum_{a\in\mathbb{Z}^n_-} \dim_K \widetilde{H}_{p-|G_a|-1}(\operatorname{lk} G_a; K)\mathbf{t}^a$$
$$= \sum_{F\in\Delta} \sum_{a\in\mathbb{Z}^n_-,\, G_a=F} \dim_K \widetilde{H}_{p-|F|-1}(\operatorname{lk} F; K)\mathbf{t}^a$$
$$= \sum_{F\in\Delta} \sum_{a\in\mathbb{N}^n,\, \operatorname{supp}(a)=F} \dim_K \widetilde{H}_{p-|F|-1}(\operatorname{lk} F; K)\left(\mathbf{t}^{-1}\right)^a$$
$$= \sum_{F\in\Delta}\Bigg(\dim_K \widetilde{H}_{p-|F|-1}(\operatorname{lk} F; K) \sum_{\substack{a\in\mathbb{N}^n\\ \operatorname{supp}(a)=F}} \left(\mathbf{t}^{-1}\right)^a\Bigg).$$

Since

$$\sum_{a\in\mathbb{N}^n,\ \operatorname{supp}(a)=F} \mathbf{t}^{-a} = \prod_{x_j\in F} \frac{t_j^{-1}}{1-t_j^{-1}}$$

we obtain the desired formula. This proof was adapted from [23]. □

DEFINITION 4.24. A simplicial complex Δ is *Cohen–Macaulay* over a field K (C–M for short) if $K[\Delta]$ is a Cohen–Macaulay ring.

As a consequence of Hochster's theorem and the Grothendieck vanishing theorem one obtains the following Cohen–Macaulay criterion [97].

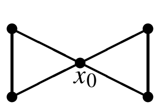
COROLLARY 4.25 (Reisner). *Let Δ be a simplicial complex. If K is a field, then Δ is Cohen–Macaulay over K if and only if $\widetilde{H}_i(\operatorname{lk}(F); K) = 0$ for all $F \in \Delta$ and all $i < \dim \operatorname{lk}(F)$.*

PROOF. Let Δ be a simplicial complex of dimension d. By Theorem 4.18 we have that Δ is C–M if and only if $H^i(C^\star) = 0$ for $i < d+1$.

($\Rightarrow$) If Δ is C–M, then by Hochster theorem $\widetilde{H}_{i-|F|-1}(\operatorname{lk}(F); K) = 0$ for $F \in \Delta$ and $i < d+1$. By Corollary 4.4 we get that Δ is pure. If $F \in \Delta$, there is a face F_1 of dimension d containing F, since $F_1 \setminus F \in \operatorname{lk}(F)$ it follows that $\dim \operatorname{lk}(F) = |F_1 \setminus F| - 1 = d - |F|$. Hence $\widetilde{H}_i(\operatorname{lk}(F); K) = 0$ for all $F \in \Delta$ and all $i < \dim \operatorname{lk}(F)$.

($\Leftarrow$) Using $\dim \operatorname{lk}(F) \leqslant d - |F|$, the hypothesis $\widetilde{H}_i(\operatorname{lk}(F); K) = 0$ for $F \in \Delta$ and $i < \dim \operatorname{lk}(F)$ implies $\widetilde{H}_{i-|F|-1}(\operatorname{lk}(F); K) = 0$ for $F \in \Delta$ and $i < d+1$. Hence by Hochster theorem $H^i(C^\star) = 0$ for $i < d+1$. Thus Δ is Cohen–Macaulay. □

EXAMPLE 4.26. Let Δ be the 2-dimensional simplicial complex:



Then $\operatorname{lk}\{x_0\}$ consist of two disjoint lines, hence $\operatorname{rank}(\widetilde{H}_0(\operatorname{lk}\{x_0\})) = 1$ and Δ is not Cohen–Macaulay.

Simplicial spheres. Next we give more applications of Hochster's theorem and Reisner's theorem.

Let Δ be a simplicial complex on the vertex set $V = \{x_1, \ldots, x_n\}$ and let $\{e_1, \ldots, e_n\}$ be the canonical basis in $\mathbb{R}^n$. Given a face $F \in \Delta$ define

$$|F| = \operatorname{conv}\{e_i \mid x_i \in F\},$$

where "conv" is the convex hull. Define the *geometric realization* of Δ as

$$|\Delta| = \bigcup_{F\in\Delta} |F|.$$

Then Δ is a topological space with the usual induced topology of $\mathbb{R}^n$. Note that there is a canonical isomorphism $\widetilde{H}_i(\Delta, A) \cong \widetilde{H}_i(|\Delta|; A)$ for all i.

DEFINITION 4.27. A simplicial complex Δ of dimension d is a *simplicial sphere* if $|\Delta| \cong S^d$. In this case Δ is called a triangulation of S^d.

THEOREM 4.28. *If Δ is a simplicial sphere of dimension d, then*

$$\widetilde{H}_i(\operatorname{lk} F; K) \cong \begin{cases} K & \text{for } i = \dim \operatorname{lk} F, \\ 0 & \text{otherwise.} \end{cases}$$

PROOF. Set $X = |\Delta| \cong S^d$. Let F be a face in Δ of dimension j, and let $p \in |F|^o$ (= relative interior of $|F|$). From [93, Lemma 63.1] we have

$$H_i\big(X, X \setminus \{p\}; K\big) \cong \begin{cases} \widetilde{H}_{i-j-1}(\operatorname{lk} F; K) & \text{if } \operatorname{lk} F \neq \emptyset,\ \forall\, i, \\ K & \text{if } \operatorname{lk} F = \emptyset \text{ and } i = j, \\ 0 & \text{if } \operatorname{lk} F = \emptyset \text{ and } i \neq j. \end{cases}$$

On the other hand using that $X \cong S^d$ one has that $\forall p \in X$:

$$\widetilde{H}_i(X; K) \cong \begin{cases} H_i(X, X \setminus \{p\}; K) = 0 & \text{if } i < \dim X = d, \\ H_i(X, X \setminus \{p\}; K) \cong K & \text{if } i = \dim X = d. \end{cases}$$

Hence Δ is pure and $d - j - 1 = \dim \operatorname{lk} F$. The required formula follow readily. □

PROPOSITION 4.29. *If Δ is a simplicial sphere and K is any field, then $K[\Delta]$ is a Cohen–Macaulay ring and $\tilde{\chi}(\operatorname{lk} F) = (-1)^{\dim \operatorname{lk} F}$ for all $F \in \Delta$.*

PROOF. Use Reisner's theorem and Theorem 4.28. See also [23]. □

LEMMA 4.30. *Let Δ be a simplicial complex on $V = \{x_1, \ldots, x_n\}$. Then*

$$F\big(K[\Delta], t_1^{-1}, \ldots, t_n^{-1}\big) = \sum_{F\in\Delta} (-1)^{\dim F} \tilde{\chi}(\operatorname{lk} F) \prod_{x_i \in F} \frac{t_i}{1 - t_i}.$$

PROOF. Recall that $F(K[\Delta],\mathbf{t})=\sum_{F\in\Delta}\prod_{x_i\in F}\frac{t_i}{1-t_i}$. Hence

$$\begin{aligned}F\big(K[\Delta],t_1^{-1},\dots,t_n^{-1}\big)&=\sum_{F\in\Delta}(-1)^{\dim F+1}\prod_{x_i\in F}\left(1+\frac{t_i}{1-t_i}\right)\\&=\sum_{F\in\Delta}(-1)^{\dim F+1}\sum_{G\subset F}\prod_{x_i\in G}\frac{t_i}{1-t_i}\\&=\sum_{G\in\Delta}\Bigg(\sum_{\substack{F\in\Delta\\ G\subset F}}(-1)^{\dim F+1}\Bigg)\prod_{x_i\in G}\frac{t_i}{1-t_i}.\end{aligned}$$

Let $\mathcal{F}=\{F\in\Delta\mid G\subset F\}$, then the map $F\mapsto F'=F\setminus G$ establish a bijection between $\mathcal{F}$ and $\operatorname{lk}G$. Since $(-1)^{|F|}=(-1)^{|F'|-|G|}$ we have

$$\sum_{\substack{F\in\Delta\\ G\subset F}}(-1)^{\dim F+1}=\sum_{F\in\operatorname{lk}G}(-1)^{\dim F-\dim G}=(-1)^{\dim G}\tilde{\chi}(\operatorname{lk}G).$$

From these identities the lemma follows. This proof was adapted from the book of Bruns and Herzog [23]. □

COROLLARY 4.31. *If Δ is a simplicial sphere, then $K[\Delta]$ is Gorenstein.*

PROOF. It follows from [113, Theorem 5.1] and Theorem 4.28. □

THEOREM 4.32. *Let Δ be a simplicial complex of dimension d. If Δ is a simplicial sphere, then its h-vector satisfies $h_i=h_{d+1-i}$ for all i.*

PROOF. It follows from Lemma 4.30, see [23] for details. This theorem can also be shown using that $K[\Delta]$ is Gorenstein, because the h-vector of a standard Gorenstein K-algebra is symmetric [131, Corollary 4.3.10]. □

Generators of least degree in Gorenstein ideals. First we present a result on the number of generators of a graded Gorenstein ideal. Recall that an ideal I of a ring R is Gorenstein if R/I is a Gorenstein ring.

THEOREM 4.33 [90]. *If I is a graded Gorenstein ideal of height $g\geqslant 3$ and initial degree $p\geqslant 2$, then*

$$\nu(I_p)\leqslant\nu_0=\binom{p+g-1}{g-1}-\binom{p+g-3}{g-1}.$$

If R/I is a Gorenstein ring, it is an interesting problem to find optimal bounds for the number of generators of least degree that occur as generators of the syzygy modules of I.

There is computational evidence that supports the following conjecture of M. Miller and R. Villarreal:

CONJECTURE 4.34 [131]. *Let Δ be a simplicial sphere and let $I = I_\Delta$ be the Stanley–Reisner ideal of Δ. If I has initial degree p and height g, then*

$$\nu(I) \leqslant \binom{p+g-1}{g-1} - \binom{p+g-3}{g-1}.$$

4.2. *The upper bound conjectures*

A *polytope* $P \subset \mathbb{R}^s$ is the convex hull of a finite set of points $\mathcal{A} = \{v_1, \dots, v_q\}$ in $\mathbb{R}^s$, that is, P is the set of all *convex combinations*:

$$P = \operatorname{conv}(\mathcal{A}) = \{a_1 v_1 + \cdots + a_q v_q \mid a_i \geqslant 0,\ a_1 + \cdots + a_q = 1,\ a_i \in \mathbb{R}\}.$$

Some standard references for polytopes are [18,69,89,132,133].

The set $\mathcal{A}$ is called *affinely independent* if a relation

$$a_1 v_1 + \cdots + a_q v_q = 0$$

with $a_1 + \cdots + a_q = 0$ and $a_i \in \mathbb{R}$ can only hold if $a_i = 0$ for all i.

Let $A \subset \mathbb{R}^s$. Recall that $\operatorname{aff}(A)$, the *affine space generated* by A, is the set of all *affine combinations* of points in A:

$$\operatorname{aff}(A) = \{a_1 p_1 + \cdots + a_r p_r \mid p_i \in A,\ a_1 + \cdots + a_r = 1,\ a_i \in \mathbb{R}\}.$$

The affine space $\operatorname{aff}(A)$ can be represented as $\operatorname{aff}(A) = x_0 + V$, where $x_0 \in \mathbb{R}^s$ and V is a (unique) linear subspace of $\mathbb{R}^s$. The *dimension* of A is defined as $\dim A = \dim_{\mathbb{R}} V$.

Let $y \in \mathbb{R}^s \setminus \{0\}$ and $a \in \mathbb{R}$. Define the hyperplane

$$H(y, a) = \{x \in \mathbb{R}^s \mid \langle x, y\rangle = a\},$$

where $\langle\, ,\, \rangle$ is the usual inner product in $\mathbb{R}^s$. The two *closed halfspaces* bounded by $H(y, a)$ are

$$H^+(y, a) = \{x \in \mathbb{R}^s \mid \langle x, y\rangle \geqslant a\} \quad \text{and} \quad H^-(y, a) = H^+(-y, -a).$$

Let P be a polytope. A *proper face* of P is a subset $F \subset P$ such that there is a *supporting hyperplane* $H(y, a)$ satisfying $F = P \cap H(y, a) \neq \emptyset$, $P \not\subset H(y, a)$ and $P \subset H^+(y, a)$. The *improper faces* of P are P itself and the empty face. The faces of dimension zero are called the *vertices* of P. Note that P is in fact the convex hull of its vertices according to the Krein–Milman theorem [132].

PROPOSITION 4.35. *Let $P \subset \mathbb{R}^s$ be a polytope. Then P is a compact convex subset of $\mathbb{R}^s$ with finitely many faces and any face of P is a polytope.*

Let P be a polytope of dimension $d+1$ and let f_i be the number of faces of dimension i of P. The *f-vector* of P is the vector $f(P)=(f_0,\ldots,f_d)$. In 1893 Poincaré proved the *Euler characteristic formula*

$$\sum_{i=0}^{d}(-1)^i f_i = 1+(-1)^d, \tag{11}$$

see [18, Theorem 16.1]. In [114] there are some historical comments about this formula. It follows from Proposition 4.29 that this formula also holds for simplicial spheres. Are there any optimal bounds for the entries of $f(P)$? As it will be seen there is a positive answer to this question which is valid in a more general setting.

In order to formulate the upper bound theorem for simplicial spheres and the upper bound theorem for convex polytopes we need to introduce some results on *cyclic polytopes*, the reader is referred to [18] for a detailed discussion on this topic.

Consider the *monomial curve* $\Gamma \subset \mathbb{R}^{d+1}$ given parametrically by

$$\Gamma = \bigl\{\bigl(\tau,\tau^2,\ldots,\tau^{d+1}\bigr) \mid \tau\in\mathbb{R}\bigr\}.$$

A *cyclic polytope*, denoted by $C(n,d+1)$, is the convex hull of any n distinct points in Γ such that $n>d+1$. The f-vector of $C(n,d+1)$ depends only on n and d and not on the points chosen, and $\dim C(n,d+1)=d+1$.

DEFINITION 4.36. A *q-simplex* is a polytope generated by a set of $q+1$ affinely independent points. A polytope is *simplicial* if everyone of its proper faces is a simplex.

The cyclic polytope $C(n,d+1)$ is simplicial and has the remarkable property that its f-vector satisfies:

$$f_i\bigl(C(n,d+1)\bigr)=\binom{n}{i+1} \quad \text{for } 0\leqslant i<\left\lfloor\frac{d+1}{2}\right\rfloor, \tag{12}$$

this means that $C(n,d+1)$ has the highest possible number of i-faces when i is within the specified rank. Hence for any $(d+1)$-polytope P we have

$$f_i(P)\leqslant\binom{n}{i+1}=f_i\bigl(C(n,d+1)\bigr),$$

for all $0\leqslant i<\lfloor(d+1)/2\rfloor$. The following upper bound conjecture for convex polytopes was posed by Motzkin [92] in 1957 and proved by P. McMullen [88] in 1970. It was motivated by the performance of the simplex algorithm in linear programming.

UBCP CONJECTURE. *Let P be a convex polytope of dimension $d+1$ with n vertices, then $f_i(P)\leqslant f_i(C(n,d+1))$ for $0\leqslant i\leqslant d$.*

Note that a simplicial complex is a natural generalization of a simplicial polytope. The f-vector of a simplicial complex Δ of dimension d is defined as the $(d+1)$-tuple

$$f(\Delta) = \big(f_0(\Delta), \dots, f_d(\Delta)\big),$$

where $f_i(\Delta)$ is the number of i-simplices in Δ. In particular $f_{-1} = 1$.

The *boundary complex* $\Delta(P)$ of a simplicial polytope of dimension $d+1$ is the abstract simplicial complex whose vertices are the vertices of P and whose faces are those sets of vertices that span a proper face of P, note $\dim \Delta(P) = d$. Thus the geometric realization $|\Delta(P)|$ is homeomorphic to a d-sphere S^d. Notice that P and $\Delta(P)$ have the same f-vector, and we define the *h-vector* of P as the h-vector of $\Delta(P)$.

REMARK 4.37. There are simplicial d-spheres with $d = 3$ and $n = 8$ which are not the boundary complex of a simplicial $(d+1)$-polytope [68]. Moreover it follows from a result of Steinitz (a graph is the 1-skeleton of a 3-polytope in $\mathbb{R}^3$ if and only if it is a 3-connected planar graph) that such an example does not exist for $d = 2$, see [133, Theorem 4.1].

THEOREM 4.38 [88]. *The h-vector $(h_0, \dots, h_{d+1})$ of a simplicial polytope of dimension $d+1$ satisfies the Dehn–Sommerville equations $h_i = h_{d+1-i}$ for $0 \leqslant i \leqslant d+1$.*

PROOF. It follows from Theorem 4.32. □

PROPOSITION 4.39. *If $h = (h_0, \dots, h_{d+1})$ is the h-vector of a cyclic polytope $C(n, d+1)$, then*

$$h_k = \binom{n-d+k-2}{k}, \quad 0 \leqslant k \leqslant \lfloor (d+1)/2 \rfloor.$$

PROOF. Let $0 \leqslant k \leqslant \lfloor (d+1)/2 \rfloor$. One has

$$\begin{aligned} h_k &= \sum_{i=0}^{k} (-1)^{k-i} \binom{d+1-i}{k-i} f_{i-1} = \sum_{i=0}^{k} (-1)^{2(k-i)} \binom{k-d-2}{k-i} \binom{n}{i} \\ &= \binom{n-d+k-2}{k}. \end{aligned}$$

The second equality follows from Equation (3). □

Since a cyclic polytope $C(n, d+1)$ is simplicial of dimension $d+1$ its boundary complex $\Delta(C(n, d+1))$ is a d-sphere. In 1964 V. Klee pointed out that the upper bound conjecture for polytopes may as well be made for simplicial spheres, this conjecture was proved by R. Stanley [111] in 1975 using techniques from commutative and homological algebra. Next we prove the *upper bound theorem for simplicial spheres*.

THEOREM 4.40 [111]. *Let Δ be simplicial complex of dimension d with n vertices. If $|\Delta| \cong S^d$, then $f_i(\Delta) \leqslant f_i(\Delta(C(n, d+1)))$ for $i = 0, \dots, d$.*

PROOF. Consider the following four conditions:

(a) $h_i(\Delta) \leqslant h_i(\Delta(C(n,d+1)))$ for $0 \leqslant i \leqslant \lfloor (d+1)/2 \rfloor$.

(b) $h_i(\Delta(C(n,d+1))) = h_{d+1-i}(\Delta(C(n,d+1)))$ for $0 \leqslant i \leqslant d+1$.

(c) $h_i(\Delta) \leqslant \binom{i+n-d-2}{i}$ all i.

(d) $h_i(\Delta) = h_{d+1-i}(\Delta)$ for $0 \leqslant i \leqslant d+1$.

We claim that (a), (b) and (d) imply the theorem. To show it notice that if $\lfloor (d+1)/2 \rfloor < i \leqslant d+1$, then $0 \leqslant d+1-i \leqslant \lfloor (d+1)/2 \rfloor$ and

$$h_i(\Delta) = h_{d+1-i}(\Delta) \leqslant h_{d+1-i}\big(\Delta\big(C(n,d+1)\big)\big) = h_i\big(\Delta\big(C(n,d+1)\big)\big),$$

hence $h_i(\Delta) \leqslant h_i(\Delta(C(n,d+1)))$ for $0 \leqslant i \leqslant d+1$. Therefore

$$\begin{aligned} f_{k-1}(\Delta) &= \sum_{i=0}^{d+1} \binom{d+1-i}{k-i} h_i(\Delta) \\ &\leqslant \sum_{i=0}^{d+1} \binom{d+1-i}{k-i} h_i\big(\Delta\big(C(n,d+1)\big)\big) \\ &= f_{k-1}\big(\Delta\big(C(n,d+1)\big)\big) \quad \text{for } 1 \leqslant k \leqslant d+1, \end{aligned}$$

and the proof of the claim is completed.

By Proposition 4.39 (c) $\Rightarrow$ (a), and by Lemma 4.38 (b) is satisfied due to the fact that $C(n,d+1)$ is a simplicial polytope. Hence the theorem is reduced to prove (c) and (d). To complete the proof notice that (c) follows from Lemma 3.10, and (d) follow from Theorem 4.32. $\square$

As a particular case of the upper bound theorem for spheres we obtain the *upper bound theorem for convex polytopes*:

THEOREM 4.41 [88]. *Let P be a convex polytope of dimension d with n vertices, then $f_i(P) \leqslant f_i(C(n,d))$ for $0 \leqslant i \leqslant d-1$.*

PROOF. By pulling the vertices P can be transformed into a simplicial polytope with the same number of vertices as P and at least as many faces of higher dimension, see [89] and [133, Lemma 8.24]. Hence we may assume that P is simplicial. Since P is simplicial the boundary complex $\Delta(P)$ is a simplicial sphere, that is, $|\Delta(P)| \cong S^{d-1}$. Using Theorem 4.40 we obtain

$$f_i(P) = f_i\big(\Delta(P)\big) \leqslant f_i\big(\Delta\big(C(n,d)\big)\big) = f_i\big(C(n,d)\big),$$

for $i = 0, \ldots, d$. $\square$

THEOREM 4.42. *If P is a simplicial d-polytope with n vertices, then*

$$f_j(P) \leqslant \varphi_{d-j-1}(d,n) \quad (j = 1, \ldots, d-1),$$

where

$$\varphi_{d-j-1}(d,n) = \sum_{i=0}^{\lfloor d/2 \rfloor} \binom{i}{d-j-1}\binom{n-d+i-1}{i} + \sum_{i=0}^{\lfloor (d-1)/2 \rfloor} \binom{d-i}{d-j-1}\binom{n-d+i-1}{i}.$$

PROOF. See [18, Corollary 18.3]. □

5. Monomial subrings

In this section we will introduce monomial rings (respectively semigroups) using integral matrices. We will describe the integral closure of those rings (respectively semigroups) and study their normality in some cases of interest arising from unimodular matrices. Then we introduce toric ideals and study systems of binomials. The Ehrhart ring will be compared with some other monomial subrings. The important notion of elementary vector and its relation to normality will be examined.

5.1. *Elementary integral vectors or circuits*

The notion of elementary integral vector occurs in convex analysis [102] and in the theory of toric ideals of graphs [117,127].

If $\alpha \in \mathbb{R}^q$, its *support* is defined as

$$\mathrm{supp}(\alpha) = \{i \mid \alpha_i \neq 0\}.$$

Note that $\alpha = \alpha_+ - \alpha_-$, where α_+ and α_- are two nonnegative vectors with disjoint support.

DEFINITION 5.1. Let N be a linear subspace of $\mathbb{Q}^q$. An *elementary vector* of N is a nonzero vector α in N whose support is minimal with respect to inclusion, i.e., $\mathrm{supp}(\alpha)$ does not properly contain the support of any other nonzero vector in N.

The concept of an elementary vector arises in graph theory when N is the kernel of the incidence matrix of a graph G. The reader is referred to [102, Section 22] for a precise interpretation of the vectors in N as flows of G which are conservative at every vertex.

LEMMA 5.2. *If N is a linear subspace of $\mathbb{Q}^q$ and α, β are two elementary vectors of N with the same support, then $\alpha = \lambda\beta$ for some $\lambda \in \mathbb{Q}$.*

PROOF. If $i \in \mathrm{supp}(\alpha)$, then one can write $\alpha_i = \lambda\beta_i$ for some scalar λ. Since $\mathrm{supp}(\alpha - \lambda\beta) \subsetneq \mathrm{supp}(\alpha)$, one concludes $\alpha - \lambda\beta = 0$, as required. □

DEFINITION 5.3. Let N be a linear subspace of $\mathbb{Q}^q$. An *elementary integral vector* or *circuit* of N is an elementary vector of N with relatively prime integral entries.

COROLLARY 5.4. *If N is a linear subspace of $\mathbb{Q}^q$, then the number of circuits of N is finite.*

PROOF. It follows from Lemma 5.2. □

DEFINITION 5.5. Two vectors $\alpha = (\alpha_i)$ and $\beta = (\beta_i)$ in $\mathbb{Q}^q$ are in *harmony* if $\alpha_i \beta_i \geqslant 0$ for every i.

LEMMA 5.6 [101]. *Let N be a linear subspace of $\mathbb{Q}^q$. If $0 \neq \alpha \in N$, then there is an elementary vector $\gamma \in N$ in harmony with α such that* $\mathrm{supp}(\gamma) \subset \mathrm{supp}(\alpha)$.

THEOREM 5.7 [101]. *If N is a vector subspace of $\mathbb{Q}^q$ and $\alpha \in N \setminus \{0\}$, then α can be written as*

$$\alpha = \sum_{i=1}^{r} \beta_i$$

for some elementary vectors $\beta_1, \ldots, \beta_r$ of N with $r \leqslant \dim N$ such that
(i) $\beta_1, \ldots, \beta_r$ *are in harmony with α,*
(ii) $\mathrm{supp}(\beta_i) \subset \mathrm{supp}(\alpha)$ *for all i, and*
(iii) $\mathrm{supp}(\beta_i)$ *is not contained in the union of the supports of $\beta_1, \ldots, \beta_{i-1}$ for all $i \geqslant 2$.*

PROOF. It follows from Lemma 5.6 by induction on the number of elements in the support of α. □

COROLLARY 5.8. *If N is a linear subspace of $\mathbb{Q}^q$, then the circuits of N generate N as a $\mathbb{Q}$-vector space.*

PROPOSITION 5.9 [46]. *Let A be an $n \times q$ integral matrix and let α be a nonzero vector in* $\ker(A)$. *If A has rank n, then α is an elementary vector of* $\ker(A)$ *if and only if there is $0 \neq \lambda \in \mathbb{Q}$ such that*

$$\alpha = \lambda \sum_{k=1}^{n+1} (-1)^k \det[v_{i_1}, \ldots, v_{i_{k-1}}, v_{i_{k+1}}, \ldots, v_{i_{n+1}}] e_{i_k}, \qquad (*)$$

for some column vectors v_{i_j} of A. Here e_k is the kth unit vector in $\mathbb{R}^q$.

PROOF. ($\Leftarrow$) Let A' be the submatrix of A consisting of the column vectors $v_{i_1}, \ldots, v_{i_{n+1}}$. Since A' has order $n \times (n+1)$ and has rank n, $\ker(A')$ is generated by a single nonzero vector. It follows readily that the vector α given by Equation $(*)$ is an elementary vector.

($\Rightarrow$) Let $\operatorname{supp}(\alpha) = \{j_1, \ldots, j_r\}$ and let $v_1, \ldots, v_q$ be the column vectors of A. Since $\alpha \in \ker(A)$ one has

$$\alpha = \lambda_{j_1} e_{j_1} + \cdots + \lambda_{j_r} e_{j_r} \quad \text{and} \quad \lambda_{j_1} v_{j_1} + \cdots + \lambda_{j_r} v_{j_r} = 0, \tag{13}$$

where $0 \neq \lambda_{j_k} \in \mathbb{Q}$ and v_{j_k} are columns of A for all $k = 1, \ldots, r$. By the minimality of the support of α one may assume that $v_{j_1}, \ldots, v_{j_{r-1}}$ are linearly independent. Thus using that $\operatorname{rank}(A) = n$, there are column vectors $v_{j_{r+1}}, \ldots, v_{j_{n+1}}$ such that the set

$$\mathcal{A}' = \{v_{j_1}, \ldots, v_{j_{r-1}}, v_{j_{r+1}}, \ldots, v_{j_{n+1}}\}$$

is linearly independent. One can write

$$\mathcal{A} = \mathcal{A}' \cup \{v_{j_r}\} = \{v_{i_1}, \ldots, v_{i_{n+1}}\}$$

such that $1 \leqslant i_1 < \cdots < i_{n+1} \leqslant q$. Consider the vector

$$\beta = \sum_{k=1}^{n+1} (-1)^k \det[v_{i_1}, \ldots, v_{i_{k-1}}, v_{i_{k+1}}, \ldots, v_{i_{n+1}}] e_{i_k}. \tag{14}$$

Note that in Equations (13) and (14) the coefficient of e_{j_r} is nonzero, hence there are scalars $0 \neq \lambda$ and c_{i_k} in $\mathbb{Q}$ such that

$$\alpha - \lambda\beta = \sum_{i_k \neq j_r} c_{i_k} v_{i_k}.$$

Since the support of $\alpha - \lambda\beta$ is strictly contained in the support of β and since β is an elementary vector of $\ker(A)$, one obtains $\alpha - \lambda\beta = 0$, as required. □

THEOREM 5.10. *Let M be an $n \times q$ integral matrix and ψ the map*

$$\psi : \mathbb{Z}^q \to \mathbb{Z}^n$$

given by $\psi(\alpha) = M(\alpha)$. Then $\ker(\psi)$ *is generated as a $\mathbb{Z}$-module by the elementary integral vectors of* $\ker(\psi)$.

PROOF. It follows from Proposition 5.9 and the fundamental theorem of finitely generated Abelian groups. See [131, Theorem 8.4.10] for details. □

5.2. *Normality of rings and semigroups*

Various results about the normality of monomial subrings will be discussed below. We will use some recent techniques based on linear programming and graph theory, which have proven to be useful to study the normality.

Normal semigroups. Let $A = (a_{ij})$ be an integral matrix of order $n \times q$ with nonzero distinct columns and let $\mathcal{A} = \{v_1, \dots, v_q\}$ be the set of column vectors of A. The *integral closure* or *normalization* of the semigroup

$$\mathbb{N}\mathcal{A} = \mathbb{N}v_1 + \cdots + \mathbb{N}v_q \subset \mathbb{Z}^n,$$

associated to A, is defined as:

$$\overline{\mathbb{N}\mathcal{A}} = \mathbb{Z}\mathcal{A} \cap \mathbb{R}_+\mathcal{A}.$$

Here $\mathbb{R}_+\mathcal{A}$ denotes the *rational polyhedral cone* in $\mathbb{R}^n$ given by

$$\mathbb{R}_+\mathcal{A} = \left\{ \sum_{i=1}^{q} \lambda_i v_i \;\middle|\; \lambda_i \in \mathbb{R}_+ \text{ for all } i \right\},$$

where $\mathbb{R}_+$ is the set of nonnegative real numbers, and $\mathbb{Z}\mathcal{A}$ is the subgroup of $\mathbb{Z}^n$ generated by $\mathcal{A}$.

By Gordan's lemma [54] there are $\gamma_1, \dots, \gamma_r \in \mathbb{Z}^n$ such that

$$\overline{\mathbb{N}\mathcal{A}} = \mathbb{N}\mathcal{A} + \mathbb{N}\gamma_1 + \cdots + \mathbb{N}\gamma_r.$$

There is an algorithm [25,26] to compute $\gamma_1, \dots, \gamma_r$.

DEFINITION 5.11. The semigroup $\mathbb{N}\mathcal{A}$ is said to be *normal* if $\overline{\mathbb{N}\mathcal{A}} = \mathbb{N}\mathcal{A}$.

Normal semigroup rings. Let $R = K[x_1, \dots, x_n]$ be a polynomial ring over a field K. There is a one to one correspondence

$$a = (a_1, \dots, a_n) \leftrightarrow x^a = x_1^{a_1} \cdots x_n^{a_n},$$

between $\mathbb{N}^n$ and the semigroup of monomials of R. If the matrix A has nonnegative entries under this correspondence

$$\mathcal{A} = \{v_1, \dots, v_q\} \leftrightarrow F = \{x^{v_1}, \dots, x^{v_q}\}.$$

We associate to A the *semigroup ring* or *monomial subring* given by

$$K[F] = K\big[\{x^a \mid a \in \mathbb{N}\mathcal{A}\}\big] \subset R,$$

that is, $K[F]$ is the subring of R generated by F over K.

DEFINITION 5.12. The *integral closure* or *normalization* of $K[F]$, denoted by $\overline{K[F]}$, is the set of all elements in its field of fractions which are integral over $K[F]$. If $K[F] = \overline{K[F]}$ the ring $K[F]$ is said to be *normal*.

A key result in the theory of monomial subrings is the following theorem of M. Hochster.

THEOREM 5.13 [77]. *If $K[F]$ is normal, then $K[F]$ is Cohen–Macaulay.*

PROOF. See [16]. □

The next result is a description of the normalization of a monomial subring. For the computation of normalizations of affine normal domains see [39,122].

THEOREM 5.14. $\overline{K[F]} = K[\{x^a \mid a \in \mathbb{Z}\mathcal{A} \cap \mathbb{R}_+\mathcal{A}\}]$.

PROOF. See [131, Chapter 7]. □

REMARK 5.15. Thus the normalization of $K[F]$ is again a monomial subring. Observe that because of the description above it is possible to use *Normaliz* [25,26] to compute the normalization of $K[F]$.

COROLLARY 5.16. *$\mathbb{N}\mathcal{A}$ is normal if and only if $K[F]$ is normal.*

DEFINITION 5.17. An integral matrix A is *t-unimodular* if all the nonzero $r \times r$ minors of A have absolute value equal to t, where $r \neq 0$ is the rank of A, if $t = 1$ we say that A is *unimodular*.

For simplicity we keep the notation introduced above throughout the rest of this section.

If $(M, +)$ is an Abelian group its torsion subgroup, denoted by $T(M)$, is the set of all x in M such that $px = 0$ for some $0 \neq p \in \mathbb{N}$.

Let $b \in \mathbb{Z}^n$ be a fixed column vector such that $r = \operatorname{rank}(A) = \operatorname{rank}([A\, b])$. By a classical result of Kronecker [104, p. 51] the system $Ax = b$ has an integral solution if and only if $\Delta_r(A) = \Delta_r([A\, b])$, where $\Delta_i(B)$ denotes the greatest common divisor of all the nonzero $i \times i$ minors of B.

To give some variants of Kronecker's theorem [50] note that there is a positive integer k such that

$$kb = \lambda_1 v_1 + \cdots + \lambda_q v_q \quad (\lambda_i \in \mathbb{Z}\ \forall i). \tag{$*$}$$

Therefore there is a canonical epimorphism of finite groups

$$\varphi : T\big(\mathbb{Z}^n/\mathbb{Z}\mathcal{A}\big) \to T\big(\mathbb{Z}^n/\mathbb{Z}\mathcal{B}\big) \quad \big(\alpha + \mathbb{Z}\mathcal{A} \overset{\varphi}{\longmapsto} \alpha + \mathbb{Z}\mathcal{B}\big),$$

where $\mathcal{B} = \mathcal{A} \cup \{b\}$.

LEMMA 5.18 (Kronecker). *The following conditions are equivalent*:

(i) *φ is injective.*

(ii) *$b \in \mathbb{Z}\mathcal{A}$.*

(iii) *The groups $\mathbb{Z}^n/\mathbb{Z}\mathcal{A}$ and $\mathbb{Z}^n/\mathbb{Z}\mathcal{B}$ have the same invariant factors.*

(iv) *The matrices* $[A\,0]$ *and* $[A\,b]$ *have the same Smith normal form.*
(v) $\Delta_r(A) = \Delta_r([A\,b])$, *where* $r = \mathrm{rank}(A)$.

PROOF. It follows basically from the fundamental theorem of finitely generated Abelian groups [81, Chapter 3]. See [50] for details. □

REMARK 5.19. Recall that $\mathbb{R}_+\mathcal{A}$ can be expressed as an intersection of closed half spaces through the origin [33], that is, there is a rational matrix C such that

$$\mathbb{R}_+\mathcal{A} = \left\{x \in \mathbb{R}^n \mid Cx \leqslant 0\right\}.$$

Thus this representation together with Lemma 5.18 yield a membership test to decide when a given α in $\mathbb{Z}^n$ belongs to $\overline{\mathbb{N}\mathcal{A}} = \mathbb{Z}\mathcal{A} \cap \mathbb{R}_+\mathcal{A}$.

PROPOSITION 5.20. *If A is a t-unimodular matrix and $v_{i_1}, \dots, v_{i_r}$ is a $\mathbb{Q}$-basis for the column space of A, then*

$$\mathbb{Z}\mathcal{A} = \mathbb{Z}v_{i_1} \oplus \cdots \oplus \mathbb{Z}v_{i_r}.$$

PROOF. For each j one has $\Delta_r([v_{i_1} \cdots v_{i_r}]) = \Delta_r([v_{i_1} \cdots v_{i_r}\ v_j]) = t$, then the result is a consequence of Lemma 5.18. □

The next result follows from [117, Remark 8.10 and Proposition 13.15]. Note that the proof given in loc. cit. basically works in general, not only for homogeneous ideals. The proof below, in contrast to that of [117], is direct and do not make any use of Gröbner bases techniques.

THEOREM 5.21. *If A is a t-unimodular matrix, then $\mathbb{N}\mathcal{A}$ is normal.*

PROOF. Let $b \in \mathbb{Z}\mathcal{A} \cap \mathbb{R}_+\mathcal{A}$. By Carathéodory's theorem [54, Theorem 2.3] there are $v_{i_1}, \dots, v_{i_r}$ linearly independent columns of A, where r is the rank of A, such that

$$b \in \mathbb{R}_+v_{i_1} + \cdots + \mathbb{R}_+v_{i_r}. \tag{15}$$

On the other hand by Proposition 5.20 one has

$$b \in \mathbb{Z}v_{i_1} + \cdots + \mathbb{Z}v_{i_r}. \tag{16}$$

Since $v_{i_1}, \dots, v_{i_r}$ are linearly independent, comparing the coefficients of b with respect to the two representations given by (15) and (16) one derives $b \in \mathbb{N}\mathcal{A}$. This proof is due to C. Escobar, I. Gitler and R. Villarreal [50]. □

The material in this section is related to covering properties. For a discussion on the existence of unimodular (Hilbert) covers of rational cones see [20] and the references there.

COROLLARY 5.22 [117]. *If A is a t-unimodular matrix with entries in $\mathbb{N}$, then $K[F]$ is a normal domain.*

PROOF. It follows from Theorem 5.21. □

REMARK 5.23. There is a converse to Corollary 5.22 and Theorem 5.21 that holds for Lawrence ideals, that is, defining ideals of toric subvarieties of $\mathbb{P}^1 \times \mathbb{P}^1 \times \cdots \times \mathbb{P}^1$, see [9, Theorem 1.2].

Next we present a sufficient condition for the normality of $K[F]$ in terms of the elementary integral vectors of the kernel of the matrix A.

THEOREM 5.24 [15]. *If the entries of the matrix A are nonnegative and each elementary integral vectors of* $\ker(A)$ *has a positive or negative part with entries consisting of* 0*'s and* 1*'s, then $K[F]$ is normal.*

DEFINITION 5.25. The *toric ideal P* of the monomial subring $K[F]$ is the kernel of the epimorphism

$$\varphi : B = K[t_1, \ldots, t_q] \to K[F] \quad (t_i \overset{\varphi}{\longmapsto} x^{v_i})$$

of K algebras, where B is a polynomial ring in the t_i variables.

COROLLARY 5.26. *If the toric ideal P of $K[F]$ is a principal ideal generated by a binomial of the form $t_1 \cdots t_s - t_{s+1}^{a_{s+1}} \cdots t_q^{a_q}$, then $K[F]$ is normal.*

A nice property of toric ideals is that they are binomial ideals, that is, they are generated by polynomials of the form $t^{\alpha_+} - t^{\alpha_-}$ [46]. Furthermore the reduced Gröbner basis of a toric ideal consist of binomials [117].

DEFINITION 5.27. A monomial subring $K[x^{v_1}, \ldots, x^{v_q}]$ is *homogeneous* if $v_1, \ldots, v_q$ lie on an affine hyperplane of $\mathbb{R}^n$ not containing the origin.

THEOREM 5.28 [117]. *Let $<$ be a monomial order of $K[t_1, \ldots, t_q]$. If $K[F]$ is homogeneous and the initial ideal of its toric ideal P is square-free, then $K[F]$ is normal.*

Normality of the Rees algebra. A matrix A is called *totally unimodular* if each $i \times i$ minor of A is 0 or ± 1 for all $i \geqslant 1$. Examples of totally unimodular matrices include incidence matrices of hypergraphs without odd cycles [11, Chapter 5], incidence matrices of directed graphs [104, p. 274] and network matrices [104, Chapter 19].

DEFINITION 5.29. The *Rees algebra* of the ideal $I = (x^{v_1}, \ldots, x^{v_q})$ is the K-subring:

$$\mathcal{R}(I) = K\left[x^{v_1}t, \ldots, x^{v_q}t, x_1, \ldots, x_n\right] \subset R[t],$$

where t is a new variable.

The normality of Rees algebras has been studied by several authors [19,55,72,99,107, 123,128]. In our present situation it is a fact that the normality of $\mathcal{R}(I)$ is equivalent to the normality of the ideal I [123].

THEOREM 5.30 [107]. *If $\mathcal{R}(I)$ is normal and I is generated by monomials of the same degree, then $K[F]$ is normal.*

Next we will use linear programming techniques to study the normality of certain Rees algebras. Our main references for linear programming are [104,132]. In the proof below we use standard notation and terminology as described in those references, in particular note that if $a = (a_i)$ and $b = (b_i)$ are two vectors in $\mathbb{R}^m$ then $a \geqslant b$ means that $a_i \geqslant b_i$ for all i.

THEOREM 5.31 [15]. *Let $A = (a_{ij})$ be a $\{0, 1\}$ totally unimodular matrix of order $n \times q$ with nonzero column vectors $v_1, \ldots, v_q$. If I is the ideal of R generated by the monomials $\mathbf{x}^{v_1}, \ldots, \mathbf{x}^{v_q}$, then the Rees algebra $\mathcal{R}(I)$ of I is a normal domain.*

PROOF. We define $\mathcal{A}' = \{(v_1, 1), \ldots, (v_q, 1), (e_1, 0), \ldots, (e_n, 0)\}$, where e_i is the ith unit vector in $\mathbb{R}^n$. According to Theorem 5.14 one has

$$\overline{\mathcal{R}(I)} = K\left[\left\{\mathbf{x}^{\alpha} t^{b} \mid (\alpha, b) \in \mathbb{Z}\mathcal{A}' \cap \mathbb{R}_+\mathcal{A}'\right\}\right] \subset R[t].$$

Note $\mathbb{Z}\mathcal{A}' = \mathbb{Z}^{n+1}$. Let $\mathbf{x}^{\alpha} t^{b} \in \overline{\mathcal{R}(I)}$. One can write

$$(\alpha, b) = \lambda_1(v_1, 1) + \cdots + \lambda_q(v_q, 1) + \mu_1(e_1, 0) + \cdots + \mu_n(e_n, 0),$$

where λ_i and μ_j are nonnegative rational numbers for all i, j. Hence one obtains the equalities

$$\begin{aligned} \alpha &= \lambda_1 v_1 + \cdots + \lambda_q v_q + \mu_1 e_1 + \cdots + \mu_n e_n, \\ b &= \lambda_1 + \cdots + \lambda_q. \end{aligned}$$

Let A' be the $n \times (q + n)$ matrix

$$A' = (A\, e_1 \cdots e_n)$$

obtained from A by adjoining the column vectors $e_1, \ldots, e_n$. Since total unimodularity is preserved under the operation of adding columns of unit vectors, the matrix A' is totally unimodular. Consider the linear program:

$$\begin{aligned} &\text{Maximize } x_1 + \cdots + x_q \qquad\qquad (*) \\ &\text{Subject to} \\ &A'x = \alpha \text{ and } x \geqslant 0, \end{aligned}$$

where x is the column vector $x = (x_1, \ldots, x_q, x_{q+1}, \ldots, x_{q+n})$. Since the column vector

$$c = (c_1, \ldots, c_{n+q}) = (\lambda_1, \ldots, \lambda_q, \mu_1, \ldots, \mu_n)$$

satisfies

$$A'c = \alpha, \quad c \geqslant 0 \quad \text{and} \quad c_1 + \cdots + c_q = b,$$

one concludes that the linear program $(*)$ has an optimal value greater or equal than b, which is attained at a vertex x_0 of the rational polytope

$$P = \{x \in \mathbb{R}^{n+q} \mid A'x = \alpha \text{ and } x \geqslant 0\},$$

see [132, Theorem 4.1.6]. Observe that by [104, Theorem 19.2] all the vertices of P have integral entries because A' is totally unimodular. Thus x_0 is a vector with nonnegative integral entries:

$$x_0 = (\eta_1, \ldots, \eta_q, \delta_1, \ldots, \delta_n)$$

such that $\eta_1 + \cdots + \eta_q \geqslant b$. There are integers $\varepsilon_1, \ldots, \varepsilon_q$ such that

$$0 \leqslant \varepsilon_i \leqslant \eta_i \ \forall i \quad \text{and} \quad \sum_{i=1}^{q} \varepsilon_i = b.$$

Therefore

$$\begin{aligned}\mathbf{x}^\alpha t^b &= (\mathbf{x}^{v_1})^{\eta_1} \cdots (\mathbf{x}^{v_q})^{\eta_q} x_1^{\delta_1} \cdots x_n^{\delta_n} t^b \\ &= [(t\mathbf{x}^{v_1})^{\varepsilon_1} \cdots (t\mathbf{x}^{v_q})^{\varepsilon_q}][(\mathbf{x}^{v_1})^{\eta_1-\varepsilon_1} \cdots (\mathbf{x}^{v_q})^{\eta_q-\varepsilon_q} x_1^{\delta_1} \cdots x_n^{\delta_n}],\end{aligned}$$

and $\mathbf{x}^\alpha t^b \in \mathcal{R}(I)$, as required. □

5.3. *Ehrhart rings*

Consider the lattice polytope

$$P = \operatorname{conv}(v_1, \ldots, v_q) \subset \mathbb{R}^n,$$

where conv($\mathcal{A}$) denotes the convex hull of $\mathcal{A}$. We are interested in comparing the following monomial subrings:

- the *Ehrhart ring* of P: $A(P) = K[x^\alpha t^i \mid \alpha \in \mathbb{Z}^n \cap iP] \subset R[t]$, where t is a new variable, and
- the *monomial subring*: $K[Ft] = K[x^{v_1}t, \ldots, x^{v_q}t] \subset R[t]$.

A first observation is that $A(P)$ is a normal domain and $K[Ft] \subset A(P)$ is an integral extension [23], thus $\overline{K[Ft]} = A(P)$ if and only if $A(P)$ is contained in the field of fractions of $K[Ft]$. Thus it is natural to try to compare $A(P)$ with $\overline{K[Ft]}$.

There are various sufficient conditions for the occurrence of the equality

$$\overline{K[Ft]} = A(P),$$

see [51]. Those conditions are given in terms of Rees algebras, minors, unimodular matrices, and unimodular coverings.

Let us introduce some more relevant features of Ehrhart rings. The ring $A(P)$ is a finitely generated K-algebra [23] and it is a graded K-algebra with ith component given by

$$A(P)_i = \sum_{\alpha \in \mathbb{Z}^n \cap iP} Kx^\alpha t^i,$$

as a consequence, the Hilbert function of $A(P)$:

$$h(i) = \left|\mathbb{Z}^n \cap iP\right| = c_d i^d + \cdots + c_1 i + c_0 \quad (i \gg 0)$$

is a polynomial function of degree $d = \dim(P)$ such that $d!c_d$ is an integer, which is the *multiplicity* of $A(P)$. The polynomial $c_d x^d + \cdots + c_1 x + c_0 \in \mathbb{Q}[x]$ is called the *Ehrhart polynomial* of P. By [115] the relative volume of P is:

$$\mathrm{vol}(P) = \lim_{i \to \infty} \frac{|\mathbb{Z}^n \cap iP|}{i^d}.$$

Hence vol(P) is the leading coefficient of the Ehrhart polynomial of P. For this reason $d!c_d$ is often referred to as the *normalized volume* of P.

For more details on Ehrhart rings consult [23,73,115] and the references there. For the actual computation of the generators of the Ehrhart ring and for the computation of the Ehrhart polynomial the best way is to use *Normaliz* [25]. In practice one can compute relative volumes of lattice polytopes using [25,29,30,40].

If $v_1, \ldots, v_q$ lie on an affine hyperplane not containing the origin, then

$$K[F] = K\left[x^{v_1}, \ldots, x^{v_q}\right] \simeq K[Ft],$$

as graded algebras, in this case we say that $K[F]$ is *homogeneous*. There is $x_0 \in \mathbb{Q}^n$ such that $\langle v_i, x_0\rangle = 1$ for all i. Therefore $K[F]$ is a standard graded algebra with the grading

$$K[F]_i = \sum_{|c|=i} K\left(x^{v_1}\right)^{c_1} \cdots \left(x^{v_q}\right)^{c_q}, \quad \text{where } |c| = c_1 + \cdots + c_q,$$

and $K[F]$ has a well defined Hilbert polynomial $h(x)$. The *multiplicity* $e(K[F])$ of $K[F]$ is the leading coefficient of $h(x)$ times $d!$, where $d+1$ is the Krull dimension of $K[F]$. Note $d = \dim(P)$, see [131, Chapter 7] for details. A standard reference for homogeneous subrings and their connections to convex polytopes is [117].

For convenience we will keep the notation introduced above. The vector v_i will be regarded as a column vector when using matrix notation. Recall that $\Delta_r(B)$ denotes the greatest common divisor of all the nonzero $r \times r$ minors of a matrix B.

THEOREM 5.32 [51]. *If the matrix*

$$B = \begin{pmatrix} v_1 & \dots & v_q \\ 1 & \dots & 1 \end{pmatrix}$$

has rank r and $\Delta_r(B) = 1$, then $\overline{K[Ft]} = A(P)$.

PROOF. Since $K[Ft] \subset A(P)$ is an integral extension of rings and $A(P)$ is normal one has $\overline{K[Ft]} \subset A(P)$.

For the other containment it suffices to prove that $A(P)$ is contained in the field of fractions of $K[Ft]$. Let $x^\alpha t^i \in A(P)_i$, that is, $\alpha \in \mathbb{Z}^n \cap iP$ and $i \in \mathbb{N}$. Hence the system of equations

$$By = \begin{pmatrix} \alpha \\ i \end{pmatrix} = \alpha' \tag{17}$$

has a real solution. Hence by Gaussian elimination the system has a rational solution y. Here we regard α as a column vector. Thus the augmented matrix $[B\,\alpha']$ has rank r. Observe that in general $\Delta_r([B\,\alpha'])$ divides $\Delta_r(B)$, so in this case, they are equal. Using Kronecker's theorem one derives that the linear system (17) has an integral solution. Therefore $x^\alpha t^i$ is in the field of fractions of $K[Ft]$, as required. □

COROLLARY 5.33 [51]. *Assume $v_1, \dots, v_q$ lie in a hyperplane not containing the origin and $d = \dim(P)$. If the matrix*

$$B = \begin{pmatrix} v_1 & \dots & v_q \\ 1 & \dots & 1 \end{pmatrix}$$

has rank r and $\Delta_r(B) = 1$, then $e(K[F]) = d!\mathrm{vol}(P)$.

PROOF. By Theorem 5.32 one obtains $\overline{K[Ft]} = A(P)$. On the other hand using [117, Theorem 4.16] one has the equality $e(K[Ft]) = e(\overline{K[Ft]})$. □

THEOREM 5.34 [51]. *If $\mathcal{A} = \{\alpha_0, \dots, \alpha_m\} \subset \mathbb{Z}^n$ is a set of vectors lying on an affine hyperplane not containing the origin and $P = \mathrm{conv}(\mathcal{A})$, then*

$$\mathrm{vol}(P) = \left|T\left(\mathbb{Z}^n/(\alpha_1 - \alpha_0, \dots, \alpha_m - \alpha_0)\right)\right| \lim_{i\to\infty} \frac{|\mathbb{Z}\mathcal{A} \cap iP|}{i^d},$$

where $d = \dim(P)$ and $T(M)$ denotes the torsion subgroup of M.

COROLLARY 5.35 [51]. *If $\mathcal{A} = \{v_1, \ldots, v_q\} \subset \mathbb{N}^n$ and $P = \operatorname{conv}(\mathcal{A})$, then the multiplicities of $A(P)$ and $K[Ft]$ are related by*

$$e\big(A(P)\big) = \big|T\big(\mathbb{Z}^n/(v_2 - v_1, \ldots, v_q - v_1)\big)\big| e\big(K[Ft]\big).$$

The next result is the converse of Theorem 5.32.

COROLLARY 5.36 [51]. *If the matrix*

$$B = \begin{pmatrix} v_1 & \ldots & v_q \\ 1 & \ldots & 1 \end{pmatrix}$$

has rank r and $\overline{K[Ft]} = A(P)$, then $\Delta_r(B) = 1$.

PROOF. By hypothesis $e(K[Ft]) = \operatorname{vol}(P)(r-1)!$. Thus from Corollary 5.35 one obtains that $\mathbb{Z}^n/(v_2 - v_1, \ldots, v_q - v_1)$ is torsion-free. Hence the group $\mathbb{Z}^{n+1}/((v_1, 1), \ldots, (v_q, 1))$ is also torsion-free, as required. $\square$

5.4. *Toric ideals*

The aim is to examine when a given set of binomials defines a toric ideal up to radical.

Systems of binomials in toric ideals. Let $A = (a_{ij})$ be an integral matrix of order $n \times q$ with non negative entries, as before we assume that the columns of A are distinct and nonzero. The *toric set Γ* determined by the matrix A is the set in the affine space $\mathbb{A}_K^q$ given parametrically by

$$t_i = x_1^{a_{1i}} \cdots x_n^{a_{ni}} \quad (i = 1, \ldots, q).$$

Consider the graded epimorphism of K-algebras:

$$\varphi : B = K[t_1, \ldots, t_q] \to K[F] \quad \big(t_i \stackrel{\varphi}{\longmapsto} x^{v_i}\big),$$

where $v_i = (a_{1i}, \ldots, a_{ni})$ is the ith column of A,

$$x^{v_i} = x_1^{a_{1i}} \cdots x_n^{a_{ni}},$$

and $F = \{x^{v_1}, \ldots, x^{v_q}\}$. As before B is a polynomial ring in the t_i variables.

To emphasize the role of the toric set Γ we set $K[\Gamma] := K[F]$. The kernel of φ, denoted by P, is the *toric ideal* of the monomial subring $K[\Gamma]$. If $n = 1$ and $\gcd(a_{11}, \ldots, a_{1q}) = 1$, then Γ is called a *monomial curve*.

DEFINITION 5.37. We say that Γ is an *affine toric variety* if Γ is the zero set of the toric ideal P associated with A.

In [100] there is a characterization of toric sets which are affine toric varieties in terms of the existence of certain roots in the base field K and a vanishing condition.

Closely related to the map φ is the homomorphism

$$\psi : \mathbb{Z}^q \to \mathbb{Z}^n$$

determined by the matrix A. Indeed, we have

$$\varphi\big(t^\alpha\big) = x^{\psi(\alpha)}$$

for all $\alpha \in \mathbb{N}^q$. As a consequence, a binomial $g = t^\alpha - t^\beta$ belongs to $P = \ker(\varphi)$ if and only if $\hat{g} = \alpha - \beta$ belongs to $\ker(\psi)$.

The following two results can be proved elegantly using the notion of simple component of a polynomial with respect to a subgroup of $\mathbb{Z}^q$. This notion was introduced and studied by Shalom Eliahou [47,48].

PROPOSITION 5.38 [49]. *If $g_1, \dots, g_r$ is a set of binomials generating the toric ideal P of $K[\Gamma]$, then $\hat{g}_1, \dots, \hat{g}_r$ generate* $\ker(\psi)$.

DEFINITION 5.39. Let I be an ideal of a ring R and $f \in R$, the *saturation* of I with respect to f is

$$\big(I : f^\infty\big) = \bigcup_{i=1}^{\infty} \big(I : f^i\big) = \big\{r \in R \mid rf^i \in I,\ \text{for some } i \geqslant 1\big\}.$$

If R is a polynomial ring the saturation can be computed using Gröbner bases and the equality

$$\big(I : f^\infty\big) = (I, 1 - tf) \cap R,$$

where t is a new variable. See [36, Chapter III, Theorem 2].

PROPOSITION 5.40 [49]. *Let $g_1, \dots, g_r$ be a set of binomials in the toric ideal P of $K[\Gamma]$ and $I = (g_1, \dots, g_r)$. If* $\operatorname{char}(K) = p \neq 0$ (*respectively* $p = 0$)*, then the following two conditions are equivalent*:

(a$_1$) $P = \operatorname{rad}(I : z^\infty)$, *where $z = t_1 \cdots t_q$ and $(I : z^\infty) = \bigcup_{i \geqslant 1}(I : z^i)$.*

(a$_2$) $p^u \ker(\psi) \subset \langle \hat{g}_1, \dots, \hat{g}_r \rangle$ *for some $u \in \mathbb{N}$ (respectively* $\ker(\psi) = \langle \hat{g}_1, \dots, \hat{g}_r \rangle$).

In order to characterize when a given set of binomials defines a toric ideal set theoretically we need:

LEMMA 5.41 [46]. *Let R be a ring and let $t_1, \dots, t_q \in R$. If I is an ideal of R, then the radical of I satisfies*

$$\sqrt{I} = \sqrt{\big(I : (t_1 \cdots t_q)^\infty\big)} \cap \sqrt{(I, t_1)} \cap \cdots \cap \sqrt{(I, t_q)}.$$

PROPOSITION 5.42 [131]. *Let $\mathfrak{p}$ be a prime ideal of R and let $t_1, \ldots, t_q$ be a sequence in $R \setminus \mathfrak{p}$. If $I \subset \mathfrak{p}$ is an ideal, then $\sqrt{I} = \mathfrak{p}$ if and only if*

(a) $\mathfrak{p} = \sqrt{(I : (t_1 \cdots t_q)^\infty)}$, *and*

(b) $\sqrt{(I, t_i)} = \sqrt{(\mathfrak{p}, t_i)}$, *for all i.*

PROOF. It follows from Lemma 5.41. □

THEOREM 5.43 [49]. *Let Γ be a toric set and $g_1, \ldots, g_r$ a set of binomials in the toric ideal P of $K[\Gamma]$. Set $I = (g_1, \ldots, g_r)$. If* $\operatorname{char}(K) = p \neq 0$ (*respectively* $\operatorname{char}(K) = 0$), *then* $\operatorname{rad}(I) = P$ *if and only if*

(a) $p^u \ker(\psi) \subset \langle \hat{g}_1, \ldots, \hat{g}_r \rangle$ *for some $u \in \mathbb{N}$* (*respectively* $\ker(\psi) = \langle \hat{g}_1, \ldots, \hat{g}_r \rangle$),

(b) $\operatorname{rad}(I, t_i) = \operatorname{rad}(P, t_i)$ *for all i.*

PROOF. It is a consequence of Proposition 5.42 and Proposition 5.40. □

Given an ideal I of a polynomial ring we denote its *zero set* or *variety* by $V(I)$. Recall that $V(I)$ is the set of points α in affine space such that $f(\alpha) = 0$ for all $f \in I$.

An important consequence of Theorem 5.43 is the following effective criterion for curves. For some other applications of this result see [2,3].

COROLLARY 5.44 [49]. *Let $g_1, \ldots, g_r$ be a set of binomials in the ideal P of the monomial curve Γ. Set $I = (g_1, \ldots, g_r)$. If* $\operatorname{char}(K) = p \neq 0$ (*respectively* $\operatorname{char}(K) = 0$), *then* $\operatorname{rad}(I) = P$ *if and only if*

(a) $p^m \ker(\psi) \subset \langle \hat{g}_1, \ldots, \hat{g}_r \rangle$ *for some $m \in \mathbb{N}$* (*respectively* $\ker(\psi) = \langle \hat{g}_1, \ldots, \hat{g}_r \rangle$),

(b) $V(g_1, \ldots, g_r, t_i) = \{0\}$, *for all i.*

COROLLARY 5.45. *Let $g_1, \ldots, g_r$ be a set of binomials in a toric ideal P. If* $\operatorname{char}(K) = p \neq 0$ (*respectively* $\operatorname{char}(K) = 0$) *and* $p^m \ker(\psi) \subset G = \langle \hat{g}_1, \ldots, \hat{g}_r \rangle$ *for some $m \geqslant 0$* (*respectively* $\ker(\psi) = G$), *then*

$$V(g_1, \ldots, g_r) \subset V(P) \cup V(t_1 \cdots t_q).$$

PROOF. By Proposition 5.40, there is a monomial t^δ and an integer N such that $t^\delta P^N \subset I$. It follows that

$$V(I) \subset V\bigl(P^N\bigr) \cup V\bigl(t^\delta\bigr) \subset V(P) \cup V(t_1 \cdots t_q),$$

as required. □

DEFINITION 5.46. Let A be an integral matrix with entries in $\mathbb{N}$. The ideal

$$I = \bigl(\bigl\{t^{\alpha_+} - t^{\alpha_-} \mid \alpha \text{ is a circuit of } \ker(A)\bigr\}\bigr)$$

is called the *circuit ideal* of A.

PROPOSITION 5.47 [46]. *If I is the circuit ideal of the matrix A and P is the toric ideal determined by A, then*

$$\mathrm{rad}(I) = P.$$

PROOF. We proceed by induction on q, the number of columns of A. If $q = 1$ or $q = 2$ the result is clear because in those cases P is a principal ideal. According to Theorems 5.43 and 5.10 it suffices to prove

$$\mathrm{rad}(I, t_i) = \mathrm{rad}(P, t_i) \tag{$*$}$$

for all i. For simplicity of notation we assume $i = 1$. Let $\gamma_1, \ldots, \gamma_r$ be the elementary integral vectors of $\ker(A)$ and let

$$f_i = t^{(\gamma_i)_+} - t^{(\gamma_i)_-}.$$

For convenience set $f_0 = f_{r+1} = 0$. If $\mathfrak{q}$ is a prime ideal containing (I, t_1), then after relabeling $t_1, \ldots, t_q$ and $f_1, \ldots, f_r$ there are $t_1, \ldots, t_s \in \mathfrak{q}$ and $0 \leqslant k \leqslant r$ such that

(i) $(f_{k+1}, \ldots, f_r) \subset (t_1, \ldots, t_s)$, and

(ii) $t_i \notin \bigcup_{i=1}^{k} \mathrm{supp}(f_i)$ $\forall i = 1, \ldots, s$.

We claim that $\mathfrak{q}$ contains (P, t_1). Take $f = t^\alpha - t^\beta \in P$, where $\mathrm{supp}(\alpha)$ and $\mathrm{supp}(\beta)$ are disjoint. According to Theorem 5.7 one can write

$$p(\alpha - \beta) = n_1\gamma_1 + \cdots + n_r\gamma_r \tag{$**$}$$

for some $p \in \mathbb{N}_+$ and $n_i \in \mathbb{N}$, such that for each i either $n_i = 0$ or $n_i > 0$, γ_i is in harmony with $\alpha - \beta$, and $\mathrm{supp}(\gamma_i) \subset \mathrm{supp}(\alpha - \beta)$.

Case (I): If $t_i \in \mathrm{supp}(f) = \mathrm{supp}(t^\alpha) \cup \mathrm{supp}(t^\beta)$ for some $1 \leqslant i \leqslant s$, then $t_i \in \mathrm{supp}(f_j)$ for some j such that $n_j > 0$. Thus by (ii) one has $j \geqslant k + 1$ and $f_j \in (t_1, \ldots, t_s)$. From $(**)$ it follows that $f \in (t_1, \ldots, t_s)$.

Case (II): If $t_i \notin \mathrm{supp}(f)$ for all $i = 1, \ldots, s$. Consider the maps

$$\varphi' = \varphi|_{K[t_{s+1}, \ldots, t_q]} \quad \text{and} \quad \psi' = \psi|_{\mathbb{Z}e_{s+1} + \cdots + \mathbb{Z}e_q}$$

obtained by restriction of φ and ψ respectively. Note that $\gamma_1, \ldots, \gamma_k$ are the circuits of $\ker(\psi')$. Hence by induction one has

$$\mathrm{rad}(I') = P' = K[t_{s+1}, \ldots, t_q] \cap P,$$

where I' is a circuit ideal generated by $f_1, \ldots, f_k$. Since $f \in P'$, then there is $m \geqslant 1$ such that $f^m \in I' \subset \mathfrak{q}$. Thus in this case $f \in \mathfrak{q}$.

Altogether $\mathfrak{q}$ contains (P, t_1). As $\mathrm{rad}(P, t_1)$ is the intersection of the minimal primes of (P, t_1) one concludes $\mathrm{rad}(P, t_1) \subset \mathrm{rad}(I, t_1)$. Since the other containment is clear one derives the equality. This proof is due to C. Escobar and R. Villarreal. $\square$

Curves in positive characteristic. Let I be an ideal of a ring R the *arithmetical rank* of I, denoted by $r = \operatorname{ara}(I)$, is the least positive integer r such that there are $f_1, \dots, f_r$ with

$$\operatorname{rad}(f_1, \dots, f_r) = \operatorname{rad}(I).$$

By Krull's principal ideal theorem [5] one has $\operatorname{ara}(I) \geqslant \operatorname{ht}(I)$. If equality occurs I is called a *set-theoretic complete intersection*. See [120] for the related notion of binomial arithmetical rank.

Let K be a field and P the associated toric ideal of a monomial curve Γ in the affine space $\mathbb{A}_K^q$.

In characteristic zero it is an open problem whether P is a set theoretic complete intersection, several authors have studied this problem, see for instance [7,47,85,91,119]. The case $q = 3$ is treated in [17].

If $\operatorname{char}(K) = p > 0$, then using [35, Theorem 2] and [35, Remark 1] it follows that P is a set-theoretic complete intersection. The following result of T.T. Moh [91] shows that P is indeed generated up to radical by $q - 1$ *binomials*.

THEOREM 5.48 [91]. *Let $A = (a_1, \dots, a_q)$ be an integral matrix of order $1 \times q$ and let P be the toric ideal of $K[x^{a_1}, \dots, x^{a_q}]$. If $\operatorname{char}(K) = p > 0$ and $a_i > 0$ for all i, then $P = \operatorname{rad}(g_1, \dots, g_{q-1})$ for some binomials $g_1, \dots, g_{q-1}$.*

PROOF. See [3] for a proof that uses Corollary 5.44. $\square$

6. Monomial subrings of graphs

Here we study monomial subrings associated to graphs. Several aspects of those subrings and their toric ideals have been studied in the literature [41,43,74,106–108,123,124,127]. In this section we describe the integral closure of those subrings and give some applications.

The edge subring $k[G]$ associated to a graph G is the monomial subring generated by the monomials corresponding to the edges of G, where k is a field. The description of the integral closure of $k[G]$ will be given in terms of special circuits of the graph (see Theorem 6.11). This description links the normality property of $k[G]$ with the combinatorics of the graph G (see Proposition 6.19). The incidence matrix of G plays an important role because its rank can be interpreted in graph theoretical terms.

6.1. *The subring associated to a graph*

Let G be a graph on the vertex set $V = \{v_1, \dots, v_n\}$ and

$$R = k[x_1, \dots, x_n] = \bigoplus_{i=0}^{\infty} R_i$$

a polynomial ring over a field k with the standard grading. To simplify notation sometimes one identifies the indeterminate x_i with the vertex v_i. The *monomial subring* or *edge subring* of the graph G is the k-subalgebra

$$k[G] = k\big[\{x_i x_j \mid v_i \text{ is adjacent to } v_j\}\big] \subset R.$$

If $F = \{f_1, \ldots, f_q\}$ is the set of monomials $x_i x_j$ such v_i is adjacent to v_j, the elements in $k[G]$ are polynomial expressions in F with coefficients in k.

As $k[G]$ is a standard k-algebra with the normalized grading

$$k[G]_i = k[G] \cap R_{2i},$$

there is a graded epimorphism of k-algebras

$$\varphi : B = k[t_1, \ldots, t_q] \to k[G], \quad t_i \mapsto f_i,$$

where B is a polynomial ring graded by $\deg(t_i) = 1$ for all i. The kernel of φ, denoted by $P(G)$, is a graded ideal of B called the *toric ideal* of $k[G]$ with respect to $f_1, \ldots, f_q$.

A dimension formula. Let us describe the *cycle space* of a graph G over the two element field $\mathbb{Z}_2$. Let C_0 and C_1 denote the vector spaces over $\mathbb{Z}_2$ of 0-chains and 1-chains respectively. Recall that a 0-chain of G is a formal linear combination

$$\sum a_i x_i$$

of the vertices $x_1, \ldots, x_n$ of G, and a 1-chain is a formal linear combination

$$\sum b_i f_i$$

of the edges $f_1, \ldots, f_q$ of G, where $a_i \in \mathbb{Z}_2$ and $b_i \in \mathbb{Z}_2$. The boundary operator ∂ is the linear transformation defined by

$$\partial(f_k) = \partial(\{x_i, x_j\}) = x_i + x_j.$$

The *cycle space* $\mathcal{Z}(G)$ of G over $\mathbb{Z}_2$ is by definition equal to $\ker(\partial)$. The elements in $\mathcal{Z}(G)$ are called *cycle vectors*.

REMARK 6.1. If G is connected, then $\dim_{\mathbb{Z}_2} \mathcal{Z}(G) = q - n + 1$, see [70].

The next result shows how the Krull dimension of $k[G]$ is related to the cycle space of its graph.

PROPOSITION 6.2 [127]. *Let G be a connected graph and let P be the toric ideal of the edge subring $k[G]$. Then*

$$\operatorname{ht}(P) = \dim_{\mathbb{Z}_2} \mathcal{Z}_e(G),$$

where $\mathcal{Z}_e(G)$ is the subspace of $\mathcal{Z}(G)$ of all even cycle vectors of G.

COROLLARY 6.3. *If G is a connected graph with n vertices, then*

$$\dim(k[G]) = \begin{cases} n & \text{if } G \text{ is not bipartite,} \\ n-1 & \text{otherwise.} \end{cases}$$

Incidence matrix of a graph. Let G be a simple graph with vertex set $V = \{x_1, \ldots, x_n\}$ and edge set $E = \{z_1, \ldots, z_q\}$, where every edge z_i is an unordered pair of distinct vertices $z_i = \{x_{i_j}, x_{i_k}\}$.

The *incidence matrix* $M_G = [b_{ij}]$ of G is the $n \times q$ matrix defined by

$$b_{ij} = \begin{cases} 1 & \text{if } x_i \in z_j, \\ 0 & \text{if } x_i \notin z_j. \end{cases}$$

Note that each column of M_G has exactly two 1's and the rest of its entries equal to zero. If $z_i = \{x_{i_j}, x_{i_k}\}$ define $\alpha_i = e_{i_j} + e_{i_k}$, where e_i is the ith canonical vector in $\mathbb{R}^n$. Thus the columns of M_G are precisely the vectors $\alpha_1, \ldots, \alpha_q$ regarded as column vectors.

REMARK 6.4. Let A be a square submatrix of M_G. In [66] it is shown that either $\det(A) = 0$ or $\det(A) = \pm 2^k$, for some integer k such that $0 \leqslant k \leqslant \tau_0$, where τ_0 is the maximum number of vertex disjoint odd cycles in G. Moreover for any such value of k there exists a minor equal to $\pm 2^k$.

Recall that a graph G is *bipartite* if all its cycles are of even length. Thus any tree and in particular any point is a bipartite graph. The number of bipartite connected components of G will be denoted by c_0, and the number of nonbipartite connected components will be denoted by c_1. Thus $c = c_0 + c_1$ is the total number of components of G.

THEOREM 6.5 [66]. *If G is a graph with n vertices and M_G its incidence matrix, then* $\operatorname{rank}(M_G) = n - c_0$.

PROOF. After permuting the vertices we may assume that the incidence matrix is "diagonal":

$$M_G = \operatorname{diag}(M_1, \ldots, M_c),$$

where $G_1, \ldots, G_c$ are the connected components of G and M_i is the incidence matrix of G_i. Since the rank of M_i is the dimension of $k[G_i]$ the result follows from Corollary 6.3. □

THEOREM 6.6 [66]. *If G is a graph with n vertices and M_G its incidence matrix, then there are invertible integral matrices U, V such that the Smith normal form of M_G is*:

$$S = U M_G V = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix},$$

where $D = \operatorname{diag}(1, \ldots, 1, 2, \ldots, 2)$, $n - c$ is the number of 1's and c_1 is the number of 2's. In particular the invariant factors of M_G are either 1 or 2.

COROLLARY 6.7. *Let G be a graph with vertices $x_1, \dots, x_n$ and $\mathcal{A}_G$ the set of all vectors $e_i + e_j$ such that $\{x_i, x_j\}$ is an edge of G. Then*

$$\mathbb{Z}^n / \mathbb{Z}\mathcal{A}_G \simeq \mathbb{Z}^{n-r} \times \mathbb{Z}_2^{c_1} = \mathbb{Z}^{c_0} \times \mathbb{Z}_2^{c_1},$$

where $r = n - c_0$ is the rank of M_G.

PROOF. It follows at once from the fundamental structure theorem of finitely generated modules over a principal ideal domain. See [81, Chapter 3]. $\square$

Rees algebras of bipartite graphs. An integral matrix A is called *totally unimodular* if each $i \times i$ minor of A is 0 or ± 1 for all $i \geqslant 1$. Recall that the bipartite simple graphs are characterized as those graphs whose incidence matrix is totally unimodular [104, Chapter 19].

Let A be an integral matrix with entries in $\{0, 1\}$. We consider the matrix C obtained from A by first adding a row of 1's to the matrix A and then adding the canonical vector $e_1, \dots, e_n$ as column vectors. In general, the first operation does not preserve total unimodularity even if one assumes that all the columns of A have exactly k entries equal to 1. The case $k = 2$, which is the interesting case here, is treated below.

THEOREM 6.8 [50]. *Let G be a simple bipartite graph with n vertices and q edges and let $A = (a_{ij})$ be its incidence matrix. If $e_1, \dots, e_n$ are the first n unit vectors in $\mathbb{R}^{n+1}$ and C is the matrix*

$$C = \begin{pmatrix} a_{11} & \dots & a_{1q} & e_1 & \dots & e_n \\ \vdots & \vdots & \vdots & & & \\ a_{n1} & \dots & a_{nq} & & & \\ 1 & \dots & 1 & & & \end{pmatrix}$$

obtained from A by adjoining a row of 1's and the column vectors $e_1, \dots, e_n$, then C is totally unimodular.

PROOF. It suffices to prove that the matrix B obtained from A by adding a row of 1's is totally unimodular.

Let $\{1, \dots, m\}$ and $\{m + 1, \dots, n\}$ be a bipartition of G. Let C' be the matrix obtained by deleting the last $n - m$ columns from C. It suffices to show that C' is totally unimodular. First one successively subtracts the rows $1, 2, \dots, m$ from the row $n + 1$. Then one reverses the sign in the rows $m + 1, \dots, n$. These elementary row operations produce a new matrix C''. The matrix C'' is the incidence matrix of a directed graph, hence it is totally unimodular [104, p. 274]. As the last m column vectors of C'' are

$$e_1 - e_{n+1}, \dots, e_m - e_{n+1},$$

one can successively pivot on the first nonzero entry of $e_i - e_{n+1}$ for $i = 1, \dots, m$ and reverse the sign in the rows $m + 1, \dots, n$ to obtain back the matrix C'. Here a pivot on the entry c'_{st} means transforming column t of C'' into the sth unit vector by elementary row operations. Since pivoting preserves total unimodularity [95, Lemma 2.2.20], C' is totally unimodular, and hence so is C. This proof is due to Bernd Sturmfels. $\square$

EXAMPLE 6.9 (*Truemper*). Consider the matrices

$$V = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} A \\ 1\,1\,1\,1 \end{pmatrix}.$$

The matrix V has determinant -2, A is totally unimodular, but B is not because V is a submatrix of B. Thus Theorem 6.8 does not extend to incidence matrices of k-hypergraphs with $k > 2$.

There is another family of $\{0, 1\}$ matrices preserving total unimodularity when adjoining a row or column consisting of 1's, see [121, Lemma 12.3.4].

COROLLARY 6.10 [107]. *Let $R = k[x_1, \ldots, x_n]$ be a polynomial ring over a field k. If G is a bipartite graph with vertices $x_1, \ldots, x_n$ and K is a field, then the Rees algebra*

$$\mathcal{R} = K\big[\{x_i x_j t \mid x_i \text{ is adjacent to } x_j\} \cup \{x_1, \ldots, x_n\}\big] \subset R[t]$$

is a normal domain. Here t denotes a new variable.

PROOF. It follows from Theorem 6.8 and Theorem 5.21. $\square$

A generalization of Corollary 6.10 is shown in Theorem 5.31 using linear programming techniques, but this generalization does not give any special information about the matrix C that defines the Rees algebra.

6.2. *The integral closure of an edge subring*

Let G be a graph and let $k[G]$ be the edge subring of G over an arbitrary field k. Our goal here is to unfold a construction for the integral closure or normalization $\overline{k[G]}$ of $k[G]$, for this purpose the underlying graph theoretic aspects are very helpful.

6.2.1. *A combinatorial description of normalizations.* Let G be a graph with vertex set $V = \{x_1, \ldots, x_n\}$ and let $R = k[x_1, \ldots, x_n]$ be a polynomial ring over a field k.

The set $\{f_1, \ldots, f_q\}$ will denote the set of all monomials $x_i x_j$ in R such that $\{x_i, x_j\}$ is an edge of G. Thus

$$k[G] = k[f_1, \ldots, f_q] \subset R.$$

One has the following general description of the integral closure: $\overline{k[G]}$ is a monomial subalgebra generated by monomials

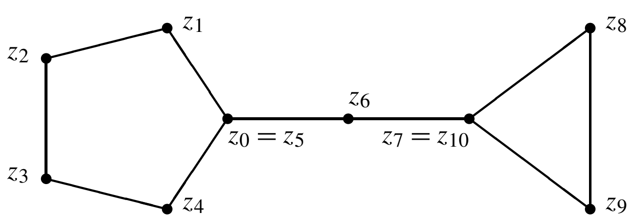
$$f = x_1^{\beta_1} \cdots x_n^{\beta_n} = \mathbf{x}^{\beta}, \quad \beta \in \mathbb{N}^n,$$

with the following two properties (see Theorem 5.14):

- $f = \prod f_i^{a_i}$, $a_i \in \mathbb{Z}$,
- $f^m = \prod f_i^{b_i}$, $m, b_i \in \mathbb{N}$ and $m \geqslant 1$.

The first condition asserts that f lies in the field of fractions of $k[G]$, and the second asserts that f is integral over $k[G]$.

To explain the description of $\overline{k[G]}$ we begin by giving some monomials in the integral closure of $k[G]$. A *bow tie* of a graph G is an induced subgraph w of G consisting of two edge disjoint odd cycles with at most one common vertex



$$Z_1 = \{z_0, z_1, \ldots, z_r = z_0\} \quad \text{and} \quad Z_2 = \{z_s, z_{s+1}, \ldots, z_t = z_s\},$$

joined by a path $\{z_r, \ldots, z_s\}$. In this case one sets $M_w = z_1 \cdots z_r z_{s+1} \cdots z_t$. We observe that Z_1 and Z_2 are allowed to intersect and that only the variables in the cycles occur in M_w, not those in the path itself.

If w is a bow tie of a graph G, as above, then M_w is in the integral closure of $k[G]$. Indeed if $f_i = z_{i-1} z_i$, then

$$z_1^2 \cdots z_r^2 = f_1 \cdots f_r, \qquad z_s^2 \cdots z_{t-1}^2 = f_{s+1} \cdots f_t, \tag{18}$$

which together with the identities

$$M_w = \prod_{i \text{ odd}} f_i \prod_{\substack{i \text{ even} \\ r < i \leqslant s}} f_i^{-1} \quad \text{and} \quad M_w^2 = f_1 \cdots f_r f_{s+1} \cdots f_t$$

gives $M_w \in \overline{k[G]}$.

THEOREM 6.11 [108]. *Let G be a graph and let k be a field. Then the integral closure $\overline{k[G]}$ of $k[G]$ is generated as a k-algebra by the set*

$$\mathcal{B} = \{f_1, \ldots, f_q\} \cup \{M_w \mid w \text{ is a bow tie}\},$$

where $f_1, \ldots, f_q$ denote the monomials defining the edges of G.

There is a version of Theorem 6.11, due to Hibi and Ohsugi [74], that allows loops in the graph G.

As an immediate consequence of Theorem 6.11 one has the following full characterization of when $k[G]$ is normal.

COROLLARY 6.12. *Let G be a connected graph. Then $k[G]$ is normal if and only if for any two edge disjoint odd cycles Z_1, Z_2 with at most one common vertex either Z_1 and Z_2 have a common vertex, or Z_1 and Z_2 are connected by an edge.*

DEFINITION 6.13. A graph G is said to satisfy the *odd cycle condition* if every two vertex disjoint odd cycles of G can be joined by an edge in G.

The *odd cycle condition* has come up in the literature in connection with the normality of edge subrings [74,107,108] and the description of the circuits of a graph [102,127]; it also occurred before [60].

6.3. *The equations of the edge cone*

Let G be a graph on the vertex set $V = \{v_1, \ldots, v_n\}$ and $R = k[x_1, \ldots, x_n]$ a polynomial ring over a field k. The *edge cone* of G is the cone $\mathbb{R}_+\mathcal{A} \subset \mathbb{R}^n$ spanned by the set $\mathcal{A}$ of all vectors $e_i + e_j$ such that v_i is adjacent to v_j, where e_i denotes the ith unit vector.

Next we present a combinatorial description of the *facets* (faces of maximal dimension) of $\mathbb{R}_+\mathcal{A}$ when $k[G]$ has dimension n.

6.3.1. *The irreducible representation of the edge cone.* The computation of the *equations* defining the facets of the edge cone of a graph G is important to determine some invariants of $k[G]$, see [129]. Here we use those equations to represent the edge cone as an intersection of halfspaces. The vertex set and edge set of a graph G will be denoted by $V(G)$ and $E(G)$ respectively.

The facets of the edge cone. Let us introduce some more terminology and fix some notation. To ease the reading we recall some of the relevant concepts in polyhedral geometry. See [18,23,131,132].

We set $\mathcal{A}_G$ (or simply $\mathcal{A}$ if G is understood) equal to the set $\{\alpha_1, \ldots, \alpha_q\}$ of row vectors of the transpose of the incidence matrix M_G of G. The *edge cone* $\mathbb{R}_+\mathcal{A}$ of G is defined as the *cone generated* by $\mathcal{A}$, that is:

$$\mathbb{R}_+\mathcal{A} = \left\{ \sum_{i=1}^{q} a_i\alpha_i \;\middle|\; a_i \in \mathbb{R}_+ \text{ for all } i \right\}.$$

If $a \in \mathbb{R}^n$, $a \neq 0$, then the set H_a will denote the *hyperplane* of $\mathbb{R}^n$ through the origin with normal vector a, that is,

$$H_a = \{x \in \mathbb{R}^n \mid \langle x, a\rangle = 0\}.$$

This hyperplane determines two *closed half-spaces*

$$H_a^+ = \{x \in \mathbb{R}^n \mid \langle x, a\rangle \geqslant 0\} \quad \text{and} \quad H_a^- = \{x \in \mathbb{R}^n \mid \langle x, a\rangle \leqslant 0\}.$$

A subset $F \subset \mathbb{R}^n$ is a *proper face* of the edge cone of G if there is a supporting hyperplane H_a such that

$$F = \mathbb{R}_+\mathcal{A} \cap H_a \neq \emptyset, \quad \mathbb{R}_+\mathcal{A} \subset H_a^-, \quad \text{and} \quad \mathbb{R}_+\mathcal{A} \not\subset H_a.$$

DEFINITION 6.14. A proper face F of the edge cone is a *facet* if

$$\dim F = \dim \mathbb{R}_+\mathcal{A} - 1.$$

Let A be an *independent set* of vertices of G, that is, no two vertices in A are adjacent. The supporting hyperplane of the edge cone defined by

$$\sum_{v_i \in A} x_i = \sum_{v_i \in N(A)} x_i$$

will be denoted by H_A. Here $N(A)$ is the *neighbor set* of A consisting of the vertices of G that are adjacent to some vertex in A.

To determine whether H_A defines a facet of $\mathbb{R}_+\mathcal{A}$ consider the subgraph $L = L_1 \cup L_2$, where L_1 is the subgraph of G with vertex set and edge set

$$V(L_1) = A \cup N(A) \quad \text{and} \quad E(L_1) = \{z \in E(G) \mid z \cap A \neq \emptyset\}$$

respectively, and $L_2 = \langle S\rangle$ is the subgraph of G induced by $S = V \setminus V(L_1)$. The vectors in $\mathcal{A} \cap H_A$ correspond precisely to the edges of L.

THEOREM 6.15 [129]. *Let G be a graph with vertex set $V = \{v_1, \ldots, v_n\}$ and M its incidence matrix. If* $\operatorname{rank}(M) = n$, *then F is a facet of $\mathbb{R}_+\mathcal{A}$ if and only if either*

(a) *$F = H_{e_i} \cap \mathbb{R}_+\mathcal{A}$ for some i, where all the connected components of $G \setminus \{v_i\}$ are nonbipartite graphs, or*

(b) *$F = H_A \cap \mathbb{R}_+\mathcal{A}$ for some independent set $A \subset V$ such that L_1 is a connected bipartite graph, and the connected components of L_2 are nonbipartite graphs.*

PROPOSITION 6.16. *If $C \neq \mathbb{R}^n$ is a polyhedral cone of dimension n, then there is a unique irreducible representation*

$$C = H_{a_1}^+ \cap \cdots \cap H_{a_r}^+, \quad \text{where } a_i \in \mathbb{R}^n \setminus \{0\}.$$

PROOF. See [132]. □

COROLLARY 6.17 [129]. *Let G be a connected nonbipartite graph with vertex set $V = \{v_1, \dots, v_n\}$. Then a vector $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is in $\mathbb{R}_+\mathcal{A}$ if and only if x is a solution of the system of linear inequalities*

$$\begin{aligned} -x_i &\leqslant 0, \quad i = 1, \dots, n \\ \sum_{v_i \in A} x_i - \sum_{v_i \in N(A)} x_i &\leqslant 0, \quad \text{for all independent sets } A \subset V. \end{aligned}$$

PROOF. Let

$$\mathbb{R}_+\mathcal{A} = H_{b_1}^- \cap \cdots \cap H_{b_m}^-$$

be the irreducible representation of the edge cone as an intersection of closed half-spaces. By Proposition 6.16 the set $H_{b_i} \cap \mathbb{R}_+\mathcal{A}$ is a facet of $\mathbb{R}_+\mathcal{A}$, and we may apply Theorem 6.15. □

PROPOSITION 6.18 [129]. *Let G be a connected nonbipartite graph with n vertices. If $k[G]$ is a normal domain, then a monomial $x_1^{\beta_1} \cdots x_n^{\beta_n}$ belongs to $k[G]$ if and only if the following two conditions hold*:

(i) *$\beta = (\beta_1, \dots, \beta_n)$ is in the edge cone of G, and*

(ii) *$\sum_{i=1}^n \beta_i$ is an even integer.*

PROOF. Set $\mathcal{A} = \mathcal{A}_G = \{\alpha_1, \dots, \alpha_q\}$. Assume $\beta \in \mathbb{R}_+\mathcal{A}$ and $\deg(x^\beta)$ even. We proceed by induction on $\deg(x^\beta)$. Using Corollary 6.7 one has the isomorphism

$$\mathbb{Z}^n/(\alpha_1, \dots, \alpha_q) \simeq \mathbb{Z}_2,$$

hence $2\beta \in \mathbb{R}_+\mathcal{A} \cap \mathbb{Z}\mathcal{A} = \mathbb{N}\mathcal{A}$ and one may write

$$2\beta = 2\sum_{i=1}^q s_i\alpha_i + \sum_{i=1}^q \varepsilon_i\alpha_i, \quad s_i \in \mathbb{N} \text{ and } \varepsilon_i \in \{0, 1\},$$

by induction one may assume $\sum_{i=1}^q s_i\alpha_i = 0$. Therefore from the equality above one concludes that the subgraph whose edges are defined by the set $\{\alpha_i \mid \varepsilon_i = 1\}$ is an edge disjoint union of cycles $Z_1, \dots, Z_r$. By induction one may further assume that all the Z_i's are odd cycles. Note $r \geqslant 2$, because $\deg(x^b)$ is even. As G is connected, using Corollary 6.12 it follows that $x^\beta \in k[G]$. The converse is clear because $k[G]$ is normal. □

Let G be a connected nonbipartite graph with $k[G]$ normal. The last results can be used to give conditions for a graph to have a perfect matching.

PROPOSITION 6.19. *Let G be a connected graph with n vertices. If n is even and $k[G]$ is normal of dimension n, then $x_1 \cdots x_n$ is in $k[G]$ if and only if $|A| \leqslant |N(A)|$ for every independent set of vertices A of G.*

PROOF. ($\Rightarrow$) Since $k[G]$ is normal:

$$\mathbb{R}_+\mathcal{A}_G \cap \mathbb{Z}\mathcal{A}_G = \mathbb{N}\mathcal{A}_G.$$

Hence $a = (1, \ldots, 1) = (a_1, \ldots, a_n)$ is in $\mathbb{R}_+\mathcal{A}_G$. Using Corollary 6.17 we get that the vector a satisfies the inequalities:

$$|A| = \sum_{v_i \in A} a_i \leqslant \sum_{v_i \in N(A)} a_i = \big|N(A)\big|$$

for every independent set of vertices A of G, as required.

($\Leftarrow$) First we use Corollary 6.17 to conclude that a is in the edge cone, then apply Proposition 6.18 to get $a \in k[G]$. □

A pairing off of all the vertices of a graph G is called a *perfect matching*. Thus a graph G has a perfect matching if and only if G has an even number of vertices and there is a set of mutually independent lines covering (containing) all the vertices of G. Two lines of a graph G are called *independent* if they do not have a common vertex.

COROLLARY 6.20 (Generalized marriage theorem). *Let G be a graph. If G is connected with an even number of vertices and satisfying the odd cycle condition, then the following are equivalent*:

(a) *G has a perfect matching.*
(b) *$|A| \leqslant |N(A)|$ for all A independent set of vertices of G.*

PROOF. The proof follows using [14, Theorem 7 of Chapter III] together with Proposition 6.19. □

Acknowledgment

The author thanks the *Consejo Nacional de Ciencia y Tecnología* for its generous support.

References

[1] W.W. Adams and P. Loustaunau, *An Introduction to Gröbner Bases*, GSM, Vol. 3, American Mathematical Society (1994).

[2] A. Alcántar, *Rees algebras, Noether normalizations and critical binomials*, PhD Thesis, CINVESTAV-IPN (2000).

[3] A. Alcántar, E. Reyes and L. Zárate, *On affine toric varieties of positive characteristic*, Aportaciones Matemáticas, Serie Comunicaciones **27** (2000), 133–140.

[4] A. Alcántar and R. Villarreal, *Critical binomials of monomial curves*, Comm. Algebra **22** (1994), 3037–3052.

[5] M.F. Atiyah and I.G. Macdonald, *Introduction to Commutative Algebra*, Addison-Wesley, Reading, MA (1969).

[6] K. Baclawski, *Cohen–Macaulay ordered sets*, J. Algebra **63** (1980), 226–258.

[7] M. Barile, M. Morales and A. Thoma, *On simplicial toric varieties which are set-theoretic complete intersections*, J. Algebra **226** (2000), 880–892.

[8] H. Bass, *On the ubiquity of Gorenstein rings*, Math. Z. **82** (1963), 8–28.

[9] D. Bayer, S. Popescu and B. Sturmfels, *Syzygies of unimodular Lawrence ideals*, J. Reine Angew. Math. **534** (2001), 169–186.

[10] D. Bayer and M. Stillman, *Computation of Hilbert functions*, J. Symbolic Comput. **14** (1992), 31–50.

[11] C. Berge, *Hypergraphs Combinatorics of Finite Sets*, Mathematical Library, Vol. 45, North-Holland, Amsterdam (1989).

[12] A. Bigatti, *Computation of Hilbert–Poincaré series*, J. Pure Appl. Algebra **119** (1997), 237–253.

[13] L.J. Billera and C.W. Lee, *Sufficiency of McMullen's condition for f-vectors of simplicial polytopes*, Bull. Amer. Math. Soc. **2** (1980), 181–185.

[14] B. Bollobás, *Modern Graph Theory*, Graduate Texts in Math., Vol. 184, Springer-Verlag, New York (1998).

[15] V. Bonanzinga, C. Escobar and R. Villarreal, *On the normality of Rees algebras associated to totally unimodular matrices*, Results Math. **41** (3/4) (2002), 258–264.

[16] L.A. Borisov, *String cohomology of a toroidal singularity*, Preprint alg-geom/9802052 (1998).

[17] H. Bresinsky, *Monomial space curves in $\mathbb{A}^3$ as set-theoretic complete intersections*, Proc. Amer. Math. Soc. **75** (1979), 23–24.

[18] A. Brøndsted, *Introduction to Convex Polytopes*, Graduate Texts in Math., Vol. 90, Springer-Verlag (1983).

[19] P. Brumatti, A. Simis and W.V. Vasconcelos, *Normal Rees algebras*, J. Algebra **112** (1988), 26–48.

[20] W. Bruns and J. Gubeladze, *Normality and covering properties of affine semigroups*, J. Reine Angew. Math. **510** (1999), 161–178.

[21] W. Bruns, J. Gubeladze and N.V. Trung, *Normal polytopes, triangulations, and Koszul algebras*, J. Reine Angew. Math. **481** (1997), 123–160.

[22] W. Bruns and J. Herzog, *Semigroup rings and simplicial complexes*, J. Pure Appl. Algebra **122** (1997), 185–208.

[23] W. Bruns and J. Herzog, *Cohen–Macaulay Rings*, rev. edn, Cambridge University Press, Cambridge (1997).

[24] W. Bruns and T. Hibi, *Cohen–Macaulay partially ordered sets with pure resolutions*, Eur. J. Comb. **19** (1998), 779–785.

[25] W. Bruns and R. Koch, *Normaliz, a program to compute normalizations of affine semigroups*, Available via anonymous ftp from ftp.mathematik.Uni-Osnabrueck.DE/pub/osm/kommalg/software.

[26] W. Bruns and R. Koch, *Computing the integral closure of an affine semigroup*, Effective Methods in Algebraic and Analytic Geometry, 2000, Kraków, Univ. Iagel. Acta Math. **39** (2001), 59–70.

[27] W. Bruns, W.V. Vasconcelos and R. Villarreal, *Degree bounds in monomial subrings*, Illinois J. Math. **41** (1997), 341–353.

[28] B. Buchberger, *An algorithmic method in polynomial ideal theory*, Recent Trends in Mathematical Systems Theory, N.K. Bose, ed., Reidel, Dordrecht (1985), 184–232.

[29] B. Büeler and A. Enge, *VINCI: An implementation of algorithms for computing the volume of a polytope of full dimension*, Available via ftp from ftp.ifor.math.ethz.ch/pub/volume/.

[30] B. Büeler, A. Enge and K. Fukada, *Exact volume computation for convex polytopes: a practical study*, Polytopes – Combinatorics and Computation (Oberwolfach, 1997), DMV Sem., 29, Birkhäuser, Basel (2000), 131–154.

[31] A. Campillo and P. Gimenez, *Syzygies of affine toric varieties*, J. Algebra **225** (2000), 142–161.

[32] A. Capani, G. Niesi and L. Robbiano, *CoCoA: A system for doing computations in commutative algebra*, Available via anonymous ftp from lancelot.dima.unige.it (1998).

[33] V. Chvátal, *Linear Programming*, W.H. Freeman and Company, New York (1983).

[34] A. Conca and E. De Negri, *M-sequences, graph ideals, and ladder ideals of linear type*, J. Algebra **211** (1999), 599–624.

[35] R. Cowsik and M. Nori, *Affine curves in characteristic p are set-theoretic complete intersections*, Invent. Math. **45** (1978), 111–114.

[36] D. Cox, J. Little and D. O'Shea, *Ideals, Varieties, and Algorithms*, Springer-Verlag (1992).

[37] V.I. Danilov, *The geometry of toric varieties*, Russian Math. Surveys **33** (1978), 97–154.

[38] W. Decker, G. Greuel and G. Pfister, *Primary decomposition; algorithms and comparisons*, Algorithmic Algebra and Number Theory (Heidelberg, 1997), Springer-Verlag, Berlin (1999) 187–220.

[39] T. De Jong, *An algorithm for computing the integral closure*, J. Symbolic Comput. **26** (1998), 273–277.

[40] D. Delfino, A. Taylor, W.V. Vasconcelos, R.H. Villarreal and N. Weininger, *Monomial ideals and the computation of multiplicities*, Commutative Rings Theory and Applications, Proceedings of the Fez Conference, M. Fontana et al., eds, Lecture Notes Pure Appl. Math., Vol. 231, Marcel Dekker, New York (2003), 87–106.

[41] J.A. de Loera, B. Sturmfels and R.R. Thomas, *Gröbner bases and triangulations of the second hypersimplex*, Combinatorica **15** (1995), 409–424.

[42] E. De Negri, *Toric rings generated by special stable sets of monomials*, Math. Nachr. **203** (1999), 31–45.

[43] L. Doering and T. Gunston, *Algebras arising from planar bipartite graphs*, Comm. Algebra **24** (1996), 3589–3598.

[44] D. Eisenbud, *Commutative Algebra with a View Toward Algebraic Geometry*, Graduate Texts in Math., Vol. 150, Springer-Verlag (1994).

[45] D. Eisenbud, C. Huneke and W.V. Vasconcelos, *Direct methods for primary decomposition*, Invent. Math. **110** (1992), 207–235.

[46] D. Eisenbud and B. Sturmfels, *Binomial ideals*, Duke Math. J. **84** (1996), 1–45.

[47] S. Eliahou, *Courbes monomiales et algèbre de Rees symbolique*, PhD Thesis, Université de Genève (1983).

[48] S. Eliahou, *Idéaux de définition des courbes monomiales*, Complete Intersections, S. Greco and R. Strano, eds, Lecture Notes in Math., Vol. 1092, Springer-Verlag, Heidelberg (1984), 229–240.

[49] S. Eliahou and R. Villarreal, *On systems of binomials in the ideal of a toric variety*, Proc. Amer. Math. Soc. **130** (2002), 345–351.

[50] C. Escobar, I. Gitler and R. Villarreal, *A note on unimodular matrices and the normality of certain Rees semigroups*, Preprint.

[51] C. Escobar, J. Martínez-Bernal and R. Villarreal, *A comparison of Ehrhart rings with some other monomial subrings*, Reporte Técnico No. 303, CINVESTAV-IPN (2001).

[52] M. Estrada and R. Villarreal, *Cohen–Macaulay bipartite graphs*, Arch. Math. **68** (1997), 124–128.

[53] E.G. Evans and P. Griffith, *Syzygies*, London Math. Soc., Lecture Note Series, Vol. 106, Cambridge University Press, Cambridge (1985).

[54] G. Ewald, *Combinatorial Convexity and Algebraic Geometry*, Graduate Texts in Math., Vol. 168, Springer-Verlag, New York (1996).

[55] S. Faridi, *Normal ideals of graded rings*, Comm. Algebra **28** (2000), 1971–1977.

[56] R. Fröberg, *A study of graded extremal rings and of monomial rings*, Math. Scand. **51** (1982), 22–34.

[57] R. Fröberg, *Rings with monomial relations having linear resolutions*, J. Pure Appl. Algebra **38** (1985), 235–241.

[58] R. Fröberg, *A note on the Stanley–Reisner ring of a join and of a suspension*, Manuscripta Math. **60** (1988), 89–91.

[59] R. Fröberg, *On Stanley–Reisner rings*, Topics in Algebra, Part 2, S. Balcerzyk et al., eds, Polish Scientific Publishers (1990), 57–70.

[60] D.R. Fulkerson, A.J. Hoffman and M.H. McAndrew, *Some properties of graphs with multiple edges*, Canad. J. Math. **17** (1965), 166–177.

[61] W. Fulton, *Algebraic Curves; An Introduction to Algebraic Geometry*, W.A. Benjamin, New York (1969).

[62] W. Fulton, *Introduction to Toric Varieties*, Princeton University Press (1993).

[63] A. Garsia, *Combinatorial methods in the theory of Cohen–Macaulay rings*, Adv. Math. **38** (1980), 229–266.

[64] P. Gianni, B. Trager and G. Zacharias, *Gröbner bases and primary decomposition of polynomial ideals*, J. Symbolic Comput. **6** (1988), 149–167.

[65] D. Grayson and M. Stillman, *Macaulay*2, Available via anonymous ftp from math.uiuc.edu (1996).
[66] J. Grossman, D.M. Kulkarni and I. Schochetman, *On the minors of an incidence matrix and its Smith normal form*, Linear Algebra Appl. **218** (1995), 213–224.
[67] A. Grothendieck, *Local Cohomology*, Lecture Notes in Math., Vol. 41, Springer-Verlag, Berlin (1967).
[68] B. Grünbaum, *Diagrams and Schlegel diagrams*, Abstract 625-112, Notices Amer. Math. Soc. **12** (1965), 578.
[69] B. Grünbaum, *Convex Polytopes*, J. Wiley (1967).
[70] F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA (1972).
[71] J. Herzog, T. Hibi and H. Ohsugi, *Combinatorial pure subrings*, Osaka J. Math. **37** (2000), 745–757.
[72] J. Herzog, A. Simis and W.V. Vasconcelos, *Arithmetic of normal Rees algebras*, J. Algebra **143** (1991), 269–294.
[73] T. Hibi, *Algebraic Combinatorics on Convex Polytopes*, Carslaw Publications, Australia (1992).
[74] T. Hibi and H. Ohsugi, *Normal polytopes arising from finite graphs*, J. Algebra **207** (1998), 409–426.
[75] D. Hilbert, *Über die Theorie von algebraischen Formen*, Math. Ann. **36** (1890), 473–534.
[76] P.J. Hilton and U. Stammbach, *A Course in Homological Algebra*, GTM, Vol. 4, Springer-Verlag (1971).
[77] M. Hochster, *Rings of invariants of tori, Cohen–Macaulay rings generated by monomials, and polytopes*, Ann. of Math. **96** (1972), 318–337.
[78] M. Hochster, *Cohen–Macaulay rings, combinatorics and simplicial complexes*, Ring Theory II, Lecture Notes in Pure and Applied Math., Vol. 26, Dekker, New York (1977), 171–223.
[79] S. Hoşten and G.G. Smith, *Monomial ideals*, Computations in Algebraic Geometry with Macaulay 2, D. Eisenbud, et al., eds, Preprint.
[80] C. Huneke, *Hyman Bass and ubiquity: Gorenstein rings*, Contemp. Math. **243** (1999), 55–78.
[81] N. Jacobson, *Basic Algebra I*, 2nd edn, W. H. Freeman and Company, New York (1996).
[82] I. Kaplansky, *Commutative Rings*, University of Chicago Press (1974).
[83] J. Kruskal, *The number of simplices in a complex*, Mathematical Optimization Techniques, R. Bellman, ed., University of California Press (1963), 251–278.
[84] G. Lyubeznik, *The minimal non-Cohen–Macaulay monomial ideals*, J. Pure Appl. Algebra **51** (1988), 261–266.
[85] G. Lyubeznik, *A survey of problems and results on the number of defining equations*, Commutative Algebra, M. Hochster et al., eds, MSRI Publications, Vol. 15, Springer-Verlag (1989), 375–390.
[86] H. Matsumura, *Commutative Algebra*, Benjamin-Cummings, Reading, MA (1980).
[87] H. Matsumura, *Commutative Ring Theory*, Cambridge Stud. Adv. Math., Vol. 8, Cambridge University Press (1986).
[88] P. McMullen, *The maximum numbers of faces of a convex polytope*, Mathematika **17** (1970), 179–184.
[89] P. McMullen and G. Shephard, *Convex Polytopes and the Upper Bound Conjecture*, LMS Lect. Note Ser., Vol. 3, Cambridge University Press (1971).
[90] M. Miller and R. Villarreal, *A note on generators of least degree in Gorenstein ideals*, Proc. Amer. Math. Soc. **124** (1996), 377–382.
[91] T.T. Moh, *Set-theoretic complete intersections*, Proc. Amer. Math. Soc. **94** (1985), 217–220.
[92] T.S. Motzkin, *Comonotone curves and polyhedra, Abstract*, Bull. Amer. Math. Soc. **63** (1957), 35.
[93] J.R. Munkres, *Elements of Algebraic Topology*, Addison-Wesley (1984).
[94] M. Newman, *Integral Matrices*, Pure Appl. Math., Vol. 45, Academic Press, New York (1972).
[95] J. Oxley, *Matroid Theory*, Oxford University Press, Oxford (1992).
[96] S. Popescu, *Monom.m2 – a series of routines for monomial ideals*, http://www.math.columbia.edu/~psorin/m2/monom.m2.
[97] G. Reisner, *Cohen–Macaulay quotients of polynomial rings*, Adv. Math. **21** (1976), 31–49.
[98] C. Rentería and R. Villarreal, *Koszul homology of Cohen–Macaulay rings with linear resolutions*, Proc. Amer. Math. Soc. **115** (1992), 51–58.
[99] G. Restuccia and R. Villarreal, *On the normality of monomial ideals of mixed products*, Comm. Algebra **29** (8) (2001), 3571–3580.
[100] E. Reyes, R. Villarreal and L. Zárate, *A note on affine toric varieties*, Linear Algebra Appl. **318** (2000), 173–179.
[101] R.T. Rockafellar, *The elementary vectors of a subspace of* R^N, Combinatorial Mathematics and its Applications, Proc. Chapel Hill Conf., Univ. North Carolina Press (1969), 104–127.

[102] R.T. Rockafellar, *Convex Analysis*, Princeton University Press (1970).
[103] J.J. Rotman, *An Introduction to Homological Algebra*, Academic Press (1979).
[104] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley, New York (1986).
[105] T. Shimoyama and K. Yokoyama, *Localization and primary decomposition of polynomial ideals*, J. Symbolic Comput. **22** (1996), 247–277.
[106] A. Simis, *Topics in Rees algebras of special ideals*, Commutative Algebra, Salvador, W. Bruns and A. Simis, eds, Lecture Notes in Math., Vol. 1430, Springer-Verlag (1990), 98–114.
[107] A. Simis, W.V. Vasconcelos and R. Villarreal, *On the ideal theory of graphs*, J. Algebra **167** (1994), 389–416.
[108] A. Simis, W.V. Vasconcelos and R. Villarreal, *The integral closure of subrings associated to graphs*, J. Algebra **199** (1998), 281–289.
[109] D.E. Smith, *On the Cohen–Macaulay property in commutative algebra and simplicial topology*, Pacific J. Math. **141** (1990), 165–196.
[110] R. Stanley, *Cohen–Macaulay rings and constructible polytopes*, Bull. Amer. Math. Soc. **81** (1975), 133–135.
[111] R. Stanley, *The upper bound conjecture and Cohen–Macaulay rings*, Stud. Appl. Math. **54** (1975), 135–142.
[112] R. Stanley, *Hilbert functions of graded algebras*, Adv. Math. **28** (1978), 57–83.
[113] R. Stanley, *Combinatorics and Commutative Algebra*, Birkhäuser, Boston (1983).
[114] R. Stanley, *The number of faces of simplicial polytopes and spheres*, Discrete Geometry and Convexity, J.E. Goodman et al., eds, Ann. New York Acad. Sci., Vol. 440 (1985), 212–223.
[115] R. Stanley, *Enumerative Combinatorics I*, Wadsworth-Brooks/Cole, Monterey, CA (1986).
[116] B. Sturmfels, *Gröbner bases of toric varieties*, Tôhoku Math. J. **43** (1991), 249–261.
[117] B. Sturmfels, *Gröbner Bases and Convex Polytopes*, University Lecture Series, Vol. 8, American Mathematical Society, RI (1996).
[118] Y. Takayama, *Normality of affine semigroup rings generated by* 2*-dimensional cone type simplicial complexes*, Comm. Algebra **29** (2001), 1499–1512.
[119] A. Thoma, *On the set-theoretic complete intersection problem for monomial curves in $\mathbb{A}^n$ and $\mathbb{P}^n$*, J. Pure Appl. Algebra **104** (1995), 333–344.
[120] A. Thoma, *On the binomial arithmetical rank*, Arch. Math **74** (2000), 22–25.
[121] K. Truemper, *Matroid Decomposition*, Academic Press (1992).
[122] W.V. Vasconcelos, *Computing the integral closure of an affine domain*, Proc. Amer. Math. Soc. **113** (1991), 633–638.
[123] W.V. Vasconcelos, *Arithmetic of Blowup Algebras*, London Math. Soc., Lecture Note Series, Vol. 195, Cambridge University Press, Cambridge (1994).
[124] W.V. Vasconcelos, *Computational Methods in Commutative Algebra and Algebraic Geometry*, Springer-Verlag (1998).
[125] R. Villarreal, *Cohen–Macaulay graphs*, Manuscripta Math. **66** (1990), 277–293.
[126] R. Villarreal, *Monomial algebras*, Aportaciones Matemáticas, Serie Investigación **6** (1992), 69–75.
[127] R. Villarreal, *Rees algebras of edge ideals*, Comm. Algebra **23** (1995), 3513–3524.
[128] R. Villarreal, *Normality of subrings generated by square free monomials*, J. Pure Appl. Algebra **113** (1996), 91–106.
[129] R. Villarreal, *On the equations of the edge cone of a graph and some applications*, Manuscripta Math. **97** (1998), 309–317.
[130] R. Villarreal, *Rees algebras of complete bipartite graphs*, XV Escola de Álgebra, Canela, Rio Grande do Sul, M. Ferrero and A.M.S. Doering, eds, Sociedad Brasileira de Matemática, Mat. Contemp. **16** (1999), 281–289.
[131] R. Villarreal, *Monomial Algebras*, Monographs and Textbooks in Pure and Applied Mathematics, Vol. 238, Marcel Dekker, New York (2001).
[132] R. Webster, *Convexity*, Oxford University Press, Oxford (1994).
[133] G.M. Ziegler, *Lectures on Polytopes*, Graduate Texts in Math., Vol. 152, Springer-Verlag (1994).

Section 3B
Associative Rings and Algebras

This Page Intentionally Left Blank

Whitehead Modules

Paul C. Eklof

Department of Mathematics, University of California at Irvine, CA 92697-3875, USA
E-mail: peklof@math.uci.edu

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

0. Introduction

Throughout, R will denote an associative ring with 1, and "module" will mean left R-module. $\mathbb{Z}$ will denote the ring of integers. The word "group" will always mean Abelian group (i.e. $\mathbb{Z}$-module). For any cardinal κ and any module M, M^κ (resp. $M^{(\kappa)}$) denotes the direct product (resp. direct sum) of κ copies of M. We will write $\operatorname{Ext}(M, N)$ for $\operatorname{Ext}^1_R(M, N)$. See, for example, [8,63] or [95] for basic information on Ext. We recall that $\operatorname{Ext}(\bigoplus_{i\in I} M_i, N) \cong \prod_{i\in I} \operatorname{Ext}(M_i, N)$ and $\operatorname{Ext}(M, \prod_{i\in I} N_i) \cong \prod_{i\in I} \operatorname{Ext}(M, N_i)$.

A *Whitehead group* is an Abelian group M such that $\operatorname{Ext}(M, \mathbb{Z}) = 0$. More generally, a module M is called a *Whitehead module* if $\operatorname{Ext}(M, R) = 0$. This is equivalent to any one, or all, of the following conditions:

- Every short exact sequence $0 \to R \to L \xrightarrow{\pi} M \to 0$ splits, i.e. there is a homomorphism $\rho: M \to L$ such that $\pi \circ \rho = id_M$.
- For any (or all) projective presentations $0 \to K \to F \to M \to 0$ (where F is projective), the induced homomorphism $\operatorname{Hom}(F, R) \to \operatorname{Hom}(K, R)$ is surjective.
- If E is the injective envelope of R, every homomorphism of M into E/R can be lifted to a homomorphism from M into E.

It is easy to see that every projective module is a Whitehead module. The Whitehead problem asks whether the converse is true. According to Ehrenfeucht [18], J.H.C. Whitehead raised this problem for Abelian groups during a visit to Warsaw in May 1952. Ehrenfeucht showed that the answer was 'yes' for countable $\mathbb{Z}$-modules; independently Stein [115] had already proved this in 1951. (See [89] for more on the history and significance of the Whitehead problem.)

For uncountable $\mathbb{Z}$-modules, partial results were obtained by Rotman [94], Nunke [88], Griffith [52], and Chase [9,11] before 1973 when Shelah [102] proved that the problem was undecidable in ordinary set theory (Zermelo–Frankel set theory with the axiom of choice, denoted ZFC). He later proved [104,106] that the problem is undecidable even assuming the Generalized Continuum Hypothesis (GCH) – that $2^{\aleph_\alpha} = \aleph_{\alpha+1}$ for every ordinal α.

The Whitehead problem can be generalized, not only to rings other than $\mathbb{Z}$ but also by placing modules other than R in the second argument of Ext. We will make use of some non-standard terminology and call a module M a W_κ-module if $\operatorname{Ext}(M, R^{(\kappa)}) = 0$. For any *finite* κ, a W_κ-module is the same as a Whitehead module (which we will also call a W-module). If $\lambda \geqslant \kappa$ then every W_λ-module is a W_κ-module. If R is hereditary and M is a W_κ-module, then $\operatorname{Ext}(M, N) = 0$ for every module N which is generated by a subset of size κ.

The generalized Whitehead problem (for κ) asks whether every W_κ-module is projective. Also, one may ask, for a given ring R, what are the *test modules* (if any) *for projectivity*, that is, the modules N such that for every module M, $\operatorname{Ext}(M, N) = 0$ implies that M is projective (cf. [121]). Another form of the problem asks for a characterization of the *Baer modules* over a domain R, that is the modules M such that $\operatorname{Ext}(M, T) = 0$ for all torsion modules T.

In this article we shall survey the known answers to these different forms of the Whitehead problem. We begin in Section 1 with the 'classical' – that is pre-Shelah – results; the next sections are organized not historically, but according to the tools used. Section 2 discusses stationary sets and their role in analyzing whether modules are Whitehead or projective. Sections 3 and 4 deal with tools for proving that Whitehead modules are projective. Section 5 reviews the most important method of constructing counterexamples, that is non-projective modules M such that (for a given N), $\mathrm{Ext}(M, N) = 0$. (The methods of Sections 3 and 5 require hypotheses stronger than the axioms of ZFC; these hypotheses are mutually contradictory, but separately consistent with ZFC.) Section 6 deals with Baer modules. In Sections 7 and 8 we restrict to $\mathbb{Z}$-modules; in Section 7 we consider splitters, that is, groups A such that $\mathrm{Ext}(A, A) = 0$; in Section 8 we study the structure of $\mathrm{Ext}(A, \mathbb{Z})$ as an Abelian group. Finally, in Section 9, we define the notion of a slender module, and prove some basic facts, including the fact that Whitehead $\mathbb{Z}$-modules are slender.

1. Classical results

If R is (left) hereditary, then a submodule of a Whitehead module is a Whitehead module. In particular, for $R = \mathbb{Z}$ (or any p.i.d.), this implies that a Whitehead group is torsion-free, because otherwise it contains a subgroup isomorphic to $\mathbb{Z}/p\mathbb{Z}$ for some prime p, but $\mathrm{Ext}(\mathbb{Z}/p\mathbb{Z}, \mathbb{Z}) \neq 0$ since $0 \to \mathbb{Z} \xrightarrow{p} \mathbb{Z} \to \mathbb{Z}/p\mathbb{Z} \to 0$ is not split.

To prove that countable Whitehead groups are free, we use a characterization known as Pontryagin's criterion. We recall that the *rank* of an Abelian group is the size of a maximal linearly independent subset. Also, a subgroup B of an Abelian group A is a *pure* subgroup if for all integers $n \in \mathbb{Z}$, $nB \supseteq nA \cap B$; if A is torsion-free, this is equivalent to saying that A/B is torsion-free.

LEMMA 1.1 [91]. *If A is a countable torsion-free Abelian group such that every subgroup of finite rank is free, then A is free.*

PROOF. Write $A = \{a_n \colon n \in \omega\}$ and let B^m be the subgroup of A generated by $\{a_n \colon n < m\}$ and $B^m_* = \{a \in A \colon ka \in B_m \text{ for some } k \neq 0\}$. Then B^m_* is a finite rank subgroup of A, hence free, and thus finitely-generated. Therefore B^{m+1}_*/B^m_* is finitely-generated; moreover it is torsion-free because B^m_* is pure in A; so by the Fundamental Theorem of finitely-generated Abelian groups, B^{m+1}_*/B^m_* is free. Since this is true for all $m \in \omega$, and $A = \bigcup_{m \in w} B^m_*$, it follows that A is free. □

We can now prove what is sometimes called the Theorem of Stein–Serre (cf. [63, §III.6]).

THEOREM 1.2. *Every countable Whitehead group is free.*

PROOF. (Following [94]) By Lemma 1.1, it suffices to show that any Whitehead group A of finite rank is free; this we do by induction on the rank, n, of A. First suppose $n = 1$. If A is not free then A contains a copy, Z', of $\mathbb{Z}$ such that A/Z' is torsion and not finitely-generated. The exact sequence $0 \to Z' \to A \to A/Z' \to 0$ induces an exact sequence

$$0 = \mathrm{Hom}(A, \mathbb{Z}) \to \mathrm{Hom}(Z', \mathbb{Z}) \to \mathrm{Ext}(A/Z', \mathbb{Z}) \to \mathrm{Ext}(A, \mathbb{Z}) = 0.$$

The first term is zero since A has rank 1 but is not free. Therefore $\mathrm{Ext}(A/Z', \mathbb{Z})$ is isomorphic to $\mathrm{Hom}(Z', \mathbb{Z}) \cong \mathbb{Z}$. Thus A/Z' must be indecomposable. The only indecomposable torsion groups are the groups $Z(p^\infty)$, the p^n-th complex roots of unity under multiplication, where p is a prime. But if $A/Z' \cong Z(p^\infty)$, then for any other prime q, multiplication by q is an automorphism of A/Z' and hence also an automorphism of $\mathrm{Ext}(A/Z', \mathbb{Z})$ and thus of $\mathbb{Z}$, a contradiction.

Now suppose $n > 1$ and choose a pure subgroup H of A of rank $n-1$, so that A/H has rank 1. By induction H is free, so it is enough to show that A/H is a W-group. But we have an exact sequence

$$\mathrm{Hom}(H, \mathbb{Z}) \to \mathrm{Ext}(A/H, \mathbb{Z}) \to \mathrm{Ext}(A, \mathbb{Z}) = 0$$

from which it follows that $\mathrm{Ext}(A/H, \mathbb{Z})$ is finitely-generated. However, since A/H is torsion-free, $\mathrm{Ext}(A/H, \mathbb{Z})$ is divisible (since for every $n \in \mathbb{Z}$, the injection $A/H \to A/H$ which is multiplication by n induces a surjection $\mathrm{Ext}(A/H, \mathbb{Z}) \to \mathrm{Ext}(A/H, \mathbb{Z})$ which is also multiplication by n). Therefore $\mathrm{Ext}(A/H, \mathbb{Z})$ must be zero. $\square$

This result extends to Whitehead modules of countable rank over any p.i.d. R which is not a complete discrete valuation ring. (See [42]. If R is a complete discrete valuation ring, then every torsion-free R-module is a Whitehead module, because R is pure-injective as a module over itself.)

An Abelian group is called *κ-free* if every subgroup of cardinality $< \kappa$ is free. Since a subgroup of a Whitehead group is Whitehead, we have:

COROLLARY 1.3. *Every Whitehead group is $\aleph_1$-free.*

This result was strengthened in several ways. An Abelian group A is called *separable* (or *locally free*) if every finite subset is contained in a free summand of A. By Pontryagin's Criterion, every separable group is $\aleph_1$-free. Rotman [94] proved:

THEOREM 1.4. *Every Whitehead group is separable.*

Since subgroups of Whitehead groups are Whitehead, it follows that every Whitehead group is *hereditarily separable*, that is, every subgroup is separable.

Griffith [52] introduced the notion of *κ-coseparable*: A is κ-coseparable if and only if every subgroup B of A with the property that A/B is generated by fewer than κ elements contains a direct summand D of A such that A/D is generated by fewer than κ elements. With the help of work of Chase [9], he showed:

THEOREM 1.5. *Let A be an $\aleph_1$-free group.*

(i) $\mathrm{Ext}(A, \mathbb{Z})$ *is torsion-free if and only if A is separable and $\aleph_0$-coseparable. In particular, Whitehead groups are $\aleph_0$-coseparable.*

(ii) *A is a $W_{\aleph_0}$-group if and only if A is $\aleph_1$-coseparable.*

Nunke and Rotman independently proved that every Whitehead group is *slender*: see Section 9.

An Abelian group A is called *strongly* $\aleph_1$*-free* if every countable subset of A is contained in a countable free subgroup B such that A/B is $\aleph_1$-free. In one of the deepest results on Whitehead's Problem prior to Shelah's work, Chase [11] proved the following:

THEOREM 1.6. *Assuming that* $2^{\aleph_0} < 2^{\aleph_1}$, *if* $\operatorname{Ext}(A, \mathbb{Z})$ *is torsion, then A is strongly* $\aleph_1$*-free. In particular, Whitehead groups are strongly* $\aleph_1$*-free.*

For an elementary proof, see [65]. A strengthening of this result is given in Section 3 (Theorem 3.7). Chase's theorem requires the assumption on cardinal arithmetic: Shelah showed ([102,105]; see also [22, Chap. 8]) that it is consistent with ZFC that there are Whitehead groups of cardinality $\aleph_1$ which are not strongly $\aleph_1$-free.

2. Stationary sets and Γ-invariants

We begin by reviewing some basic notions from set theory which will be used in this and succeeding sections. (See, for example, [76] or [70] for more information.) An ordinal number γ will always be regarded as the set of its predecessors: $\gamma = \{\nu\colon \nu < \gamma\} = \{\nu\colon \nu \in \gamma\}$. A cardinal number is an initial ordinal, that is, one which is not equal in size to (cannot be put in one–one correspondence with) any predecessor. An ordinal is a *limit ordinal* if has no immediate predecessor, or, equivalently, is not of the form $\mu + 1$ $(= \mu \cup \{\mu\})$; an ordinal of the form $\mu + 1$ is called a *successor ordinal*. An infinite cardinal is always a limit *ordinal*, but it may be a successor *cardinal*, that is, have an immediate predecessor as cardinal, or equivalently, equal $\aleph_{\alpha+1}$ for some ordinal α; we will write λ^+ for $\aleph_{\alpha+1}$ if $\lambda = \aleph_\alpha$.

The cofinality of a limit ordinal γ, denoted $\operatorname{cof}(\gamma)$, is the size of the smallest subset of γ whose supremum is γ. An infinite cardinal κ is called *regular* if $\operatorname{cof}(\kappa) = \kappa$ and *singular* otherwise. Every successor cardinal is regular. The first infinite cardinal is $\aleph_0$; it is a regular limit cardinal. The next limit cardinal is $\aleph_\omega$; it is singular since $\operatorname{cof}(\aleph_\omega) = \aleph_0$. It is not possible to prove in ZFC that there are regular limit cardinals larger than $\aleph_0$. In fact, it is impossible to prove that the assumption of their existence is relatively consistent with ZFC but, as far as we know, it is consistent to assume that they exist; they are called (weakly) *inaccessible* cardinals.

Let γ be a limit ordinal; a subset C of γ is called a *club* (closed unbounded subset) in γ if C is unbounded, i.e. $\sup C = \gamma$ and C is closed, i.e. for every subset X of C, $\sup X \in C \cup \{\gamma\}$. A subset S of γ is called *stationary* in γ if for every club C in γ, $S \cap C \neq \emptyset$. Every club, or set containing a club, is stationary because the intersection of two clubs is also a club (cf. Lemma 2.2 below). Assuming the axiom of choice, there are many other stationary sets. In fact, for every regular cardinal κ there are 2^κ stationary subsets $\{S_\nu\colon \nu < 2^\kappa\}$ of κ such that for any $\mu \neq \nu$ and any club C in κ, $C \cap S_\mu \neq C \cap S_\nu$.

An important fact about stationary sets (perhaps explaining their name) is the following, known as Fodor's lemma or the pressing-down lemma.

LEMMA 2.1. *Let S be a stationary subset of a regular uncountable cardinal κ and $f : S \to \kappa$ such that $f(\nu) < \nu$ for all $\nu \in S$. Then for some $\gamma \in \kappa$, $\{\nu \in S\colon f(\nu) = \gamma\}$ is stationary in κ.*

It is a useful analogy to think of clubs as sets of measure 1 and stationary sets as sets of non-zero measure. In fact, this analogy can be made precise. If κ is a regular uncountable cardinal and S is a subset of κ, let $\widetilde{S} = \{S \cap C\colon\ C \text{ is a club in } \kappa\}$. Let $D(\kappa) = \{\widetilde{S}\colon\ S \subseteq \kappa\}$. Define $\widetilde{S_1} \leqslant \widetilde{S_2}$ iff $S_1 \cap C \subseteq S_2 \cap C$ for some club C. Then $D(\kappa)$ with this ordering is a Boolean algebra with least element $0 = \tilde{\emptyset} = \{S \subseteq \kappa\colon\ S \text{ is not stationary in } \kappa\}$ and greatest element $1 = \tilde{\gamma} = \{S \subseteq \kappa\colon\ S \text{ contains a club in } \kappa\}$. Moreover, the "$D(\kappa)$-valued measure" defined by $\mu(S) = \widetilde{S}$ is $<\kappa$-additive. (See [29, §II.4] for more on this.)

Given a regular uncountable cardinal κ, a module N, and a module M generated by κ elements, we shall assign to M an element of $D(\kappa)$, denoted $\Gamma_{\kappa,N}(M)$, which is an invariant of M (i.e. depends only on the isomorphism type of M) and measures how far $\operatorname{Ext}(M, N)$ is from being zero. We shall also define another invariant of M, denoted $\Gamma_\kappa(M)$, which measures how far M is from being projective. The first step in defining these invariants is to choose a *κ-filtration* of M, that is a continuous chain $\{M_\alpha\colon\ \alpha < \kappa\}$ of submodules of M each of which is generated by fewer than κ elements. Here "chain" means that $\alpha < \beta < \kappa$ implies that $M_\alpha \subseteq M_\beta$ and "continuous" means that for every limit ordinal $\sigma < \kappa$, $M_\sigma = \bigcup_{\alpha<\sigma} M_\alpha$. Clearly M has κ-filtrations, in fact many of them; but any two agree on a "set of measure 1":

LEMMA 2.2. *If κ is a regular uncountable cardinal and $\{M_\alpha\colon\ \alpha < \kappa\}$ and $\{M'_\alpha\colon\ \alpha < \kappa\}$ are κ-filtrations of M, then $\{\alpha < \kappa\colon\ M_\alpha = M'_\alpha\}$ is a club in κ.*

PROOF. Let $C = \{\alpha < \kappa\colon\ M_\alpha = M'_\alpha\}$. Using the continuity of the filtrations, it is easy to check that C is closed. To see that C is unbounded, given any $\gamma < \kappa$, inductively define a sequence $\gamma = \alpha_0 < \alpha_1 < \cdots$ such that $M_{\alpha_{2n}} \subseteq M'_{\alpha_{2n+1}} \subseteq M_{\alpha_{2n+2}}$ for all $n \in \omega$. (Here we use the regularity of κ and the small size of the members of a κ-filtration.) Then $\sup\{\alpha_n\colon\ n \in \omega\} \in C$. $\square$

For simplicity assume that R is hereditary. (For the right definitions in general see [23] or [29, §IV.1].) Given a κ-filtration $\{M_\alpha\colon\ \alpha < \kappa\}$ of M, define

$$E_\kappa(M) = \{\alpha < \kappa\colon\ \exists\beta > \alpha \text{ s.t. } M_\beta/M_\alpha \text{ is not projective}\}$$

and

$$E_{\kappa,N}(M) = \bigl\{\alpha < \kappa\colon\ \exists\beta > \alpha \text{ s.t. } \operatorname{Ext}(M_\beta/M_\alpha, N) \neq 0\bigr\}.$$

These sets depend on the chosen κ-filtration, but $\Gamma_\kappa(M) = \widetilde{E_\kappa(M)}$ and $\Gamma_{\kappa,N}(M) = \widetilde{E_{\kappa,N}(M)}$ can be shown, using Lemma 2.2, to be invariants of M. Note, in particular, that $\Gamma_{\aleph_1,\mathbb{Z}}(M) = \Gamma_{\aleph_1}(M)$ for any Abelian group M, since countable Whitehead groups are free (and projective $\mathbb{Z}$-modules are free).

PROPOSITION 2.3. *Assume R is hereditary, κ is a regular uncountable cardinal and M is generated by κ elements.*

(i) $\Gamma_\kappa(M) = 0\ (= \tilde{\emptyset})$ *if and only if $M = L \oplus P$ where P is projective and L is generated by $<\kappa$ elements.*

(ii) *If* $\Gamma_{\kappa,N}(M) = 0$, *then there is a submodule* L *of* M *generated by fewer than* κ *elements such that* $\operatorname{Ext}(M/L, N) = 0$.

PROOF. (i) If C is a club in κ such that $C \cap E_\kappa(M) = \emptyset$, then $\{M_\alpha\colon \alpha \in C\}$ is a κ-filtration of M with that property that for any $\alpha \in C$ if α_+ denotes the next largest member of C, then M_{α_+}/M_α is projective. Let γ be the least element of C and $L = M_\gamma$. Then M/L is projective, in fact, isomorphic to the direct sum of all the M_{α_+}/M_α ($\alpha \in C$); so $M \cong L \oplus (M/L)$.

To prove the converse, suppose that $M = L \oplus P$, as described. Since every projective P is a direct sum of countably-generated projectives [74], there is a κ-filtration $\{M'_\alpha\colon \alpha < \kappa\}$ of M such that $M'_0 = L$ and for all $\alpha < \gamma$, M'_γ/M'_α is projective. But then, calculated with respect to this filtration, $E_\kappa(M) = \emptyset$, so $\Gamma_\kappa(M) = 0$.

The proof of (ii) is similar to that of (i), using the fact that if a module $\overline{M}$ ($= M/L$) has a κ-filtration $\{\overline{M}_\alpha\colon \alpha < \kappa\}$ such that $\overline{M}_0 = 0$ and $\operatorname{Ext}(\overline{M}_{\alpha+1}/\overline{M}_\alpha, N) = 0$ whenever $\alpha < \kappa$, then $\operatorname{Ext}(\overline{M}, N) = 0$. (See [21, Thm. 1.2] or [41, Lemma 2.1, p. 74] for this fact.) □

There is no converse to (ii) provable in ZFC even for the case $N = R = \mathbb{Z}$. This is the key to the independence of Whitehead's problem, as we will see.

3. Diamond and weak diamond

Gödel introduced the axiom of constructibility, usually denoted $\mathrm{V} = \mathrm{L}$, in his proof that the axiom of choice and the continuum hypothesis are relatively consistent with ordinary set theory. (By Gödel's second incompleteness theorem, the axioms of ZFC cannot be proved to be consistent, but his proof shows that the axiom of choice and the continuum hypothesis do not introduce inconsistencies where there were none before.) We will not state the axiom of constructibility here since what we employ are certain combinatorial consequences of $\mathrm{V} = \mathrm{L}$, the "diamond" principles, which were discovered by Jensen [72]. These can, in fact, be proved relatively consistent with ZFC by other means (by forcing) as well.

DEFINITION 3.1. Let κ be a regular cardinal and S a stationary subset of κ. By $\diamondsuit_\kappa(S)$ is denoted the assertion that there is a sequence $\langle W_\alpha\colon \alpha \in S\rangle$ of sets such that $W_\alpha \subseteq \alpha$ for all $\alpha \in S$ and such that for every subset X of κ, $\{\alpha \in S\colon X \cap \alpha = W_\alpha\}$ is stationary in κ.

Such a sequence $\langle W_\alpha\colon \alpha \in S\rangle$ will be called a $\diamondsuit_\kappa(S)$-*sequence*. The assertion $\diamondsuit_\kappa(S)$ is rather remarkable: note that the sequence is independent of X and yet "predicts" what $X \cap \alpha$ will be, and is correct quite often – on a "non-zero measure" subset. Yet all of these assertions $\diamondsuit_\kappa(S)$ (for all regular uncountable κ and stationary S) are simultaneously consistent with ZFC (but not provable in ZFC). In fact, they are all consequences of $\mathrm{ZFC} + \mathrm{V} = \mathrm{L}$. (It is also consistent with ZFC that $\diamondsuit_{\aleph_1}(S)$ is false for all $S \subseteq \aleph_1$; it is also consistent that $\diamondsuit_{\aleph_1}(S)$ holds for some subsets S of $\aleph_1$ and fails for others [104]. However, if $\widetilde{S}_1 = \widetilde{S}_2$, then $\diamondsuit_\kappa(S_1)$ if and only if $\diamondsuit_\kappa(S_2)$.)

By a simple coding argument, $\diamondsuit_\kappa(S)$ also implies the following.

LEMMA 3.2. *For any sets A and B of size κ and κ-filtrations $\{A_\alpha\colon \alpha < \kappa\}$ and $\{B_\alpha\colon \alpha < \kappa\}$ of A and B, respectively, there is a sequence $\langle H_\alpha\colon \alpha \in S\rangle$ of functions $H_\alpha : A_\alpha \to B_\alpha$ such that for every function $F : A \to B$, $\{\alpha \in S\colon F \restriction A_\alpha = H_\alpha\}$ is stationary in κ.*

PROOF. There is a bijection $\theta : \kappa \to A \times B$ such that there is a club $C \subseteq \kappa$ such that for $\alpha \in C$, $\theta \restriction \alpha$ is a bijection of α with $A_\alpha \times B_\alpha$. For $\alpha \in C \cap S$, let $H_\alpha = \theta[W_\alpha]$ if the latter is a function on A_α and otherwise let H_α be arbitrary. Given any function $F : A \to B$, regarded as a subset of $A \times B$, let $X = \theta^{-1}[F]$. There is a club D such that for $\alpha \in D$, $F[A_\alpha] \subseteq B_\alpha$. Then $C \cap D$ is a club, so there is $\alpha \in C \cap D \cap S$ such that $X \cap \alpha = W_\alpha$; we have $\theta[X \cap \alpha] = F \cap (A_\alpha \times B_\alpha) = F \restriction A_\alpha$ so $F \restriction A_\alpha = \theta[W_\alpha] = H_\alpha$. $\square$

THEOREM 3.3. *Assume $\diamondsuit_\kappa(S)$. Let R be an hereditary ring of cardinality $< \kappa$ and N a module of cardinality $\leqslant \kappa$. If M is a module of cardinality κ such that* $\operatorname{Ext}(M, N) = 0$, *then $\widetilde{S} \nsubseteq \Gamma_{\kappa,N}(M)$.*

PROOF. We sketch the proof of the contrapositive when N has cardinality $< \kappa$ (cf. [21]). Suppose that $\{M_\alpha\colon \alpha < \kappa\}$ is a κ-filtration of M such that $S \subseteq \{\alpha\colon \exists\beta > \alpha \text{ s.t. } \operatorname{Ext}(M_\beta/M_\alpha, N) \neq 0)\}$. Without loss of generality, we can assume that if $\alpha \in S$, then $\operatorname{Ext}(M_{\alpha+1}/M_\beta, N) \neq 0$. Also, we can assume that $\operatorname{Ext}(M_\alpha, N) = 0$ for all $\alpha < \kappa$ since otherwise we can immediately conclude that $\operatorname{Ext}(M, N) \neq 0$, because R is hereditary. We will use $\diamondsuit_\kappa(S)$ to build a short exact sequence

$$\mathcal{E} : 0 \to N \to L \xrightarrow{\pi} M \to 0$$

which does not split. We build it as the union of sequences

$$\mathcal{E}_\alpha : 0 \to N \to L_\alpha \xrightarrow{\pi_\alpha} M_\alpha \to 0$$

where the underlying set of L_α is $N \times M_\alpha$. By Lemma 3.2, there is a sequence $\langle H_\alpha\colon \alpha \in S\rangle$ such that each H_α is a function: $M_\alpha \to N \times M_\alpha$ and for every (set) function $F : M \to N \times M$, $\{\alpha \in S\colon F \restriction M_\alpha = H_\alpha\}$ is stationary. The crucial case is when $\alpha \in S$, $\mathcal{E}_\alpha$ has been defined and H_α is a homomorphism which is a splitting of π_α. We will define $\mathcal{E}_{\alpha+1} : 0 \to N \to L_{\alpha+1} \to M_{\alpha+1} \to 0$ extending $\mathcal{E}_\alpha$ but such that H_α does not extend to a splitting of $\mathcal{E}_{\alpha+1}$. This will suffice since any possible splitting $F : M \to L$ of $\mathcal{E}$ will restrict to H_α for some $\alpha \in S$ and the construction will have ruled out the existence of F.

To define $\mathcal{E}_{\alpha+1}$ we use the fact that $\operatorname{Ext}(M_{\alpha+1}/M_\alpha, N) \neq 0$ in the exact sequence

$$\begin{aligned} \operatorname{Hom}(M_{\alpha+1}, N) &\to \operatorname{Hom}(M_\alpha, N) \\ &\to \operatorname{Ext}(M_{\alpha+1}/M_\alpha, N) \to \operatorname{Ext}(M_{\alpha+1}, N) = 0 \end{aligned}$$

implies that some homomorphism $\psi_\alpha : M_\alpha \to N$ does not extend to a homomorphism: $M_{\alpha+1} \to N$. Now $L_\alpha = N \oplus \operatorname{im}(H_\alpha)$, since H_α is a splitting of π_α. Extend H_α to an isomorphism: $M_{\alpha+1} \to M'_{\alpha+1} \supseteq \operatorname{im}(H_\alpha)$ and let $L'_{\alpha+1} = N \oplus M'_{\alpha+1}$. Let θ_α be the embedding: $L_\alpha \to L'_{\alpha+1} : n + H_\alpha(m) \mapsto n + \psi_\alpha(m) + H_\alpha(m)$. Then we have a commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & N & \longrightarrow & L'_{\alpha+1} & \xrightarrow{\pi_{\alpha+1}} & M_{\alpha+1} & \longrightarrow & 0 \\ & & \| & & \uparrow{\scriptstyle \theta_\alpha} & & \uparrow & & \\ 0 & \longrightarrow & N & \longrightarrow & L_\alpha & \xrightarrow{\pi_\alpha} & M_\alpha & \longrightarrow & 0 \end{array}$$

There is no splitting $\rho : M_{\alpha+1} \to L'_{\alpha+1}$ of $\pi_{\alpha+1}$ such that $\rho \restriction M_\alpha = \theta_\alpha \circ H_\alpha$, since if there were, then by composing with projection on N we would get an extension of ψ_α to $M_{\alpha+1}$. Finally, by a standard set-theoretic argument we can replace $L'_{\alpha+1}$ by a module with underlying set $N \times M_{\alpha+1}$ which actually contains L_α as submodule (so that θ_α becomes inclusion). □

COROLLARY 3.4. *Assume* V = L. *Then every Whitehead group of cardinality* $\aleph_1$ *is free.*

PROOF. Since V = L implies $\diamondsuit_{\aleph_1}(S)$ for every stationary subset S of $\aleph_1$, the theorem implies that if A is a Whitehead group of cardinality $\aleph_1$, then $\Gamma_{\aleph_1,\mathbb{Z}}(A) = 0$. So A has an $\aleph_1$-filtration $\{A_\alpha\colon\ \alpha < \omega_1\}$ such that for every $\alpha < \omega_1$, $\operatorname{Ext}(A_{\alpha+1}/A_\alpha, \mathbb{Z}) = 0$. But $A_{\alpha+1}/A_\alpha$ is a *countable* Whitehead group, and hence free, by Theorem 1.2. Therefore A is free. □

The same argument in fact proves:

COROLLARY 3.5. *Assume* V = L. *Let κ be a regular uncountable cardinal such that every Whitehead group of cardinality $< \kappa$ is free. Then every Whitehead group of cardinality κ is free.*

By results of Gregory and Shelah [51,108], GCH implies that for every cardinal $\kappa = \lambda^+$ such that λ is uncountable, $\diamondsuit_\kappa(\kappa)$ holds (although $\diamondsuit_\kappa(S)$ may fail for other stationary subsets, S, of κ). This is not true for $\kappa = \aleph_1$, but Devlin and Shelah [12] discovered that a weak form of diamond is implied by CH (or even weak CH, that is, $2^{\aleph_0} < 2^{\aleph_1}$). In general we define:

DEFINITION 3.6. $\Phi_\kappa(S)$ is the assertion that given any family of functions F_α from the powerset $\mathcal{P}(\alpha)$ to $\{0, 1\}$ ($\alpha \in S$), there is a function $\rho : S \to \{0, 1\}$ such that for every subset X of κ, $\{\alpha \in S\colon\ F_\alpha(X \cap \alpha) = \rho(\alpha)\}$ is stationary in κ.

Devlin and Shelah showed that weak CH implies $\Phi_{\aleph_1}(\aleph_1)$. It is easy to see that $\diamondsuit_\kappa(S)$ implies $\Phi_\kappa(S)$. It is less obvious, but true, that the weaker hypothesis $\Phi_\kappa(S)$ is sufficient to prove Theorem 3.3. (See [12, §6].) In particular therefore we have:

THEOREM 3.7. *Assuming that* $2^{\aleph_0} < 2^{\aleph_1}$, *if B is a Whitehead group of cardinality $\aleph_1$, then* $\Gamma_{\mathbb{Z}}(B) \neq 1$.

This is, in fact, a strong form of Chase's result, Theorem 1.6, that Whitehead groups are strongly $\aleph_1$-free. Indeed, if A is a Whitehead group of arbitrary cardinality which is not strongly $\aleph_1$-free, there is a countable subgroup B_0 which is not contained in any countable subgroup H such that A/H is $\aleph_1$-free. One can then inductively define a continuous chain of countable subgroups $\{B_\nu\colon\ \nu < \omega_1\}$ such that for all ν, $B_{\nu+1}/B_\nu$ is not free; but then $B = \bigcup_{\nu<\omega_1} B_\nu$ is a Whitehead group of cardinality $\aleph_1$ with $\Gamma_{\mathbb{Z}}(B) = 1$.

REMARK. Megibben used the diamond principles to solve Crawley's problem – a problem in Abelian p-group theory analogous to Whitehead's problem – for p-groups of cardinality $\aleph_1$, and showed the problem is undecidable in ZFC; see [80]. The solution for groups of arbitrary size was given by Mekler and Shelah [84,85]. Mekler [82] showed that the problem is undecidable in ZFC + GCH.

4. Singular compactness

Using Theorem 1.2 and Corollary 3.5 we can show, by induction, that the axiom of constructibility, V = L, implies that every Whitehead group of cardinality $\aleph_n$ is free, for all $n \in \omega$. To deal with a singular cardinal such as $\aleph_\omega$ we need a different argument, which was also supplied by Shelah [103]. The following "singular compactness theorem" is provable in ZFC. (The cases of λ of cofinality $\aleph_0$ or $\aleph_1$ for $R = \mathbb{Z}$ were proved earlier by Hill [60,59].)

THEOREM 4.1. *Let λ be a singular cardinal. Let R be an hereditary ring and M a module which is generated by a set of size λ. Suppose that every submodule of M which can be generated by a set of size $< \lambda$ is projective. Then M is projective.*

COROLLARY 4.2. *Assume* V = L. *Then every Whitehead group is free.*

PROOF. We prove by induction on cardinals κ that every Whitehead group of cardinality κ is free. For $\kappa = \aleph_0$ this is Theorem 1.2. Assume that κ is uncountable and that the result has been proved for all cardinals $< \kappa$. If κ is regular, use Corollary 3.5. (Note that a subgroup of a Whitehead group is Whitehead.) If κ is singular, use Theorem 4.1. □

More generally one can prove (cf. [121, Thm. 3.13]):

COROLLARY 4.3. *Assume* V = L. *Let R be a left hereditary non-left perfect ring and $\mu = 2^{\mathrm{card}(R)}$. Then every W_μ-module is projective.*

(For left perfect rings, there are, provably in ZFC, test modules for projectivity: see [121, Prop. 1.4].)

The version of the singular compactness theorem stated above is actually a very special case of a much more general theorem, which applies to an abstract notion of "free", defined axiomatically. The theorem says that if an object has singular cardinality λ and if "most" of its subobjects of cardinality $< \lambda$ are free, then the given object is free. In particular, the singular compactness theorem applies to the usual notion of "free" in a variety; "most" can be taken to mean “all” if the variety has the property that subalgebras of free algebras are free. Otherwise “most” must be appropriately defined to make the theorem true, but non-vacuous; in particular, “most” subalgebras of a free algebra should be free. (See [64] for a good exposition.) For example, here is a version for modules over general (non-hereditary) rings:

THEOREM 4.4. *Let λ be a singular cardinal. Let R be any ring and M a module which is generated by a set of size λ. Suppose that for all infinite $\kappa < \lambda$, there is a set $\mathcal{C}_\kappa$ of free submodules each generated by κ elements of M which is closed and unbounded, that is, every subset of M of size κ is contained in a member of $\mathcal{C}_\kappa$ and the union of a well-ordered chain of length at most κ of members of $\mathcal{C}_\kappa$ belongs to $\mathcal{C}_\kappa$. Then M is free.*

Note that a free module F which is generated by a set of size λ satisfies the hypothesis of the theorem; indeed, if B is a basis of F and $\kappa < \lambda$, let $\mathcal{C}_\kappa$ consist of all submodules of F which are generated by a subset of B of size at most κ.

Even for modules, versions of the singular compactness theorem for abstract notions of “free” can be useful. For example, the following result uses such a theorem, and its proof, as well as Theorem 3.3 for the case of regular κ. (See [22, Thm. 5.5], where the proof is for $R = \mathbb{Z}$ but extends easily to any hereditary ring; or [5, Thm. 3.1], where the proof for Whitehead modules extends easily to the general case.)

THEOREM 4.5. *Assume* $\mathrm{V} = \mathrm{L}$. *Let R be an hereditary ring, N an R-module and* $\mu = \mathrm{card}(N) + \mathrm{card}(R) + \aleph_0$. *An R-module M of cardinality $\kappa > \mu$ satisfies* $\mathrm{Ext}(M, N) = 0$ *if and only if it is the union of a continuous well-ordered chain $\{M_\alpha\colon \alpha < \kappa\}$ such that $M_0 = 0$ and for all $\alpha < \kappa$, $M_{\alpha+1}/M_\alpha$ is of cardinality $\leqslant \mu$ and satisfies* $\mathrm{Ext}(M_{\alpha+1}/M_\alpha, N) = 0$.

In particular, the question of characterizing the Whitehead R-modules is reduced to the question of characterizing the “small” Whitehead R-modules. Thus, if one can show that all Whitehead R-modules of cardinality $\leqslant \mathrm{card}(R) + \aleph_0$ are projective, then it follows (assuming $\mathrm{V} = \mathrm{L}$) that all Whitehead R-modules are projective. This is the case, for example, if R is a p.i.d. of cardinality at most $\aleph_1$ which is not a complete discrete valuation ring, or if R is a countable valuation domain: see [5, §§5, 6].

For more on abstract versions of the singular compactness theorem for modules see [29, §IV.3].

5. Uniformization

Originally Shelah proved the consistency with ZFC of the existence of a non-free Whitehead group by using Martin’s axiom plus the negation of the continuum hypothesis (CH). (See [102] or the exposition in [20, §7].) However in order to prove consistency of the same conclusion with ZFC + CH (and even with GCH), he introduced the technique of uniformization. (See [24, §§1,2] for more on the history.) The method has great power; it can be used to prove that for any non-perfect ring R and any R-module N, it is consistent with ZFC + GCH that there is a non-projective M such that $\mathrm{Ext}(M, N) = 0$. First we need some definitions.

DEFINITIONS. Let S be a subset of κ consisting of limit ordinals of cofinality $\aleph_0$. For any $\delta \in S$, a *ladder* on δ is a strictly increasing function $\eta_\delta : \omega \to \delta$ whose range is cofinal in δ. A *ladder system* on S is an indexed family $(\eta_\delta\colon\ \delta \in S)$ where each η_δ is a ladder on δ.

If μ is a cardinal (finite or infinite), a *μ-coloring* of $(\eta_\delta\colon\ \delta \in S)$ is an indexed family $c = (c_\delta\colon\ \delta \in S)$ of functions $c_\delta : \omega \to \mu = \{\nu\colon\ \nu < \mu\}$. (Think of $c_\delta(n)$ as being the color of the "n-th rung" $\eta_\delta(n)$ of the ladder η_δ.) A *uniformization* of c is a pair (h, h^*) of functions such that $h : \kappa \to \mu$, $h^* : S \to \omega$ and for all $\delta \in S$, $c_\delta(n) = h(\eta_\delta(n))$ for all $n \geqslant h^*(\delta)$. Say that $(\eta_\delta\colon\ \delta \in S)$ has the *μ-uniformization property* iff every μ-coloring of $(\eta_\delta\colon\ \delta \in S)$ has a uniformization.

By Lemma 2.1 we cannot hope to have h^* be constantly zero. (By considering the function $f(\delta) = \eta_\delta(0)$ we see that there are $\delta_1 \neq \delta_2$ in S with $\eta_{\delta_1}(0) = \eta_{\delta_2}(0)$; define $c_{\delta_1}(0) \neq c_{\delta_2}(0)$.) If $\diamondsuit_\kappa(S)$ is true, then no ladder system on S has the 2-uniformization property; in fact, the diamond sequence gives a prediction of what $h \restriction \delta$ will be for $\delta \in S$, and then the coloring c_δ can be chosen to prevent h from being a uniformization. The same conclusion follows also from weak diamond $\Phi_\kappa(S)$. Since CH implies $\Phi_{\aleph_1}(\aleph_1)$ for uncountable successor cardinals, it follows that CH implies that no ladder system on a club $S \subseteq \aleph_1$ has the 2-uniformization property. On the other hand, we have the following consistency results [104,32]:

THEOREM 5.1.
- (i) *It is consistent with* ZFC + GCH *that there is a stationary subset $S \subseteq \aleph_1$ which does not contain a club* (*i.e.* $\aleph_1 - S$ *is stationary*) *such that every ladder system on S has the $\aleph_0$-uniformization property.*
- (ii) *It is consistent with* ZFC + GCH *that for all $\kappa = \lambda^+$ where λ is a singular cardinal of cofinality $\aleph_0$, there is a stationary subset S of κ consisting of limit ordinals of cofinality $\aleph_0$ and a ladder system η on S which has the μ-uniformization property for every $\mu < \lambda$.*

Using ladder systems with uniformization properties we can construct non-projective Whitehead modules for any ring which is not left perfect.

THEOREM 5.2. *Let R be a ring which is not left perfect. Suppose that for some regular uncountable cardinal κ there is a stationary subset S of κ consisting of limit ordinals of cofinality $\aleph_0$ and a ladder system η on S which has the μ-uniformization property for some cardinal μ. Then there is a non-projective module M such that* $\mathrm{Ext}(M, N) = 0$ *for all modules N of cardinality $\leqslant \mu$.*

PROOF. Since R is not left perfect, there is an infinite descending chain of principal right ideals of R (cf. [4]). Let $a_0R \supset a_0a_1R \supset a_0a_1a_2R \supset \cdots \supset a_0a_1 \ldots a_nR \supset \cdots$ be such a chain. Let F be the free R-module on the basis $\{z_{\delta,n}\colon\ \delta \in S,\ n \in \omega\} \cup \{x_\alpha\colon\ \alpha \in \kappa\}$. Let K be the submodule of F generated by $\{w_{\delta,n}\colon\ \delta \in S,\ n \in \omega\}$ where

$$w_{\delta,n} = a_n z_{\delta,n+1} - z_{\delta,n} - x_{\eta_\delta(n)}.$$

It is easy to check that $\{w_{\delta,n}\colon\ \delta \in S,\ n \in \omega\}$ is a basis of K. Let $M = F/K$. First let us show that $\mathrm{Ext}(M, N) = 0$ for all modules N of cardinality $\leqslant \mu$. It suffices to show that every homomorphism ψ from K into N extends to a homomorphism φ from F to N. As a first naive attempt, we might try to define $\varphi(z_{\delta,n}) = 0$ for all δ, n and $\varphi(x_{\eta_\delta(n)}) = \psi(w_{\delta,n})$. However, we may have $\eta_\delta(n) = \eta_\gamma(m)$ for $(\delta, n) \neq (\gamma, m)$ yet $\psi(w_{\delta,n}) \neq \psi(w_{\gamma,m})$. This is exactly the problem that uniformization will solve.

Given ψ, define a μ-coloring c of η by: $c_\delta(n) = \psi(w_{\delta,n})$. (We can assume that the underlying set of N is a subset of μ.) If (h, h^*) is a uniformization of c, we define the extension, φ, of ψ to F as follows: $\varphi(x_\alpha) = h(\alpha)$, $\varphi(z_{\delta,n}) = 0$ if $n \geqslant h^*(\delta)$, and for $n < h^*(\delta)$, $\varphi(z_{\delta,n})$ is defined by downward induction according to the formula

$$\varphi(z_{\delta,n}) = a_n\varphi(z_{\delta,n+1}) - \varphi(x_{\eta_\delta(n)}) - \psi(w_{\delta,n}).$$

Finally we must show that M is not projective. For hereditary rings, it suffices, by Proposition 2.3, to show that $\Gamma_\kappa(M) \neq 0$. In fact we show that $\Gamma_\kappa(M) \supseteq \widetilde{S}$. We use the κ-filtration $\{M_\alpha\colon\ \alpha < \kappa\}$ where M_α is the submodule generated by $\{z_{\delta,n} + K\colon\ \delta \in S, \delta < \alpha, n \in \omega\} \cup \{x_\nu + K\colon\ \nu < \alpha\}$. Then for $\delta \in S$, $M_{\delta+1}/M_\delta$ is not projective by a result of Bass [4, Lemma 1.3]. For non-hereditary rings the argument is similar, see [120]. □

As an immediate consequence of Theorems 5.1 and 5.2, we have the following.

COROLLARY 5.3.
(i) *It is consistent with* ZFC + GCH *that there is a non-free Whitehead group of cardinality* $\aleph_1$.
(ii) *It is consistent with* ZFC + GCH *that for every cardinal* μ*, there is a non-free Whitehead* W_μ*-group.*
(iii) *It is consistent with* ZFC + GCH *that for every ring* R *which is not left perfect, there is no test module for projectivity.*

For Abelian groups of cardinality $\aleph_1$, there is an exact equivalence between the uniformization problem and the Whitehead problem; the following is a theorem of ZFC (cf. [106, Thm. 3.9], [33, §6]).

THEOREM 5.4. *For* μ *equal to* 2 *or* $\aleph_0$*, there is a non-free* W_μ*-group of cardinality* $\aleph_1$ *if and only there is a ladder system on a stationary subset of* $\aleph_1$ *which has the* μ*-uniformization property.*

For cardinals κ larger than $\aleph_1$, there is also a (more complicated) combinatorial equivalent to the existence of a κ-free, non-free Whitehead group of cardinality κ: see [33].

Models (of set theory, ZFC) which have, or fail to have, certain uniformization properties are created by the method of iterated forcing, especially proper forcing (cf. [109]). To derive a property of W-groups in the model, it is sometimes just enough to know that a uniformization property holds or fails; other times it is necessary to use stronger properties of the model. By these methods one can prove, for example, that each of the following is (separately) consistent with ZFC + GCH:

- There are non-free W-groups of cardinality $\aleph_1$ and every W-group of cardinality $\aleph_1$ is a $W_{\aleph_0}$-group [102].
- There are non-free W-groups of cardinality $\aleph_1$ but every $W_{\aleph_0}$-group of cardinality $\aleph_1$ is free ([106], [29, Thm. XIII.3.9]).
- For any $\aleph_1$-free Abelian groups A and B of cardinality $\aleph_1$, if A is a W-group and $\Gamma_{\aleph_1}(A) = \Gamma_{\aleph_1}(B)$, then B is a W-group [104,81].
- There are $\aleph_1$-free Abelian groups A and B of cardinality $\aleph_1$, such that A is a W-group and $\Gamma_{\aleph_1}(A) = \Gamma_{\aleph_1}(B)$, but B is not a W-group [106].
- Every W-group (of arbitrary cardinality) is free, but there are non-free hereditarily separable groups of cardinality $\aleph_1$ [31].
- There are non-free W-groups and the smallest non-free W-group is of cardinality $> \aleph_{\omega_1+174}$ (or any given cardinal) [32].
- There is an (inaccessible) cardinal κ such that there are non-free κ-free groups of cardinality κ and every κ-free group of cardinality κ is a Whitehead group [112].

Lest one think that "anything is possible" with Whitehead groups, we conclude with the following, which is proved using uniformizations:

THEOREM 5.5 [30].
(i) *It is* not *consistent with* ZFC *that* every *W-group of cardinality* $\aleph_1$ *is strongly* $\aleph_1$-*free.*
(ii) *It is* not *consistent with* ZFC *that* every *strongly* $\aleph_1$-*free group of cardinality* $\aleph_1$ *is a W-group and none is a* $W_{\aleph_0}$-*group.*

6. Baer modules

In 1936 Baer [2] proved that if M is a countable Abelian group such that $\mathrm{Ext}(M, T) = 0$ for all torsion Abelian groups T, then M is free. Thirty years later Griffith [53] finally showed that the same holds for all uncountable Abelian groups M. These results were extended to modules over Dedekind domains by Nunke [86] and Grimaldi [54]. All these proofs use heavily the fact that the ring in question is hereditary. Kaplansky [75] raised the question of characterizing Baer modules over arbitrary domains and proved the following:

LEMMA 6.1. *Let R be any integral domain. Let B be a Baer module, that is,* $\mathrm{Ext}(B, T) = 0$ *for every torsion module T. Then B is flat, hence torsion-free, and has projective dimension* $\leqslant 1$.

Eklof and Fuchs [25] proved that Baer modules over arbitrary valuation domains are free. Since valuation domains, other than the discrete rank one valuation domains, are not hereditary, an approach different from Griffith's is needed, and, in fact, the method used is that employed in Sections 3 and 4 above. After the case of Baer modules of countable rank is proved, by methods special to valuation domains, the Baer modules of uncountable rank κ are handled by an induction on κ. The case of singular κ uses a version of the singular compactness theorem [25, Thm. 11] (where "most" submodules are free means that all pure submodules L of B generated by $< \kappa$ elements and with pr.dim.$(B/L) \leqslant 1$ are free). The case of regular κ uses an analog of Theorem 3.3, but one which is provable in ZFC (cf. [25, Lemma 9] and [26, Lemma 9]):

THEOREM 6.2. *Let R be an integral domain and B a Baer module generated by κ elements. If $\{B_\alpha: \alpha < \kappa\}$ is a κ-filtration of B such that* pr.dim.$(B_{\alpha+1}/B_\alpha) \leqslant 1$ *for all $\alpha < \kappa$, then $\{\alpha < \kappa: \exists \beta > \alpha$ s.t. B_β/B_α is not a Baer module$\}$ is not stationary in κ.*

The essential reason why this result (as opposed to Theorem 3.3) can be proved in ZFC is that the notion of Baer module is not relative to a single fixed T (in the hypothesis that $\text{Ext}(B, T) = 0$) but relative to a class of modules (the torsion modules) which is closed under arbitrary direct sums.

A result about Baer modules over arbitrary domains in the spirit of Theorem 4.5 (but provable in ZFC) was proved in [26]:

THEOREM 6.3. *Let R be an integral domain. An R-module M is a Baer module if and only if it is the union of a continuous well-ordered chain $\{M_\alpha: \alpha < \sigma\}$ such that $M_0 = 0$ and for all $\alpha + 1 < \sigma$, $M_{\alpha+1}/M_\alpha$ is a countably generated Baer module.*

7. Splitters

In this section and the next we confine ourselves to Abelian groups. A group A is called a *splitter* if $\text{Ext}(A, A) = 0$. This terminology is due to Schultz [101], but the condition that a module M satisfy $\text{Ext}(M, M) = 0$ (sometimes with additional requirements) has arisen in different settings and been given different names in the literature: exceptional, stone, Schur module, open brick: see the “Dictionary” in Ringel [93, p. 351]. Every (partial) tilting and every (partial) cotilting module is a splitter; moreover, if R is a hereditary finite-dimensional algebra, then every splitter is a partial tilting module in the sense of Bongartz (cf. [46,7,56]).

Hausen [57] proved that a countable torsion-free splitter is free as a module over its *nucleus*, nuc(A), the subring of $\mathbb{Q}$ generated by $\{p^{-1}: p$ is a prime such that $pA = A\}$. (Note that $\text{Ext}_{\mathbb{Z}}(A, A) = \text{Ext}_{\text{nuc}(A)}(A, A)$.) A torsion-free group A of arbitrary size will be a splitter if it is free as a nuc(A)-module or if it is cotorsion. (Cotorsion means that $\text{Ext}(B, A) = 0$ for every torsion-free group; for a torsion-free group this is equivalent to being pure-injective, or algebraically compact.) The question whether all splitters are either free over their nucleus or cotorsion was recently answered in the negative by Göbel and Shelah [45]:

THEOREM 7.1. *For any proper subring R of $\mathbb{Q}$, there is a torsion-free splitter A of cardinality $2^{\aleph_0}$ with nucleus R which is neither R-free nor cotorsion. Moreover, there exist arbitrarily large indecomposable splitters.*

Surprisingly perhaps, in view of the previous sections, this is settled in ZFC. An alternative proof of the first part is given in [37]. The latter proof generalizes easily to show that there are $\aleph_1$-free splitters of cardinality $2^{\aleph_1}$ which are not free and not cotorsion. On the other hand, Göbel and Shelah [44] have proved:

THEOREM 7.2. (CH) *Every $\aleph_1$-free splitter of cardinality $\aleph_1$ is free.*

For more information, see [29, §XVI.3].

8. The structure of Ext

If A is not a Whitehead group, the question arises what $\mathrm{Ext}(A, \mathbb{Z})$ could be. If A is a torsion-free group, then $\mathrm{Ext}(A, \mathbb{Z})$ is divisible (cf. the proof of Theorem 1.2). Therefore, for any group A, the short exact sequence $0 \to tA \to A \to A/tA \to 0$ induces a split exact sequence

$$0 = \mathrm{Hom}(tA, \mathbb{Z}) \to \mathrm{Ext}(A/tA, \mathbb{Z}) \to \mathrm{Ext}(A, \mathbb{Z}) \to \mathrm{Ext}(tA, \mathbb{Z}) \to 0$$

since $\mathrm{Ext}(A/tA, \mathbb{Z})$ is injective. So $\mathrm{Ext}(A, \mathbb{Z}) \cong \mathrm{Ext}(tA, \mathbb{Z}) \oplus \mathrm{Ext}(A/tA, \mathbb{Z})$, and the problem of determining the structure of $\mathrm{Ext}(A, \mathbb{Z})$ breaks up into two cases: torsion and torsion-free A.

In the case that A is a torsion group, the short exact sequence $0 \to \mathbb{Z} \to \mathbb{R} \to \mathbb{R}/\mathbb{Z} \to 0$ induces the exact sequence

$$0 = \mathrm{Hom}(A, \mathbb{R}) \to \mathrm{Hom}(A, \mathbb{R}/\mathbb{Z}) \to \mathrm{Ext}(A, \mathbb{Z}) \to \mathrm{Ext}(A, \mathbb{R}) = 0$$

and hence $\mathrm{Ext}(A, \mathbb{Z}) \cong \mathrm{Hom}(A, \mathbb{R}/\mathbb{Z}) = \mathrm{Char}(A)$. The structure of such a group is known (cf. [67,38]). It is a reduced group (no non-zero divisible subgroups) and isomorphic to a product of copies of J_p, the additive group of the p-adic integers, and copies of the finite cyclic groups $Z(p^n)$ (where p ranges over the primes and n over the positive integers). All such products are realizable as $\mathrm{Ext}(A, \mathbb{Z})$ for some torsion A. (See [40, §47].)

So we concentrate on the structure of $\mathrm{Ext}(A, \mathbb{Z})$ when A is torsion-free. It is a divisible group, so it is of the form

$$\mathbb{Q}^{(\nu_0(A))} \oplus \bigoplus_{p \text{ prime}} Z(p^\infty)^{(\nu_p(A))}$$

for some cardinals $\nu_p(A)$ ($p = 0$ or a prime) determined by A. For countable A the values of the cardinals which occur were determined by Jensen [71]: if A is not free, then $\nu_0(A) = 2^{\aleph_0}$ and for each prime p, $\nu_p(A)$ is either finite or $2^{\aleph_0}$.

For uncountable A, the analysis used in the proof of Theorem 3.3 was strengthened by Hiller, Huber and Shelah [61] to prove the following theorem. (It is proved under weaker hypotheses – but still not in ZFC – in [27].)

THEOREM 8.1. *Assume* $\mathrm{V} = \mathrm{L}$. *If A is a torsion-free group which is not free and B is a subgroup of A of minimal cardinality such that A/B is free, then $\nu_0(A) = 2^{\mathrm{card}(B)}$ and for all primes p, $\nu_p(A) \leqslant \nu_0(A)$.*

In particular, $\nu_0(A)$ is non-zero; thus, the result may be regarded as a strengthening of Corollary 3.4. Again, the proof of the theorem is by induction on the cardinality, κ, of B and uses a version of the singular compactness theorem for singular κ and the diamond principles $\diamondsuit_\kappa(S)$ for regular κ. One application is to answer a question of Kan and Whitehead [73]: assuming $\mathrm{V} = \mathrm{L}$, there is no topological space X and integer $n \geqslant 2$ such that $H^n(X, \mathbb{Z}) \cong \mathbb{Q}$. (See [62].)

The theorem (and the result regarding cohomology) is independent of ZFC + GCH: Shelah [107] proved that for any countable divisible group D, it is consistent with ZFC + GCH that there is a torsion-free A with $\operatorname{Ext}(A, \mathbb{Z}) \cong D$ (that is, it is consistent that any finite or countable values for $\nu_0(A)$ and the $\nu_p(A)$ can be realized).

Chase [11] proved in 1963 that the continuum hypothesis implies that there is a torsion-free group A of cardinality $\aleph_1$ such that $\operatorname{Ext}(A, \mathbb{Z})$ is torsion-free, i.e. $\nu_p(A) = 0$ for all primes p (and $\nu_0(A) = 2^{\aleph_1}$). This result was generalized by Sageev and Shelah [99] as follows. (See [28] for another proof.)

THEOREM 8.2. *Assume* CH. *For any sequence of cardinals $\langle \kappa_p\colon\ p \text{ a prime}\rangle$ such that for each p, $0 \leqslant \kappa_p \leqslant \aleph_1$ or $\kappa_p = 2^{\aleph_1}$, there is a torsion-free group A of cardinality $\aleph_1$ such that $\nu_0(A) = 2^{\aleph_1}$ and for each prime p, $\nu_p(A) = \kappa_p$.*

The result is extended in [83] to larger cardinals under the stronger assumption of $\mathrm{V} = \mathrm{L}$; the proof requires proving first new strong forms of the diamond principles.

THEOREM 8.3 [83]. *Assume* $\mathrm{V} = \mathrm{L}$. *For any regular cardinal λ less than the first weakly compact cardinal and any sequence of cardinals $\langle \kappa_p\colon\ p \text{ a prime}\rangle$ such that for each p, $0 \leqslant \kappa_p \leqslant \lambda^+$, there is a torsion-free group A of cardinality λ such that $\nu_0(A) = \lambda^+$ and for each prime p, $\nu_p(A) = \kappa_p$.*

We will not give the formal definition of a weakly compact cardinal here (see, for example, [70, p. 325] or [29, p. 29]) but simply note that if κ is weakly compact, κ is inaccessible and there are κ inaccessible cardinals below it; for such cardinals there are further restrictions on the p-ranks of Ext:

THEOREM 8.4 [98]. *If λ is a weakly compact cardinal and A is a torsion-free group of cardinality λ, then for each prime p, $\nu_p(A) < \lambda$ or $\nu_p(A) = 2^\lambda$. (In other words, $\nu_p(A) = \lambda$ is impossible.)*

The same result is proved in [55] for certain singular cardinals λ.

Considerations from algebraic topology regarding co-Moore spaces suggest investigating the structure of $\operatorname{Ext}(A, \mathbb{Z})$ when $\operatorname{Hom}(A, \mathbb{Z}) = 0$. (See [61, Part II]; also [50] and [35].) We have the following elementary lemma:

LEMMA 8.5 [61, Prop. 2]. *If A is torsion-free and* $\operatorname{Hom}(A, \mathbb{Z}) = 0$*, then for every prime p, $\nu_p(A)$ is finite or of the form 2^{μ_p} for some infinite cardinal μ_p.*

There are also the following restrictions and possibilities:

THEOREM 8.6 [35].

(i) *Assume* GCH. *For any torsion-free group A of uncountable cardinality, if* $\mathrm{Hom}(A, \mathbb{Z}) = 0$ *and* $\nu_0(A) < 2^{|A|}$, *then for each prime p, $\nu_p(A) = 2^{|A|}$.*

(ii) *It is consistent with* ZFC + GCH *that for any cardinal $\rho \leqslant \aleph_1$, there is a torsion-free group A of cardinality $\aleph_1$ such that* $\mathrm{Hom}(A, \mathbb{Z}) = 0$ *and* $\nu_0(A) = \rho$.

9. Slender modules

The notion of slenderness is due to Łoś (see [39, §47]). Given a product $\prod_{i\in I} M_i$ of modules, $j \in I$ and $a \in M_j$, let $\lambda_j(a)$ be the element of $\prod_{i\in I} M_i$ such that $\lambda_j(a)(j) = a$ and $\lambda_j(a)(i) = 0$ for all $i \neq j$. Obviously, λ_j is an embedding of M_j into $\prod_{i\in I} M_i$. An R-module S is said to be *slender* if for every homomorphism $\theta : R^\omega \to S$, $\theta(\lambda_n(1)) = 0$ for all but finitely many $n \in \omega$. Then a stronger property holds:

THEOREM 9.1. *Suppose S is slender. For any family $\{M_i\colon i \in I\}$ of R-modules, and any homomorphism $\varphi : \prod_{i\in I} M_i \to S$, there is a finite subset J of I such that for all $i \notin J$, $\varphi(\lambda_i(a)) = 0$ for all $a \in M_i$.*

PROOF. Suppose this is false; let $\varphi : \prod_{i\in I} M_i \to S$ be a counterexample. Then there is an infinite subset $\{i_n\colon n \in \omega\}$ of I such that for every $n \in \omega$, there exists $a_n \in M_{i_n}$ such that $\varphi(\lambda_{i_n}(a_n)) \neq 0$. Define $\psi : R^\omega \to \prod_{i\in I} M_i$ by

$$\psi\big((r_n)_{n\in\omega}\big)(i) = \begin{cases} r_m a_m & \text{if } i = i_m \text{ for some } m, \\ 0 & \text{otherwise} \end{cases}$$

and let $\theta = \varphi \circ \psi$. Then $\theta(\lambda_n(1)) = \varphi(\lambda_{i_n}(a_n)) \neq 0$ for every $n \in \omega$, which contradicts the fact that S is slender. $\square$

THEOREM 9.2. *If S is slender, then for any countable family $\{M_n\colon n \in \omega\}$ of modules and any homomorphism $\varphi : \prod_{n\in\omega} M_n \to S$, there exists a finite subset J of ω such that $\varphi \restriction \prod_{n\notin J} M_n$ is identically zero.*

PROOF. Here $\prod_{n\notin J} M_n = \{x \in \prod_{n\in\omega} M_n\colon\ x(n) = 0 \text{ for all } n \in J\}$. Suppose that $\varphi : \prod_{n\in\omega} M_n \to S$ is a counterexample. Then for every $m \in \omega$ there exists $x^{(m)} \in \prod_{n\geqslant m} M_n$ such that $\varphi(x^{(m)}) \neq 0$. Define $\theta : R^\omega \to S$ by $\theta((r_n)_{n\in\omega}) = \varphi(\sum_{n\in\omega} r_n x^{(n)})$, where $\sum_{n\in\omega} r_n x^{(n)}$ is the element of $\prod_{k\in\omega} M_k$ defined by

$$\left(\sum_{n\in\omega} r_n x^{(n)}\right)(k) = \sum_{n\in\omega} r_n x^{(n)}(k) = \sum_{n\leqslant k} r_n x^{(n)}(k).$$

Then θ is clearly a homomorphism such that for all $n \in \omega$, $\theta(\lambda_n(1)) = \varphi(x^{(n)}) \neq 0$, which is a contradiction of the definition of slenderness. $\square$

Theorem 9.2 does not hold in this form for families $\{M_i\colon i \in I\}$ over arbitrarily large index sets I. However, it is true if the cardinality of I is less than the first measurable cardinal (if any). (See [39, Thm. 47.2], [40, Thm. 94.4], or [29, Cor. III.3.3].) There is a generalization due to Eda [15,17] which holds for arbitrary index sets:

THEOREM 9.3. *Let S be a slender module. Then for any set I and any family $\{M_i\colon i \in I\}$ of R-modules and any homomorphism $\varphi : \prod_{i\in I} M_i \to S$ of R-modules, there are ω_1-complete ultrafilters $D_1, \ldots, D_n$ on I such that for all $x \in \prod_{i\in I} M_i$, if for all $k = 1, \ldots, n$ $\{i \in I\colon x(i) \neq 0\} \notin D_k$, then $\varphi(x) = 0$.*

This reduces to Theorem 9.2 when the size of I is less than the first measurable cardinal because in that case every ω_1-complete ultrafilter is principal, that is consists exactly of all sets which have a fixed element of I as a member. (See also [29, Thm. III.3.2] for the proof.)

It is well-known that for any family $\{M_i\colon i \in I\}$ of modules and any module N, there is a natural isomorphism between $\mathrm{Hom}(\bigoplus_{i\in I} M_i, N)$ and $\prod_{i\in I} \mathrm{Hom}(M_i, N)$ and between $\mathrm{Hom}(N, \prod_{i\in I} M_i)$ and $\prod_{i\in I} \mathrm{Hom}(N, M_i)$. There is in general no analogous isomorphism involving $\mathrm{Hom}(\prod_{i\in I} M_i, N)$, but when N is slender and the index set I is not too large, we do get one:

THEOREM 9.4. *Let S be slender and $\{M_i\colon i \in I\}$ a family of modules such that I is countable or, more generally has size less than the first measurable cardinal. Then the map Φ which takes $\vec{g} = (g_i)_{i\in I} \in \bigoplus_{i\in I} \mathrm{Hom}(M_i, S)$ to the element $\Phi(\vec{g}) \in \mathrm{Hom}(\prod_{i\in I} M_i, S)$ defined by*

$$\Phi(\vec{g})(m_i)_{i\in I} = \sum_{i\in I} g_i(m_i)$$

is an isomorphism.

PROOF. It is easy to check that Φ is a monomorphism; this does not require the slenderness of S. To prove that Φ is surjective, consider any $\varphi \in \mathrm{Hom}(\prod_{i\in I} M_i, S)$ and let J be as in Theorem 9.2. For each $i \in I$ define $g_i : M_i \to S$ by: $g_i(a) = \varphi(\lambda_i(a))$. Note that g_i is identically zero if $i \notin J$, so $\vec{g} = (g_i)_{i\in I}$ belongs to $\bigoplus_{i\in I} \mathrm{Hom}(M_i, S)$. It is then easy to check that $\Phi(\vec{g}) = \varphi$. □

In the general case, where I may be measurable, Theorem 9.3 yields an isomorphism of $\mathrm{Hom}(\prod_{i\in I} M_i, S)$ with $\bigoplus_D \mathrm{Hom}(\prod_{i\in I} M_i/D, S)$, where D ranges over the ω_1-complete ultrafilters on I and $\prod_{i\in I} M_i/D$ is the ultraproduct with respect to D.

Consideration of the relation between $\mathrm{Hom}(N, \bigoplus_{i\in I} M_i)$ and $\bigoplus_{i\in I} \mathrm{Hom}(N, M_i)$ leads to the notions of *dually slender* and *almost dually slender* (also known as *Fuchs-44*) modules. See [97,43], or [29, pp. 62, 82f] for precise definitions and results.

A direct sum of slender modules is slender: see [39,58,77,13], [41, §XIV.7].

An important fact, due to Specker [114], is that $\mathbb{Z}$ is a slender module. (See also [19, 123].) This is a consequence of the following theorem of Sasaida [100].

THEOREM 9.5. *A torsion-free reduced Abelian group of cardinality $< 2^{\aleph_0}$ is slender.*

PROOF. Suppose $\theta : \mathbb{Z}^\omega \to A$ is a counterexample, where A is a reduced torsion-free group of cardinality $< 2^{\aleph_0}$. Without loss of generality, $\theta(\lambda_n(1)) \neq 0$ for every $n \in \omega$. Since A is reduced, $\bigcap_{m\in\omega} m!A = 0$, so, since A is torsion-free, we can inductively choose m_n such that $\theta(m_n!\lambda_n(1)) \notin m_{n+1}!A$. Now the set $X = \{x \in \mathbb{Z}^\omega\colon\ x(n) \in \{0, m_n!\}\}$ has cardinality $2^{\aleph_0}$, so there exists $x_1 \neq x_2$ in X such that $\theta(x_1) = \theta(x_2)$. Consider $y = x_1 - x_2$: $\theta(y) = 0$ but there is a least n such that $y(n) \neq 0$. Then $y(n) = \pm m_n!$, so $\theta(y(n)\lambda_n(1)) \notin m_{n+1}!A$. On the other hand, for some $y' \in \mathbb{Z}^\omega$, $\theta(y(n)\lambda_n(1)) = \theta(y) - m_{n+1}!\theta(y') = -m_{n+1}!\theta(y')$, which is a contradiction. □

More generally we have the following. (See Heinlein [58]; or Lady [77] with the correction in Mader [78]; or see [29, §III.2].)

THEOREM 9.6.
(a) *A countable integral domain which is not a field is slender as a module over itself.*
(b) *A principal ideal domain is slender as a module over itself if and only if it is not a complete discrete valuation ring.*

Nunke gave a characterization of which Abelian groups are slender:

THEOREM 9.7 (Nunke [87]). *An Abelian group A is slender if and only if it is reduced and torsion-free and does not contain a copy of $\mathbb{Z}^\omega$ or J_p (the additive group of the p-adic integers) for any prime p.*

For any Abelian group A, let A^* denote $\mathrm{Hom}(A, \mathbb{Z})$, called the *dual* of A. There is a natural map from A to its double dual A^{**}, which takes $a \in A$ to the element $\check{a} \in A^{**}$ given by $\check{a}(f) = f(a)$ for all $f \in A^*$. A is said to be *reflexive* if this natural map is an isomorphism.

THEOREM 9.8 (Specker [114]). *$\mathbb{Z}^\omega$ and $\mathbb{Z}^{(\omega)}$ are reflexive.*

PROOF. In fact $(\mathbb{Z}^\omega)^*$ is naturally isomorphic to $\mathbb{Z}^{(\omega)}$ by Theorem 9.4, because $\mathbb{Z}$ is slender. Moreover, as a special case of the comments preceding Theorem 9.4, $(\mathbb{Z}^{(\omega)})^*$ is naturally isomorphic to $\mathbb{Z}^\omega$. One may check that the composition of the isomorphisms is the map described. □

This result holds as well for $\mathbb{Z}^{(\kappa)}$ and $\mathbb{Z}^\kappa$, when κ is less than the first measurable cardinal. For larger κ it is false; however for any cardinal κ, $(\mathbb{Z}^{(\kappa)})^{**}$ is free, by Eda's result. (See [29, III.3.8].)

We conclude with a result about Whitehead groups.

THEOREM 9.9 [88]. *Every Whitehead group is slender.*

PROOF. If A is a Whitehead group, A is $\aleph_1$-free (Theorem 1.2), so A is reduced and torsion-free and does not contain a copy of J_p. It remains, by Theorem 9.7, to prove that A does not contain a copy of $\mathbb{Z}^\omega$. (Note that Specker also proved that $\mathbb{Z}^\omega$ is $\aleph_1$-free, in fact, separable.) We do this by showing that $\mathrm{Ext}(\mathbb{Z}^\omega, \mathbb{Z}) \neq 0$. Now the exact sequence $0 \to \mathbb{Z}^\omega \xrightarrow{p} \mathbb{Z}^\omega \to \mathbb{Z}^\omega/p\mathbb{Z} \to 0$ induces an exact sequence

$$\mathrm{Hom}(\mathbb{Z}^\omega, \mathbb{Z}) \to \mathrm{Ext}(\mathbb{Z}^\omega/p\mathbb{Z}^\omega, \mathbb{Z}) \to \mathrm{Ext}(\mathbb{Z}^\omega, \mathbb{Z}) \xrightarrow{p} \mathrm{Ext}(\mathbb{Z}^\omega, \mathbb{Z})$$

and the kernel of the last map is the p-torsion subgroup of $\mathrm{Ext}(\mathbb{Z}^\omega, \mathbb{Z})$. This subgroup is non-zero, in fact uncountable, since $\mathrm{Hom}(\mathbb{Z}^\omega, \mathbb{Z})$ is countable (by Theorem 9.2) but $\mathrm{Ext}(\mathbb{Z}^\omega/p\mathbb{Z}^\omega, \mathbb{Z}) \cong \mathrm{Ext}((\mathbb{Z}/p\mathbb{Z})^{(2^{\aleph_0})}, \mathbb{Z})$ is uncountable because it is the direct product of $2^{\aleph_0}$ copies of $\mathrm{Ext}(\mathbb{Z}/p\mathbb{Z}, \mathbb{Z})$. □

In fact, Nunke proved that $\mathrm{Ext}(\mathbb{Z}^\omega, \mathbb{Z})$ is the direct sum of $2^{2^{\aleph_0}}$ copies of $\mathbb{Q} \oplus \mathbb{Q}/\mathbb{Z}$.

References

[1] F.W. Anderson and K.R. Fuller, *Rings and Categories of Modules*, 2nd edn, Springer-Verlag, Berlin (1992).
[2] R. Baer, *The subgroup of the elements of finite order of an Abelian group*, Ann. of Math. **37** (1936), 766–781.
[3] S. Balcerzyk, *On groups of functions defined on Boolean algebras*, Fund. Math. **50** (1962), 347–367.
[4] H. Bass, *Finitistic dimension and a homological generalization of semiprimary rings*, Trans. Amer. Math. Soc. **95** (1960), 466–488.
[5] T. Becker, L. Fuchs and S. Shelah, *Whitehead modules over domains*, Forum Math. **1** (1989), 53–68.
[6] A. Blass and C. Laflamme, *Consistency results about filters and the number of inequivalent growth types*, J. Symbolic Logic **54** (1989), 50–56.
[7] K. Bongartz, *Tilted algebras*, Representations of Algebras (Puebla, 1980), Lecture Notes in Math., Vol. 903, Springer, Berlin (1981), 26–38.
[8] H. Cartan and S. Eilenberg, *Homological Algebra*, Princeton Univ. Press, Berlin (1956).
[9] S.U. Chase, *Locally free modules and a problem of Whitehead*, Illinois J. Math. **6** (1962), 682–699.
[10] S.U. Chase, *On direct sums and products of modules*, Pacific J. Math. **12** (1962), 847–854.
[11] S.U. Chase, *On group extensions and a problem of J.H.C. Whitehead*, Topics in Abelian Groups, Scott, Foresman and Co. (1963), 173–197.
[12] K.J. Devlin and S. Shelah, *A weak version of $\diamond$ which follows from $2^{\aleph_0} < 2^{\aleph_1}$*, Israel J. Math. **29** (1978), 239–247.
[13] R. Dimitrić, *Slender modules over domains*, Comm. Algebra **11** (1983), 1685–1700.
[14] M. Dugas and B. Zimmermann-Huisgen, *Iterated direct sums and products of modules*, Abelian Group Theory, Lecture Notes in Math., Vol. 874, Springer-Verlag, Berlin (1981), 179–193.
[15] K. Eda, *A Boolean power and a direct product of Abelian groups*, Tsukuba J. Math. **6** (1982), 187–193.
[16] K. Eda, *Almost-slender groups and Fuchs-44-groups*, Comment. Math. Univ. St. Paul **32** (1983), 131–135.
[17] K. Eda, *On a Boolean power of a torsion free Abelian group*, J. Algebra **82** (1983), 84–93.
[18] A. Ehrenfeucht, *On a problem of J.H.C. Whitehead concerning Abelian groups*, Bull. Acad. Polon. Sci. III **3** (1955), 127–128.
[19] A. Ehrenfeucht and J. Łoś, *Sur les produits cartésiens des groupes cycliques infinis*, Bull. Acad. Polon. Sci., Sér. Sci. Math. **2** (1954), 261–263.
[20] P.C. Eklof, *Whitehead's problem is undecidable*, Amer. Math. Monthly **83** (1976), 775–788.
[21] P.C. Eklof, *Homological algebra and set theory*, Trans. Amer. Math. Soc. **227** (1977), 207–225.
[22] P.C. Eklof, *Set Theoretic Methods in Homological Algebra and Abelian Groups*, Les Presses de L'Université de Montréal (1980).

[23] P.C. Eklof, *Set-theoretic methods: the uses of Gamma invariants*, Abelian Groups, The Proceedings of the Curaçao Conference on Torsion-Free Abelian Groups, Lecture Notes in Pure Appl. Math., Vol. 146, Marcel Dekker, New York (1993), 143–153.

[24] P.C. Eklof, *Set theory generated by Abelian group theory*, Bull. Symbolic Logic **3** (1997), 1–16.

[25] P.C. Eklof and L. Fuchs, *Baer modules over valuation domains*, Ann. Mat. Pura Appl. (IV) **150** (1988), 363–374.

[26] P.C. Eklof, L. Fuchs and S. Shelah, *Baer modules over domains*, Trans. Amer. Math. Soc. **322** (1990), 547–560.

[27] P.C. Eklof and M. Huber, *On the rank of* Ext, Math. Z. **174** (1980), 159–185.

[28] P.C. Eklof and M. Huber, *On the p-ranks of* $\mathrm{Ext}(A, G)$, *assuming* CH, Abelian Group Theory, Lecture Notes in Math., No. 874, Springer-Verlag, Berlin (1981), 93–108.

[29] P.C. Eklof and A.H. Mekler, *Almost Free Modules: Set-Theoretic Methods*, rev. edn, North-Holland, Amsterdam (2002).

[30] P.C. Eklof, A.H. Mekler and S. Shelah, *Uniformization and the diversity of Whitehead groups*, Israel J. Math. **80** (1992), 301–321.

[31] P.C. Eklof, A.H. Mekler and S. Shelah, *Hereditarily separable groups and monochromatic uniformization*, Israel J. Math. **88** (1994), 213–235.

[32] P.C. Eklof and S. Shelah, *On Whitehead modules*, J. Algebra **142** (1991), 492–510.

[33] P.C. Eklof and S. Shelah, *A combinatorial principle equivalent to the existence of non-free Whitehead groups*, Contemp. Math. **171** (1994), 79–98.

[34] P.C. Eklof and S. Shelah, *A non-reflexive Whitehead group*, J. Pure Appl. Algebra **156** (2001), 199–214.

[35] P.C. Eklof and S. Shelah, *The structure of* $\mathrm{Ext}(A, \mathbb{Z})$ *and* GCH: *possible co-Moore spaces*, Math. Z. **239** (2002), 143–157.

[36] P.C. Eklof and J. Trlifaj, *Covers induced by* Ext, J. Algebra **231** (2000), 640–651.

[37] P.C. Eklof and J. Trlifaj, *How to make* Ext *vanish*, Bull. London Math. Soc. **33** (2001), 41–51.

[38] L. Fuchs, *On character groups of discrete Abelian groups*, Acta Math. Acad. Sci. Hungar. **10** (1959), 133–140.

[39] L. Fuchs, *Abelian Groups*, Pergamon Press, Oxford (1960).

[40] L. Fuchs, *Infinite Abelian Groups*, Vols. I and II, Academic Press, New York (1970, 1973).

[41] L. Fuchs and L. Salce, *Modules over Valuation Domains*, Marcel Dekker, Oxford (1985).

[42] O. Gerstner, L. Kaup and H.G. Weidner, *Whitehead-Moduln abzählbaren Ranges über Hauptidealringen*, Arch. Math. (Basel) **20** (1969), 503–514.

[43] R. Göbel, S.V. Rychkov and B. Wald, *A general theory of slender groups and Fuchs 44-groups*, Abelian Group Theory, Lecture Notes in Math., Vol. 874, Springer-Verlag, Berlin (1981), 194–201.

[44] R. Göbel and S. Shelah, *Almost free splitters*, Colloq. Math. **8** (1999), 193–221; Correction: Colloq. Math. **88** (2001), 155–158.

[45] R. Göbel and S. Shelah, *Cotorsion theories and splitters*, Trans. Amer. Math. Soc. **352** (2001), 5357–5379.

[46] R. Göbel and J. Trlifaj, *Cotilting and a hierarchy of almost cotorsion groups*, J. Algebra **224** (2000), 110–122.

[47] R. Göbel and J. Trlifaj, *Large indecomposable roots of* Ext, J. Pure Appl. Algebra **157** (2001), 241–246.

[48] R. Göbel and B. Wald, *Wachstumstypen und schlanke Gruppen*, Symposia Math. **23** (1979), 201–239.

[49] R. Göbel and B. Wald, *Martin's Axiom implies the existence of certain slender groups*, Math. Z. **172** (1980), 107–121.

[50] M. Golasiński and D.L. Gonçalves, *On co-Moore spaces*, Math. Scand. **83** (1998), 42–52.

[51] J. Gregory, *Higher Souslin trees and the generalized continuum hypothesis*, J. Symbolic Logic **41** (1976), 663–671

[52] P. Griffith, *Separability of torsion-free groups and a problem of J.H.C. Whitehead*, Illinois J. Math. **12** (1968), 654–659.

[53] P. Griffith, *A solution to the splitting mixed problem of Baer*, Trans. Amer. Math. Soc. **139** (1969), 261–269.

[54] R. Grimaldi, *Baer and UT-modules over domains*, Ph.D. Thesis, New Mexico State Univ. (1972).

[55] R. Grossberg and S. Shelah, *On the structure of* $\mathrm{Ext}_p(G, \mathbb{Z})$, J. Algebra **121** (1989), 117–128.

[56] D. Happel and C.M. Ringel, *Tilted algebras*, Trans. Amer. Math. Soc. **274** (1982), 399–443.

[57] J. Hausen, *Automorphismen gesättige Klassen abzählbarer abelscher Gruppen*, Studies on Abelian Groups, Springer, Berlin (1968), 147–181.
[58] G. Heinlein, *Vollreflexive Ringe und schlanke Moduln*, Doctoral dissertation, Erlangen (1971).
[59] P. Hill, *On the freeness of Abelian groups: a generalization of Pontryagin's theorem*, Bull. Amer. Math. Soc. **76** (1970), 1118–1120.
[60] P. Hill, *A special criterion for freeness*, Symposia Math. XIII, Academic Press, New York (1974), 311–314.
[61] H. Hiller, M. Huber and S. Shelah, *The structure of* $\mathrm{Ext}(A, \mathbb{Z})$ *and* $\mathrm{V} = \mathrm{L}$, Math. Z. **162** (1978), 39–50.
[62] H. Hiller and S. Shelah, *Singular cohomology in L*, Israel J. Math. **26** (1977), 313–319.
[63] P.J. Hilton and U. Stammbach, *A Course in Homological Algebra*, 2nd edn, Springer-Verlag, Berlin (1997).
[64] W. Hodges, *In singular cardinality, locally free algebras are free*, Algebra Universalis **12** (1981), 205–220.
[65] M. Huber, *A simple proof for a theorem of Chase*, Rend. Sem. Mat. Univ. Padova **74** (1985), 45–49.
[66] M. Huber and R.B. Warfield, Jr., *On the torsion subgroup of* $\mathrm{Ext}(A, G)$, Arch. Math. (Basel) **32** (1979), 5–9.
[67] A. Hulanicki, *Algebraic structure of compact Abelian groups*, Bull. Acad. Polon. Sci. **6** (1958), 71–73.
[68] A.V. Ivanov, *A problem on Abelian groups*, Math. USSR-Sb. **34** (1978), 461–474.
[69] A.V. Ivanov, *Direct and complete direct sums of Abelian groups*, Vestnik Moskov. Univ. Ser. I Mat. Mekh. **37** (1979), 96 (in Russian).
[70] T. Jech, *Set Theory*, Academic Press, New York (1978).
[71] C.U. Jensen, *Les Foncteurs Dérivés de lim et leurs Applications en Théorie des Modules*, Lecture Notes in Math., Vol. 254, Springer-Verlag, Berlin (1972).
[72] R.B. Jensen, *The fine structure of the constructible hierarchy*, Ann. Math. Logic **4** (1972), 229–308.
[73] D. Kan and G. Whitehead, *On the realizability of singular cohomology groups*, Proc. Amer. Math. Soc. **12** (1961), 24–25.
[74] I. Kaplansky, *Projective modules*, Ann. of Math. (2) **68** (1958), 372–377.
[75] I. Kaplansky, *The splitting of modules over integral domains*, Arch. Math. **13** (1962), 341–343.
[76] K. Kunen, *Set Theory: An Introduction to Independence Proofs*, North-Holland, Amsterdam (1980).
[77] E.L. Lady, *Slender rings and modules*, Pacific J. Math. **49** (1973), 397–406.
[78] A. Mader, *Groups and modules that are slender as modules over their endomorphism rings*, Abelian Groups and Modules, CISM Courses and Lectures, Vol. 287, Springer-Verlag, Berlin (1984), 315–327.
[79] M. Magidor and S. Shelah, *When does almost free imply free? (For groups, transversals, etc.),* J. Amer. Math. Soc. **7** (1994), 769–830.
[80] C. Megibben, *Crawley's problem on the unique ω-elongation of p-groups is undecidable*, Pacific J. Math. **107** (1983), 205–212.
[81] A.H. Mekler, *On Shelah's Whitehead groups and* CH, Rocky Mountain J. Math. **12** (1982), 271–278.
[82] A.H. Mekler, *Proper forcing and Abelian groups*, Abelian Group Theory, Lecture Notes in Math., Vol. 1006, Springer-Verlag, Berlin (1983), 285–303.
[83] A. Mekler, A. Roslanowski and S. Shelah, *On the p-rank of* Ext, Israel J. Math. **112** (1999), 327–356.
[84] A.H. Mekler and S. Shelah, *ω-elongations and Crawley's problem*, Pacific J. Math. **121** (1986), 121–132.
[85] A.H. Mekler and S. Shelah, *The solution to Crawley's problem*, Pacific J. Math. **121** (1986), 133–134.
[86] R.J. Nunke, *Modules of extensions over Dedekind rings*, Illinois J. Math. **3** (1959), 222–241.
[87] R.J. Nunke, *On direct products of infinite cyclic groups*, Proc. Amer. Math. Soc. **13** (1962), 66–71.
[88] R.J. Nunke, *Slender groups*, Acta Sci. Math. (Szeged) **23** (1962), 67–73.
[89] R.J. Nunke, *Whitehead's problem*, Abelian Group Theory, Lecture Notes in Math., Vol. 616, Springer-Verlag, Berlin (1977), 240–250.
[90] R.J. Nunke and J.J. Rotman, *Singular cohomology groups*, J. London Math. Soc. **37** (1962), 301–306.
[91] L.S. Pontryagin, *The theory of topological commutative groups*, Ann. of Math. **35** (1934), 361–388.
[92] G.A. Reid, *Almost free Abelian groups*, Lecture Notes, Tulane University (1967), Unpublished.
[93] C.M. Ringel, *The braid group action on the set of exceptional sequences of a hereditary Artin algebra*, Abelian Group Theory and Related Topics, Contemporary Math. **171** (1994), 339–352.
[94] J.J. Rotman, *On a problem of Baer and a problem of Whitehead in Abelian groups*, Acad. Sci. Hungar. **12** (1961), 245–254.
[95] J.J. Rotman, *An Introduction to Homological Algebra*, Academic Press, New York (1979).
[96] S.V. Rychkov, *On cartesian products of Abelian groups*, Soviet Math. Dokl. **21** (1980), 747–748.

[97] S.V. Rychkov and B. Thomé, *Slender groups and related concepts*, Comm. Algebra **14** (2) (1986), 333–387.
[98] G. Sageev and S. Shelah, *Weak compactness and the structure of* $\mathrm{Ext}(A, \mathbb{Z})$, Abelian Group Theory, Lecture Notes in Math., Vol. 874, Springer-Verlag, Berlin (1981), 87–92.
[99] G. Sageev and S. Shelah, *On the structure of* $\mathrm{Ext}(A, \mathbb{Z})$ *in* ZFC^{+}, J. Symbolic Logic **50** (1985), 302–315.
[100] E. Sasaida, *Proof that every countable and reduced torsion-free Abelian group is slender*, Bull. Acad. Polon. Sci. **7** (1959), 143–144.
[101] P. Schultz, *Self-splitting groups*, Preprint series of the University of Western Australia at Perth (1980).
[102] S. Shelah, *Infinite Abelian groups, Whitehead problem and some constructions*, Israel J. Math. **18** (1974), 243–256.
[103] S. Shelah, *A compactness theorem for singular cardinals, free algebras, Whitehead problem and transversals*, Israel J. Math. **21** (1975), 319–349.
[104] S. Shelah, *Whitehead groups may not be free even assuming* CH*, I*, Israel J. Math. **28** (1977), 193–203.
[105] S. Shelah, *On uncountable Abelian groups*, Israel J. Math. **32** (1979), 311–330.
[106] S. Shelah, *Whitehead groups may not be free even assuming* CH*, II*, Israel J. Math. **35** (1980), 257–285.
[107] S. Shelah, *The consistency of* $\mathrm{Ext}(G, Z) = \mathbb{Q}$, Israel J. Math. **39** (1981), 74–82.
[108] S. Shelah, *Models with second order properties, III. Omitting types for* $L(Q)$, Arch. Math. Logik **21** (1981), 1–11.
[109] S. Shelah, *Proper Forcing*, Lecture Notes in Math., Vol. 940, Springer-Verlag, Berlin (1982).
[110] S. Shelah, *Diamonds, uniformization*, J. Symbolic Logic **49** (1984), 1022–1033.
[111] S. Shelah, *Incompactness in regular cardinals*, Notre Dame J. Formal Logic **26** (1985), 195–228.
[112] S. Shelah, *Not collapsing cardinals* $\leqslant \kappa$ *in* $(< \kappa)$*-support iterations*, Preprint.
[113] S. Shelah, *Not collapsing cardinals* $\leqslant \kappa$ *in* $(< \kappa)$*-support iterations, Part II*, Preprint.
[114] E. Specker, *Additive Gruppen von Folgen ganzer Zahlen*, Portugaliae Math. **9** (1950), 131–140.
[115] K. Stein, *Analytische Funktionen mehrerer komplexer Veränderlichen zu vorgegebenen Periodizitätsmoduln und das zweite Cousinsche Problem*, Math. Ann. **123** (1951), 201–222.
[116] J. Trlifaj, *Whitehead property of modules*, Czech. Math. J. **36** (1986), 467–475.
[117] J. Trlifaj, *Lattices of orthogonal theories*, Czech. Math. J. **39** (1989), 595–603.
[118] J. Trlifaj, *Associative rings and the Whitehead property of modules*, Algebra Berichte, Vol. 63, Verlag R. Fischer (1990).
[119] J. Trlifaj, *Von Neumann regular rings and the Whitehead property of modules*, Comment. Math. Univ. Carolinae **31** (1990), 621–625.
[120] J. Trlifaj, *Non-perfect rings and a theorem of Eklof and Shelah*, Comment. Math. Univ. Carolinae **32** (1991), 27–32.
[121] J. Trlifaj, *Whitehead test modules*, Trans. Amer. Math. Soc. **348** (1996), 1521–1554.
[122] C.A. Weibel, *An Introduction to Homological Algebra*, Cambridge Stud. Adv. Math., Vol. 38, Cambridge Univ. Press (1994).
[123] E.C. Zeeman, *On direct sums of free cycles*, J. London Math. Soc. **30** (1955), 195–212.

This Page Intentionally Left Blank

Flat Covers

Edgar E. Enochs

Department of Mathematics, University of Kentucky, Lexington, KY 40506-0027, USA
E-mail: enochs@ms.uky.edu

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1. Background

Historically there are various categorical settings where one considers a class of objects in a category which have some special properties and then asks if there is a universal map from (to) one of these special objects to (from) a given object of the category. So, for example, given any topological space X the Stone–Čech compactification of X is a right universal map $\phi : X \to \check{X}$ into a compact space $\check{X}$. This means that if $X \to Y$ is any continuous map into a compact space Y, then there is a unique continuous map $\check{X} \to Y$ so that the following diagram commutes

$$\begin{tikzcd} X \arrow[r] \arrow[dr] & \check{X} \arrow[d] \\ & Y \end{tikzcd}$$

Given any group G, $G \to G/G'$ is a right universal map into an Abelian group (where G' is the commutator subgroup). If A is an Abelian group and $t(A)$ its torsion subgroup, the canonical injection $t(A) \to A$ is a left universal map from a torsion Abelian group into A.

There are situations where a weaker notion than that of a universal morphism comes into play. For example, in the category of fields the algebraic closure $K \subset \Omega$ (thought of as a morphism $K \to \Omega$) is such that

$$\begin{tikzcd} K \arrow[r] \arrow[dr] & \Omega \arrow[d, dotted] \\ & \Omega' \end{tikzcd}$$

can be completed to a commutative diagram whenever Ω' is also algebraically closed. However, the morphism $\Omega \to \Omega'$ is far from unique in general. But the uniqueness of $K \to \Omega$ is guaranteed by the fact that

$$\begin{tikzcd} K \arrow[r] \arrow[dr] & \Omega \arrow[d, dotted] \\ & \Omega \end{tikzcd}$$

can be completed to a commutative diagram only by automorphisms of Ω.

Analogously, we have the universal covering space $U \to X$ of any connected locally Euclidean topological space X. Here U is simply connected and $U \to X$ has properties dual to those of $K \to \Omega$ above where the special of objects consists of the simply connected locally Euclidean spaces.

Baer [2] prove that every module M can be embedded in an injective module E. This means that if E' is any other injective module then the diagram

$$\begin{tikzcd} M \arrow[r] \arrow[dr] & E \arrow[d, dotted] \\ & E' \end{tikzcd}$$

can be completed to a commutative diagram. Eckmann and Schöpf [6] proved that M can be embedded as an essential submodule of an injective module and noted that if $M \subset E_1$ and $M \subset E_2$ are two injective envelopes of M then

$$\begin{tikzcd} M \arrow[r] \arrow[dr] & E_1 \arrow[d, dotted] \\ & E_2 \end{tikzcd}$$

can be completed to a commutative diagram only by isomorphisms. This is implied by the property that

$$\begin{tikzcd} M \arrow[r] \arrow[dr] & E \arrow[d, dotted] \\ & E \end{tikzcd}$$

can be completed only by automorphisms when M is an essential submodule of the injective module E. If $M \subset E$ is such an embedding, they call E an injective envelope of M.

The notion dual to that of an essential embedding (in the category of R-modules for some R) is that of a surjective morphism with a superfluous kernel. And then the notion dual to that of an injective envelope is a projective cover. Projective covers are also unique up to isomorphism.

In his University of Chicago thesis (see [3]), Bass characterized those rings R for which every left R-module has a projective cover. Using the notion of a flat module as defined by Serre in his GAGA article [25], he proved that these rings are precisely those for which every flat left R-module is also projective. As we will see in the next section, when the definition of a projective cover is given in a different but equivalent form (which leads to the notion of a flat cover), Bass' result naturally suggests the conjecture that every module has a flat cover.

2. Covers and envelopes

In [26], Sonner weakened the notion of a left and right universal morphism and gave the categorical definition of a weak left and weak right universal morphism. This eventually suggested even weaker notions, those of covers and envelopes.

DEFINITION 2.1. If $\mathcal{C}$ is a category and $\mathcal{F}$ a class of objects of $\mathcal{C}$, by an $\mathcal{F}$-pre-envelope of an object X of $\mathcal{C}$ we mean a morphism $\phi : X \to F$ with $F \in \mathcal{F}$ such that for any object $G \in \mathcal{F}$, $\mathrm{Hom}(F, G) \to \mathrm{Hom}(X, G)$ is surjective (i.e. the obvious diagrams can be completed to commutative diagrams). If furthermore any morphism $f : F \to F$ such that $f \circ \phi = \phi$ is an automorphism of F, then $\phi : X \to F$ is called an $\mathcal{F}$-envelope of X. The dual notions are those of an $\mathcal{F}$-precover and an $\mathcal{F}$-cover of X.

We see that if $\phi_1 : X \to F_1$ and $\phi_2 : X \to F_2$ are two $\mathcal{F}$-envelopes of X, then any $f : F_1 \to F_2$ such that $\phi_2 \circ f = \phi_1$ (and such exist) is an isomorphism. So envelopes and covers are unique up to isomorphism when they exist.

We will be concerned with covers and envelopes in the category of left R-modules over some ring R. Here covers and envelopes are often named after the associated class. So a flat cover of a module is an $\mathcal{F}$-cover where $\mathcal{F}$ is the class of flat modules. We will let ${}_R\mathrm{Mod}$ denote the category of left R-modules over the ring R.

DEFINITION 2.2. If $\mathcal{F}$ is a class of objects of ${}_R\mathrm{Mod}$, then a morphism $\phi : F \to M$ in ${}_R\mathrm{Mod}$ is said to be a special $\mathcal{F}$-precover of M if $F \in \mathcal{F}$, ϕ is surjective and if $\mathrm{Ext}^1(G, \ker(\phi)) = 0$ for all $G \in \mathcal{F}$. The dual notion is that of a special $\mathcal{F}$-pre-envelope.

We note that a special $\mathcal{F}$-precover $F \to M$ is in fact an $\mathcal{F}$-precover, for $\mathrm{Hom}(G, F) \to \mathrm{Hom}(G, M) \to \mathrm{Ext}^1(G, \ker(\phi)) = 0$ is exact for all $G \in \mathcal{F}$.

We will see that for nice classes $\mathcal{F}$, all $\mathcal{F}$-covers are special. Recall that we say $\mathcal{F}$ is closed under extensions if whenever $0 \to F' \to F \to F'' \to 0$ is exact with $F', F'' \in \mathcal{F}$ then $F \in \mathcal{F}$.

THEOREM 2.3. *If $\mathcal{F}$ is a class of objects in ${}_R\mathrm{Mod}$ which is closed under extensions and if $\phi : F \to M$ is an $\mathcal{F}$-cover then if $A \subset B$ is a submodule of B such that $B/A \in \mathcal{F}$ then any diagram*

$$\begin{tikzcd} A \arrow[r] \arrow[d] & B \arrow[d] \arrow[dl, dotted] \\ F \arrow[r] & M \end{tikzcd}$$

can be completed to a commutative diagram.

PROOF. We form the pushout of

$$\begin{tikzcd} A \arrow[r] \arrow[d] & B \\ F & \end{tikzcd}$$

and get the commutative diagram

$$\begin{tikzcd} 0 \arrow[r] & A \arrow[r] \arrow[d] & B \arrow[r] \arrow[d] & G \arrow[r] \arrow[d, equal] & 0 \\ 0 \arrow[r] & F \arrow[r] & H \arrow[r] & G \arrow[r] & 0 \end{tikzcd}$$

with exact rows. Since $F, G \in \mathcal{F}$ we get $H \in \mathcal{F}$. By the pushout property we get a morphism $H \to M$ with $F \to H \to M$ the original cover ϕ. But $H \to M$ can be factored $H \to F \xrightarrow{\phi} M$. But then $F \to H \to F$ is an automorphism of F, say f. Then we check that $B \to H \to F \xrightarrow{f^{-1}} F$ is the desired morphism. $\square$

The next corollary is usually called Wakamatsu's lemma.

COROLLARY. $\operatorname{Ext}^1(G, \ker(\phi)) = 0$ *for all* $G \in \mathcal{F}$.

PROOF. Let $0 \to S \to P \to G \to 0$ be exact with P projective. Let $S \to \ker(\phi)$ be any morphism. Then

$$\begin{tikzcd} S \arrow[r, hook] \arrow[d] & P \arrow[d] \arrow[dl, dotted] \\ F \arrow[r] & M \end{tikzcd}$$

can be completed to a commutative diagram. But then $P \to F$ has its image in $\ker(\phi)$ and so $S \to \ker(\phi)$ has an extension $P \to \ker(\phi)$. This gives that $\operatorname{Ext}^1(G, \ker(\phi)) = 0$. This result shows the importance of studying the module C such that $\operatorname{Ext}^1(F, C) = 0$ for all $F \in \mathcal{F}$ when considering the question of the existence of $\mathcal{F}$-covers and precovers. $\square$

3. Cotorsion modules

A module C is said to be a cotorsion if $\operatorname{Ext}^1(F, C) = 0$ for all left R-modules, or equivalently if all short exact sequences $0 \to C \to U \to F \to 0$ with F flat split. We note that any such sequence is pure exact, for if M is a right R-module then

$$0 = \operatorname{Tor}_1(M, F) \to M \otimes C \to M \otimes C \to M \otimes U \to M \otimes F \to 0$$

is exact. A module P is pure injective if every pure short exact sequence $0 \to P \to U \to V \to 0$ splits. From the above we see that all pure injective modules are cotorsion. This gives us a plentiful supply of cotorsion modules.

The class of cotorsion modules is closed under extensions, products and summands. We have the following change of ring result.

LEMMA 3.1. *If $R \to R'$ is any ring homomorphism and if C is a cotorsion left R'-module, then C is a cotorsion left R-module.*

PROOF. Given the flat left R-module F and an exact sequence $0 \to S \to P \to F \to 0$ with P projective, let $S \to C$ be R-linear. Then since F is flat the sequence is pure exact, so $0 \to R' \otimes S \to R' \otimes P \to R' \otimes F \to 0$ is exact and $R' \otimes F$ is a flat R'-module. Since the R'-linear extension $R' \otimes S \to C$ can be extended to $R' \otimes P \to C$, we see that $S \to C$ has an R-linear extension $P \to C$. Hence C is a cotorsion R-module. $\square$

LEMMA 3.2. *If R is a commutative ring, M is an R-module and E is an injective R-module then* $\operatorname{Hom}_R(M, E)$ *is pure injective.*

PROOF. If $S \subset N$ is pure then $0 \to S \otimes M \to N \otimes M$ is exact. So since E is injective, $\operatorname{Hom}(N \otimes M, E) \to \operatorname{Hom}(S \otimes M, E) \to 0$ is exact, i.e.

$$\operatorname{Hom}\big(N, \operatorname{Hom}(M, E)\big) \to \operatorname{Hom}\big(S, \operatorname{Hom}(M, E)\big) \to 0$$

is exact. So $\operatorname{Hom}(M, E)$ is pure injective. □

From the above we get that if R is commutative, Noetherian and if $P \in \operatorname{Spec}(R)$, then any Matlis reflexive $\widehat{R}_P$-module is cotorsion as an R-module. So all finitely generated and all Artinian $\widehat{R}_P$-modules are cotorsion R-modules.

EXAMPLES. If R is a commutative local ring (so Noetherian) with residue field k then $\widehat{R} \cong \operatorname{Hom}(E(k), E(k))$ (by Matlis [21]).

So $\widehat{R}$ is a cotorsion R-module. More generally, if R is commutative Noetherian and $P \in \operatorname{Spec}(R)$, then by the same type reasoning $\widehat{R}_P$ is a cotorsion R-module. By Ishikawa [20], if E, E' are injective over a commutative Noetherian R then $\operatorname{Hom}(E, E')$ is a flat R-module. So we see that for such R, each $\widehat{R}_P$ is both cotorsion and flat as an R-module.

For a commutative, Noetherian R and for $P \in \operatorname{Spec}(R)$, $\operatorname{Hom}(E(R/P), E(R/P)^{(I)})$ is the P-adic completion of the free $\widehat{R}_P$-module $\operatorname{Hom}(E(R/P), E(R/P))^{(I)} = \widehat{R}_P^{(I)}$ (for any set I). Such a module is denoted T_P and is uniquely determined by the cardinality of the base of the free module.

4. Set theoretic homological algebra

Let λ be an ordinal number and let $(M_\alpha)_{\alpha<\lambda}$ be a family of submodules of M. We say the family is a (well ordered) chain of submodules if $\alpha \leqslant \beta < \lambda$ implies $M_\alpha \subset M_\beta$. We say the chain is furthermore continuous if $M_\beta = \bigcup_{\alpha<\beta} M_\alpha$ for each limit ordinal $\beta < \lambda$.

THEOREM 4.1 (see [7]). *Let M, N be modules and suppose that M is the union of a continuous chain $(M_\alpha)_{\alpha \leqslant \lambda}$ of submodules. If* $\operatorname{Ext}^1(M_0, N) = 0$ *and if* $\operatorname{Ext}^1(M_{\alpha+1}/M_\alpha, N) = 0$ *whenever $\alpha + 1 \leqslant \lambda$, then* $\operatorname{Ext}^1(M, N) = 0$.

PROOF. We prove by transfinite induction on $\alpha \leqslant \lambda$ that $\operatorname{Ext}^1(M_\alpha, N) = 0$. If $\alpha = 0$, this is part of the hypotheses. Suppose $\beta \leqslant \lambda$ and that $\operatorname{Ext}^1(M_\alpha, N) = 0$ for $\alpha < \beta$. If β is not a limit ordinal, let $\beta = \alpha + 1$. Then since $\operatorname{Ext}^1(M_\alpha, N) = 0$ and $\operatorname{Ext}^1(M_{\alpha+1}/M_\alpha, N) = 0$ we get that $\operatorname{Ext}^1(M_{\alpha+1}, N) = 0$, i.e. that $\operatorname{Ext}^1(M_\beta, N) = 0$.

Now assume that β is a limit ordinal. Let $0 \to N \xrightarrow{f} U \xrightarrow{g} M_\beta \to 0$ be a short exact sequence. We must prove that this sequence splits, that is we should prove there is a section $s : M_\beta \to U$ (so with $g \circ s = \mathrm{id}_{M_\beta}$). For $\alpha < \beta$ we have the exact sequence By hypothesis each of these sequences splits and so has a section s_α. We use transfinite induction again to prove that we can find compatible such sections, that is, such that if $\alpha \leqslant \alpha' < \beta$ then $s'_\alpha | M_\alpha = s_\alpha$. If we have compatible sections for all α with $\alpha < \tau \leqslant \beta$ where τ is a limit ordinal, then let s_τ agree with each s_α on M_α for $\alpha < \tau$. Then $s_\tau | M_\alpha = s_\alpha$ for $\alpha < \tau$.

So the problem reduces to arguing that given a section s_α where $\alpha + 1 \leqslant \lambda$, there is a section $s_{\alpha+1}$ that agrees with s_α on M_α. Since by hypothesis, $\mathrm{Ext}^1(M_{\alpha+1}, N) = 0$, there is a section $s'_{\alpha+1}$ for $g_{\alpha+1}$.

We have the linear $s_\alpha - (s'_{\alpha+1} | M_\alpha) : M_\alpha \to U$. Since $\mathrm{Ext}^1(M_{\alpha+1}/M_\alpha, N) = 0$, this map has an extension $t : M_{\alpha+1} \to U$. But then if $s_{\alpha+1} = s'_{\alpha+1} + t$, we see that $s_{\alpha+1}$ is the desired section.

In what follows we will use the current convention that the cardinal numbers are a subclass of the class of ordinal numbers and that every ordinal number β is the set of ordinal numbers α with $\alpha < \beta$. If we want to emphasize that we are considering an infinite cardinal number, we will use the notation $\aleph$ or $\aleph_\alpha$. And then $\aleph_{\alpha+1}$ will denote the successor cardinal to $\aleph_\alpha$.

We need the following result about ordinal numbers. Given any ordinal u there is a limit ordinal λ with $\mu < \lambda$ so that if $(\beta_\alpha)_{\alpha<\mu}$ is any family of ordinals with each $\beta_\alpha < \lambda$, then $\sup \beta_\alpha < \lambda$. It is easy to find such a λ. For example, if $\mu \leqslant \aleph_\alpha$ for the infinite cardinal $\aleph_\alpha$, let $\lambda = \aleph_{\alpha+1}$. Then for such a family $(\beta_\alpha)_{\alpha<\mu}$ we have

$$\mathrm{card}\left(\sum_{\alpha<\mu} \beta_\alpha\right) \leqslant \aleph_\alpha \cdot \mathrm{card}(\mu) \leqslant \aleph_\alpha^2 = \aleph_\alpha < \lambda$$

and so $\sum_{\alpha<\mu} \beta_\alpha < \lambda$. Hence $\sup_{\alpha<\mu} \beta_\alpha < \lambda$. Now given a set X, let $\mathrm{card}(X) = \mu$ and let λ be as above. Suppose a set Y is the union of a chain $(Y_\alpha)_{\alpha<\lambda}$ of subsets of Y. Then if $f : X \to Y$ is any function, there is an $\alpha < \lambda$ such that $f(X) \subset Y_\alpha$. We can suppose $X = \mu$. Then for $\alpha < \mu$ (so $\alpha \in \mu$) there is a least ordinal $\beta_\alpha < \lambda$ such that $f(\alpha) \in Y_{\beta_\alpha}$. Since $\sup_{\alpha<\mu}(\beta_\alpha) < \lambda$ we see that there is such an $\alpha < \lambda$. We will use this result for modules in the following way. Let M be a module and let $0 \to S \to P \to M \to 0$ be exact with P projective. Let N be the continuous union of a chain of submodules $(N_\alpha)_{\alpha<\lambda}$ for some limit ordinal λ. Suppose λ is such that any function $S \to N$ can be factored $S \to N_\alpha \to N$ for some $\alpha < \lambda$ (so in particular this holds for any linear $S \to N$). Then suppose that any morphism $S \to N_\alpha$ admits an extension $P \to N_{\alpha+1}$. Then we see that any morphism $S \to N$ admits an extension $P \to N$. And so $\mathrm{Ext}^1(M, N) = 0$. □

5. Flat precovers

In this section we prove that every module has a flat precover. To do so, we first argue that modules have special cotorsion envelopes. For this argument we need the fact that there is a single flat module G which is a test module for cotorsion, i.e. such that C is cotorsion if and only if $\mathrm{Ext}^1(G, C) = 0$.

We recall that a submodule $S \subset M$ of a module M (over R) is a pure submodule if and only if for every finitely generated $T \subset R^n$ (for any $n \geqslant 1$), any linear $T \to S$ has an extension $R^n \to M$ if and only if there is an extension $R^n \to S$. Using this, it is easy to prove that if $\aleph$ is an infinite cardinal with $\operatorname{card}(R) \leqslant \aleph$ and if $x \in M$, there is a pure submodule $S \subset M$ with $x \in S$ and $\operatorname{card}(S) \leqslant \aleph$.

The essential step is then as follows. Let $U \subset M$ has $\operatorname{card}(U) \leqslant \aleph$. For each $T \subset R^n$ and $T \to U$ such that there is an extension $R^n \to M$, make a choice of such an extension $\phi\colon R^n \to M$. Then letting U' be the submodule generated by U and all such $\phi(R^n)$, we have $U \subset U'$, $\operatorname{card}(U') \leqslant \aleph$ and that if some $T \to U$ has an extension $R^n \to M$ (for some finitely generated $T \subset R^n$, $n \geqslant 1$) then there is an extension $R^n \to U'$. Repeating the procedure with U' replacing U, we get $U' \subset U''$ and then $U'' \subset U'''$ etcetera. If $S = \bigcup_{n\geqslant 1}^{(n)}$ we see that $U \subset S$, $\operatorname{card}(S) \leqslant \aleph$ and $V \subset M$ is pure (any $T \to S$ can be factored $T \to U^{(n)} \to S$ for some $n \geqslant 1$ since T is finitely generated). Applying this with $U = Rx$ we get the desired result.

Now let F be a flat R-module. If $x \in F$ let $x \in F_0 \subset F$ with $F_0 \subset F$ a pure submodule such that $\operatorname{card}(F_0) \leqslant \aleph$. So F_0 and F/F_0 are flat. If $y + F_0 \subset F/F_0$, let $F_1/F_0 \subset F/F_0$ be pure with $y + F_0 \in F_1/F_0$ (i.e. $y \in F_1$) and with $\operatorname{card}(F_1/F_0) \leqslant \aleph$. Then $F_1 \subset F$ is pure and so F/F_1 is flat. Continuing in this manner while letting $F_\beta = \bigcup_{\alpha<\beta} F_\alpha$ when β is a limit ordinal and when we already have the F_α for $\alpha < \beta$ we see that we have the next result.

PROPOSITION 5.1. *If* $\operatorname{card}(R) \leqslant \aleph$ *where* $\aleph$ *is an infinite cardinal, then every flat left* R*-module* F *is the union of a continuous chain of pure submodules* $(F_\alpha)_{\alpha<\lambda}$ *(for some ordinal* λ*) where* $\operatorname{card}(F_0) \leqslant \aleph$ *and* $\operatorname{card}(F_{\alpha+1}/F_\alpha) \leqslant \aleph$ *whenever* $\alpha + 1 < \lambda$.

We know that if C is a module and if $\operatorname{Ext}^1(F_0, C) = 0$ and $\operatorname{Ext}^1(F_{\alpha+1}/F_\alpha, C) = 0$ where $\alpha + 1 < \lambda$ then $\operatorname{Ext}^1(F, C) = 0$.

Now note that each of F_0 and $F_{\alpha+1}/F_\alpha$ is flat and has cardinality at most $\aleph$. From this it follows that a module C is cotorsion if and only if $\operatorname{Ext}^1(H, C) = 0$ for all flat modules with $\operatorname{card}(H) \leqslant \aleph$. Letting G be a direct sum of a set of representatives of such modules H, we see that G is flat and that the single G tests whether a module C is cotorsion. This just means that C is cotorsion if and only if $\operatorname{Ext}^1(G, C) = 0$.

THEOREM 5.2. *For any ring* R*, every left* R*-module* N *has a special cotorsion pre-envelope.*

PROOF. Let G be as above. Let $C_0 = N$. We use transfinite induction to construct a continuous chain of modules $(C_\alpha)_{\alpha<\lambda}$ for $\lambda > 0$ any ordinal number. We only need to construct $C_{\alpha+1}$, given C_α. We let $0 \to S \to P \to G \to 0$ be exact with P projective. We have the pushout diagram

$$\begin{tikzcd}
S^{(\operatorname{Hom}(S,C_\alpha))} \arrow[r] \arrow[d] & C_\alpha \arrow[d] \\
P^{(\operatorname{Hom}(S,C_\alpha))} \arrow[r] & C_{\alpha+1}
\end{tikzcd}$$

where the top horizontal morphism is the evaluation morphism. Then $C_\alpha \subset C_{\alpha+1}$ and any linear $S \to C_\alpha$ has an extension $P \to C_{\alpha+1}$.

Now assume that λ is as in the last section with respect to S and $C = \bigcup_{\alpha<\lambda} C_\alpha$, i.e. that every linear $S \to C$ factors $S \to C_\alpha \to C$ for some $\alpha < \lambda$. But then we see that $\mathrm{Ext}^1(G, C) = 0$. But this means that C is cotorsion. Now we note that by the pushout diagram above we get that

$$C_{\alpha+1}/C_\alpha \cong (P/S)^{(\mathrm{Hom}(S, C_\alpha))} = G^{(\mathrm{Hom}(S, C_\alpha))}.$$

So since G is flat, $C_{\alpha+1}/C_\alpha$ is flat. But this easily implies $C/M = C/C_0$ is a flat module. And so

$$0 \to N \to C \to \frac{C}{N} \to 0$$

gives the desired special cotorsion pre-envelope of N.

The above proof is essentially that given in Eklof and Trlifaj [8]. They in turn credit Göbel and Shelah [15] with the idea for the method of construction of the cotorsion module C in the theorem. A similar argument can be found in Quillen [23] (also see Hovey [19] and Hirschhorn [18]). But in fact Baer [2] used a related construction in his original proof that every module is a submodule of an injective module. □

THEOREM 5.3. *Every module has a flat precover.*

PROOF. Let M be any module. Let $0 \to S \to P \to M \to 0$ be exact with P projective. From this we know there is an exact sequence $0 \to S \to C \to F \to 0$ with C cotorsion and F flat. Forming a pushout of

$$\begin{tikzcd} S \arrow[r] \arrow[d] & P \\ C & \end{tikzcd}$$

we get the commutative diagram

$$\begin{tikzcd} & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & \\ 0 \arrow[r] & S \arrow[r] \arrow[d] & P \arrow[r] \arrow[d] & M \arrow[r] \arrow[d, equal] & 0 \\ 0 \arrow[r] & C \arrow[r] \arrow[d] & G \arrow[r] \arrow[d] & M \arrow[r] \arrow[d] & 0 \\ 0 \arrow[r] & F \arrow[r] \arrow[d] & F \arrow[r] \arrow[d] & 0 & \\ & 0 & 0 & & \end{tikzcd}$$

with exact rows and columns. Since P and F are flat, we see that G is flat. Since P and F are flat, we see that G is flat. Since C is cotorsion, the exact sequence $0 \to C \to G \to M \to 0$ shows that $G \to M$ is a special flat precover.

The argument above was first used by Salce [24] in a different but related setting. □

6. Flat covers

In general proving the existence of a cover (or envelope) a certain type breaks up into two distinct parts. These are to prove that a precover exists and then to prove that the existence of a precover implies the existence of a cover. In this section we sketch a proof that the existence of a flat precover implies the existence of a flat cover in the category of left R-modules. The basic fact that we need is that the class of flat modules is closed under direct limits. We will also use the trivial but useful fact that if $\phi : F \to M$ is a flat precover and if $F \to G \to M$ is a factorization of ϕ where G is flat, then $G \to M$ is also a flat precover.

The method of proof consists of a sequence of constructions by transfinite induction.

LEMMA 6.1. *Given a flat precover $\phi : F \to M$, there exists a flat cover $\psi : G \to M$ and an $f : F \to G$ with $\psi \circ f = \phi$, such that if $\psi' : G' \to M$ is any other flat precover and $g : G \to G'$ any morphism such that $\psi' \circ g = \psi$, then* $\ker(g \circ f) = \ker(f)$.

PROOF. If not, given any ordinal we can use transfinite induction to construct an inductive system $((F_\alpha), (f_{\beta\alpha}))$ (for all $\alpha < \lambda$ and all $\alpha \leqslant \beta < \lambda$) along with a map (ϕ_α) of the system to M such that $F_0 = F$, $\phi_0 = \phi$, such that $\phi_\alpha : F_\alpha \to M$ is a flat precover for each α, such that $\ker(F \to F_\alpha) = \ker(F_0 \to F_\alpha) \nsubseteq \ker(F \to F_{\alpha+1})$ when $\alpha + 1 < \lambda$ and such that $F_\beta = \varinjlim_{\alpha<\beta}$ when $\beta < \lambda$ is a limit ordinal (and such that $f_{\beta\alpha} : F_\alpha \to F_\beta$ is that given by the map of each F_α into the limit).

But then we see that $\ker(F \to F_\alpha)_{\alpha<\lambda}$ is a strictly increasing chain of submodules of F. This implies $\operatorname{card}(F) \geqslant \operatorname{card}(\lambda)$. But since λ is arbitrary, this is impossible. □

LEMMA 6.2. *If M has a flat precover, then M has a flat precover $\phi : F \to M$ with the property that if $\psi : G \to M$ is any flat precover and if $f : F \to G$ is such that $\psi \circ f = \phi$ then f is an injection.*

PROOF. We use the preceding lemma and construct an inductive system $((F_n), (f_{mn}))$ $(n \leqslant m < \omega)$ and a map (ϕ_n) of the system into M such that each $\phi_n : F_n \to M$ is a flat precover and such that for each $n < w$, $f_{n+1,n} : F_n \to F_{n+1}$ satisfies the conclusion of the preceding lemma.

Then if $F = \varinjlim F_n$ and if $\phi : F = \varinjlim F_n \to M$ is given by the morphisms $\phi_n : F_n \to M$, then it is easy to see that $\phi : F \to M$ is the desired precover. □

LEMMA 6.3. *If $\phi : F \to M$ is a flat precover such that every $f : F \to F$ with $\phi \circ f = \phi$ is an injection, then $\phi : F \to M$ is a flat cover.*

PROOF. Suppose $f : F \to F$ is such that $\phi \circ f = \phi$ and that f is not surjective (but is injective).

Again we let λ be an ordinal and we construct an inductive system $((F_\alpha), (f_{\beta\alpha}))$ $(\alpha \leqslant \beta \leqslant \lambda)$ and a map (ϕ_α) of the system to M. But in this case each $F_\alpha = F$ and each $\phi_\alpha = \phi$, and when $\alpha + 1 \leqslant \lambda$, $f_{\alpha+1,\alpha} = f$.

When $\beta \leqslant \lambda$ is a limit ordinal we note that $\varinjlim_{\alpha<\beta} F_\alpha \to M$ can be factored $\varinjlim F_\alpha \to F \xrightarrow{\phi} M$, so with $F_\beta = F$ and $\alpha' < \beta$, $P_{\beta\alpha'}$ is given by the composition $F_{\alpha'} \to \varinjlim F_\alpha \to F$.

But now we see that when $\alpha < \beta \leqslant \lambda$ we have $\operatorname{im}(F_\alpha \to F_\lambda) \nsubseteq \operatorname{im}(F_\beta \to F_\lambda)$. But this implies $\operatorname{card}(F) = \operatorname{card}(F_\lambda) \geqslant \operatorname{card}(\lambda)$ and this is again impossible for all λ. □

From the above and the previous section we see that every module has a flat cover. El Bashir [5] using a different approach proved a result that then he and Bican argued implied that modules have flat covers.

The existence of flat covers of modules was conjectured in 1981 in [9]. In 1995 Xu [28] made the first major step toward settling the conjecture. He proved that the conjecture holds for commutative, Noetherian rings of finite Krull dimension. His proof was more structural than that given in this chapter and subsequently led to his definition of the dual Bass numbers [29,14].

Let R be commutative Noetherian and let $P \in \operatorname{Spec}(R)$. Then $\widehat{R}_P$ is a flat R-module and $\widehat{R}_P \to k(P)$ (with $k(P)$ the residue field of $\widehat{R}_P$) has a cotorsion (as an R-module) kernel. Hence $\widehat{R}_P \to k(P)$ is a (special) flat precover. But any R-linear $\widehat{R}_P \to \widehat{R}_P$ is also $\widehat{R}_P$-linear so we see that $\widehat{R}_P \to k(P)$ is a flat cover. So, for examples, $\widehat{Z}_p \to Z/(p)$ is a flat cover over Z and $\phi : k[[x_1, \ldots, x_n]) \to k$ $(\phi(S) = S(0, 0, \ldots, 0))$ is a flat cover over $k[x_1, \ldots, x_n]$ where k is a field.

More generally given any T_P (see Section 3) we have that $T_P \to T_P/PT_P$ is a flat cover over R. This result can be used to prove a structure theorem.

THEOREM 6.4. *If R is commutative, Noetherian and F is a flat and cotorsion module then F is uniquely up to isomorphism a product* $\prod T_P (P \in \operatorname{Spec}(R))$ (*see Xu* [30]).

This result will be used in the next section.

7. Dual Bass numbers

If R is commutative, Noetherian and

$$0 \to M \to E^0(M) \to E^1(M) \to \cdots$$

is a minimal injective resolution of the R-module M, then by Matlis each $E^n(M)$ is uniquely up to isomorphism the direct sum of copies of the $E(R/P)$ where $P \in \operatorname{Spec}(R)$. The cardinality of the number of copies of $E(R/P)$ in such a decomposition of $E^n(R)$ is called a Bass number of M and is denoted $\mu_n(P, M)$. These invariants of modules are an important tool in homological commutative algebra. On the other hand, if M is finitely generated, the Betti numbers $\beta(P, M)$ are defined by using the minimal projective resolution of M_P over the local ring R_P. If M is not finitely generated, Xu defines invariants of M which are in some sense dual to the Bass numbers of M as follows.

We know that M has a minimal flat resolution

$$\cdots \to F_1(M) \to F_0(M) \to M \to 0$$

constructed by taking successive flat covers. By Wakamatsu's lemma each $\ker(F_0(M) \to M)$, $\ker(F_i(M) \to F_{i-1}(M))$ for $i \geqslant 1$ is cotorsion. But this implies that each $F_i(M)$ for $i \geqslant 1$ is both cotorsion and flat ($F_0(M)$ will not in general be cotorsion).

Hence each $F_i(M)$ $(i \geqslant 1)$ is uniquely a product $\prod T_P$ $(P \in \operatorname{Spec}(R))$. The cardinal numbers $\pi_n(P, M)$ are defined as the rank (possibly infinite) of the free R_P-module whose completion is T_P.

It is known that the Bass numbers can be computed by the formula $\mu_i(P, M) = \dim_{k(P)} \operatorname{Ext}^i_{R_P}(k(P), M_P)$. The dual Bass numbers are in turn given by

$$\prod_i (P, M) = \dim_{k(P)} \operatorname{Tor}_i^{R_P}\big(k(P), \operatorname{Hom}_R(R_P, M)\big)$$

when M is cotorsion. Here we see that the localization $M_P = R_P \otimes_R M$ in the first formula is replaced by the "colocalization" (see Melkersson and Schenzel [22]) $\operatorname{Hom}_R(R_P, M)$ in the second. And, of course, Ext is replaced by Tor.

8. Further developments

Given a topological space X and a sheaf $\mathcal{O}$ of rings on X, by an $\mathcal{O}$-module we mean a sheaf F on X with $F(U)$ an $\mathcal{O}(U)$-module for all open $U \subset X$ subject to the condition that the restriction maps $F(U) \to F(V)$ $(V \subset U)$ are $\mathcal{O}(U)$-linear. It is known that in general there aren't enough projective $\mathcal{O}$-modules. But the argument proving the existence of flat covers of modules can be modified to prove that $\mathcal{O}$-modules have flat covers where F is flat if the functor $F \otimes_{\mathcal{O}} -$ is exact (see [12]).

If X is a scheme with $\mathcal{O}$ its structure sheaf (see Hartshorne [17] or Grothendieck and Dieudonné [16]) and if G is a quasi-coherent $\mathcal{O}$-module, it is not known whether G has a quasi-coherent flat cover.

If Q is a quiver and R a ring, then a representation of Q in the category of left R-modules is said to be flat if it is the direct limit of projective representations. Then it is not known whether flat covers exist in this category. However, if Q has no path of the form

$$\cdots \to \circ \to \circ \to \circ$$

then it is known that flat covers exist [12].

Flat covers also exist in the category of complexes of R-modules for any ring R [10].

References

[1] M. Auslander and S. Smalø, *Preprojective modules over Artin rings*, J. Algebra **66** (1980), 61–122.

[2] R. Baer, *Abelian groups which are direct summands of every containing group*, Bull. Amer. Math. Soc. **46** (1940), 800–806.

[3] H. Bass, *Finitistic dimension and a homological generalization of semi-primary rings*, Trans. Amer. Math. Soc. **95** (1960), 466–488.

[4] H. Bass, *On the ubiquity of Gorenstein rings*, Math. Z. **82** (1963), 8–28.

[5] L. Bican, R. El Bashir and E. Enochs, *All modules have flat covers*, Bull. London Math. Soc., to appear.

[6] B. Eckmann and A. Schöpf, *Über injektive Moduln*, Arch. Math. **4** (1953), 75–78.

[7] P.C. Eklof and A.H. Mekler, *Almost Free Modules: Set-Theoretic Methods*, North-Holland, Amsterdam (1990).

[8] P.C. Eklof and J. Trlifaj, *How to make* Ext *vanish*, Bull. Amer. Math. Soc., to appear.

[9] E. Enochs, *Injective and flat covers and resolvents*, Israel J. Math. **39** (1981), 189–209.

[10] E. Enochs and J.R. Garcia Rozas, *Flat covers of complexes*, J. Algebra **210** (1998), 86–102.

[11] E. Enochs and O. Jenda, *Relative Homological Algebra*, De Gruyter Expositions in Math., Vol. 30, Walter de Gruyter, Berlin–New York (2000).

[12] E. Enochs and L. Oyonarte, *Covers, envelopes and cotorsion theories*, Manuscript.

[13] E. Enochs and L. Oyonarte, *Flat covers and cotorsion envelopes of sheaves*, Submitted.

[14] E. Enochs and J. Xu, *On invariants dual to the Bass numbers*, Proc. Amer. Math. Soc. **125** (1997), 951–960.

[15] R. Göbel and S. Shelah, *Cotorsion theories and splitters*, Trans. Amer. Math. Soc., to appear.

[16] A. Grothendieck and J. Dieudonné, *Eléments de Géométrie Algébrique I*, Springer, Berlin (1971).

[17] R. Hartshorne, *Algebraic Geometry*, Graduate Texts in Math., Vol. 52, Springer, Berlin (1977).

[18] P. Hirschhorn, *Localization, cellularization, and homotopy colimits*, Preprint (1997).

[19] M. Hovey, *Model Categories*, Mathematical Surveys and Monographs, Vol. 63, Amer. Math. Soc. (1991).

[20] T. Ishikawa, *On injective modules and flat modules*, J. Math. Soc. Japan **17** (1965), 291–296.

[21] E. Matlis, *Injective modules over Noetherian rings*, Pacific J. Math. **8** (1958), 511–528.

[22] L. Melkersson and P. Schenzel, *The co-localization of an Artinian module*, Proc. Edinburgh Math. Soc. **38** (1995), 121–131.

[23] P. Quillen, *Homotopical Algebra*, Lecture Notes in Math., Vol. 43, Springer, Berlin (1967).

[24] L. Salce, *Cotorsion theories for Abelian groups*, Symposia Math. **23** (1979), 11–32.

[25] J.P. Serre, *Geométrie algébrique et geométrie analytique*, Ann. Inst. Fourier **6** (1956), 669–719.

[26] J. Sonner, *Universal and special problems*, Math. Z. **82** (1963), 200–211.

[27] T. Wakamatsu, *Stable equivalence of self-injective algebras and a generalization of tilting modules*, J. Algebra **134** (1990), 298–325.

[28] J. Xu, *The existence of flat covers over Noetherian rings of finite Krull dimension*, Proc. Amer. Math. Soc. **123** (1995), 27–32.

[29] J. Xu, *Minimal injective and flat resolutions of modules over Gorenstein rings*, J. Algebra **175** (1995), 451–477.

[30] J. Xu, *Flat Covers of Modules*, Lecture Notes in Math., Vol. 1634, Springer, Berlin (1996).

The Krull–Schmidt Theorem

Alberto Facchini*

Dipartimento di Matematica Pura e Applicata, Università di Padova, 35131 Padova, Italy

Contents

*Partially supported by Gruppo Nazionale Strutture Algebriche e Geometriche e loro Applicazioni of Istituto Nazionale di Alta Matematica, Italy, and by Ministero dell'Università e della Ricerca Scientifica e Tecnologica (progetto di ricerca di rilevante interesse nazionale "Nuove prospettive nella teoria degli anelli, dei moduli e dei gruppi abeliani"), Italy.

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1. Introduction

In this chapter, we shall consider unital right modules over associative rings with identity $1 \neq 0$. If M is a module and $\{M_i \mid i \in I\}$, $\{N_j \mid j \in J\}$ are two families of submodules of M such that $M = \bigoplus_{i \in I} M_i = \bigoplus_{j \in J} N_j$, we say that these two direct sum decompositions of M are *isomorphic* if there is a one-to-one correspondence $\varphi : I \to J$ such that $M_i \cong N_{\varphi(i)}$ for every $i \in I$. A module is called *indecomposable* if it cannot be expressed as a direct sum of two of its non-zero submodules. The "classical" Krull–Schmidt theorem (Theorem 2.1) asserts that any two direct sum decompositions of a module of finite length into indecomposable summands are isomorphic.

The origin of the Krull–Schmidt theorem goes back to group theory. Joseph Henry Maclagan Wedderburn [W9] proved in 1909 that any two direct product decompositions of a finite group G into indecomposable factors $G = H_1 \times \cdots \times H_r = K_1 \times \cdots \times K_s$ are isomorphic. Later, Remak [R2] showed that the indecomposable factors are centrally isomorphic (that is, $r = s$ and there is an automorphism σ of G that is the identity modulo the center $Z(G)$ of G such that, after suitable relabeling of the indices if necessary, $\sigma(H_i) = K_i$ for every i). Subsequently, Krull [K4, Fundamentalsatz] and Schmidt [S1,S2] extended this result to "generalized Abelian groups" with the acc and the dcc on admissible subgroups. In 1950 Azumaya [A3] extended the theorem to the case of arbitrary (i.e. possibly infinite) direct sums of modules with local endomorphism rings. This extension of the Krull–Schmidt theorem is the celebrated Krull–Schmidt–Azumaya theorem (Theorem 2.2).

The problem considered in the Krull–Schmidt theorem, that is, asking whether the direct sum decomposition of a module into indecomposable summands is unique up to isomorphism, is very natural and is the topic of this chapter. After having proved the theorem for the class of modules of finite length, W. Krull [K6, last sentence of the paper] asked in 1932 whether the theorem holds also for the class of Artinian modules (for the answers to this question see Section 4.2). Similar problems were subsequently posed for a number of classes of modules. If $\mathcal{C}$ is a class of right modules over a ring R, we shall say that the *Krull–Schmidt theorem holds for* $\mathcal{C}$ if the class $\mathcal{C}$ is closed under direct summands, every module $M \in \mathcal{C}$ is a direct sum of indecomposable modules, and all direct sum decompositions of M into indecomposable direct summands are isomorphic. Obviously, it would be better to say that the Krull–Schmidt *property* holds for $\mathcal{C}$, but the expression "the Krull–Schmidt theorem holds for $\mathcal{C}$" is charming, stimulating and widespread. Thus the Krull–Schmidt theorem holds for the class of modules of finite length. For another example of class of modules for which the Krull–Schmidt theorem holds, consider the ring of integers $R = \mathbf{Z}$ and the class FG-$\mathbf{Z}$ of finitely generated Abelian groups. Every finitely generated Abelian group is a direct sum of copies of $\mathbf{Z}$ and $\mathbf{Z}/\mathbf{Z}p^n$, where p is a prime and n is a natural number (these are the only indecomposable finitely generated Abelian groups), and the Krull–Schmidt theorem holds for the class FG-$\mathbf{Z}$. For another non-trivial example, note that if Mod-R denotes the class of all right R-modules, then the Krull–Schmidt theorem holds for Mod-R if and only if R is right pure-semisimple (because if R is right pure-semisimple, then every right R-module is Σ-pure-injective, and Σ-pure-injective modules are direct sums of modules with local endomorphism rings [F2, Theorem 2.29]; conversely, if the Krull–Schmidt theorem holds for Mod-R, then every right R-module is a direct sum of indecomposable modules, hence R must be right pure-semisimple [Z4, Corollary 2]).

The outline of the chapter is the following. In Section 2 we give the basic concepts and definitions: we sketch two proofs of the Krull–Schmidt–Azumaya theorem, mention refinements of decompositions and the exchange property and study endomorphism rings of modules of finite length. Notice that the validity of the Krull–Schmidt theorem, at least for finite direct sum decompositions, is a property of the endomorphism ring. That is, if M_R is a module and $E = \mathrm{End}(M_R)$ is its endomorphism ring, then any two direct sum decompositions $M_R = M_1 \oplus \cdots \oplus M_t = M'_1 \oplus \cdots \oplus M'_s$ of M_R into finitely many indecomposable direct summands are isomorphic if and only if any two direct sum decompositions $E_E = e_1 E \oplus \cdots \oplus e_t E = e'_1 E \oplus \cdots \oplus e'_s E$ of the right E-module E_E into finitely many (projective cyclic) indecomposable direct summands are isomorphic. More generally, the category $\mathrm{add}(M_R)$ of all modules isomorphic to direct summands of direct sums M_R^n of finitely many copies of M_R and the category proj-E of all finitely generated projective right E-modules are equivalent. We recall the basic properties of semiperfect rings and semilocal rings, because endomorphism rings of modules of finite length are semiperfect rings, and semiperfect rings are semilocal.

In Section 3 we survey the validity of the Krull–Schmidt theorem for modules over commutative rings. Historically, in the setting of commutative rings, the first results obtained were the theorem about elementary divisors and Steinitz' theorem about the structure of finitely generated modules over a Dedekind domain. We consider finite rank torsion-free Abelian groups, introducing the results obtained by Jónsson [J1,J2,J3] and Walker [W1]. Finite rank torsion-free Abelian groups are *almost Krull–Schmidt* $\mathbf{Z}$-modules, in the sense that they have only finitely many direct sum decompositions up to isomorphism. Then we consider finitely generated modules over commutative rings. Here the notion of Henselian ring appears naturally because of the link between the Krull–Schmidt theorem and Henselian rings that was first pointed out by Swan and Evans [E]. In this area, the main results were obtained by Vámos [V3,V4], Siddoway [S3] and Levy [L3]. The case of commutative rings whose finitely generated modules are direct sums of cyclic submodules or of submodules generated by $\leqslant n$ elements is particularly interesting. Then we turn our attention to the case of torsion-free modules over valuation domains, another area in which Henselian rings appear naturally. In Section 4 we study a second class of almost Krull–Schmidt modules, the modules whose endomorphism ring is semilocal. In this setting the main results were obtained by Camps and Dicks [CD], Herbera and Shamsuddin [HS], Facchini, Herbera, Levy and Vámos [FHLV,FH3].

In Section 5 we consider biuniform modules in general and uniserial modules in particular. Warfield [W5] proved in 1975 that every finitely presented module over a serial ring is a finite direct sum of uniserial summands and asked whether such a finite direct sum decomposition is unique up to isomorphism, i.e. whether the Krull–Schmidt theorem holds for finitely presented modules over serial rings. The negative answer to Warfield's question was given by the author [F1] in 1996. Subsequently, part of the contents of [F1] was extended to direct sums of arbitrary (possibly infinite) families of uniserial modules by Nguyen Viet Dung and the author in [DF1], and results about direct summands of serial modules were obtained in [DF2]. The three articles [F1,DF1,DF2] deal with direct sum decompositions of serial modules, but most of the results they contain hold not only for serial modules, but also for modules that are direct sums of biuniform modules, as was observed independently by Nguyen Viet Dung, Dolors Herbera and Ladislav Bican [B3]. The most recent results about direct sum decompositions into uniserial modules were obtained by Puninski [P3,P4], who constructed wonderful examples.

In Section 6 we study projective modules over semilocal rings, because this is equivalent to studying direct sum decompositions of modules whose endomorphism ring is semilocal and these modules are almost Krull–Schmidt, as we have already said above. The isomorphism classes of finitely generated projective modules over a semilocal ring form a commutative monoid that is isomorphic to a full submonoid of the additive monoid $\mathbf{N}^n$. In this setting, the main results were obtained by the author and Herbera [FH1], R. Wiegand [W10] and Yakovlev [Y2]. In Section 7 we present some results about direct sums of modules whose endomorphism rings are homogeneous semilocal. Homogeneous semilocal rings are a class of rings that properly contains the class of local rings, and therefore it is natural to ask whether the Krull–Schmidt theorem holds for the modules that are direct sums of modules whose endomorphism rings are homogeneous semilocal; see [CF,BFRR, FH3]. In Section 8 we present some nice theorems due to Levy and Odenthal [LO1] about the Krull–Schmidt theorem for modules over a semiprime module-finite k-algebra, where k is a commutative, Noetherian ring of Krull dimension 1. Section 9 contains the theorem on torsion-free Krull–Schmidt for integral group rings by Hindman, Klingler and Odenthal [HKO].

This report is not exhaustive. We have not considered the validity of the Krull–Schmidt theorem for other algebraic structures (for the classical case with regard to groups see [R4, 3.3.8] or [R6], and for modular lattices of finite length and universal algebras see [C1]). We have only mentioned the case of additive categories in Appendix A. In this framework, we cite the nice introduction of the paper [WW1] for a detailed report on the history of the Krull–Schmidt theorem for additive categories.

Finally, Appendix B is devoted to the Goldie dimension of modular lattices, because the endomorphism ring of a finite direct sum of modules with local endomorphism rings is a semilocal ring and semilocal rings are characterized as the rings for which the dual lattice of the modular lattice of their right ideals has finite Goldie dimension. The *Goldie dimension* (or, *uniform dimension*, or *uniform rank*, or simply the *rank*) of a module was introduced by Alfred Goldie in 1960. Goldie, while studying the question of when it was possible to construct a classical right quotient ring that is semisimple Artinian, realized that a necessary condition was that the module R_R had finite Goldie dimension. Dually, Varadarajan [V5] realized in 1979 the importance of the Goldie dimension of the dual lattice $\mathcal{L}(M_R)^{\mathrm{op}}$ of the lattice $\mathcal{L}(M_R)$ of all submodules of a module M_R, the so-called *dual Goldie dimension* of M_R. Thus semilocal rings are exactly the rings with finite dual Goldie dimension. In our presentation we follow Grzeszczuk and Puczyłowski [GP], who introduced in 1984 the Goldie dimension for arbitrary modular lattices with 0 and 1.

The Krull–Schmidt theorem holds for a number of classes of modules, but apart from those classes in which the indecomposable modules have local endomorphism rings and therefore the Krull–Schmidt–Azumaya Theorem 2.2 applies directly, for the other classes of modules the proof of the validity of the Krull–Schmidt theorem depends strongly on properties of the class itself. In other words, a general theory is missing. The reason of this is, in our opinion, the following. The Krull–Schmidt theorem does not hold for arbitrary modules, that is, an arbitrary module usually has non-isomorphic direct sum decompositions. The validity of the Krull–Schmidt theorem for a class of modules is an exception, and when it holds, it must be examined separately, case by case, using the particular properties of the class in all the key points of the proof. And this must be the reason why a general theory cannot exist: there cannot be a general theory that includes all these exceptions.

I wish to thank Professor Lawrence Levy, to whom Sections 8 and 9 are almost entirely due. I am also indebted to Professors David Arnold, Lee Klingler, Peter Vámos and Paolo Zanardo for reading parts of the manuscript.

2. Basic concepts and definitions

2.1. *The Krull–Schmidt–Azumaya theorem*

The well-known classical Krull–Schmidt theorem asserts that

THEOREM 2.1. *If M is a module of finite length, then any two direct sum decompositions $M = M_1 \oplus M_2 \oplus \cdots \oplus M_t = N_1 \oplus N_2 \oplus \cdots \oplus N_s$ of M into indecomposable summands M_i, N_j are isomorphic.*

Subsequently this theorem was generalized, and its generalization is usually known as the Krull–Schmidt–Azumaya theorem (Theorem 2.2). For any ring R we shall denote the Jacobson radical of R by $J(R)$. A ring R is *local* if $R/J(R)$ is a division ring. Equivalently, a ring R is local if and only if the sum of two non-invertible elements of R is non-invertible. Therefore for us a local ring will not necessarily be either commutative or Noetherian.

THEOREM 2.2 (Krull–Schmidt–Azumaya theorem [A3]). *If a module $M = \bigoplus_{i\in I} M_i$ is a direct sum of modules M_i $(i \in I)$ and the endomorphism ring of each M_i is local, then any two direct sum decompositions of M into indecomposable direct summands are isomorphic.*

A concept strictly related to the Krull–Schmidt theorem and the notion of isomorphic decompositions is the concept of isomorphic refinements. Let M be a module over a ring R. Suppose that $\{M_i \mid i \in I\}$ and $\{N_j \mid j \in J\}$ are two families of submodules of M such that $M = \bigoplus_{i\in I} M_i = \bigoplus_{j\in J} N_j$. The second decomposition is a *refinement* of the first if there is a surjective mapping $\varphi : J \to I$ such that $N_j \subseteq M_{\varphi(j)}$ for every $j \in J$. Equivalently, the decomposition $M = \bigoplus_{j\in J} N_j$ is a *refinement* of the decomposition $M = \bigoplus_{i\in I} M_i$ if there is a surjective mapping $\varphi : J \to I$ such that $\bigoplus_{j\in\varphi^{-1}(i)} N_j = M_i$ for every $i \in I$. Here we shall not consider the property of having isomorphic refinements in detail. We just recall the following result due to Warfield:

THEOREM 2.3 [W3, Theorem 1]. *If a module $M = \bigoplus_{i\in I} M_i$ is a direct sum of modules M_i $(i \in I)$ and each M_i is countably generated and has a local endomorphism ring, then* *every direct sum decomposition of M has a refinement isomorphic to the decomposition $M = \bigoplus_{i \in I} M_i$.*

We recall two different proofs of Theorem 2.2. The first, that can be found in [AF2, Theorem 12.6], makes use of the notion of decomposition that complements (maximal) direct summands. If M is a module, all complements of a direct summand M' of M are isomorphic, because $M = M' \oplus M_1 = M' \oplus M_2$ implies $M_1 \cong M/M' \cong M_2$. A direct summand M' of M is a *maximal direct summand* of M if M' has an indecomposable complement in M. A direct sum decomposition $M = \bigoplus_{i \in I} M_i$ of a module M is said to *complement direct summands* if for every direct summand M' of M there is a subset $J \subseteq I$ such that $M = M' \oplus (\bigoplus_{i \in J} M_i)$, and is said to *complement maximal direct summands* if for every maximal direct summand M' of M there exists $J \subseteq I$ with $M = M' \oplus (\bigoplus_{i \in J} M_i)$. Right perfect rings can be characterized by the property that all projective right modules have a decomposition that complements direct summands [AF1, Theorem 6]. If $M = \bigoplus_{i \in I} M_i$ is a direct sum decomposition that complements direct summands, then all the modules M_i are indecomposable, and if a module M has a direct sum decomposition $M = \bigoplus_{i \in I} M_i$ that complements maximal direct summands and all the modules M_i are indecomposable, then any two direct sum decompositions of M into indecomposable direct summands are isomorphic [AF2, Theorem 12.4]. It is possible to prove that a decomposition $M = \bigoplus_{i \in I} M_i$ in which the modules M_i have local endomorphism rings complements maximal direct summands. Theorem 2.2 follows from this. For the details of this proof, see [AF2, §12].

A second proof of the Krull–Schmidt–Azumaya theorem, which can be found in [F2, Theorem 2.12], uses the so-called *exchange property*. Let $\aleph$ be a cardinal number. An R-module M is said to have the *$\aleph$-exchange property* if for any R-module G and any two direct sum decompositions

$$G = M' \oplus N = \bigoplus_{i \in I} M_i,$$

where $M' \cong M$ and the cardinality $|I|$ of I is $\leqslant \aleph$, then there are R-submodules N_i of M_i, $i \in I$, such that $G = M' \oplus (\bigoplus_{i \in I} N_i)$. A module has the *exchange property* if it has the $\aleph$-exchange property for every cardinal number $\aleph$, and has the *finite exchange property* if it has the n-exchange property for every finite cardinal n. A module with the 2-exchange property has the finite exchange property. If $\aleph$ is a cardinal number, a module $M = M_1 \oplus M_2$ has the $\aleph$-exchange property if and only if both M_1 and M_2 have the $\aleph$-exchange property. An indecomposable module M_R has the exchange property if and only if it has the finite exchange property, if and only if its endomorphism ring $\mathrm{End}(M_R)$ is local. It is possible to prove [CJ, 4.1] that if $\aleph$ is a cardinal, M is a module with the $\aleph$-exchange property, $M = \bigoplus_{i \in I} M_i = \bigoplus_{j \in J} N_j$ are two direct sum decompositions of M with I finite and $|J| \leqslant \aleph$, then the two direct sum decompositions of M have isomorphic refinements. Making use of these properties of modules with the exchange property, it is possible to show that for a module M that is a direct sum of modules with local endomorphism rings, the endomorphism ring of *any* indecomposable direct summand of M is local. From this the Krull–Schmidt–Azumaya Theorem 2.2 follows. For the details, see [F2, §§2.1–2.4].

2.2. *Endomorphism rings of modules of finite length*

Theorems about finite direct sum decompositions of a module M_R, like Theorem 2.1, are essentially theorems about the endomorphism ring of M_R because of the one-to-one correspondence between finite direct sum decompositions of M_R and complete sets of orthogonal idempotents in its endomorphism ring $\mathrm{End}(M_R)$. More generally, let M_R be a module over an arbitrary ring R and let $E = \mathrm{End}(M_R)$. All direct sum decompositions of M_R in the category $\mathrm{Mod}\, R$ are direct sum decompositions of M_R in the full subcategory $\mathrm{add}(M_R)$ of $\mathrm{Mod}\, R$ whose objects are the modules isomorphic to direct summands of direct sums M_R^n of finitely many copies of M_R. For any ring S, we shall denote by proj-S the full subcategory of $\mathrm{Mod}\, S$ whose objects are all finitely generated projective right S-modules. Thus proj-$S = \mathrm{add}(S_S)$.

THEOREM 2.4. *Let M_R be an R-module over a ring R and let $E = \mathrm{End}(M_R)$, so that ${}_EM_R$ is a bimodule. The categories* $\mathrm{add}(M_R)$ *and* proj-E *are equivalent via the functors* $\mathrm{Hom}_R(M_R, -) : \mathrm{add}(M_R) \to$ proj-E *and* $- \otimes_E M :$ proj-$E \to \mathrm{add}(M_R)$.

As category equivalences preserve direct sum decompositions, the study of the direct sum decompositions of a module N_R belonging to $\mathrm{add}(M_R)$ can be reduced to that of the corresponding projective E-module $\mathrm{Hom}_R({}_EM_R, N_R)$. Note that in this equivalence between $\mathrm{add}(M_R)$ and proj-E the module M_R correspond to the module E_E. Hence, studying the direct sum decompositions in the category $\mathrm{add}(M_R)$ is completely equivalent to studying the direct sum decompositions of finitely generated projective right modules over the ring $E = \mathrm{End}(M_R)$.

A ring S is *semiperfect* if it has a complete set $\{e_1, e_2, \dots, e_n\}$ of orthogonal idempotents for which every e_iSe_i is a local ring. Equivalently, a ring S is semiperfect if and only if $S/J(S)$ is semisimple Artinian and idempotents lift modulo $J(S)$, if and only if every finitely generated S-module has a projective cover [F2, Theorem 3.6]. The reason why semiperfect rings naturally appear in connection with the Krull–Schmidt theorem is explained by the next proposition.

PROPOSITION 2.5 [F2, Proposition 3.14]. *If M_R is a module and $E = \mathrm{End}(M_R)$ is its endomorphism ring, then M_R is a finite direct sum of R-modules with local endomorphism rings if and only if E_E is a finite direct sum of E-modules with local endomorphism rings, if and only if E is semiperfect.*

An immediate application of Theorem 2.3 to free modules over semiperfect rings yields the following result.

PROPOSITION 2.6. *The Krull–Schmidt theorem holds for the class of all projective right modules over a semiperfect ring.*

Theorem 2.1 considers modules of finite length. It is possible to prove that *the endomorphism ring $E = \mathrm{End}(M_R)$ of a module M_R of finite length is a semiprimary ring* (that is, $E/J(E)$ is semisimple Artinian and $J(E)$ is nilpotent) *with the acc and the dcc on* *right annihilators and on left annihilators*. More generally, *if M_R is an Artinian module, then* $\mathrm{End}(M_R)/J(\mathrm{End}(M_R))$ *is semisimple Artinian* [CD, Corollary 6], $\mathrm{End}(M_R)$ *satisfies the acc on left annihilators and every nil subring of* $\mathrm{End}(M_R)$ *is nilpotent* [F5, Theorem 1.5]. Here is a sketch of the proof of what we have stated for modules of finite length. A module M_R is a *Fitting module* if for every $f \in \mathrm{End}(M_R)$ there is an integer $n \geqslant 1$ such that $M = \ker(f^n) \oplus f^n(M)$. If M_R is a module of finite length, then M_R is a Fitting module (this is Fitting's lemma [F2, Lemma 2.20]). Hence, for every $f \in J(\mathrm{End}(M_R))$, there exists an n such that $M = \ker(f^n) \oplus f^n(M)$. In particular, f^n induces an automorphism on $f^n(M)$. Let α be the endomorphism of M that is equal to the inverse of this automorphism on $f^n(M)$ and is equal to zero on $\ker(f^n)$. Then $f^n\alpha$ is the idempotent endomorphism of M that is the identity on $f^n(M)$ and is equal to zero on $\ker(f^n)$. Since it belongs to $J(\mathrm{End}(M_R))$, it must be zero, i.e. $f^n(M) = 0$. Thus $J(\mathrm{End}(M_R))$ is nil, hence nilpotent [F5, Theorem 1.5]. In order to show that $\mathrm{End}(M_R)$ satisfies the acc and the dcc on left annihilators, let $\mathcal{LA}(\mathrm{End}(M_R))$ denote the set of all left annihilators of $\mathrm{End}(M_R)$ and $\mathcal{L}(M_R)$ the set of all submodules of M_R. The mapping $\varphi : \mathcal{LA}(\mathrm{End}(M_R)) \to \mathcal{L}(M_R)$ defined by $\varphi(I) = \bigcap_{f \in I} \ker(f)$ for every $I \in \mathcal{LA}(\mathrm{End}(M_R))$ is inclusion reversing. Similarly, if $\mathcal{L}(\mathrm{End}(M_R))$ denotes the set of all left ideals of $\mathrm{End}(M_R)$, the mapping $\psi : \mathcal{L}(M_R) \to \mathcal{L}(\mathrm{End}(M_R))$ defined by $\psi(N) = \{f \in \mathrm{End}(M_R) \mid f(N) = 0\}$ is inclusion reversing. If A is a subset of $\mathrm{End}(M_R)$, then its left annihilator $I = \{f \in \mathrm{End}(M_R) \mid fg = 0$ for every $g \in A\}$ is equal to $\psi(N)$ where $N = \sum_{g \in A} g(M)$. Trivially $N \subseteq \varphi(\psi(N))$, so that $I = \psi(N) \supseteq \psi(\varphi(\psi(N))) = \psi(\varphi(I))$. Conversely $I \subseteq \psi(\varphi(I))$, so that $I = \psi(\varphi(I))$. In particular, the mapping $\varphi : \mathcal{LA}(\mathrm{End}(M_R)) \to \mathcal{L}(M_R)$ is injective. As $\mathcal{L}(M_R)$ satisfies the acc and the dcc, it follows that $\mathcal{LA}(\mathrm{End}(M_R))$ satisfies the acc and the dcc, i.e. $\mathrm{End}(M_R)$ satisfies both the acc and the dcc on left annihilators. Finally, a ring satisfies the acc and the dcc on left annihilators if and only if it satisfies the acc and the dcc on right annihilators.

There exist examples of modules M_R of finite length whose endomorphism ring is a non-Noetherian commutative ring. For instance, let $K \subseteq F$ be an extension of fields of infinite degree $[F : K]$. Let $F^* = \mathrm{Hom}_K(F, K)$ be the dual of the K-vector space F, so that F^* is a K-vector space of infinite dimension and ${}_F F^*_K$ is an F-K-bimodule. Let R be the ring of all 3×3 matrices $\left(\begin{smallmatrix} a & b & \varphi \\ 0 & a & 0 \\ 0 & 0 & k \end{smallmatrix}\right)$ with $a, b \in F$, $\varphi \in F^*$ and $k \in K$ and let M be the set of all 1×3 row matrices (c, d, ℓ) with $c, d \in F$ and $\ell \in K$. Then M is a right module over R and it is easily verified that M_R is a cyclic module of length 3 whose endomorphism ring $\mathrm{End}(M_R)$ is isomorphic to the ring of all 2×2 matrices $\left(\begin{smallmatrix} k & a \\ 0 & k \end{smallmatrix}\right)$ with $k \in K$ and $a \in F$. This is a local non-Noetherian commutative ring.

2.3. *First applications of the Krull–Schmidt–Azumaya theorem*

The first example of trivial application of the Krull–Schmidt–Azumaya theorem concerns the class of semisimple modules. As the endomorphism ring of any simple module is a division ring by Schur's lemma, the Krull–Schmidt theorem holds for the class of semisimple modules.

A second natural application of the Krull–Schmidt–Azumaya Theorem 2.2 is that of deriving from it the classical Krull–Schmidt Theorem 2.1, which concerns modules of finite length. More generally, this can be done for Fitting modules. The endomorphism ring of an indecomposable Fitting module is local [F2, Lemma 2.21], so that the Krull–Schmidt–Azumaya Theorem 2.2 applies to any module M that is a direct sum of indecomposable Fitting modules. By Fitting's lemma [F2, Lemma 2.20], every module of finite length is a Fitting module, so that Theorem 2.1 follows from Theorem 2.2.

Our third immediate application of the Krull–Schmidt–Azumaya theorem is to the class of injective modules of finite Goldie dimension. For the definition of the Goldie dimension of a module, see Appendix B. Recall that every non-invertible element φ of the endomorphism ring $\mathrm{End}(M_R)$ of an injective indecomposable R-module M_R has non-zero kernel. It follows that [F2, Lemma 2.25]:

PROPOSITION 2.7. *An injective R-module M_R is indecomposable if and only if its endomorphism ring* $\mathrm{End}(M_R)$ *is local. Therefore the Krull–Schmidt theorem holds for the class of injective modules of finite Goldie dimension.*

Any Noetherian module N_R has finite Goldie dimension, so that the injective envelope $E(N_R)$ of any Noetherian module N_R has finite Goldie dimension. Thus the Krull–Schmidt theorem holds for the class of the injective envelopes $E(N_R)$ of Noetherian modules N_R.

THEOREM 2.8 (Matlis [M1], Papp [P1]). *A ring R is right Noetherian if and only if every injective right R-module is a direct sum of indecomposable injective R-modules.*

Therefore every injective right module over a right Noetherian ring R is a direct sum of modules whose endomorphism rings are local. Hence

COROLLARY 2.9. *If the ring R is right Noetherian, the Krull–Schmidt theorem holds for the class of injective right R-modules.*

2.4. *Semilocal rings*

For the details and the proofs of the results presented in this section, see [F2].

PROPOSITION 2.10. *The following conditions are equivalent for a ring R:*
(a) *The ring $R/J(R)$ is semisimple Artinian.*
(b) *The ring $R/J(R)$ is isomorphic to a finite direct product $\prod_{i=1}^{t} M_{n_i}(D_i)$ of rings $M_{n_i}(D_i)$ of $n_i \times n_i$ matrices over division rings D_i, $i = 1, 2, \dots, t$.*
(c) *The ring $R/J(R)$ is right Artinian.*
(d) *There exists a finite set $\{I_1, I_2, \dots, I_n\}$ of maximal right ideals of R such that $J(R) = I_1 \cap I_2 \cap \cdots \cap I_n$.*
(c*), (d*) *The left–right duals of* (c), (d).

A ring R is said to be *semilocal* if it satisfies the equivalent conditions of Proposition 2.10.

EXAMPLES 2.11.

(1) By condition (d), a commutative ring is semilocal if and only if it has only finitely many maximal ideals.
(2) By condition (c), every right (or left) Artinian ring is semilocal.
(3) Every local ring is semilocal.
(4) If R is semilocal, the ring $M_n(R)$ of $n \times n$ matrices over R is semilocal for every positive integer n. If $0 \neq e \in R$ is idempotent and R is semilocal, then eRe is semilocal.
(5) Let k be a commutative ring with identity. A *k-algebra* is a ring R that is a k-module and for which multiplication is k-bilinear. Equivalently, a k-algebra can be defined as a ring R with a ring homomorphism of k into the center of R. A k-algebra is said to be *module-finite* if it is finitely generated as a k-module. A module-finite algebra over a semilocal commutative ring is semilocal [F2, p. 7].
(6) Let k be either a semilocal commutative principal ideal domain or a valuation domain (that is, a commutative integral domain whose ideals are linearly ordered under inclusion). If R is a k-algebra which is torsion-free and of finite rank over k, then R is semilocal [W7, Theorems 5.2 and 5.4].
(7) Semiperfect rings, i.e. endomorphism rings of finite direct sums of modules with local endomorphism rings (Proposition 2.5), are semilocal rings.
(8) For examples of modules with semilocal endomorphism rings, see Examples 4.2.

A semilocal ring has only finitely many simple right modules up to isomorphism. We already know that in a semilocal ring the Jacobson radical is the intersection of finitely many maximal right ideals. We shall see in the next theorem that semilocal rings are exactly those of finite dual Goldie dimension (for the definition of dual Goldie dimension of a module, see Appendix B). In [FS2] it is proved that a semilocal ring has only finitely many indecomposable projective right modules up to isomorphism. Thus being semilocal is a finiteness condition on the ring.

THEOREM 2.12 [SV1, Corollary 1.14]. *A ring R is semilocal if and only if the right R-module R_R has finite dual Goldie dimension, if and only if the left R-module ${}_RR$ has finite dual Goldie dimension. If these equivalent conditions hold, then* $\operatorname{codim}(R_R) = \operatorname{codim}({}_RR) =$ "*Goldie dimension of the semisimple R-module $R/J(R)$*".

If $e \in R$ is idempotent, the dual Goldie dimension $\operatorname{codim}(eR)$ of the right R-module eR is equal to the dual Goldie dimension $\operatorname{codim}(eRe)$ of the (right or left) eRe-module eRe [SV1, Theorems 2.3 and 2.5]. As $\operatorname{codim}(R_R) = \operatorname{codim}(eR) + \operatorname{codim}((1-e)R)$, we get that

PROPOSITION 2.13. *For any idempotent e of a ring R*

$$\operatorname{codim}(R) = \operatorname{codim}(eRe) + \operatorname{codim}\big((1-e)R(1-e)\big).$$

In particular, R is semilocal if and only if eRe and $(1-e)R(1-e)$ are both semilocal.

3. Krull–Schmidt and modules over commutative rings

3.1. *Elementary divisors and Steinitz' theorem*

The origins of the Krull–Schmidt theorem with regard to modules over commutative rings go back to the well-known theorem about elementary divisors (modules over principal ideal domains) and Steinitz' theorem (modules over Dedekind domains). By a *principal ideal domain* we mean a commutative integral domain in which every ideal is principal.

THEOREM 3.1 (Elementary divisors). *Let M_R be a finitely generated module over a principal ideal domain R. Then there exist a natural number n and a chain $I_1 \supseteq I_2 \supseteq \cdots \supseteq I_n$ of proper ideals of R such that $M \cong \bigoplus_{k=1}^n R/I_k$. The number n and the ideals I_k, $k = 1, 2, \dots, n$, are uniquely determined by M_R.*

Note that if R is a principal ideal domain and I is a non-zero proper ideal of R, then I is the product of a non-empty finite family of prime ideals of R, that is, $I = P_1^{m_1} P_2^{m_2} \dots P_t^{m_t}$ with $t \geqslant 1$, $P_1, P_2, \dots, P_t$ distinct prime ideals of R and $m_1, m_2, \dots, m_t$ positive integers. Then $I = P_1^{m_1} \cap P_2^{m_2} \cap \cdots \cap P_t^{m_t}$, so that $R/I \cong R/P_1^{m_1} \oplus R/P_2^{m_2} \oplus \cdots \oplus R/P_t^{m_t}$ by the Chinese remainder theorem and $\mathrm{End}(R/P_j^{m_j}) \cong R/P_j^{m_j}$ is a local ring. Thus Theorem 3.1 – when the terms R/I are decomposed in this way – is a prototype theorem of Krull–Schmidt type.

The extension of Theorem 3.1 to Dedekind domains is the following:

THEOREM 3.2 (Steinitz, 1912). *Let M_R be a finitely generated module over a Dedekind domain R. Then*:

(a) *$M = T \oplus F$, where T is torsion and F is projective*;

(b) *$T \cong R/Q_1^{n_1} \oplus R/Q_2^{n_2} \oplus \cdots \oplus R/Q_t^{n_t}$ for a suitable integer $t \geqslant 0$, non-zero prime ideals $Q_1, Q_2, \dots, Q_t$ and positive integers $n_1, n_2, \dots, n_t$*;

(c) *$F \cong I_1 \oplus I_2 \oplus \cdots \oplus I_m$ for a suitable integer $m \geqslant 0$ and non-zero ideals $I_1, I_2, \dots, I_m$ of R.*

Moreover, if $M = T' \oplus F'$, $T' \cong R/(Q_1')^{n_1'} \oplus R/(Q_2')^{n_2'} \oplus \cdots \oplus R/(Q_{t'}')^{n_{t'}'}$, $F' \cong I_1' \oplus I_2' \oplus \cdots \oplus I_{m'}'$ are direct sum decompositions with properties (a), (b) *and* (c), *then $T = T'$, $F \cong F'$, $t = t'$, $m = m'$, $I_1 I_2 \dots I_m \cong I_1' I_2' \dots I_m'$, and there is a permutation σ of $\{1, 2, \dots, t\}$ such that $Q_{\sigma(i)} = Q_i'$ and $n_{\sigma(i)} = n_i'$ for every $i = 1, 2, \dots, t$.*

For a proof of the results of Section 3.1 see, for instance, [SV2].

Notice that every finitely generated module over a principal ideal domain is a direct sum of cyclic modules, and every finitely generated module over a Dedekind domain is a direct sum of modules generated by $\leqslant 2$ elements. As far as part (c) of Theorem 3.2 is concerned, note that two ideals I and J of an integral domain R with field of fractions K are isomorphic if and only if there exists a non-zero element α of K such that $I = \alpha J$. Here is a description of the torsion-free (= projective) part of the modules described in Theorem 3.2.

COROLLARY 3.3. *Let R be a Dedekind domain and let $I_1, I_2, \ldots, I_t, I'_1, I'_2, \ldots, I'_s$ be non-zero ideals of R. Then $I_1 \oplus I_2 \oplus \cdots \oplus I_t \cong I'_1 \oplus I'_2 \oplus \cdots \oplus I'_s$ if and only if $t = s$ and $I_1 I_2 \ldots I_t \cong I'_1 I'_2 \ldots I'_s$. In particular, $I_1 \oplus I_2 \oplus \cdots \oplus I_t \cong R^{t-1} \oplus I_1 I_2 \ldots I_t$.*

Every ideal of a Dedekind domain R is isomorphic to a direct summand of $R_R \oplus R_R$. It follows that the Krull–Schmidt theorem holds for the class of finitely generated projective modules over a Dedekind domain R if and only if R is a principal ideal domain.

For commutative rings R such that for any ideals $I_1, I_2, \ldots, I_t, I'_1, I'_2, \ldots, I'_s$ of R, if $I_1 \oplus I_2 \oplus \cdots \oplus I_t \cong I'_1 \oplus I'_2 \oplus \cdots \oplus I'_s$, then $t = s$ and after reindexing, $I_j \cong J_j$ for all j, see [GO2,GO3].

3.2. *Torsion-free Abelian groups*

In this section we consider torsion-free Abelian groups. We begin with the general case of finite rank torsion-free modules and finitely generated torsion-free modules over arbitrary commutative integral domains. For these classes the Krull–Schmidt theorem does not usually hold. Examples are easy to find. For instance, let I and J be two comaximal ideals of a commutative integral domain R, that is, two ideals of R with $I + J = R$. For instance, I and J could be two distinct maximal ideals of R. If I and J are not principal ideals, then the Krull–Schmidt theorem does not hold for the class of finite rank torsion-free R-modules (because the kernel of the canonical mapping $I \oplus J \to R$, $(i, j) \mapsto i + j$, is a direct summand). Similarly, if I and J are finitely generated ideals that are not principal, then the Krull–Schmidt theorem does not hold for the class of finitely generated torsion-free R-modules. As far as the case $R = \mathbf{Z}$ is concerned, the Krull–Schmidt theorem does not hold for the class of finite rank torsion-free Abelian groups [J1,J2], i.e. the direct sum decomposition is not unique up to isomorphism. Nevertheless, there are only finitely many direct sum decompositions up to isomorphism, as we shall see now.

A module M_R over an arbitrary ring R is said to be an *almost Krull–Schmidt module* [V4, p. 480] if it has only a finite number of direct sum decompositions up to isomorphism. Lady [L1] proved that:

LEMMA 3.4. *If R is a ring whose additive group is torsion-free of finite rank, then R_R is an almost Krull–Schmidt module.*

If E is the endomorphism ring of a finite rank torsion-free Abelian group, then the additive group of E is likewise a finite rank torsion-free Abelian group. Making use of Theorem 2.4 one obtains:

THEOREM 3.5 [L1]. *Every finite rank torsion-free Abelian group is an almost Krull–Schmidt $\mathbf{Z}$-module.*

For a characterization of the Dedekind domains of characteristic 0 over which every finite rank torsion-free module is almost Krull–Schmidt, see [FV1, Theorem 5.1]. We shall see in Lemma 4.3(c) that all modules with semilocal endomorphism rings are almost Krull–Schmidt modules.

Although the Krull–Schmidt theorem does not hold for finite rank torsion-free Abelian groups, it does hold in the category we shall describe now. We follow Walker [W1]. An Abelian group A is *bounded* if $nA = 0$ for some positive integer n. Let $\mathcal{B}$ be the class of all bounded Abelian groups. If A and B are two arbitrary Abelian groups, we say that A and B *belong to the same monogeny class* if there exist a monomorphism $A \to B$ and a monomorphism $B \to A$ (in this case we shall write $[A]_m = [B]_m$), and that they are *quasi-isomorphic* [J3] if there exist a subgroup S of A and a subgroup T of B with $A/S, B/T \in \mathcal{B}$ and $S \cong T$.

LEMMA 3.6 (Jónsson [J3], Warfield [W7, Lemma 5.1]). *If M is a finite rank torsion-free module over an integral domain R and $f : M \to M$ is a monomorphism, then there is an $r \in R$, $r \neq 0$, such that $rM \subseteq f(M)$.*

For $R = \mathbf{Z}$, that is, when M is an Abelian group, Lemma 3.6 states that every subgroup of M isomorphic to M has finite index. Two torsion-free Abelian groups A and B are quasi-isomorphic if and only if there exist a monomorphism $f : A \to B$ and a monomorphism $g : B \to A$ with $A/g(B), B/f(A) \in \mathcal{B}$; and two torsion-free Abelian groups of finite rank are quasi-isomorphic if and only if they belong to the same monogeny class.

Let $\mathcal{A}$ be the category of all Abelian groups. The class $\mathcal{B}$ is a *Serre class* of $\mathcal{A}$, that is, for every exact sequence $0 \to G' \to G \to G'' \to 0$ of Abelian groups, G belongs to $\mathcal{B}$ if and only if G' and G'' belong to $\mathcal{B}$. Therefore it is possible to construct the *quotient category* $\mathcal{A}/\mathcal{B}$ in the sense of Grothendieck [G2]. The objects of $\mathcal{A}/\mathcal{B}$ are all the objects of $\mathcal{A}$, that is, all Abelian groups, and $\mathrm{Hom}_{\mathcal{A}/\mathcal{B}}(G, H) = \varinjlim \mathrm{Hom}_{\mathbf{Z}}(G', H/H')$, where the direct system consists of all subgroups G' of G and H' of H with $G/G' \in \mathcal{B}$ and $H' \in \mathcal{B}$. The category $\mathcal{A}/\mathcal{B}$ is Abelian, and we may look at the usual direct sum decompositions of the Abelian group G in the category $\mathcal{A}$ or the coproduct decompositions of G in the category $\mathcal{A}/\mathcal{B}$ (called *quasi-decompositions* of G).

There is an equivalent way to look at the category $\mathcal{A}/\mathcal{B}$. Let $\mathcal{A}_{\mathbf{Q}}$ be the category whose objects are all Abelian groups and whose morphism groups are defined by $\mathrm{Hom}_{\mathcal{A}_{\mathbf{Q}}}(G, H) = \mathbf{Q} \otimes \mathrm{Hom}_{\mathbf{Z}}(G, H)$ (Jónsson [J3]). Every element of $\mathbf{Q} \otimes \mathrm{Hom}_{\mathbf{Z}}(G, H)$ can be written in the form $1/n \otimes f$ for some positive integer n and some $f \in \mathrm{Hom}_{\mathbf{Z}}(G, H)$. The composition of morphisms in $\mathcal{A}_{\mathbf{Q}}$ is defined by $(1/n \otimes f) \circ (1/m \otimes g) = (1/nm \otimes fg)$. Then the categories $\mathcal{A}_{\mathbf{Q}}$ and $\mathcal{A}/\mathcal{B}$ are equivalent [W1].

For every $f \in \mathrm{Hom}_{\mathbf{Z}}(G, H)$, let $\overline{f}$ denote the element of $\mathrm{Hom}_{\mathcal{A}/\mathcal{B}}(G, H)$ induced by f. There is a functor $J : \mathcal{A} \to \mathcal{A}/\mathcal{B}$ defined by $J(G) = G$ and $J(f) = \overline{f}$. The functor J has kernel $\mathcal{B}$ and is exact. If $f' \in \mathrm{Hom}_{\mathbf{Z}}(G', H/H')$ and $G/G', H' \in \mathcal{B}$, then f' induces a morphism $\overline{f'} \in \mathrm{Hom}_{\mathcal{A}/\mathcal{B}}(G, H)$. One has that $\overline{f'}$ is an epimorphism if and only if $\operatorname{coker} f' \in \mathcal{B}$, and $\overline{f'}$ is a monomorphism if and only if $\ker f' \in \mathcal{B}$. As remarked by Walker [W1, pp. 147 and 154], two torsion-free Abelian groups are quasi-isomorphic if and only if they are isomorphic in the quotient category $\mathcal{A}/\mathcal{B}$. Hence the decomposition theory of torsion-free Abelian groups in $\mathcal{A}/\mathcal{B}$ is equivalent to the quasi-decomposition theory of torsion-free Abelian groups in $\mathcal{A}$. Notice that our definition of quasi-isomorphism applies to all Abelian groups, not just torsion-free ones, but quasi-isomorphism does not correspond to isomorphism in $\mathcal{A}/\mathcal{B}$ for arbitrary Abelian groups. The equivalent categories $\mathcal{A}/\mathcal{B}$ and $\mathcal{A}_{\mathbf{Q}}$ have enough projectives and injectives, but they do not have injective envelopes and infinite coproducts.

Let us restrict our attention to the full subcategory FRTF of $\mathcal{A}_{\mathbf{Q}}$ whose objects are all finite rank torsion-free Abelian groups. The category FRTF is an additive category and *idempotents split* in FRTF, that is, if G is an object of FRTF and $e \in \mathrm{End}_{\mathcal{A}_{\mathbf{Q}}}(G)$ is an idempotent, then there exist an object H of FRTF, $p \in \mathrm{Hom}_{\mathcal{A}_{\mathbf{Q}}}(G, H)$ and $q \in \mathrm{Hom}_{\mathcal{A}_{\mathbf{Q}}}(H, G)$ such that $qp = e$ and $pq = 1_H$ (see, for instance, [A1, p. 73]). As we have already said, two torsion-free Abelian groups of finite rank are isomorphic in the category FRTF if and only if the are quasi-isomorphic (in the category $\mathcal{A}$), if and only if they are in the same monogeny class. An object G of FRTF is a coproduct $G = G_1 \amalg \cdots \amalg G_t$ in FRTF if and only if there is a monomorphism $f : G_1 \oplus \cdots \oplus G_t \to G$ in $\mathcal{A}$ with $G/\mathrm{im}(f)$ finite, or, equivalently, if and only if $[G]_m = [G_1 \oplus \cdots \oplus G_t]_m$. An object G of FRTF is indecomposable in FRTF if and only if whenever $nG \subseteq H \oplus H' \subseteq G$ for some integer $n > 0$, then $H = 0$ or $H' = 0$. In this case one says that G is *strongly indecomposable*. A finite rank torsion-free Abelian group is strongly indecomposable if and only if its endomorphism ring $\mathbf{Q} \otimes \mathrm{End}_{\mathbf{Z}}(G)$ in $\mathcal{A}_{\mathbf{Q}}$ is local [R1].

The main result of B. Jónsson's paper [J3] is the Krull–Schmidt theorem for the category FRTF (it holds by Theorem A.1 in Appendix A) and its interpretation in the language of Abelian groups is the following:

THEOREM 3.7. *Let G be a finite rank torsion-free Abelian group. Then there exist a positive integer m and strongly indecomposable subgroups $G_1, \ldots, G_t$ of G such that $mG \subseteq G_1 \oplus \cdots \oplus G_t \subseteq G$. Moreover, if k is a positive integer and $H_1, \ldots, H_s$ are strongly indecomposable subgroups of G with $kG \subseteq H_1 \oplus \cdots \oplus H_s \subseteq G$, then $t = s$ and there is a permutations σ of $\{1, 2, \ldots, t\}$ such that $[G_i]_m = [H_{\sigma(i)}]_m$ for every $i = 1, 2, \ldots, t$.*

An extension of this theorem to Abelian groups of infinite rank is given in Fuchs and Viljoen [FV2]. For the proofs of the statements presented in this Section 3.2 and further results about torsion-free Abelian groups, see [A1].

Similar constructions and results are also possible for other classes of modules. For instance, it is possible to construct a category whose objects are Noetherian right modules over a ring but whose morphisms have been modified in such a way that indecomposable objects have local endomorphism ring. Hence the Krull–Schmidt theorem holds in this new category. See [B5].

Some most remarkable results obtained in the setting of torsion-free Abelian groups appear in the papers [C2,C3,C4]. Corner proved in [C2] that every reduced, torsion-free ring of rank $n \leqslant \aleph_0$ is isomorphic to the endomorphism ring of a reduced, torsion-free group of rank $2n$. This theorem exhibits the enormous amount of pathology which can be encountered in the structure of countable, torsion-free Abelian groups. For instance, applying this result, Corner [C3] showed that for any positive integer r there is a countable torsion-free Abelian group G such that G^m is isomorphic to G^n if and only if $m \equiv n \pmod r$. Also, he proved that there exists a countable, torsion-free group which has no non-zero indecomposable direct summand [C2]. Corner's results have been greatly generalized and have found several applications.

3.3. *Henselian rings*

A local commutative ring R with maximal ideal M is said to be *Henselian* if the following condition holds:

if $f, g_0, h_0 \in R[x]$ are monic polynomials such that
(1) $f \equiv g_0 h_0 \pmod{MR[x]}$ and
(2) $g_0 R[x] + h_0 R[x] + MR[x] = R[x]$,
then there exist monic polynomials $g, h \in R[x]$ such that $f = gh$, $g \equiv g_0 \pmod{MR[x]}$ and $h \equiv h_0 \pmod{MR[x]}$.

For example, complete local commutative rings are Henselian.

PROPOSITION 3.8 (Azumaya [A4]). *A local commutative ring R is Henselian if and only if every module-finite R-algebra* (*not necessarily commutative*) *is semiperfect.*

The next result was proved by Vámos [V3, Lemma 13] and Siddoway [S3] independently.

THEOREM 3.9. *Every finitely generated indecomposable module M_R over a Henselian local commutative ring R has a local endomorphism ring* $\mathrm{End}(M_R)$.

The proof is essentially the following. If M_R is a finitely generated indecomposable module over a Henselian ring R, then $\mathrm{End}(M_R)$ is a module-finite R-algebra, hence a semiperfect ring by Proposition 3.8. Thus M_R is a direct sum of modules with local endomorphism rings (Proposition 2.5). As M_R is indecomposable, its endomorphism ring $\mathrm{End}(M_R)$ must be local.

For a ring R, we denote by FG-R the class of all finitely generated right R-modules. From the Krull–Schmidt–Azumaya Theorem 2.2 and Theorem 3.9 we immediately obtain:

COROLLARY 3.10. *If R is a Henselian local commutative ring, the Krull–Schmidt theorem holds for the class* FG-R.

The sketch of proof we have given for Theorem 3.9 also shows that if R is any module-finite algebra over a Henselian local commutative ring, then every finitely generated indecomposable R-module has a local endomorphism ring, so that the Krull–Schmidt theorem holds for the class FG-R.

From Example 4.2(5) and Lemma 4.3(c) it will follow that finitely generated modules over semilocal commutative rings (more generally, finitely generated modules over module-finite k-algebras, where k is a semilocal commutative ring [W7, Lemma 2.3]) are almost Krull–Schmidt modules.

3.4. *Finitely generated modules over valuation rings*

A *valuation ring* is a commutative ring whose ideals are linearly ordered under inclusion. A *valuation domain* is a valuation ring that is an integral domain. Vámos proved that the Krull–Schmidt theorem does not necessarily hold for the class FG-R of finitely generated modules over a valuation ring R making use of the next theorem. In order to prove it, Vámos used the fact that the Krull–Schmidt theorem does not hold for the class of torsion-free modules of finite rank over a discrete valuation domain of rank one, a class of modules that will be considered in Theorem 3.21. Recall that a subgroup H of a totally ordered Abelian group G is said to be *convex* if $0 \leqslant g \leqslant h$, $g \in G$ and $h \in H$ imply $g \in H$. If R is a valuation domain and G is its value group, then R is said to be a *discrete* valuation domain if whenever $H \subset H'$ are convex subgroups of G and there is no convex subgroup of G properly between H and H', then H'/H is order isomorphic to $\mathbf{Z}$. A discrete valuation domain R is *of rank n* if its value group G has a chain $0 = H_0 \subseteq H_1 \subset \cdots \subset H_n = G$ of convex subgroups with H_i/H_{i-1} order isomorphic to $\mathbf{Z}$ for all $i = 1, 2, \dots, n$. Thus a valuation domain R is a discrete valuation domain of rank one if and only if its value group G is order isomorphic to $\mathbf{Z}$.

THEOREM 3.11 [V3, Theorem 20]. *Let R be a valuation ring (not necessarily a domain) with a prime ideal P and an element $a \in P$ such that $V = R/P$ is a non-Henselian discrete valuation domain of rank one, $Pa \neq 0$ and P is the radical of Pa. Then the Krull–Schmidt theorem does not hold for the class* FG-R.

A commutative ring R is said to be *almost Henselian* [V3] if every proper homomorphic image of R is a local Henselian ring. Let R be an almost Henselian valuation ring. Then the endomorphism ring of every indecomposable finitely generated R-module is local. In particular, the Krull–Schmidt theorem holds for FG-R [V3, Theorem 19].

COROLLARY 3.12 [V3, Corollary 21]. *Let R be a discrete valuation ring of finite rank. The Krull–Schmidt theorem holds for the class* FG-R *if and only if R is almost Henselian.*

If R is a valuation domain and $\mathrm{FG}_{\oplus 2}$-R is the class of all modules that are direct sums of finitely many modules generated by $\leqslant 2$ elements, then the Krull–Schmidt theorem holds for $\mathrm{FG}_{\oplus 2}$-R [SZ]. For a commutative ring R, let FG_n-R denote the class of all R-modules that can be generated by $\leqslant n$ elements. If a module over a valuation domain is generated by $\leqslant 4$ elements and is not indecomposable, then it decomposes as the direct sum of a cyclic module and a module generated by $\leqslant 3$ elements or as the direct sum of two modules generated by $\leqslant 2$ elements. In both cases the direct sum decomposition as a direct sum of indecomposables is unique [SZ], so that the Krull–Schmidt theorem holds for the class FG_4-R for every valuation domain R. Zanardo [Z3] showed that there exist valuation domains R such that the Krull–Schmidt theorem does not hold for the class FG_6-R. It is not known whether the Krull–Schmidt theorem holds for the class FG_5-R for any valuation domain R.

3.5. *Finitely generated modules over subrings of* $\mathbf{Z}^n$

A nice set of examples showing a number of ways in which the Krull–Schmidt uniqueness fails for the class FG-R of Noetherian modules over suitable subrings R of the direct product $\mathbf{Z}^n$ was discovered by Levy [L3] (these rings are all commutative Noetherian rings of Krull dimension one). Levy showed that:

(1) (Non-uniqueness of the number of indecomposable direct summands) For every $m \geqslant 3$ there is a Noetherian module M over a suitable subring R of $\mathbf{Z}^{m-1}$ that has a direct sum decomposition $M = M_{1,i} \oplus M_{2,i} \oplus \cdots \oplus M_{i,i}$ into i indecomposable direct summands $M_{1,i}, M_{2,i}, \ldots, M_{i,i}$ for every $i = 2, 3, \ldots, m$.

(2) (Failure of the cancellation property) There exist Noetherian modules M, N, N' over suitable subrings R of $\mathbf{Z}^n$ such that $M \oplus N \cong M \oplus N'$, but N is not isomorphic to N'. We shall see in Lemma 4.3(a) that the cancellation property holds for modules whose endomorphism ring is semilocal.

(3) (Failure of the n-th root property) There exist Noetherian modules M, N over a suitable subring R of $\mathbf{Z}^m$ such that $M^n \cong N^n$ (for suitable $n > 1$), but M is not isomorphic to N. We shall see in Lemma 4.3(b) that the n-th root property holds for modules whose endomorphism ring is semilocal.

(4) (Failure of the interchange of localizations property) There exist four pairwise non-isomorphic indecomposable Noetherian modules H, K, M, N over a suitable subring R of $\mathbf{Z}^n$ and two maximal ideals P and Q of R such that $H \oplus K \cong M \oplus N$, $H_P \cong M_P \ncong K_P \cong N_P$ and $H_Q \cong N_Q \ncong K_Q \cong M_Q$.

In (1), (4), (5), and (7) the symbol $\mathbf{Z}$ denotes the ring of integers, but it can be any integral domain with at least four maximal ideals. In (2) and (3) also, the symbol $\mathbf{Z}$ can denote the ring of integers, but it can be any integral domain with at least four maximal ideals that satisfies an additional non-triviality condition involving non-liftability of units modulo maximal ideals to units of $\mathbf{Z}$. If $\mathbf{Z}$ denotes the polynomial ring $F[x]$, where F is any field which is not an algebraic extension of a finite field, then there are subrings R of the direct product $\mathbf{Z}^n$ that satisfy all of (1)–(7).

(5) (Indecomposability is not a local property) There exists an indecomposable Noetherian R-module M such that M_P is not indecomposable, for some subring R of $\mathbf{Z}^n$ and some maximal ideal P of R.

(6) (Projective modules of rank 1) There exists a projective R-module M of rank 1 such that $M^n \ncong R^n$ for every positive n, R a suitable subring of $\mathbf{Z}^2$. Here $\mathbf{Z}$ can be $F[x]$, F a field that is not an algebraic extension of a finite field, but $\mathbf{Z}$ cannot be the ring of the integers.

(7) (Isomorphism is not a local property) There exist two non-isomorphic Noetherian modules M and N such that $M_P \cong N_P$ for every maximal ideal P of a subring R of $\mathbf{Z}^n$.

We note that, in example (1) above, the integer m must be selected before the ring R and the R-module M are formed. If one does not insist that M be Noetherian, then this is not necessary, as shown by the following example of Osofsky which Levy obtains, as a corollary of his methods.

(8) (B. Osofsky [O]) There exists a (non-Noetherian!) module M_R that is not the direct sum of infinitely many submodules, but that is the direct sum of m indecomposable modules for every $m \geqslant 2$. In particular, M_R is not an almost Krull–Schmidt module. (Here the

ring R is an ultraproduct of subrings of $\mathbf{Z}$. The M Levy obtains is a ring, considered as a module over itself. Like every ring with identity, M cannot be written as the direct sum of an infinite number of non-zero submodules.)

Levy's examples can be used, via Theorem 2.4, to reconstruct some of the classical examples of the failure of the Krull–Schmidt theorem for torsion-free Abelian groups of finite rank mentioned in Section 3.2; see [A2].

3.6. *Commutative rings whose finitely generated modules are direct sums of cyclic submodules or of n-generated submodules*

A commutative ring R is called an FG(n)-ring if every finitely generated R-module is a direct sum of submodules generated by $\leqslant n$ elements. Thus FG(1)-rings are exactly the commutative rings over which every finitely generated module is a direct sum of cyclics. For instance, principal ideal domains are FG(1)-rings (Theorem 3.1) and Dedekind domains are FG(2)-rings (Theorem 3.2). FG(n)-rings are also called *rings of bounded module type*. The characterization of FG(1)-rings, the solution of the so-called Kaplansky's problem, was found in 1976 (for a nice presentation of this topic, the history of the problem and the proofs of the results presented in this section, see [B4,V2] or [WW2]). In order to present the characterization of FG(1)-rings, we need a number of definitions. Recall that a commutative ring R is *maximal* (or *linearly compact*) if whenever $\{I_\alpha \mid \alpha \in A\}$ is a family of ideals of R, $\{x_\alpha \mid \alpha \in A\}$ is a family of elements of R and the family of cosets $\{x_\alpha + I_\alpha \mid \alpha \in A\}$ has the finite intersection property (that is, the intersection of every finite subfamily is non-empty), then $\bigcap_{\alpha \in A} x_\alpha + I_\alpha \neq \emptyset$. A commutative ring is *almost maximal* if R/I is a maximal ring for all non-zero ideals I of R, and is *locally almost maximal* if R_M is an almost maximal ring for all maximal ideals M of R.

If R and S are valuation rings, $R \subseteq S$ and M is the maximal ideal of R, then S is an *immediate extension* of R if the embedding $R \to S$ induces an isomorphism between the value groups of R and S and an isomorphism $R/M \to S/MS$. A valuation domain R is maximal if and only if it is has no proper immediate extension. Any valuation domain R has an immediate extension $\widetilde{R}$ that is a maximal valuation domain ([K5]; see also [W2, p. 717]). The R-module $\widetilde{R}$ is the pure-injective envelope of the R-module R, so that $\widetilde{R}$ is unique (up to isomorphism) as an R-module, though, [K1], it is not unique as an R-algebra. A valuation ring that is not a domain is maximal if and only if it is almost maximal [G1]. A valuation domain is maximal if and only if it is almost maximal and complete in the valuation topology. The completion of a valuation domain R in the valuation topology will be denoted by $\widehat{R}$. If $K = K(R)$, $\widetilde{K} = K(\widetilde{R})$ are the fields of fractions of R, $\widetilde{R}$, respectively, then $\widetilde{K}$ is the completion of K as a valuated field and the degree $[\widetilde{K} : K]$ is the torsion-free rank of the R-module $\widetilde{R}$. A valuation domain R is almost maximal if and only if $\widehat{R} = \widetilde{R}$.

A commutative integral domain R is *h-local* if every non-zero element of R belongs to only finitely many maximal ideals of R and every non-zero prime ideal of R is contained in only one maximal ideal. A *Bezout ring* is a commutative ring in which every finitely generated ideal is principal. A right module M is said to be *uniserial* if its lattice of submodules is linearly ordered by set inclusion, that is, if for any submodules A and B of M either $A \subseteq B$ or $B \subseteq A$. We are ready to define torch rings. A *torch ring* is a commutative ring R

that satisfies all the following conditions: (1) the ring R is not local; (2) the ring R has a unique minimal prime ideal P and P is a non-zero uniserial R-module; (3) the ring R/P is h-local; (4) R is a locally almost maximal Bezout ring. Torch rings were first considered by Shores and R. Wiegand in [SW], and their name was suggested by Vámos in [V1]. The name refers to the ideal lattice of a torch ring R: all ideals of R either contain P or belong to the linearly ordered set of the ideals contained in P.

THEOREM 3.13 (W. Brandal, T. Shores, P. Vámos, R. Wiegand and S. Wiegand). *A commutative ring is an* FG(1)*-ring if and only if it is a finite direct product of rings of the following three types*:

(a) *maximal valuation rings*;
(b) *almost maximal Bezout domains*;
(c) *torch rings.*

THEOREM 3.14 [B4, Theorem 9.2]. *If R is an* FG(1)*-ring, the Krull–Schmidt theorem holds for the class* FG-R.

Another form of uniqueness of direct sum decompositions is given by the so-called canonical form decompositions. A *canonical form decomposition* of a module M over a commutative ring R is a decomposition of the form $M \cong R/I_1 \oplus R/I_2 \oplus \cdots \oplus R/I_t$ where $I_1 \subseteq I_2 \subseteq \cdots \subseteq I_t$ are proper ideals of R. The following result, proved by Kaplansky, partially extends Theorem 3.1.

THEOREM 3.15 [K3, Theorem 9.3]. *If a module M over a commutative ring R has a canonical form decomposition, then the canonical form decomposition is unique up to isomorphism. More generally, if R is a ring in which every one-sided ideal is two-sided, $R/I_1 \oplus R/I_2 \oplus \cdots \oplus R/I_t \cong R/I'_1 \oplus R/I'_2 \oplus \cdots \oplus R/I'_s$ and $I_1 \subseteq I_2 \subseteq \cdots \subseteq I_t$, $I'_1 \subseteq I'_2 \subseteq \cdots \subseteq I'_s$ are proper ideals of R, then $t = s$ and $I_k = I'_k$ for every $k = 1, 2, \ldots, t$.*

THEOREM 3.16 [B4, Theorem 9.5]. *Every finitely generated module over an* FG(1)*-ring has a canonical form decomposition.*

We now turn our attention to FG(n)-rings for arbitrary n. If R is an FG(n)-ring, then: (1) the localization R_M is an FG(n) valuation ring for every maximal ideal M of R (Warfield, [W4, Theorem 2]); (2) R has only finitely many minimal prime ideals (Midgarten and S. Wiegand, [MW, Theorem 1.8]); (3) R_M/P_M is complete for all maximal ideals M and prime ideals $P \subseteq M$ that are not minimal primes (Zanardo, [Z2, Theorem 4]).

THEOREM 3.17 (Warfield, [W4, Corollary 4.1]). *Let R be a commutative Noetherian* FG(n)*-ring for some $n \geqslant 1$. Then R is a finite product of Dedekind domains and valuation rings. In particular, R is an* FG(2)*-ring.*

THEOREM 3.18 (Vámos, [V3]). *If R is a valuation domain that is an* FG(n)*-ring for some $n \geqslant 1$ and is an algebra over the field* $\mathbf{Q}$ *of rational numbers, then R is an almost maximal valuation domain, i.e. an* FG(1)*-ring.*

More generally, Vámos conjectured that every FG(n) local commutative ring has to be almost maximal (as we have already said above, it is necessarily a valuation ring by [W4, Theorem 2]). Couchot proved in [C5,C6] that if every FG(n) valuation domain of Krull dimension 1 is almost maximal, then Vámos' conjecture is true for every local commutative ring.

As a corollary of Theorem 3.9, Vámos [V3, Theorem 19] obtains:

COROLLARY 3.19. *If R is an* FG(n) *valuation domain for some $n \geqslant 1$, the Krull–Schmidt theorem holds for the class* FG-R.

3.7. *Torsion-free modules over valuation domains*

In this section, R will always be a commutative integral domain, usually a valuation domain, and FRTF-R will denote the class of all finite rank torsion-free R-modules.

THEOREM 3.20 ([S3, Theorem 9], [V3, Lemma 14]). *The endomorphism ring* $\mathrm{End}(M_R)$ *of a finite rank torsion-free indecomposable module M_R over a Henselian valuation domain R is local. Hence the Krull–Schmidt theorem holds for the class* FRTF-R *whenever R is a Henselian valuation domain.*

Historically, the first example that shows that the Krull–Schmidt theorem does not hold for FRTF-$\mathbf{Z}_p$, where $\mathbf{Z}_p$ is the discrete valuation domain of the integers localized at a prime p, is due to Butler and goes back to the sixties (see [A1, Example 2.15] and [W6, p. 461]). The following theorem, due to Vámos [V3, Theorem 17], shows why $\mathbf{Z}_p$ can be used to construct examples of torsion-free modules for which Krull–Schmidt does not hold. The "if" part of the theorem was also proved by Lady.

THEOREM 3.21. *Let R be a discrete rank one valuation domain. The Krull–Schmidt theorem holds for the class* FRTF-R *if and only if R is Henselian.*

Theorem 3.21 does not hold for arbitrary valuation domains: Vámos gave examples of non-Henselian valuation domains R such that the Krull–Schmidt theorem holds for FRTF-R, and proved in [V3, Theorems 5 and 10 and last note of the paper] that

THEOREM 3.22. *If R is a non-Henselian valuation domain and the maximal immediate extension $\widetilde{R}$ of R belongs to* FRTF-R, *then the Krull–Schmidt theorem holds for the class* FRTF-R.

Notice that the maximal immediate extension $\widetilde{R}$ of a valuation domain R is always a torsion-free R-module. Hence it belongs to FRTF-R if and only if it has finite rank. May and Zanardo, extending previous results of Goldsmith and May [GM, Theorem 2], proved in [MZ, Theorem 3] that

THEOREM 3.23. *If a valuation domain R contains a prime ideal P such that R/P is not Henselian and the completion $\widehat{R/P}$ in the valuation topology of R/P is an R/P-module of torsion-free rank $\geqslant 6$, then the Krull–Schmidt theorem does not hold for* FRTF-R.

For instance, Theorem 3.23 applies to the valuation domain $R = \mathbf{Z}_p + x\mathbf{Q}[[x]]$, which is a complete valuation domain, with P the prime ideal $x\mathbf{Q}[[x]]$ of R, because the completion $\widehat{\mathbf{Z}}_p$ of $\mathbf{Z}_p$ in the valuation topology has infinite torsion-free rank as a $\mathbf{Z}_p$-module.

The following is a corollary of Theorem 3.21.

COROLLARY 3.24 [V3, Corollary 18]. *Let R be a Dedekind domain. The Krull–Schmidt theorem holds for the class* FRTF-R *if and only if R is a Henselian valuation domain.*

For a commutative integral domain R let $\mathrm{fr}(R)$ denote the supremum of the ranks of all indecomposable finite rank torsion-free R-modules. One can prove [AD] that if R is a discrete valuation domain, then $\mathrm{fr}(R) = 1, 2, 3$ or ∞.

THEOREM 3.25 ([K2, Theorem 12], [M2, Theorem 65]). *For a valuation domain R,* $\mathrm{fr}(R) = 1$ *if and only if R is a maximal valuation domain.*

More generally, the integrally closed domains with $\mathrm{fr}(R) = 1$ were determined by Matlis [M2].

THEOREM 3.26 [V3]. *Let n be a positive integer and let R be a valuation domain such that every torsion-free R-module is a direct sum of modules of rank $\leqslant n$. Then the Krull–Schmidt theorem holds for the class of torsion-free R-modules. If, moreover, R is an algebra over the field $\mathbf{Q}$ of rational numbers, then every torsion-free R-module is a direct sum of modules of rank $\leqslant 2$.*

4. Modules with a semilocal endomorphism ring

4.1. *Examples and properties of modules whose endomorphism ring is semilocal*

The aim of this section is to study modules whose endomorphism ring is semilocal. Further details and the proofs can be found in [F2, Chapter 4]. By Theorem 2.4, if a module M_R has a semilocal endomorphism ring E, the study of the direct sum decompositions of M_R reduces to the study of the idempotents of the semilocal ring E. Theorem 4.1 will allow us to give examples of modules with a semilocal endomorphism ring. For a module M_R set $\delta(M_R) = \mathrm{codim}(\mathrm{End}(M_R))$. Thus $\delta(M_R)$ is a non-negative integer if and only if $\mathrm{End}(M_R)$ is semilocal (Theorem 2.12); otherwise, $\delta(M_R) = \infty$. From Proposition 2.13 one immediately obtains that $\delta(M_R \oplus M'_R) = \delta(M_R) + \delta(M'_R)$ for any M_R, M'_R. Immediate consequences of this formula and the definition of the dimension δ are that: $\delta(M_R) = 0$ if and only if $M_R = 0$; $\delta(M_R^n) = n\delta(M_R)$; $\delta(M_R) = 1$ if and only if $\mathrm{End}(M_R)$ is local. Notice that under the conditions of the Krull–Schmidt theorem, that is, if $M_R = M_1 \oplus \cdots \oplus M_t$ is the direct sum of t modules M_i and all the endomorphism rings $\mathrm{End}(M_i)$ are local, then $\delta(M_R) = t$.

THEOREM 4.1 (D. Herbera and A. Shamsuddin [HS]). *Let M_R be a module over a ring R. Then the following statements hold*:

(a) $\delta(M_R) \leqslant \dim(M_R) + \operatorname{codim}(M_R)$.

(b) *If every injective endomorphism of M_R is bijective, then* $\delta(M_R) \leqslant \dim(M_R)$ (Camps and Dicks [CD, Theorem 5]).

(c) *If every surjective endomorphism of M_R is bijective, then* $\delta(M_R) \leqslant \operatorname{codim}(M_R)$.

EXAMPLES 4.2.

(1) Artinian modules have semilocal endomorphism rings [CD]. In fact, every injective endomorphism of an Artinian module is bijective, so that Theorem 4.1(b) can be applied. More generally, modules that are linearly compact in the discrete topology have semilocal endomorphism rings [HS].

(2) Noetherian modules of finite dual Goldie dimension have semilocal endomorphism rings. This is the dual of the previous Example 4.2(1). In particular, finitely generated right modules over semilocal right Noetherian rings have semilocal endomorphism rings.

(3) Modules of finite Goldie dimension and finite dual Goldie dimension have semilocal endomorphism rings by Theorem 4.1(a).

(4) Let R be either a semilocal commutative principal ideal domain or a valuation domain. If M_R is a finite rank torsion-free R-module, then $\operatorname{End}(M_R)$ is semilocal by Example 2.11(6).

(5) Let k be a semilocal commutative ring and R a module-finite k-algebra. The endomorphism ring of any finitely generated R-module is a semilocal ring [W7, Lemma 2.3].

(6) For further examples of modules whose endomorphism ring is semilocal, see [HS] and [F2, §4.3].

If M_R is a module and its endomorphism ring $E = \operatorname{End}(M_R)$ is semilocal, then every direct sum decomposition of M_R has finitely many direct summands, that is, if $M_R = \bigoplus_{\lambda\in\Lambda} M_\lambda$ is a decomposition of M_R as a direct sum of non-zero modules M_λ and $E = \operatorname{End}(M_R)$ is semilocal, then the index set Λ must be necessarily finite. In the next lemma we have collected some further properties of modules whose endomorphism ring is semilocal.

LEMMA 4.3. *Let M_R, N_R, N'_R be modules over an arbitrary ring R and suppose* $\operatorname{End}(M_R)$ *is semilocal. The following properties hold*:

(a) (Cancellation property) *If $M \oplus N \cong M \oplus N'$, then $N \cong N'$.*

(b) (n-th root property) *If $M^n \cong N^n$ for some positive integer n, then $M \cong N$.*

(c) *The module M_R is an almost Krull–Schmidt module.*

Part (a) of Lemma 4.3 follows from results of Bass [B1], who proved that semilocal rings have stable range 1, and Evans [E], who showed that modules whose endomorphism ring has stable range 1 cancel from direct sums. We shall not insist on the cancellation property and the n-th root property here. For a nice survey on the n-th root property, see [L2].

Notice that, by (c), not only are all direct sum decompositions of a module with semilocal endomorphism ring finite, but also there are just a finite number of such decompositions.

We conclude this section with a further result about modules M_R with a semilocal endomorphism ring, that is, modules M_R with $\delta(M_R) < \infty$.

THEOREM 4.4 [FH3, Theorem 2.3]. *Let M_R be a module over a ring R with $\delta(M_R) = n < \infty$. Let $N_1, N_2, \ldots, N_m$ be R-modules such that M_R is isomorphic to a direct summand of $\bigoplus_{i=1}^m N_i$. Then there exists a subset α of $\{1, 2, \ldots, m\}$ of cardinality $\leqslant n$ such that M_R is isomorphic to a direct summand of $\bigoplus_{j \in \alpha} N_j$.*

4.2. *Krull–Schmidt fails for Artinian modules*

Every Artinian right module decomposes as a finite direct sum of indecomposable modules, but the Krull–Schmidt theorem does not hold for the class of Artinian modules, as the following two examples that appear in [FHLV] show.

EXAMPLE 4.5 (*Non-uniqueness of the number of indecomposable summands*). Let $n \geqslant 2$ be an integer. There exists an Artinian module M over a suitable ring R that has a direct sum decomposition

$$M = M_{1,i} \oplus M_{2,i} \oplus \cdots \oplus M_{i,i}$$

into i indecomposable direct summands $M_{1,i}, M_{2,i}, \ldots, M_{i,i}$ for every $i = 2, 3, \ldots, n$.

EXAMPLE 4.6 (*Simple failure of Krull–Schmidt for Artinian modules*). There exist four indecomposable, pairwise non-isomorphic, Artinian modules M_1, M_2, M_3, M_4 over a suitable ring R such that $M_1 \oplus M_2 \cong M_3 \oplus M_4$.

These two examples were constructed showing that every module-finite algebra over a semilocal commutative Noetherian ring can be realized as the endomorphism ring of a suitable Artinian right module over a suitable ring [FHLV]. Further nice examples of Artinian modules with particular direct sum decompositions were later constructed by Yakovlev [Y1] and Pimenov–Yakovlev [PY]. Making use of the techniques of Pimenov–Yakovlev, Ringel showed in [R3] that the Krull–Schmidt theorem does not necessarily hold for the class of Artinian R-modules even when the base ring R is *local*. This answered a question posed in [F2, Problem 7, p. 268]. For instance, Ringel constructed five pairwise non-isomorphic indecomposable Artinian modules M_1, M_2, M_3, M_4, M_5 over a local ring R such that $M_1 \oplus M_2 \cong M_3 \oplus M_4 \oplus M_5$.

Notice that Warfield showed in 1969 that:

PROPOSITION 4.7 [W3, Proposition 5]. *If a ring R is either right Noetherian or commutative, then the Krull–Schmidt theorem holds for the class of all Artinian right R-modules.*

5. Biuniform modules

5.1. *Generalities*

A module A_R is called *uniform* if it is non-zero and the intersection of any two non-zero submodules of A_R is non-zero. Thus "uniform" means "of Goldie dimension one". Dually, a module A_R is *couniform* (or *hollow*) if it is non-zero and the sum of any two proper submodules of A_R is a proper submodule of A_R (equivalently, if its dual Goldie dimension is one). A module is *biuniform* if it is uniform and couniform. For example, non-zero uniserial modules (recall that a right module M is said to be *uniserial* if its lattice of submodules is linearly ordered by set inclusion) are biuniform. Sections 5.3 and 5.4 will be completely devoted to uniserial modules and direct sum decompositions of direct sums of uniserial modules. If A_R is a biuniform module over an arbitrary ring R, then its endomorphism ring $E = \mathrm{End}(A_R)$ has at most two maximal right ideals, the subset I of E whose elements are all the endomorphisms of A_R that are not injective, and the subset K of E whose elements are all the endomorphisms of A_R that are not surjective [F2, Theorem 9.1]. These two right ideals I and K are two-sided completely prime ideals of E, and every proper right ideal of E and every proper left ideal of E is contained either in I or in K. Moreover, if the ideals I and K are comparable, then E is a local ring and $I \cup K$ is its maximal ideal, whereas if I and K are not comparable, then $J(E) = I \cap K$ and $E/J(E)$ is canonically isomorphic to the direct product of the two division rings E/I and E/K.

5.2. *Weak Krull–Schmidt theorem for biuniform modules*

Two modules A and B *belong to the same monogeny class* if there exist a monomorphism $A \to B$ and a monomorphism $B \to A$. In this case, we write $[A]_m = [B]_m$. Similarly, A and B *belong to the same epigeny class* if there exist an epimorphism $A \to B$ and an epimorphism $B \to A$, and in this case we write $[A]_e = [B]_e$. Clearly, belonging to the same monogeny class and belonging to the same epigeny class are two equivalence relations.

PROPOSITION 5.1 ([F1, Proposition 1.6], [F2, Proposition 9.3]). *Two biuniform modules A and B are isomorphic if and only if $[A]_m = [B]_m$ and $[A]_e = [B]_e$.*

THEOREM 5.2 ([DF1, Theorem 3.1], [F2, Theorem 9.11]). *If $\{A_i \mid i \in I\}$ and $\{B_j \mid j \in J\}$ are two families of biuniform modules over a ring R and there exist two bijections $\sigma, \tau : I \to J$ such that $[A_i]_m = [B_{\sigma(i)}]_m$ and $[A_i]_e = [B_{\tau(i)}]_e$ for every $i \in I$, then $\bigoplus_{i \in I} A_i \cong \bigoplus_{j \in J} B_j$.*

Only a half of the implication in Theorem 5.2 can be reversed, as the next theorem and Example 5.11 show.

THEOREM 5.3 ([DF1, Theorem 3.3], [F2, Theorem 9.12]). *Let R be a ring and let $\{A_i \mid i \in I\}$, $\{B_j \mid j \in J\}$ be two families of biuniform right R-modules such that $\bigoplus_{i \in I} A_i \cong \bigoplus_{j \in J} B_j$. Then there exists a bijection $\sigma : I \to J$ such that $[A_i]_m = [B_{\sigma(i)}]_m$ for every $i \in I$.*

The next result describes when two direct sums of biuniform modules are isomorphic in the finite case.

THEOREM 5.4 (Weak Krull–Schmidt theorem for biuniform modules, [F1, Theorem 1.9]). *Let $A_1, \dots, A_t, B_1, \dots, B_s$ be biuniform modules over a ring R. Then $A_1 \oplus \cdots \oplus A_t \cong B_1 \oplus \cdots \oplus B_s$ if and only if $t = s$ and there are two permutations σ, τ of $\{1, 2, \dots, t\}$ such that $[A_i]_m = [B_{\sigma(i)}]_m$ and $[A_i]_e = [B_{\tau(i)}]_e$ for every $i = 1, 2, \dots, t$.*

We shall see in Example 5.7 that Theorem 5.4 cannot be improved, in the sense that, over suitable rings R, the Krull–Schmidt theorem does not hold for the class of the R-modules that are finite direct sums of biuniform modules.

In this setting we mention a result due to L. Diracca. Part (a) was first proved by Zanardo for modules over commutative rings [Z1, Theorem 2].

THEOREM 5.5. *Let R be an arbitrary ring.*

(a) *If $A_1, \dots, A_t, B_1, \dots, B_s$ are uniform R-modules and $[A_1 \oplus \cdots \oplus A_t]_m = [B_1 \oplus \cdots \oplus B_s]_m$, then $t = s$ and there is a permutation σ of $\{1, 2, \dots, t\}$ such that $[A_i]_m = [B_{\sigma(i)}]_m$ for every $i = 1, 2, \dots, t$.*

(b) *Dually, if $A_1, \dots, A_t, B_1, \dots, B_s$ are couniform R-modules and $[A_1 \oplus \cdots \oplus A_t]_e = [B_1 \oplus \cdots \oplus B_s]_e$, then $t = s$ and there is a permutation τ of $\{1, 2, \dots, t\}$ such that $[A_i]_e = [B_{\tau(i)}]_e$ for every $i = 1, 2, \dots, t$.*

5.3. *Uniserial modules*

A module is *serial* if it is a direct sum of uniserial submodules. In particular, serial modules of finite Goldie dimension are exactly direct sums of finitely many uniserial modules, and their endomorphism ring is semilocal (Theorem 4.1(a)). For instance, the Prüfer groups $\mathbf{Z}(p^\infty)$ are uniserial modules over the ring $\mathbf{Z}$ of integers, so that every torsion injective Abelian group is a serial $\mathbf{Z}$-module. Finite Abelian groups are serial $\mathbf{Z}$-modules. Vector spaces are serial modules. More generally, semisimple modules are serial.

A *chain ring* is a ring R such that both R_R and ${}_RR$ are uniserial modules. A ring R is *serial* if both R_R and ${}_RR$ are serial modules. Hence chain rings are serial rings, and valuation rings are exactly commutative chain rings. Semisimple Artinian rings and rings of $n \times n$ upper triangular matrices with entries in a field are serial rings.

Let R be a serial ring. Then there exists a complete set of orthogonal idempotents $\{e_1, e_2, \dots, e_n\}$ of R such that both e_iR and Re_i are uniserial modules. Moreover, all the rings e_iRe_i are chain rings. Every indecomposable projective right R-module is isomorphic to e_iR for some $i = 1, 2, \dots, n$ (although e_iR and e_jR may be isomorphic for $i \neq j$). Since serial rings are semiperfect, the Krull–Schmidt theorem holds for the class of all projective modules over a serial ring (Proposition 2.6). In particular, every projective module over a serial ring is a direct sum of indecomposables. This fact plus the fact that every indecomposable projective module is isomorphic to an e_iR imply that every projective module over a serial ring is serial. Moreover:

THEOREM 5.6 [W5]. *Every finitely presented module over a serial ring is serial.*

The next example shows that the Krull–Schmidt theorem does not hold for the class of serial modules.

EXAMPLE 5.7 [F1]. Let $n \geqslant 2$ be an integer. There exist n^2 pairwise non-isomorphic finitely presented uniserial modules $U_{i,j}$ $(i, j = 1, 2, \dots, n)$ over a suitable serial ring R satisfying the following properties:

(a) for every $i, j, k, \ell = 1, 2, \dots, n$, $[U_{i,j}]_m = [U_{k,\ell}]_m$ if and only if $i = k$;

(b) for every $i, j, k, \ell = 1, 2, \dots, n$, $[U_{i,j}]_e = [U_{k,\ell}]_e$ if and only if $j = \ell$.

Thus the n^2 modules $U_{i,j}$ $(i, j = 1, 2, \dots, n)$ can be represented as the entries of a square $n \times n$ matrix in such a way that two modules $U_{i,j}$ belong to the same monogeny class if and only if they are on the same row, and they belong to the same epigeny class if and only if they are on the same column. If σ and τ are any two permutations of $\{1, 2, \dots, n\}$, then $U_{1,1} \oplus U_{2,2} \oplus \cdots \oplus U_{n,n} \cong U_{\sigma(1),\tau(1)} \oplus U_{\sigma(2),\tau(2)} \oplus \cdots \oplus U_{\sigma(n),\tau(n)}$ by Theorem 5.4, and these two decompositions have the property that σ and τ are exactly the two permutations that mix the monogeny classes and the epigeny classes, respectively (compare with Theorem 5.4).

By Theorem 5.4 and Example 5.7 the Krull–Schmidt theorem does not hold for the class of finitely generated (respectively, finitely presented) uniserial modules over a serial ring.

In opposition to the general case of finitely presented modules over *serial* rings, the Krull–Schmidt theorem holds for finitely presented modules over *chain* rings, a result due to Puninski (see [F2, Theorem 9.19]):

THEOREM 5.8 (Puninski). *The Krull–Schmidt theorem holds for the class of finitely presented modules over chain rings.*

A module M_R over a ring R is said to be *small* if for every family $\{N_i \mid i \in I\}$ of R-modules and any homomorphism $\varphi : M_R \to \bigoplus_{i \in I} N_i$, there exists a finite subset F of I such that $\varphi(M) \subseteq \bigoplus_{i \in F} N_i$. For instance, finitely generated modules are small. A proof due to Fuchs and Salce shows that every uniserial module that is not countably generated is small (cf. [DF1, Lemma 4.2] and [FS1, Lemma 24].) An R-module M_R is called *quasi-small* if for any family $\{N_i \mid i \in I\}$ of R-modules such that M_R is isomorphic to a direct summand of $\bigoplus_{i \in I} N_i$, there exists a finite subset F of I such that M_R is isomorphic to a direct summand of $\bigoplus_{i \in F} N_i$. Small modules, modules with local endomorphism rings and Artinian modules are quasi-small ([DF1, p. 112] and [FH3, Corollary 3.3]).

THEOREM 5.9 [DF1, Theorem 4.7]. *Let $\{U_i \mid i \in I\}$ and $\{V_j \mid j \in J\}$ be two families of non-zero uniserial modules over an arbitrary ring R and suppose that $\bigoplus_{i \in I} U_i \cong \bigoplus_{j \in J} V_j$. Set $I' = \{i \in I \mid U_i$ is quasi-small$\}$ and $J' = \{j \in J \mid V_j$ is quasi-small$\}$. Then there is a bijection $\tau : I' \to J'$ such that $[U_i]_e = [V_{\tau(i)}]_e$ for every $i \in I'$.*

COROLLARY 5.10. *The following conditions on a ring R are equivalent*:

(a) *All uniserial right R-modules are quasi-small.*

(b) *For two arbitrary families $\{U_i \mid i \in I\}$ and $\{V_j \mid j \in J\}$ of uniserial right R-modules, $\bigoplus_{i \in I} U_i \cong \bigoplus_{j \in J} V_j$ if and only if there exist two bijections $\sigma : I \to J$ and $\tau : I \to J$ such that $[U_i]_m = [V_{\tau(i)}]_m$ and $[U_i]_e = [V_{\tau(i)}]_e$ for every $i \in I$.*

EXAMPLE 5.11. An interesting example discovered by Puninski [P3] shows that there exist rings R for which the equivalent conditions of Corollary 5.10 do not hold, that is, over which there exist uniserial modules that are not quasi-small. This solves a problem posed in [DF1, p. 111] (see also [F2, Problem 15 on p. 269]). Recall that a chain ring R is called a *nearly simple* chain ring if it has exactly three two-sided ideals, which must be necessarily the ideals $0 \subset J(R) \subset R$. For an example of a nearly simple chain domain, see [BBT, §6.5]. If R is a nearly simple chain domain and a, b are non-zero non-invertible elements of R, then the right R-modules R/aR and R/bR are always isomorphic [P3]. Hence, let a be a non-zero non-invertible element of R and $U_R = R/aR$. Every finitely presented right module over the nearly simple chain domain R is isomorphic to $R_R^n \oplus U^m$ for two uniquely determined non-negative integers n and m (Theorems 5.6 and 5.8). Puninski showed that there exists a suitable countably generated uniserial module V_R such that $V_R \oplus U_R^{(\aleph_0)} \cong U_R^{(\aleph_0)}$, where $U_R^{(\aleph_0)}$ denotes the direct sum of countably many copies of U_R. Since U_R is cyclic and V_R is not, the epigeny classes of U_R and V_R are different, so that in the two direct sum decompositions $V_R \oplus U_R^{(\aleph_0)} \cong U_R^{(\aleph_0)}$ there cannot be a one-to-one correspondence between the epigeny classes.

A suitable generalization of Theorem 5.4 to arbitrary families of uniserial modules is not known yet.

5.4. *Direct summands of serial modules*

Until now we have studied direct sum decompositions of a serial module into uniserial modules and the uniqueness of these decompositions up to isomorphism. Now a natural question is: Does every direct sum decomposition of a serial module M_R refine to a direct sum decomposition into uniserial direct summands? Equivalently, if N_R is a direct summand of a serial module M_R, is N_R also a serial module? Here are some partial answers to this question.

If M_R is a direct sum of uniserial modules each of which has a local endomorphism ring, then N_R is serial [F2, Corollary 2.54]. Thus the answer to our question is "yes" in this case. This happens, for instance, for any serial R-module when the base ring R is either commutative or right Noetherian.

The answer is "no", in general, for a serial module M_R of infinite Goldie dimension over an arbitrary ring R, as is shown by Puninski in [P4]. Puninski considers uniserial modules over prime coherent nearly simple chain rings that are not domains. Over such a ring R, he constructs a pure-projective module that is not a direct sum of indecomposable modules. This shows that not every direct summand of a serial module over a chain ring is serial, solving Problems 10 and 11 of [F2, pp. 268, 269]. An example of a prime coherent nearly simple chain ring R that is not a domain can be found in [D].

It is not known yet whether every direct summand of a serial module *of finite Goldie dimension* is serial [F2, Problem 9 on p. 268]. The best result in this direction is the following theorem, concerning the case of a direct sum of finitely many isomorphic uniserial modules.

THEOREM 5.12 [DF2, Theorem 2.7]. *If R is an arbitrary ring, U is a uniserial right R-module and $n \geqslant 0$ is an integer, then every direct summand of U^n is isomorphic to U^m for some $m \leqslant n$.*

Some generalizations of Theorem 5.12 can be found in [DF2].

6. Projective modules over semilocal rings

For the proofs of the results presented in this section, see [FH1] and [FH2].

A module A_S whose endomorphism ring is semilocal is an almost Krull–Schmidt module, that is, has only finitely many direct sum decompositions up to isomorphism (Lemma 4.3(c)). It is therefore natural to try to describe the direct sum decompositions of A_S. More generally, we shall consider the direct sum decompositions of the objects in the category $\mathrm{add}(A_S)$. By Theorem 2.4, this is equivalent to describing finitely generated projective modules over the semilocal ring $R = \mathrm{End}(A_S)$.

The set of isomorphism classes of finitely generated projective right modules over a ring R can be given the structure of a commutative monoid in the following way. For every right R-module P_R, let $\langle P_R \rangle$ denote the isomorphism class of P_R, that is, the class of all right R-modules isomorphic to P_R. Let $V(R)$ be the additive commutative monoid whose elements are the isomorphism classes $\langle P_R \rangle$ of all finitely generated projective right R-modules P_R, and whose addition is defined by $\langle P_R \rangle + \langle Q_R \rangle = \langle P_R \oplus Q_R \rangle$.

For example, if R is a semisimple Artinian ring, each finitely generated R-module is a finite direct sum of simple modules. It follows that $V(R)$ is the free commutative monoid having the isomorphism classes of simple R-modules as a free set of generators, that is, $V(R) \cong \mathbf{N}^n$, where n is the number of simple R-modules up to isomorphism.

If for every ring morphism $\varphi : R \to S$ we denote by $V(\varphi) : V(R) \to V(S)$ the monoid morphism defined by $V(\varphi)(\langle P_R \rangle) = \langle P \otimes_R S \rangle$, V becomes a functor from the category of associative rings with identity to the category of commutative monoids.

Notice that $V(R)$ describes both right projective R-modules and left projective R-modules, because the contravariant functor $\mathrm{Hom}_R(-, R)$ induces a duality between the full subcategory of $\mathrm{Mod}\, R$ whose objects are all finitely generated projective right R-modules and the full subcategory of R-Mod whose objects are all finitely generated projective left R-modules, so that if we use finitely generated projective left modules instead of finitely generated projective right modules in the construction of $V(R)$, we obtain a monoid canonically isomorphic to $V(R)$.

Two projective modules P_R and Q_R are *stably isomorphic* if $P_R \oplus R_R^n \cong Q_R \oplus R_R^n$ for some integer $n \geqslant 0$. We shall denote by $[P_R]$ the *stable isomorphism class* of P_R, that is, the class of all modules Q_R stably isomorphic to P_R. For a ring R, the *Grothendieck group* of R is the Abelian group

$$K_0(R) = \big\{ [P_R] - [Q_R] \mid P_R, Q_R \text{ finitely generated projective right } R\text{-modules} \big\}.$$

Here two elements $[P_R] - [Q_R]$ and $[P'_R] - [Q'_R]$ are equal if and only if $P_R \oplus Q'_R$ and $P'_R \oplus Q_R$ are stably isomorphic. The addition is defined by $([P_R] - [Q_R]) + ([P'_R] -$

$[Q'_R]) = [P_R \oplus P'_R] - [Q_R \oplus Q'_R]$. The assignment K_0 also turns out to be a functor from the category of associative rings with identity to the category of Abelian groups. The canonical monoid morphism $V(R) \to K_0(R)$ defined by $\langle P_R \rangle \mapsto [P_R]$ yields a natural transformation of functors $V \to K_0$.

For the rest of this section we suppose R is semilocal. In this case, finitely generated projective R-modules cancel from direct sums (Example 2.11(4) and Lemma 4.3(a)), that is, stable isomorphism coincides with isomorphism, i.e. $[P_R] = \langle P_R \rangle$ for every P_R. Thus the morphism $V(R) \to K_0(R)$ is an embedding, so that we may suppose $V(R) \subseteq K_0(R)$ for R semilocal. The canonical projection $\pi : R \to R/J(R)$ induces a monoid embedding $V(\pi) : V(R) \to V(R/J(R))$, and, as we have already remarked, the semisimplicity of $R/J(R)$ implies that $V(R/J(R)) \cong \mathbf{N}^n$, where n is the number of simple $R/J(R)$-modules up to isomorphism. Thus $V(R)$ is isomorphic to a submonoid of $\mathbf{N}^n$. Similarly, the canonical projection $\pi : R \to R/J(R)$ induces an embedding of Abelian groups $K_0(\pi) : K_0(R) \to K_0(R/J(R))$ and $K_0(R/J(R)) \cong \mathbf{Z}^n$, so that $K_0(R)$ is isomorphic to a subgroup of $\mathbf{Z}^n$. There is a commutative diagram

$$\begin{tikzcd} V(R) \arrow[r, "V(\pi)"] \arrow[d] & V(R/J(R)) \cong \arrow[d] & \mathbf{N}^n \arrow[d] \\ K_0(R) \arrow[r, "K_0(\pi)"] & K_0(R/J(R)) \cong & \mathbf{Z}^n \end{tikzcd}$$

in which all the arrows are embeddings. Hence the monoid $V(R)$ is isomorphic to its image in $\mathbf{Z}^n$, and the image of $V(R)$ in $\mathbf{Z}^n$ is the intersection of $\mathbf{N}^n$ with the image of the Abelian group $K_0(R)$ in $\mathbf{Z}^n$. In particular, $V(R)$ is isomorphic to the intersection of $\mathbf{N}^n$ with a subgroup of $\mathbf{Z}^n$ [FH1]. A submonoid M of the additive monoid $\mathbf{N}^n$ is a *full submonoid* if there exists a subgroup G of $\mathbf{Z}^n$ such that $M = G \cap \mathbf{N}^n$. Thus $V(R)$ is isomorphic to a full submonoid of $\mathbf{N}^n$ for every semilocal ring R. In this isomorphism, the element $\langle R_R \rangle$ of $V(R)$ corresponds to an n-tuple $u = (d_1, \dots, d_n)$ of positive integers. Conversely:

THEOREM 6.1. *Let M be a full submonoid of $\mathbf{N}^n$ and let $T : M \to \mathbf{N}^n$ be the embedding. Assume that $u \in M$ is such that $T(u) = (d_1, \dots, d_n)$ with $d_1, \dots, d_n > 0$. Then there exist a semilocal ring R and two monoid isomorphisms $g : M \to V(R)$ and $h : \mathbf{N}^n \to V(R/J(R))$ such that if $\pi : R \to R/J(R)$ denotes the canonical projection, then the diagram of commutative monoids and monoid homomorphisms*

$$\begin{tikzcd} M \arrow[r, "T"] \arrow[d, "g"'] & \mathbf{N}^n \arrow[d, "h"] \\ V(R) \arrow[r, "V(\pi)"] & V(R/J(R)) \end{tikzcd} \tag{1}$$

commutes and $g(u) = \langle R_R \rangle$.

Moreover, R can be

(a) *either a hereditary k-algebra over an arbitrary field k* [FH1, Theorem 6.1],

(b) *or the endomorphism ring of a finitely generated reflexive module M_S over a commutative Noetherian local unique factorization domain S of Krull dimension* 2 [W10, Theorem 4.1],

(c) *or the endomorphism ring of a finite rank torsion-free module over the local ring* $\mathbf{Z}_{(p)}$ *(the ring of integers* $\mathbf{Z}$ *localized at its prime ideal* (p) *for an arbitrary prime* p*)* [Y2, Theorem].

For (b), recall that if M is a module over a commutative ring S and $M^* = \mathrm{Hom}_S(M, S)$ denotes the dual module, then M is *reflexive* if the canonical mapping $M \to M^{**}$ is an isomorphism.

Theorem 6.1 describes all possible direct sum decompositions of Artinian modules, of finitely generated modules over commutative Noetherian semilocal rings, and of finite rank torsion-free $\mathbf{Z}_{(p)}$-modules. For some examples that show how the method expounded in this section allows us to solve the problem of the existence of modules with particular direct sum decompositions whenever we are considering finitely generated modules over commutative Noetherian semilocal rings, or Artinian modules over arbitrary rings, or projective modules over semilocal rings, etc., see the examples at the end of Section 2 of [F4].

THEOREM 6.2 [F3, Theorem 4.4]. *The following conditions are equivalent for a semilocal ring* R:

(a) *The Krull–Schmidt theorem holds for the class* proj-R *of all finitely generated projective right* R*-modules.*

(b) $V(R) \cong \mathbf{N}^t$ *for some* $t \geqslant 1$.

(c) *There exists a direct sum decomposition* $R_R = I_1 \oplus \cdots \oplus I_m$ *such that if* $\{I_1, \ldots, I_t\}$ *is a complete irredundant set of representatives of the isomorphism classes of* $I_1, \ldots, I_m$*, then every finitely generated projective right* R*-module is isomorphic to a direct sum of copies of* $I_1, \ldots, I_t$ *in a unique way.*

Notice that the condition "there exists a direct sum decomposition $R_R = I_1 \oplus \cdots \oplus I_m$ such that every finitely generated projective right R-module is isomorphic to a direct sum of the I_j's" is strictly weaker than condition (c); see [F3].

7. Homogeneous semilocal rings, modules whose endomorphism ring is homogeneous semilocal

In this section we shall present some results that appear in [CF] and [BFRR]. A ring R is a *homogeneous semilocal* ring if $R/J(R)$ is simple Artinian, that is, if $R/J(R)$ is isomorphic to the ring $M_n(D)$ of $n \times n$ matrices with entries in some division ring D for some positive integer n. Equivalently, a ring R is homogeneous semilocal if and only if there exists a finite set $\{I_1, I_2, \ldots, I_n\}$ of maximal right ideals of R such that $J(R) = I_1 \cap I_2 \cap \cdots \cap I_n$ and $R/I_j \cong R/I_k$ for every $j, k = 1, 2, \ldots, n$.

EXAMPLES 7.1.

(1) Every local ring is homogeneous semilocal. More generally, if R is a local ring and t is a non-negative integer, the ring $M_t(R)$ of $t \times t$ matrices with entries in R is a homogeneous semilocal ring.

(2) If a ring R is commutative, then R is homogeneous semilocal if and only if it is local.
(3) Homomorphic images of homogeneous semilocal rings are homogeneous semilocal rings.
(4) If R is a right Noetherian ring and P is a right localizable prime ideal of R, the localization of R with respect to P is a homogeneous semilocal ring.
(5) For further examples of homogeneous semilocal rings, see [CF, §§4 and 5].

A number of properties of local rings extend to homogeneous semilocal rings [CF, BFRR], and it is natural to ask whether the Krull–Schmidt theorem, which holds for direct sums of modules whose endomorphism rings are local, holds also for direct sums of modules whose endomorphism rings are homogeneous semilocal. The next theorem considers this case.

THEOREM 7.2 (Krull–Schmidt theorem for direct sums of modules with homogeneous semilocal endomorphism rings, [BFRR]). *Let $\overline{M}_R = M_1 \oplus \cdots \oplus M_t = N_1 \oplus \cdots \oplus N_s$ be two direct sum decompositions of a module $\overline{M}_R$ over a ring R. If all the R-modules M_i and N_j are indecomposable and all their endomorphism rings* $\mathrm{End}(M_i)$ *and* $\mathrm{End}(N_j)$ *are homogeneous semilocal, then the two direct sum decompositions are isomorphic.*

Notice that in order to get the uniqueness of the decompositions, it does not suffice that in the decompositions $\overline{M}_R = M_1 \oplus \cdots \oplus M_t = N_1 \oplus \cdots \oplus N_s$ only *one* decomposition $M_1 \oplus \cdots \oplus M_t$ is into indecomposable direct summands M_i whose endomorphism rings $\mathrm{End}(M_i)$ are homogeneous semilocal. In fact, for any pair of integers $t > 1$ and $s > 1$, there exists a module $\overline{M}_R$ over a suitable ring R with the property that $\overline{M}_R$ has two non-isomorphic direct sum decompositions into indecomposable direct summands $\overline{M}_R = M_1 \oplus \cdots \oplus M_t \cong N^s$ and such that the endomorphism rings $\mathrm{End}(M_i)$ are homogeneous semilocal rings for every $i = 1, \dots, t$ ([FH2, Lemma 5.5] and [BFRR, Example 3.4]).

We conclude this section with a generalization of Theorem 7.2 to arbitrary, possibly infinite, direct sums. For other generalizations of the theorem to infinite direct sums, for instance to indecomposable Artinian direct summands, see [FH3].

THEOREM 7.3 [FH3, Corollary 3.5]. *Let $\overline{M}_R = \bigoplus_{i \in I} M_i = \bigoplus_{j \in J} N_j$ be two direct sum decompositions of a module $\overline{M}_R$ into indecomposable direct summands such that all the endomorphism rings* $\mathrm{End}(M_i)$ *and* $\mathrm{End}(N_j)$ *are homogeneous semilocal. Suppose that each module M_i and each module N_j is either small, or quasi-small and countably generated. Then the two direct sum decompositions of $\overline{M}_R$ are isomorphic.*

8. Module-finite algebras in Krull dimension 1

In this section we shall present some results obtained recently by Levy and Odenthal [LO1]. Recall that a ring is *semiprime* if it has no non-zero nilpotent ideal. In Sections 8 and 9, the symbol R will denote a semiprime module-finite k-algebra, where k is a commutative Noetherian ring of Krull dimension 1. Without loss of generality, we may assume that k is contained in the center $Z(R)$, and hence k has no non-zero nilpotent elements, because any such element would generate a nilpotent ideal of R. Thus R can be any of the orders studied in integral representation theory. Without loss of generality, we shall assume that R is indecomposable as a ring; and we shall assume that R is not Artinian, since the Krull–Schmidt theorem holds for finitely generated modules over Artinian rings (Theorem 2.1).

Let Q denote the finite set of minimal prime ideals of k. Since k has no nilpotent elements, $k - \bigcup Q$ is the set of regular elements of k, that is, elements that are not zero-divisors. Let k_Q and R_Q denote the localizations of k and R respectively formed by inverting all of the elements of $k - \bigcup Q$. The ring k_Q is Artinian (because all prime ideals are maximal), has no nilpotent element, and therefore it is a direct product of fields.

Since k_Q is an Artinian ring and R_Q is module-finite over k_Q, R_Q is also an Artinian ring. Moreover, since R is semiprime, the localization R_Q is a semisimple Artinian ring. In fact, R is a *k-order in the semisimple Artinian ring R_Q* in the sense that the natural map $R \to R_Q$ is one-to-one. (The simple proof of this last fact uses the hypotheses that R is indecomposable and not Artinian. See the introduction to [LO1] for details.)

In this situation, Levy and Odenthal define a *normalization* R' of R (in R_Q) to be a maximal element of the family of subrings of R_Q that contain R and are integral over k. In the classical situation, where R is contained in a maximal k-order in R_Q, the normalization R' is easily seen to be such a maximal order. And when R is commutative, R' becomes a normalization of R in the sense of commutative ring theory; that is, the integral closure in the total quotient ring. In all cases, R' turns out to be a direct product of classical maximal orders over Dedekind domains in simple Artinian rings, and thus provides a link between Levy and Odenthal's theory and classical integral representation theory [LO2, §4].

Two finitely generated R-modules A, B *are in the same genus* if $A_M \cong B_M$ (as R_M-modules) for every maximal ideal M of k. This property is independent of the particular ring k over which R is a module-finite algebra.

We shall denote by FG-R the class of all finitely generated right R-modules, and by FGTF-R the class of all finitely generated torsion-free right R-modules, that is, the A_R for which the natural map $A \to A \otimes_k Q$ is injective. In the language of integral representation theory (when R is a classical k-order) every finitely generated torsion-free right R-module is an R-lattice. The Krull–Schmidt theorem holds for FG-R and FGTF-R only in the presence of extreme restrictions on the structure of R, as the next two results show.

THEOREM 8.1 [LO1, Theorem 1.1]. *The Krull–Schmidt theorem holds for* FG-R *if and only if the following three conditions hold.*

(a) *Every genus of finitely generated right R-modules consists of a single isomorphism class.*
(b) *Either $R = R'$ or exactly one maximal ideal N of $Z(R)$ is* singular *with respect to R; that is, $R_N \neq R'_N$ (localizations of R, R').*
(c) *If this singular N occurs, then primitive idempotents of R'_N remain primitive modulo $J(R'_N)$.*

Condition (c) is equivalent to the statement that the Krull–Schmidt theorem holds for FG-R_N, as one can see by applying the theorem to the k_N-algebra R_N, and noting that $Z(R_N) = Z(R)_N$.

THEOREM 8.2 [LO1, Theorem 1.3]. *The Krull–Schmidt theorem holds for* FGTF-R *if and only if the following conditions are satisfied.*

(a) *Every genus of finitely generated torsion-free R-modules consists of a single isomorphism class.*
(b) *Either $R = R'$ or exactly one maximal ideal N of $Z(R)$ is singular with respect to R (i.e. $R_N \neq R'_N$).*
(c) *If this singular N occurs, then either condition* (c1) *or* (c2) *holds*:
  (c1) *Primitive idempotents of R'_N remain primitive modulo $J(R'_N)$; that is, the Krull–Schmidt theorem holds for* FG-R_N*,*
  (c2) *R_N is a member of a very small class of exceptional rings.* [*See Remark* (2) *below.*]

REMARKS.

(1) [on condition (a)] It is possible for every genus of torsion-free R-modules to consist of a single isomorphism class without the same being true for genera of modules that are not torsion-free. An example is given in [KL]. In this example the Krull–Schmidt theorem holds for FGTF-R and fails for FG-R; in fact, direct-sum cancellation fails for modules that are not torsion-free.
(2) [on condition (c2)] It is interesting that these exceptional rings are "classical" in the sense that R'_N is module-finite over R_N. Moreover, R'_N is a prime ring, and has at most two maximal 2-sided ideals. See [LO1] for full details.

We remark that, when k is semilocal, the notions of genus and isomorphism class coincide. This has a number of interesting consequences. The most obvious of these is:

THEOREM 8.3. *Suppose that k is semilocal. Then the Krull–Schmidt theorem holds for the class of finitely generated right R-modules if and only if it holds for the class of finitely generated left R-modules. The same is true for the classes of finitely generated torsion-free right and left R-modules.*

It seems to be unknown whether the "semilocal" hypothesis can be dropped, because there seems to be no criterion that decides whether every genus consists of a single isomorphism class.

An often-asked question is whether indecomposable modules have local endomorphism rings whenever the Krull–Schmidt theorem holds for FG-R. This is never asked in print, of course, since the ring of integers gives an immediate counterexample; and the 6-localization of this ring is even a semilocal counterexample. These examples seem to have discouraged looking at this question seriously. In fact, a more subtle version of the question has a positive answer, after one ignores the trouble caused by maximal orders:

COROLLARY 8.4 [LO1, Corollary 2.11]. *Suppose that the Krull–Schmidt theorem holds for* FG-R*, and $R \neq R'$. Let N be the singular maximal ideal of $Z(R)$ with respect to R.*

Then the endomorphism ring of the N-localization of every indecomposable R-module is a local ring.

The analogous result for the class FGTF-R fails, because of the exceptional rings in Theorem 8.2. See [LO1, Remark 2.12(iv)].

A well-known theorem of A. Heller (implicit in the proof of [H, 2.5]) states that if k is a discrete valuation ring of characteristic zero and R is contained in a maximal order, then condition (a) of the following corollary implies that the Krull–Schmidt theorem holds for FG-R. The characteristic 0 hypothesis was removed in [CR, (30.18)], but only for torsion-free R-modules. The following corollary removes these special hypotheses, and condition (b) provides a simple explanation of condition (a) in terms of the structure of the ring R'.

COROLLARY 8.5 [LO1, Corollary 2.5]. *The following statements are equivalent, if k is semilocal.*

(a) *The $J(k)$-induced completion $\hat{k} \otimes_k V$ of every simple R_Q-module V is an indecomposable $\widehat{R}_Q$-module.*

(b) *Primitive idempotents of R' remain primitive modulo $J(R')$.*

When the conditions hold, the Krull–Schmidt theorem holds for FG-R *and $Z(R)$ is a direct product of local rings.*

Again the analogous result for the class FGTF-R fails, because of the exceptional rings [LO1, Remarks 2.12].

9. Integral group rings

It may be surprising that the Krull–Schmidt theorem holds for FGTF-$\mathbf{Z}G$ (and hence for FG-$\mathbf{Z}G$) for only finitely many integral group rings $\mathbf{Z}G$ of finite groups G. In more detail, Hindman, Klingler, and Odenthal [HKO] – building upon earlier work of Swan, Wiegand, and others – prove:

THEOREM 9.1. *If G is not dihedral of order* 16 (*the only undetermined case*), *then the Krull–Schmidt theorem holds for* FGTF-$\mathbf{Z}G$ *exactly when G is one of the following groups*:

(a) *Cyclic of prime order $p \leqslant 19$.*

(b) *Cyclic of order* 1, 4, 8, *or* 9.

(c) *Klein* 4-*group.*

(d) *Dihedral of order* 8.

This is the main result of [HKO, 1.6], except that their condition (a) can be simplified to our present statement, in view of the following two facts. (1) Let $m \not\equiv 2 \pmod 4$, and let ω_m be a primitive m-th root of unity. Then the class number of $\mathbf{Q}(\omega_m)$ is 1 if and only if m is one of the following: 1, 3, 4, 5, 7, 8, 9, 11, 12, 13, 15, 16, 17, 19, 20, 21, 24, 25, 27, 28, 32, 33, 35, 36, 40, 44, 45, 48, 60, 84. See, e.g. [W8, Theorem 11.1] (only the case $m = p$ is needed); and (2) the Krull–Schmidt theorem holds for FGTF-$\mathbf{Z}G$ when G is cyclic of order 2. See the last paragraph of [HKO, p. 3247] for this.

It seems to be unknown for which of these groups G the Krull–Schmidt theorem holds for FG-$\mathbf{Z}G$. The critical property is whether every genus of finitely generated R-modules – not just the torsion-free ones – consists of a single isomorphism class. (See Theorems 8.1 and 8.2.)

The Krull–Schmidt theorem holds for the finitely generated modules over the rings of Theorem 9.1(a). This follows from the following two facts: (1) $\mathrm{Pic}(R)$ is trivial (has order 1) since the Krull–Schmidt theorem holds for FGTF-R, because Theorem 8.2 states that every genus of torsion-free modules consists of a single isomorphism class, and (2) these rings are, by [L5], special cases of the Dedekind-like rings studied in [L4]; and for the Dedekind-like rings studied in [L4], every genus class group is a homomorphic image of $\mathrm{Pic}(R)$.

Appendix A. The Krull–Schmidt theorem for additive categories

In this appendix we mention a generalization of the Krull–Schmidt theorem to additive categories. An additive category $\mathcal{A}$ with kernels is said to satisfy a *weak Grothendieck condition* if for every index set I and every non-zero subobject A of $\coprod_{i\in I} C_i$, there exists a finite subset F of I such that $A \cap \coprod_{i\in F} C_i \neq 0$. A *small* object is an object S such that every morphism into a coproduct $f : S \to \coprod_{i\in I} C_i$ factors though $\coprod_{i\in F} C_i$ for some finite subset F of I. An object M is *finitely approximable* if for any object L and morphism $f : L \to M$, f is an isomorphism if and only if (1) f is monic, and (2) for any small object S and any morphism $g : S \to M$ there is a morphism $h : S \to L$ such that $g = fh$.

The Krull–Schmidt–Azumaya theorem generalizes to additive categories as follows.

THEOREM A.1 [B2,WW1,AHR]. *Let $\mathcal{A}$ be an additive category in which idempotent endomorphisms have kernels. Let C_i, $i \in I$, be objects of $\mathcal{A}$ with local endomorphism rings. If I is infinite, suppose that $\mathcal{A}$ has kernels and infinite coproducts and that either* (1) *the C_i are finitely approximable or* (2) *the category $\mathcal{A}$ satisfies a weak Grothendieck condition. Then any two coproduct decompositions of $\coprod_{i\in I} C_i$ into indecomposables are isomorphic.*

The finite case of this theorem is due to Bass [B2], and the infinite case to Walker and Warfield [WW1] (see [AHR]). For the details, see [WW1] and [AHR]. For an analogue of Theorem A.1 in the case where the endomorphism rings of the C_i are principal ideal domains, see [AHR].

In this setting we must mention *spectral* categories, that is, Abelian categories in which every object is injective. In these categories the theory of decompositions is particularly good [GO1,GB,P2,R5].

Appendix B. Goldie dimension of modular lattices

In this chapter we have heavily used the notion of dual Goldie dimension of a module. Since the knowledge of this notion is not so widespread among algebraists, we recall its definition in this appendix. The correct framework for the notions of Goldie dimension and dual Goldie dimension is that of modular lattices with a smallest element 0 and a greatest element 1. We shall briefly review how it is possible to define the Goldie dimension in this setting.

Let L be a modular lattice with a smallest element 0 and a greatest element 1. A finite subset $\{a_i \mid i \in I\}$ of $L \setminus \{0\}$ is said to be *join-independent* if $a_i \wedge (\bigvee_{j \neq i} a_j) = 0$ for every $i \in I$. An arbitrary subset A of $L \setminus \{0\}$ is *join-independent* if all finite subsets of A are join-independent. An element a of L is *uniform* if $a \neq 0$ and for every $x, y \in L$, $x \leqslant a$, $y \leqslant a$ and $x \wedge y = 0$ imply $x = 0$ or $y = 0$. An element $a \in L$ is *essential* in L if $a \wedge x \neq 0$ for every non-zero element $x \in L$. If a, b are elements of L, $a \leqslant b$ and a is essential in the interval $[0, b] = \{x \in L \mid x \leqslant b\}$, then a is said to be *essential* in b.

THEOREM B.1 [GP]. *Let L be a modular lattice with 0 and 1. The following conditions are equivalent*:

(a) *L does not contain infinite join-independent subsets.*

(b) *There exists a finite join-independent subset $\{a_1, a_2, \ldots, a_n\}$ of L with $a_1 \vee a_2 \vee \cdots \vee a_n$ essential in L and $a_1, a_2, \ldots, a_n$ uniform elements.*

(c) *The cardinality of all join-independent subsets of L is $\leqslant m$ for some non-negative integer m.*

(d) *For every ascending chain $a_0 \leqslant a_1 \leqslant a_2 \leqslant \cdots$ of elements of L there exists an index $i \geqslant 0$ with a_i essential in a_j for every $j \geqslant i$.*

If these equivalent conditions hold and $\{a_1, a_2, \ldots, a_n\}$ is a finite join-independent subset of L with $a_1 \vee a_2 \vee \cdots \vee a_n$ essential in L and $a_1, a_2, \ldots, a_n$ uniform elements, then any other join-independent subset of L has cardinality $\leqslant n$.

For a modular lattice L satisfying the equivalent conditions of the theorem, the integer n of the statement, that is, the greatest element of the set of the cardinalities of all join-independent subsets of $L \setminus \{0\}$, is called the *Goldie dimension* of the lattice L, denoted $\dim L$. If L does not satisfy the equivalent conditions of Theorem B.1, that is, if L contains an infinite join-independent subset, the lattice L is said to have *infinite Goldie dimension*.

The *Goldie dimension* $\dim(M_R)$ of a module M_R is the Goldie dimension of its lattice $\mathcal{L}(M_R)$ of submodules. As the dual (= opposite) L^{op} of any modular lattice L is a modular lattice, the *dual Goldie dimension* $\operatorname{codim}(L) = \dim(L^{\mathrm{op}})$ of a modular lattice L is always defined [GP] (it is possibly infinite). In particular, the *dual Goldie dimension* $\operatorname{codim}(M_R)$ of a module M_R is defined as the Goldie dimension of the dual lattice $\mathcal{L}(M_R)^{\mathrm{op}}$. Thus a module M has finite dual Goldie dimension $\operatorname{codim}(M_R) = n$ if and only if there is a coindependent set $\{N_1, N_2, \ldots, N_n\}$ of proper submodules of M with $N_1 \cap N_2 \cap \cdots \cap N_n$ superfluous in M and M/N_i couniform for $i = 1, 2, \ldots, n$.

An injective module has finite Goldie dimension n if and only if it is the direct sum of n indecomposable modules. More generally, let $E(M_R)$ denote the injective envelope of a module M_R. If $E(M_R)$ is the direct sum of n indecomposable modules $E_1, E_2, \ldots, E_n$, the number n depends only on M_R by Theorem 2.2 and Proposition 2.7 and coincides with the Goldie dimension of M_R. Otherwise, that is, if $E(M_R)$ is not a direct sum of finitely many indecomposable modules, then M_R has infinite Goldie dimension.

References

[AF1] F.W. Anderson and K.R. Fuller, *Modules with decompositions that complement direct summands*, J. Algebra **22** (1972), 241–253.

[AF2] F.W. Anderson and K.R. Fuller, *Rings and Categories of Modules*, 2nd edn, Springer-Verlag, New York (1992).

[A1] D. Arnold, *Finite Rank Torsion-free Abelian Groups and Rings*, Lecture Notes in Math., Vol. 931, Springer-Verlag, Berlin–Heidelberg–New York (1982).

[A2] D. Arnold, *Butler groups, representations of finite posets, and modules over multiple pullback rings*, Methods in Module Theory (Colorado Springs, CO, 1991), G. Abrams, J. Haefner and K.M. Rangaswamy, eds, Lecture Notes in Pure Appl. Math., Vol. 140, Dekker, New York (1993), 1–18.

[AD] D. Arnold and M. Dugas, *Indecomposable modules over Nagata valuation domains*, Proc. Amer. Math. Soc. **122** (1994), 689–696.

[AHR] D. Arnold, R. Hunter and F. Richman, *Global Azumaya theorems in additive categories*, J. Pure Appl. Algebra **16** (1980), 223-242.

[A3] G. Azumaya, *Corrections and supplementaries to my paper concerning Krull–Remak–Schmidt's theorem*, Nagoya Math. J. **1** (1950), 117–124.

[A4] G. Azumaya, *On maximally central algebras*, Nagoya Math. J. **2** (1951), 119–150.

[BFRR] F. Barioli, A. Facchini, F. Raggi and J. Ríos, *Krull–Schmidt theorem and homogeneous semilocal rings*, Comm. Algebra **29** (2001), 1649–1658.

[B1] H. Bass, *K-theory and stable algebra*, Publ. Math. I.H.E.S. **22** (1964), 5–60.

[B2] H. Bass, *Algebraic K-theory*, Benjamin, New York (1968).

[BBT] C. Bessenrodt, H.H. Brungs and G. Törner, *Right chain rings, Part 1*, Schriftenreihe des Fachbereichs Math., Vol. 181, Universität Duisburg (1990).

[B3] L. Bican, *Weak Krull–Schmidt theorem,* Comment. Math. Univ. Carolin. **39** (1998), 633–643.

[B4] W. Brandal, *Commutative rings whose finitely generated modules decompose*, Lecture Notes in Math., Vol. 723, Springer-Verlag, Berlin (1979).

[B5] G. Brookfield, *A Krull–Schmidt theorem for Noetherian modules*, J. Algebra **251** (2002), 70–79.

[CD] R. Camps and W. Dicks, *On semilocal rings*, Israel J. Math. **81** (1993), 203–211.

[C1] P.M. Cohn, *Universal Algebra*, Math. Appl., Vol. 6, D. Reidel Publishing Co., Dordrecht–Boston, MA (1981).

[CF] R. Corisello and A. Facchini, *Homogeneous semilocal rings*, Comm. Algebra **29** (2001), 1807–1819.

[C2] A.L.S. Corner, *Every countable reduced torsion-free ring is an endomorphism ring*, Proc. London Math. Soc. (3) **13** (1963), 687–710.

[C3] A.L.S. Corner, *On a conjecture of Pierce concerning direct decompositions of Abelian groups*, Proc. Colloq. Abelian Groups, Tihany (Hungary), 1963, Akadémiai Kiadó, Budapest (1964), 43–48.

[C4] A.L.S. Corner, *Endomorphism rings of torsion-free Abelian groups*, Proc. Internat. Conf. Theory of Groups, Austral. Nat. Univ. Canberra, 1965 (1967), 59–69.

[C5] F. Couchot, *Anneaux locaux à type de module borné*, Ark. Mat. **34** (1996), 65–72.

[C6] F. Couchot, *Commutative local rings of bounded module type*, Comm. Algebra **29** (2001), 1347–1355.

[CJ] P. Crawley and B. Jónsson, *Refinements for infinite direct decompositions of algebraic systems,* Pacific J. Math. **14** (1964), 797–855.

[CR] C.W. Curtis and I. Reiner, *Methods of Representation Theory, with Applications to Finite Groups and Orders*, Vol. I, Wiley-Interscience, New York (1981).

[D] N. Dubrovin, *The rational closure of group rings in left ordered groups*, Schriftenreihe des Fachbereichs Math., Vol. 254 , Universität Duisburg (1994).

[DF1] N.V. Dung and A. Facchini, *Weak Krull–Schmidt for infinite direct sums of uniserial modules,* J. Algebra **193** (1997), 102–121.

[DF2] N.V. Dung and A. Facchini, *Direct summands of serial modules,* J. Pure Appl. Algebra **133** (1998), 93–106.

[E] E.G. Evans, Jr., *Krull–Schmidt and cancellation over local rings,* Pacific J. Math. **46** (1973), 115–121.

[F1] A. Facchini, *Krull–Schmidt fails for serial modules*, Trans. Amer. Math. Soc. **348** (1996), 4561–4575.

[F2] A. Facchini, *Module Theory. Endomorphism Rings and Direct Sum Decompositions in Some Classes of Modules*, Progress in Math., Vol. 167, Birkhäuser, Basel (1998).

[F3] A. Facchini, *Failure of the Krull–Schmidt theorem for Artinian modules and serial modules*, Infinite Length Modules, H. Krause and C.M. Ringel, eds, Trends in Math., Birkhäuser, Basel (2000).

[F4] A. Facchini, *Krull–Schmidt theorem and semilocal endomorphism rings*, Ring Theory and Algebraic Geometry, A. Granja, J.A. Hermida Alonso and A. Verschoren, eds, Marcel Dekker, New York (2001), 187–201.

[FH1] A. Facchini and D. Herbera, *K_0 of a semilocal ring*, J. Algebra **225** (2000), 47–69.

[FH2] A. Facchini and D. Herbera, *Projective modules over semilocal rings*, Algebra and Its Applications, D.V. Huynh, S.K. Jain and S.R. López-Permouth, eds, Amer. Math. Soc., Providence, RI (2000), 181–198.

[FH3] A. Facchini and D. Herbera, *Two results on modules whose endomorphism ring is semilocal*, Preprint (2000).

[FHLV] A. Facchini, D. Herbera, L.S. Levy and P. Vámos, *Krull–Schmidt fails for Artinian modules*, Proc. Amer. Math. Soc. **123** (1995), 3587–3592.

[F5] J.W. Fisher, *Nil subrings of endomorphism rings of modules,* Proc. Amer. Math. Soc. **34** (1972), 75–78.

[FS1] L. Fuchs and L. Salce, *Uniserial modules over valuation rings*, J. Algebra **85** (1983), 14–31.

[FV1] L. Fuchs and P. Vámos, *The Jordan–Zassenhaus theorem and direct decompositions*, J. Algebra **230** (2000), 730–748.

[FV2] L. Fuchs and G. Viljoen, *Local quasi-isomorphisms of torsion free Abelian groups*, Math. Scand. **33** (1973), 205–212.

[FS2] K.R. Fuller and W.A. Shutters, *Projective modules over non-commutative semilocal rings*, Tôhoku Math. J. **27** (1975), 303–311.

[GO1] P. Gabriel and U. Oberst, *Spektralkategorien und reguläre Ringe im von-Neumannschen Sinn*, Math. Z. **92** (1966), 389–395.

[G1] D.T. Gill, *Almost maximal valuation rings*, J. London Math. Soc. **4** (1971), 140–146.

[GO2] P. Goeters and B. Olberding, *Unique decompositions into ideals for Noetherian domains*, J. Pure Appl. Algebra **165** (2001), 169–182.

[GO3] P. Goeters and B. Olberding, *Krull–Schmidt for ideals and modules over integral domains*, to appear in Proc. of the Second Honolulu Conference on Abelian Groups and Modules (2002).

[GM] B. Goldsmith and W. May, *The Krull–Schmidt problem for modules over valuation domains*, J. Pure Appl. Algebra **140** (1999), 57–63.

[GB] K.R. Goodearl and A.K. Boyle, *Dimension theory for nonsingular injective modules*, Mem. Amer. Math. Soc. **177** (1976).

[G2] A. Grothendieck, *Sur quelques points d'Algèbre homologique*, Tôhoku Math. J. **9** (1957), 119–221.

[GP] P. Grzeszczuk and E.R. Puczyłowski, *On Goldie and dual Goldie dimensions*, J. Pure Appl. Algebra **31** (1984), 47–54.

[H] A. Heller, *On group representations over a valuation ring*, Proc. Nat. Acad. Sci. USA **47** (1961), 1194–1197.

[HS] D. Herbera and A. Shamsuddin, *Modules with semi-local endomorphism ring*, Proc. Amer. Math. Soc. **123** (1995), 3593–3600.

[HKO] P. Hindman, L. Klingler and C.J. Odenthal, *On the Krull–Schmidt–Azumaya theorem for integral group rings*, Comm. Algebra **26** (1998), 3743–3758.

[J1] B. Jónsson, *On unique factorization problem for torsion-free Abelian groups*, Bull. Amer. Math. Soc. **51** (1945), 364.

[J2] B. Jónsson, *On direct decompositions of torsion-free Abelian groups*, Math. Scand. **5** (1957), 230–235.

[J3] B. Jónsson, *On direct decomposition of torsion-free Abelian groups*, Math. Scand. **7** (1959), 361–371.

[K1] I. Kaplansky, *Maximal fields with valuations,* Duke Math. J. **9** (1942), 303–321.

[K2] I. Kaplansky, *Modules over Dedekind rings and valuation rings,* Trans. Amer. Math. Soc. **72** (1952), 327–340.

[K3] I. Kaplansky, *Elementary divisors and modules,* Trans. Amer. Math. Soc. **66** (1949), 464–491.

[KL] L. Klingler and L.S. Levy, *Direct-sum cancellation: modules versus lattices*, Comm. Algebra **18** (1990), 1857–1868.

[K4] W. Krull, *Über verallgemeinerte endliche Abelsche Gruppen,* Math. Z. **23** (1925), 161–196.

[K5] W. Krull, *Allgemeine Bewertungstheorie,* J. Reine Angew. Math. **167** (1932), 160–196.

[K6] W. Krull, *Matrizen, Moduln und verallgemeinerte Abelsche Gruppen im Bereich der ganzen algebraischen Zahlen,* Beiträge zur Algebra, Vol. 19, Heidelberger Akademie der Wissenschaften **2** (1932), 13–38.

[L1] E.L. Lady, *Summands of finite rank torsion-free Abelian groups*, J. Algebra **32** (1974), 51–52.

[L2] T.-Y. Lam, *Modules with isomorphic multiples and rings with isomorphic matrix rings. A survey*, Monographies de L'Enseignement Mathématique, Vol. 35, Enseign. Math. (1999).

[L3] L.S. Levy, *Krull–Schmidt uniqueness fails dramatically over subrings of* $\mathbf{Z} \oplus \cdots \oplus \mathbf{Z}$, Rocky Mountain J. Math. **13** (1983), 659–678.

[L4] L.S. Levy, *Modules over Dedekind-like rings*, J. Algebra **93** (1985), 1–116.

[L5] L.S. Levy, $\mathbf{Z}G_n$*-modules,* G_n *cyclic of square-free order* n, J. Algebra **93** (1985), 354–375.

[LO1] L.S. Levy and C.J. Odenthal, *Krull–Schmidt theorems in dimension 1*, Trans. Amer. Math. Soc. **348** (1996), 3391–3455.

[LO2] L.S. Levy and C.J. Odenthal, *Package deal theorems and splitting orders, in dimension 1*, Trans. Amer. Math. Soc. **348** (1996), 3457–3503.

[M1] E. Matlis, *Injective modules over Noetherian rings*, Pacific J. Math. **8** (1958), 511–528.

[M2] E. Matlis, *Torsion-free Modules*, University Press, Chicago (1972).

[MZ] W. May and P. Zanardo, *Non-Henselian valuation domains and the Krull–Schmidt property for torsion-free modules*, Forum Math. **10** (1998), 249–257.

[MW] B. Midgarten and S. Wiegand, *Commutative rings of bounded module type*, Comm. Algebra **9** (1981), 1001–1025.

[O] B.L. Osofsky, *A remark on the Krull–Schmidt–Azumaya theorem*, Canad. Math. Bull. **13** (1970), 501–505.

[P1] Z. Papp, *On algebraically closed modules*, Publ. Math. Debrecen **6** (1959), 311–327.

[PY] K.I. Pimenov and A.V. Yakovlev, *Artinian modules over a matrix ring*, Infinite Length Modules, H. Krause and C.M. Ringel, eds, Trends in Math., Birkhäuser, Basel (2000).

[P2] N. Popescu, *Abelian Categories with Applications to Rings and Modules*, Academic Press, London–New York, 1973.

[P3] G. Puninski, *Some model theory over a nearly simple uniserial domain and decompositions of serial modules*, J. Pure Appl. Algebra **163** (2001), 319–337.

[P4] G. Puninski, *Some model theory over an exceptional uniserial ring and decompositions of serial modules*, J. London Math. Soc. (2) **64** (2001), 311–326.

[R1] J.D. Reid, *A note on torsion-free Abelian groups of infinite rank*, Proc. Amer. Math. Soc. **13** (1962), 222–225.

[R2] R. Remak, *Über die Zerlegung der endlichen Gruppen in direkte unzerlegbare Faktoren*, J. Reine Angew. Math. **139** (1911), 293–308.

[R3] C.M. Ringel, *Krull–Remak–Schmidt fails for Artinian modules over local rings*, Algebras and Representation Theory **4** (2001), 77–86.

[R4] D.J.S. Robinson, *A Course in the Theory of Groups*, Graduate Texts in Math., Vol. 80, Springer-Verlag, New York (1982).

[R5] J.-E. Roos, *Locally distributive spectral categories and strongly regular rings*, Reports of the Midwest Category Seminar, Lecture Notes in Math., Vol. 47, Springer-Verlag, Berlin (1967), 156–181.

[R6] J.J. Rotman, *An Introduction to the Theory of Groups*, Graduate Texts in Math., Vol. 148, Springer-Verlag, New York (1995).

[SZ] L. Salce and P. Zanardo, *On two-generated modules over valuation domains*, Arch. Math. (Basel) **46** (1986), 408–418.

[SV1] B. Sarath and K. Varadarajan, *Dual Goldie dimension, II,* Comm. Algebra **7** (1979), 1885–1899.

[S1] O. Schmidt, *Über die Zerlegung endlicher Gruppen in direkte unzerlegbare Faktoren,* Universitetskije Izvestija, Kiew (1912).

[S2] O. Schmidt, *Über unendliche Gruppen mit endlicher Kette,* Math. Z. **29** (1928), 34–41.

[SV2] D.W. Sharpe and P. Vámos, *Injective Modules*, Cambridge University Press, Cambridge (1972).

[SW] T. Shores and R. Wiegand, *Rings whose finitely generated modules are direct sums of cyclics,* J. Algebra **32** (1974), 152–172.

[S3] M.F. Siddoway, *On endomorphism rings of modules over Henselian rings*, Comm. Algebra **18** (1990), 1323–1335.

[V1] P. Vámos, *The decomposition of finitely generated modules and fractionally self-injective rings*, J. London Math. Soc. (2) **16** (1977), 209–220.

[V2] P. Vámos, *Sheaf-theoretical methods in the solution of Kaplansky's problem*, Applications of Sheaves, M.P. Fourman, C.J. Mulvey and D.S. Scott, eds, Proc. Res. Sympos. Appl. Sheaf Theory to Logic, Algebra and Anal., Univ. Durham, Durham, 1977, Lecture Notes in Math., Vol. 753, Springer-Verlag, Berlin (1979), 732–738.

[V3] P. Vámos, *Decomposition problems for modules over valuation domains*, J. London Math. Soc. (2) **41** (1990), 10–26.

[V4] P. Vámos, *The holy grail of algebra: seeking complete sets of invariants*, Abelian Groups and Modules, A. Facchini and C. Menini, eds, Math. and Its Appl., Vol. 343, Kluwer, Dordrecht (1995), 475–483.

[V5] K. Varadarajan, *Dual Goldie dimension*, Comm. Algebra **7** (1979), 565–610.

[WW1] C.L. Walker and R.B. Warfield, Jr., *Unique decomposition and isomorphic refinement theorems in additive categories,* J. Pure Appl. Algebra **7** (1976), 347–359.

[W1] E.A. Walker, *Quotient categories and quasi-isomorphisms of Abelian groups*, Proc. Colloq. Abelian Groups, Tihany (Hungary), 1963, Akadémiai Kiadó, Budapest (1964), 147–162.

[W2] R.B. Warfield, Jr., *Purity and algebraic compactness for modules,* Pacific J. Math. **28** (1969), 699–719.

[W3] R.B. Warfield, Jr., *A Krull–Schmidt theorem for infinite sums of modules*, Proc. Amer. Math. Soc. **22** (1969), 460–465.

[W4] R.B. Warfield, Jr., *Decomposability of finitely presented modules*, Proc. Amer. Math. Soc. **25** (1970), 167–172.

[W5] R.B. Warfield, Jr., *Serial rings and finitely presented modules*, J. Algebra **37** (1975), 187–222.

[W6] R.B. Warfield, Jr., *Large modules over Artinian rings*, Representation Theory of Algebras (Proc. Conf., Temple Univ., Philadelphia, PA, 1976), Lecture Notes in Pure Appl. Math., Vol. 37, Dekker, New York (1978), 451–463.

[W7] R.B. Warfield, Jr., *Cancellation of modules and groups and stable range of endomorphism rings*, Pacific J. Math. **91** (1980), 457–485.

[W8] L.C. Washington, *Introduction to Cyclotomic Fields*, 2nd edn, Graduate Texts in Math., Vol. 83, Springer-Verlag, New York (1997).

[W9] J.H.M. Wedderburn, *On the direct product in the theory of finite groups*, Ann. Math. **10** (1909), 173–176.

[W10] R. Wiegand, *Direct-sum decompositions over local rings*, J. Algebra **240** (2001), 83–97.

[WW2] R. Wiegand and S. Wiegand, *Commutative rings whose finitely generated modules are direct sums of cyclics*, Abelian Group Theory, D. Arnold, R. Hunter and E. Walker, eds, Proc. Second New Mexico State Univ. Conf., Las Cruces, NM, 1976, Lecture Notes in Math., Vol. 616, Springer-Verlag, Berlin (1977), 406–423.

[Y1] A.V. Yakovlev, *On direct sum decompositions of Artinian modules*, Algebra i Analiz **10** (1998), 229–238.

[Y2] A.V. Yakovlev, *On direct decompositions of p-adic groups*, Algebra i Analiz **12** (2000), 217–223.

[Z1] P. Zanardo, *Quasi-isomorphisms of finitely generated modules over valuation domains*, Ann. Mat. Pura Appl. (4) **151** (1988), 109–123.

[Z2] P. Zanardo, *Valuation domains of bounded module type*, Arch. Math. (Basel) **53** (1989), 11–19.

[Z3] P. Zanardo, *The Krull–Schmidt property for finitely generated modules over valuation domains*, J. Algebra **237** (2001), 533–561.

[Z4] B. Zimmermann-Huisgen, *Rings whose right modules are direct sums of indecomposable modules,* Proc. Amer. Math. Soc. **77** (1979), 191–197.

This Page Intentionally Left Blank

Coherent Rings and Annihilator Conditions in Matrix and Polynomial Rings

Carl Faith[1]

Department of Mathematics, Rutgers University, 110 Frelinghuysen Rd., Piscataway, NJ 08554-8019, USA
E-mail: carlfaith@aol.com

In memory of a loving and courageous friend, Ahmad Shamsuddin (1951–2001); and to Nisrine, his courageous and loving widow—"Es hat so kurz gedauert."

Contents

[1]Author's current mailing address: 199 Longview Dr., Princeton, NJ 08540, USA.

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

Abstract

This is an exposition of various topics, indicated by the title, and related concepts. Some of this is well known, and some not, e.g., papers published in 1999 and 2000. Indications of proofs are given in some instances, but in many others, e.g., those of Kerr, Camillo–Guralnick, Cedó–Herbera, Roitman, and Antoine–Cedó on the ascending chain conditions for annihilators over polynomial rings, and of Carson, Camillo, Chase, Fieldhouse, Harris, McCarthy, Pillay, Soublin and Vasconcelos, among others, on the question of when the polynomial ring is coherent, proofs are omitted. In short, this is a leisurely stroll through the park with a good view of the architecture of coherency, and associated problems.

1. Introduction

An annihilator right ideal of a ring R is the right annihilator $S^{\perp}$ of a subset S of R. Kerr [75] gave an example of a commutative ring R with acc on right annihilators, denoted $acc\perp$, such that the 2×2 matrix ring R_2 does not have $acc\perp$, but the polynomial ring $R[X]$ does. In this example, R is a *Goldie ring*, i.e., satisfies both $acc\perp$ and the acc on direct summands ($= acc\oplus$).

REMARK 1.0. $acc\oplus$ is equivalent to the requirement that R has no infinite set of independent right ideals, i.e., R has *finite Goldie dimension* as a right R-module. Cf. [40], 3.13s and 16.9B, or [79].

R is a *Kerr ring* if $R[X]$ is an $acc\perp$ ring. In [76] Kerr gave an example of a commutative Goldie ring of characteristic 2 whose polynomial ring is not Goldie. Since $R[X]$ preserves $acc\oplus$ by a theorem of Shock [95], the substance of Kerr's theorem is that $R[X]$ does not preserve $acc\perp$. However:

THEOREM 1.1 (Camillo and Guralnick [14], Roitman [90]). *If R is an $acc\perp$ algebra over an uncountable field k then $R[X]$, preserves $acc\perp$ for any set X of commuting variables, hence R is Kerr.*

Theorem 1.1 follows from the more general:

THEOREM 1.2 (Camillo and Guralnick [14]). *Let R be a ring containing an uncountable set V in the center of R so that if $u \neq v$ are elements of V, then $u - v$ is not a zero-divisor. Then if P is a property of rings that holds in a ring S iff P holds in any countable subring, then P holds in R iff $R[X]$ preserves P for any set X of commuting variables.*

REMARK. The method of proof is to show that any countable subring R_0 of $R[X]$ embeds in R.

Roitman (ibid.) provided the *coup de grace*:

THEOREM 1.3 (Roitman [91]). *If k is any countable field, there exists an $acc\perp$ algebra over k that is not Kerr.*

Noting that R in Roitman's example is not Goldie, Antoine and Cedó generalized Kerr's construction to prove:

THEOREM 1.3′ (Antoine and Cedó [3]). *For each finite field k there exists a commutative Goldie algebra R over k that is not Kerr.*

1.1. *Some positive results on Kerr rings*

THEOREM 1.4 (Faith [35–38]). *If R is a commutative acc$\perp$ ring then R is a Kerr ring under any of the following conditions*:

(1) *R is a uniform ring, i.e., has Goldie dimension* 1.

(2) *R is Goldie and its quotient ring $Q(R)$ has nil Jacobson radical.*

(3) *$Q(R)$ is finitely embedded.*

(4) *R is reduced.*

In these cases, $Q(R)$ is Artinian; *and $R = Q(R)$ in case* (3). (Cf. Remark 1.15A below.)

REMARK 1.5A. In case (1) of Theorem 1.4, then $Q(R)$ is QF. In case (4), $Q(R)$ is semisimple Artinian, hence QF.

REMARK 1.5B. Any subring of a Noetherian ring R is Kerr by the Hilbert basis theorem and the fact that any subring of an *acc*$\perp$ ring satisfies *acc*$\perp$.

COROLLARY 1.6 (Ibid.). *Any Goldie commutative algebra R that is either algebraic over a field k, or has vector space dimension over k strictly less than the cardinal of k, is a Kerr ring.*

Cedó and Herbera proved that the property of being a Kerr ring is quite restrictive:

THEOREM 1.7 (Cedó and Herbera [18]). *There exists a Kerr ring R such that for any integer n, the polynomial ring in n variables $R[X_1, \ldots, X_n]$ is Kerr but that in $n + 1$ variables over R is not Kerr.*

REMARK 1.7$'$. The ring R in this construction is not Goldie, i.e., does not have finite Goldie dimension.

An ideal P is an *associated prime* of R if P is a prime ideal and $P = x^{\perp}$ for some $x \in R$.

THEOREM 1.8 (Brewer and Heinzer [9]). *If R is a commutative ring, the contraction map $R[X] \to R$ sending $P \to P \cap R$ is bijection between associated primes of $R[X]$ and R.*

REMARK 1.9. The author, [41], gave a new direct proof using ideas of Shock [95]. Annin [2] has generalized Theorem 1.8 to modules over noncommutative R, and to skew polynomial rings.

COROLLARY 1.10. *If R satisfies the acc* (*dcc*) *on associated primes, then so does $R[X]$.*

REMARK 1.11. The same holds true for the set of all prime ideals, i.e., acc is preserved and reflected by $R[X]$. This is an exercise in Kaplansky [71], p. 52, Ex. 2; cf. ibid., Theorem 39, p. 26. The converse of Corollary 1.10 obviously holds.

Below, for any ideal I of a ring R, $I^{\perp}$ denotes the right, and ${}^{\perp}I$ denotes the left annihilator of I in R.

A ring R is *right Kasch* if R satisfies the equivalent conditions:

(K1) If m is a maximal right ideal of R, then ${}^{\perp}m \neq 0$.

(K2) If m is a maximal right ideal of R, then $R/m \hookrightarrow R$.

REMARK 1.12. A right Kasch ring R coincides with its maximal right quotient ring $Q^r_{\max}(R)$. (See, e.g., Stenström [99], Lam [79], Corollary 13.24, or [40], sup. 12.0A.) Any Noetherian commutative ring R has Kasch $Q(R)$ as the next result indicates.

THEOREM 1.13 (Faith [36], Small). *If R is a commutative acc$\perp$ ring, then the classical quotient ring $Q = Q(R)$ is a semilocal Kasch ring.*

REMARK 1.14. For the attribution to Small, see [36], or [40], Remarks, sup., 16.32, or [79, Theorem 8.31, p. 283].

THEOREM 1.15 (Faith [37]). *If R is a commutative Goldie ring and $Q = Q(R)$ has nil Jacobson radical, then Q is Artinian.*

REMARK 1.15A. (1) Any finitely embedded commutative *acc*$\perp$ ring is Artinian (Faith [35]). Moreover, R is then QF iff R has square-free socle (ibid.). For background on *QF* rings, see, e.g., [31], [40], or [79]. (2) Any finitely embedded two-sided Noetherian ring is Artinian (Ginn and Moss [42]).

COROLLARY 1.16. *If R is a commutative Goldie ring such that either Q is an algebraic algebra over a field, or else Q is an algebra of dimension $< \operatorname{card} k$, then Q is Artinian.*

PROOF. [37], Theorem 2.5 and Corollary 2.7. □

COROLLARY 1.17 (Faith [38]). *If R is a uniform commutative acc$\perp$ ring, then Q is QF.*

REMARK 1.17A. Shizhong [94] proved this for a subdirectly irreducible *acc*$\perp$ commutative ring, in which case R is *QF*. Note that Corollary 1.16 implies Corollary 1.6; and Corollary 1.17 implies Theorem 1.4(1).

DEFINITION 1.17B. (1) A ring R is *right $\aleph_0$-injective* (respectively f.g.-*injective*; respectively *p-injective*) if every mapping $f : I \to R$ of a countably generated (respectively finitely generated (= f.g.); respectively principal) right ideal I extends to $R \to R$, equivalently there exists $a \in R$ so that $f(x) = ax$ for all $x \in I$.

(2) A right R-module M is *FP-injective* provided that every mapping $f : S \to M$ of a f.g. submodule S of a free module F extends to $F \to M$. (Because S is f.g., it is sufficient to assume F is a f.g. free module in this definition.)

(3) A ring is right *FP*-injective (or right self-FP-injective) provided that R is an *FP*-injective right R-module.

REMARK 1.17C. Obviously $\aleph_0$-injective $\Rightarrow$ *FP*-injective $\Rightarrow$ f.g.-injective $\Rightarrow$ p-injective. However, R is right *FP*-injective iff every full matrix ring R_n is right p-injective. (See Theorems 5.0A, 5.0B and 5.1 in Section 5.)

COROLLARY 1.18. *If R is a commutative acc$\perp$ ring, and if Q is f.g.-injective (e.g., FP or self-injective), then R is QF.*

PROOF. [38], Theorem 3. □

2. Coherent rings

Following Chase [19], a ring R is *right coherent* if all finitely generated (= f.g.) right ideals of R are finitely presented (= f.p.)

Note: any f.g. projective module is f.p.

THEOREM 2.0A (Chase [19]). *The following are equivalent statements on a ring R.*
(1) *R is right coherent.*
(2) *Any product of copies of R is a flat left R-module.*
(3) *Any f.g. submodule of a free right R-module is finitely presented.*
(4) *The intersection of two f.g. right ideals is f.g., and $a^\perp$ is f.g. for all $a \in R$ ($= R$ is right pseudo-coherent. Cf. Section 2.3).*

DEFINITION. A ring R is *von Neumann regular* (= VNR) iff R satisfies the following equivalent conditions, which are right-left symmetric.
(a) Every principal right ideal is generated by an idempotent.
(b) Every right R-module is flat.
(c) Every f.g. right ideal is generated by an idempotent.

Note: (a) $\Leftrightarrow$ (b) is due to Auslander [4] and Harada [59]. See, e.g., [30], 11.24, p. 434, or [79], 4.21, p. 128.

EXAMPLES OF RIGHT COHERENT RINGS 2.0B (Ibid.).
(1) R right Noetherian.
(2) R right semihereditary, e.g., any VNR ring is right and left semihereditary, hence coherent.
(3) The polynomial ring $k[X]$ in any set X of commuting variables over a field k.
(4) The free algebra $k\langle X\rangle$ on a finite set X over a field k. (This ring is hereditary.)
(5) The endomorphism ring of any free module F over a right coherent ring R.
(6) If R is right coherent, and I is an ideal f.g. as a right ideal, then R/I is right coherent.

REMARK 2.0C. A local right (semi) hereditary ring is a *right* (*semi*) *fir*, i.e., any f.g. right ideal is free of unique rank. This follows since: (1) every projective module over a local ring is free by Kaplansky's theorem; and (2) The unique decomposition, or Krull–Schmidt–Azumaya theorem, holds for direct decompositions of modules having local endomorphism rings.

THEOREM 2.1 (Menal [85]). (1) *A ring R is right coherent iff every principal right ideal in every $n \times n$ matrix ring R_n has a f.g. right annihilator. In this case, every f.g. right ideal if R_n has f.g. right annihilator in R_n, and R_n is right coherent*; (2) *If F is a free right R-module of infinite rank, then $A = \operatorname{End} F_R$ is right* (*left*) *coherent iff every principal right* (*left*) *ideal of A has f.g. right* (*left*) *annihilator.*

REMARKS 2.2.

(1) The proof of (1) is by Morita theory, and (2) follows since $A \approx A_n$ for all $n \geqslant 1$;
(2) Lenzing [80] (respectively Menal [85]) characterizes the condition that A is right (respectively left) coherent in several different ways, involving the cardinalities of the sets of relations and generators of R. (Cf. Theorem 6.4 below.) Lenzing (loc. cit.) also gives an example of noncoherent (incoherent?) ring R such that R^{ω} is flat. This illuminates condition (2) in Theorem 2.0A.

THEOREM 2.3 (Matlis [82]). *A commutative ring R is coherent iff* $\operatorname{Hom}_R(E, F)$ *is a flat R-module for any two injective modules E and F. In this case,* $\operatorname{End}_R E$ *is faithfully flat for any injective cogenerator E.*

EXAMPLE 2.4 (Soublin [97]). (1) There exists a coherent commutative ring $R[X]$ is not coherent.

The specific example: $R = \mathbb{Q}^{\omega}[\![u, v]\!]$ in commuting variables u and v. Note, there exists a ring R such that $R[X]$ is coherent but $R[\![X]\!]$ is not (ibid.);

(2) There exists a noncommutative VNR ring R such that $R[X]$ is neither right nor left coherent. See Remark 2.6A and Theorem 2.7.

THEOREM 2.5 (McCarthy [84]). *If R is a commutative VNR ring, then the polynomial ring $R[X]$ is semihereditary.*

THEOREM 2.6 (Camillo [12], Pillay [89]). *If the polynomial ring $R[X]$ in a variable X is right or left semihereditary, then R is VNR.*

REMARK 2.6A. Camillo's contribution is for R commutative. In his Math. Review of Pillay [89], C.U. Jensen remarks that S. Jøndrup has shown (unpublished) that there exists a VNR ring R such that $R[X]$ is neither right nor left coherent. (Cf. Example 2.4(2).)

However, assuming that $R[X]$ is coherent, then one gets a positive result:

THEOREM 2.7 (Fieldhouse [50]). *A polynomial ring $R[X]$ in a single variable X is right semihereditary iff $R[X]$ is right coherent and R is VNR.*

REMARK 2.7A. This implies Theorem 2.6.

DEFINITION 2.7B. (1) A nilpotent element x in R has *index* n if n is the least positive integer such that $x^n = 0$. (In particular, x has index 1 iff $x = 0$.) The *index of an ideal* I of R is the supremum of the indices of all nilpotent elements of I. If this supremum is finite, then I is said to be of *bounded* or *finite index*. This also applies to the ideal $I = R$.

(2) A ring R is *reduced* if R has index 1. A VNR ring R is said to be *Abelian* if R is reduced. One shows that a VNR ring R is Abelian iff every idempotent of R is central. See, e.g., Goodearl [55, Theorem 3.2, p. 26] for this and equivalent formulations.

THEOREM 2.7C (Kaplansky [69]). *Let R be a VNR ring, and $n \geqslant 0$. Then $R \approx A_n$ for an Abelian VNR ring A iff every primitive factor ring of R has bounded index n.*

PROOF. See Goodearl, loc. cit., p. 76, Theorem 7.14. □

REMARK 2.7C$'$. For a related theorem of Armendariz–Steinberg, see [40], p. 95, Theorem 4.5. For theorems on VNR rings of bounded index, see Goodearl, loc. cit., Chapter 7.

THEOREM 2.8A (Carson [16]). *If R is a VNR ring of finite index, then $R[X]$ is coherent for any set X of commuting variables.*

PROOF. The proof reduces to the case R is Abelian using Kaplansky's theorem 2.7C and 2.1(1). □

REMARK 2.8B. By Theorem 2.7, $R[X]$ in Theorem 2.8A is then semihereditary.

THEOREM 2.8C (Carson [15]). *Let R be an algebraic algebra over a field k. Further, assume R is reduced* ($= R$ *has no nilpotent elements* $\neq 0$). *If R is* (1) *commutative or* (2) *if k is the field $\mathbb{R}$ of real numbers, then $R[X]$ is coherent* (*on both sides*) *for any set X of commuting variables.*

PROOF. By a theorem of Jacobson [66, Proposition 1, p. 210], any algebraic algebra R over a field is π-regular in the sense that for every $A \in R$ there exists $x \in R$ and an integer n so that $a^n x a^n = a^n$. This implies that $e = xa^n$ (or $a^n x$) is an idempotent and $ea^n = a^n$. Moreover, e is central when R is reduced. In this case R is *biregular* in the sense that every principal two-sided ideal (a) is generated by a central idempotent e. Since the idempotent e is central $ea^n = a^n$ implies $ea = a$.

Next assume (1) that R is commutative. Since $e \in (a)$ in the above, we have $y \in R$ so that $e = ay$, hence $a = ea$ implies that $a = aya$, hence R is VNR in case (1), and the theorem follows from, e.g., Theorem 2.5, that in fact $R[X]$ is semihereditary.

For (2) refer to Carson [15]. □

THEOREM 2.9A. *If R is a commutative semihereditary ring, then $R[X]$ is coherent for a finite set of commuting variables.*

PROOF. See Vasconcelos [104], pp. 85–86, esp. Proposition 8.2(b). □

REMARK 2.9B. (1) For the proof of Vasconcelos' proposition just cited, one needs to know that a semihereditary ring R has VNR quotient ring $Q(R)$. See Endo's Theorem 2.12A below.

(2) Vasconcelos, ibid., attributes Theorem 2.9A essentially to Gruson and Raynaud [57].

(3) Theorem 2.5 is a corollary.

(4) Cf. Example 2.4 and Remark 2.6A.

THEOREM 2.10 (Greenberg and Vasconcelos; see Vasconcelos [104], p. 88). *If R is a coherent commutative ring of global dimension* 2, *then $R[X_1, \dots, X_n]$ is coherent.*

DEFINITION 2.11A. A ring R is a *right p.p.*, or *PP*, ring (also called *right Rickhart*) if R satisfies the equivalent conditions:

(1) Every principal right ideal xR is projective,

(2) For every $x \in R$, $x^{\perp}$ is generated by an idempotent.

PROPOSITION 2.11B. *If every $n \times n$ matrix ring R_n over R is right PP, then R is right coherent.*

PROOF. See Theorem 2.1(1). □

DEFINITION 2.11C. (1) A ring R is *right Bezout* provided each f.g. right ideal is principal; (2) A ring R is a *right chain* ring if the right ideals of R are linearly ordered by inclusion. Any right chain ring is right Bezout. *Valuation ring* (= VR) is a variant term for chain ring.

PROPOSITION 2.11D. *Any right Bezout right PP ring is semihereditary. In particular any right Bezout domain is semihereditary.*

PROOF. Any f.g. right ideal in a Bezout ring is principal hence projective in a PP ring. Any integral domain is PP, hence a Bezout domain is semihereditary. □

THEOREM 2.12A (Endo [26]). *A commutative ring R is a PP ring iff R has a VNR quotient ring $Q(R)$ and R_m is an integral domain for every maximal* (*or prime*) *ideal m.*

REMARK. This generalizes Theorem 1.5.

THEOREM 2.12B (Jøndrup [68]). *If R is commutative, then R is PP iff the polynomial ring $R[X]$ is PP.*

REMARK 2.12E. (1) See the author's book [40, p. 140] for a number of results on PP rings and Bezout rings; (2) In Theorem 2.12B, note that $Q(R[X])$ is VNR by Theorem 2.12A.

THEOREM 2.13 (Brewer, Rutter and Watkins [10]). *The following are equivalent conditions on a commutative VNR ring R and its power series ring $S = R[[X]]$:*

(1) *S is semihereditary.*

(2) *S is coherent.*

(3) *S is a Bezout PP ring.*
(4) *R is $\aleph_0$-injective and PP.*

REMARK 2.14. Herbera [63] extended this theorem to noncommutative rings in various ways. In particular if $R[\![X]\!]$ is right semihereditary then R must be VNR. Cf. Hirano, Hung and Kim [64], and Karamzadeh and Koochakapoor [72]; (2) By Theorem 2.12A, $Q(S)$ in Theorem 2.13 is VNR.

THEOREM 2.15 (Harris [61]). *Let R be a commutative ring.*
(1) *If R is semilocal, then R is coherent iff every local ring R_m at a maximal ideal m is coherent.*
(2) *There exists a noncoherent R such that R_m is Noetherian, hence coherent, for all m.*

Couchot characterized commutative coherent rings locally:

THEOREM 2.16A (Couchot [23]). *A commutative ring R is coherent iff for each maximal ideal m*: (1) *R_m is a coherent ring*; *and* (2) *every FP-injective R-module M localizes to an FP-injective R_m-module* (*cf. Section* 5).

REMARK 2.16B. (1) According to 2.15 and 2.16A, the condition (1) of 2.16A holds in a semilocal coherent ring; (2) Localizations of injective modules over a commutative ring are injective when R is Noetherian, but may not be otherwise. See Dade [24] or consult [40], Theorems 3.17A–B.

DEFINITION 2.17A. A ring R is a *Baer ring* if every right annihilator is generated by an idempotent. Then the same is true of left annihilators.

REMARK 2.17A′. In this definition we are following Kaplansky [70], Berberian [8], and [79]. Elsewhere, e.g., in papers of Kist [76] and Facchini [27], Baer rings are defined as *PP* rings.

PROPOSITION 2.17B. *Any right Baer ring R is right PP* (= *Rickart*).

PROOF. See 2.11A(2). □

2.1. *Π-coherent rings and star-rings*

A *right annulet* of a ring R is an annihilator right ideal (see Section 1). A right annulet I of R is *finitely annihilated* (= FA) provided I is the right annihilator $L^{\perp}$ of a finite subset L of R, equivalently, R/I embeds in a free R-module.

A right R-module M is *torsionless* provided that M satisfies the equivalent conditions:
(1) M embeds in a product of copies of R.
(2) The canonical mapping of M into its double dual $M^{\star\star}$ is injective. (Cf. [40], 1.5.)

Thus for any right annulet I, the cyclic module R/I is *torsionless*, since R/I embeds in a product of copies of R. R is said to be a *right FA-ring* if every cyclic torsionless R-module R/I embeds in a free R-module, equivalently, every right annulet is FA.

A ring R is *right Π-coherent* if every f.g. torsionless right R-module is f.p. A Π-coherent ring is one that is both left and right (Π)-coherent.

2.2. *Star rings*

A ring R is a *left star-ring*, notation: *left $\star$-ring*, if the R-dual module $M^\star = \mathrm{Hom}_R(M, R)$ of any f.g. left R-module M is f.g. as a canonical right R-module. If this is true for all finitely presented left R-modules M, then R is said to be a *left weak $\star$-ring*.

EXAMPLE 2.17C. (1) if A is any left R-module, and K is a submodule, then

$$(A/K)^\star \approx \mathrm{ann}_{A^\star} K = B$$

where the inverse isomorphism sends $f \in B$ onto the element $\bar{f} \in (A/K)^\star$ such that

$$\bar{f}(a + K) = f(a) \quad \forall a \in A.$$

(See [30, p. 378].) In particular, for $A = R$, we see that $(R/K)^\star \approx K^\perp$. In case K is f.g., R/K is f.p., so if R is left weak $\star$, then $K^\perp$ is f.g. for all f.g. K. Since the Morita correspondence (category isomorphism),

$$R\text{-mod} \underset{X \mapsto X^n}{\rightsquigarrow} R_n\text{-mod}$$

sends f.p. left R-modules generated by n elements onto f.p. cyclic left R_n-modules, one obtains:

THEOREM 2.18 (Camillo [13]). *The following on equivalent conditions on a ring R*:

(1) *R is a left weak $\star$-ring.*
(2) *$L^\perp$ is f.g. for all f.g. left ideals L of R_n for all n.*
(3) *R is right coherent.*

Cf. Menal's Theorem 2.1, and Camillo [13]. Similarly one proves:

THEOREM 2.19 (Camillo [13]). *For a ring R, the following are equivalent conditions*:

(1) *R is a left $\star$-ring.*
(2) *$L^\perp$ is f.g. for all left ideals L of R_n for every n.*
(3) *R is right Π-coherent.*

COROLLARY 2.20. (1) *If R is a coherent ring, then every left or right FA annulet of R_n is f.g. for all n*; (2) *If R is Π-coherent, then every left or right annulet of R_n is f.g. for all n.*

COROLLARY 2.21. *A left coherent right Π-coherent ring is left Π-coherent.*

PROOF. See [34], Corollary 2.5B. □

COROLLARY 2.22. *A left coherent left $\star$-ring is left Π-coherent.*

PROOF. *Ibid*. Corollary 2.5C. □

2.3. *Pseudo-coherent rings*

As stated in Theorem 2.0A, a ring R is (right) pseudo-coherent provided that $x^{\perp}$ is f.g. for all $x \in R$. If this holds for all $n \times n$ full matrix rings over R, then R is (right) *matrix pseudo-coherent*.

EXAMPLE 2.23. (1) If R is right matrix-pseudo-coherent, then R is right coherent. (See Theorem 2.1(1).)

(2) If R is right Rickart (= PP) then R is right pseudo-coherent. (See Definition 2.11A.)

(3) Any Baer ring is right and left PP, hence right and left pseudo-coherent (see Proposition 2.17B).

(4) If R is a commutative VNR but not self-injective, then R is coherent, hence matrix-coherent, but R is not matrix-Baer by Theorem 3.6 (below).

(5) A right valuation ring R is right pseudo-coherent iff R is right coherent. (If I is any f.g. right ideal, then $I = xR$ is principal, hence I is f.p. iff $x^{\perp}$ is f.g.)

PROPOSITION 2.23A. *If R is a right pseudo-coherent right Kasch ring then every maximal right ideal of R is finitely generated* (*and $R = Q^r_{\max}(R)$*).

PROOF. If m is a maximal right ideal, then m is an annihilator right ideal (= right annulet) since $R/m \hookrightarrow R$. Since m is maximal, m is a principal annihilator right ideal, hence f.g. since R is pseudo-coherent. The parenthetical remark follows from Remark 1.12. □

COROLLARY 2.23B. *If R is a commutative Kasch* 0*-dimensional pseudo-coherent ring, then R is Artinian.*

PROOF. In a zero-dimensional ring every prime ideal is maximal, and hence f.g. by Proposition 2.23A. By a theorem of I.S. Cohen [21], finite generation of prime ideals in a commutative ring implies Noetherian; and in a zero dimensional ring this implies Artinian. □

THEOREM 2.24. *If R is a commutative pseudo-coherent acc$\perp$ ring, then each of the* (*finitely many*) *maximal ideals of $Q(R)$ is f.g.*

PROOF. By Theorem 1.12, $Q = Q(R)$ is semilocal Kasch. Since R pseudo-coherent implies that Q is pseudo-coherent, the theorem follows from Proposition 2.23A. □

3. FGTF rings

A ring R is *right FGTF* if every f.g. torsionless right R-module embeds in a free R-module.

3.1. *Matrix-P rings*

Let P be a property of rings. Then R is said to be *matrix-P* if every $n \times n$ matrix ring R_n satisfies P.

THEOREM 3.1 (Faith [34]). *A ring R is right FGTF iff R is right matrix-FA, that is, every right annulet of R_n is finitely annihilated for all n.*

PROOF. As stated in the Example *sup* Theorem 2.18, under the Morita correspondence, n-generated right R-modules convert to cyclic R_n-modules. A cyclic module R_n/I is torsionless iff I is an annulet, and embeds in a free module iff I is *FA*. □

PROPOSITION 3.2 (Faith [34], Proposition 1.1). *A right $\star$-ring R is right FGTF.*

PROOF. The equivalence of right $\star$-rings and left Π-coherent rings, and Theorem 2.18(2) shows that "right star" is Morita invariant. In any ring A, and right ideal I, the A-dual of A/I is ${}^{\perp}I$. Applied to this case, we conclude that in any matrix ring R_n over R that any left annulet is f.g., hence every right ideal is FA, so R is right FGTF. □

THEOREM 3.3 (Faith [34]). *The following are equivalent conditions on a ring R*:
(1) *R is left star.*
(2) *R is right Π-coherent.*
(3) *R is right coherent and right FGTF.*
Furthermore, a left coherent right Π-coherent ring R is left Π-coherent, hence left FGTF, and right star.

PROOF. (1) $\Leftrightarrow$ (2) by 2.19. Obviously (2) implies R is right coherent, and by 2.18(2) and 3.1, R is right FGTF. Conversely, assuming (3), by 2.18(2), f.g. left ideals of R_n have f.g. right annihilators, hence right FGTF implies by 3.1 that all right annulets of R_n are f.g., so R is right Π-coherent (see 2.19). Thus, (3) $\Leftrightarrow$ (2).

For the last part, assuming R is right Π-coherent, then R is left star, hence left FGTF by Proposition 3.2. Then left coherency implies left Π-coherent by (3) $\Rightarrow$ (2) above, hence right star by (2) $\Rightarrow$ (1). □

COROLLARY 3.3A. *For a coherent ring, the following are equivalent*:
(1) *R is right or left star.*
(2) *R is right or left Π-coherent.*
(3) *R is right or left FGTF.*

In this case, R is both right and left star, Π-coherent and FGTF. In particular, one-sided annulets of every matrix ring are finitely generated and finitely annihilated.

3.2. *Baer rings*

Recall that a ring R is a *Baer ring* if every right annulet is generated by an idempotent. (In this case, every left annulet is generated by an idempotent.)

EXAMPLE. Any right self-injective VNR ring is a Baer ring (see Theorem 3.5 below), hence is a matrix-Baer ring (see Theorem 3.4).

THEOREM 3.4 [34, 2.2-3]. *For a VNR ring R, the following are equivalent*:

(1) *R is right FGTF.*
(2) *R is left FGTF.*
(3) *R is a right $\star$-ring.*
(4) *R is a left $\star$-ring.*
(5) *R is right Π-coherent.*
(6) *R is left Π-coherent.*
(7) *R is a matrix-Baer ring.*

PROOF. R is coherent, hence (1)–(6) are equivalent by Corollary 3.3A. Obviously, (1) $\Rightarrow$ (7) by Theorem 2.20(2). Conversely, assuming (7), since R_n is VNR, and annihilator right ideals are f.g., each is generated by an idempotent, so R is Baer. □

REMARK 3.4A. Not every Baer ring is matrix-Baer (Berberian [8]; Cf. 3.6 below).

3.3. *Nonsingular, co-nonsingular and Utumi rings*

A *complement submodule* S of an R-module M is one which is maximal w.r.t. $S \cap T = 0$ for some submodule T; equivalently, S is a maximal essential extension of a submodule S_0 of M. In this case, S has no essential extensions in M. Furthermore, if M is injective, or quasi-injective, a submodule S is a complement iff S is a direct summand of M. Cf. [40, pp. 52–53], and Theorem 3.9D on p. 60, *ibid.*

A ring R is *right nonsingular*, if $I^{\perp} = 0$ for any essential right ideal I. In this case the maximal right quotient ring $Q = Q^r_{\max}(R)$ is a right self-injective VNR ring. Furthermore, any complement right ideal of Q is an annihilator right ideal (see [31, Chapter 19], or [40, Chapter 12]). We record these facts in the following:

THEOREM 3.5 (R.E. Johnson and Y. Utumi). *The maximal right quotient ring $Q^r_{\max}(R)$ of a right nonsingular ring R is a right self-injective Baer VNR ring. Thus, any right self-injective VNR ring is Baer.*

A right and left nonsingular ring R is said to be *Utumi* provided that $Q^r_{\max}(R)$ coincides with $Q^{\ell}_{\max}(R)$, the maximal left quotient ring. An Abelian *VNR* is Utumi.

THEOREM 3.5′ (Utumi [100,101]). *A right and left nonsingular ring R is Utumi iff every complement right or left ideal is an annihilator.*

THEOREM 3.6 [34, Theorem 3.2]. *The following are equivalent conditions on an Utumi VNR*:

(1) *R is right FGTF.*
(2) *R is left FGTF.*
(3) *R is right self-injective.*
(4) *R is left self-injective.*
(5) *R is a matrix-Baer ring.*
(6) *R is a right and left $\star$-ring.*
(7) *R is Π-coherent.*

REMARK 3.7. This indicates the well-known facts that: (1) VNR's are not in general Baer rings; and (2) Baer rings are not in general matrix Baer rings; (3) even commutative VNR rings are not in general Π-coherent. In particular, coherent does not imply Π-coherent; (4) Moreover, Kobayashi [78] showed that any right self-injective VNR ring is a right $\star$-ring. Thus, by Theorem 3.4, a right but not left self-injective VNR ring R is a right and left $\star$-ring, right and left FGTF, matrix-Baer ring that is *not* left self-injective.

4. Matrix-*acc*$\perp$ rings

We need the following characterization of *acc*$\perp$ rings:

THEOREM 4.1 (Faith [29]). *A ring R satisfies acc$\perp$ iff for each right ideal I there is a f.g. ideal $I_1 \subseteq I$ such that I_1 and I have the same left annihilators*; *notation*: ${}^{\perp}I_1 = {}^{\perp}I$.

We let $\perp acc$ denote the acc on left annulets of R.

COROLLARY 4.2. *Any $\perp$ acc ring R is a right FA-ring. Thus, any matrix-$\perp$acc ring R is right FGTF.*

PROOF. This follows from Theorems 3.1 and 4.1. □

REMARK 4.3. Any subring of left Noetherian ring is a matrix $\perp acc$ ring, hence right FGTF.

The next result, stated without reference in Camillo [13], is non-trivial.

PROPOSITION 4.4 (Gilmer and Heinzer [51]). *If R is a commutative Noetherian ring, then for any set X of commuting indeterminates, the ring $A = R[X]$ localized at the set S of polynomials of content* 1 *is a Noetherian ring R_W, where $W = A\backslash S$, hence A is matrix-$\perp$acc and matrix-acc$\perp$.*

COROLLARY 4.5. *If R is a Noetherian commutative ring, then for any finite or infinite set X the polynomial ring $A = R[X]$ is FGTF.*

PROOF. Apply Proposition 4.4 and Remark 4.3. □

PROPOSITION 4.6. *If R is a right coherent right FA-ring, then all right annulets are $f.g$.*

PROOF. If I is a right annulet, then by Theorem 4.1 $acc\bot$ implies that $I = L^{\bot}$ for a f.g. left ideal L, hence R/I embeds in a free R-module F. Since R is right coherent, then R/I is f.p., so I is f.g. □

Proposition 4.6 gives another proof of (3) $\Rightarrow$ (2) of Theorem 3.3:

COROLLARY 4.7. *Any right coherent right FGTF ring is right Π-coherent.*

PROOF. Theorem 2.18(1) and Proposition 4.6 imply that every right annulet of every matrix ring is f.g., so R is right Π-coherent by 2.18(2). □

COROLLARY 4.8 (Camillo [13, Proposition 4]). *Any right coherent matrix-$\bot$acc ring R is right Π-coherent.*

PROOF. R is right FGTF by Corollary 4.7, so Corollary 4.2 applies. □

COROLLARY 4.9. *A coherent matrix-$\bot$acc ring is Π-coherent on both sides, hence a $\star$-ring, and FGTF, on both sides.*

PROOF. Follows from Corollary 4.9, Theorem 2.20 and Corollary 2.21. Cf. Corollary 3.3A. □

The next theorem goes back to Chase [19] for a field R.

THEOREM 4.10. *If R is right Noetherian, the polynomial ring $R[X]$ in any set X of variables is right coherent.*

PROOF. See Camillo [13, Proposition 5, p. 74]. □

THEOREM 4.11 (Camillo, ibid.). *If R is a semiprime right and left Noetherian ring, then the polynomial ring $A = R[X]$ in any set X of variables is Π-coherent.*

PROOF. Goldie's theorem states that R has a semisimple Artinian classical quotient ring Q. By a theorem of Small and Pillay (see, e.g., [40], p. 56, 3.6A, p. 62, 3.13f, and p. 78, 3.55B), A has a semisimple Artinian quotient ring, hence A is matrix-$acc\bot$ and $\bot acc$. Thus Corollary 4.8 applies. □

THEOREM 4.12 (Camillo [13, p. 75]). *If R is Noetherian, and an algebra over an uncountable field, then the polynomial ring $A = R[X]$ in any set X of variables is Π-coherent.*

5. FP-injective rings

As stated in 1.17B, a ring R is *right p-injective* iff for all $a \in R$ every mapping $f : aR \to R$ extends to a mapping $R \to R$; equivalently, there exists $r \in R$ such that $f(x) = rx \; \forall x \in R$. By Ikeda and Nakayama [65], R is right p-injective iff every principal left ideal of R is a left annulet.

A ring R is *right FP-injective* iff every mapping $f : S \to F$ of a f.g. submodule S of a free right R-module F extends to $F \to F$. (Cf. Theorem 5.1.)

THEOREM 5.0A (Stenström [98], Jain [67]). *A ring R is left FP-injective iff every f.p. right R-module is torsionless.*

REMARK. Stenström proved this for two-sided FP-injective rings.

THEOREM 5.0B. *A ring R is left FP-injective iff every f.g. right annulet of every matrix ring R_n is finitely annihilated* (= *FA*).

PROOF. This follows from 5.0A via Morita theory. The Morita correspondence $X \mapsto X^n$ maps n-generated right R-modules onto cyclic right R_n-modules. Similarly, if X is f.p., then X^n will be a f.p. right R_n-module. Consequently, f.p. torsionless right R-modules embed in free modules iff every f.p. cyclic torsionless right R_n-module, that is, R_n/I, where I is a f.g. right annulet of R_n, embeds in a free R_n-module. But this happens iff every such I is FA. □

THEOREM 5.1 (Ikeda and Nakayama [65] and Puninski, cited by Nicholson and Yousif [87]). *A ring R is right self-FP-injective iff every $n \times n$ matrix ring R_n is right p-injective, equivalently all principal left ideals of R_n are annulets.*

In this case all f.g. left ideals of R_n are left annulets.

PROOF. Using Theorem 5.0B, this is proved via (1) Morita theory in the same way as Menal's Theorem 2.1(1) is proved, and (2) using the Ikeda–Nakayama result that right p-injectivity of R is equivalent to principal left ideals being annulets. □

REMARK 5.2. (1) The Ikeda–Nakayama result [65] is the "equivalently" statement in Theorem 5.1, and related results (see, e.g., [31, Proposition 23.21, p. 189]); (2) Any ring R can be embedded in a right and left FP-injective ring, by a theorem of Menal and Vámos. The proof depends on a theorem of Eklof and Sabbagh on embedding any ring into an existential closed (= EC) ring. See [40, pp. 128–129] for bibliography citations; (3) The author and Facchini [28] classified commutative rings such that every factor ring R/I has FP-injective $Q(R/I)$. These rings are necessarily arithmetical (= locally valuation rings), and conversely if R is semilocal. See 6.6.

6. IF rings

A ring R is (*right*) *IF* if all injective right R-modules are flat. An *IF ring* is one which is both right and left IF. A ring R is *right FPF* if every f.p. right module embeds in a free module.

THEOREM 6.0A (Colby [22], Würfel [105]). *A ring R is right IF iff every finitely presented ($=$f.p.) right R-module embeds in a free R-module, i.e., iff R is right FPF.*

THEOREM 6.0B (Colby [22]). *The following are equivalent on a ring R*:
(1) $R' = \mathrm{Hom}_{\mathbb{Z}}(R, \mathbb{Q}/\mathbb{Z})$ *is a flat right R-module.*
(2) *R is right IF.*
(3) *Every f.p. right R-module has flat injective hull.*

REMARK 6.0C. (1) By the Bourbaki–Lambek theorem (see, e.g., [30, Theorem 5.60, p. 281]) a left R-module M is flat iff the character module $M' = \mathrm{Hom}(M, \mathbb{Q}/\mathbb{Z})$ is an injective *right* R-module. Since R is flat, this says that R' is an injective right R-module, in fact, $F = R'$ is an injective cogenerator of R_R, that is, R_R has a flat injective cogenerator. Theorem 6.0B follows from this; (2) (Rutter [92]) An injective right R-module E is flat iff every f.g. submodule embeds in a flat module.

THEOREM 6.1A. *The following are equivalent conditions on a ring R*:
(1) *R is right IF.*
(2) *Every f.g. right ideal of every matrix ring R_n is a finitely annihilated* ($=$ *FA*) *right annulet.*
(3) *Every f.p. cyclic right R_n-module embeds in a free R_n-module, for every n.*
In particular, a right IF ring R is left FP-injective.

PROOF. (1)–(3) follow by Morita theory in the same way as Theorem 5.0B. The last statement follows from the same theorem, or Theorem 5.0A. $\square$

DEFINITION. A ring R is right *FGF* if every f.g. right R-module embeds in a free module. (Cf. Chapter 8.)

THEOREM 6.1B (Jain [67], Rutter [92]). *Any right FGF ring is right IF.*

PROOF. Theorem 6.0A. $\square$

THEOREM 6.1C (Jain [67], Xue [107]). *For a right IF ring, the following are equivalent*:
(1) *R is QF.*
(2) *R has acc$\perp$* (*or dcc$\perp$*).
(3) *R has $\perp$acc or $\perp$dcc.*

PROOF. Clearly (1) $\Rightarrow$ (2) $\Leftrightarrow$ (3), so assume (3). We first assume that R is right Noetherian. Then R is right FGF by Theorem 6.0A, since every f.g. right module is f.p. Then, R is QF by [32], Theorem 3.5C.

Next, assume R has $acc\perp$, equivalently $\perp acc$. By (2) of Theorem 6.1A, then R is right Noetherian, hence QF.

Finally $dcc\perp$ implies via Theorem 6.1A again that R satisfies the *dcc* on f.g. right ideals, hence is left perfect. In this case we appeal to Jain [67]. □

REMARK. Xue extended Jain's result that (1) ⇔ (2).

THEOREM 6.1D (Colby [22], Theorem 4.7). *A left FP-injective left coherent ring R is right IF.*

REMARK 6.1E. Colby, ibid., p. 247 (also Xue [107, p. 476]) asks if the converse is true. By Theorem 6.1A, the question reduces to the Colby (ibid.) (Xue, ibid., p. 479) question: is a right IF ring R left coherent? By Theorem 6.1F below the answer is yes assuming two-sided IF. By Menal's Theorem 2.1(1), R is left coherent iff every principal left ideal Sa in any matrix ring $S = R_n$ has f.g. right S-dual module $(S/Sa)^\star \approx a^\perp$. Since left $\star$-rings have this property, we see that Colby's question is true if R right IF implies R is a left $\star$-ring. By Xue, ibid. Corollary 3.4, this is equivalent to asking if R is left FGTF. (Xue's corollary states that R left FP-injective is a left $\star$-ring iff R is left FGTF.)

The conditions coherent, FP-injective, IF etc., without the qualifiers right or left mean both right and left coherent, FP-injective, IF, etc.

A ring R is FPTF if every f.p. torsionless right or left R-module embeds in a free module. Cf. Theorem 6.0A.

THEOREM 6.1F (Chen and Ding [20]). *A ring R is IF iff R is coherent and FPTF.*

PROOF. Theorems 6.0A, 6.1A and 2.1(1). (See Remark 6.1E.) □

"IF THEOREM" 6.2 (Colby [22], Gomez Pardo and Gonzales [53], Saroj Jain [67], Matlis [83] and Würfel [105]). *The following are equivalent*:

(1) *R is IF.*
(2) *R is coherent and f.g. one-sided ideals are annulets.*
(3) *Annihilation defines a duality on the sets of f.g. right and f.g. left ideals.*
(4) *R is coherent and every one-sided annulet is f.g., and every f.g. one-sided ideal is an annulet.*
(5) *R is coherent and FP-injective.*

COROLLARY 6.3. *A ring R is right IF iff R is right matrix-IF.*

PROOF. This follows from 6.2 via the matrix characterizations of coherence and FP-injectivity stated in Sections 2 and 5, e.g., 2.1 and 5.1. □

THEOREM 6.4 (Faith and Walker [43], Menal [85], Sandomierski [93]). *Let F be a free right R-module of infinite rank, and let $A = \operatorname{End} F_R$. Then, the following are equivalent*:

(1) *A is left IF.*

(2) *A is left FP-injective as a left A-module.*
(3) *A is right self-injective.*
(4) *R is quasi-Frobenius* (*QF*).

PROOF. See Menal [85]. □

Spec R is the set of all prime ideals of R. *mSpec* R is the set of all maximal ideals of R.

THEOREM 6.5 (Matlis [83]). (1) *A commutative coherent ring R is IF iff R_m is IF $\forall m \in$ mspec R*; (2) *A commutative domain R is Prüfer iff R/I is IF $\forall f.g.$ ideals I.*

COROLLARY. *A commutative coherent ring R is IF iff R_P is IF $\forall P \in$ spec R.*

PROOF. See Corollary 3.5D of [39]. □

6.1. *Fractionally self FP-injective rings*

A commutative ring R has a *property P fractionally* provided that the quotient ring $Q(R/I)$ has property P for every ideal I.

A commutative ring R is *fractionally self-FP-injective* (= FSFPI) provided that for every ideal I, the quotient ring $Q(R/I)$ is an FP-injective ring.

THEOREM 6.6 (Facchini and Faith [28]). *A commutative FSFPI ring R is an arithmetic ring, equivalently, R_m is a VR for all maximal ideals. The converse holds for arithmetic rings which are either* (1) *semilocal,* (2) *fractionally semilocal,* (3) *zero-dimensional,* (4) 1*-dimensional domains, or* (5) *fractionally Kasch.*

COROLLARY 6.6A (Facchini). *Any VR is FSFPI.*

REMARK 6.7. This was communicated to the author at the Conference in Ring Theory, Universitat de Barcelona, fall '89, and subsequently anthologized by the author and Pillay [49].

THEOREM 6.8 [28]. *An arithmetical or coherent commutative FP-injective ring R is locally FP-injective. Conversely, any coherent locally FP-injective ring is FP-injective.*

PROOF. Ibid., Corollary 18. □

6.1.1. *Fractionally IF rings.* We restrict ourselves to commutative rings and abbreviate *fractionally IF rings* by *FIF*.

THEOREM 6.9. *A ring R is FIF iff R is fractionally coherent and FSFPI.*

PROOF. Clear from Theorem 6.2(5). □

REMARK 6.10. A sufficient condition that R be fractionally coherent is that R/I be coherent for every ideal I.

COROLLARY 6.11. *An arithmetic ring R satisfying one of* (1)–(5) *of Theorem* 6.6 *is FIF iff R is fractionally coherent.*

For the background in PF rings needed below, see Section 9, esp. Theorem 9.4.

EXAMPLE 6.12A. A non-Noetherian coherent Kasch almost maximal valuation PF ring R such that R/I is coherent for every ideal I.

Let R be the split-null or trivial extension (A, U) of the quasi-cyclic p-group $U = \mathbb{Z}_{p^\infty}$ by its endomorphism ring A, the ring of p-adic integers. Then R is a non-Artinian PF valuation ring (Osofsky [88]). Every ideal I of R either: (1) contains $D = (0, U)$, or (2) is properly contained in D. Then R/I is a factor ring of A in case (1) $R/I \approx R$ qua ring in case (2).

First consider case (2). Since $U/S \approx U$ for any subgroup $S \neq U$, then R/I is the split-null extension $(A, U/S) \approx R$.

Case (1). Since A is a PID, then every proper factor ring is QF. Thus all images of R are either QF or PF. Since A is a complete local ring, then every factor ring of R is linearly compact (= l.c.), that is, R is a maximal valuation ring. (This also follows from the fact that R is PF, hence a Morita ring, hence l.c.)

Every ideal I of R is principal except the ideal $D = (0, U)$ which is divisible by every element $x = (a, u)$ where $0 \neq a \in A$ and $u \in U$. The annihilator of x is $(0, W)$, where W is the annihilator of a in U, a finite cyclic subgroup of U. Conversely, the annihilator in R of any $0 \neq (o, u) \in D$ is a principal ideal containing D. It follows that R is pseudo-coherent, which in a VR implies coherent. Since the socle V of R is simple, and R is a VR, then R is Kasch. (This also follows from the fact that R is PF.)

Finally, for every ideal I of R, R/I is either a factor ring of A, hence Noetherian, whence coherent, or $R/I \approx R$, which is coherent.

EXAMPLE 6.12B. The example in 6.12A is FIF. This follows from 6.10 and 6.11.

THEOREM 6.13 [33]. *If R is a coherent commutative VR, and $J = \operatorname{rad} R$ is a principal ideal (p), then $P = \bigcap_{n=1}^{\infty} (p^n)$ is a prime ideal such that $\overline{R} = R/P$ is a discrete valuation domain* (= *DVD*). *If $P \neq P^2$, then either*

(1) $P/P^2 \approx Q(\overline{R})$, *or*

(2) $E = P/P^2$ *is the least injective cogenerator of $\overline{R}$ and $R/P^2 \approx (\overline{R}, E)$, the split null extension of E by $\overline{R}$.*

Moreover, in case (2), *R/P^2 is PF and a VR.*

PROOF. It is well-known that P is prime ideal, and $\overline{R}$ is DVD. Since P is divisible by p^n for every n, then $E = P/P^2$ is an injective $\overline{R}$-module. Since $\overline{R}$ is a PID and E is uniform, then E is indecomposable, and either torsion-free, whence (1) holds, or torsion, in which case (2) holds. See ibid. Theorem 4.4 and Corollary 4.6 for further details. □

7. Σ-injective modules

An injective right R-module E is *Σ-injective* if every direct sum of copies of E is injective.

THEOREM 7.1 (Faith [29]). *An injective right R-module E is Σ-injective iff for each right ideal I there exists a f.g. right ideal I_1 such that I and I_1 have the same annihilator in E, equivalently, R satisfies the acc on annihilators of subsets of E.*

An R-module M has an injective resolution

$$0 \to M \to E_0 \to \cdots \to E_n \to \cdots$$

where E_i is injective $\forall i \geqslant 0$.

DEFINITION 7.2. Following Beck [7], the *Noetherian depth n.d.M* of M is the largest integer n such that E_i is Σ-injective for $i \leqslant n$. If E_0 is not Σ-injective, then set $n.d.M = -1$; and if every E_n is Σ-injective, put $n.d.M = \infty$.

THEOREM 7.3A (Beck [7]). *The following are equivalent conditions on a commutative ring R:*

(1) $n.d.R \geqslant 0$.
(2) *R has a flat embedding in a Noetherian ring T.*
(3) *R has just finitely many associated primes $P_1, \ldots, P_r$, every zero divisor is contained in some P_i, and R_{P_i} is Noetherian $\forall i$.*
(4) *$Q(R)$ is Noetherian.*

PROOF. See Beck, ibid., Theorem 5.1, for the first three equivalences, and Corollary 3.10 for (1) $\Leftrightarrow$ (3). □

THEOREM 7.3B (Beck [7]). *If R is a commutative ring, then R has Σ-injective injective hull $E(R)$ iff R has Noetherian quotient ring $Q(R)$.*

PROOF. We give a sketch of a proof:

First, assume that $Q = Q(R)$ is Noetherian. Then $E = E(Q)$ is Σ-injective, e.g., by Theorem 7.1 above. (Furthermore, a theorem of Cartan–Eilenberg–Bass states that R is right Noetherian iff every injective right R-module is Σ-injective.)

Q is flat as an R-module, and injective as an R-module. In fact, every injective Q-module is an injective R-module (e.g., see [40, Proposition 4.C(3) and Theorem 7.35]). It follows from this that $Q = E(R)$ and is Σ-injective.

The converse uses the fact that by Theorem 1.8, Q is Kasch, hence $E(R) = E(Q)$ is an injective cogenerator over Q. Any right ideal of a ring satisfies the double annihilator condition with respect to a cogenerator (see, e.g., [40, 3.3(3)]). Moreover, Q satisfies acc on ideals by Σ-injectivity (see Theorem 7.1). Thus, Q is Noetherian. □

THEOREM 7.3C. *If R is a commutative ring with Σ-injective hull, then R is Kerr, and so is $R[X]$.*

PROOF. By Theorem 7.3B, $Q = Q(R)$ is Noetherian, hence R is Kerr by Remark 1.5B. It follows that $Q(R[X])$ is Noetherian, so $R[X]$ is Kerr too. □

DEFINITION. A ring R is *right Goldie* provided R satisfies *acc*$\perp$ and *acc*$\oplus$.

REMARK 7.4. $|\operatorname{Ass} R| < \infty$ for any *acc*$\oplus$ ring. See, e.g., [40, Theorem 16.18, p. 235], for commutative R. However, this is true for any right Goldie ring, since $|\operatorname{Ass}|$ is bounded by u-dim R. See Lam [79], Exercise 6.2.

THEOREM 7.5. *If R is a commutative Goldie ring, and if R_P is Noetherian for all $P \in$* Ass R*, then $Q(R)$ is Noetherian, hence R and $R[X]$ are Kerr rings.*

PROOF. By Remark 1.5B, R and $R[X]$ are Kerr whenever R is a subring of a Noetherian ring, so it suffices to prove $Q = Q(R)$ is Noetherian. But this follows since *acc*$\perp$ implies that every annulet $\neq R$ is contained in maximal annihilator, and maximal annihilators are associated primes. Thus by Remark 7.4, (3) of 7.3A is satisfied, so Q is Noetherian. □

THEOREM 7.6 (Cailleau [11]). *An injective right R-module E is Σ-injective iff E is a direct sum of indecomposable Σ-injective modules.*

COROLLARY 7.7. *If M is finitely generated right R-module, and if M has Σ-injective hull $E(M)$, then M is finite dimensional, that is, M satisfies acc$\oplus$.*

THEOREM 7.8 (Goursaud-Valette [56]). *If R has a faithful Σ-injective right R-module E, then R satisfies acc$\oplus$.*

THEOREM 7.9 (Faith [29]). *If R has Σ-injective hull as a right R-module, then R is right Goldie.*

PROOF. Follows from Theorem 7.1 to obtain *acc*$\perp$ in R. Then one can apply the argument loc.cit. that states that an integral domain R with $E(R)$ Σ-injective is a right Ore domain, or apply either Corollary 7.7 or Theorem 7.8 to obtain R has *acc*$\oplus$. □

COROLLARY (Ibid.). *An integral domain R is right Ore iff R has Σ-injective hull in mod-R.*

THEOREM 7.10 (Faith [29]). *A semiprime ring R is right Goldie iff R has Σ-injective hull in mod-R.*

PROOF. This follows from Goldie's theorem and the above results. □

8. When finitely generated modules embed in free modules: FGF rings

R is *right FGF* (respectively *CF*) if all f.g. (respectively cyclic) right R-modules embed in free modules. Any right CF ring R is right Kasch, hence $R = Q^r_{\max}(R)$.

REMARK 8.1. Any right FGF ring is IF by the Jain–Rutter theorem 6.1B.

THEOREM 8.2 (Faith and Walker [43]). *A ring R is right and left CF iff R is Quasi-Frobenius* ($=$ *QF*).

A conjecture of the author's is that $FGF \Rightarrow QF$. This has been verified in a number of special cases, and we propose to discuss some of these. The next theorem is due to a number of people, e.g., (FG 1–4) and (FG 6–9) is due to the author, (FG 5) to Tol'skaya and Björk independently, while (FG 10) is due to Gómez Pardo and Guil Asensio [54]. See the author's book [40, pp. 200–201], for further attributions and bibliographical details.

THEOREM 8.3. *Let R be right FGF ring. The following are equivalent*:

(FGF 1) *R is QF.*
(FGF 2) *R is a subring of a right Noetherian ring.*
(FGF 3) *R is semilocal with essential right socle,*
(FGF 4) *R has finite essential right socle.*
(FGF 5) *R is right self-injective.*
(FGF 6) *R is left CF.*
(FGF 7) *R is left Kasch.*
(FGF 8) *R satisfies acc$\perp$.*
(FGF 9) *R satisfies dcc$\perp$.*
(FGF 10) *Every countably generated right R-module has a maximal submodule.*

COROLLARY. 8.3′ *Any right coherent right FGF ring is QF.*

PROOF. Every right ideal I is an annulet and *FA*, hence f.g. by Proposition 4.6. Thus, R is right Noetherian so (FGF 2) applies. □

THEOREM 8.4 (Faith and Huynh [44]). *If every factor ring of R is right FGF, then R is QF, in fact uniserial.*

REMARK 8.5. See [40] and [44] for related theorems.

THEOREM 8.6 (Gomez Pardo and Guil Asensio [54]). *If α is a cardinal $\geqslant |R|$, and if every α generated right R-module embeds in a free module, the R is QF.*

REMARK. This verified a conjecture of Menal [85].

8.1. *Johns rings*

As stated, R is *right Johns* if R is a right Noetherian right D-ring ($=$ every right ideal is an annihilator); R is *strongly right Johns* if R is a right Noetherian right matrix-D ring, i.e., matrix-Johns.

THEOREM 8.7 (Rutter [92]). *For a right Artinian ring, the following are equivalent*:
(1) *R is QF.*
(2) *R_n is right D-ring for all n (= R is a right matrix-D ring).*
(3) *R is strong right Johns.*

THEOREM 8.8 (Faith–Menal theorem [46]). *The following are equivalent*:
(1) *R is strongly right Johns.*
(2) *R is left FP-injective and right Noetherian.*
(3) *Every finitely generated right R-module is Noetherian torsionless.*
In this case R is right FPF, i.e., every finitely generated faithful right ideal generates mod-R. (*Cf.* 9.4.)

REMARK. See [48] for the background on non-commutative, and [49] for commutative, FPF rings.

THEOREM 8.9 ([46]). *Let R be strongly right Johns. The following are equivalent*:
(1) *R is QF.*
(2) *R is semilocal.*
(3) *R has finite left Goldie dimension.*
(4) *R is left Noetherian.*
(5) *R_n is right FA for all n; equivalently, R is right FGF.*
(6) $J = \operatorname{rad} R = X^{\perp}$ *for a finite subset X of R.*

REMARK 8.10. (1) The semilocal case is due to Johns; (2) The first condition in (5) was stated in the cited paper, but since every f.g. right R module is torsionless by 8.8(3), then (5) implies by Theorem 3.2 that R is right FGF, and conversely.

9. FGT rings, D-rings and PF rings

We introduce a concept FGT that appears in Section 8, esp. 8.8.

A ring R *right FGT* if all f.g. right R-modules are torsionless. A ring R is a *right cogenerator ring* provided that satisfies the equivalent conditions:
(C1) R a cogenerator of mod-R.
(C2) Every right R-module is torsionless.
(C3) R contains the injective hull of every simple right R-module.

EXAMPLES 9.0. (1) Any right cogenerator ring is right FGT; (2) Any right FGF ring is right FGT.

PROPOSITION 9.1. *Every right FGT ring R is right Kasch, hence coincides with $Q^r_{\max}(R)$, the maximal right quotient ring of R.*

PROOF. If I is a maximal right ideal of R, then $V = R/I$ is torsionless, hence V embeds in a product of copies of R, hence $V \hookrightarrow R$. Thus, R is Kasch, hence $R = Q^r_{\max}(R)$ by Remark 1.12. $\square$

THEOREM 9.2. *A ring R is right FGT iff every right ideal of every matrix ring R_n is a right annulet.*

PROOF. Same proof as Theorem 5.1, and the other results of this type cited earlier. □

A ring R is a *right D-ring* if every right ideal is a right annulet. A *D-ring* is a two-sided *D-ring*. D denotes "dual" in this definition. Any right D-ring is right Kasch. (See proof of 9.1.)

THEOREM 9.3 (Hajarnavis and Norton [58]). *Let R be a D-ring. Then*:

(1) *R is semilocal.*
(2) *R is right and left $\aleph_0$-injective.*
(3) *Every f.g. R-module has finite Goldie and dual Goldie dimension.*
(4) *R/J^ω is Noetherian, where J is the Jacobson radical.*

COROLLARY 9.3A (Ibid.). *A D-ring is QF iff J is transfinite nilpotent, that is, $J^\alpha = 0$ for some ordinal α, where J^β is defined by transfinite induction for each ordinal β.*

Note: For a number of related results on when D-rings are QF, see [44] (in this volume).

THEOREM 9.4 (Azumaya [5], Osofsky [88], Utumi [102]). *A ring R is said to be* right PF *if R satisfies the equivalent conditions*:

(1) *R is right self-injective, semiperfect, and has essential right socle.*
(2) *R is right self-injective with finite, essential, right socle.*
(3) *R is right self-injective and right Kasch.*
(4) *R is an injective cogenerator of mod-R.*
(5) *Every faithful right R-module generates mod-R.*

REMARK 9.4A. (1) There exists a right but not left PF ring (Dischinger and Müller [25]).
(2) Azumaya initiated the concept (5) which he called "upper distinguished". The PF rings generalize QF rings, and Osofsky [88] gave the first example of a *PF* ring not *QF*;
(3) Any right PF ring is a right D-ring. (See [40], $3.3'(1)$.)

THEOREM 9.5 (Kato [73]). *The following are equivalent conditions on a ring R*:

(1) *R is (right and left) PF.*
(2) *R is a (right and left) cogenerator ring.*
(3) *R is right PF and left self-injective.*
(4) *Every 2-generated (right or left) R-module is torsionless.*

THEOREM 9.6. *The following are equivalent conditions on a ring R*:

(1) *R is PF.*
(2) *R is a matrix-D-ring.*
(3) *R is an FGT ring, i.e., f.g. left or right R-modules are torsionless.*

PROOF. Theorems 9.5 and 9.2. □

REMARK 9.7. By Morita's theorems, if R is PF, then $\mathrm{Hom}_R(\ ,R)$ induces a Morita duality. See, e.g., Xue [106], also Ánh [1], who proves that any commutative linearly compact ring has a Morita duality.

10. Some questions

We single out some questions that arose in the text, implicitly or explicitly.

If R is commutative, then the mapping of each matrix a of R_n onto its transpose is an involution of R_n. It follows that R is then matrix-$acc\perp$ iff R is matrix-$\perp acc$. Moreover, by definition, Π-coherent is right-left symmetric over a commutative ring R. However, several questions below are for not necessarily commutative rings:

(Q1) The first question is about what conditions on R are necessary and sufficient that the polynomial, matrix, and power series rings satisfy the *acc* on annihilators. See Section 1, esp. 1.1, 1.4, 1.6 and 1.7.

(Q2) Similarly, we know that unlike matrix rings, polynomial and power series rings over a coherent ring R are not in general coherent. We seek necessary and sufficient conditions in order that coherency is inherited by these rings over R. See Section 2, esp. 2.1, and 2.4–2.6.

(Q3) If R is a matrix-$acc\perp$ ring, is R Kerr?

(Q4) If R is matrix-$acc\perp$ and matrix-$\perp acc$, is R Kerr?

REMARK 10.1. By Soublin's example 2.4, coherency of a commutative ring is not necessarily preserved by the polynomial ring. Thus, by Theorem 2.1(1), the property that f.g. right (left) ideals of all matrix rings have f.g. right (left) annihilators is not preserved over the polynomial ring.

(Q5) If R is right Π-coherent, equivalently, if all right annulets of R_n are f.g. $\forall n$, is the polynomial ring $R[X]$ Π-coherent or even coherent?

REMARK 10.2.

(Q6) Camillo [13] gives sufficient conditions under which $R[X]$ is Π-coherent for a Noetherian ring R, where X an arbitrary set of commuting variables, e.g., when R is semiprime, or when R is an algebra over an uncountable field. The latter result involves the Camillo–Guralnick theorem 1.1. Cf. Theorems 4.10–4.12.

(Q7) Are right FGF rings QF? (See Section 8.)

(Q8) Are strongly right Johns' rings QF? (See Theorem 8.9.)

(Q9) (Menal's conjecture [85]) If α is any infinite cardinal, and every α-generated right R-module embeds in a free R-module is R QF?

This is a variation on a conjecture of Menal. Theorem 8.6 shows the truth of (Q9) when $\alpha \geqslant |R|$.

(Q10) Is a Noetherian matrix right D-ring necessarily QF? Cf. Theorem 8.8 and Theorem 8.9 for some necessary and sufficient conditions for an affirmative answer.

Acknowledgment

I thank Barbara L. Miller at the University of California, Riverside, for her patience and skill in putting this article into $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-TEX. I am deeply indebted to Scott Annin of the University of California at Berkeley, and Alberto Facchini of Padova University, for a number of corrections and suggestions. I also thank Professor A.B. Carson for his communications.

References

[1] P.N. Ánh, *Morita duality for commutative rings*, Comm. Algebra **18** (1990), 1981–1988.
[2] S. Annin, *Associated primes over skew polynomial rings*, Comm. Algebra **30** (2002), 2511–2528.
[3] R. Antoine and F. Cedó, *Polynomial rings over Goldie rings*, J. Algebra **237** (2001), 262–272.
[4] M. Auslander, *On regular group rings*, Proc. Amer. Math. Soc. **8** (1957), 390–405.
[5] G. Azumaya, *Completely faithful modules and self-injective rings*, Nagoya Math. J. **27** (1966), 697–708.
[6] G. Azumaya, *A characterization of coherent rings in terms of finite matrix functors*, Second Japan–Asia Symp. on Ring Theory and the Twentieth Symp. on Ring Theory at Okayama Univ., 1995, Okayama Univ. Press, Okayama (1996), 1–3.
[7] I. Beck, *Σ-injective modules*, J. Algebra (1972), 232–249.
[8] S.I. Berberian, *Baer $\star$-Rings*, Grundl. Math. Wiss., Vol. 195, Springer, Berlin (1972).
[9] J.W. Brewer and W. Heinzer, *Associated primes of principal ideals*, Duke Math. J. **41** (1974), 1–7.
[10] J.W. Brewer, E.A. Rutter and J.J. Watkins, *Coherence and weak global dimension of $R[\![X]\!]$ when R is von Neumann regular*, J. Algebra **46** (1977), 278–289.
[11] A. Cailleau, *Une caractérization des modules Σ-injectifs*, C. R. Acad. Sci. Paris **270** (1970), 1391–1394.
[12] V.P. Camillo, *Semihereditary polynomial rings*, Proc. Amer. Math. Soc. **45** (1974) 173–174.
[13] V.P. Camillo, *Coherence for polynomial rings*, J. Algebra **132** (1990), 72–76.
[14] V.P. Camillo and R. Guralnick, *Polynomial rings are often Goldie*, Proc. Amer. Math. Soc. **98** (1986), 567–568.
[15] A.B. Carson, *Coherence of polynomial rings over semisimple algebraic algebras*, Proc. Amer. Math. Soc. **34** (1972), 20–24.
[16] A.B. Carson, *Coherent polynomial rings over regular rings of finite index*, Pacific J. Math. **74** (1978), 326–332.
[17] F. Cedó, *Zip rings and Mal'cev domains*, Comm. Algebra **19** (1991), 1983–1991.
[18] F. Cedó and D. Herbera, *The maximum condition on annihilators for polynomial rings*, Proc. Amer. Math. Soc. **126** (1998), 2541–2548.
[19] S.U. Chase, *Direct products of modules*, Trans. Amer. Math. Soc. **97** (1960), 457–473.
[20] J. Chen and N. Ding, *Characterizations of coherent rings*, Comm. Algebra **27** (1999), 2491–2501.
[21] I.S. Cohen, *Commutative rings with restricted minimum condition*, Duke Math. J. **17** (1950), 27–42.
[22] R.R. Colby, *Rings which have flat injectives*, J. Algebra **35** (1975), 239–252.
[23] F. Couchot, *Anneaux auto-fp-injectifs*, C. R. Acad. Sci. Paris Sér. A-B **284** (1977), A579–A582.
[24] E.C. Dade, *Localization of injective modules*, J. Algebra **69** (1981), 416–425.
[25] F. Dischinger and W. Müller, *Left PF is not right PF*, Comm. Algebra **14** (1986), 1223–1227.
[26] S. Endo, *Note on p.p. rings*, Nagoya Math. J. **7** (1960), 167–170 (A supplement to Hattori's paper).
[27] A. Facchini, *Commutative rings whose finitely embedded modules have injective dimension* $\leqslant 1$, J. Algebra **77** (1982), 467–483.
[28] A. Facchini and C. Faith, *FP-injective quotient rings and elementary divisor rings*, Commutative Ring Theory, P.-J. Cahen et al., eds, Proc. of the II International Conf. (Fes̀), Lectures in Pure and Appl. Math., Vol. 125, Marcel Dekker, Basel (1997), 293–302.
[29] C. Faith, *Rings with ascending condition on annihilators*, Nagoya Math. J. **7** (1966), 169–191.
[30] C. Faith, *Algebra I: Rings, Modules and Categories*, Grundl. Math. Wiss., Vol. 190, Springer, Berlin (1973).
[31] C. Faith, *Algebra II: Ring Theory*, Grundl. Math. Wiss., Vol. 191, Springer, Berlin (1976).
[32] C. Faith, *Embedding modules in projectives*, Advances in Non-Commutative Ring Theory, P. Fleury, ed., Lecture Notes in Math., Vol. 951, Springer, Berlin (1982), 21–40.

[33] C. Faith, *Linearly compact injective modules and a theorem of Vámos*, Publ. Mat. **30** (1986), 127–148.
[34] C. Faith, *Embedding torsionless modules in projectives*, Publ. Math. **34** (1990), 379–387.
[35] C. Faith, *Finitely embedded commutative rings*, Proc. Amer. Math. Soc. **112** (1991), 657–659; Addendum **118** (1993), 331.
[36] C. Faith, *Annihilators, associated primes, and Kasch–McCoy commutative rings*, Comm. Algebra **19** (1991), 1867–1892.
[37] C. Faith, *Polynomial rings over Goldie–Kerr commutative rings*, Proc. Amer. Math. Soc. **120** (1994), 989–993.
[38] C. Faith, *Polynomial rings over Goldie–Kerr commutative rings II*, Proc. Amer. Math. Soc. **124** (1996), 341–344.
[39] C. Faith, *Commutative rings with acc on irreducible ideals*, Abelian Groups, Module Theory and Topology, Dikranjan and Salce, eds, Lecture Notes in Pure and Applied Algebra, Vol. 201, Marcel Dekker, Basel (1998), 157–178.
[40] C. Faith, *Rings and Things and a Fine Array of Twentieth Century Associative Algebra*, Surv. Amer. Math. Soc., Vol. 65, Providence (1999).
[41] C. Faith, *Associated primes in commutative polynomial rings*, Comm. Algebra **28** (2000), 3983–3986.
[42] C. Faith, *Ahmad Shamsuddin (1951–2001); 40th Memorial Eulogy*, May 16, 2001.
[43] C. Faith and E.A. Walker, *Direct sum decompositions of injective modules*, J. Algebra **5** (1967), 203–221.
[44] C. Faith and D.V. Huynh, *When self-injective rings are QF: A report*, J. Algebra App. **I** (2002), 75–1005.
[45] C. Faith and P. Menal, *A counterexample to a conjecture of Johns*, Proc. Amer. Math. Soc. **116** (1992), 21–26.
[46] C. Faith and P. Menal, *The structure of Johns rings*, Proc. Amer. Math. Soc. **120** (1994), 1071–1081; Erratum **125** (1997), 127.
[47] C. Faith and P. Menal, *A new duality theorem for semisimple modules and characterization of Villamayor rings*, Proc. Amer. Math. Soc. **123** (1995), 1635–1638.
[48] C. Faith and S. Page, *FPF Ring Theory: Faithful Modules and Generators of Mod-R*, Lecture Notes of the London Math. Soc., Vol. 88, Cambridge Univ. Press, Cambridge, New York and Melbourne (1984).
[49] C. Faith and P. Pillay, *Classification of FPF Rings*, Notas de Matemática, Vol. 4, Univ. de Murcia, Murcia (1990).
[50] D. Fieldhouse, *Semihereditary polynomial rings*, Publ. Math. Debrecen **25** (1978), 211.
[51] R. Gilmer and W. Heinzer, *The Noetherian property for quotient rings of infinite polynomial rings*, Proc. Amer. Math. Soc. **76** (1979), 1–7.
[52] S.M. Ginn and P.B. Moss, *Finitely embedded modules over Noetherian rings*, Bull. Amer. Math. Soc. **81** (1975), 709–710.
[53] J.L. Gómez Pardo and N.R. Gonzáles, *On some properties of IF rings*, Questiones Arithmeticae **48** (1987), 40–52.
[54] J.L. Gómez Pardo and P.A. Guil Asensio, *Embeddings in free modules and Artinian rings*, J. Algebra **198** (1997), 608–617.
[55] K.R. Goodearl, *Von Neumann Regular Rings*, Pitman, London, San Francisco, Melbourne (1979). Second edition with a 40-page appendix and problem updates, Krieger Publ., San Francisco (1991).
[56] J.M. Goursaud and J. Valette, *Sur l'enveloppe des anneaux de groupes réguliers*, Bull. Soc. Math. France **103** (1975), 91–102.
[57] L. Gruson and M. Raynaud, *Critéres de platitude et de projectivité*, Invent. Math. **13** (1971), 1–89.
[58] C.R. Hajarnavis and N.C. Norton, *On dual rings and their modules*, J. Algebra **93** (1985), 253–266.
[59] M. Harada, *A note on the dimension of modules and algebras*, J. Inst. Polytech, Osaka City Univ. **7** (1956), 17–28.
[60] M. Harris, *Some results on coherent rings*, Proc. Amer. Math. Soc. **17** (1966), 474–479.
[61] M. Harris, *Some results on coherent rings II*, Glasgow Math. J. **8** (1967), 123–126.
[62] M. Hazewinkel (ed.), *Handbook of Algebra, Vols. 1–3*, Elsevier Science, Amsterdam (1996, 2000 and 2003).
[63] D. Herbera, Ph.D. Thesis, Univ. Autónoma de Barcelona, 08193 Bellaterra, Spain (1991) (in Catalán).
[64] Y. Hirano, C.Y. Hung, and J.Y. Kim, *On strongly bounded and duo rings*, Comm. Alg. **23** (1995), 2199–2214.

[65] M. Ikeda and T. Nakayama, *On some characteristic properties of quasi-Frobenius and regular rings*, Proc. Amer. Math. Soc. **5** (1954), 15–19.

[66] N. Jacobson, *Structure of Rings*, Colloquium Pub., Vol. 37, Amer. Math. Soc., Providence, RI (1955), revised (1964).

[67] Saroj Jain, *Flat and FP-injectivity*, Proc. Amer. Math. Soc. **41** (1973), 437–442.

[68] S. Jøndrup, *p.p. rings and finitely generated ideals*, Proc. Amer. Math. Soc. **28** (1971), 431–435.

[69] I. Kaplansky, *Topological representations of algebras II*, Proc. Amer. Math. Soc. **68** (1950), 62–75.

[70] I. Kaplansky, *Rings of Operators*, Benjamin, Addison-Wesley, Reading, MA (1968).

[71] I. Kaplansky, *Commutative Rings*, Allyn and Bacon, Boston, rev. edn, Univ. of Chicago (1974).

[72] O.A.S. Karamzadeh and A.A. Koochakapoor, *On $\aleph_0$-self-injectivity of strongly regular rings*, Comm. Algebra **27** (1999), 1501–1513.

[73] T. Kato, *Self-injective rings*, Tohoku Math. J. **19** (1967), 485–494.

[74] T. Kato, *Torsionless modules*, Tohoku Math. J. **20** (1968), 234–243.

[75] J.W. Kerr, *An example of a Goldie ring whose matrix ring is not Goldie*, J. Algebra **61** (1979), 590–592.

[76] J.W. Kerr, *The polynomial ring over a Goldie ring need not be Goldie*, J. Algebra **134** (1990), 131–159.

[77] J. Kist, *Two characterizations of commutative Baer rings*, Pacific J. Math. **50** (1974), 125–134.

[78] S. Kobayashi, *A note on regular self-injective rings*, Osaka J. Math. **21** (1984).

[79] T.Y. Lam, *Lectures on Modules and Rings*, Springer, New York (1999).

[80] H. Lenzing, *Über kohärente Ringe*, Math. Z. **114** (1970), 201–212.

[81] H. Lenzing, *Die Kohärenz unendlicher Matrizxenringe*, Rings, Modules and Radicals (Proc. Colloq. Keszthely, 1971), Colloq. Math. Soc. János Bolyai, Vol. 6, North-Holland, Amsterdam (1973), 329–336.

[82] E. Matlis, *Commutative coherent rings*, Canad. J. Math. **34** (1982), 1240–1244.

[83] E. Matlis, *Commutative semicoherent and semiregular rings*, J. Algebra **95** (1985), 343–372.

[84] P.J. McCarthy, *The ring of polynomials over a von Neumann regular ring*, Proc. Amer. Math. Soc. **39** (1973), 253–254.

[85] P. Menal, *On the endomorphism ring of a free module*, Publ. Math. **27** (1982), 141–154.

[86] M. Nagata, *On the fourteenth problem of Hilbert*, Proc. Internat. Congress Math. (ICM), Cambridge Univ. Press, London (1960, 1958), 459–463.

[87] W.K. Nicholson and M.F. Yousif, *Principally injective rings*, J. Algebra **174** (1995), 66–93.

[88] B.L. Osofsky, *A generalization of quasi-Frobenius rings*, J. Algebra **4** (1966), 373–387.

[89] P. Pillay, *On semihereditary polynomial rings*, Proc. Amer. Math. Soc. **78** (1980), 533–536.

[90] M. Roitman, *On Mori domains and commutative rings with $CC^{\perp}$, I*, J. Pure Appl. Algebra **58** (1989), 247–258; *II*, ibid. **61** (1989), 63–77.

[91] M. Roitman, *On polynomial extensions over countable fields*, J. Pure Appl. Algebra **64** (1990), 315–328.

[92] E.A. Rutter, *A characterization of QF-1 rings*, Pacific J. Math. **51** (1974) 533–536.

[93] F.L. Sandomierski, *Some examples of right self-injective rings which are not left self-injective*, Proc. Amer. Math. Soc. **26** (1970), 373–387.

[94] P. Shizhong, *Commutative quasi-Frobenius rings*, Comm. Algebra **19** (1991), 663–667.

[95] R.C. Shock, *Polynomial rings over finite dimensional rings*, Pacific J. Math. **42** (1972), 251–258.

[96] J. Soublin, *Anneaux cohérents*, C. R. Acad. Sci. Ser. A-B **267** (1968), A183–A186.

[97] J. Soublin, *Un anneau cohérent dont l'anneau des polynômes n'est pas cohérent*, C. R. Acad. Sci. Ser. A-B **267** (1968), A241–243.

[98] B. Stenström, *Coherent rings, and FP-injective modules*, J. London Math. Soc. **2** (1970), 323–329.

[99] B. Stenström, *Rings of Quotients*, Grundl. Math. Wiss., Vol. 217, Springer, Berlin (1975).

[100] Y. Utumi, *On rings of which any one-sided quotient ring is two-sided*, Proc. Amer. Math. Soc. **14** (1963), 141–147.

[101] Y. Utumi, *A note on rings of which any one-sided ring is two-sided*, Proc. Japan Acad. **39** (1963), 287–288.

[102] Y. Utumi, *Self-injective rings*, J. Algebra **6** (1967), 56–64.

[103] P. Vámos, *Classical rings*, J. Algebra **34** (1975), 114–129.

[104] W. Vasconcelos, *Rings of Dimension Two*, Marcel Dekker, Basel (1976).

[105] T. Würfel, *Über absolut reine Ringe*, J. Reine Angew. Math. **262/263** (1973), 381–391.

[106] W. Xue, *Rings with Morita Duality*, Lecture Notes in Math., Vol. 1523, Springer, Berlin (1992).

[107] W. Xue, *Rings related to quasi-Frobenius rings*, Algebra Colloq. **5** (1998) 471–480.

Hamilton's Quaternions

T.Y. Lam

University of California, Berkeley, CA 94720, USA
E-mail: lam@math.berkeley.edu

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1. A spark flashed forth

From a course in modern algebra, we learned that a division ring is a ring K with identity $1 \neq 0$ in which every nonzero element has an inverse. These rings deserve our close attention because in some sense they are the most "perfect" algebraic systems: in them, we can add, subtract, multiply, and divide (by nonzero elements), and with respect to these four basic operations, all of the usual rules of arithmetic hold – except possibly the commutativity of multiplication. If the commutative law for multiplication does hold in a division ring K, we say that K is a *field*. In other words, a division ring K is a field iff its multiplicative group K^* is abelian.

The first example of a division ring that is not a field was discovered by the great Irish mathematician Sir William Rowan Hamilton (1805–1865). This is the division ring of real quaternions generated by four basis elements $1, i, j, k$ over the real numbers $\mathbb{R}$, with the famous Hamiltonian relations

$$i^2 = j^2 = k^2 = ijk = -1. \tag{1.1}$$

Throughout this chapter, we shall denote this division ring by $\mathbb{H}$. It is easy to check from (1.1) that

$$ij = k = -ji, \quad jk = i = -kj, \quad \text{and} \quad ki = j = -ik. \tag{1.2}$$

In particular, the elements i, j, k pairwise anticommute. Using the rules (1.2), we can easily multiply out any pair of quaternions $a + bi + cj + dk$ and $a' + b'i + c'j + d'k$ by using the two distributive laws.

Hamilton's original motivation (coming from physics) was to find an algebraic formalism for the points (x_1, x_2, x_3) in 3-space, in generalization of the formalism of the complex numbers $\mathbb{C}$ as pairs of real numbers. His initial desire was thus to find a 3-dimensional "hypercomplex system" with all the right properties (e.g., containing $\mathbb{C}$, and having $x_1^2 + x_2^2 + x_3^2$ as a multiplicative norm function). Hamilton worked off and on without success on this problem in the period 1830–1843. Two decades later, reminiscing on this frustrating experience, he was to write in one of his letters to his son Archibald H. Hamilton:

> "*Every morning…, on my coming down to breakfast, your (then) little brother William Edwin, and yourself, used to ask me, 'Well, Papa, can you multiply triplets?'. Whereto I was always obliged to reply, with a sad shake of the head: 'No, I can only add and subtract them'.*"

Of course, nowadays, any student of modern algebra would be able to see that Hamilton's effort was doomed from the start on several grounds. First, the system Hamilton tried to find was to contain properly the complex field $\mathbb{C}$. By the transitivity formula for dimensions, we know that such a system can only have even real dimension. Second, the existence of a composition formula for sums of three squares would quickly contradict the fact that such sums are not closed under multiplication in $\mathbb{Z}$ and in $\mathbb{Q}$. For a detailed discussion on the history of this problem, see K.O. May's article [Ma].

The breakthrough finally came on Monday, October 16, 1843. On that day, as Hamilton took a walk with his wife along the Royal Canal in Dublin, on his way to a council meeting of the Royal Irish Academy, a sudden flash of genius led him to the *four*-dimensional system $\mathbb{H}$ with basis elements $1, i, j, k$ multiplied according to the laws (1.1) (and their consequences (1.2)). In Hamilton's own words (quoted from [Kl, p. 779]):

> "*I then and there felt the galvanic circuit of thought closed, and the sparks which fell from it were the fundamental equations between i, j and k, exactly such as I have used them ever since.*"

With Irish exuberance (borrowing a phrase of E.T. Bell), Hamilton took out his pocket knife, and carved his fundamental equations on a stone of the Brougham Bridge. Eqs. (1.1) and (1.2) have since appeared in countless books in mathematics, history of mathematics, as well as in many items commemorating the life and work of Sir William Hamilton, including a series of Irish postal stamps issued in 1983, one and a half century after Hamilton's discovery of the quaternions.

To see that $\mathbb{H}$ is a division ring, note that every quaternion $q = a + bi + cj + dk \in \mathbb{H}$ has a "norm"

$$\mathrm{N}(q) = a^2 + b^2 + c^2 + d^2 \in \mathbb{R}, \tag{1.3}$$

and a conjugate $\bar{q} = a - bi - cj - dk \in \mathbb{H}$. A quick calculation shows that $\mathrm{N}(q) = q\bar{q} = \bar{q}q$, so if $q \neq 0$, we have $\mathrm{N}(q) \neq 0$, and q^{-1} is given by $\bar{q}/\mathrm{N}(q)$. However, $\mathbb{H}$ is not a field since $ij \neq ji$ (and $jk \neq kj$, $ki \neq ik$). Thus, $\mathbb{H}$ gave historically the very first example of a *noncommutative* division ring.

Of Hamilton's monumental discovery of the quaternions, J.J. Sylvester wrote the following thoughtful passage:

> "*In Quaternions the example has been given of Algebra released from the yoke of the commutative principle of multiplication – an emancipation somewhat akin to Lobachevsky's of Geometry from Euclid's noted empirical axiom.*"

Sylvester's referral to Lobachevsky's non-Euclidean geometry provided an interesting analogy for Hamilton's invention of the quaternions. This analogy becomes even somewhat uncanny if we also bring Gauss into the picture. It is well reported in history books in mathematics that, while Nikolai Ivanovich Lobachevsky discovered non-Euclidean geometry in 1826 and published his findings between 1829–1837, and János Bolyai made similar discoveries in the period 1825–1833, Carl Friedrich Gauss (1777–1855) seemed to have been aware of the independence of Euclid's parallel postulate as early as 1799, and by 1813 he had arrived at the rudiments of a logically consistent "anti-Euclidean geometry". However, always cautious about how his work would be received, and fearful of the ridicule of his revolutionary ideas by his contemporaries, Gauss never published his findings. As for the quaternions, a rather similar situation prevailed, although it is not as widely reported. There is no doubt that Hamilton deserved full credit for his discovery of the quaternions in 1843, but again Gauss had already a good "sighting" of the quaternion system around 1819–1823. In a short note [Ga] from his diary, in working with transformations of spaces, Gauss came up with a way of composing real quadruples. Given (a, b, c, d) and $(\alpha, \beta, \gamma, \delta)$ in $\mathbb{R}^4$, he associated the composite quadruple (A, B, C, D) given by

$$\begin{aligned}(a\alpha - b\beta - c\gamma - d\delta,\ & a\beta + b\alpha - c\delta + d\gamma,\\ & a\gamma + b\delta + c\alpha - d\beta,\ a\delta - b\gamma + c\beta + d\alpha).\end{aligned} \tag{1.4}$$

To see that this is no less than the composition of quaternions, one need only transpose the second and third coordinates in Gauss's notation. That is, if we express the quadruples (a, b, c, d) and $(\alpha, \beta, \gamma, \delta)$ as $a + bj + ci + dk$ and $\alpha + \beta j + \gamma i + \delta k$ in Hamilton's notation, Gauss's composition amounts *exactly* to the multiplication of two arbitrary quaternions! Gauss wrote, in an almost matter-of-fact fashion:

> "*Wir bezeichnen allgemein die Combination* a, b, c, d *durch* (a, b, c, d) *und schreiben*
>
> $$(a, b, c, d)(\alpha, \beta, \gamma, \delta) = (A, B, C, D).$$
>
> *Es ist also* $(a, b, c, d)(\alpha, \beta, \gamma, \delta)$ *nicht mit* $(\alpha, \beta, \gamma, \delta)(a, b, c, d)$ *zu verwechseln. ... Ferner bezeichne man die Combination* (a, b, c, d) *durch einen Buchstaben, z.B. g, und dann die Combination* $(a, -b, -c, -d)$ *durch* g'. *Es ist also*
>
> $$gg' = g'g = (aa + bb + cc + dd, 0, 0, 0)."$$

Although quadruples of the form (A, B, C, D) in (1.4) had certainly appeared earlier, notably in Euler's 4-square identity, Gauss took the unmistakable step of viewing (A, B, C, D) in (1.4) as a *composition* of (a, b, c, d) and $(\alpha, \beta, \gamma, \delta)$, and explicitly noted the failure of the commutative law. And of course, the last remark in the quotation above was tantamount to the realization that, under such a composition, the quadruples (a, b, c, d)'s form a division algebra over the real numbers. Although Gauss did not introduce the i, j, k notation as Hamilton did, it would be difficult to deny, on the basis of what Gauss had written (so succinctly!) in his diary, that he had essentially "come up" with the quaternion system as early as 1819. But again, Gauss never published anything about this discovery, so the "official" birth of the quaternions was to wait another quarter of a century, until Hamilton took his famous walk along the Royal Canal on that eventful October day in 1843.

2. Great expectations [1]

Hamilton noted his discovery of the quaternions on the Council Books of the Royal Academy on the very same October day, and obtained leave to read a paper on quaternions to

[1] Our pilfering of the title of Charles Dickens's famous novel here is not purely accidental. Dickens, who lived from 1812 to 1870, was almost an exact contemporary of Hamilton: it seems remarkable that the life spans of these two great men differed only by a translation of about six years. *Great Expectations* was written in 1861 – four years before Hamilton's death. And, if the reader is curious as to what Charles Dickens was up to in 1843, the answer is that he wrote *A Christmas Carol*, which has also endured the times very well, and remained a delightful part of the Christmas tradition in countries around the world.

a general meeting of the Academy on November 13, 1843. After this, he was to devote much of the remaining 22 years of his life to the study of quaternions and their applications. Besides the many papers he published on quaternions (one hundred and nine in all, according to Crowe [Cr, p. 41]), his main writings in the subject survived in his two books [H_1] and [H_2] (the latter published posthumously, first in 1866).

Hamilton had great expectations for the wide applicability of his quaternions to mathematics, physics, and astronomy. At one point, he even went so far as to compare his discovery to that of Newton on fluxions in the 17th century (see [Cr, p. 30]). In later years, his zeal in converting his contemporaries to the quaternionic thinking virtually verged on obsession. As it turned out, the controversy over the usefulness of quaternions was to last for several generations after Hamilton's death. Supporters of Hamilton touted the quaternions as one of the greatest inventions in the 19th century, "fitted to be of the greatest use in all parts of science", whereas antagonists of Hamilton's theory attacked the quaternions as "an unmixed evil to those who have touched them in anyway".[2] The only mathematical theory I can think of that has evoked among its practitioners such intense controversy and vigorous debate is perhaps, again, non-Euclidean geometry!

As is always the case, the truth is somewhere in between. In retrospect, Hamilton's original expectations were largely overblown, and indeed, a considerable part of the subsequent work on quaternions by his disciples and followers did not survive the time test. Nevertheless, ever since that October day in 1843, the quaternions have assumed their permanent place in mathematics as the first noncommutative system to be studied. Their bold introduction affirmed the supreme freedom of mathematical thought, and opened the floodgate to the fruitful study of algebraic systems satisfying different sets of axioms, not all of which reflect the familiar properties of the real (or the complex) numbers. For instance, much of the work in hypercomplex systems (the theory of algebras) in late 19th century and the early part of the 20th century was rightfully a direct descendant of Hamilton's theory of quaternions. (For a survey on this line of work, see [Hap].) Theoretically, to every concept making sense for the reals and the complexes, there is a quaternionic analogue. Whether such a quaternionic analogue is truly worthwhile of study can only be seen in time. For instance, the efforts of Hamilton's followers in developing a quaternionic theory of functions (complete with differentiation, integration, and theorems of Gauss, Green and Stokes) are now largely forgotten. But in contemporary topology, there is certainly still a healthy interest in quaternionic manifolds. A recent search on the MathSciNet for papers with the word "quaternion" in their titles turned up 2073 entries. Quaternions did not become a main tool in physics as Hamilton had hoped. But late in the 20th century, there seemed to have been a revival of interest in the quaternions among some physicists, as they tried to formulate quaternionic generalizations of the postulates of quantum mechanics, quantum field theory, gravitational theory, and elementary particle symmetries. For some work in this area, see, for instance, [Ad].

---

[2] For more complete original quotations and attributions to their sources, we refer the readers to [KR_1], [Al_2] and [Cr, Ch. 2].

3. Matrix models

The main purpose of this section is to discuss a matrix model of the quaternion algebra $\mathbb{H}$. This discussion will also help us to solidify our understanding of the quaternions before we return to look at other historical aspects pertaining to them.

Nowadays every college student in mathematics knows that matrices do not commute under multiplication, so the idea of a noncommutative system comes as no surprise. But matrices were not yet in the mathematical vocabulary in 1843. They were introduced only in 1855 by Arthur Cayley (1821–1895) who subsequently published his famous memoir [Ca] on the subject in 1858. For a detailed discussion on the invention of the theory of matrices, see [Kl, Ch. 33, §4].

From the viewpoint of matrices, the complex field $\mathbb{C}$ can be realized as a subalgebra of the matrix algebra $\mathbb{M}_2(\mathbb{R})$ as follows. We view $\mathbb{C}$ as a vector space over $\mathbb{R}$ with basis $\{1, -i\}$, and "identify" a complex number $a+bi$ $(a, b \in \mathbb{R})$ with the left multiplication map by $a+bi$ on $\mathbb{C}$ (the Cayley representation). With respect to the basis $\{1, -i\}$, this linear map has a matrix $\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$, so $\mathbb{C}$ may be identified with the algebra of all such matrices in $\mathbb{M}_2(\mathbb{R})$.

We can use a similar procedure to realize the quaternions as a certain real subalgebra of 2×2 *complex* matrices. To this end, let us identify the complex numbers $\mathbb{C}$ with $\mathbb{R} \oplus \mathbb{R}i$ in $\mathbb{H}$, and work with $\mathbb{H}$ as a 2-dimensional right vector space over $\mathbb{C}$ with basis $\{1, -j\}$. By the Cayley representation again, every quaternion q induces a $\mathbb{C}$-linear left multiplication map on $\mathbb{H}$, which can be expressed by a 2×2 complex matrix $L(q) \in \mathbb{M}_2(\mathbb{C})$. For instance, from $i \cdot 1 = i$ and $i \cdot (-j) = (-j)(-i)$, we have

$$\begin{aligned} L(i) &= \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad \text{and similarly,} \\ L(j) &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad L(k) = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}. \end{aligned} \tag{3.1}$$

From these, we compute easily that, for a quaternion $q = a + bi + cj + dk$,

$$\begin{aligned} L(q) = \begin{pmatrix} a+bi & c+di \\ -c+di & a-bi \end{pmatrix} &= \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix}, \\ \text{where } \alpha = a+bi, \ \text{and } \beta &= c+di. \end{aligned} \tag{3.2}$$

Since the Cayley representation is faithful ($L(v) = 0 \Rightarrow 0 = L(v)(1) = v$), we obtain an $\mathbb{R}$-algebra isomorphism

$$\mathbb{H} \cong \left\{ \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix} : \alpha, \beta \in \mathbb{C} \right\} \subseteq \mathbb{M}_2(\mathbb{C}), \tag{3.3}$$

which is given in many textbooks to describe (and sometimes to define) the division ring of real quaternions. Note that this "Cayley model" of $\mathbb{H}$ makes it unnecessary to verify the associative law for multiplication (since matrix multiplication is always associative). Also, $L(i), L(j)$ and $L(k)$ in (3.1) are now just three concrete anticommuting matrices over $\mathbb{C}$ (with squares $-I_2$ and product I_2): they no longer have the aura of mystery of the quaternions i, j and k.

The matrix model for $\mathbb{H}$ in (3.3) has several wonderful features. First, in this model, quaternionic conjugation $q \mapsto \bar{q}$ corresponds to the "conjugate transpose" operation * on $\mathbb{M}_2(\mathbb{C})$. More precisely,

$$L(\bar{q}) = \begin{pmatrix} a-bi & -(c+di) \\ c-di & a+bi \end{pmatrix} = \begin{pmatrix} a+bi & c+di \\ -c+di & a-bi \end{pmatrix}^* = L(q)^*. \tag{3.4}$$

Secondly, the quaternionic norm corresponds to the determinant on $\mathbb{M}_2(\mathbb{C})$; namely,

$$\mathrm{N}(q) = a^2+b^2+c^2+d^2 = \det\bigl(L(q)\bigr). \tag{3.5}$$

The fact that "det" is multiplicative implies the same for the quaternion norm,[3] and this gives essentially the four-square identity of Euler. Thirdly, it is easy to verify that the matrices $L(1), L(i), L(j), L(k)$ are linearly independent not only over $\mathbb{R}$, but also over $\mathbb{C}$. Thus, they form a $\mathbb{C}$-basis for $\mathbb{M}_2(\mathbb{C})$. From this, we get a $\mathbb{C}$-algebra isomorphism

$$\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H} \cong \mathbb{M}_2(\mathbb{C}). \tag{3.6}$$

Here, $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H}$ denotes the "scalar extension" of the quaternion algebra from $\mathbb{R}$ to $\mathbb{C}$. In other words, it is the quaternion algebra formed over the complexes; in Hamilton's work, this is called the *algebra of biquaternions.*[4] As a $\mathbb{C}$-algebra, it turns out to be isomorphic to $\mathbb{M}_2(\mathbb{C})$. (In modern parlance, we say that $\mathbb{H}$ "splits" over $\mathbb{C}$.)

It has been said that most of the applications of quaternions to physics are actually applications of biquaternions. There is perhaps considerable truth in this statement. Note that the matrices $L(i), L(j), L(k)$ in (3.1) are *unitary* matrices (that is, complex matrices U with $UU^* = I$). If we multiply them by the scalar $-i$, we get the following three *Hermitian* matrices (that is, complex matrices H such that $H^* = H$):

$$\sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{3.7}$$

These are the famous *Pauli spin matrices*, which are used by physicists in the study of the quantum mechanical motion of a spinning electron. The following relations among the Pauli spin matrices are familiar to all quantum physicists (see, e.g., [Me, p. 545]):

$$\begin{aligned} &\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = I, \qquad \sigma_x\sigma_y\sigma_z = iI, \\ &\sigma_x\sigma_y = -\sigma_y\sigma_x = i\sigma_z, \qquad \sigma_y\sigma_z = -\sigma_z\sigma_y = i\sigma_x, \qquad \sigma_z\sigma_x = -\sigma_x\sigma_z = i\sigma_y. \end{aligned}$$

These are to be compared with the original Hamiltonian relations (1.1) and (1.2).

[3] Of course, we don't really have to use the matrix model to prove the multiplicativity of N. Since conjugation on $\mathbb{H}$ is easily seen to be an involution, we have $\mathrm{N}(pq) = (pq)\overline{pq} = pq \cdot \bar{q}\,\bar{p} = p\mathrm{N}(q)\bar{p} = \mathrm{N}(p)\mathrm{N}(q)$.

[4] The word "biquaternion", however, may have a different meaning in other contexts. For instance, William Kingdon Clifford (1845–1879) used the same word for the elements in one of his hypercomplex number systems. In this chapter, we have, unfortunately, no space to discuss Clifford's contributions.

In the above, we have viewed $\mathbb{H}$ as a (right) $\mathbb{C}$-space so as to keep the matrices at size 2×2. If we simply view $\mathbb{H}$ as an $\mathbb{R}$-vector space with basis $\{1, i, j, k\}$, the Cayley representation would have given the 4×4 real matrix model:

$$\mathbb{H} \cong \left\{ \begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix} : a, b, c, d \in \mathbb{R} \right\} \subseteq \mathbb{M}_4(\mathbb{R}), \tag{3.8}$$

with the above matrix corresponding to the quaternion $q = a + bi + cj + dk$. Here, the quaternionic conjugate $q \mapsto \bar{q}$ corresponds to the ordinary transpose on $\mathbb{M}_4(\mathbb{R})$, while the determinant of the matrix above is now $\mathrm{N}(q)^2$.

4. Related mathematical discoveries

To place Hamilton's discovery of the quaternions in the proper perspective of 19th (and 20th) century mathematics, it would be incumbent on us to describe several other important mathematical events which took place after 1843, and which turned out to have had a direct impact on the role of quaternions in mathematics (and physics). We itemize several of these relevant developments in the following.

• **Cayley Algebra.** Only a few months after Hamilton's discovery of the quaternions, his friend John T. Graves invented the 8-dimensional hypercomplex system $\mathbb{O}$, now known as the *algebra of octonions* (or *octaves*). Like $\mathbb{H}$, $\mathbb{O}$ is a real division algebra (every nonzero element has an inverse), and is not commutative. But unlike $\mathbb{H}$, $\mathbb{O}$ is no longer associative! Thus, a second bondage to the existing fundamental laws of algebra was quickly broken following on the heels of the first.

Although Graves made his discovery in 1843–44, his results were not published until 1848. In the mean time, Arthur Cayley made the same discovery and published it in 1845. From then on, the octonions also became known as the *Cayley numbers.* The quickest construction of the Graves–Cayley algebra results from defining the following multiplication on the additive group $\mathbb{O} := \mathbb{H} \times \mathbb{H}$:

$$\begin{aligned} &(p_1, q_1)(p_2, q_2) = (p_1 p_2 - \bar{q}_2 q_1,\ q_2 p_1 + q_1 \bar{p}_2), \\ &\qquad \text{where } (p_i, q_i) \in \mathbb{H} \times \mathbb{H}. \end{aligned} \tag{4.1}$$

It is straightforward to check that, with this multiplication, $\mathbb{O}$ is indeed an $\mathbb{R}$-algebra, except for the lack of associativity, which can be seen from, say:

$$\begin{aligned} &\big[(i,0)(j,0)\big](0,k) = (k,0)(0,k) = (0,-1), \\ &(i,0)\big[(j,0)(0,k)\big] = (i,0)(0,-i) = (0,1). \end{aligned}$$

The definition of the multiplication in (4.1) is directly inspired by the following view of quaternion multiplications. Every quaternion

$$q = a + bi + cj + dk = (a + bi) + (c + di)j \in \mathbb{H}$$

can be "identified" with the pair $(\alpha, \beta) \in \mathbb{C}^2$, where $\alpha = a + bi$ and $\beta = c + di$ constitute the first row of the representing matrix $L(q)$ in (3.2). Taking the product of two such matrices and reading off the first row, we see that the multiplication of the quaternions as pairs of complex numbers is given by

$$(\alpha_1, \beta_1)(\alpha_2, \beta_2) = (\alpha_1\alpha_2 - \beta_1\bar{\beta}_2,\ \alpha_1\beta_2 + \beta_1\bar{\alpha}_2). \tag{4.2}$$

If we rewrite the right-hand side as $(\alpha_1\alpha_2 - \bar{\beta}_2\beta_1,\ \beta_2\alpha_1 + \beta_1\bar{\alpha}_2)$ (here the α's and the β's commute), we get exactly the formula (4.1). The point is, however, that (4.1) *has to be carefully adapted from* (4.2), since the p's and the q's no longer commute. If one makes a wrong choice in the orders of the factors in (4.1), one will not get the Cayley algebra $\mathbb{O}$.

Every Cayley number $x = (p, q) \in \mathbb{O}$ has a norm $\mathrm{N}(x) = \mathrm{N}(p) + \mathrm{N}(q)$, and a conjugate $\bar{x} = (\bar{p}, -q)$. One has again $\mathrm{N}(x) = x\bar{x} = \bar{x}x$, so if $x \neq 0$, it has an inverse $\bar{x}/\mathrm{N}(x)$, just as in the case of quaternions.[5] This shows that $\mathbb{O}$ is a (nonassociative) division algebra. Unfortunately, the "duplication" process *cannot* be used on the Cayley numbers again to get a 16-dimensional division algebra. Our luck runs out at dimension 8!

Although $\mathbb{O}$ is not associative, it can be shown to satisfy the following two weaker versions of the associative law:

$$x(xy) = (xx)y, \qquad (xy)y = x(yy), \tag{4.3}$$

for any $x, y \in \mathbb{O}$. An algebra satisfying these two laws is called an *alternative algebra*. Along with such algebras, various other kinds of nonassociative algebras have been studied, notably, the Lie algebras, and the Jordan algebras. Even within the framework of associative division algebras, one can try to get new objects by weakening some other axioms. For instance, if one keeps only one distributive law $x(y + z) = xy + xz$ and discards the other ($(y + z)x = yx + zx$), one gets a system called a *near field*. The investigations of these diverse systems were all made possible by Hamilton's pioneering discovery of the quaternions.

For more information on Cayley numbers, see [Ba], [Bl] and [W].

• **Frobenius' Theorem.** In a *tour de force* in 1877, F.G. Frobenius (1849–1917) proved that, up to isomorphisms, the only finite-dimensional real (associative) division algebras are $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{H}$ [Fr]. The same theorem was also proved independently (if just a little bit later) by C.S. Peirce, the son of Benjamin Peirce, who published the result in an Appendix to his father's long memoir [Pe] on associative algebras in 1881. This is a truly marvelous theorem as far as the quaternions are concerned, since it showed that, beyond $\mathbb{R}$ and $\mathbb{C}$, $\mathbb{H}$

[5] As the reader can perhaps guess, $\mathrm{N}(xy) = \mathrm{N}(x)\mathrm{N}(y)$ also holds for Cayley numbers, so that we get an 8-square identity. However, the proof for multiplicativity given for quaternions in footnote 3 no longer works, since the associative law does not hold here! We invite the reader to supply a correct proof for the case of Cayley numbers.

is in fact *the only* finite-dimensional real division algebra that can exist. This wonderful affirmation of the "cosmic" role of the quaternions in mathematics would seem to have at least partly fulfilled Hamilton's "great expectations", although unfortunately, it did not seem to have mattered very much in the great quaternion debate still raging on in the last decades of the 19th century. The anti-quaternionists were bent on arguing that quaternions were useless, whether they were shown to have any cosmic significance or not.

Nowadays, mathematicians certainly have a greater appreciation for uniqueness theorems. Frobenius' theorem was perhaps one of the earliest classification results of its kind proved in algebra. This theorem has been imitated many times over in the 20th century in different contexts of classification. Most notable examples are: Zorn's theorem (1933) on alternative real division algebras, the Gel'fand–Mazur theorem (1938) for commutative Banach division algebras, and Hopf's theorem (1940) on finite-dimensional commutative (but not necessarily associative) real division algebras. Hopf, a pioneer in applying topology to algebra, was the first one to see the possibility of proving results on division algebras from the geometry of spheres and projective spaces. His program in this direction was ultimately completed by Milnor and Kervaire. who proved independently in 1958 that *the unit sphere* S^{n-1} *is parallelizable*[6] *only if* $n = 1, 2, 4, 8$. An algebraic implication of this topological theorem is that the only possible finite dimensions of a (possibly noncommutative and nonassociative) real division algebra are $1, 2, 4$ or 8. This is a powerful expansion of Frobenius's 1877 theorem; however, up to this time, no purely algebraic proof of it is known. (This could very well be the most painful instance of the "topological thorn in the flesh of algebra" that Koecher and Remmert spoke about in the introduction to their article [KR_2, p. 223].) For an exposition on the topological methods used in the classification of division algebras, the reader may consult Hirzebruch's article [Hi]. For a modern formulation of a proof of Frobenius's theorem, see [PS, (3.4)], or [La_2, (13.12)].

• **Vector Analysis.** The fast-paced progress in physics (especially in mechanics, electricity and magnetism) in the second half of the 19th century clearly called for a vectorial theory for the 3-space that could serve as a sound mathematical foundation for the study of the physical quantities. This was indeed very much on Hamilton's mind when he first set out to find a product operation for triplets. After Hamilton invented the quaternions $\mathbb{H}$, the space of "pure quaternions"

$$\mathbb{H}_0 := \mathbb{R}i \oplus \mathbb{R}j \oplus \mathbb{R}k \tag{4.4}$$

would seem to be a natural candidate for a model for the 3-space. However, for any $\mathbf{u} = u_1 i + u_2 j + u_3 k \in \mathbb{H}_0$, one has $\mathbf{u}^2 = -(u_1^2 + u_2^2 + u_3^2)$, which is a negative real number. It seemed counter-intuitive to physicists that the square of a vector should be a negative quantity, and this no doubt hindered the wide acceptance of the quaternion system by the physicists.

Only one year after Hamilton's discovery, Hermann Grassmann (1809–77) published his "Ausdehnungslehre 1844", which purported to provide not only a firm foundation for the linear algebra of space, but in fact a whole new discipline of mathematics. In retrospect,

[6] A differentiable manifold is said to be *parallelizable* if it has a trivial tangent bundle.

Grassmann's book was truly a ground-breaking masterpiece, and one of the most brilliant mathematical works of the 19th century. But his "linear extension theory" (complete with a myriad of all kinds of vector products) was couched in such general and abstract terms, and so laden with metaphysical overtones, that it came off as almost totally incomprehensible to his contemporaries. Gauss politely declined to read "Ausdehnungslehre", and Möbius openly admitted that he had not managed to get through more than "a few sheets" (see [Cr, p. 80]). The book essentially met with complete silence, and sold so poorly that by 1864 its publisher saw fit to shred about 600 of its remaining copies! A second edition of the book met with a somewhat better reception, but the ultimate recognition of Grassmann's genius would have to wait until the 20th century.

Grassmann's work certainly did not succeed in providing physicists with the vectorial system they felt comfortable to use in physical theories. This task was accomplished around the 1880s by Josiah Willard Gibbs (1839–1903) and Oliver Heaviside (1850–1925), whose system of "modern vector analysis" relatively quickly gained acceptance and popularity. Central among the concepts in this theory are the *inner product* and the *cross product* of vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$, which we shall denote by $\langle \mathbf{u}, \mathbf{v} \rangle$ and $\mathbf{u} \times \mathbf{v}$. The former is a number, and the latter is a vector. The cross product is sometimes also called the outer product or the vector product. Grassmann himself had certainly defined inner and outer products before; however, his outer product for two vectors was not a vector, but an oriented area.

To a modern reader, the connection between quaternions and vector analysis is easy to discern. In view of (4.4), we can identify $\mathbb{H}_0$ with $\mathbb{R}^3$ (calling its elements "vectors"), and express $\mathbb{H}$ as a direct sum $\mathbb{R} \oplus \mathbb{H}_0$. Thus, every quaternion q is uniquely a sum $u_0 + \mathbf{u}$, where $u_0 \in \mathbb{R}$, and $\mathbf{u} \in \mathbb{H}_0 = \mathbb{R}^3$. Hamilton called u_0 the "scalar part", and $\mathbf{u}$ the "vector part", of the quaternion q, and explicitly used the notations: $S.q = u_0$, and $V.q = \mathbf{u}$. (This may well have been the very first usage of the words "scalar" and "vector" in the mathematical literature.) For any two "vectors" $\mathbf{u}, \mathbf{v} \in \mathbb{H}_0$, the quaternion product of $\mathbf{u}$ and $\mathbf{v}$ is easily computed to be

$$\mathbf{uv} = -\langle \mathbf{u}, \mathbf{v} \rangle + \mathbf{u} \times \mathbf{v}. \tag{4.5}$$

Thus, in Hamilton's notation, $S.\mathbf{uv} = -\langle \mathbf{u}, \mathbf{v} \rangle$, and $V.\mathbf{uv} = \mathbf{u} \times \mathbf{v}$. From these, we see that $\mathbf{uv}$ is in general not a "vector"; in fact, it is a vector iff $\mathbf{u}, \mathbf{v}$ are *perpendicular* in $\mathbb{R}^3$, in which case we have $\mathbf{uv} = \mathbf{u} \times \mathbf{v}$. From (4.5), we also see that the inner and outer products for vectors can be quickly retrieved from the quaternion products, namely:

$$\langle \mathbf{u}, \mathbf{v} \rangle = -(\mathbf{uv} + \mathbf{vu})/2 \quad \text{and} \quad \mathbf{u} \times \mathbf{v} = (\mathbf{uv} - \mathbf{vu})/2. \tag{4.6}$$

The advantages of the inner and outer products are clear and many. First and foremost, they are well grounded in physics. Second, the mixed triple product

$$\langle \mathbf{u} \times \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \times \mathbf{w} \rangle \tag{4.7}$$

corresponds to $\det(\mathbf{u}, \mathbf{v}, \mathbf{w})$, and thus detects linear dependence. Third, the cross product is a closed binary operation on $\mathbb{R}^3$. So we do get a multiplication for triplets! Under the cross product, the elements $\{i, j, k\}$ multiply with each other exactly like the quaternions, but now

$$i \times i = j \times j = k \times k = 0. \tag{4.8}$$

It can be checked that the cross product turns $\mathbb{R}^3$ into a *Lie algebra*. From the Cayley model (3.3), one can further show that this is isomorphic to the Lie algebra of all 2×2 skew-Hermitian matrices with trace zero over $\mathbb{C}$ (under the Lie bracket $[A, B] = (AB - BA)/2$).

Central to the Gibbs–Heaviside vector calculus were the three operators "grad", "div", and "curl". However, just like the inner and outer products, these operators were also more or less already "embedded" in Hamilton's quaternion system. Indeed, Hamilton was the first one to introduce (and to name) the "nabla" operator

$$\nabla = i\frac{\partial}{\partial x} + j\frac{\partial}{\partial y} + k\frac{\partial}{\partial z}, \tag{4.9}$$

which, when applied to a differentiable scalar function $f(x, y, z)$, yields the gradient field grad$f$. If we apply "nabla" instead to a vector field $\mathbf{F} = u(x, y, z)i + v(x, y, z)j + w(x, y, z)k$, then the use of quaternion multiplications leads to

$$\begin{aligned}\nabla\mathbf{F} &= -\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) + \left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}\right)i \\ &\quad + \left(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}\right)j + \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)k \\ &= -\operatorname{div}\mathbf{F} + \operatorname{curl}\mathbf{F},\end{aligned}$$

so "div" and "curl" come into the picture simultaneously.[7] If we apply this formula to a gradient field $\mathbf{F} = \operatorname{grad} f$, then the i, j, k terms will all cancel out (because of the interchangeability of the order of partial differentiations), and we will be left with:

$$\nabla^2 f = -\left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}\right); \tag{4.10}$$

that is, we get the *negative* of the famous Laplace operator "Δ" in potential theory.

On the basis of (4.5), (4.6), and the above discussions on "nabla", a liberal thinker today could very well say that vector algebra in $\mathbb{R}^3$ and the quaternions are "interchangeable systems". But they were definitely *not* the same in the eye of 19th century mathematicians, and it is a fact that Hamilton himself never explicitly introduced inner, outer products, or "div" and "curl". Gibbs had always maintained that these constructs were best studied on their own, and that it would not add anything to work with them as a part of the quaternionic structure. Heaviside held a similar view, priding himself on the fact that he never had to use a single quaternion in his own work. But ironically, it is perhaps still the following quotation from Heaviside [Cr, p. 192] that best summed up the situation:

[7] Note that, if we had used the cross products instead, then in view of (4.8), there would be no scalar term in this calculation, and we would have gotten the more familiar calculus formula $\nabla \times \mathbf{F} = \operatorname{curl}\mathbf{F}$.

"The invention of quaternions must be regarded as a most remarkable feat of human ingenuity. Vector analysis, without quaternions, could have been found by any mathematician by carefully examining the mechanics of the Cartesian mathematics; but to find out quaternions required a genius."

• **Composition of Sums of Squares.** We have pointed out before that the multiplicativity of the norm on the quaternions and on the Cayley numbers gave natural interpretations (and derivations) for the 4- and 8-square identities, just as the multiplicativity of complex number norms did for the 2-square identity. *How about n-square identities for other values of n?* Adolf Hurwitz (1859–1919) took up this question in 1888 and arrived at its complete solution [Hu], showing that, if there exists an identity

$$\left(x_1^2+\cdots+x_n^2\right)\left(y_1^2+\cdots+y_n^2\right)=z_1^2+\cdots+z_n^2, \tag{4.11}$$

where $z_1,\ldots,z_n \in \mathbb{R}[\mathbf{x},\mathbf{y}]$, then $n=1,2,4,8$! Later, Hurwitz and Radon considered the more general problem of finding, for a given n, the largest possible m for which there exists an identity

$$\left(x_1^2+\cdots+x_m^2\right)\left(y_1^2+\cdots+y_n^2\right)=z_1^2+\cdots+z_n^2, \tag{4.12}$$

where $z_1,\ldots,z_n \in \mathbb{R}[\mathbf{x},\mathbf{y}]$. The answer here is that $m=\rho(n)$, where ρ is the Hurwitz–Radon function, defined by

$$\rho(n)=8a+2^b \quad \text{for } n=2^{4a+b}n_0 \ (n_0=\text{odd},\ b\in\{0,1,2,3\}). \tag{4.13}$$

By easy arithmetic, we have $\rho(n)\leqslant n$, with equality iff $n=1,2,4,8$. Thus, the Hurwitz–Radon theorem contains Hurwitz's original theorem in 1888.

The work of Hurwitz–Radon is of great significance to topologists, since the square identities (4.12) for $m=\rho(n)$ can be used to construct readily $\rho(n)-1$ linearly independent tangent vector fields on the unit sphere S^{n-1}. Much later (in 1962), using deep tools from topology, J. Frank Adams proved that this is indeed the best construction possible; that is, S^{n-1} *cannot admit* $\rho(n)$ *linearly independent tangent vector fields* [A, Thm.1.1]. Adams's famous solution of the Vector Field Problem for Spheres implied, in particular, the aforementioned Milnor–Kervaire theorem that the only parallelizable spheres are S^0, S^1, S^3, and S^7.

The study of sums-of-squares identities prompted by the existence of the quaternions and Cayley numbers have led to a lot of fruitful mathematics, both in algebra and in topology. However, the following original form of the sums-of-squares problem proposed by Hurwitz [Hu] in 1888 has remained unsolved to this date: *for a given integer n, what is the smallest k for which* $(x_1^2+\cdots+x_n^2)(y_1^2+\cdots+y_n^2)$ *is a sum of k squares in* $\mathbb{R}[\mathbf{x},\mathbf{y}]$? For a thorough survey on this Hurwitz problem and many other related problems, see Shapiro's monograph [Sh].

• **Fundamental Theorem of Algebra.** This important theorem says that $\mathbb{C}$ is an algebraically closed field; that is, any non-constant polynomial in $\mathbb{C}[x]$ has a zero in $\mathbb{C}$. Indeed, the main point of passing from $\mathbb{R}$ to $\mathbb{C}$ is exactly to get an "algebraic closure" of the reals. In this spirit, one may ask: *Is there an analogue of the Fundamental Theorem of Algebra for the quaternions?* To address this question properly, one must, however, first re-examine the notion of a "polynomial". The "usual" form of a polynomial over $\mathbb{H}$ is a (finite) expression of the kind $f(x) = \sum_i a_i x^i$, where $a_i \in \mathbb{H}$, and we can formally define the evaluation of f at $q \in \mathbb{H}$ to be $f(q) := \sum_i a_i q^i$. However, since $\mathbb{H}$ is not commutative, the above kind of polynomials is clearly not sufficiently general. For instance, one might legitimately consider a degree 2 expression of the shape $axbxc$ (where $a, b, c \in \mathbb{H}$), whose evaluation at $q \in \mathbb{H}$ would be $aqbqc$ (which *cannot* be reduced to $abcq^2$ due to the lack of commutativity).

To formulate a Fundamental Theorem of Algebra for $\mathbb{H}$, we shall, therefore, redefine a *monomial of degree* n to be an expression of the form

$$m(x) = a_0 x a_1 x \cdots a_{n-1} x a_n, \quad \text{where } a_i \in \mathbb{H}. \tag{4.14}$$

The evaluation $m(q)$ is then defined as $a_0 q a_1 q \cdots a_{n-1} q a_n$. A "polynomial" $f(x)$ is redefined to be a (finite) sum of monomials $m_i(x)$, and $f(q)$ is accordingly defined to be $\sum_i m_i(q)$. The "degree" of f is the highest degree of the monomials actually appearing in f. Notice that, if $\deg(f) = n$, the nth degree part of $f(x)$ is now a sum of degree n monomials, instead of a single monomial. (Of course, the same holds for other homogeneous parts of f.)

In 1944, building upon earlier work of Niven, Eilenberg and Niven [EN] proved the following beautiful result.

(4.15) FUNDAMENTAL THEOREM OF ALGEBRA FOR QUATERNIONS. *Let $f(x)$ be a polynomial* (*in the generalized sense above*) *of degree $n \geqslant 1$ over $\mathbb{H}$, whose n-th degree part consists of a single monomial. Then the evaluation map $q \mapsto f(q)$ from $\mathbb{H}$ to $\mathbb{H}$ is surjective. In particular, f has a zero on $\mathbb{H}$.*

The only known proof of this theorem (so far) is again by an appeal to topology. We exploit the topology of the spheres, and think of S^4 as a compactification of $\mathbb{H} = \mathbb{R}^4$ by adding a "point at infinity". With f given as in (4.15), one observes that the evaluation map $q \mapsto f(q)$ can be extended to a *continuous* mapping $\hat{f} : S^4 \to S^4$ by taking $\hat{f}(\infty) = \infty$. This map can be shown to be of degree n in the *topological* sense; that is, its induced map in the 4th homology group multiplies a generator by n). Since $n \geqslant 1$, this map is onto, and therefore so is the original evaluation map $q \mapsto f(q)$ from $\mathbb{H}$ to $\mathbb{H}$.

In this theorem, the hypothesis on the highest degree term of $f(x)$ is a subtle one. Without this hypothesis, the theorem fails. For instance, for any nonzero $a \in \mathbb{H}$, any additive commutator $aq - qa$ has real part zero, and therefore the polynomial $f(x) = ax - xa + 1$ cannot have a zero. Here, the top degree part $ax - xa$ of f is not a single monomial.

It follows from (4.15), in particular, that any "ordinary" polynomial $f(x) = a_n x^n + \cdots + a_1 x + a_0$ ($a_i \in \mathbb{H}$, $a_n \neq 0$) always has a zero in $\mathbb{H}$, provided that $n \geqslant 1$. This special case of (4.15) is known as the Niven–Jacobson theorem. It can be proved without too much difficulty by pure algebra; see, for instance, [La$_2$, (16.14)]. From this theorem (and an appropriate Remainder Theorem), it follows that any polynomial $f \in \mathbb{H}[x]$ (as above) can be factored into a product of linear polynomials.

5. The rotation groups SO(3) and SO(4)

One of the most beautiful features of the quaternions is the role they play in the understanding and representation of the rotations of the low-dimensional Euclidean spaces. This piece of mathematics also goes back to Hamilton; we shall give an exposition of it here. Using modern notations, we write O(3) and SO(3) for the orthogonal group and the special orthogonal group of $\mathbb{R}^3 = \mathbb{H}_0$ (with respect to the Euclidean norm), that is, the group of norm-preserving linear automorphisms and its (normal) subgroup of automorphisms of determinant 1. Let us consider the group homomorphism

$$\varphi : \mathbb{H}^* \to \mathrm{O}(3), \quad \text{defined by} \quad \varphi(q)(\mathbf{v}) = q\mathbf{v}q^{-1} \ (\forall\, \mathbf{v} \in \mathbb{H}_0).^{8} \tag{5.1}$$

Here, $\varphi(q) \in \mathrm{O}(3)$ since $\mathrm{N}(q\mathbf{v}q^{-1}) = \mathrm{N}(\mathbf{v})$ for all "vectors" $\mathbf{v} \in \mathbb{H}_0$. It is easy to see that $\mathbb{R}$ is the center of $\mathbb{H}$; this implies readily that the kernel of φ is $\mathbb{R}^*$. In the following, we'll show that the image of φ is SO(3).

To accomplish our goal, we'll need some notations. For the rest of this paper, we'll write $\mathbb{H}_1$ for the multiplicative group of quaternions of norm 1. (We'll call these the *unit quaternions.*) For any vector $\mathbf{u} \in \mathbb{H}_0 \cap \mathbb{H}_1$, we denote by $\tau_{\mathbf{u}}$ the reflection of $\mathbb{H}_0$ with respect to the plane (through the origin) with unit normal $\mathbf{u}$, and by $\rho_{\mathbf{u}}^{\theta}$ the rotation of the 3-space $\mathbb{H}_0$ about the vector $\mathbf{u}$ (anti-clockwise) by the angle θ.

(5.2) THEOREM.

(1) *If* $\mathbf{u} \in \mathbb{H}_0 \cap \mathbb{H}_1$, *then* $\varphi(\mathbf{u}) = -\tau_{\mathbf{u}} = \rho_{\mathbf{u}}^{\pi}$.

(2) *Let* $\mathbf{u_1}, \mathbf{u}_2$ *be vectors in* $\mathbb{H}_0 \cap \mathbb{H}_1$, *making an angle* $\theta \in (0, \pi)$, *and let* $\mathbf{u}$ *be the unit vector in the direction of the cross product* $\mathbf{u}_1 \times \mathbf{u}_2$. *Then* $\tau_{\mathbf{u}_2}\tau_{\mathbf{u}_1} = \rho_{\mathbf{u}}^{2\theta}$.

(3) *For any* $\mathbf{u} \in \mathbb{H}_0 \cap \mathbb{H}_1$ *and any angle* θ, $\varphi(\cos\theta + (\sin\theta)\mathbf{u}) = \rho_{\mathbf{u}}^{2\theta}$.

(4) (Rodrigues' formula for rotations) *For* $\mathbf{u}$ *as in* (3), *and any vector* $\mathbf{v} \in \mathbb{H}_0$:

$$\rho_{\mathbf{u}}^{2\theta}(\mathbf{v}) = (\cos 2\theta)\mathbf{v} + (\sin 2\theta)(\mathbf{u} \times \mathbf{v}) + (1 - \cos 2\theta)\langle \mathbf{v}, \mathbf{u}\rangle \mathbf{u}.$$

(5) $\varphi(\mathbb{H}^*) = \varphi(\mathbb{H}_1) = \mathrm{SO}(3)$. *Thus, we have an exact sequence*

$$1 \to \mathbb{R}^* \to \mathbb{H}^* \xrightarrow{\varphi} \mathrm{SO}(3) \to 1,$$

though φ *is not a split epimorphism.*

PROOF. (1) Here, $\mathbf{u}^2 = -\mathbf{u}\bar{\mathbf{u}} = -\mathrm{N}(\mathbf{u}) = -1$, so (4.6) yields

$$\mathbf{uvu} = -\big(\mathbf{vu} + 2\langle \mathbf{v}, \mathbf{u}\rangle\big)\mathbf{u} = \mathbf{v} - 2\langle \mathbf{v}, \mathbf{u}\rangle \mathbf{u} \quad (\forall\, \mathbf{v} \in \mathbb{H}_0). \tag{5.3}$$

Thus, $\varphi(\mathbf{u})(\mathbf{v}) = -\mathbf{uvu} = 2\langle \mathbf{v}, \mathbf{u}\rangle \mathbf{u} - \mathbf{v} = -\tau_{\mathbf{u}}(\mathbf{v}) = \rho_{\mathbf{u}}^{\pi}(\mathbf{v})$, as claimed.

(2) This is a well-known geometric fact, for which we'll only give a brief proof. Note that $\mathbf{u}$ is on the line of intersection of the two planes (through the origin) normal to the

[8] We leave to the reader the easy task of checking that $\mathbf{v} \in \mathbb{H}_0 \Rightarrow q\mathbf{v}q^{-1} \in \mathbb{H}_0$.

vectors $\mathbf{u}_1$ and $\mathbf{u}_2$. With this observation, the proof can be then reduced to checking that, in the plane, the composition of reflections with respect to two lines intersecting at an angle θ is a rotation (anti-clockwise) by 2θ about the point of intersection. This is easily verified by drawing a picture.

(3) Without loss of generality, we may assume that the vector $\mathbf{u}$ is as in (2) above. We can thus carry out the proof here using the notations in (2). By the definition of inner products and cross products, we have $\langle \mathbf{u}_1, \mathbf{u}_2\rangle = \cos\theta$, and $\mathbf{u}_1 \times \mathbf{u}_2 = (\sin\theta)\mathbf{u} = -\mathbf{u}_2 \times \mathbf{u}_1$. Therefore, by (4.5),

$$-\mathbf{u}_2\mathbf{u}_1 = \langle \mathbf{u}_1, \mathbf{u}_2\rangle - \mathbf{u}_2 \times \mathbf{u}_1 = \cos\theta + (\sin\theta)\mathbf{u}. \tag{5.4}$$

From this, and from (1), (2) above, it follows that

$$\begin{aligned}\varphi\big(\cos\theta + (\sin\theta)\mathbf{u}\big) &= \varphi(-\mathbf{u}_2\mathbf{u}_1) = \varphi(\mathbf{u}_2)\varphi(\mathbf{u}_1)\\ &= (-\tau_{\mathbf{u}_2})(-\tau_{\mathbf{u}_1}) = \tau_{\mathbf{u}_2}\tau_{\mathbf{u}_1} = \rho_{\mathbf{u}}^{2\theta}.\end{aligned}$$

(4) Using (3), we can give the following quaternionic derivation of Rodrigues' formula:

$$\begin{aligned}\rho_{\mathbf{u}}^{2\theta}(\mathbf{v}) &= (\cos\theta + \mathbf{u}\sin\,\theta)\mathbf{v}(\cos\theta - \mathbf{u}\sin\theta)\\ &= (\cos^2\theta)\mathbf{v} + \cos\theta\sin\theta(\mathbf{uv} - \mathbf{vu}) - (\sin^2\theta)\mathbf{uvu}\\ &= (\cos^2\theta)\mathbf{v} + 2\sin\theta\cos\theta(\mathbf{u}\times\mathbf{v}) - (\sin^2\theta)\big(\mathbf{v} - 2\langle\mathbf{v},\mathbf{u}\rangle\mathbf{u}\big) \quad \text{(by (5.3))}\\ &= (\cos 2\theta)\mathbf{v} + (\sin 2\theta)(\mathbf{u}\times\mathbf{v}) + 2\sin^2\theta\langle\mathbf{v},\mathbf{u}\rangle\mathbf{u}\\ &= (\cos 2\theta)\mathbf{v} + (\sin 2\theta)(\mathbf{u}\times\mathbf{v}) + (1-\cos 2\theta)\langle\mathbf{v},\mathbf{u}\rangle\mathbf{u}.\end{aligned}$$

(5) Note that, for any $\mathbf{u} \in \mathbb{H}_0 \cap \mathbb{H}_1$, $\cos\theta + (\sin\theta)\mathbf{u}$ is a unit quaternion. Since any element in SO(3) is a rotation $\rho_{\mathbf{u}}^{2\theta}$, (3) above shows that $\mathrm{SO}(3) \subseteq \varphi(\mathbb{H}_1)$, and of course, we have $\varphi(\mathbb{H}_1) = \varphi(\mathbb{H}^*)$. The inclusion above will be an equality if we can show that *every* unit quaternion $q = a + bi + cj + dk$ can be expressed in the form $\cos\theta + (\sin\theta)\mathbf{u}$, where $\mathbf{u} \in \mathbb{H}_0 \cap \mathbb{H}_1$, and $\theta \in [0, \pi]$. Indeed, since $a^2 + b^2 + c^2 + d^2 = 1$, there exists a unique $\theta \in [0, \pi]$ such that $\cos\theta = a$. If $q \neq \pm 1$, then $\theta \in (0, \pi)$, so $b^2 + c^2 + d^2 = 1 - a^2 = \sin^2\theta > 0$, and we have

$$\begin{aligned}&q = \cos\theta + \mathbf{u}\sin\theta, \quad \text{uniquely for}\\ &\quad \mathbf{u} = (\sin\theta)^{-1}(bi + cj + dk) \in \mathbb{H}_0 \cap \mathbb{H}_1.\end{aligned} \tag{5.5}$$

This expression is called the *polar form of a unit quaternion* of the unit quaternion q. It is uniquely determined if $q \neq \pm 1$. (If $q = \pm 1$, we take $\theta = 0$ or π, and take $\mathbf{u}$ to be any unit vector.)

If φ splits, there would exist a subgroup $G \subset \mathbb{H}^*$ giving a complement to the central subgroup $\mathbb{R}^*$, and hence $\mathbb{H}^* = \mathbb{R}^* \times G$. This implies that G contains all commutators in $\mathbb{H}^*$; but then $-1 = i^{-1}j^{-1}ij \in G$, a contradiction. □

For a little bit of history: Olinde Rodrigues (1795–1851) was of Spanish ancestry, though he was born in Bordeaux. An amateur mathematician, he spent his life in France as an economist, banker, social reformer, and railroad tycoon. Besides the rotation formula in (4) above, his claim to fame was the "other" Rodrigues formula expressing the Legendre polynomials in terms of the higher derivatives of $(x^2-1)^n$, which apparently came from his doctoral dissertation. For more information on Rodrigues' life and work (especially his contribution to the theory of rotations), see [Al$_1$,Al$_2$].

Coming back to (5.2), it is of interest to note that the last part of this theorem can be used to give a rational parametrization of SO(3). In fact, if $q = a+bi+cj+dk \in \mathbb{H}^*$, then $q^{-1} = (a-bi-cj-dk)/n$ with $n = a^2+b^2+c^2+d^2$. The orthogonal transformation $\varphi(q)$ on $\mathbb{H}_0$ given by

$$\mathbf{v} \mapsto q\mathbf{v}q^{-1} = n^{-1}(a+bi+cj+dk)\mathbf{v}(a-bi-cj-dk) \tag{5.6}$$

has thus a matrix with entries of the form r_{ij}/n, where r_{ij} are real quadratic forms in a, b, c, d. After the calculation is carried out, we see that this matrix is

$$n^{-1}\cdot\begin{pmatrix} a^2+b^2-c^2-d^2 & -2ad+2bc & 2ac+2bd \\ 2ad+2bc & a^2-b^2+c^2-d^2 & -2ab+2cd \\ -2ac+2bd & 2ab+2cd & a^2-b^2-c^2+d^2 \end{pmatrix}. \tag{5.7}$$

Since $\varphi(\mathbb{H}^*) = \mathrm{SO}(3)$, (5.7) gives a rational parametrization of SO(3). This rational parametrization of SO(3) is exactly that discovered by Euler (using Euler angles) in 1770. (It is true that Euler did not work with matrices *per se*; however, the three columns of an SO(3) matrix are nothing more than a triad of unit vectors with a left-hand orientation.)

A nice way to think of the group $\mathbb{H}_1$ is the following. Using the matrix model (3.3), we can interpret $\mathbb{H}_1$ as the group of complex matrices $\begin{pmatrix} \alpha & \beta \\ -\bar\beta & \bar\alpha \end{pmatrix}$ of determinant 1. This is exactly the special unitary group SU(2). Making this identification, and restricting the homomorphism φ to $\mathbb{H}_1$, we get a new exact sequence

$$1 \to \{\pm I_2\} \to \mathrm{SU}(2) \xrightarrow{\varphi} \mathrm{SO}(3) \to 1, \tag{5.8}$$

where, again, φ is nonsplit. Here, we have a 2-fold fibration of compact Lie groups. The best way to understand the topological structure of $\mathrm{SU}(2) \approx \mathbb{H}_1$ is to think of it as the 3-sphere S^3 in the Euclidean 4-space $\mathbb{R}^4 = \mathbb{H}$. It is well known that S^3 is simply connected (that is, it has a trivial fundamental group), so the same holds for SU(2). Thus, via the epimorphism φ, SU(2) provides a (2-fold) *universal covering* of SO(3). This is called the *spin covering* of SO(3), and for this reason SU(2) is often called a "spin group" (and denoted by Spin(3)). The spin covering is useful for many purposes; for instance, a continuous representation of Spin(3) that is not trivial on $\{\pm 1\}$ can be considered as a double-valued "spin representation" of SO(3) (and all of this has generalizations to the higher dimensions).

Using the topological model S^3 for SU(2), we see, in addition, that the rotation group SO(3) can be identified with the real projective space $\mathbb{RP}^3$. (Note that, under the covering φ in (5.8), we are identifying a matrix in SU(2) with its negative. This corresponds to identifying a unit quaternion with its negative, which translates into identifying antipodal points on S^3.) The physicists, however, have always preferred working with SU(2) (to working with S^3) since they also have to deal with the higher unitary groups SU(3) and SU(4) in elementary particle physics.

As it turns out that, the quaternions can also be used to describe the group SO(4). The possibility of this was already observed by Cayley. We use $\mathbb{H}$ as a model for the 4-space, and the main point is that the usual inner product on $\mathbb{R}^4$ can be expressed by the pairing

$$(p, q) \mapsto B(p, q) := (p\bar{q} + q\bar{p})/2 \in \mathbb{R} \tag{5.9}$$

on $\mathbb{H}$. (This fact can be checked by an easy direct calculation.) Given this, we can already use quaternion multiplications to construct a large family of isometries of $\mathbb{R}^4$. For, if $x, y \in S^3 = \mathbb{H}_1$, the map $q \mapsto xq\bar{y}$ is an isometry on $(\mathbb{H}, B)$, as

$$\mathrm{N}(xq\bar{y}) = \mathrm{N}(x)\mathrm{N}(q)\mathrm{N}(\bar{y}) = \mathrm{N}(q).$$

Moreover, $q \mapsto xq$ and $q \mapsto q\bar{y}$ both have determinant 1, so the isometry constructed above is in SO(4). Cayley's beautiful result says that *all* isometries of $\mathbb{R}^4 = (\mathbb{H}, B)$ arise in this way.

(5.10) THEOREM. *Let $\psi : S^3 \times S^3 \to \mathrm{SO}(4)$ be defined by $\psi(x, y)(q) = xq\bar{y}$ (for all $q \in \mathbb{H}$). Then ψ is a group homomorphism, and we have an exact sequence*

$$1 \to \big\{\pm(1, 1)\big\} \to S^3 \times S^3 \xrightarrow{\psi} \mathrm{SO}(4) \to 1. \tag{5.11}$$

PROOF. Here, we continue to identify $\mathbb{H}_1$ with S^3, so the latter becomes a group (and so does $S^3 \times S^3$). It is routine to check that ψ is a group homomorphism. To compute its kernel, suppose $\psi(x, y) = 1$, where $x, y \in \mathbb{H}_1$. Then $1 = \psi(x, y)(1) = x\bar{y}$ implies that $x = \bar{y}^{-1} = y$, and we need to have

$$q = \psi(x, y)(q) = xq\bar{y} = xq\bar{x} = xqx^{-1} \quad \text{for all } q \in \mathbb{H}.$$

Since the center of $\mathbb{H}$ is $\mathbb{R}$, this means that $x \in \mathbb{R} \cap \mathbb{H}_1 = \{\pm 1\}$. Thus, $(x, y) = \pm(1, 1)$, as desired.

It only remains to prove that ψ is *onto*. The proof of this becomes quite easy if we assume the Cartan–Dieudonné theorem on isometries. According to (a weak version of) this theorem, any isometry on a finite-dimensional nonsingular symmetric bilinear space (over a field of characteristic $\neq 2$) is a product of hyperplane reflections (see [La$_1$, p. 27]). Thus, our group SO(4) is generated by "pair-products" of hyperplane reflections[9] $\tau_a\tau_b$, where a, b range over $\mathbb{H}_1 = S^3$. Therefore, it suffices to show that $\tau_a\tau_b \in \mathrm{im}(\psi)$. Now, using the pairing B in (5.9), we can explicitly compute the effect of τ_b as follows. For any $q \in \mathbb{H}$:

$$\tau_b(q) = q - 2B(q, b)b = q - (q\bar{b} + b\bar{q})b = -b\bar{q}b. \tag{5.12}$$

[9]As before, we use the notation τ_b for hyperplane reflections. Thus, here, τ_b denotes the reflection of $\mathbb{R}^4$ with respect to the 3-dimensional subspace orthogonal to b in the usual inner product.

Therefore, $\tau_a\tau_b(q) = \tau_a(-b\bar{q}b) = -a\overline{(-b\bar{q}b)}a = a\bar{b}q\bar{b}a$. Since this holds for all $q \in \mathbb{H}$, and $\bar{b}a = \overline{\bar{a}b}$, we have $\tau_a\tau_b = \psi(a\bar{b}, \bar{a}b)$, as desired. $\square$

In many ways, the group SO(4) turns out to be exceptional in the series of the orthogonal groups SO(n) ($n \geqslant 3$). In this series, SO(n) is a simple group when n is odd, and has a unique nontrivial normal subgroup given by its center $\{\pm I\}$ when n is even and $\neq 4$ (see Appendix II in [Di]). But for SO(4), the exact sequence (5.11) implies, remarkably, that SO(4) has a pair of nontrivial normal subgroups $\psi(S^3 \times \{1\})$ and $\psi(\{1\} \times S^3)$ (both isomorphic to S^3) intersecting at its center $\{\pm I\}$. This exceptional behavior of SO(4) (exhibited above by quaternion constructions) may have been one of the sources of the rather peculiar role played by 4-space in mathematics. As Dieudonné wrote in [Di, p. 172], "The claim that four-dimensional spaces are quite exceptional is no idle talk."

6. Finite groups of quaternions

For any field K, it is well known that any finite subgroup of the multiplicative group K^* is cyclic, so nothing much remains to be said about these finite subgroups. In the case of a *division ring* K, finding the finite subgroups of K^* is a much more interesting problem. Since $\mathbb{H}$ was the first known noncommutative division ring, it would seem particularly natural to try to find all finite subgroups of $\mathbb{H}^*$. However, this problem was not satisfactorily solved until 1940, when Coxeter came to the scene.

In his classical paper [Co], Coxeter not only determined the finite subgroups of $\mathbb{H}^*$, but also put these groups in the general context of what he called the *binary polyhedral groups*. Before stating this classification result, let us first introduce two general families of groups after Coxeter.

For integers ℓ, m, n such that $2 \leqslant \ell \leqslant m \leqslant n$, let (ℓ, m, n) and $\langle \ell, m, n\rangle$ denote, respectively, the following two groups defined by generators and relations:

$$\langle R, S, T \mid R^\ell = S^m = T^n = RST = 1\rangle; \tag{6.1}$$

$$\langle R, S, T \mid R^\ell = S^m = T^n = RST\rangle. \tag{6.2}$$

Note that the element RST is central in $\langle \ell, m, n\rangle$ (being a power of each of the generators), and so $\langle \ell, m, n\rangle / \langle RST\rangle \cong (\ell, m, n)$. Coxeter has proved that (ℓ, m, n) is finite exactly in the following four cases:

$$\text{(a) } (2,2,n), \quad \text{(b) } (2,3,3), \quad \text{(c) } (2,3,4), \quad \text{and} \quad \text{(d) } (2,3,5). \tag{6.3}$$

These groups are, respectively:

(a) the *dihedral groups* of order $2n$;
(b) the *tetrahedral group* of order 12, isomorphic to the alternating group A_4;
(c) the *octahedral group* of order 24, isomorphic to the symmetric group S_4; and
(d) the *icosahedral group* of order 60, isomorphic to the alternating group A_5.

These are the symmetry groups of the regular n-gon and the regular tetrahedron, octahedron and icosahedron, respectively, so they are subgroups of SO(3). It is well known that, up to conjugation, these "(regular) polyhedral groups" and the cyclic groups (generated by rotations by the angle $2\pi/n,\ n \geqslant 1$) are, in fact, *all finite subgroups* of SO(3).

In the four cases in (6.3), Coxeter showed that the element RST has order 2 in $\langle \ell, m, n\rangle$, so the latter groups are also finite, with order twice that of (ℓ, m, n). The resulting finite groups $\langle \ell, m, n\rangle$ are called the *binary polyhedral groups*, since they are 2-fold coverings of the (regular) polyhedral groups. Note, for instance,

$$\begin{aligned}\langle 2,2,n\rangle &= \langle R,S,T \mid R^2 = S^2 = T^n = RST\rangle \\ &= \langle S,T \mid (ST)^2 = S^2 = T^n\rangle \\ &= \langle S,T \mid T^{2n} = 1,\ S^2 = T^n,\ S^{-1}TS = T^{-1}\rangle;\end{aligned}$$

this is called the *generalized quaternion* (or *dicyclic*) *group* of order $4n$.

The name "generalized quaternion group" suggests that these groups (for $n \geqslant 2$) are strongly related to the quaternions. Indeed, a straightforward calculation shows that the two unit quaternions

$$s = j, \quad \text{and} \quad t = \cos(\pi/n) + i\sin(\pi/n) \tag{6.4}$$

generate a subgroup of $\mathbb{H}_1$ that is isomorphic to $\langle 2,2,n\rangle$ (by the isomorphism $s \leftrightarrow S,\ t \leftrightarrow T$). In the simplest case $n = 2$, we get the group generated by $\{i, j\}$, which is the ordinary quaternion group $\{\pm 1, \pm i, \pm j, \pm k\}$, familiar to all students in elementary group theory!

The next important step is to observe that the other three finite groups in the family $\langle \ell, m, n\rangle$, that is, the binary tetrahedral, octahedral, and icosahedral groups, can likewise be realized as subgroups of $\mathbb{H}_1$. For instance, we can check that the following set of 24 quaternions

$$\left\{\pm 1,\ \pm i,\ \pm j,\ \pm k,\ (\pm 1 \pm i \pm j \pm k)/2\right\} \subset \mathbb{H}_1 \subset \mathbb{H}^* \tag{6.5}$$

(where the signs are arbitrarily chosen) constitute a model of the binary tetrahedral group $\langle 2,3,3\rangle$, with an isomorphism given by

$$R \leftrightarrow i, \quad S \leftrightarrow (1 + i + j - k)/2, \quad \text{and} \quad T \leftrightarrow (1 + i + j + k)/2. \tag{6.6}$$

In the same spirit, we can also construct triplets of elements in $\mathbb{H}_1$ that generate subgroups of $\mathbb{H}_1$ providing models for the binary octahedral and binary icosahedral groups. We shall only give the "answers" here:

$$R \leftrightarrow (i + j)/\sqrt{2}, \quad S \leftrightarrow (1 + i + j + k)/2, \quad \text{and} \quad T \leftrightarrow (1 + i)/\sqrt{2}, \tag{6.7}$$

$$R \leftrightarrow i, \quad S \leftrightarrow (1 + \tau i - \tau^{-1}k)/2, \quad \text{and} \quad T \leftrightarrow (\tau + i + \tau^{-1}j)/2, \tag{6.8}$$

where, in (6.8), τ denotes the "golden ratio" $(1 + \sqrt{5})/2$. These explicit constructions came from [Co], except that, curiously enough, there was a mistake in [Co, p. 371] in the modeling of the binary icosahedral group! Our proposed correction is based on (7.6) (3) below.[10]

A more conceptual way to see that the binary polyhedral groups occur as subgroups of $\mathbb{H}_1$ is to use the spin covering $\varphi : \mathbb{H}_1 \to \mathrm{SO}(3)$ with kernel $\{\pm 1\}$ (constructed in §5). Fixing copies of the finite polyhedral groups in SO(3), we can form the preimages of these groups under the covering φ. The resulting groups in $\mathbb{H}_1$, twice the size of the polyhedral groups, turn out to be isomorphic to the binary polyhedral groups, though this statement would require a careful proof.

Given the above constructions, and assuming the fact stated earlier on the classification of finite subgroups in SO(3), it is now relatively easy to prove the following beautiful result of Coxeter.

(6.9) COXETER'S THEOREM. *Up to conjugation, the models of the binary polyhedral groups constructed above, and the cyclic groups* $\langle \cos(2\pi/n) + i \sin(2\pi/n)\rangle$ *(for* $n \geqslant 1$*), are all the finite subgroups of* $\mathbb{H}^*$.

PROOF (*Sketch*). Let G be a finite subgroup of $\mathbb{H}^*$. Since every element in G has finite order, G must lie in $\mathbb{H}_1$. If G is cyclic, it is not hard to show that G is conjugate to $\langle \cos(2\pi/n) + i \sin(2\pi/n)\rangle$ (for some n). If G is not cyclic, then $|G|$ must be even. (If $|G|$ was odd, we would have $G \cong \varphi(G) \subseteq \mathrm{SO}(3)$. But the only odd-order subgroups of SO(3) are cyclic.) By considering an element of order 2 in G, we see that $G \supseteq \{\pm 1\}$. Then $\varphi(G)$ is conjugate to a polyhedral group in SO(3). Taking preimages under the spin covering $\varphi : \mathbb{H}_1 \to \mathrm{SO}(3)$, we can argue easily that G is conjugate to a (fixed) binary polyhedral group in $\mathbb{H}_1$. □

With this theorem proven, one ambitious question to ask is: *how about finite subgroups of division rings in general?* It turned out that this difficult problem also has a satisfactory answer: in his paper [Am] in 1955, S.A. Amitsur furnished a complete classification for the finite groups that are embeddable in the multiplicative groups of division rings. There is no need for us to reproduce Amitsur's list of groups here (which, of course, includes all finite cyclic groups and binary polyhedral groups). Let us just content ourselves by mentioning the rather remarkable fact that all finite groups in Amitsur's list turned out to be *solvable* – with the sole exception of the binary icosahedral group! (See [Am, p. 384]; also, cf. [Br, p. 102].) This group has order 120, and is isomorphic to the special linear group $\mathrm{SL}_2(\mathbb{F}_5)$. The latter group is a double covering of $\mathrm{PSL}(\mathbb{F}_5) = \mathrm{SL}_2(\mathbb{F}_5)/\{\pm 1\}$, which is the icosahedral group A_5 (the smallest nonabelian simple group). It is interesting that the *only* nonsolvable finite group embeddable in a division ring appears already in Hamilton's system of real quaternions.

7. Exponentials, De Moivre's formula, and other information

In this final section, we shall assemble a few additional facts which are pertinent to understanding the structure of the division ring of quaternions. This will be followed by some parting pointers to the literature.

[10] For nice "geometric pictures" of these constructions via the regular solids, see [Vi, pp. 15–16].

First, we note that every quaternion $q = a + bi + cj + dk$ satisfies a quadratic equation over $\mathbb{R}$. For, if we write $\mathrm{Tr}(q) = q + \bar{q} = 2a$ (the trace of q), then

$$q^2 - \mathrm{Tr}(q)q + \mathrm{N}(q) = q^2 - (q + \bar{q})q - q\bar{q} = 0. \tag{7.1}$$

Thus, q satisfies $q^2 - 2aq + (a^2 + b^2 + c^2 + d^2) = 0$. If $q \notin \mathbb{R}$, this is clearly the minimal equation of q over $\mathbb{R}$, so the quadratic extension $\mathbb{R}[q]$ must be isomorphic to the complex field $\mathbb{C}$. It is easy to check further that: (1) after identifying $\mathbb{R}[q]$ with $\mathbb{C}$, the norm of q is exactly the squared modulus of q as a complex number, and (2) the centralizer of q in $\mathbb{H}$ is exactly the field $\mathbb{R}[q]$.

One thing we can do with the above information is to define an exponential function on the quaternions. This can be done exactly in the usual way: for any $q \in \mathbb{H}$, we simply define

$$\mathrm{e}^q = \sum_{n=0}^{\infty} \frac{q^n}{n!} = 1 + \frac{q}{1!} + \frac{q^2}{2!} + \frac{q^3}{3!} + \cdots \in \mathbb{H}. \tag{7.2}$$

By the standard methods, it can be shown that this series is (absolutely) convergent (with respect to the Euclidean metric on $\mathbb{H}$), so its sum lies in $\mathbb{H}$ as indicated. But actually, this verification is not strictly necessary. For, if x is not a real number, then $\mathbb{R}[q]$ is just a copy of $\mathbb{C}$ as we have pointed out above. Therefore, we know that the series in (7.2) is (absolutely) convergent in any case, and hence already $\mathrm{e}^q \in \mathbb{R}[q]$.

One of the principal properties of the ordinary exponential function is $\mathrm{e}^p \cdot \mathrm{e}^q = \mathrm{e}^{p+q}$. If $pq = qp \in \mathbb{H}$, then they are contained in a common quadratic extension of $\mathbb{R}$, so this equality will indeed hold. But if $pq \neq qp$, we can no longer prove (and thus should not expect) the same equality. For instance, $\mathrm{e}^{\pi i} \cdot \mathrm{e}^{\pi j} = (-1)(-1) = 1$, but $\mathrm{e}^{\pi(i+j)}$ is definitely not 1. The failure of the equality $\mathrm{e}^p \cdot \mathrm{e}^q = \mathrm{e}^{p+q}$ on $\mathbb{H}$ is a serious drawback, and may have been the principal roadblock to the development of a really useful theory of functions on the quaternions.

Nevertheless, the exponential function on $\mathbb{H}$ leads to nice ways of representing quaternions. For any $\mathbf{u} \in \mathbb{H}_0 \cap \mathbb{H}_1$, we have $1 = \mathbf{u}\bar{\mathbf{u}} = -\mathbf{u}^2$, so $\mathbf{u}^2 = -1$. Thus, for any real angle θ:

$$\begin{aligned}
\mathrm{e}^{\theta\mathbf{u}} &= 1 + \frac{\theta\mathbf{u}}{1!} - \frac{\theta^2}{2!} - \frac{\theta^3\mathbf{u}}{3!} + \frac{\theta^4}{4!} + \frac{\theta^5\mathbf{u}}{5!} - \frac{\theta^6}{6!} - \frac{\theta^7\mathbf{u}}{7!} + \cdots \\
&= \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots\right) + \mathbf{u}\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots\right) \\
&= \cos\theta + \mathbf{u}\sin\theta.
\end{aligned}$$

This is entirely to be expected. For, we may identify $\mathbb{R}[\mathbf{u}]$ with $\mathbb{C}$ by identifying $\mathbf{u}$ with the complex number i, so the above equation could have been obtained from Euler's formula $\mathrm{e}^{i\theta} = \cos\theta + i\sin\theta$ for the complex numbers. Using this method, we also obtain without further proof the following extension of the usual "De Moivre formula" to unit quaternions:

$$(\cos\theta + \mathbf{u}\sin\theta)^n = \cos n\theta + \mathbf{u}\sin n\theta \quad \text{(for any integer } n). \tag{7.3}$$

Any quaternion $q \in \mathbb{H}^*$ can be written in the form[11] $\rho \cdot (\cos\theta + \mathbf{u}\sin\theta)$, where $\rho = \mathrm{N}(q)^{1/2}$, and $\theta \in [0, \pi]$ is called the *polar angle of a quaternion* of q. (Hamilton, who had a penchant for inventing strange names, called $\cos\theta + \mathbf{u}\sin\theta$ the *versor* of q.) Thus, we can represent q in the exponential form $\rho \mathrm{e}^{\theta\mathbf{u}}$, in exactly the same way as we represent complex numbers. We can then compute e^q as follows. Writing $a = \rho\cos\theta$ and $b = \rho\sin\theta$, we have $q = a + b\mathbf{u}$, and thus,

$$\mathrm{e}^q = \mathrm{e}^{a+b\mathbf{u}} = \mathrm{e}^a\mathrm{e}^{b\mathbf{u}} = \mathrm{e}^a(\cos b + \mathbf{u}\sin b). \tag{7.4}$$

From this, it follows, for instance, that e^q has norm e^{2a}, and "versor" $\cos b + \mathbf{u}\sin b$.

The (generalized) De Moivre formula (7.3) is extremely useful in analyzing the multiplicative structure of $\mathbb{H}^*$. From this formula for q^n, we can, for instance, deduce the following nice fact, which enables us to compute quickly the multiplicative order of any nonzero quaternion q via $\mathrm{N}(q)$ and the real part of q.

(7.5) THEOREM. *A nonzero quaternion q has finite multiplicative order iff* $\mathrm{N}(q) = 1$ *and the polar angle θ of q (in radian measure) is a rational multiple of 2π. In this case, the order of q is the smallest positive integer n such that $n\theta \in 2\pi\mathbb{Z}$.*

The explicit examples below will illustrate the considerable power of (7.5).

(7.6) EXAMPLES.

(1) Let $\alpha = (\pm i \pm j \pm k)/\sqrt{3}$ (where the signs are arbitrary). Then α has norm 1 and polar angle $90° = \pi/2$. Thus, $\alpha^2 = -1$, and α has order 4. (The same holds for any unit pure quaternion!)

(2) Let $\beta := (1 \pm i \pm j \pm k)/2$, which has norm 1. Its polar angle is $60° = \pi/3$. Thus, $\beta^3 = -1$, and β has order 6. On the other hand, any quaternion $(-1 \pm i \pm j \pm k)/2$ has norm 1 and polar angle $120° = 2\pi/3$, so it has order 3.

(3) Let τ denote the golden ratio $(1 + \sqrt{5})/2$, which is the positive root of $\tau^2 = \tau + 1$. From $\tau - \tau^{-1} = 1$, one gets $\tau^2 + \tau^{-2} = 3$. Thus, any quaternion of the form $\gamma := (\tau \pm i \pm \tau^{-1}j)/2$ has norm 1. Furthermore, since $\tau = 2\cos 36°$, γ has polar angle $36° = \pi/5$. From this, we see that $\gamma^5 = -1$, and the order of γ is 10. On the other hand, if, say, $\delta := (1 \pm \tau i \pm \tau^{-1}k)/2$, then $\delta^3 = -1$, and δ has order 6.

The observations above enable us to check quickly that the constructions in (6.6), (6.7) and (6.8) give us well-defined quaternion models for the binary tetrahedral, octahedral, and icosahedral groups.

We record below, largely without proof, some of the other properties of quaternions that are useful to know.

- A quaternion $q \notin \mathbb{R}$ is a pure quaternion iff $q^2 \in \mathbb{R}$.
- Two quaternions in $\mathbb{H}$ are conjugate iff they have the same norm and the same trace, iff they have the same minimal polynomial over $\mathbb{R}$. In particular, two pure quaternions

[11] The case when $q \in \mathbb{R}^*$ is a little exceptional. In this case, as we have noted earlier, we take θ to be 0 or π depending on whether q is positive or negative, and take $\mathbf{u}$ to be any unit vector.

q_1, q_2 are conjugate iff $q_1^2 = q_2^2$, so a typical conjugacy class of $\mathbb{H}^*$ in the pure quaternions is the set of zeros (in $\mathbb{H}$) of a polynomial $x^2 + r$ (where $r \in \mathbb{R}_+^*$). Each of these zero sets is *uncountable* (contrary to the case over fields where any polynomial of degree n has at most n roots).

- Every $q \in \mathbb{H}_1$ can be written as a *single* commutator $x^{-1}y^{-1}xy$, where $x, y \in \mathbb{H}_1$. In particular, the commutator subgroups of $\mathbb{H}^*$ and $\mathbb{H}_1$ are both equal to $\mathbb{H}_1$.
- (*Stereographic projection*) Taking 1 as the "north pole" of the sphere $S^3 = \mathbb{H}_1$, we can use a stereographic projection to map $\mathbb{H}_1 \setminus \{1\}$ onto the space of pure quaternions $\mathbb{H}_0$. This map works out to be $q \mapsto \frac{1+q}{1-q}$, with inverse map $p \mapsto \frac{p-1}{p+1}$ for any pure quaternion $p \in \mathbb{H}_0$. Both of these are conformal mappings.
- Any quaternion is a product of two pure quaternions. (This can be easily deduced from the proof of (5.2)(3).)
- Every linear endomorphism of $\mathbb{H}$ as a real vector space is expressible in the form $q \mapsto \sum_i x_i q y_i$, for suitable quaternions $x_i, y_i \in \mathbb{H}$.
- Every automorphism σ of $\mathbb{H}$ as a ring is an inner automorphism. (This is proved by first observing that $\sigma(\mathbb{R}) = \mathbb{R}$, since σ must take the center of $\mathbb{H}$ onto itself. Thus, σ is an $\mathbb{R}$-automorphism, and the Skolem–Noether theorem gives the desired conclusion.)[12]

We close by mentioning some additional references on quaternions, especially on topics that we have had no space to touch upon in this article. On the life and work of Sir William Rowan Hamilton, a good modern reference is [Ha]. For more historical backgrounds related to the discovery of quaternions, see [Cr,Kl,Wa$_1$,Wa$_2$]. For a survey on determinants of quaternionic matrices, see [As]. For general discussions on the use of quaternions in physics, see [AJ,Va,W]. For more information on rotations as related to quaternions, with applications to astronomy and aerospace sciences, see [Ku].

References

[A] J.F. Adams, *Vector fields on spheres*, Ann. of Math. **72** (1962), 603–632.

[Ad] S.L. Adler, *Quaternion Quantum Mechanics and Quantum Fields*, Oxford Univ. Press, New York–Oxford (1995).

[Al$_1$] S.L. Altmann, *Rotations, Quaternions, and Double Groups*, Oxford Univ. Press (1986).

[Al$_2$] S.L. Altmann, *Hamilton, Rodrigues, and the quaternion scandal*, Math. Mag. **62** (1989), 291–308.

[Am] S.A. Amitsur, *Finite subgroups of division rings*, Trans. Amer. Math. Soc. **80** (1955), 361–386.

[AJ] R. Anderson and G.C. Joshi, *Quaternions and the heuristic role of mathematical structures in physics*, Phys. Essays **6** (1993), 308–319.

[As] H. Aslaksen, *Quaternionic determinants*, Math. Intelligencer **18** (1996), 57–65.

[Ba] J.C. Baez, *The octonions*, Bull. Amer. Math. Soc. (N.S.) **39** (2002), 145–205.

[Bl] F. van der Blij, *History of the octaves*, Simon Stevin **34** (1961), 106–125.

[Br] R. Brown, *Frobenius Groups and Classical Maximal Orders*, Mem. Amer. Math. Soc., No. 717, Amer. Math. Soc., Providence, RI (2000).

[Ca] A. Cayley, *A memoir on the theory of matrices*, London Philos. Trans. **148** (1858), 17–37. (Collected Works, Vol. 2, 475–496.)

[12] In contrast to this, however, the automorphism group of $\mathbb{C}$ is more complicated. Contrary to a claim once made by Dedekind, this automorphism group contains more elements than just the identity map and the complex conjugation. The point is that an automorphism of $\mathbb{C}$ need not take $\mathbb{R}$ to $\mathbb{R}$: this shows a subtle difference between the complex numbers and the quaternions. Dedekind's claim would have been correct if we consider only $\mathbb{R}$-automorphisms on $\mathbb{C}$.

[Co] H.S.M. Coxeter, *The binary polyhedral groups, and other generalizations of the quaternion groups*, Duke Math. J. **7** (1940), 367–379.

[Cr] M.J. Crowe, *A History of Vector Analysis*, Univ. of Notre Dame Press (1967); Reprinted by Dover Publications, Inc., New York (1985).

[Di] J. Dieudonné, *Linear Algebra and Geometry*, Hermann, Houghton Mifflin Co., New York (1969).

[EN] S. Eilenberg and I. Niven, *The "fundamental theorem of algebra" for quaternions*, Bull. Amer. Math. Soc. **50** (1944), 246–248.

[Fr] F.G. Frobenius, *Über lineare Substitutionen und bilineare Formen*, Crelle J. **84** (1878), 1–63; Ges. Abhandl., Vol. 1, Springer-Verlag (1968), 343–405.

[Ga] C.F. Gauss, *Mutation des Raumes*, Carl Friedrich Gauss Werke, Achter Band, König. Gesell. Wissen. Göttingen (1900), 357–361. (This note appeared to be an updated version of an entry in Gauss's diary; editor Stäckl of the volume placed the date of the entry at around 1819.)

[H_1] W.R. Hamilton, *Lectures on Quaternions*, Hodges and Smith, Dublin (1853).

[H_2] W.R. Hamilton, *Elements of Quaternions* (1866); 2nd edn. (2 volumes), C.J. Jolly, ed., Longman, Green & Co., London (1899–1901); reprinted by Chelsea (1969).

[Ha] T.L. Hankins, *Sir William Rowan Hamilton*, Johns Hopkins Univ. Press, Baltimore (1980).

[Hap] D. Happel, *Klassifikationstheorie endlich-dimensionaler Algebren in der Zeit von 1880 bis 1920*, L'Enseig. Math. **26** (1980), 91–102.

[Hi] F. Hirzebruch, *Division algebras and topology*, Numbers, H.-D. Ebbinghaus et al., eds, Readings in Math., Springer, Berlin–Heidelberg–New York (1991), 281–302.

[Hu] A. Hurwitz, *Über die Komposition der quadratischen Formen von beliebig vielen Variablen*, Nach. der könig. Gesell. Wissen. Göttingen Math.-Phys. Kl. (1898), 309–316; Math. Werke **2** (1933), 565–571.

[Kl] M. Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford Univ. Press, Oxford–New York–Toronto (1972).

[KR_1] M. Koecher and R. Remmert, *Hamilton's quaternions*, Numbers, H.-D. Ebbinghaus et al., eds, Graduate Texts in Math., Vol. 123 (Readings in Mathematics), Springer, Berlin–Heidelberg–New York (1991), 189–220.

[KR_2] M. Koecher and R. Remmert, *The isomorphism theorems of Frobenius, Hopf and Gelfand-Mazur*, Numbers, H.-D. Ebbinghaus et al., eds, Graduate Texts in Math., Vol. 123 (Readings in Mathematics), Springer, Berlin–Heidelberg–New York (1991), 221–247.

[Ku] J. Kuipers, *Quaternions and Rotation Sequences*, Princeton Univ. Press, Princeton, NJ (1999).

[La_1] T.Y. Lam, *The Algebraic Theory of Quadratic Forms* (2nd printing, with revisions), Benjamin–Addison-Wesley, Reading, MA (1980).

[La_2] T.Y. Lam, *A First Course in Noncommutative Rings*, 2nd edn., Graduate Texts in Math., Vol. 131, Springer, Berlin–Heidelberg–New York (2001).

[Ma] K.O. May, *The impossibility of a division algebra of vectors in three-dimensional space*, Amer. Math. Monthly **73** (1966), 289–291.

[Me] A. Messiah, *Quantum Mechanics*, Vol. 2, North-Holland, J. Wiley & Sons, Inc., New York (1965).

[Pe] B. Peirce, *Linear associative algebras*, Amer. J. Math. **4** (1881), 97–229.

[PS] G. Pickert and H.-G. Steiner, *Komplexe Zahlen und Quaternionen*, Kapitel 8 in *Grundzüge der Mathematik I*, Behnke, Bachmann, Fladt, Süss, eds, Vandenhoeck & Ruprecht, Göttingen (1962).

[Sh] D. Shapiro, *Compositions of Quadratic Forms*, de Gruyter Expos. in Math., Vol. 33, Berlin (2000).

[Va] M.A. Vandyck, *Quaternions in physics, Victorian curiosities or enlightening language? New insights 150 years after Hamilton's discovery*, Rev. Questions Sci. **172** (2002), 293–307.

[Vi] M.-F. Vigneras, *Arithmétique des Algèbres de Quaternions*, Lecture Notes in Math., Vol. 800, Springer, Berlin–Heidelberg–New York (1980).

[Wa_1] B.L. van der Waerden, *Hamiltons Entdeckung der Quaternionen*, erweitierte Fassung eines Vortrags 26 Juni, 1973, Vandenhoeck & Ruprecht, Göttingen (1973).

[Wa_2] B.L. van der Waerden, *Hamilton's discovery of quaternions*, Math. Magazine **49** (1976), 227–234.

[W] J.P. Ward, *Quaternions and Cayley Numbers*, Mathematics and its Applications, Vol. 403, Kluwer Academic Publishers, Dordrecht (1997).

Group Rings

S.K. Sehgal

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1
E-mail: S.Sehgal@ualberta.ca

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1. Introduction

In 1837, Sir William Rowan Hamilton[1] gave the first formal theory of complex numbers, defining them as ordered pairs of real numbers, just as is done nowadays, thus ending almost three hundred years of discussions regarding their legitimacy.

He then came to consider elements of the form $\alpha = a + bi + cj + dk$, which he called *quaternions*, where the coefficients a, b, c, d represent real numbers and i, j, k are formal symbols called *basic units*. It was obvious to him that two such elements should be added componentwise.

The main difficulty was to define the product of two such elements in a reasonable way. Since this product should have the usual properties of a multiplication, such as the distributive law, it would actually be enough to decide how to multiply the symbols i, j, k among themselves. This again demanded considerable effort since from the very beginning Hamilton also assumed implicitly that the product should be commutative. This was perfectly reasonable since he did not know, at the time, that he was about to discover the first noncommutative algebra structure in the history of mathematics. Finally, in October 1843 he discovered the fundamental laws of the product of quaternions:

$$i^2 = j^2 = k^2 = ijk = -1,$$

which also imply the well-known formulas:

$$\begin{aligned} ij &= k = -ji, \\ jk &= i = -kj, \\ ki &= j = -ik \end{aligned}$$

and guarantee that they form a (noncommutative) field: every nonzero element has an inverse.

The very next day, he presented an extensive memoir on quaternions to the Royal Irish Academy. The discovery of quaternions came as a shock to mathematicians of the time for several reasons. Among other things, it opened the possibilities of new extensions of the field of complex numbers, precisely when the recent discovery of the so-called fundamental theorem of algebra seemed to indicate that the need for new extensions had come to an end.

In 1845, Sir Arthur Cayley introduced a new set of numbers, the *octonions*, which can be defined as the set of elements of the form $a_0 + a_1e_1 + a_2e_2 + \cdots + a_7e_7$, where the coefficients a_i, $0 \leqslant i \leqslant 7$, are real numbers and the symbols e_i, $1 \leqslant i \leqslant 7$, are the *basic units*. Once again, the sum of two such elements is defined componentwise and the product is first defined on the basic units according to certain rules and then extended distributively. A striking fact about octonions is that they form a noncommutative field which is not even associative. Octonions are also known as *Cayley numbers*.

Hamilton himself realized that it was possible to extend this construction even further and he first defined *biquaternions*, which are again elements of the form $\alpha = a + bi + cj + dk$, where the coefficients a, b, c, d are now assumed to be complex numbers. Soon

[1] See [Kli, p. 775].

afterwards he introduced the *hypercomplex systems*. These are sets of elements of the form $\alpha = a_1e_1 + a_2e_2 + \cdots + a_ne_n$, where the sum is once again defined componentwise and multiplication is defined by establishing the values of the products of basic units pairwise. Since the product of units e_i pairwise must again be an element of this set, it must be of the form:

$$e_ie_j = \sum_{k=1}^{n} \gamma_k(i, j)e_k.$$

In other words, to establish an algebra structure on this set, it is enough to choose conveniently values of the coefficients $\gamma_k(i, j)$, called the *structural constants* of the system.

These facts were the first steps in the development of ring theory. Soon, many new systems were discovered and the need for a classification was felt. In a paper read in 1870 and published in lithographed form in 1871, entitled *Linear Associative Algebras*, Benjamin Peirce[2] gave a classification of algebras known at the time and determined 162 algebras of dimension less than or equal to 6. As tools of his method of classification, B. Peirce introduced some very important ideas in ring theory, such as the notions of idempotent and nilpotent elements, and the use of idempotents to obtain a decomposition of a given algebra.

During that century, important developments were taking place in the theory of nonassociative algebras. Following the work of S. Lie and W. Killing in the study of Lie groups and Lie algebras, A. Study and G. Scheffers introduced in the period 1889–1898 the concepts of simple and semisimple algebras (though using another terminology) – basic notions for the development of structure theory.

Inspired by these results both T. Molien and E. Cartan obtained, independently, important results regarding the structure theory of finite-dimensional real or complex algebras, introducing in this context the notions of simple and semisimple algebras and characterizing simple algebras as full matrix algebras.

All this work culminated in the beautiful theorems of J.H.M. Wedderburn describing the structure of finite-dimensional algebras over arbitrary fields, using techniques related to the existence of idempotent elements, as suggested by the earlier work of B. Peirce.

Now, let us turn to group theory. It is well-known that the attention was focused on permutations by the work on algebraic equations of J.L. Lagrange in 1770, followed by P. Ruffini and N.H. Abel. In his classical work of 1830, E. Galois was the first to consider groups and subgroups of permutations, using the term *group* in its modern sense – though restricted to permutations – and introducing such concepts as those of normal subgroup, solvable group, etc.

A. Cauchy was a pioneer in understanding the relevance of permutation groups as an independent subject. He wrote a series of interesting papers about them, in the period 1844–1846.

Influenced by Cauchy's work, Arthur Cayley recognized that the notion of a group could be formulated in a more abstract setting and gave the first definition of an *abstract group*

[2] The paper was finally published in the American Journal of Mathematics in 1881, then edited by J.J. Sylvester, with notes and addenda by Charles Sanders Peirce, son of the author and better known for his fundamental work in philosophy, see [Pei].

in 1854 [Cay]. This paper is considered by several authors (e.g., N. Bourbaki [Bou] or M. Kline [Kl]) as the beginning of abstract group theory. It is a relatively short work, but it contains a number of important features:

- Gives an abstract definition of a group, in multiplicative notation.
- Introduces the *table* of an operation.
- Shows that there exist two nonisomorphic groups of order four, giving explicit examples.
- Shows that there exist two nonisomorphic groups of order six, one of them being commutative and the other isomorphic to S_3, the group of permutations of three elements.
- Shows that the order of every element is a divisor of the order of the group.

It is interesting to note that it is in this very same paper that the notion of a *group ring* appears for the first time. Almost at the end of the article, Cayley states that if we consider the elements of a (finite) group as "basic units" of a hypercomplex system or, in his own words: "*if we consider them as quantities … such as the quaternions imaginaries*" then the group's product actually defines the product of these hypercomplex quantities.

Explicitly, given a finite group $G = \{g_1, g_2, \dots, g_n\}$, consider all elements of the form:

$$x_1 g_1 + x_2 g_2 + \cdots + x_n g_n$$

where $x_1, x_2, \dots, x_n$ are either real or complex numbers; then the product of two such elements $\alpha = \sum_{i=1}^{n} x_i g_i$, $\beta = \sum_{i=1}^{n} y_i g_i$ is given by:

$$\alpha\beta = \sum_{i,j} (x_i y_j)(g_i g_j).$$

Though this is precisely the definition of a group ring – in this particular case – this paper had no immediate influence on the mathematicians of the time and group rings remained unknown for still quite some time. They were introduced again by Theodor Molien when he realized that this was a natural setting in which to apply some of his earlier criteria for semisimplicity. In his doctoral thesis, submitted to the University of Yurev in Estonia, in 1892 (and published in 1893 [Mol1]), he tried to give a structure theory for hypercomplex systems.

His criteria were based on certain bilinear forms which, in turn, depended on the structural constants of the system. He realized that if the basic units formed a group these would be particularly simple (since in that case they assume only the values 0 and 1), so he was naturally led to work with group rings.

In a subsequent paper [Mol2], Molien obtained important results "*relating the representability of a given discrete group in the form of a homogeneous linear substitution group*". In this way, he discovered some of the basic results in the theory of complex representations of finite groups, including the orthogonality relations for group characters.

The connection between group representation theory and the structure theory of algebras – which is obtained through group rings – was widely recognized after the most influential paper by Emmy Noether [No1], some joint work of hers with Richard Brauer [BN] and Brauer's paper [Br1], giving the subject a new impulse.

Later, the subject gained importance of its own after the inclusion of questions on group rings in Kaplansky's famous list of problems [Kap2,Kap4]. Let us remember that G. Higman [Hi1,Hi2] investigated and raised important questions about units of integral group rings in 1940. Those questions originated from his investigations of the Whitehead torsion in topology – an important invariant of polyhedrons (Higman was a student of J.H.C. Whitehead). Also, Magnus [M] and Zassenhaus [Zas1] studied dimension subgroups of free groups. Other important facts to stimulate the area were the paper by Connell [Con] considering ring-theoretical questions about group rings, the inclusion of chapters on group rings in the books on ring theory by Lambek [Lam] and Ribenboim [Rib] as well as the publication of the first book entirely devoted to the subject, by Passman [P9].

Several new books on the subject have been published in recent years by Bovdi [Bov2], Karpilovsky [Kar], Passi [Pas2], Roggenkamp and Taylor [RT], Sehgal [Se10,Se13] and Polcino Milies and Sehgal [PSe4]. Passman's "The algebraic structure of group rings" [P21] is truly classic and encyclopedic. For more details on the history of group rings see [Haw1,Haw2,Haw3,Po4].

The subject of group rings has developed so much in breadth and depth that it is not possible to do it justice in an article like the present one due to the constraints of time and space. Accordingly, we have picked only highlights from a small number of topics to report. The glaring omissions include semiprimitivity, primitivity and ideal theory. For these see [P21,P22,Ros1].

We wish to express our gratitude to Antonio Giambruno, Klaus Hoechsmann, Ted Hurley, Ernst Kleinert, Yuanlin Li, Gabi Nebbe, Mike Parmenter, Cesar Polcino Milies, Jürgen Ritter and Al Weiss for their generous help in the preparation of this article. Finally, we are grateful to Vivian Spak for her excellent typing and cheerful assistance.

The object of our report, a group ring RG of a group G over a unital ring R is simply the set of all finite sums $\alpha = \sum \alpha(g)g$, $\alpha(g) \in R$, $g \in G$ (also written $\sum \alpha_g g$, sometimes) with equality and addition componentwise and multiplication induced from the group multiplication:

$$\Big(\sum \alpha(g)g\Big)\Big(\sum \beta(h)h\Big) = \sum \alpha(g)\beta(h)gh = \sum \gamma(x)x$$

with $\gamma(x) = \sum_{gh=x} \alpha(g)\beta(h)$. The set $\{g\colon\ \alpha(g) \neq 0\}$ is called the support of α and is denoted by $\operatorname{supp}(\alpha)$. Another way of looking at RG is as the set of functions from G to R with almost all values equal to zero with pointwise addition and convolution as multiplication. In fact, there is a natural and useful extension of the concept.

DEFINITION 1.1. Let G be a group and R a ring with identity. Suppose that we are given a function $\rho : G \times G \to \mathcal{U}(R)$, the unit group of R, called a factor system and an automorphism σ_g of R for each $g \in G$. Suppose that ρ and σ satisfy the following properties for each $g, h, \ell \in G$, $a \in R$:

(i) $\rho_{g,h}\rho_{gh,\ell} = \sigma_g(\rho_{h,\ell})\rho_{g,h\ell}$ and

(ii) $\rho_{h,g}\sigma_{hg}(a) = \sigma_h(\sigma_g(a))\rho_{h,g}$.

Then by the crossed product $R(G,\rho,\sigma)$ of G over R with factor system ρ and automorphisms σ, we understand the set of finite sums

$$\Bigl\{\sum a_i\bar{g}_i\colon a_i\in R,\ g_i\in G\Bigr\},$$

where $\bar{g}_i$ is a symbol corresponding to g_i. Equality and addition are defined componentwise and for $g,h\in G$, $a\in R$ we have

$$\bar{g}\bar{h}=\rho_{g,h}\overline{gh},\qquad \bar{g}a=\sigma_g(a)\bar{g}. \tag{$*$}$$

$R(G,\rho,\sigma)$ is easily seen to be a ring if we extend $(*)$ distributively.

As a special case, if $\rho_{gh}=1$ for all $g,h\in G$ we get the skew group ring $R(G,1,\sigma)$ and we denote it by $R_\sigma(G)$. In addition, if we have $\sigma_g=I$ for all $g\in G$ we get the group ring RG.

The crossed products arise naturally. For example, if G has a normal subgroup N then picking a fixed set S of coset representatives $\{\bar{g}_v\}$ of N in G we can write every $\alpha\in RG$ as $\alpha=\sum a_g\bar{g}$, $a_g\in RN$, $\bar{g}\in S$. Then for $\bar{g},\bar{h}\in S$, $a\in RN$, we write

$$\bar{g}\bar{h}=\rho_{g,h}\overline{gh},\qquad \bar{g}a=\sigma_g(a)\bar{g}.$$

It follows that

$$RG=(RN)(G/N,\rho,\sigma).$$

We shall denote the crossed product by $R*G$, for simplicity. This concept will be needed a few times during this exposition. Throughout, $\boldsymbol{Z}$, $\boldsymbol{Q}$, $\boldsymbol{R}$, $\boldsymbol{C}$ will denote the rational integers, the rational numbers, the real numbers and the complex numbers. One more bit of standard notation: the map $\varepsilon: RG\to R$, $\sum a_g g\to\sum a_g$ is called the augmentation map. Its kernel is denoted by $\Delta_R(G)$. Also, the kernel of the natural projection $RG\to R(G/N)$ where $N\lhd G$ is denoted by $\Delta_R(G,N)$. The subscript R is usually omitted.

2. Idempotents

For an arbitrary associative ring Λ the group $K_0(\Lambda)$ consists of formal differences $[P]-[Q]$ of pairs (of isomorphism classes) of finitely generated projective Λ-modules modulo the relation $[P\oplus Q]=[P]+[Q]$. When $\Lambda=\boldsymbol{Z}G$, the group $K_0(\boldsymbol{Z}G)$ arises in topology, as the set of obstacles to finiteness of an infinite cell complex with the fundamental group G.

Units in $\boldsymbol{Z}G$ can be used to construct examples of projective $\boldsymbol{Q}G$-modules in the following way. Suppose u is an invertible element with $u^n=1$ for some $n>1$. Then the element $e=\frac{1}{n}(1+u+\cdots+u^{n-1})$ satisfies the relation $e^2=e$, i.e., it is an idempotent. Then $\boldsymbol{Q}Ge$ is a left projective $\boldsymbol{Q}G$-module.

Our first observation about idempotent elements in KG is the following

THEOREM 2.1. *Let K be a field of characteristic zero and let G be a finite group. Suppose $e = \sum e(g)g$ is an idempotent $\neq 0, 1$ in KG. Then*
(i) $e(1) \in \boldsymbol{Q}$,
(ii) $0 < e(1) < 1$.

PROOF. Let $\rho : KG \to M_{|G|}(K)$ be the regular representation; it maps α to the matrix $\rho(\alpha)$ of the left multiplication by α with respect to the natural basis G of KG. The eigenvalues of the idempotent matrix $\rho(e)$ are zeros and ones. So the trace of $\rho(e)$ is a natural number r equal to its rank. Also,

$$\operatorname{tr}\rho(e) = \sum_g e(g)\operatorname{tr}\rho(g).$$

But $\operatorname{tr}\rho(g) = 0$ if $g \neq 1$ and $\operatorname{tr}\rho(1) = |G|$ as multiplication by g fixes an $h \in G$ if and only if $g = 1$. We obtain that

$$\operatorname{tr}\rho(e) = e(1)|G| = r.$$

It follows that $e(1) = r/|G|$ with $0 \leqslant r \leqslant |G|$. Also we have:

$$e(1) = 0 \ \Leftrightarrow\ e = 0 \quad \text{and} \quad e(1) = 1 \ \Leftrightarrow\ \rho(e) = I,\ e = 1. \qquad \square$$

We shall see that this theorem extends beautifully to infinite groups.

DEFINITION 2.2. The first coefficient $\gamma(1)$ of $\gamma = \sum \gamma(g)g \in RG$ is called the trace of γ and is denoted by $\operatorname{tr}(\gamma)$ (or $\operatorname{tr}\gamma$).

It is easy to see that tr has trace like properties, especially, $\operatorname{tr}(\alpha\beta) = \operatorname{tr}(\beta\alpha)$.

The first part of the last theorem was extended to arbitrary groups by A. Zalesskii [Za4]. He used techniques from algebraic number theory including deep results like the Frobenius density theorem.

THEOREM 2.3. *Let K be a field and G an arbitrary group. Let $e = \sum e(g)g$ be an idempotent in KG. Then $\operatorname{tr} e = e(1)$ belongs to the prime subfield of K.*

PROOF. At first, we assume that $\operatorname{char}(K) = p > 0$. Let $e = \sum e(g)g = e^2 \in KG$. Since the support of e is finite we can pick a natural number n so that the p-elements in $\operatorname{supp}(e)$ have order dividing p^n. Then

$$e = e^{p^n} = \sum e(g)^{p^n} g^{p^n} + \lambda, \quad \lambda \in [KG, KG].$$

Here, $[KG, KG] \subset KG$ is the linear space spanned by all the Lie products $[\gamma, \mu] = \gamma\mu - \mu\gamma$ with $\gamma, \mu \in G$. Thus

$$e(1) = \sum_{g^{p^n}=1} e(g)^{p^n} = \Bigl(\sum_{g^{p^n}=1} e(g)\Bigr)^{p^n}.$$

By the same argument,

$$e(1) = \sum_{g^{p^{n+1}}=1} e(g)^{p^{n+1}} = \left(\sum_{g^{p^n}=1} e(g) \right)^{p^{n+1}} = e(1)^p.$$

It follows that $e(1)$ belongs to the prime subfield $\boldsymbol{F}_p \subseteq K$.

The $\mathrm{char}(K) = 0$ case is deduced from the characteristic p case and the following result, for which we refer the reader to [P21, p. 48].

(∗) If $R = \boldsymbol{Z}[x_1, \ldots, x_m]$ is a finitely generated commutative integral domain such that $x_1 \notin \boldsymbol{Q}$ then there exists a field F of characteristic $p > 0$ and a homomorphism $\varphi : R \to F$ so that $\varphi(x_1) \notin \boldsymbol{F}_p$.

Given $e = e^2 = \sum e(g)g \in KG$ with $\mathrm{char}(K) = 0$, we want to prove $e(1) \in \boldsymbol{Q}$. Suppose not. Let $R = \boldsymbol{Z}[e(g)\colon\ g \in G]$. Let φ be the homomorphism given by $(*)$ with $\varphi(e(1)) \notin \boldsymbol{F}_p$. Extend φ to $- : RG \to FG$. Then $\bar{e} \in FG$, $(\bar{e})^2 = \bar{e}$, $\bar{e}(1) = \varphi(e(1)) \notin \boldsymbol{F}_p$, contradicting the first part. □

The second part of the next theorem for infinite groups was proved by Kaplansky. Passman gave an elementary proof of the same.

THEOREM 2.4 (Kaplansky). *If G is an arbitrary group and K is a field of characteristic* 0 *then the trace of an idempotent $e \in KG$, $e \neq 0, 1$, lies strictly between* 0 *and* 1.

PROOF. See [Kap3] or [P5, p. 97]. □

REMARK. A striking consequence of the last theorem is that the integral group ring $\boldsymbol{Z}G$ has no idempotents other than 0 and 1.

We have seen that, in characteristic zero, if $e = \sum e(g)g \in KG$ is an idempotent then $e(1)$ must be rational. One cannot say much about the other coefficients $e(g)$, $g \neq 1$. They need not even be algebraic (see [Se10, p. 12], for an example). However, it gets more interesting if we consider

$$\tilde{e}(g) = \sum_{h \sim g} e(h),$$

the sum of coefficients of e over the conjugacy class of g in G.

The fact that the coefficients $e(g)$ are not algebraic but the $\tilde{e}(g)$ are, was proved by H. Bass. We have

THEOREM 2.5 (Bass). *Let K be a field of characteristic zero and $e = \sum e(g)g = e^2 \in KG$. Then $E = \boldsymbol{Q}(\tilde{e}(g), g \in G)$ is a finite normal extension of $\boldsymbol{Q}$ with Abelian Galois group.*

PROOF. See [Bas3] or [Se10, p. 14]. □

We have an analogue of Kaplansky's theorem:

THEOREM 2.6 (Weiss). *Let e be an idempotent in KG where* $\operatorname{char}(K)=0$. *Let $g \in G$ be an element with c conjugates. Then*

$$\left|\tilde{e}(g)\right|^2 \leqslant c.$$

PROOF. See [We1]. □

By a famous theorem of Kronecker and Weber the extension E of Theorem 2.5 is contained in a cyclotomic field. Bass [Bas3] has conjectured that it should be possible to be more precise about the cyclotomic field. He confirmed the conjecture for polycyclic-by-finite groups.

CONJECTURE 2.7. *$E = \boldsymbol{Q}(\tilde{e}(g), g \in G)$ is contained in $\boldsymbol{Q}(\zeta_\ell)$ where ζ_ℓ is a primitive ℓth root of unity and ℓ is the least common multiple of all finite orders $o(g)$ of the group elements $g \in G$, for which $\tilde{e}(g) \neq 0$.*

In the same vein there is the

THEOREM 2.8 (Cliff–Sehgal). *Let KG be the group ring of a polycyclic-by-finite group G over a field K of characteristic* 0. *Let $e = \sum e(g)g$ be a nontrivial idempotent in KG. Write the rational number $e(1) = r/s$ with $(r,s) = 1$. If a prime p divides s then there exists a non-identity element $g \in G$ of p-power order with $\tilde{e}(g) \neq 0$.*

PROOF. See [CS1] or [Se10, p. 21]. □

Accordingly, we have the

PROBLEM 2.9. Can the last theorem be extended to more general groups?

Theorem 2.6 has been extended by Passi and Passman [PP] as follows.

THEOREM 2.10. *Let $\alpha = \lambda_1 e_1 + \cdots + \lambda_n e_n \in \boldsymbol{C}G$ with $e_1, \ldots, e_n$ orthogonal idempotents and $\lambda_i \in \boldsymbol{C}$. Let $\mathcal{C}_g$ be the conjugacy class of g and let T be a set of representatives of conjugacy classes of G. Then*

$$\sum_{g \in T} \frac{1}{|\mathcal{C}_g|} \left|\tilde{\alpha}(g)\right|^2 \leqslant |\lambda_1|^2 \operatorname{tr}(e_1) + |\lambda_2|^2 \operatorname{tr}(e_2) + \cdots + |\lambda_n|^2 \operatorname{tr}(e_n)$$

with equality if and only if α is central. Furthermore, if $\lambda = \max_i |\lambda_i|$, then $\sum \frac{1}{|\mathcal{C}_g|} |\alpha(g)|^2 \leqslant |\lambda|^2$ with equality if and only if α is central, $|\lambda_i| = \lambda$ for all i with $e_i \neq 0$ and $1 = e_1 + \cdots + e_n$.

For a detailed survey of these trace functions and the Bass conjecture we refer the reader to [Pas3].

Idempotent ideals of $\mathbf{Z}G$

If e is a central idempotent then $I = e\mathbf{Z}G$ satisfies $I = I^2$, i.e., it is an idempotent ideal. More generally, we can address the question of existence of idempotent ideals in integral group rings. If H is a perfect group, i.e., $H = H'$ then $(\Delta(H))^2 = \Delta(H)$, thanks to the identity

$$\begin{aligned}(x, y) - 1 &= x^{-1}y^{-1}xy - 1 = x^{-1}y^{-1}(xy - yx)\\ &= x^{-1}y^{-1}\big((x-1)(y-1) - (y-1)(x-1)\big).\end{aligned}$$

Thus, whenever a group G contains a normal perfect subgroup H then $\Delta(G, H)$ is an idempotent ideal of $\mathbf{Z}G$. It was conjectured by Akasaki [Ak1,Ak2], who also obtained partial results, and proved by Roggenkamp [Rog] that for finite groups G, $\mathbf{Z}G$ contains a nontrivial idempotent ideal if and only if G is not solvable. Patrick Smith [Sm5] extended this result to polycyclic-by-finite groups G. We have the

PROBLEM 2.11. Characterize groups G such that $\mathbf{Z}G$ has no nontrivial idempotent ideals.

In this connection we should keep in mind (see [Gr2]) that if G is an Abelian torsion divisible group then $\Delta^2(G) = \Delta^4(G) \neq 0$.

3. Units of ZG

The theory of units of integral group rings has developed a great deal in the last twenty five or thirty years. We begin with a classical result of Graham Higman [Hi2]. In analogy with the fact that the torsion units of the ring of cyclotomic integers $\mathbf{Z}[\zeta]$ are of the form $\pm\zeta^i$, we have that the torsion units of the integral group ring of a finite Abelian group A are simply $\pm a$, $a \in A$. Let us fix notation and denote by $\mathcal{U}(\mathbf{Z}G)$ (respectively $\mathcal{U}_1\mathbf{Z}(G)$) the units of $\mathbf{Z}G$ (respectively units of augmentation one). Clearly, $\mathcal{U}(\mathbf{Z}G) = \pm\mathcal{U}_1(\mathbf{Z}G)$. Our first lemma is

LEMMA 3.1 [Hi1,Hi2]. *Let* $u = \sum u(g)g$ *be a unit of finite order in* $\mathbf{Z}G$ *where* G *is a finite group. Then*

$$u(1) \neq 0 \quad \Rightarrow \quad u = u(1) = \pm 1.$$

PROOF. Denote by $\rho(u)$ the matrix of the regular representation with respect to the basis G. Then $\rho(u)$ can be diagonalized. Its eigenvalues are roots of unity $\zeta_1, \zeta_2, \ldots, \zeta_n$ where $n = |G|$. We have trace $\rho(u) = \sum \zeta_i$. Also,

$$\operatorname{tr}\rho(u) = \sum u(g)\operatorname{tr}\rho(g) = u(1)\operatorname{tr}\rho(1) = nu(1).$$

We have $nu(1) = \sum_1^n \zeta_i$. Consider the absolute value of both sides to conclude that all ζ_i's are equal to ζ_1. Thus $u(1) = \zeta_1 = \pm 1$. We conclude that $\rho(u) = \pm I$, $u = \pm 1$. $\square$

THEOREM 3.2 [Hi1,Hi2]. *Let A be a finite Abelian group. Then $\mathcal{U}(\boldsymbol{Z}A) = \pm A \times F$ where F is a torsion free Abelian group of finite rank.*

PROOF. (i) First we see that all torsion units are trivial. If $u \in \boldsymbol{Z}A$ is of finite order let a_0 be an element in the support of u. Then considering $\mu = a_0^{-1}u$ we have $\mu(1) \neq 0$ and μ is torsion. It follows by the lemma above that $\mu = \pm 1$ and thus $u = \pm a_0$.

(ii) It remains to prove that $\mathcal{U}(\boldsymbol{Z}A)$ is finitely generated. We know that $\boldsymbol{Q}A$ is a direct sum of cyclotomic fields:

$$\boldsymbol{Q}A = \sum\nolimits^{\oplus} \boldsymbol{Q}(\zeta), \quad \text{and} \quad \boldsymbol{Z}A \hookrightarrow \sum\nolimits^{\oplus} \boldsymbol{Z}[\zeta] = \boldsymbol{M}. \tag{$*$}$$

Since $\boldsymbol{Z}A$ is an order contained in the maximal order $\boldsymbol{M}$, the additive index $[\boldsymbol{M}\colon \boldsymbol{Z}A]$ is finite. Thus the multiplicative index $(\mathcal{U}(\boldsymbol{M})\colon \mathcal{U}(\boldsymbol{Z}A))$ is also finite (see [Se13, p. 19]). But $\mathcal{U}(\boldsymbol{M})$ is finitely generated as each $\mathcal{U}\boldsymbol{Z}[\zeta]$ is. Thus $\mathcal{U}(\boldsymbol{Z}A)$ is finitely generated and we are done. □

REMARK 3.3. In fact analogous to, and an application of, the Dirichlet's unit theorem is the more precise statement: the rank of F equals $\frac{1}{2}(n + 1 + n_2 - 2\ell)$ where $n = |A|$, $n_2 =$ the number of elements of order 2 in A and $\ell =$ the number of cyclic subgroups of A.

The next result follows easily from $(*)$. The obvious units $\pm g,\ g \in G$ of $\boldsymbol{Z}G$ are called trivial units.

THEOREM 3.4. *Let A be a finite Abelian group. Then $\mathcal{U}(\boldsymbol{Z}A)$ is trivial if and only if A has exponent dividing* 4 *or* 6.

In general we have

THEOREM 3.5 (G. Higman). *Let G be a finite non-Abelian group. Then $\mathcal{U}(\boldsymbol{Z}G)$ is trivial if and only if $G = K_8 \times E$ where K_8 is the quaternion group of order* 8 *and E is an elementary Abelian* 2*-group.*

PROOF. See [Se13, p. 8]. □

The group K_8 consists of the "Hamilton quaternion units" $\{\pm i, \pm j, \pm k, \pm 1\}$ with the well known multiplications. Abstractly, it is isomorphic to $\langle x, y\colon\ x^4 = 1 = y^4,\ x^2 = y^2,\ yx = x^{-1}y\rangle$.

In general it is very difficult to describe $\mathcal{U}(\boldsymbol{Z}G)$ for a finite group G. We do know, thanks to Siegel [Si] and Borel and Harish-Chandra [BH] that it is a finitely generated, even a finitely presented group.

For $G = S_3$ a matrix description was given for $\mathcal{U}(\boldsymbol{Z}G)$ by Hughes and Pearson [HP]. Marciniak and Sehgal [MS6] gave a presentation of the same. For description of units of other examples of groups see [AH1,AH2,AH4,Kl1,Kl3,Po1,RS1,Fe2,JL1,SL,SLD].

A slightly improved version of the first part of Theorem 3.2 is given by

THEOREM 3.6 [Hi2]. *Let G be a finite group. Then all torsion central units of $\boldsymbol{Z}G$ are trivial.*

The proof is the same as in Theorem 3.2. The regular representation also gives

THEOREM 3.7 [Se8]. *Suppose that a central element $\gamma = \sum \gamma(g)g$ of the integral group ring $\boldsymbol{Z}G$ satisfies the equation $\gamma^2 = m\gamma$ with $m \neq 0$. Suppose $\gamma(1) \neq 0$. Then $\gamma = \widehat{N} = \sum_{x \in N} x$ for some normal subgroup N of G.*

There is an analogue of Lagrange's theorem of group theory as follows:

THEOREM 3.8. *Let G be a finite group. Suppose $u \in \mathcal{U}_1(\boldsymbol{Z}G)$ is a unit of order n. Then n divides $|G|$.*

PROOF. Write $e = (1 + u + \cdots + u^{n-1})/n$. Then $e^2 = e$ is an idempotent of $\boldsymbol{Q}G$. Also, $\operatorname{tr} e = e(1) = 1/n$. Moreover, the matrix trace of the regular representation of e equals $\operatorname{tr} \rho(e) = e(1)|G| = \frac{1}{n}|G| = s$, the rank of $\rho(e)$. Thus $sn = |G|$ and n divides $|G|$ as claimed. $\square$

It is also an interesting fact that G is a maximal finite group of units in $\mathcal{U}_1(\boldsymbol{Z}G)$.

THEOREM 3.9. *Any finite group of units in $\mathcal{U}_1(\boldsymbol{Z}G)$ is linearly independent and its order is a divisor of $|G|$.*

PROOF. Suppose $\{u_1, u_2, \ldots, u_n\} = H$ is a finite subgroup of $\mathcal{U}_1(\boldsymbol{Z}G)$. Suppose $\sum c_i \cdot u_i = 0$ for some $c_i \in \boldsymbol{Z}$. Then multiplying by u_1^{-1} we get $\sum c_i(u_1^{-1}u_i) = 0$. This says

$$c_1 \cdot 1 + c_2 u_1^{-1} u_2 + \cdots + c_n u_1^{-1} u_n = 0.$$

Computing the trace of both sides, we obtain $c_1 = 0$. Similarly $c_i = 0$ for all i. This proves linear independence of the elements of H. The proof of divisibility of $|G|$ by $|H|$ is the same as in the last theorem. $\square$

Most of the above results can be extended to infinite groups as we see now. We begin with a result which extends Lemma 3.1.

LEMMA 3.10 [Bas3,P3]. *Let u be a torsion unit of $\boldsymbol{Z}G$ where G is an arbitrary group. Then*

$$u(1) \neq 0 \quad \Rightarrow \quad u = u(1) = \pm 1.$$

PROOF. Suppose $u^n = 1$. Consider the algebra $\boldsymbol{C}\langle u\rangle$. The map $x \to u$ from the polynomial ring $\boldsymbol{C}[x]$ to the subalgebra $C\langle u\rangle \subset \boldsymbol{C}G$ is an epimorphism whose kernel is a principal ideal $f(x)\boldsymbol{C}[x]$, where $f(x)$ is a divisor of $x^n - 1$. Since $f(x)$ has distinct roots, $\boldsymbol{C}\langle u\rangle$ is a direct sum of fields:

$$\boldsymbol{C}\langle u\rangle = \sum\nolimits^{\oplus} \boldsymbol{C}. \tag{$*$}$$

Write $1 = \sum e_i$, a sum of orthogonal idempotents in the above decomposition. Taking the trace we get

$$1 = \sum e_i(1) = \sum r_i/s, \quad r_i, s \in \boldsymbol{Z},\ 0 < r_i < s,$$

by theorems of Zalesskii (2.3) and Kaplansky (2.4). Note that we have $s = \sum r_i$. In the decomposition $(*)$, we can write

$$u = \sum \zeta_i \cdot e_i, \quad \zeta_i \in \boldsymbol{C},\ \zeta_i^n = 1.$$

Taking the trace of the last equation we obtain

$$0 \neq a = u(1) = \sum \zeta_i \operatorname{tr}(e_i) = \sum \zeta_i r_i/s$$

which gives $sa = \sum \zeta_i r_i$, a sum of s roots of unity. By considering the absolute value of both sides which must be a multiple of s and also less than or equal to s, we conclude that all ζ_i are equal to ζ_1. We get $sa = \zeta_1 \cdot \sum r_i = \zeta_1 s$, $a = \zeta_1 = \pm 1$. Also, $u = \sum \zeta_1 e_i = \zeta_1$, $\sum e_i = \zeta_1 = \pm 1$, as claimed. □

This result has several important consequences.

COROLLARY 3.11. *For an arbitrary group G every central torsion unit of $\boldsymbol{Z}G$ is trivial.*

PROOF. Write $u = \sum u(g)g$. Then $u^* = \sum u(g)g^{-1}$ is also central and of finite order. Then $uu^* = (\sum_g (u(g))^2) \cdot 1 + \cdots$ is a torsion unit with nonzero trace. It follows that $uu^* = \pm 1$. Since $\varepsilon(uu^*) = 1$ we must have $uu^* = 1$, consequently, $\sum (u(g))^2 = 1$. It follows that there is a unique $u(g_0) \neq 0$. Thus $u = \pm g_0$, as claimed. □

REMARK 3.12. Another proof of the above corollary is contained in [PSe3].

COROLLARY 3.13 [BMS]. *Suppose that $u = \sum u(g)g \in \boldsymbol{Z}G$ is a torsion unit with a central element z in the support of u. Then $u = \pm z$.*

PROOF. See [Se13, p. 286]. □

Another application is the next result of Zassenhaus [Zas3] and Sehgal [Se8].

THEOREM 3.14. *Let $u = \sum u(g)g \in \mathcal{U}_1(\boldsymbol{Z}G)$ be an element of order p^n where p is a prime and G is arbitrary. Then there exists an element of order p^n in the support of u.*

PROOF. We have $u^{p^m} \equiv \sum u(g)^{p^m} g^{p^m} + \lambda \pmod{p\mathbf{Z}G}$ with $\lambda \in [\mathbf{Z}G, \mathbf{Z}G]$. Note that $\lambda(1) = 0$. We conclude that for all m

$$u^{p^m}(1) \equiv \sum_{g^{p^m}=1} u(g) \pmod{p}.$$

Thus

$$1 = u^{p^n}(1) \equiv \sum_{g^{p^n}=1} u(g) \equiv \sum_{g^{p^{n-1}}=1} u(g) + \sum_{o(g)=p^n} u(g) \pmod{p}.$$

The first sum on the right-hand side is zero as it is congruent to the trace of $u^{p^{n-1}} \neq 1$. It follows that there exists a $g \in G$ with $o(g) = p^n$ and $u(g) \neq 0$. □

COROLLARY 3.15. *$\mathcal{U}_1(\mathbf{Z}G)$ is torsion free if and only if G is torsion free.*

For torsion free groups there is the famous and fascinating question of Kaplansky:

THE UNIT CONJECTURE 3.16. *The unit group of KG, where K is a field and G is a torsion free group, is trivial, namely, all units are of the form kg, $k \in K^\times$, $g \in G$.*

As of now (Fall 2000) we cannot say much about this problem; only a few easy, nevertheless, useful cases have been solved.

THEOREM 3.17. *Let G be an ordered group and K a field. Then the unit group $\mathcal{U}(KG)$ is trivial.*

PROOF. Let $\gamma, \mu \in KG$ with $\gamma\mu = 1$. If $\operatorname{supp}(\gamma)$ or $\operatorname{supp}(\mu)$ has only one element there is nothing to prove. So let us assume the contrary and write

$$\begin{aligned} \gamma &= c_1 g_1 + c_2 g_2 + \cdots + c_r g_r \quad \text{with } g_1 < g_2 < \cdots < g_r,\ r \geqslant 2, \\ \mu &= b_1 h_1 + b_2 h_2 + \cdots + b_s h_s \quad \text{with } h_1 < h_2 < \cdots < h_s,\ s \geqslant 2. \end{aligned}$$

Since $g_1 < g_i$ for all $i > 1$ we have $g_1 h_1 < g_1 h_j < g_i h_j$ for all j. Thus $g_1 h_1$ is the smallest element among the products $\{g_i h_j\}$ and by symmetry we conclude that $g_r h_s$ is the largest of this set. Since $\gamma\mu = 1$, both these elements cannot survive in the product. We have a contradiction, proving the theorem. □

A slight modification of this argument proves the theorem for right ordered groups (see [P21, p. 588]). The first example of an ordered group is a torsion free Abelian group. Also, torsion free nilpotent groups can be ordered. Moreover, a group G having a subnormal series

$$G = G_0 \triangleright G_1 \triangleright G_2 \triangleright \cdots \triangleright G_n = 1$$

with each G_i/G_{i+1} torsion free Abelian can be right ordered. Thus group algebras KG of these groups G also have trivial units. This was proved by Bovdi [Bov1]. Beyond this, there are no results about this conjecture. One reason for this could be that the unit conjecture is stronger than the zero divisor conjecture. On the latter question we have a great deal of progress. We state

THEOREM 3.18. *If G is a torsion free group and K is a field then*

$$\mathcal{U}(KG) \text{ trivial} \quad \Rightarrow \quad KG \text{ has no zero divisors.}$$

PROOF. Suppose that $\mathcal{U}(KG) = K^{\times} \cdot G$. Suppose there exists nonzero elements α and β in KG with $\alpha\beta = 0$. We know that KG is a prime ring. So $\beta(KG)\alpha \neq 0$ which means there is an $x \in G$ with $\gamma = \beta x \alpha \neq 0$. Clearly, $\gamma^2 = 0$ and $(1-\gamma)(1+\gamma) = 1$. Therefore, by assumption $1 - \gamma = kg$ for some $k \in K$, $g \in G$, $\gamma = 1 - kg$ which contradicts $\gamma^2 = 0$. □

The easiest supersolvable, torsion free, not right orderable group is given by

$$\Gamma = \langle x, y \mid x^{-1}y^2x = y^{-2},\ y^{-1}x^2y = x^{-2} \rangle.$$

Then $A = \langle x^2, y^2, (xy)^2 \rangle$ is a normal free Abelian subgroup of rank 3 with $\Gamma/A \simeq V_4$, the Klein four group. Moreover, Γ is torsion free but not right orderable. For the proof, see [P21, p. 606]. It is known (see Formanek's theorem (11.14)) that $K\Gamma$ has no zero divisors.

PROBLEM 3.19. Let K be a field and let Γ be the above group. Are the units of $K\Gamma$ trivial?

Let us return for a moment to ordered groups. Let

$$K[\![x]\!] = \left\{ \sum_s^{\infty} a_i x^i \;\middle|\; a_i \in K,\ s \in \mathbf{Z} \right\}$$

be the ring of Laurent series in x over a field K. In fact, $K[\![x]\!]$ is a division ring and it contains the group ring $K\langle x \rangle$ of the infinite cyclic group $\langle x \rangle$. In this way, considering the formal sums $\sum_{g \in G} a_g g$, $a_g \in K$, where G is an ordered group and $K[\![G]\!]$ is the ring of those formal sums which have well ordered support, Malcev [Mal1] and Neumann [Ne1] proved that $K[\![G]\!]$ is a division ring (see also, [P21, p. 601]). We have

THEOREM 3.20 (Malcev–Neumann). *The group algebra of an ordered group over a field can be embedded in a division ring.*

It is not known whether this 'old' theorem can be extended to right ordered groups. More generally, we have the

PROBLEM 3.21. Let G be a torsion free group and K a field. Can KG be embedded in a division ring?

There are two more observations regarding the unit conjecture which, hopefully, will have some applications in the future.

THEOREM 3.22 (Cliff–Sehgal). *Let G be a torsion free group. Then $G \lhd \mathcal{U}(\boldsymbol{Z}G) \Leftrightarrow \mathcal{U}(\boldsymbol{Z}G) = \pm G$.*

PROOF. See [CS3]. □

The next observation, due to Marciniak and Sehgal [MS6], concerns crystallographic groups. Recall that a crystallographic group is a discrete, co-compact subgroup of the group of Euclidean isometries $\mathrm{ISOM}(n)$ of $\boldsymbol{R}^n$. In 1910 Bieberbach characterized those groups in algebraic terms: a group Γ is crystallographic if and only if it has a finite index, normal subgroup A isomorphic to $\boldsymbol{Z}^n$ and A is maximal among the Abelian subgroups in Γ.

By its very nature, the group Γ acts by isometries on $\boldsymbol{R}^n$. In the next lemma we observe that this action extends naturally to an action of the unit group $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$.

LEMMA 3.23. *There exists an action of $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ on $\boldsymbol{R}^n$ by affine transformations extending the original isometric action of Γ.*

PROOF. Let $\mathrm{Aff}(n)$ denote the group of invertible affine transformations of $\boldsymbol{R}^n$. By a well known trick one can embed this group into $GL_{n+1}(\boldsymbol{R})$: one sends the map $x \mapsto Ax + b$ to the matrix

$$\begin{bmatrix} A & b \\ 0\ldots0 & 1 \end{bmatrix}.$$

Let $M_{n+1}(\boldsymbol{R})$ be the ring of all matrices of degree $n+1$. We then have the following sequence of multiplicative embeddings:

$$\Gamma < \mathrm{ISOM}(n) < \mathrm{Aff}(n) < GL_{n+1}(\boldsymbol{R}) \subset M_{n+1}(\boldsymbol{R}).$$

The above composition of embeddings can be extended by linearity to a commutative square of multiplicative maps:

$$\begin{array}{ccc} \Gamma & \longrightarrow & \mathrm{ISOM}(n) \\ \downarrow & & \downarrow \\ \boldsymbol{Z}\Gamma & \longrightarrow & M_{n+1}(\boldsymbol{R}). \end{array}$$

Notice that a unit $u = \sum u_g g \in \mathcal{U}_1(\boldsymbol{Z}\Gamma) \subset \boldsymbol{Z}\Gamma$ maps to an invertible, integral matrix which has the bottom row filled with zeros except the last position, where we find

$\sum u_g \cdot 1 = 1$. Thus the image of $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ is contained in $\mathrm{Aff}(n) \subset M_{n+1}(\boldsymbol{R})$ and we obtain the desired diagram

$$\begin{tikzcd} \Gamma \arrow[r] \arrow[d] & \mathrm{ISOM}(n) \arrow[d] \\ \mathcal{U}_1(\boldsymbol{Z}\Gamma) \arrow[r] & \mathrm{Aff}(n) \end{tikzcd}$$

which concludes the proof. □

Usually, an action of an infinite group on Euclidean space can be quite nasty. Luckily, here this is not the case. We will give now another description of the above action from which it will be evident.

It is well known that every crystallographic group Γ is a finite index subgroup in another crystallographic group Γ^* which is a split extension. To achieve this, it is enough to squeeze the lattice $A = \boldsymbol{Z}^n$ by the factor $|G|$, where $G = \Gamma/A$:

$$\begin{tikzcd} 1 \arrow[r] & A \arrow[r] \arrow[d, "\cdot|G|"] & \Gamma \arrow[r] \arrow[d] & G \arrow[r] \arrow[d, "="] & 1 \\ 1 \arrow[r] & A^* \arrow[r] & \Gamma^* \arrow[r] & G \arrow[r] & 1. \end{tikzcd}$$

As $\mathcal{U}_1(\boldsymbol{Z}\Gamma) \subseteq \mathcal{U}_1(\boldsymbol{Z}\Gamma^*)$, it is enough to define an action of $\mathcal{U}_1(\boldsymbol{Z}\Gamma^*)$ on $\boldsymbol{R}^n$ by affine transformations.

Notice that the group $\mathcal{U}_1(\boldsymbol{Z}\Gamma^*)$ acts naturally on the set $\boldsymbol{Z}A^*$. In fact, let u be a unit from $\mathcal{U}_1(\boldsymbol{Z}\Gamma^*)$. We write it as $u = \sum u_g g$ with $g \in G$, $u_g \in \boldsymbol{Z}A^*$, and define the action by setting $u(x) = \sum u_g \cdot x^g$ for $x \in \boldsymbol{Z}A^*$. Here by x^g we mean the linear extension to all of $\boldsymbol{Z}A^*$ of the action of $g \in G$ on A^* (by an inner automorphism in Γ).

From the definition of the action it is easy to see that if $I \subset \boldsymbol{Z}A^*$ is a G-invariant ideal and $x \in I$ then also $u(x) \in I$. In particular, this applies to the augmentation ideal $\Delta(A^*)$ and all its powers. Hence the group $\mathcal{U}_1(\boldsymbol{Z}\Gamma^*)$ acts on the $\boldsymbol{Z}$-module $\Delta(A^*)/\Delta(A^*)^2 \simeq A^*$.

The above action can be extended by linearity from A^* to $A^* \otimes_{\boldsymbol{Z}} \boldsymbol{R} \simeq \boldsymbol{R}^n$. By looking at the action of translations $a \in A$ and rotations $g \in G$ separately, it is not difficult to see that, in fact, we have defined the same action of $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ as before.

COROLLARY 3.24. *As the action of $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ on $\boldsymbol{R}^n$ is induced from an action on a lattice $\boldsymbol{Z}^n$, it is simplicial with respect to a locally finite triangulation of $\boldsymbol{R}^n$.*

REMARK 3.25. It is also interesting to look at the action of $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ on the higher factors $\Delta(A^*)^n/\Delta(A^*)^{n+1}$. This leads to an "infinitesimal" study of the action in the spirit of Taylor's expansion with respect to the Fox derivatives, as practiced in combinatorial group theory.

We are going to show that for crystallographic groups Γ the unit conjecture is strongly related to the properties of the above action of $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ on the Euclidean space.

To simplify the arguments we will restrict our attention to oriented crystallographic groups only. Those are groups which act on $\boldsymbol{R}^n$ by isometries which preserve orientation.

Recall that a group G acts on a set X freely if from $g(x) = x$ it follows that $g = e$. In other words, e is the only element which has fixed points.

THEOREM 3.26. *Let Γ be a torsion free, oriented crystallographic group. All units in $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ are trivial if and only if $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ acts freely on $\boldsymbol{R}^n$.*

PROOF. It is well known that a torsion free crystallographic group acts freely on $\boldsymbol{R}^n$. Hence, if $\mathcal{U}_1(\boldsymbol{Z}\Gamma) = \Gamma$ then $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ acts freely on $\boldsymbol{R}^n$ as well.

Assume now that the above described action of $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ on $\boldsymbol{R}^n$ is free. Even though Γ preserves orientation of $\boldsymbol{R}^n$, there is no a priori reason for $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ to do the same. Let then $\mathcal{U}_0 \subset \mathcal{U}_1(\boldsymbol{Z}\Gamma)$ be the subgroup consisting of those (affine) transformations which preserve the orientation of $\boldsymbol{R}^n$. By assumption, $\Gamma \subset \mathcal{U}_0$.

As $\mathcal{U}_0$ acts freely and simplicially on a triangulation coming from a lattice, there exists $\varepsilon > 0$ such that for every $g \in \mathcal{U}_0$ and $x \in \boldsymbol{R}^n$, $\operatorname{dist}(x, g(x)) \geqslant \varepsilon$. In particular, $\mathcal{U}_0$ acts properly discontinuously on the Euclidean space. Therefore the natural map $p : \boldsymbol{R}^n \to \boldsymbol{R}^n/\mathcal{U}_0$ is a covering and $\widetilde{M}^n = \boldsymbol{R}^n/\mathcal{U}_0$ is a smooth, connected, orientable manifold of dimension n.

Dividing $\boldsymbol{R}^n$ first by Γ and then by $\mathcal{U}_0$ leads to a factorization of p through a compact, orientable manifold $M^n = \boldsymbol{R}^n/\Gamma$. It follows that the manifold $\widetilde{M}^n$ is compact as well. In particular, the covering $p : M^n \to \widetilde{M}^n$ has a finite number of sheets, say k. From the elementary covering theory we know that

$$k = \left|\pi_1\left(\widetilde{M}^n\right) : \pi_1\left(M^n\right)\right| = |\mathcal{U}_0 : \Gamma|.$$

We are going to prove that $k = 1$.

It is well known from topology that $H_n(M^n, \mathbf{Z}) \simeq H_n(\widetilde{M}^n, \mathbf{Z}) \simeq \mathbf{Z}$ and our covering p induces a homomorphism $p_* : H_n(M^n, \mathbf{Z}) \to H_n(\widetilde{M}^n, \mathbf{Z})$ which is a multiplication by k.

Now, both manifolds are covered by $\boldsymbol{R}^n$ and hence they are $K(\pi, 1)$ spaces. In particular, the homology of those manifolds coincides with the homology of their fundamental groups. Thus $p_* : H_n(\Gamma, \mathbf{Z}) \to H_n(\mathcal{U}_0, \mathbf{Z})$, $p_*(x) = k \cdot x$, is induced by the inclusion $\Gamma < \mathcal{U}_0$.

However, group homology is really an augmented group ring invariant, not just a group invariant. So let us look carefully at the inclusion $i : \boldsymbol{Z}\Gamma \to \boldsymbol{Z}\mathcal{U}_0$. Consider the map $s : \boldsymbol{Z}\mathcal{U}_0 \to \boldsymbol{Z}\Gamma$ which is the linear extension of the natural embedding $\mathcal{U}_0 < \mathcal{U}_1(\boldsymbol{Z}\Gamma) \subset \boldsymbol{Z}\Gamma$. Both i and s are augmented ring homomorphisms. Moreover, $s \circ i$ is the identity on $\mathbb{Z}\Gamma$ as can be easily seen by looking at the images of elements of Γ. Thus the map $i : \boldsymbol{Z}\Gamma \to \mathcal{U}_0$ has a right inverse s. Consequently, the map $p_* : H_n(\Gamma, \mathbf{Z}) \to H_n(\mathcal{U}_0, \mathbf{Z})$ has a right inverse s_*. But a multiplication by $k > 0$ on $\boldsymbol{Z}$ has a right inverse only if $k = 1$. Therefore we have $\Gamma = \mathcal{U}_0$.

As $|\mathcal{U}_1(\boldsymbol{Z}\Gamma) : \mathcal{U}_0| \leqslant 2$, we have already proved that $|\mathcal{U}_1(\boldsymbol{Z}\Gamma) : \Gamma| \leqslant 2$. We will show now that in that case $\mathcal{U}_1(\boldsymbol{Z}\Gamma) = \Gamma$. To this end take any unit $u = \sum u_g g \in \mathcal{U}_1(\boldsymbol{Z}\Gamma)$. Then we have $u^* u = g \in \Gamma$. As $0 \neq \operatorname{tr}(u^* u) = \operatorname{tr}(g)$, so $g = 1$. Hence $u^* u = 1$ and $1 = \operatorname{tr}(u^* u) = \sum u_g^2$. Thus $u_g \neq 0$ can happen for a single integer u_g only and hence $u \in \Gamma$. $\square$

Let A be an Abelian group and T the torsion subgroup. Then the units of $\boldsymbol{Z}A$ are just the obvious ones:

THEOREM 3.27 (Sehgal). $\mathcal{U}(\boldsymbol{Z}A) = \mathcal{U}(\boldsymbol{Z}T) \cdot A$.

PROOF. See [Se10, p. 194] or [Se4]. □

One ingredient in the proof is the fact that $\boldsymbol{Z}A$ has no nilpotent elements. This is obvious for finite A as $\boldsymbol{Z}A$ is contained in a direct sum of fields. Finitely generated Abelian groups are residually finite and thus their group rings (and hence those of arbitrary Abelian A) cannot have nilpotent elements.

Suppose that G is arbitrary and $\boldsymbol{Z}G$ has no nilpotent elements. Let $e = e^2 \in \boldsymbol{Q}G$. Then $ex(1-e)$ for $x \in \boldsymbol{Q}G$ is a square-zero element of $\boldsymbol{Q}G$. Consequently, $ex(1-e) = 0$ and $ex = exe$. Similarly $xe = exe$ and e is central.

Now, if $t \in G$ is an element of finite order n then $\hat{t}/n = \frac{1}{n}(\sum_0^{n-1} t^i)$ is an idempotent. The centrality of this idempotent implies that $\langle t \rangle$ is normal in G. Thus the torsion elements of G form a subgroup T which is Abelian or Hamiltonian: $T = A \times E \times K_8$ where A is Abelian with every element of odd order, E is an elementary Abelian 2-group and K_8 is the quaternion group of order 8. We have the following theorem, proved for finite groups by Pascaud [Pa] and in general by Sehgal [Se9].

THEOREM 3.28. *Suppose that $\boldsymbol{Z}G$ has no nilpotent elements. Then*

(1) *every idempotent of $\boldsymbol{Q}G$ is central, every finite subgroup of G is normal and thus the torsion elements of G form a subgroup T;*

(2) *if T is not Abelian, $T = A \times E \times K_8$ and the order of* 2 mod m *is odd for every element of order m in A.*

Conversely, if G satisfies (1) *and* (2) *and G/T is nilpotent then $\boldsymbol{Z}G$ has no nilpotent elements.*

REMARKS 3.29. It is clear that, in the absence of an answer to the unit conjecture, we need an assumption on G/T for the converse. For convenience we have assumed that G/T is nilpotent. It could be easily replaced by 'ordered' or generalizations thereof. The proof is dependent on a number theoretic result about representing -1 as a sum of squares in cyclotomic fields, namely,

LEMMA 3.30. *Let $\zeta = \zeta_m$ be a primitive mth root of unity. Then $-1 = x^2 + y^2$ has a solution in $\boldsymbol{Q}(\zeta)$ if and only if the order of* 2 mod m *is even.*

For the proof, see, for example, [Mos].

Theorem 3.27 can be extended to some noncommutative situations:

THEOREM 3.31 (Sehgal–Zassenhaus). *Suppose that every idempotent of $\boldsymbol{Q}G$ is central so the torsion elements of G form a group T. If G/T is nilpotent then $\mathcal{U}(\boldsymbol{Z}G) = \mathcal{U}(\boldsymbol{Z}T) \cdot G$.*

PROOF. See [Se10, p. 194]. □

As remarked above, the restriction G/T 'nilpotent' can be relaxed a little.

We have mentioned earlier that the unit group of the integral group ring of a finite group is finitely generated. This is no longer true even for finitely generated nilpotent groups. We give an easy example.

THEOREM 3.32. *Let $\Gamma = D_8 \times C_\infty$ where D_8 is the dihedral group of order* 8 *and C_∞ is the infinite cyclic group. Then the group of normalized units $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ is not finitely generated.*

We fix the presentation $D_8 = \langle x, y\colon\ x^4 = 1,\ y^2 = 1,\ yxy^{-1} = x^{-1}\rangle$. We write $z = x^2 \in D_8$. Easy calculations show that $\mathcal{U}_1\boldsymbol{Z}\Gamma = V \rtimes \Gamma$, where $V = \{u \in \mathcal{U}_1(\boldsymbol{Z}\Gamma)\colon\ u = 1 + a(z-1),\ a \in \boldsymbol{Z}\Gamma \text{ with } \varepsilon(a) = \text{even}\}$.

LEMMA 3.33. *If $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$ is finitely generated then V is finitely generated as well.*

PROOF. VD_8 is a subgroup of $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$. As C_∞ is central in $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$, so VD_8 is normal in $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$. Because also $(VD_8) \cap C_\infty = 1$, it follows that $\mathcal{U}_1(\boldsymbol{Z}\Gamma) = (VD_8) \times C_\infty$. Therefore VD_8 is a homomorphic image of $\mathcal{U}_1(\boldsymbol{Z}\Gamma)$, and hence it is finitely generated. Finally, V is the kernel of the epimorphism $VD_8 \to D_8$, hence it is of finite index in VD_8. Therefore V is also finitely generated. $\square$

Let us write $R = \boldsymbol{Z}[C_\infty] = \boldsymbol{Z}[t, t^{-1}]$.

LEMMA 3.34. *The group V is isomorphic to a subgroup of*

$$SL_2(R) \cap \begin{bmatrix} 1+2R & 4R \\ 2R & 1+2R \end{bmatrix}.$$

PROOF. We select a Wedderburn isomorphism

$$\phi : \boldsymbol{Q}D_8 \xrightarrow{\simeq} \boldsymbol{Q}\big[D_8/\langle z\rangle\big] \oplus M_2(\boldsymbol{Q})$$

as follows. For any $a \in \boldsymbol{Q}D_8$ we set $\phi(a) = (\phi_1(a), \phi_2(a))$ where ϕ_1 is induced by the natural epimorphism $D_8 \to D_8/\langle z\rangle$ and ϕ_2 is the linear representation

$$\phi_2(x) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \qquad \phi_2(y) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

After tensoring with $\boldsymbol{Q}C_\infty$ we obtain

$$\phi \otimes \mathrm{id} : \boldsymbol{Q}\Gamma \xrightarrow{\simeq} \boldsymbol{Q}\big[\Gamma/\langle z\rangle\big] \oplus M_2(\boldsymbol{Q}C_\infty).$$

Now we have $V \subset \mathcal{U}_1\boldsymbol{Z}\Gamma \subset \boldsymbol{Q}\Gamma$. From the definition of V it follows that $\phi_1(V) = \langle 1\rangle$. Therefore we have an embedding $\phi_2 : V \to GL_2(\boldsymbol{Q}C_\infty)$.

We shall describe this embedding explicitly. Any unit $u \in V$ can be written in the form $u = 1 + (a_0 + a_1x + a_2y + a_3xy)(z-1)$ with $a_i \in R$ and $\sum \varepsilon(a_i)$ even. Then

$$\phi_2(u) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \left(a_0 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + a_1 \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} + a_2 \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \right.$$
$$\left. + a_3 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right) \cdot \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$$

and hence

$$\phi_2(u) = \begin{bmatrix} 1 - 2(a_0 + a_2) & 2(a_1 - a_3) \\ -2(a_1 + a_3) & 1 - 2(a_0 - a_2) \end{bmatrix}.$$

In this way we have embedded V into the subgroup $GL_2(R) \subseteq GL_2(\boldsymbol{Q}C_\infty)$.

After we conjugate with $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ we obtain a new embedding ψ with

$$\psi(u) = \begin{bmatrix} 1 - 2(a_0 + a_1 + a_2 + a_3) & 4(a_1 + a_2) \\ -2(a_1 + a_3) & 1 - 2(-a_0 + a_1 + a_2 + a_3) \end{bmatrix}$$
$$\in \begin{bmatrix} 1 + 2R & 4R \\ 2R & 1 + 2R \end{bmatrix}.$$

Let us write $W = \psi(V) \subseteq GL_2(R)$. It remains to show that each matrix from W has determinant one.

To this end, take any matrix $A \in W$. After we augment the entries of A, we obtain an invertible integral matrix of the form

$$\varepsilon(A) = \begin{bmatrix} 1 + 4\alpha & 4\beta \\ 2\gamma & 1 + 4\delta \end{bmatrix},$$

as $2 | \sum \varepsilon(a_i)$. Therefore $\det(\varepsilon(A)) \equiv 1 \pmod 4$ and $\det(\varepsilon(A)) = \pm 1$. It follows that $\det(\varepsilon(A)) = 1$.

Let $d \in R$ be the determinant of the matrix A. We have already checked that $\varepsilon(d) = 1$. As d is a unit of R of augmentation 1, it is just a group element: $d \in C_\infty$. It is easy to see from the determinant formula that the coefficient of 1 in d is odd and hence $d = 1$. Thus ψ embeds V into $SL_2(R)$, as desired. $\square$

Let us denote by W the image $\psi(V) \subseteq SL_2(R)$. We exhibit now a sequence of matrices from W.

LEMMA 3.35. *For any integer k the matrix*

$$A_k = \begin{bmatrix} 1 & 0 \\ -2t^k & 1 \end{bmatrix}$$

belongs to W.

PROOF. Notice that the element $n = (x - y)(z - 1) \in \mathbf{Z}D_8$ is nilpotent. In fact,

$$\begin{aligned} n^2 &= (x-y)^2(z-1)^2 = (z - xy - yx + 1)(2 - 2z) \\ &= -2\big(z + 1 - yx^{-1} - yx\big)(z-1) \\ &= -2(1 - yx^{-1})(z+1)(z-1) = 0. \end{aligned}$$

Therefore the elements $u_k = 1 + t^k n$ are units and clearly they belong to V. It is easy to see that $\psi(u_k) = A_k$. □

PROOF OF THEOREM 3.32. Suppose that the group $\mathcal{U}_1(\mathbf{Z}\Gamma)$ is finitely generated. From Lemma 3.33 it follows that V, and hence W, is finitely generated.

Let $\mu : R \to \overline{R}$ be the reduction modulo 4, i.e., $\overline{R} = (\mathbf{Z}/4\mathbf{Z})[C_\infty]$. Let $\overline{W}$ be the image $\mu(W) \subseteq SL_2(\overline{R})$. Clearly, this group is also finitely generated.

We show that $\overline{W}$ is Abelian. In fact, it consists of elements of the form $\left[\begin{smallmatrix} 1+2a & 0 \\ 2c & 1+2d \end{smallmatrix}\right]$ with $a, c, d \in \overline{R}$. Hence, when we multiply two elements of $\overline{W}$ we obtain

$$\begin{bmatrix} 1+2a & 0 \\ 2c & 1+2d \end{bmatrix} \cdot \begin{bmatrix} 1+2a' & 0 \\ 2c' & 1+2d' \end{bmatrix} = \begin{bmatrix} 1+2(a+a') & 0 \\ 2(c+c') & 1+2(d+d') \end{bmatrix}$$

and the result does not depend on the order of matrices.

Thus, $\overline{W}$ is a finitely generated Abelian group. However, it contains an infinite elementary Abelian 2-group $\langle \mu(A_k) \mid k \in \mathbf{Z} \rangle \subseteq \overline{W}$. This contradiction proves that $\mathcal{U}_1(\mathbf{Z}\Gamma)$ is not finitely generated. □

Surprisingly, it is possible to classify nilpotent groups G whose unit groups $\mathcal{U}(\mathbf{Z}G)$ are finitely generated. Only very special infinite nilpotent groups have this property. The result is

THEOREM 3.36 (Wiechecki). *Let G be a nilpotent group. Then $\mathcal{U}(\mathbf{Z}G)$ is finitely generated if and only if either G is finite or G is finitely generated and there are no nilpotent elements in $\mathbf{Z}G$.*

Normal complements

We shall now study extensions of the theorem of Higman to finite noncommutative groups. There are two parts to this theorem. The first says that the torsion units are trivial. This we shall generalize to noncommutative groups G later in Section 8. The second part, which we wish to generalize now, concerns the direct decomposition. In the noncommutative situation the question is whether there exists a normal torsion free group N so that we have a semidirect product $\mathcal{U}_1(\mathbf{Z}G) = N \rtimes G$. In other words, the question posed by Dennis [Den2] is:

(a) Does the inclusion $G \to \mathcal{U}_1(\mathbf{Z}G)$ split?

(b) If a splitting exists, is its kernel torsion free?

An affirmative answer to both these questions has applications to the famous isomorphism problem:

(ISO) Does the ring isomorphism $\boldsymbol{Z}G \simeq \boldsymbol{Z}H$ imply the group isomorphism $G \simeq H$?

REMARK. If we are given an isomorphism $\sigma : \boldsymbol{Z}G \to \boldsymbol{Z}H$ it may not preserve augmentation. It is possible to modify this isomorphism to $\mu : \boldsymbol{Z}G \to \boldsymbol{Z}H$ by setting $\mu(g) = (\varepsilon(g))^{-1}\sigma(g)$, so that $\varepsilon(\mu(\alpha)) = \varepsilon(\alpha)$ for all α. We call such an isomorphism normalized. All our isomorphisms σ will be normalized.

PROPOSITION 3.37. (i) *An affirmative answer to* (a) *and* (b) *above implies an affirmative answer to* (ISO).

(ii) *For finite nilpotent groups, an affirmative answer to* (a) *above implies an affirmative answer to* (ISO).

PROOF. Let $\theta : \boldsymbol{Z}G \to \boldsymbol{Z}H$ be an isomorphism. We can assume θ to be augmentation preserving and that G and H are nilpotent together. Write $\mathcal{U}_1(\boldsymbol{Z}H) = N \rtimes H$. For $g \in G$ if $\theta(g) = nh$, $n \in N$, $h \in H$ we define $\alpha(g) = h$. Then $\alpha : G \to H$ is a group homomorphism.

(i) In this case, $\operatorname{Ker}\alpha = \{g \in G\colon\ \theta(g) \in N\} = 1$ as N is torsion free. Thus $G \simeq H$ as they both have the same orders.

(ii) Since G is nilpotent, its centre $\mathfrak{z} \neq 1$. Consequently, if $\operatorname{Ker}\alpha \neq 1$ then $\operatorname{Ker}\alpha \cap \mathfrak{z} \neq 1$. Let z be a nontrivial element in $\operatorname{Ker}\alpha \cap \mathfrak{z}$. Then $\theta(z)$ being a torsion element in the centre of $\boldsymbol{Z}H$ belongs to H. Thus $\theta(z) \in H \cap N = 1$ and therefore $z = 1$, a contradiction. Hence $G \simeq H$. $\square$

The questions were answered affirmatively for circle groups of nilpotent rings by Sandling [San6] and Passman and Smith [PSm] who also handled groups having an Abelian subgroup of index 2. The latter was greatly extended by Cliff, Sehgal and Weiss [CSW]. We begin with an easy way of obtaining a normal complement in some cases. We need the following result which is called the Whitcomb argument.

LEMMA 3.38. *Suppose an element γ of $\boldsymbol{Z}G$ satisfies $\gamma \equiv g$* mod $\Delta(G, A)$ *where $g \in G$ and $A \lhd G$. Then $\gamma \equiv ga_0$* mod $\Delta(G)\Delta(A)$ *for a suitable $a_0 \in A$.*

PROOF. We have

$$\begin{aligned}
\gamma &= g + \delta, \quad \delta \in \Delta(G, A) \\
&\equiv g + \sum \alpha(a-1), \quad \alpha \in \boldsymbol{Z}G,\ a \in A \\
&\equiv g + \sum n(a-1) \bmod \bigl(\Delta(G)\Delta(A)\bigr), \quad \varepsilon(a) = n \in \boldsymbol{Z},\ a \in A \\
&\equiv g + (a_0 - 1) \bmod \bigl(\Delta(G)\Delta(A)\bigr) \quad \text{for some } a_0 \in A \\
&\equiv ga_0 \bmod \bigl(\Delta(G)\Delta(A)\bigr).
\end{aligned}$$

The last two congruences are consequences of the identity

$$xy - 1 = (x-1) + (y-1) + (x-1)(y-1). \qquad \square$$

The first elementary case where we can find a normal complement to G in $\mathcal{U}_1(\boldsymbol{Z}G)$ is when $\boldsymbol{Z}(G/A)$ has trivial units with A Abelian normal.

LEMMA 3.39. *Suppose that $A \lhd G$ is Abelian and G/A is Abelian of exponent* 2, 3, 4 *or* 6. *Then* $\mathcal{U}_1(\boldsymbol{Z}G) = N \rtimes G$ *where* $N = \mathcal{U}(1 + \Delta(G)\Delta(A))$.

PROOF. Since $\boldsymbol{Z}\overline{G}$, $\overline{G} = G/A$, has only trivial units by (3.4), for any unit u of $\boldsymbol{Z}G$ we have $\bar{u} = \bar{g}$, $g \in G$. Thus $u \equiv g \bmod(\Delta(G, A))$. It follows by the Whitcomb argument that $u \equiv g_0 \bmod(\Delta(G)\Delta(A))$, $g_0 \in G$. We have seen that ug_0^{-1} is a unit of $1 + \Delta(G)\Delta(A)$. Since $\Delta(G)\Delta(A)$ is an ideal of $\boldsymbol{Z}G$, N is normal in $\mathcal{U}(\boldsymbol{Z}G)$. Then a standard argument gives $G \cap N = 1$. $\square$

Let us see how we can attempt to extend the above arguments to the metacyclic group

$$G = \langle a \rangle \rtimes \langle x \rangle, \qquad o(x) = m, \qquad a^x = a^j, \qquad o(j) \equiv m \bmod\bigl(o(a)\bigr).$$

We write $\langle a \rangle = A$ and $\langle x \rangle = X$. It can be proved that

$$\mathcal{U}_1(\boldsymbol{Z}X) = X \times V, \quad V \subseteq \mathcal{U}_*(\boldsymbol{Z}X),$$

the $*$-symmetric units of $\boldsymbol{Z}X$. The Whitcomb argument, as in the last lemma, gives

$$\mathcal{U}_1(\boldsymbol{Z}G) = VG\mathcal{U}\bigl(1 + \Delta(G)\Delta(A)\bigr).$$

The difficulty is that V may not be normalized by $G(1 + \Delta(G)\Delta(A))$. We try to modify $\Delta(G)\Delta(A)$ to J so that G normalizes $V \bmod(1 + J)$. We wish to find an integer n (relatively prime to $o(a)$) such that the ideal

$$J_n = \bigl\langle (x - n)\Delta(A) + \Delta(A)^2 \bigr\rangle$$

has the property that V is normal in G modulo $1 + J_n$. Since $v^* = v$ for all $v \in V$ it is enough to verify that J_n contains

$$\gamma = a\bigl(x + x^{-1}\bigr)a^{-1} - \bigl(x + x^{-1}\bigr).$$

Writing $a^x = xax^{-1}$ we have

$$\begin{aligned}\gamma &= x\bigl(a^{x-1} - 1\bigr) + x^{-1}\bigl(a^{x^{-1}-1} - 1\bigr) = x\bigl(a^{j-1} - 1\bigr) + x^{-1}\bigl(a^{j^{-1}-1} - 1\bigr)\\ &= n\bigl(a^{j-1} - 1\bigr) + n^{-1}\bigl(a^{j^{-1}-1} - 1\bigr) \pmod{J_n}\\ &= a^{n(j-1)+n^{-1}(j^{-1}-1)} - 1 \pmod{J_n}.\end{aligned}$$

It suffices to find n such that

$$n(j-1) + n^{-1}\bigl(j^{-1} - 1\bigr) = 0 \ \bigl(\bmod\, o(a)\bigr),$$

equivalently,

$$\left(n^2 j-1\right)(j-1)=0 \left(\bmod o(a)\right).$$

It is enough to find n so that $n^2 j = 1 \pmod{o(a)}$. This we can do if $m = o(x)$ is an odd number $2k+1$, namely, take $n = j^k$.

We observe that $\mathcal{U}_1(\boldsymbol{Z}G) = VG\mathcal{U}(1+\Delta(G)\Delta(A))$ implies by an easy computation that $\mathcal{U}_1(\boldsymbol{Z}G) = VG\mathcal{U}(1+J_n)$ for any n. Thus for $n = j^k$, $\mathcal{U}_1(\boldsymbol{Z}G) = GV\mathcal{U}(1+J_n) = GN$ where N is normal in $\mathcal{U}_1(\boldsymbol{Z}G)$.

It is an easy exercise to check that J_{j^ℓ} is the kernel of the homomorphism

$$\phi_\ell : \Delta(G,A) \to A, \quad \phi_\ell\big(y(b-1)\big) = b^{y^\ell}, \quad b \in A,\ y \in X.$$

In general, let G be metabelian with $A \triangleleft G$ and both A and G/A Abelian. For any $k \in \boldsymbol{Z}$ we have a homomorphism of Abelian groups

$$\phi_k : \Delta(G,A) \to A, \quad \phi_k\big(x(a-1)\big) = a^{x^k},$$

where $a \in A$ and x is a coset representative of $G \bmod A$. Then we define $I_k = I_k(G,A) = \operatorname{Ker}\phi_k$. It is easy to see that

$$\Delta(G,A)^2 \subseteq I_k \subseteq \Delta(G,A), \qquad I_0 = \Delta(G)\Delta(A).$$

The next difficult theorem was considered for $k=0$ by Jackson [Jak], although his proof is not complete.

THEOREM 3.40 (Cliff–Sehgal–Weiss). *Let $A \triangleleft G$ with A and G/A both finite Abelian. For any $k \in \boldsymbol{Z}$, $\mathcal{U}(1+I_k)$ is torsion free.*

The most general positive answer known for the question of Dennis is

THEOREM 3.41 (Cliff–Sehgal–Weiss). *Let G be a finite group having an Abelian normal subgroup A, such that either*
(a) *G/A is Abelian of exponent dividing 4 or 6, or*
(b) *G/A is Abelian of odd order.*
Then G has a torsion free complement in $\mathcal{U}_1(\boldsymbol{Z}G)$.

We have already seen the complement in case (a). For case (b) a normal complement N in $\mathcal{U}_1(\boldsymbol{Z}G)$ is an extension of $\mathcal{U}(1+I_s)$ by a finitely generated free Abelian group where $s = \frac{1}{2}((G:A)-1)$.

The last theorem is not necessarily true if $(G:A)$ is even. Counterexamples were given by Roggenkamp and Scott [RoS1,RoS2,RoS3]. For instance, if $G = C_{41} \rtimes C_8$ or $C_{74} \rtimes C_8$ or $C_{241} \rtimes C_{10}$ then $G \to \mathcal{U}_1(\boldsymbol{Z}G)$ does not split (see also [Se13, p. 182]).

We have seen that if G has a normal subgroup A with both A and G/A Abelian then $\mathcal{U}(1+\Delta(G)\Delta(A))$ is torsion free. One can ask if this is still the case when G/A is not supposed to be Abelian. An easy induction argument (see [Se12, p. 182]) proves

THEOREM 3.42. *Let A be a normal Abelian subgroup of a finite nilpotent group G. Then $\mathcal{U}(1+\Delta(G)\Delta(A))$ is torsion free.*

This result is also true for infinite nilpotent groups as seen by the same argument. Moreover, if G is nilpotent class 2, we have

THEOREM 3.43 (Ritter–Sehgal). *If G is an arbitrary nilpotent group of class 2 then G has a torsion free normal complement in $\mathcal{U}_1(\boldsymbol{Z}G)$.*

PROOF. See [Se13, p. 181]. □

We state the still open questions (2000):

PROBLEM 3.44. Let G be a finite p-group. Does the natural map $G \hookrightarrow \mathcal{U}_1(\boldsymbol{Z}G)$ split?

PROBLEM 3.45. Let A be a normal Abelian subgroup of a finite group G. Is $\mathcal{U}(1+\Delta(G)\Delta(A))$ torsion free?

Added in proof: This problem has an affirmative answer as proved by Marciniak and Sehgal (J. Group Theory, to appear).

4. Isomorphism problem

A natural question is, to what extent does a group ring KG, where K is a field or a suitable ring, determine G. We notice, immediately, that if G is finite Abelian and $K=\boldsymbol{C}$, the complex number field, then $\boldsymbol{C}G$ is a direct sum of $|G|$ copies of $\boldsymbol{C}$. Thus if H is another Abelian group of the same order as G then $\boldsymbol{C}G \simeq \boldsymbol{C}H$ as rings but G need not be isomorphic to H. However, if we take for K the prime field $\boldsymbol{Q}$ then we have

THEOREM 4.1 (Perlis–Walker). *If G is a finite Abelian group then*

$$\boldsymbol{Q}G \simeq \boldsymbol{Q}H \quad \Rightarrow \quad G \simeq H.$$

A proof can be worked out from the structure theorem (3.3).

However, if we let G and H be the two non-Abelian nonisomorphic groups of order p^3 then $\boldsymbol{Q}G \simeq \boldsymbol{Q}H$. In fact, we have

THEOREM 4.2 [P3]. *There exists a set of at least $p^{2/27\cdot(n^3-23n^2)}$ nonisomorphic p-groups of order p^n that have isomorphic group algebras over all fields of characteristic not equal to p.*

PROOF. See [P21, p. 658]. □

This result was improved by Dade [Da] who constructed two metabelian groups of order p^3q^6 whose group algebras are isomorphic over all fields. These groups can be easily described and we do so below.

Let p and q be distinct primes with $q \equiv 1 \bmod p^2$. There are infinitely many such pairs. Since q is odd let Q_1, Q_2 be two non-Abelian groups of order q^3:

$$Q_1 = \langle x_1, y_1, z_1 \mid x_1^q = y_1^q = z_1^q = 1,\ (x_1, y_1) = z_1, z_1 \text{ central}\rangle, \quad \text{and}$$
$$Q_2 = \langle x_2, y_2, z_2 \mid x_2^q = z_2^q = 1,\ y_2^q = z_2,\ (x_2, y_2) = z_2,\ z_2 \text{ central}\rangle.$$

Let $\langle u_1\rangle$ be cyclic of order p^2 and $\langle u_2\rangle$ cyclic of order p. Since $p|(q-1)$ and the unit group $\mathcal{U}(\boldsymbol{Z}/q^2\boldsymbol{Z})$ has order $q(q-1)$, there exists an integer n with $n \not\equiv 1 \bmod q^2$, $n^p \equiv 1 \bmod q^2$. This implies that $n \not\equiv 1 \bmod q$. Using this integer we define an action of u_i on Q_j by

$$x_j^{u_i} = x_j, \qquad y_j^{u_i} = y_j^n, \qquad z_j^{u_i} = z_j^n.$$

This is an automorphism of Q_j because $Q_j = \langle y_j, z_j\rangle \rtimes \langle x_j\rangle$ and the nth power map commutes with the action of x_j. Furthermore, this automorphism has order p. Thus we can define the groups

$$G_1 = Q_1\langle u_1\rangle \times Q_2\langle u_2\rangle,$$
$$G_2 = Q_1\langle u_2\rangle \times Q_2\langle u_1\rangle$$

which are both metabelian of order p^3q^6.

THEOREM 4.3 [Da]. *Let p, q, G_1, G_2 be as above. Then G_1 is not isomorphic to G_2 but for all fields K, KG_1 is isomorphic to KG_2 as a K-algebras.*

PROOF. See [P21, p. 661]. □

Thus the only isomorphism questions of interest are:

(ISO) Does the ring isomorphism $\boldsymbol{Z}G \simeq \boldsymbol{Z}H$ imply the group isomorphism $G \simeq H$?

which we have already met in Section 3, and the modular isomorphism problem,

(MISO) Let $\boldsymbol{F}$ be a field of p-elements and let P be a finite p-group. Does $\boldsymbol{F}P \simeq \boldsymbol{F}Q$ imply $P \simeq Q$?

We shall now give a report on these questions.

Since $\boldsymbol{Z}G \simeq \boldsymbol{Z}H$ implies $\boldsymbol{Q}G \simeq \boldsymbol{Q}H$ it follows by Theorem 4.1 that finite Abelian groups are characterized by their integral group rings. This also follows by the theorem of Higman:

$$\boldsymbol{Z}G \simeq \boldsymbol{Z}H \quad\Rightarrow\quad \mathcal{U}(\boldsymbol{Z}G) \simeq \mathcal{U}(\boldsymbol{Z}H) \quad\Rightarrow\quad \pm G \simeq \pm H \quad\Rightarrow\quad G \simeq H.$$

The first real progress is due to Whitcomb [Wh], who proved

THEOREM 4.4. *Let G be a finite metabelian group. Then*

$$\mathbf{Z}G \simeq \mathbf{Z}H \quad \Rightarrow \quad G \simeq H.$$

PROOF. Suppose $\theta : \mathbf{Z}G \to \mathbf{Z}H$ is an isomorphism. For $g \in G$, let $\theta(g) = u \in \mathbf{Z}H$. Then u is a unit of finite order. First, it is proved that if G is metabelian then so is H. Let $H' = A$. Consider the epimorphism $\mathbf{Z}H \to \mathbf{Z}\overline{H} = \mathbf{Z}(H/H')$. Then $\bar{u}$ is a torsion unit of the Abelian group ring $\mathbf{Z}\overline{H}$. Consequently, by Higman's theorem $\bar{u} = \bar{h}$, $h \in H$. Thus

$$u = h + \delta, \quad \delta \in \Delta(H, A).$$

We conclude by Whitcomb's argument (3.38) that

$$u = h_0 \bmod \Delta(H)\Delta(A).$$

It can be checked that h_0 is unique and that $g \to h_0$ is an isomorphism. For details see [Se10, p. 102]. □

A breakthrough on the isomorphism problem was provided by Roggenkamp and Scott [RoS5, p. 102] who proved

THEOREM 4.5. *Let G be a finite nilpotent group. Then*

$$\mathbf{Z}G \simeq \mathbf{Z}H \quad \Rightarrow \quad G \simeq H.$$

Subsequently, A. Weiss (see Section 8) proved a stronger result.
An easy consequence is a result of Sehgal, Sehgal and Zassenhaus [SSZ].

THEOREM 4.6. *Suppose that G is a finite group which is an extension of an Abelian group A by a nilpotent group B. Suppose that* $(|A|, |B|) = 1$. *Then* $\mathbf{Z}G \simeq \mathbf{Z}H \Rightarrow G \simeq H$.

Roggenkamp and Scott [RoS5] announced (4.6) without the assumption of $(|A|, |B|) = 1$. A proof can be found in [RT, p. 112].
It is known from the classification of finite simple groups that, but for a few exceptions, different simple groups have different orders. It follows easily then, see [San9, Theorem 4.2], that we have

THEOREM 4.7. *Let G be a finite simple group. Then*

$$\mathbf{Z}G \simeq \mathbf{Z}H \quad \Rightarrow \quad G \simeq H.$$

Recently, Martin Hertweck [Hew] has given a counterexample to (ISO). We give a very brief account.

The counterexample

Let X be a semidirect product $Q \rtimes P$ with Q a normal Sylow 97-subgroup and P a Sylow 2-subgroup of X. The latter is a semidirect product

$$P = \big(\langle u\rangle \times \langle v\rangle \times \langle w\rangle\big) \rtimes \big(\langle a\rangle \times \langle b\rangle \times \langle c\rangle\big),$$

where u, v, w, a, b and c have order $32, 4, 8, 128, 2$ and 8, respectively. The operation of a is given by $u^a = u$, $v^a = u^{16}v$ and $w^a = u^4 w$. The operation of the elements b and c is given by $x^b = x^{-1}$ and $x^c = x^5$ for all $x \in \langle u, v, w\rangle$.

Construction of Q

Let $D = (\langle z\rangle \times \langle y\rangle) \rtimes \langle x\rangle \cong C_{97}^{(2)} \rtimes C_{97}$ with $y^x = zy$ and $z^x = z$. An automorphism $\delta \in \mathrm{Aut}(D)$ of order 64 is defined by $z\delta = z^{-19}$, $y\delta = x$ and $x\delta = y^{19}$. Let $R = D^{(2)}$ and $\rho \in \mathrm{Aut}(R)$ with $(d_1, d_2)\rho = (d_2, d_1\delta)$, an automorphism of order 128. Let M be an elementary Abelian group of order 97^4. The group Q is the direct product of the normal subgroups $R^{(4)}$ and M of X. The elements u, v, w, b, c centralize M and a operates faithfully on M. M can be thought of as the additive group of the finite field $\boldsymbol{F}_{97^4}$ with a acting as multiplication by a fixed root of unity of order 128 in the field.

The operation of P on $R^{(4)}$ is given by: u, v centralize $R^{(4)}$ and

$$\begin{aligned}
(r_1, r_2, r_3, r_4)^a &= (r_1\rho, r_2\rho, r_3\rho, r_4\rho),\\
(r_1, r_2, r_3, r_4)^w &= \big(r_4\rho^{64}, r_1, r_2, r_3\big),\\
(r_1, r_2, r_3, r_4)^b &= \big(r_1, r_4\rho^{64}, r_3\rho^{64}, r_2\rho^{64}\big),\\
(r_1, r_2, r_3, r_4)^c &= \big(r_1, r_2\rho^{64}, r_3, r_4\rho^{64}\big).
\end{aligned}$$

Write $G = Q \rtimes (\langle u\rangle \times \langle v\rangle \times \langle w\rangle) \rtimes (\langle a\rangle \times \langle b\rangle)$. Then $X = G \rtimes \langle c\rangle$. We have the

THEOREM 4.8 (Hertweck).

(a) *There is an automorphism τ of G which is not inner and a unit $t \in \mathcal{U}_1(\boldsymbol{Z}G)$ so that $g \xrightarrow{\tau} g^t$ for all $g \in G$.*

(b) *In $\boldsymbol{Z}X$, the element c inverts t.*

(c) *The subgroup $Y = \langle G, tc\rangle$ of $\mathcal{U}_1(\boldsymbol{Z}X)$ has the same order as X but is not isomorphic to X.*

(d) *The order of X is $2^{21} \cdot 97^{28}$. The group X has a normal Sylow 97-subgroup and the derived length of X is 4.*

Thus $\boldsymbol{Z}X \simeq \boldsymbol{Z}Y$ but $X \not\simeq Y$. Moreover, we see from (a) that there is an element t in the normalizer, $N_{\mathcal{U}_1(\boldsymbol{Z}G)}(G)$, of G in $\mathcal{U}_1(\boldsymbol{Z}G)$ which is not of the form $h\alpha$, $h \in G$, $\alpha \in \mathfrak{z}(\boldsymbol{Z}G)$. This disproves the normalizer conjecture. In fact, Hertweck also gave a smaller, metabelian counterexample of order $2^{25} \cdot 97^2$ to this conjecture.

THE NORMALIZER CONJECTURE 4.9. $N_{\mathcal{U}_1(\boldsymbol{Z}G)}(G) = G\mathfrak{z}(\mathcal{U}_1(\boldsymbol{Z}G))$.

However, this conjecture is known to be true for p-groups (Coleman [Col2]) and also for those groups which have normal Sylow 2-groups (Jackowski and Marciniak [JM]).

Modular isomorphism problem

Let us now turn to the modular isomorphism problem:

(MISO) Let P be a finite p-group and let $\boldsymbol{F}$ be a field of p elements. Does $\boldsymbol{F}P \simeq \boldsymbol{F}Q$ imply $P \simeq Q$?

For simplicity we have taken $\boldsymbol{F}$ to be the prime field of p elements. The problem could be stated for any finite extension of $\boldsymbol{F}$. Unfortunately, very little has been done on this problem. Here are the few results that have been obtained so far. Throughout this section $\boldsymbol{F}$ will be the finite field of p-elements.

THEOREM 4.10 (Deskins [Des]). *Let G be a finite Abelian p-group. Then $\boldsymbol{F}G \simeq \boldsymbol{F}H$ implies $G \simeq H$.*

PROOF. It follows from the cardinality of $\boldsymbol{F}G$ that $|H| = |G|$. Let $G_i = \{g \in G\colon g^{p^i} = 1\}$ and $\Gamma_i = \{\gamma \in \boldsymbol{F}G\colon \gamma^{p^i} = 0\}$. Then $\Gamma_i = \Delta(G, G_i)$ is a vector space over $\boldsymbol{F}$ of dimension $(G:1) - (G:G_i)$. Since under the isomorphism $\boldsymbol{F}G \to \boldsymbol{F}H$, Γ_i maps to its counterpart in $\boldsymbol{F}H$ it follows that $|G_i| = |H_i|$ for all i. Thus $G \simeq H$. $\square$

There are two easy consequences of this result.

COROLLARY 4.11. *If G is a finite p-group then $\boldsymbol{F}G \simeq \boldsymbol{F}H$ implies that $G/G' \simeq H/H'$.*

PROOF. Since $\Delta(G, G')$ is the smallest ideal I of $\boldsymbol{F}G$ with the property that $\boldsymbol{F}G/I$ is commutative it follows that under the isomorphism of group algebras $\boldsymbol{F}G \to \boldsymbol{F}H$, $\Delta(G, G')$ is mapped to $\Delta(H, H')$. Thus $\boldsymbol{F}(G/G') \simeq \boldsymbol{F}G/\Delta(G, G') \simeq \boldsymbol{F}H/\Delta(H, H') \simeq \boldsymbol{F}(H/H')$. It follows from the last theorem that $G/G' \simeq H/H'$. $\square$

The next result was proved by Ward [War] and also by Sehgal [Se1].

COROLLARY 4.12. *Let G be a finite p-group. Then*

$$\boldsymbol{F}G \simeq \boldsymbol{F}H \quad \Rightarrow \quad \mathfrak{z}(G) \simeq \mathfrak{z}(H).$$

PROOF. Recall that

$$[\boldsymbol{F}G, \boldsymbol{F}G] = \left\{\alpha = \sum \alpha(g)g \in \boldsymbol{F}G\colon \tilde{\alpha}(g) = \sum_{x \sim g} \alpha(x) = 0 \text{ for all } g \in G\right\}.$$

Then

$$I = [\boldsymbol{F}G, \boldsymbol{F}G] \cap \mathfrak{z}(\boldsymbol{F}(G)) = \left\{\sum a_x C_x\colon\ C_x = \sum_{y\sim x} y,\ a_x \in \boldsymbol{F},\ x \notin \mathfrak{z}(G)\right\}.$$

It is easy to see that I is an ideal in the centre, $\mathfrak{z}(\boldsymbol{F}G)$, of $\boldsymbol{F}G$. Moreover, $\mathfrak{z}(\boldsymbol{F}(G))/I \simeq \boldsymbol{F}(\mathfrak{z}(G))$. Thus we have $\boldsymbol{F}(\mathfrak{z}(G)) \simeq \boldsymbol{F}(\mathfrak{z}(H))$. The result follows from the theorem. □

The last theorem was extended to

THEOREM 4.13 (Berman and Mollov [BM], May [May]). *Let G and H be Abelian groups. Then $\boldsymbol{F}G \simeq \boldsymbol{F}H \Rightarrow G$ and H have the same Ulm invariants.*

PROOF. See [DuS] or [Se10, p. 85]. □

It follows from this theorem that countable Abelian p-groups are determined by their group algebras over $\boldsymbol{F} = \boldsymbol{F}_p$, see [Se10, p. 86].

A positive answer to (MISO) exists for groups of order $\leqslant p^4$ due to Passman [P4] and for groups of order 32 due to Makasikis [Mak] and Michler, Newman and O'Brien [MNO]. The next result was proved by Bagiński [Ba] for $p > 3$ and extended by Sandling [San11] to all p.

THEOREM 4.14. *If G is a metacyclic p-group and $\boldsymbol{F}G \simeq \boldsymbol{F}H$ then $G \simeq H$.*

Baginski and Caranti [BaC] answered (MISO) in the affirmative for p-groups of maximal class of order $\leqslant p^{p+1}$ for odd primes p, having a maximal Abelian subgroup. Carlson [Car] handled all 2-groups of maximal class.

Sandling [San10] proved

THEOREM 4.15. *Let G be a finite p-group of nilpotent class two. Suppose that G' is of exponent p. Then, $\boldsymbol{F}G \simeq \boldsymbol{F}H \Rightarrow G \simeq H$.*

To present the last result of this section we need to recall the definition of the Brauer–Jennings–Zassenhaus M-series of a group G:

$$M_{n,p}(G) = \prod_{ip^j \geqslant n} (G_i)^{p^j}.$$

Here, $\{G_i\}$ is the lower central series of G. We shall drop the subscript p and simply write $M_n(G)$ for $M_{n,p}(G)$ as p is fixed for us. The first part was proved by Dieckman [Di] and Hill [Hil] for finite p-groups, whereas Passi and Sehgal [PS1] proved the next theorem for all groups.

THEOREM 4.16. *Suppose $\boldsymbol{F}G \simeq \boldsymbol{F}H$. Then*

(1) *$M_n(G)/M_{n+1}(G) \simeq M_n(H)/M_{n+1}(H)$ for all n.*
(2) *$M_n(G)/M_{n+2}(G) \simeq M_n(H)/M_{n+2}(H)$ for all n.*

PROOF. See [Se10, p. 113]. □

5. Large groups of units of integral group rings of finite nilpotent groups

Let G be a finite group and $\mathcal{U}(\boldsymbol{Z}G)$ the group of units of the integral group ring $\boldsymbol{Z}G$. It is an interesting and difficult question to describe $\mathcal{U}(\boldsymbol{Z}G)$ by giving generators and relations. But this is to be expected as it is not possible to even find fundamental units in rings of algebraic integers. So as in number theory one should be satisfied to find generators up to finite index of $\mathcal{U}(\boldsymbol{Z}G)$. We report on the recent progress on the problem for the case of nilpotent groups G. First, we record the various recipes for construction of units and then we state the results.

5.1. *Examples*

5.1.1. *Trivial units.* These are the units $\pm g,\ g \in G$, which we have already seen.

5.1.2. *Bass cyclic units.* To introduce our second example we recall the notion of "cyclotomic units". These are elements of $\boldsymbol{Z}[\zeta]$ where ζ is a primitive nth root of unity and are of the form

$$\alpha = \left(1-\zeta^i\right)/(1-\zeta) = 1+\zeta+\cdots+\zeta^{i-1}, \quad \text{where } (i,n)=1.$$

The inverse of α is given by

$$(1-\zeta)/\left(1-\zeta^i\right) = \left(1-\zeta^{ij}\right)/\left(1-\zeta^i\right) = 1+\zeta^i+\zeta^{2i}+\cdots+\zeta^{(j-1)i} \in \boldsymbol{Z}[\zeta],$$

where $ij \equiv 1 \pmod n$.

Now let a be an element of order n in a group G. Then the element of $\boldsymbol{Z}\langle a\rangle$ analogous to a cyclotomic unit, namely, $\beta = 1+a+\cdots+a^{i-1},\ (i,n)=1$, is not invertible in $\boldsymbol{Z}\langle a\rangle$ as the augmentation $\varepsilon(\beta)=i>1$. Remember that the augmentation map $\varepsilon(\sum c_i g_i)=\sum c_i \in \boldsymbol{Z}$ is a ring homomorphism and maps units to 1 or -1. We need to be a little more clever.

The rational group algebra $\boldsymbol{Q}\langle a\rangle$ is a direct sum of fields:

$$\boldsymbol{Q}\langle a\rangle \simeq \sum_{d\mid n}^{\oplus} \boldsymbol{Q}\left(\zeta^d\right), \quad \text{where } a \to \sum^{\oplus} \zeta^d.$$

Clearly, $\boldsymbol{Z}\langle a\rangle$ injects into

$$\sum_{d\mid n}^{\oplus} \boldsymbol{Z}\left[\zeta^d\right] = \boldsymbol{M},$$

the unique maximal order. Since $\boldsymbol{Z}\langle a\rangle \subset \boldsymbol{M}$ are orders, an element of $\boldsymbol{Z}\langle a\rangle$ is a unit if and only if it has an inverse in $\boldsymbol{M}$ [Se13, p. 19]. Thus to check if an element of $\boldsymbol{Z}\langle a\rangle$ is a unit it suffices to produce its inverse in $\boldsymbol{M}$. Let us consider the element β above. Its image under any projection $\boldsymbol{Z}\langle a\rangle \to \boldsymbol{Z}[\zeta^d]$ is a cyclotomic unit except when $d=n$. Thus we need to modify β to get around the augmentation obstruction. For technical reasons, let k be a fixed multiple of $|G|$ and $\varphi(|G|)$ with φ denoting Euler's function. Since $(i,n)=1$, $i^{\varphi(n)}\equiv 1 \pmod n$ and so $i^k \equiv 1 \pmod n$. Then the element

$$u = \left(1+a+\cdots+a^{i-1}\right)^k + \left(\left(1-i^k\right)/n\right)\hat{a}, \quad \hat{a}=\sum_1^n a^i,$$

belongs to $\boldsymbol{Z}\langle a\rangle$. Moreover, $\varepsilon(u)=1$, u is invertible in $\boldsymbol{M}$ and hence $u\in\mathcal{U}(\boldsymbol{Z}\langle a\rangle)$. These units are called *Bass cyclic units* of $\boldsymbol{Z}G$. We denote by $\mathcal{B}_1=\mathcal{B}_1(G)$ the group generated by them.

5.1.3. *Hoechsmann units.* Again let $C=\langle a\rangle$ be of order n. We modified the element $\beta = 1+a+\cdots+a^{i-1}$, $(i,n)=1$, to obtain the Bass cyclic unit. One could take a quotient of two elements of equal augmentation. In this spirit let

$$v = 1+a^j+\cdots+a^{j(i-1)}/1+a+\cdots+a^{(i-1)}, \quad (i,n)=1,\ (j,n)=1,$$

be an element of $\boldsymbol{Q}\langle a\rangle$. It is easily seen that $v(\zeta^d)$ is a unit of $\boldsymbol{Z}[\zeta^d]$ for all $d|n$. Thus v is a unit of $\boldsymbol{M}$. In fact,

$$v = \left(1+a^j+\cdots+a^{j(i-1)}\right)\left(1+a^i+\cdots+a^{(\ell-1)i}\right)+\left((1-i\ell)/n\right)\hat{a},$$

where $i\ell\equiv 1 \pmod n$ is an element of $\boldsymbol{Z}\langle a\rangle$ [Se13, p. 34]. Hence v is a unit of $\boldsymbol{Z}\langle a\rangle$. All these units are called *Hoechsmann units* and we denote by $\mathcal{H}(C)$ the group generated by them.

It is not hard to see that $\mathcal{H}(C)$ contains $\mathcal{B}_1(C)$ but is much larger. Modulo trivial units, the index of one in the other grows at a more than exponential rate with n. If n equals 101, for example, this index is 2×10^{96}, while $\mathcal{H}(C)$ equals all of $\mathcal{U}(\boldsymbol{Z}C)$ (see [Ho5]).

5.1.4. *Alternating units.* If n is odd it is possible to fix the augmentation difficulties in a straightforward manner. Let $(i,2n)=1$ then

$$u = 1-a+a^2-\cdots+a^{i-1}$$

clearly has augmentation equal to one. Moreover, $u(\zeta)=1-\zeta+\zeta^2+\cdots+\zeta^{i-1}=1-\zeta+(-\zeta)^2+\cdots+(-\zeta)^{i+1}$ is a cyclotomic unit. Thus u is a unit of $\boldsymbol{Z}C$. We call these type of units alternating units and denote the group generated by them by $\mathcal{A}(C)$.

5.1.5. *The group $\mathcal{C}_1$.* Before introducing these units, which have only one non-identity component in the decomposition

$$\boldsymbol{Q}C = \boldsymbol{Q}\langle a\rangle = \sum_{d|n}^{\oplus} \boldsymbol{Q}\left(\zeta^d\right), \qquad (*)$$

we recall the formulae for the primitive idempotents of $\boldsymbol{Q}C$. Let $\langle d\rangle$ be the Sylow p-subgroup of C for a prime dividing n. Then it can be checked directly that the primitive idempotents of $\boldsymbol{Q}\langle d\rangle$ are given by

$$e_{p0}=\hat{d},\ e_{p1}=d^p-\hat{d},\ \dots,\ e_{pn_p}=\hat{d}^{\ell}-1$$

with $n=\prod_p p^{n_p}$ and $\ell=p^{n_p-1}$. It follows that the primitive idempotents of $\boldsymbol{Q}C$ are given by $e=\prod_p e_{pi_p},\ 0\leqslant i_p\leqslant n_p$.

Let $\alpha\in\boldsymbol{Z}\langle a\rangle$ and $e=e^2\in\boldsymbol{Q}\langle a\rangle$ be a primitive idempotent as described above. Then in the decomposition $(*)$ we have

$$\begin{aligned}&\alpha e=(0,\dots,0,\bar{\alpha},0,\dots,0),\\&1-e=(1,\dots,1,0,1,\dots,1),\\&\alpha e+(1-e)=1+(\alpha-1)e=(1,\dots,1,\bar{\alpha},1,\dots,1).\end{aligned}$$

Let us specialize α to elements of the form $1+a+\cdots+a^{i-1},\ (i,n)=1$. Then $1+(\alpha-1)e$ is a unit of the maximal order $\boldsymbol{M}$. It follows from [Rei, p. 379] that $nM\subseteq\boldsymbol{Z}\langle a\rangle$ and thus there is a fixed number k depending only on n so that $(1+(\alpha-1)e)^k\in\mathcal{U}\boldsymbol{Z}\langle a\rangle$ [Se13, p. 19]. It is easy to prove that $\mathcal{C}_1=\langle(1+(\alpha-1)e)^k\colon\ \alpha=1+a+\cdots+a^{i-1},\ (i,n)=1\rangle$ is a subgroup of finite index in $\mathcal{U}\boldsymbol{Z}\langle a\rangle$.

We have given some recipes for writing down explicitly units in Abelian group rings. The next example is a noncommutative one.

5.1.6. *Bicyclic units.* For an element $a\in G$ of order n recall that $\hat{a}=\sum_{i=1}^n a^i$. Then, obviously, $(a-1)\hat{a}=\hat{a}(a-1)=0$. For a pair of elements $a,b\in G$ consider the element $\eta=(a-1)b\hat{a}$. It is nilpotent, in fact, $\eta^2=0$. Thus $(1+\eta)(1-\eta)=1$. We have found units

$$u_{a,b}=1+(a-1)b\hat{a},\quad a,b\in G.$$

Similarly, $u'_{a,b}=1+\hat{a}\ b(a-1)$ is also invertible. We call $\mathcal{B}_2=\langle u_{a,b}\colon\ a,b\in G\rangle$ the group of bicyclic units. We set $\mathcal{B}'_2=\langle u'_{a,b}\colon\ a,b\in G\rangle$. Clearly, $u_{a,b}=1$ if and only if b normalizes $\langle a\rangle$. Thus $\mathcal{B}_2=1$ if and only if all subgroups of G are normal in G. A non-Abelian group with this property is called Hamiltonian. It turns out that a Hamiltonian group can be written as $K_8\times E\times O$ where $E^2=1$, O is an odd Abelian group and K_8 is the quaternion group of order 8. In this situation there are other useful units.

5.1.7. *The group $\mathcal{B}_3$.* These units can be constructed whenever G has a homomorphic image $K_8\times C_p$, with p an odd prime and $\boldsymbol{Q}(K_8\times C_p)$ contains nilpotent elements. For simplicity, we shall illustrate only the case $G=K_8\times C_p$. We have $K_8=\langle a,b\colon\ a^2=b^2=z,\ z^2=1,\ a^b=a^{-1}\rangle$. Let

$$\boldsymbol{H}(\boldsymbol{Q})=\boldsymbol{Q}\dotplus\boldsymbol{Q}i\dotplus\boldsymbol{Q}j\dotplus\boldsymbol{Q}k$$

be the rational Hamiltonian (skew) field. Then

$$\boldsymbol{Q}K_8 = 4\boldsymbol{Q} \oplus \boldsymbol{H}(\boldsymbol{Q}).$$

Moreover,

$$\begin{aligned}\boldsymbol{Q}(K_8 \times C_p) &= \big(4\boldsymbol{Q} \oplus \boldsymbol{H}(\boldsymbol{Q})\big) \otimes \big(\boldsymbol{Q} \oplus \boldsymbol{Q}(\zeta_p)\big)\\ &= 4\boldsymbol{Q} \oplus 4\boldsymbol{Q}(\zeta_p) \oplus \boldsymbol{H}(\boldsymbol{Q}) \oplus \boldsymbol{H}\big(\boldsymbol{Q}(\zeta_p)\big)\end{aligned}$$

where ζ_p is a primitive pth root of unity. We have assumed that $\boldsymbol{H}(\boldsymbol{Q}(\zeta_p))$ splits. Therefore, it is possible to find explicitly $x', y' \in \boldsymbol{Z}[\zeta_p]$ so that ${x'}^2 + {y'}^2 = -1$. Let x, y be chosen in $\boldsymbol{Z}(C_p)$ so that $x(\zeta_p) = x'$, $y(\zeta_p) = y'$. Let e be the idempotent corresponding to $\boldsymbol{H}(\boldsymbol{Q}(\zeta_p))$. Then e is given by $e = \frac{1-z}{2} \cdot (1 - \hat{c})$ where $\langle c \rangle = C_p$. Furthermore, $(x^2 + y^2 + 1)e = 0$. Define $\eta = |G|(ya + b + xab)e$, $\eta' = |G|(ya - b + xab)e$. Then

$$\eta^2 = |G|^2\big(xy(1+z)b + y(1+z)ab + x(1+z)a\big)e = 0.$$

Similarly, $(\eta')^2 = 0$. Therefore, $(\eta\ \boldsymbol{Z}G\eta)^2 = 0 = (\eta'\boldsymbol{Z}G\eta')^2$. We have units $1 + \eta x\eta$ and $1 + \eta' x\eta'$, $x \in \boldsymbol{Z}G$. We define

$$\mathcal{B}_3 = \langle 1 + \eta g\eta,\ 1 + \eta' g\eta',\ g \in G\rangle.$$

These units can be extended to the general situation [GS4].

5.2. *Generators of large subgroups*

We shall say that a subgroup is large if it is of finite index. In many cases it is possible to find very nice explicit generators of large subgroups of $\mathcal{U}(\boldsymbol{Z}G)$. We shall present these results below.

Let C be cyclic of order n. Then Bass [Bas1] has proved

THEOREM 5.1. *The subgroup $\mathcal{B}_1$ of the Bass cyclic units in $\boldsymbol{Z}C$ is free and of finite index in $\mathcal{U}(\boldsymbol{Z}C)$.*

PROOF. See [Bas1] or [Se13, p. 45]. □

We shall see now that this solves the problem for the Abelian case. Let $\mathcal{M}(A) = \prod_{C \subseteq A} \mathcal{B}_1(C)$ denote the product of the Bass cyclic units where C runs over the cyclic subgroups of the finite Abelian group A. We have

THEOREM 5.2 (Bass–Milnor). *$\mathcal{M}(A)$ is of finite index in $\mathcal{U}(\boldsymbol{Z}A)$ and the product is direct.*

PROOF. See [Se13, p. 63]. □

The above result says, in particular, that the units of cyclic subgroups generate a large subgroup of $\mathcal{U}(\boldsymbol{Z}A)$. A lot more can be said about this index. For instance, if A is an elementary Abelian p-group and p is a regular prime, then the index is equal to 1 (see [Se13] or [HoS1]). Klaus Hoechsmann has extensively studied arithmetic questions about $\mathcal{U}(\boldsymbol{Z}A)$ and obtained wonderful results, especially (but not exclusively) for p-groups; these include the elimination of the word "elementary" in the preceding sentence, as well as relations between indices in the unit group and orders in the ideal class group of $\boldsymbol{Z}A$. The general aim of this work is the explicit construction of as much of $\mathcal{U}(\boldsymbol{Z}A)$ as possible. Much of it depends on studying logarithms of units in the p-adic group ring $\boldsymbol{Z}_pA$, and requires a close examination of the group of "constructible" units, generated by the $\mathcal{H}(C)$ as C runs over the cyclic subgroups of A. For a survey see [Ho6].

In another direction we have

THEOREM 5.3 (Bass–Milnor). *Let G be a finite group. The image of $\langle\mathcal{U}(\boldsymbol{Z}C)\rangle_{C\subseteq G}$ under the natural map $j:\mathcal{U}(\boldsymbol{Z}C)\to K_1(\boldsymbol{Z}G)$ as C runs over the cyclic subgroups of G is of finite index.*

PROOF. See [Bas1]. □

A straightforward extension of Theorem 5.3, namely that for a finite G, $(\mathcal{U}(\boldsymbol{Z}G):\langle\mathcal{U}(\boldsymbol{Z}C)\rangle_{C\subseteq A}) < \infty$ is not true as seen by the following example.

EXAMPLE 5.4. Let $G=\langle a,b\colon a^3=1=b^2,\ a^b=a^{-1}\rangle$ be the symmetric group on three letters. Then every cyclic subgroup of G has order $1,2$ or 3. Thus $\mathcal{U}(\boldsymbol{Z}C)=\pm C$ for all $C\subseteq G$. Thus $\langle\mathcal{U}(\boldsymbol{Z}C)\rangle_{C\subseteq G}=\pm G$ whereas $\mathcal{U}(\boldsymbol{Z}G)$ contains elements of infinite order, for example, $u_{b,a}=1+(b-1)a(b+1)$.

However, if one considers the normal subgroup of $\mathcal{U}(\boldsymbol{Z}G)$ generated by all $\mathcal{U}(\boldsymbol{Z}C)$ then we do get a subgroup of finite index as was proved by Kleinert [Kl4].

THEOREM 5.5. *But for a few exceptions, we have that the normal closure of $\langle\mathcal{U}(\boldsymbol{Z}C)\rangle_{C\subseteq G}$ in $\mathcal{U}(\boldsymbol{Z}G)$ is of finite index in $\mathcal{U}(\boldsymbol{Z}G)$.*

PROOF. See [Se13, p. 116]. □

The exceptions referred to above arise from the failure of the congruence subgroup theorem. We shall discuss this soon. Let us establish some notation first. Let

$$\boldsymbol{Q}G=S_1\oplus\cdots\oplus S_t$$

be the decomposition of $\boldsymbol{Q}G$ into its simple Wedderburn components and let π_i be the projections. We identify S_i with $(D_i)_{n_i\times n_i}$, the ring of $n_i\times n_i$ matrices over the division ring D_i. Suppose $n_i\geqslant 2$. Let $\mathcal{O}_i$ be a maximal order in D_i and $\mathfrak{p}$ an ideal in $\mathcal{O}_i$. Let $E(\mathfrak{p})$ be the $\mathfrak{p}$-elementary matrices and SL_i the matrices of reduced norm one contained in $(\mathcal{O}_i)_{n_i\times n_i}$. We shall assume that in our case, G satisfies the following conditions:

(i) $(SL_i : E(\mathfrak{p})) < \infty$ for all $\mathfrak{p} \lhd \mathcal{O}_i$.

(ii) Any subgroup H of SL_i normalized by a subgroup of finite index in SL_i contains $E(\mathfrak{p})$ for some $0 \neq \mathfrak{p} \lhd \mathcal{O}_i$.

We say that the component $(D_i)_{n_i \times n_i}$ satisfies (C.S.T.). If all components have this property, we shall say that G satisfies (C.S.T.). In fact, it is a result of Bass–Milnor–Serre–Vaserstein that (i) and (ii) always hold if $n_i \geqslant 3$ and for $n_i = 2$ they hold provided that $D_i \neq$ the rational field or an imaginary quadratic field or a definite quaternion algebra. We shall further assume that if $n_i = 1$, and D_i is not commutative then D_i is a totally definite quaternion algebra, namely, the group of units of $\mathcal{O}_i$ of reduced norm 1 is finite. If G is nilpotent of odd order then all the D_i's are commutative and also $n_i \neq 2$. Our first main result for noncommutative groups is

THEOREM 5.6 (Ritter–Sehgal). *If G is nilpotent of odd order then $\langle \mathcal{B}_1, \mathcal{B}_2 \rangle$ is of finite index in $\mathcal{U}(\boldsymbol{Z}G)$.*

PROOF. See [RS8] or [Se13]. □

REMARK 5.7. In fact, it is proved in [RS8] that if G is a nilpotent group, satisfying (C.S.T.), for whose Sylow 2-subgroups the D_i's are commutative then $\langle \mathcal{B}_1, \mathcal{B}_2 \rangle$ is a large subgroup of $\mathcal{U}(\boldsymbol{Z}G)$. If G has no subhomomorphic image K_8, the quaternion group of order 8, then the second condition above holds [Se13, p. 106].

Now, some words about the strategy of proof of a result as above. By using Theorem 5.3 it follows that in order to prove that $\langle \mathcal{B}_1, H \rangle$ is of finite index in $\mathcal{U}(\boldsymbol{Z}G)$ it suffices to show that H contains a subgroup W_i of finite index in SL_i for all i with $n_i \geqslant 2$ [Se13, p. 123]. To produce W_i we prove $\pi_i(H)$ contains a subgroup of finite index in SL_i. Under the assumption of (C.S.T.) it suffices to produce an $E(\mathfrak{p})$ in $\pi_i(H)$.

5.3. *A counterexample*

Theorem 5.6 is not true for all finite nilpotent groups. Counterexamples were provided by Ritter and Sehgal [Se13, p. 617]. For instance, if

$$G = \langle a, b\colon a^4 = 1 = b^4,\ a^b = a^{-1} \rangle, \quad \text{a group of order 32,}$$

then we have $(\mathcal{U}(\boldsymbol{Z}G) : \langle \mathcal{B}_1, \mathcal{B}_2 \rangle) = \infty$.

We return to the discussion of the case when G has a homomorphic image K_8. Remember that $\boldsymbol{Q}K_8 = 4\boldsymbol{Q} \oplus \boldsymbol{H}(\boldsymbol{Q})$ has no nilpotent elements and moreover, $\mathcal{U}(\boldsymbol{Z}K_8) = \pm K_8$. Further, if ζ is a primitive pth root of unity, $\boldsymbol{H}(\boldsymbol{Q}(\zeta))$ is not always a division ring. The result here is the

THEOREM 5.8. *$\boldsymbol{H}(\boldsymbol{Q}(\zeta))$ is not a division ring $\Leftrightarrow$ $o(2)$ mod p is even $\Leftrightarrow$ $x^2 + y^2 = -1$ has a solution in $\boldsymbol{Q}(\zeta)$.*

PROOF. See [Se13, p. 173]. □

REMARK 5.9. If $p \equiv 3 \pmod 8$ then certainly $o(2)$ mod p is even as $2^{(p-1)/2} \equiv (2/p) \bmod p \equiv (-1)^{(p^2-1)/8} \equiv -1 \pmod p$. Here is a list of a few orders of 2 mod p.

p	3	5	7	11	13	17	19	23
$o(2)$	2	4	3	10	12	8	18	11

Recall that $\boldsymbol{H}(\boldsymbol{Q}(\sqrt{-1})) \simeq (\boldsymbol{Q}(i))_{2\times 2}$ by the map

$$i \to \begin{bmatrix} \sqrt{-1} & 0 \\ 0 & -\sqrt{-1} \end{bmatrix}, \qquad j \to \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

This can be extended to $\boldsymbol{H}(\boldsymbol{Q}(\zeta)) \cong (\boldsymbol{Q}(\zeta))_{2\times 2}$ if $\boldsymbol{Q}(\zeta)$ has elements x, y satisfying $x^2 + y^2 = -1$ by the map

$$i \to \begin{bmatrix} x & y \\ y & -x \end{bmatrix}, \qquad j \to \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Jespers and Leal extended Theorem 5.6 to prove

THEOREM 5.10. (a) *If G is nilpotent satisfying* (C.S.T.) *and has no homomorphic image $K_8 \times C_p$ where p is an odd prime then $\langle \mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_2' \rangle$ is a large subgroup of $\mathcal{U}(\boldsymbol{Z}G)$.*

(b) *If G is nilpotent satisfying* (C.S.T.) *and all odd primes dividing $|G|$ are* $\equiv 3 \pmod 8$ *then there is a finite set $\mathcal{B}_4$* (*explicitly defined*) *so that $\langle \mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_2', \mathcal{B}_4 \rangle$ is a large subgroup of $\mathcal{U}(\boldsymbol{Z}G)$.*

PROOF. See [JL2] and [JL4]. □

This was further extended by Giambruno and Sehgal [GS4] who proved the following

THEOREM 5.11. *Let G be a finite nilpotent group such that for each odd prime p dividing $|G|$, the order of* 2 $(\bmod\ p)$ *is even. If G satisfies* (C.S.T.) *then $\langle \mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_2', \mathcal{B}_3 \rangle$ is a large subgroup of $\mathcal{U}(\boldsymbol{Z}G)$.*

In order to complete the study for nilpotent groups one needs to know units in cyclotomic quaternions. We suggest the

PROBLEM 5.12. Let ζ be a root of unity so that the Hamiltonian quaternions $\boldsymbol{H}(\boldsymbol{Q}(\zeta))$ do not split. Find explicit generators of a large subgroup of $\mathcal{U}(R)$ where

$$R = \boldsymbol{Z}[\zeta] \dotplus \boldsymbol{Z}[\zeta]i \dotplus \boldsymbol{Z}[\zeta]j \dotplus \boldsymbol{Z}[\zeta]k.$$

REMARK 5.13. In the strategy for exhibiting generators of large subgroups it is possible to allow one simple component of $\boldsymbol{Q}G$ which doesn't satisfy (C.S.T.). This is used to prove that if G is dihedral then $\langle\mathcal{B}_1, \mathcal{B}_2\rangle$ is a large subgroup [Se13, p. 125]. The same result is also true for $G = S_n$ [Se13, p. 146]. It is possible to find generators for $\mathcal{U}(RG)$ where $R = \boldsymbol{Z}[\zeta]$ for a suitable root of unity ζ. This was done by Ritter and Sehgal, and Jespers and Leal (see [Se12, p. 154]). We do not know (Fall 2000) how to descend to $\boldsymbol{Z}G$. See also [NS1] and [NS2]. For real progress, henceforth, new ideas are needed.

We close this section with an important result of Jespers and Leal [JL2] and a problem.

THEOREM 5.14. *Suppose that G is a finite group satisfying* (C.S.T.) *and having no non-Abelian homomorphic image which is fixed point free. Then* $(\mathcal{U}(\boldsymbol{Z}G) : \langle\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_2'\rangle)$ *is finite.*

In this instance Jespers and Leal [JL6] (see [Jes3]) have also given an estimate of the index.

PROBLEM 5.15. Is $\mathcal{B}_2$ torsion free?

Added in proof: This problem has now been answered negatively by Olivieri and del Rio, Proc. Amer. Math. Soc. (to appear).

Bhandari and Ritter have used the approach of this section to produce generators of large subgroups of units of orders in finite-dimensional algebras over local or global algebraic number fields (see [BR]).

6. Central units

We know (3.11) that the torsion central units of $\boldsymbol{Z}G$ are trivial for arbitrary groups G. Let G be finite. Then the rational group algebra $\boldsymbol{Q}G$ is a direct sum of matrix rings over division rings:

$$\boldsymbol{Q}G = \sum^{\oplus} M_n(D).$$

The centre of $\boldsymbol{Q}G$ is a direct sum of the character fields

$$\mathfrak{z}(\boldsymbol{Q}G) = \sum_{\chi}^{\oplus} \boldsymbol{Q}_{\chi}.$$

Then the centre of the integral group ring, $\mathfrak{z}(\boldsymbol{Z}G)$ is an order in $\mathfrak{z}(\boldsymbol{Q}G)$. Also, $\mathfrak{z}(\boldsymbol{Z}G)$ is a suborder of the maximal order

$$\boldsymbol{M} = \sum^{\oplus} \mathfrak{o}_{\chi},$$

where $\mathfrak{o}_{\chi}$ is the ring of algebraic integers of $\boldsymbol{Q}_{\chi}$. It follows that $\mathcal{U}\mathfrak{z}(\boldsymbol{Z}G)$ is of finite index in $\boldsymbol{M}^{\times}$. Accordingly, all central units of $\boldsymbol{Z}G$ are trivial if and only if $\boldsymbol{Q}_{\chi}$ is rational or imaginary quadratic for all χ. This is the case, for example, if $G = S_m$, the symmetric group. Finite groups with trivial central units have been classified by Ritter and Sehgal.

THEOREM 6.1 [RS7]. *Let G be a finite group. All central units of $\mathbf{Z}G$ are trivial if and only if for every $x \in G$ and every natural number j, relatively prime to $|G|$, $x^j \sim x$ or $x^j \sim x^{-1}$.*

As we see from this result, generally speaking, there exist central units of infinite order. They are hard to find. Aleev [Al] constructed all central units of $\mathbf{Z}A_5$ and $\mathbf{Z}A_6$. Also, Li and Parmenter [LiPa1] independently constructed all central units of $\mathbf{Z}A_5$.

Ritter and Sehgal [RS7] gave a recipe for generating a subgroup of finite index in $\mathfrak{z}\mathcal{U}(\mathbf{Z}G)$ for all finite groups G. Their formula involves the absolutely irreducible characters of G. For nilpotent groups (and some additional groups) Jespers, Parmenter and Sehgal [JPS] gave a construction for central units as follows.

Let G be a finite group which is nilpotent of class n. Let $\mathfrak{z}_i$ be the ith centre of G. Thus $\mathfrak{z}_n = G$. For any $x \in G$ let $b \in \mathbf{Z}\langle x\rangle$ be a Bass cyclic unit. We define $b_{(1)} = b$ and by induction for $2 \leqslant i \leqslant n$, $b_{(i)} = \prod_{g \in \mathfrak{z}_i} b_{(i-1)}^g$, a product of conjugates of $b_{(i-1)}$. By induction, one proves that $b_{(i)}$ is central in $\mathbf{Z}\langle \mathfrak{z}_i, x\rangle$ and that the conjugates commute making the product independent of the order. Then $b_{(n)}$ is central in $\mathbf{Z}G$. We have

THEOREM 6.2. *Let G be a finite nilpotent group of class n. Then*

$$\langle b_{(n)}\colon \ b \text{ a Bass cyclic unit of } \mathbf{Z}G\rangle$$

is a subgroup of finite index in $\mathfrak{z}(\mathcal{U}(\mathbf{Z}G))$.

For infinite groups, similar to the Abelian groups (3.27), we have a decomposition of central units.

THEOREM 6.3 (Polcino Milies and Sehgal [PSe3]). *Let G be any group and $\phi = \phi(G)$ the FC-subgroup of G. Let $T = T(\phi)$ be the torsion subgroup of ϕ. Then every central unit μ of $\mathbf{Z}G$ can be written in the form $\mu = wg$, with $w \in \mathbf{Z}T$ and $g \in \phi$. Moreover, $wg = gw$.*

This result was proved for finitely generated nilpotent groups by Jespers, Parmenter and Sehgal [JPS]. For these groups they also gave a finite set of generators of a large subgroup of the centre of the unit group. Polcino Milies and Sehgal [PSe3] extended this to those groups G whose FC-subgroup is finitely generated. The question of trivial central units remains open for infinite groups.

PROBLEM 6.4. Classify groups whose integral group rings have trivial central units.

We can ask about the higher terms in the upper central series of $\mathcal{U} = \mathcal{U}(\mathbf{Z}G)$. The next result was obtained by Arora, Hales and Passi [AHP] for finite groups and for torsion groups by Li [Li].

THEOREM 6.5. *If G is a torsion group then $\mathfrak{z}_2 = \mathfrak{z}_3$ where $\mathfrak{z}_i$ denotes the ith centre of $\mathcal{U}(\boldsymbol{Z}G)$.*

Moreover for the second centre we have

THEOREM 6.6 (Arora and Passi [AP]). *Let G be a finite group. Then $\mathfrak{z}_2\mathcal{U}(\boldsymbol{Z}G) = T \cdot \mathfrak{z}_1(\mathcal{U}(\boldsymbol{Z}G))$ where $T = G \cap \mathfrak{z}_2(\mathcal{U}\boldsymbol{Z}G)$ is the torsion subgroup of $\mathfrak{z}_2(\mathcal{U}(\boldsymbol{Z}G))$.*

The above result was extended to torsion groups by Li and Parmenter.

THEOREM 6.7 [LiPa2]. *Let G be a torsion group. Then $\mathfrak{z}_2 = T \cdot \mathfrak{z}_1$. Moreover, $\mathfrak{z}_2(\mathcal{U}(\boldsymbol{Z}G)) \neq \mathfrak{z}_1(\mathcal{U}(\boldsymbol{Z}G))$ if and only if (G has an Abelian normal subgroup A of index 2 which has an element a of order 4 and $g^2 = a^2$, $g^{-1}hg = h^{-1}$ for all $g \in G \backslash A$, $h \in A$). In that case, either G is a Hamiltonian 2-group and $\mathfrak{z}_2(\mathcal{U}(\boldsymbol{Z}G)) = \mathcal{U}(\boldsymbol{Z}G) = \pm G$ or $\mathfrak{z}_2(\mathcal{U}(\boldsymbol{Z}G)) = \langle a \rangle \mathfrak{z}_1(\mathcal{U}(\boldsymbol{Z}G))$ where a is any element of the type defined above.*

7. Explicit free subgroups of units

In this section we ask if there exist (noncyclic) free groups of units in group algebras KG or the integral group ring $\boldsymbol{Z}G$. We shall understand 'noncyclic free' when we say 'free' in this section. A moment's reflection reveals that if G is finite noncommutative then this is so, almost always, as seen below. We have

$$\boldsymbol{Q}G = \sum^{\oplus} (D_i)_{n_i \times n_i}, \quad \text{a direct sum of matrix rings over division rings,}$$

where either at least one $n_i > 1$ or at least one D_i is noncommutative. It is a consequence of the famous Tits' alternative that a division ring which is finite-dimensional over its centre contains a (noncommutative) multiplicative free group (see Gonçalves [Go1]). Also, the matrices $\left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)$ generate a free group. It follows that the rational group algebras of finite, noncommutative, groups contain free groups of units. Let us see what happens if we work in the integral group ring. We do know (3.5) that $\mathcal{U}(\boldsymbol{Z}K_8) = \pm K_8$ which, of course, being finite contains no free groups. But if we do have a matrix ring with $n_i > 2$ in the decomposition in $\boldsymbol{Q}G$ above then the elements $(0, 0, \ldots, e_{12}, \ldots, 0)$ and $(0, 0, \ldots, e_{21}, \ldots, 0)$ belong to $\boldsymbol{Q}G$. Here, e_{ij} denote matrix units. By removing denominators we find $m \geqslant 2$ so that

$$\alpha = (0, 0, \ldots, 0, me_{12}, \ldots, 0) \quad \text{and} \quad (0, \ldots, 0, me_{21}, \ldots, 0) = \beta$$

are elements of $\boldsymbol{Z}G$. Then $\langle 1 + \alpha, 1 + \beta \rangle$ is a free group contained in $\boldsymbol{Z}G$ as follows from

THEOREM 7.1 (Sanov). *The matrices $\left(\begin{smallmatrix} 1 & \lambda \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & 0 \\ \lambda & 1 \end{smallmatrix}\right)$ of $GL_2(\boldsymbol{C})$ with $|\lambda| \geqslant 2$ generate a free group.*

What we have seen is

THEOREM 7.2. *If G is finite and $\boldsymbol{Z}G$ contains nilpotent elements then $\mathcal{U}(\boldsymbol{Z}G)$ contains a free subgroup.*

If there are no nilpotent elements in $\boldsymbol{Z}G$ then all subgroups of G are normal. Furthermore, there is the following

THEOREM 7.3 (Hartley and Pickel [HP]). *Let G be a finite non-Abelian group. Then $\mathcal{U}(\boldsymbol{Z}G)$ contains a free group if and only if G is not a Hamiltonian 2-group.*

REMARK. Finite groups G whose unit groups $\mathcal{U}(\boldsymbol{Z}G)$ contain a free subgroup or a direct product of free groups or a free product of Abelian groups as subgroup of finite index have been classified by Jespers [Jes2], Jespers, Leal and Del Rio [JLD], Leal and Del Rio [LD] and Jespers and Leal [JL5]. For a detailed account see the very nice article of Jespers [Jes3].

As we have seen there are plenty of free groups of units in $\boldsymbol{Z}G$ and also in $\boldsymbol{Q}G$. How do we find these units? It is not practical to lift the transvection matrices back to $\boldsymbol{Z}G$. Luckily, we find the desired units the first place we look. The units u and u^* generate a free group when $u = 1 + (x-1)y\hat{x}$ is a nontrivial bicyclic unit. Of course, $u \neq 1$ if and only if y does not normalize $\langle x\rangle$. Just remember $(x-1)y\hat{x}$ is a square-zero element. The idea of the proof is quite simple as seen below.

Let us pick two nonzero nilpotent elements $a, b \in \boldsymbol{Z}G$ satisfying $a^2 = b^2 = 0$. Consider the left multiplication by ab in the space $\boldsymbol{C}G$. Obviously, it is a linear transformation.

Suppose this transformation has a nonzero eigenvalue λ. Then there exists a nonzero vector $v \in \boldsymbol{C}G$ such that $ab \cdot v = \lambda v$. Consider also another vector $w = (1/\sqrt{\lambda})b \cdot v$. These two vectors span a subspace $V = \mathrm{lin}_{\boldsymbol{C}}(v, w) \subset \boldsymbol{C}G$.

Now, V is invariant under the left multiplication by both a and b. In fact, it is easy to check that we have $a \cdot v = 0$, $a \cdot w = \sqrt{\lambda} \cdot v$, $b \cdot v = \sqrt{\lambda} \cdot w$ and $b \cdot w = 0$. The above calculation also shows that v, w are linearly independent and hence V is a complex plane. Left multiplications by the units $1 + a, 1 + b \in \mathcal{U}_1(\boldsymbol{Z}G)$ are linear endomorphisms of V corresponding, with respect to the basis $\{v, w\}$, to the matrices $\left(\begin{smallmatrix} 1 & \sqrt{\lambda} \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & 0 \\ \sqrt{\lambda} & 1 \end{smallmatrix}\right)$. We know (7.1) that such pairs of matrices generate a free subgroup in $GL_2(\boldsymbol{C})$, provided $\sqrt{|\lambda|} \geqslant 2$. In particular, we then have an epimorphism of $\langle 1 + a, 1 + b\rangle \subseteq \mathcal{U}_1(\boldsymbol{Z}G)$ onto a free subgroup of $GL_2(\boldsymbol{C})$, i.e., we obtain an explicit construction of a free subgroup of integral group ring units.

We are thus left with the following question: how to pick a pair a, b of nilpotent elements to ensure a large eigenvalue for the left multiplication by ab?

In fact, this is easy: we can pick b arbitrarily and take $a = b^* = \sum \bar{b}_g g^{-1}$. The star operator is strongly related to the inner product: $(a, b) = \mathrm{tr}(a^*b)$ where tr denotes the coefficient at 1. In particular, for any triple $c, a, b \in \boldsymbol{C}G$ the relation $(c^*a, b) = (a, cb)$ holds. As $c := b^*b$ satisfies $c = c^*$, it follows that left multiplication by c is a symmetric operator with respect to $(.\,,.)$.

Consider the linear subspace $W = \mathrm{lin}_{\boldsymbol{C}}\{1, c, c^2, \ldots\} \subseteq \boldsymbol{C}G$. Clearly, W is invariant under $T(x) = c \cdot x$. If, in addition, $\dim(W) = n < \infty$ (which is always the case when G is finite) then W has a unitary basis $\{v_1, \dots, v_n\}$ of eigenvectors for T. Let $T(v_i) = \lambda_i v_i$. We show that at least one of the numbers $|\lambda_i|$ exceeds 4.

To this end, write $1 \in W$ as a combination of the basis vectors: $1 = \sum t_i v_i$. Then $c = T(1) = \sum t_i \lambda_i v_i$. Let $\mu = \max\{|\lambda_i|\}$. Then, by the oldest theorem of geometry, the Pythagoras theorem, we have:

$$\|c\|^2 = \sum |t_i \lambda_i|^2 \leqslant \mu^2 \sum |t_i|^2 = \mu^2 \cdot \|1\|^2 = \mu^2.$$

Hence

$$\mu \geqslant \|c\| \geqslant |c_1| = \mathrm{tr}(b^* b) = \|b\|^2 \geqslant 4.$$

The last inequality is easy to see as c has integral coefficients and has augmentation zero. If $\dim(W) = \infty$ then the above argument does not work but the situation is even simpler: the monoid generated by a and b has no other relations but $a^2 = b^2 = 0$, hence may be mapped to matrices so that $a \to \left(\begin{smallmatrix} 0 & 2 \\ 0 & 0 \end{smallmatrix}\right)$, $b \to \left(\begin{smallmatrix} 0 & 0 \\ 2 & 0 \end{smallmatrix}\right)$ (see Marciniak and Sehgal [MS5]). In any case, we have

THEOREM 7.4 (Marciniak and Sehgal). *Let G be an arbitrary group. If $b \in \mathbf{Z}G$ satisfies $b^2 = 0$, $b \neq 0$ then the subgroup $\langle 1 + b, 1 + b^* \rangle$ of $\mathcal{U}_1(\mathbf{Z}G)$ is free.*

COROLLARY 7.5. *Let G be an arbitrary group. Suppose that $u = 1 + (x-1)y\hat{x}$, $x, y \in G$, $o(x) < \infty$, is a bicyclic unit $\neq 1$. Then the subgroup $\langle u, u^* \rangle \subseteq \mathcal{U}_1(\mathbf{Z}G)$ is free.*

By using (7.3) we have

THEOREM 7.6. *Let G be an arbitrary group. Suppose $\mathcal{U}(\mathbf{Z}G)$ contains no free subgroups. Then* (i) *the torsion elements of G form a subgroup T which is Abelian or a Hamiltonian 2-group. Conversely, if we assume that* (ii) *G contains no free subgroups and* (iii) *G/T is right ordered then $\mathcal{U}(\mathbf{Z}G)$ contains no free subgroups.*

It is easy now to classify groups whose rational group algebras contain no free subgroups of units under the assumption that G/T is right ordered. We do need an assumption like this for the converse as we do not know if the unit conjecture (3.16) is true. The result for $\boldsymbol{Q}G$ is similar to the last one with T Abelian.

Characteristic p

Let us turn to the modular case and ask if there are explicit units generating a free group. If G is finite then K cannot be algebraic. In fact, Gonçalves and Passman [GP1] gave the following analogue of (7.5).

THEOREM 7.7. *Let K be a field of characteristic $p > 0$ containing an element t transcendental over its prime subfield. Let G be a group which has two elements x, y such* *that x has finite order n, y does not normalize $\langle x\rangle$, and the subgroup $\langle x, y^{-1}xy\rangle$ has no p-torsion. If we let*

$$a=(1-x)y\hat{x}, \qquad b=\hat{x}y^{-1}\bigl(1-x^{\delta}\bigr) \quad \textit{where } \delta=(-1)^p,$$

then $\mathcal{U}(KG)$ contains

$$\bigl\langle 1+ta,\ 1+tbab,\ 1+t(1-b)aba(1+b)\bigr\rangle \simeq \mathbf{Z}_p * \mathbf{Z}_p * \mathbf{Z}_p.$$

COROLLARY 7.8. *If G is a non-Abelian torsion p'-group and K is not algebraic over its prime subfield F_p, then $\mathcal{U}(KG)$ contains a free group.*

In [GP2] the authors handle the general situation. Furthermore, Gonçalves [Go1] has proved the

THEOREM 7.9. *Let G be a finite group and K a field of characteristic $p>0$. Then $\mathcal{U}(KG)$ does not contain a free subgroup if and only if one of the following conditions occurs*:

(i) *G is Abelian,*

(ii) *K is algebraic over $\boldsymbol{F}_p$,*

(iii) *$S_p(G)$, the Sylow p-subgroup of G, is normal in G, and $G/S_p(G)$ is Abelian.*

8. Zassenhaus conjectures

We know by a theorem of Graham Higman (3.2) that all torsion units of the integral group ring $\boldsymbol{Z}A$ of a finite Abelian group A are trivial; being of the form $\pm a$, $a\in A$. If a finite group G is non-Abelian then, of course, along with $\pm g$, all conjugates $\alpha^{-1}g\alpha$, $\alpha\in\mathcal{U}(\boldsymbol{Z}G)$, $g\in G$, are also of finite order. These are not all the torsion units as can be seen in $\boldsymbol{Z}S_3$ (see Hughes and Pearson [HP]). Accordingly, in the mid 1960's Hans Zassenhaus made several conjectures. The basic idea is that all torsion units are the obvious ones if you allow conjugation by units in $\boldsymbol{Q}G$. We describe below these conjectures and their relationship to the isomorphism problem (ISO) of Section 4. The groups G in this section are finite and all automorphisms of $\boldsymbol{Z}G$ are normalized with respect to ε_G, the augmentation associated to the group basis G.

(ZC1) $u\in\mathcal{U}_1(\boldsymbol{Z}G)$, $o(u)<\infty \Rightarrow u \underset{\boldsymbol{Q}G}{\sim} g$ for some $g\in G$.

(ZC2) $\boldsymbol{Z}G=\boldsymbol{Z}H$, $\varepsilon_G(H)=1 \Rightarrow \alpha^{-1}H\alpha=G$ for some $\alpha\in\mathcal{U}(\boldsymbol{Q}G)$.

(ZC3) H a finite subgroup of $\mathcal{U}_1(\boldsymbol{Z}G)\Rightarrow \alpha^{-1}H\alpha\subseteq G$ for some $\alpha\in\mathcal{U}(\boldsymbol{Q}G)$.

(Aut) $\theta\in \mathrm{Aut}(\boldsymbol{Z}G)\Rightarrow \exists\ \beta\in\mathrm{Aut}(G)$, $\alpha\in\mathcal{U}(\boldsymbol{Q}G)$ such that $\theta=I_\alpha\circ\beta$, where I_α is conjugation by α.

Let us make a few observations. We have already seen (3.9) that every finite subgroup H of $\mathcal{U}_1(\boldsymbol{Z}G)$ is linearly independent and the order of H divides $|G|$. Also, we know

LEMMA 8.1. *If H is a subgroup of $\mathcal{U}_1(\boldsymbol{Z}G)$ with $|H|=|G|$ then $\boldsymbol{Z}G=\boldsymbol{Z}H$.*

PROOF. Since H is linearly independent we have $\boldsymbol{Q}G = \boldsymbol{Q}H$. It follows that $\boldsymbol{Z}H$ is a suborder of $\boldsymbol{Z}G$ and consequently $n\boldsymbol{Z}G \leqslant \boldsymbol{Z}H$ for some natural number n. We shall prove that $n = 1$. Let $g \in G$. Then $ng = \sum z_i h_i$, $z_i \in \boldsymbol{Z}$, $h_i \in H$. It is enough to show that each z_i is divisible by n. We have

$$ngh_i^{-1} = z_i + \sum_{j \neq i} z_j \left(h_j h_i^{-1}\right).$$

Since by (3.1) $(h_j h_i^{-1})(1) = 0$, the coefficient of one on the right-hand side is z_i whereas it is a multiple of n from the left-hand side. It follows that $n | z_i$ and we are done. $\square$

We see that (ZC3) concerns any finite subgroup of $\mathcal{U}_1(\boldsymbol{Z}G)$, (ZC2) deals with maximal order finite subgroups whereas (ZC1) has to do with cyclic subgroups of $\mathcal{U}_1(\boldsymbol{Z}G)$. Clearly,

$$\text{(ZC3)} \quad \Rightarrow \quad \text{(ZC1) and (ZC2).} \tag{$*$}$$

Moreover,

$$\text{(ZC2)} \quad \Rightarrow \quad \text{(ISO).} \tag{$**$}$$

PROOF. Suppose that $\theta : \boldsymbol{Z}H \to \boldsymbol{Z}G$ is an isomorphism. Then $\boldsymbol{Z}G = \boldsymbol{Z}H^{\theta}$. Therefore, by (ZC2), $H^{\theta} = \alpha^{-1} G \alpha$ for a suitable $\alpha \in \boldsymbol{Q}G$. We deduce that $H \simeq H^{\theta} = \alpha^{-1} G \alpha \simeq G$. $\square$

Further we have the implications:

$$\text{(ZC2)} \quad \Rightarrow \quad \text{(Aut).} \tag{$***_1$}$$

$$\text{(Aut)} + \text{(ISO)} \quad \Rightarrow \quad \text{(ZC2).} \tag{$***_2$}$$

PROOF. (1) Suppose (ZC2). Let $\theta \in \operatorname{Aut}(\boldsymbol{Z}G)$. Then $\boldsymbol{Z}G = \boldsymbol{Z}G^{\theta}$. By (ZC2), $G^{\theta} = \alpha^{-1} G \alpha$ for some $\alpha \in \boldsymbol{Q}G$. Then

$$\theta(g) = \alpha^{-1} g_1 \alpha, \quad g_1 \in G.$$

Clearly, $g \to g_1$ is an automorphism and we have (Aut).

(2) Suppose (Aut) and (ISO). Let $\boldsymbol{Z}G = \boldsymbol{Z}H$. Then by (ISO) there is an isomorphism $G \xrightarrow{\theta} H$. By (Aut) there exists $\alpha \in \boldsymbol{Q}G$ and $\sigma \in \operatorname{Aut}(G)$ such that $\theta(g) = \alpha^{-1} g^{\sigma} \alpha$. Consequently, $H = \alpha^{-1} G \alpha$. $\square$

The conjugating element may be found in any larger field, as it is only a matter of solving some linear equations.

LEMMA 8.2. *Let $K \geqslant k$ be infinite fields. Suppose that H_1 and H_2 are two finite subgroups of units in kG. Then, for their conjugacy, we have*

$$H_1 \sim H_2 \text{ in } KG \quad \Rightarrow \quad H_1 \sim H_2 \text{ in } kG.$$

PROOF. See [Se13, p. 208]. □

(ZC2) was proved by Roggenkamp and Scott [RoS5] for nilpotent groups. They also gave a counterexample to (ZC2) which appears in a modified form in a paper of Klingler [Kli]. A. Weiss [We2,We3] proved (ZC3) (and hence (ZC1) and (ZC2)) for nilpotent groups.

Returning to (ZC1), it remains open in general. Besides nilpotent groups it is known to be true for certain split metacyclic groups as proved by Polcino Milies, Ritter and Sehgal [PRS].

THEOREM 8.3. *If $G = \langle a\rangle \rtimes \langle b\rangle$ is the semidirect product of two cyclic groups of relatively prime orders then* (ZC1) *holds for G.*

The proof consists of looking at the absolutely irreducible representations ρ and proving that for a torsion unit $u \in \mathcal{U}_1(\boldsymbol{Z}G)$, $\rho(u) \sim \rho(g)$ for a fixed $g \in G$. A special case is based on a Hilbert 90 argument, first used by Bhandari and Luthar [BL1] in this connection.

LEMMA 8.4. *Let G be a split extension $A \rtimes X$ where A is a normal p-group and X any group. Let $u \in \mathcal{U}_1(\boldsymbol{Z}G)$ be a unit of the form $u = vw$, $v \in \mathcal{U}(1 + \Delta(G, A))$, $w \in \mathcal{U}(\boldsymbol{Z}X)$. If u has finite order s not divisible by p then u is conjugate to w in $\boldsymbol{Q}G$.*

PROOF. The split exact sequence

$$1 \to A \to G \to X \to 1$$

induces the split exact sequence

$$1 \to \mathcal{U}\big(1 + \Delta(G, A)\big) \to \mathcal{U}_1(\boldsymbol{Z}G) \to \mathcal{U}_1(\boldsymbol{Z}X) \to 1$$

and we have $\mathcal{U}_1(\boldsymbol{Z}G) = \mathcal{U}(1 + \Delta(G, A)) \rtimes \mathcal{U}_1(\boldsymbol{Z}X)$. Writing $v^w = wvw^{-1}$ we have

$$1 = u^s = v \cdot v^w \cdot v^{w^2} \cdots v^{w^{s-1}} \cdot w^s.$$

It follows that $w^s = 1$ and $v \cdot v^w \cdots v^{w^{s-1}} = v^{1+w+\cdots+w^{s-1}} = 1$. It should be noted that the elements in the last product do not commute. Writing

$$z = 1 + v + v^{1+w} + \cdots + v^{1+w+\cdots+w^{s-2}}$$

we get

$$wzw^{-1} = 1 + v^w + v^{w+w^2} + \cdots + v^{w+w^2+\cdots+w^{s-1}}.$$

Thus

$$vwzw^{-1} = v + v^{1+w} + v^{1+w+w^2} + \cdots + v^{1+w+\cdots+w^{s-1}} = z.$$

Now, since $v \equiv 1 \pmod{\Delta(G, A)}$ it follows that $z \equiv s \bmod \Delta(G, A)$. It can be proved that z is invertible in $\boldsymbol{Q}G$ and we can conclude that $z^{-1}vwz = w$. □

Theorem 8.3 was strengthened to (ZC3) by Valenti [V2].

(Aut) was proved for symmetric groups S_n by Peterson [Pet1]. Kimmerle [RT] proved that the class of groups satisfying (Aut) is closed under direct products. He also considered certain variations of the (Aut) conjecture (see [RT]). (Aut) is proved for wreath products H wr S_n where H is Abelian or a p-group by Giambruno, Sehgal and Valenti [GSV1, GSV2] and when H is S_k for some k by Valenti [V1]. For some families of simple groups also (ZC2) has been proved (see [BKRW,BHK,BK]). But it remains open for A_n.

There are several interesting extensions of the above conjectures. Suppose that A is a normal subgroup of index n in G. Then $\boldsymbol{Z}G$ is a right $\boldsymbol{Z}A$-module. By left multiplication $\boldsymbol{Z}G$ can be represented by $n \times n$ matrices over $\boldsymbol{Z}A$. Any torsion unit u of $\boldsymbol{Z}G$ which is mapped by the natural homomorphism $G \to G/A$ to $1 \in \boldsymbol{Z}(G/A)$ gives rise to a torsion matrix $U \in SGL_n(\boldsymbol{Z}A)$. Here, $SGL_n(\boldsymbol{Z}A)$ denotes the subgroup of the general linear group $GL_n(\boldsymbol{Z}A)$ consisting of the matrices U which are mapped by the augmentation homomorphism $\boldsymbol{Z}A \to \boldsymbol{Z}$, when applied to each entry, to the identity matrix. Thus (ZC1) translates to the question about diagonalization of U in $GL_n(\boldsymbol{Q}A)$. We have then the

PROBLEM 8.5. Let $U \in SGL_n(\boldsymbol{Z}G)$, where G is a finite group, be a torsion matrix. Is U conjugate in $(\boldsymbol{Q}G)_{n\times n}$ to a matrix $\mathrm{diag}(g_1, \dots, g_n)$, $g_i \in G$?

This was answered positively by Weiss [We2] for p-groups. Cliff and Weiss [CW2] provide an explicit example of a matrix $U \in SGL_6(\boldsymbol{Z}(C_6 \times C_6))$ such that $U^6 = I$ which cannot be diagonalized. They also prove that such a matrix U exists for finite nilpotent G and some n if and only if G has at least two noncyclic Sylow p-subgroups.

However, it was proved by Luthar and Passi [LP2] that if $n = 2$ and G is finite Abelian then U is conjugate in $(\boldsymbol{Q}G)_{n\times n}$ to $\mathrm{diag}(g_1, g_2)$. This has been extended to $n \leqslant 5$ by Marciniak and Sehgal [MS7], bridging the gap between 2 and 6.

We see that there is a relationship between n and $|G|$ for a counterexample to exist. In this connection there are two results that we mention.

THEOREM 8.6 (Marciniak, Ritter, Sehgal and Weiss [MRSW]). *Let A be Abelian and U a torsion matrix in $SGL_n(\boldsymbol{Z}A)$. If $n < p$ for all prime divisors p of $|A|$ then*

$$U \sim \mathrm{diag}(a_1, \dots, a_n), \quad a_i \in A,$$

in $(\boldsymbol{Q}A)_{n\times n}$.

THEOREM 8.7 (Lee and Sehgal [LS2]. *Let A be a finite Abelian group. Suppose that either*

(1) *a has at most one noncyclic Sylow subgroup, or*
(2) *if q_1 and q_2 are the two smallest (distinct) primes such that the Sylow q_1 and q_2-subgroups are noncyclic then $q_1 + q_2 > \frac{n^2+n-8}{4}$.*

Then any torsion matrix $U \in SGL_n(\boldsymbol{Z}A)$ is conjugate in $GL_n(\boldsymbol{Q}A)$ to a diagonal matrix with group elements in the diagonal.

Part (1) is, of course, [CW2]. See [LS2] for details.

The problem remains open for nilpotent G and $n < 6$.

Zassenhaus conjectures for infinite groups

The problems raised by Zassenhaus in his conjectures in the context of finite groups are also very important for the study as well as for the applications of group rings $\boldsymbol{Z}G$ of infinite groups G. The knowledge of units of finite multiplicative order leads to a better understanding of idempotents and, more generally, projective $\boldsymbol{Z}G$-modules. Also, the study of automorphisms of any algebraic object is nowadays a standard research tool.

The Zassenhaus conjectures, as stated for finite groups, also make perfect sense for infinite groups.

(Aut) If $\sigma \in \mathrm{Aut}(\boldsymbol{Z}G)$ is an augmentation preserving automorphism then $\sigma = \tau_\gamma \circ \alpha$ where $\alpha \in \mathrm{Aut}(G)$ and τ_γ is conjugation by a rational unit $\gamma \in \boldsymbol{Q}G$. Thus, we have $\sigma(g) = \gamma^{-1} g^\alpha \gamma$ for all $g \in G$.

(ZC1) Every torsion unit of augmentation one in $\boldsymbol{Z}G$ is conjugate in $\boldsymbol{Q}G$ to a group element $g \in G$.

The automorphisms

An easy counterexample to (Aut) for infinite groups is due to Sehgal and Zalesskii [Se13, p. 279]. The idea is as follows.

Let S_n denote the group of all permutations of the set $\{1, 2, \ldots, n\}$. For any $n > 2$ there exists a unit in $\boldsymbol{Z}S_n$ such that $\tau_u \neq \tau_g$ for all $g \in S_n$. In fact, if $\tau_u = \tau_g$ then ug^{-1} is central in the group $\mathcal{U}$ of units in $\boldsymbol{Z}S_n$. However, it is easy to prove that the centre of $\mathcal{U}$ is finite, while the group $\mathcal{U}$ itself is infinite.

Fix now $n > 6$ and set $G_i = S_n$ for all $i \geqslant 1$. Consider the restricted direct product $G = \dot{\prod} G_i$. For any i pick a unit v_i in $\boldsymbol{Z}G_i$, as in the last paragraph. It defines an inner automorphism τ_{v_i} of $\boldsymbol{Z}G_i$. Then the automorphism

$$\varphi : \boldsymbol{Z}G \to \boldsymbol{Z}G, \quad \varphi = \prod_{i \geqslant 1} \tau_{v_i},$$

cannot be presented in the form required by the (Aut) conjecture.

In fact, suppose that $\varphi = \tau_\gamma \circ \alpha$ for some $\alpha \in \mathrm{Aut}(G)$ and a rational unit $\gamma \in \boldsymbol{Q}G$. As γ has finite support, there exists j such that τ_γ centralizes all subgroups G_i for $i > j$. Then for $i > j$ the map $\alpha_{|G_i}$ preserves G_i and coincides with τ_{v_i}. However, it is well known that for $n \neq 6$ all automorphisms of S_n are inner; in particular $\tau_{v_i} = \alpha_{|G_i} = \tau_g$ for some $g \in S_n$ – a contradiction with the choice of the unit v_i.

It is clear that instead of S_n we could use in the above construction any group H such that $\mathrm{Aut}(H) = \mathrm{Inn}(H)$ and the group of units of $\boldsymbol{Z}H$ is infinite with finite centre.

The (ZC1) conjecture for infinite groups

Here we ask whether a given torsion unit $u \in \boldsymbol{Z}G$ of augmentation one can be rationally conjugated to a group element $g \in G$. We start with a description of a weaker *unique trace* condition, which however can be easily checked with u alone, without using an auxiliary unit from $\boldsymbol{Q}G$.

Let $u = \sum u(g)g$ be any element of $\boldsymbol{Z}G$, where $g \in G$ and $u(g) \in \boldsymbol{Z}$. For any element $h \in G$ consider the number $\tilde{u}(h) = \sum_{g \sim h} u(g)$, where by $g \sim h$ we mean that g is conjugate to h in G. We call the numbers $\tilde{u}(h)$ *the traces* of $u \in RG$.

It is easy to check that conjugate group ring elements have all traces equal. In particular, if $u = \gamma^{-1} g_0 \gamma$ for some $g_0 \in G$ then $\tilde{u}(g_0) = 1$, $\tilde{u}(g) = 0$ for g not conjugate to g_0. In other words: if u is conjugate to a group element then it has only one trace not equal to zero. The next result shows that the above condition is very close to (ZC1).

THEOREM 8.8 (see [Se13, p. 238]). *For an arbitrary group G consider the following conditions*:

(a) *The* (ZC1) *conjecture holds for G.*

(b) *For any torsion unit $u = \sum u(g)g \in \mathcal{U}_1(\boldsymbol{Z}G)$ there exists a unique* (*up to conjugacy*) *group element g_0 such that $\tilde{u}(g_0) \neq 0$.*

Condition (a) *always implies condition* (b). *If the group G is finite then the two conditions are equivalent.*

Motivated by the above result, we say that G is a *UT*-group (a *unique trace* group) if for every torsion unit $u \in \mathcal{U}_1(\boldsymbol{Z}G)$ there exists a unique (up to conjugacy) element g_0 of G such that $\tilde{u}(g_0) \neq 0$.

If we want to extend the positive results about (ZC1) to a class of infinite groups we must first prove that this class consists of *UT*-groups. Here we have a result of this type.

THEOREM 8.9 (Bovdi, Marciniak and Sehgal [BMS]). *Every nilpotent group is a UT-group.*

The idea of the proof is as follows. From Weiss' results [We2,We3] we know that all finite nilpotent groups are *UT*-groups. Let G be an infinite nilpotent group. Without loss of generality we may assume that G is finitely generated. Then G is residually finite and conjugacy separable (i.e., every pair of conjugacy classes remains distinct in a finite homomorphic image of G). In particular, G has a finite homomorphic image $\overline{G}$ such that: (i) the support of u maps injectively to $\overline{G}$; (ii) the conjugacy classes in G, which intersect the support of u, map to distinct conjugacy classes in $\overline{G}$.

Because the image of u in $\boldsymbol{Z}\overline{G}$ has one nonzero trace only, the same holds for u, by conditions (i) and (ii).

Returning to (ZC1) let us recall that no counterexample is known for finite groups G. For infinite nilpotent groups there is a counterexample due to Marciniak and Sehgal [MS3].

The nilpotent group used here is $G = H \times D_8$, where $H = (H_0) \rtimes_\sigma \boldsymbol{Z}$ with $H_0 = \boldsymbol{Z} \times \boldsymbol{Z}$, $\sigma \in \operatorname{Aut}(H_0)$ given by $\sigma(x_1, x_2) = (x_1 + x_2, x_2)$ and $D_8 = \langle x, y \colon x^4 = 1 = y^2,\ x^y = x^{-1} \rangle$ is the dihedral group.

To describe the unit $u \in \mathcal{U}_1(ZG)$ which is not conjugate to any group element we need a few auxiliary elements in $\boldsymbol{Z}G$. Let $t \in H$ generate $H/(\boldsymbol{Z} \times \boldsymbol{Z})$, let $X_0 = (0, 1) - 1 \in \boldsymbol{Z}[\boldsymbol{Z} \times \boldsymbol{Z}]$ and let $X_i = t^i X_0 t^{-i}$ for $i \geqslant 1$. Define the following elements in $\boldsymbol{Z}G$:

$$\alpha = 2t^{-2} X_1 X_2, \qquad \gamma = 2t^{-2} X_0 X_2 + t^{-1} X_0,$$
$$\beta = 2t^{-2} X_0 X_3, \qquad \delta = 2t^{-2} X_1 X_3 - t^{-1} X_2.$$

THEOREM 8.10. *Consider the element*

$$u = y + \bigl[(\alpha - \beta) + (\alpha + \beta)y + (\gamma + \delta)x + (\gamma - \delta)yx\bigr] \cdot \bigl(x^2 - 1\bigr) \in \boldsymbol{Z}G.$$

Then $u^2 = 1$ and u is not conjugate to a trivial unit in KG for any field K of characteristic zero.

For the proof we identify

$$KG \simeq KH \otimes \boldsymbol{Q}D_8 \xrightarrow{\simeq} K\bigl[H \times D_8/\langle x^2\rangle\bigr] \oplus M_2(KH).$$

Our unit $u \in \boldsymbol{Z}G \subset KG$ is mapped to the pair (y, U) where $y \in D_8$ and $U \in M_2(KH)$. It is easy to calculate that $U^2 = I$ and hence $u^2 = 1$ as well.

To prove the second part of the statement it is enough to show that U is not conjugate to any matrix with K-coefficients. Equivalently, we prove the same about the idempotent matrix $E = (I + U)/2$. Our statement then translates to the condition that the projective KH-module P which is the kernel of the multiplication $KH \oplus KH \xrightarrow{\cdot E} KH \oplus KH$ is stably free of rank one but not free. This can be verified by a standard identification of P with a left ideal of KH which is not principal.

For some infinite groups (ZC1) is known to be true.

THEOREM 8.11. *In the following cases* (ZC1) *is true.*

(i) *Levin and Sehgal* [LeS1]:

$$G = D_\infty = C_2 * C_2 = \langle t, z\colon z^2 = 1,\ t^z = t^{-1}\rangle;$$

(ii) *Bovdi, Marciniak and Sehgal* [BMS]

$$G = T \times A \quad \textit{where } T \textit{ is finite nilpotent and } A \textit{ is torsion free Abelian.}$$

(iii) *Dokuchaev and Sobral Singer* [DSS]:

$$G = C_1 * \cdots * C_n \quad \textit{where } C_i \textit{ is cyclic of prime order } p_i.$$

Moreover, when $G = C_{m_1} * C_{m_2} * \cdots * C_{m_k}$ (ZC1) is true provided we allow the conjugating element to be found in a very large ring containing $\boldsymbol{Q}G$, see Lichtman and Sehgal [LiS].

There are more positive results concerning (ZC1) for infinite groups. We need only to adjust the conjecture slightly. To find the right adjustment, recall that the above example was based on the observation that the group algebra $\boldsymbol{Q}H$ has a finitely generated projective module P which is not free. The group H is a torsion free, finitely generated nilpotent group. It is well known that $\widetilde{K}_0(\boldsymbol{Q}H)=0$ for such groups. Therefore our projective module P is *stably free*. In our case: $P\oplus \boldsymbol{Q}H\simeq \boldsymbol{Q}H\oplus \boldsymbol{Q}H$. The next theorem easily follows.

THEOREM 8.12 [MS3]. *The unit* $u\in \boldsymbol{Z}[H\times D_8]$, *constructed in Theorem* 8.10, *has an additional property*: *the matrices* $\left(\begin{smallmatrix} u & 0\\ 0 & -1\end{smallmatrix}\right)$, $\left(\begin{smallmatrix} y & 0\\ 0 & -1\end{smallmatrix}\right)$ *are conjugate in* $GL_2(\boldsymbol{Q}G)$.

Motivated by this result, we introduce the following

DEFINITION 8.13. Let R be a K-algebra over a field K. Two matrices $A,B\in M_n(R)$ are *stably conjugate* ($A\sim_s B$) if there exist roots of unity $\xi_1,\dots,\xi_d\in K$ so that

$$A\oplus \operatorname{diag}(\xi_1,\dots,\xi_d)\sim B\oplus \operatorname{diag}(\xi_1,\dots,\xi_d)\quad \text{in } M_{n+d}(R).$$

For any (associative) K-algebra R we have the *Bass rank map*

$$r: M_n(R)\to R/[R,R],\quad r(A)=\operatorname{Tr}(A)\ (\operatorname{mod}[R,R]).$$

The map r is K-linear and satisfies the standard trace condition $r(AB)=r(BA)$. When restricted to the set of idempotent matrices it defines a map $r: K_0(R)\to R/[R,R]$. The rank map is related to stable conjugation by the following criterion.

THEOREM 8.14 [BMS]. *Let $R=KG$ and let K have enough roots of unity. Suppose the rank map $r: K_0(R)\to R/[R,R]$ is injective. For a matrix $U\in M_n(R)$ satisfying $U^d=I$ the following conditions are equivalent*:

(i) $U\sim_s \operatorname{diag}(g_1,\dots,g_n)$ *for some* $g_i\in G$.

(ii) *There exists a diagonal matrix $D=(g_1,\dots,g_n)$ with $g_i\in G$ so that $r(U^k)=r(D^k)$ for all $k\in N$.*

The rank map $r: K_0(KG)\to KG/[KG,KG]$ is injective for finitely generated nilpotent groups G. Also, when a matrix U augments to identity and d is a power of a prime number then it is possible to find a diagonal matrix D satisfying (ii) in the above criterion.

THEOREM 8.15 [MS2]. *Let G be nilpotent. Suppose that $U\in SGL_n(\mathbb{Z}G)$ satisfies $U^{p^m}=I$. Then $U\sim_s \operatorname{diag}(g_1,\dots,g_n)$, $g_i\in G$.*

The notion of stable conjugacy of matrices can be generalized to pairs of representations $\phi,\psi: H\to SGL_n(\boldsymbol{Z}G)$ for an arbitrary finite group H. Let $\rho: H\to GL_d(\boldsymbol{Z})\subseteq GL_d(\boldsymbol{Z}G)$ be the regular representation, where $d=|H|$.

DEFINITION 8.16. Representations ϕ and ψ are *stably conjugate* over KG if there exist a natural number k and a matrix $Y\in GL_{n+kd}(KG)$ such that $(\phi\oplus\rho^k)(h)=Y\cdot(\psi\oplus\rho^k)(h)\cdot Y^{-1}$ holds for all $h\in H$. We write then $\phi\sim_s\psi$.

Notice that when the group H is cyclic and the field K contains a primitive dth root of unity then the regular representation ρ can be diagonalized and we get back the definition of stable conjugation of matrices.

PROPOSITION 8.17 [MS2]. *Suppose that the rank map for KG is injective. If H is a finite group and K is its splitting field then for any representations $\phi, \psi : H \to GL_n(KG)$ the following conditions are equivalent*:

(i) $r \cdot \phi = r \cdot \psi$,

(ii) $\phi \sim_s \psi$.

Let $\delta_n(G)$ denote the group of all diagonal matrices $\operatorname{diag}(g_1, \dots, g_n) \in SGL_n(\mathbb{Z}G)$, $g_i \in G$.

THEOREM 8.18 [MS2]. *Let G be a nilpotent group and let P be a finite p-group. For each representation $\phi : P \to SGL_n(\boldsymbol{Z}G)$ there exists a diagonal representation $\psi : P \to \delta_n(G) \subseteq SGL_n(\boldsymbol{Z}G)$ such that ϕ and ψ are stably conjugate over $\boldsymbol{C}G$.*

Returning to group ring units, as an application, we obtain the following result.

PROPOSITION 8.19. *If G is a nilpotent group then every finite subgroup $H \subseteq \mathcal{U}_1(\boldsymbol{Z}G)$ is isomorphic to a subgroup of G.*

Some weaker versions of the Zassenhaus conjectures have also been studied. For these we refer the reader to [Do] and [RT].

9. Dimension subgroups and related questions

"*There is a question in group theory called the 'dimension subgroup problem'. Though rather technical to state, it has an important status in group theory. There has been much work on it, and, in particular, several erroneous proofs have been published*". So wrote Irving Kaplansky in the 1974 yearbook of the Encyclopedia Brittanica [Kap5]. We introduce this problem now. Let R be an integral domain and G any group. By the nth dimension subgroup of G over R we understand

$$D_{n,R}(G) = G \cap \left(1 + \Delta_R^n(G)\right) = \left\{g \in G\colon\ g - 1 \in \Delta^n(G)\right\}.$$

We drop the subscript R whenever it is convenient to do so. Identification of $D_{n,\boldsymbol{Z}}(G)$ is the dimension subgroup problem. The problem was first introduced by Magnus [M] who proved that for a free group F, $D_{n,\boldsymbol{Z}}(F) = \gamma_n(F)$, the nth term of the lower central series of F. Zassenhaus [Zas1] computed $D_{n,K}(F)$ when K is a field of characteristic p. By using the identities

$$\begin{aligned} xy - 1 &= (x-1)(y-1) + (x-1) + (y-1), \\ xyx^{-1} - 1 &= x(y-1)x^{-1}, \end{aligned}$$

it follows that $D_{n,R}(G)$ is a normal subgroup of G, giving us a normal series

$$G = D_{1,R}(G) \supseteq D_{2,R}(G) \supseteq \cdots \supseteq D_{n,R}(G) \supseteq \cdots.$$

Moreover, $\gamma_n(G)$, the nth term of the lower central series of G, is contained in $D_{n,R}(G)$ because $(D_{i,R}(G))$ is a central series as seen below.

LEMMA 9.1.
(i) $(D_i(G), D_j(G)) \subseteq D_{i+j}(G)$ *for all* $i, j \geqslant 1$.
(ii) *If the characteristic of* R *is* (*a prime*) p *then* $(D_i(G))^p \subseteq D_{ip}(G)$.

PROOF. (i) follows from the identity

$$(x, y) - 1 = x^{-1}y^{-1}(xy - yx) = x^{-1}y^{-1}\big[(x-1)(y-1) - (y-1)(x-1)\big].$$

(ii) is a consequence of $g^p - 1 = (g-1)^p$ in characteristic p.
To compute $D_{n,R}(G)$ for any integral domain R it suffices to know the answer for $R = \mathbf{Z}$ and for prime fields as was proved by Sandling [San3] and Parmenter [Par1]. □

THEOREM 9.2. *Let* R *be an integral domain and* G *any group. We have*:
(i) *if the characteristic of* R *is* p *then* $D_{n,R}(G) = D_{n,\mathbf{Z}/p\mathbf{Z}}(G)$,
(ii) *if the characteristic of* R *is zero then*

$$D_{n,R}(G) = \prod_{p\in\mathcal{U}(R)} \big\{g \in G\colon\ g^{p^k} \in D_{n,\mathbf{Z}}(G) \text{ for some } k\big\}.$$

It follows that for fields of characteristic p we have $D_{n,K}(G) = D_{n,\mathbf{Z}/p\mathbf{Z}}(G)$ whereas for a field K of characteristic zero we have

$$D_{n,K}(G) = \sqrt{D_{n,\mathbf{Z}}(G)} = \big\{g \in G\colon\ g^k \in D_{n,\mathbf{Z}}(G) \text{ for some } k\big\} = D_{n,\mathbf{Q}}(G).$$

Also, $D_{n,K}(G)$ can be completely described in terms of the structure of G. We need a couple of definitions. Let p be a fixed prime. We define the Brauer–Jennings–Zassenhaus M-series $\{M_{n,p}(G)\}_{n\geqslant 1}$ of a group G inductively by $M_{1,p}(G) = G$ and for $n \geqslant 2$

$$M_{n,p}(G) = \big(G, M_{n-1,p}(G)\big)M_i^p(G),$$

where i is the smallest integer satisfying $ip \geqslant n$. This is then the minimal central series of G with the property:

$$x \in M_{n,p}(G) \quad \Rightarrow \quad x^p \in M_{np,p}(G) \quad \text{for all } n \geqslant 1. \tag{$*$}$$

The Lazard series $\{L_{n,p}(G)\}_{n\geqslant 1}$, of G is given by

$$L_{n,p}(G) = \prod_{ip^j \geqslant n} \gamma_i(G)^{p^j}.$$

Then it is easy to see that

PROPOSITION 9.3. *For any group G we have*
(i) $D_{n,\boldsymbol{Q}}(G) \supseteq \sqrt{\gamma_n(G)}$, *the isolator of* $\gamma_n(G)$,
(ii) $D_{n,\boldsymbol{Z}/p\boldsymbol{Z}}(G) \supseteq M_{n,p}(G) \supseteq L_{n,p}(G)$.

PROOF. (i) Suppose that $g \in \sqrt{\gamma_n(G)}$, i.e., $g^m \in \gamma_n(G)$ for some $m > 0$. Thus $g^m - 1 \in \Delta^n_{\boldsymbol{Q}}(G)$. It follows from the equation

$$g^m - 1 = m(g-1) + \binom{m}{2}(g-1)^2 + \cdots + \binom{m}{m-1}(g-1)^{m-1}$$

that $g - 1 \in \Delta^n_{\boldsymbol{Q}}(G) + (g-1)^2 \boldsymbol{Q}G$. Hence $g - 1 \in \Delta^n_{\boldsymbol{Q}}(G)$ and $g \in D_{n,\boldsymbol{Q}}(G)$.

(ii) Let $g \in \gamma_i(G)$, $ip^j \geqslant n$, then $g^{p^j} \in M_{ip^j,p}(G) \subseteq M_{n,p}(G)$. Thus

$$L_{n,p}(G) \subseteq M_{n,p}(G).$$

The containment $M_{n,p}(G) \subseteq D_{n,\boldsymbol{Z}/p\boldsymbol{Z}}(G)$ follows by induction on n in view of Lemma 9.1(ii). □

We have proved the trivial part of the next theorem, a complete proof of which is given in Passman [P21].

THEOREM 9.4. *Let G be any group and n a natural number.*
(i) (Hall–Jennings) $D_{n,\boldsymbol{Q}}(G) = \sqrt{\gamma_n(G)}$.
(ii) (Jennings–Lazard–Zassenhaus) $D_{n,\boldsymbol{Z}/p\boldsymbol{Z}}(G) = M_{n,p}(G) = L_{n,p}(G)$.

It remains to evaluate the integral dimension subgroups $D_{n,\boldsymbol{Z}}(G)$. It was thought for a long time that $D_{n,\boldsymbol{Z}}(G) = \gamma_n(G)$. This is the *dimension subgroup conjecture*. Clearly, $D_1(G) = G$. Also, $D_2(G) = \gamma_2(G)$.

PROPOSITION 9.5. *For any group G, $D_{2,\boldsymbol{Z}}(G) = \gamma_2(G)$. Moreover, if $N \triangleleft G$ then*

$$g - 1 \in \Delta(G)\Delta(G,N) \quad \Rightarrow \quad g \in N'.$$

PROOF (*Sandling*). We first prove that $D_2(G) = \gamma_2(G) = G'$. By factoring with G' we may assume that G is Abelian. The additive group $\boldsymbol{Z}G^+$ of $\boldsymbol{Z}G$ is free Abelian with a basis $\{g\colon g \in G\}$. We define a map $\varphi : \boldsymbol{Z}G^+ \to G$ by $\varphi(g) = g$. Then $(x-1)(y-1) = xy - x - y + 1 \xrightarrow{\varphi} xy \cdot x^{-1} \cdot y^{-1} \cdot 1 = 1$. Therefore $\Delta^2(G) \subseteq \operatorname{Ker}\varphi$. Also, if $g \neq 1$ then $\varphi(g-1) = g \cdot 1^{-1} = g \neq 1$. Therefore, $g - 1 \notin \operatorname{Ker}\varphi$ and so $g - 1 \notin \Delta^2(G)$. We have proved that $D_2(G) = 1$, establishing the first assertion. For the second claim see [Se13, p. 29]. □

It is a result of G. Higman and D. Rees that $D_3(G) = \gamma_3(G)$ for all G. For a proof see Passi [Pas2] and [San4]. Since $\gamma_n(G) \subseteq D_n(G)$ for all n and all G let $\overline{G} = G/\gamma_n(G)$. To prove the dimension subgroup conjecture for any n and any class of groups closed under homomorphisms it suffices to prove for this class that $\gamma_n(G) = 1$ implies $D_n(G) = 1$. In particular, it suffices to consider nilpotent groups. We may also suppose that our group is finitely generated and therefore by a theorem of Gruenberg [Gr1] we can suppose that G is a finite p-group. We have

THEOREM 9.6 (Passi [Pas1]). *If G is a finite group of odd order then $D_4(G) = \gamma_4(G)$.*

In addition it is known (see [Mor]) that for any p-group G, $D_n(G) = \gamma_n(G)$ for $1 \leqslant n \leqslant p$. Also, for groups G of exponent p we have $D_n(G) = \gamma_n(G)$ for all n (see Cohn [Coh1] and Losey [Los1]).

The conjecture was refuted by Rips [Rip] in 1972 by constructing a 2-group G of order 2^{38}, nilpotency class 3 ($\gamma_4(G) = 1$) and with $D_4(G) \neq 1$. Inspired by Rips, Gupta [Gu3] constructed finite 2-groups $G_n = G$ for all $n \geqslant 4$ satisfying $D_n(G) \neq \gamma_n(G)$. However, Losey [Los4] proved that $D_4(G)/\gamma_4(G)$ is of exponent dividing 2 for all groups G. Sjögren [Sj] vastly extended this result by giving an exponent, independent of G, for all factors $D_n(G)/\gamma_n(G)$.

SJÖGREN'S THEOREM 9.7. *Let b_m be the ℓ.c.m. $\{1, 2, \ldots, m\}$, let $c_1 = c_2 = 1$ and $c_n = \prod_{i=1}^{n-2} b_i^{\binom{n-2}{i}}$ for $n \geqslant 3$. Then for any group G the factor group $D_n(G)/\gamma_n(G)$ has exponent dividing c_n.*

In view of Rips' counterexample, Sandling [San4] proposed the study of 'Lie dimension subgroups', a concept we explain below. For any group ring RG over a commutative unital ring R we define the Lie powers $RG^{[n]}$ and $RG^{(n)}$ inductively by

$RG^{[1]} = RG^{(1)} = RG$,

$RG^{[m+1]} = [RG^{[m]}, RG]$, the additive group generated by the Lie products,

$[x, y] = xy - yx$ with $x \in RG^{[m]}$ and $y \in RG$, and

$RG^{(m+1)} = [RG^{(m)}, RG]RG$, the associative ideal generated by the corresponding Lie products $[x, y]$ with $x \in RG^{(m)}$, $y \in RG$. Then the nth Lie dimension subgroup $D_{(n)}(G)$ is defined to be

$$D_{(n)}(G) = G \cap \bigl(1 + \bigl(\Delta_{\mathbf{Z}}(G)\bigr)^{(n)}\bigr).$$

The restricted nth Lie dimension subgroup $D_{[n]}(G)$ is given by

$$D_{[n]}(G) = G \cap \bigl(1 + \bigl(\Delta_{\mathbf{Z}}(G)\bigr)^{[n]}\bigr).$$

The Lie dimension subgroups over any ring are defined similarly. It follows from the identity

$$(x, y) - 1 = x^{-1}y^{-1}xy - 1 = x^{-1}y^{-1}\bigl((x-1)(y-1) - (y-1)(x-1)\bigr)$$

that $\gamma_n(G) \subseteq D_{(n)}(G)$. Moreover, it is nontrivial to prove that $\gamma_n(G) \subseteq D_{[n]}(G)$ (see Gupta and Levin [GL]). Thus we have

$$\gamma_n(G) \subseteq D_{[n]}(G) \subseteq D_{(n)}(G) \subseteq D_n(G).$$

Sandling [San4] investigated whether $\gamma_n(G) = D_{(n)}(G)$. This is the Lie dimension subgroup conjecture (problem). He proved

THEOREM 9.8. *Let G be any group. Then*
(i) $D_{(n)}(G) = \gamma_n(G)$ *for* $n \leqslant 6$,
(ii) $D_{(n)}(G) = \gamma_n(G)$ *if G is metabelian.*

It was proved by Hurley and Sehgal [HoS1] by means of a counterexample that $D_{(n)}(G) \neq \gamma_n(G)$ for $n \geqslant 9$ (the group depending on n). Subsequently, Gupta and Tahara [GT] proved that the Lie dimension subgroup conjecture indeed holds for $n = 7$ and 8. The restricted Lie dimension subgroup conjecture is also false [HoS1].

For Lie dimension subgroups over fields results analogous to the ordinary dimension subgroups hold. Parmenter, Passi and Sehgal [PPSe] have proved that over integral domains $\boldsymbol{R}$ the Lie dimension subgroups can be computed in terms of those over $\boldsymbol{Q}$, $\boldsymbol{Z}/p\boldsymbol{Z}$ and $\boldsymbol{Z}$. In particular, we have that over fields the Lie dimension subgroups depend only on the characteristic and the group. The result is

THEOREM 9.9 (Passi and Sehgal [PS2]).
(a) *For all* $n \geqslant 2$, $D_{(n),\boldsymbol{Q}}(G) = \sqrt{\gamma_n(G)} \cap \gamma_2(G)$.
(b) *For all* $n \geqslant 1$

$$D_{(n+1),\boldsymbol{Z}/p\boldsymbol{Z}}(G) = M_{(n+1),p} = \prod_{ip^j \geqslant n+p^j} \gamma_i(G)^{p^j}.$$

Related to the dimension subgroups are Fox subgroups. Let F be free and let N be a normal subgroup of F. Then we define $F(n, N) = F \cap (1 + (\Delta_{\boldsymbol{Z}} F)^n \Delta_{\boldsymbol{Z}}(N))$ to be the nth Fox subgroup of F relative to N. The determination of these groups is the Fox problem which was introduced by Fox [Fox] in 1953, in connection with his free *differential* calculus. It follows from Proposition 9.5 that $F(2, N) = N'$. The (integral) Fox problem was solved by Hurley [Hu] and Yunus [Yu]. See also Gupta [Gu2, p. 54]. The modular Fox subgroups, namely, when $\boldsymbol{Z}$ is replaced by $\boldsymbol{Z}/p\boldsymbol{Z}$, have been determined by Hurley and Sehgal [HoS2].

In closing we mention that Gupta has announced that for all groups G, $D_n(G)/\gamma_n(G)$ has a 2-power exponent, thus proving that for odd finite groups $D_n(G) = \gamma_n(G)$ and solving the dimension subgroup problem. A publication is eagerly awaited.

Added in proof: This will appear in J. Group Theory.

10. Identities

The theory of polynomial identity rings, called $\boldsymbol{PI}$ rings, has been extensively developed, see Rowen [Row1,Row2]. In fact, the concept was introduced by M. Dehn in 1922.

DEFINITION 10.1. Let $F\langle z_1, z_2, \ldots, z_n\rangle$ be the polynomial ring over the field F in the noncommuting variables $z_1, \ldots, z_n$. An F-algebra A is said to be a $\boldsymbol{PI}$ algebra or to satisfy a polynomial identity if there exists a nonzero polynomial

$$f(z_1, \ldots, z_n) \in F\langle z_1, \ldots, z_n\rangle \quad \text{with } f(a_1, \ldots, a_n) = 0 \text{ for all } a_i \in A.$$

We also write $A \in \boldsymbol{PI}$.

Thus any commutative algebra satisfies the identity $f(z_1, z_2) = z_1z_2 - z_2z_1$. Kaplansky [Kap1] initiated the study of $\boldsymbol{PI}$ group rings. Partial results on the classification of group algebras satisfying a polynomial identity were obtained by Amitsur [Am3] and Smith [Smm]. A complete classification was given by Isaacs and Passman [IP] for characteristic zero and Passman [P11] for characteristic p. A group G is said to be p-Abelian if G' is a finite p-group. Also G is 0-Abelian if G is Abelian.

THEOREM 10.2 (Isaacs–Passman, Passman). *A group algebra FG of characteristic $p \geqslant 0$ satisfies a $\boldsymbol{PI}$ if and only if G has a p-Abelian subgroup of finite index.*

Let $*$ be the natural involution on FG, $\gamma = \sum \gamma(g)g \to \gamma^* = \sum \gamma(g)g^{-1}$. Let us denote by

$$(FG)^+ = \{\gamma \in FG\colon\ \gamma^* = \gamma\} \quad \text{and} \quad (FG)^- = \{\gamma \in FG\colon\ \gamma^* = -\gamma\},$$

the sets of symmetric and skew symmetric elements respectively. We investigate whether certain identities on these and similar subsets control identities on the whole group ring.

10.1. *Lie identities*

Write $[x, y]$ for the Lie product $xy - yx$. We say that FG is Lie nilpotent if

$$\underbrace{[FG,\ FG, \ldots, FG]}_{n} = 0$$

for some n. Suppose for a moment, that G is finite and $p > 0$. Let FG be Lie nilpotent. Then for $x,\ y \in G$,

$$[x, \underbrace{y, \ldots, y}_{p^k}] = 0$$

for a fixed k with $p^k > n$. It follows that $[x, y^{p^k}] = 0$. Therefore, y^{p^k} is central in G for all $y \in G$. Thus by Schur's theorem [Se10, p. 39], the commutator group G' is a p-group. Moreover, since G/centre is a p-group, G is nilpotent. We have proved:

(1) *If G is finite and FG is Lie nilpotent with F of characteristic $p > 0$, then G is nilpotent and p-Abelian.*

Conversely,

(2) *Suppose F has characteristic $p > 0$. If G is a finite nilpotent p-Abelian group, then FG is Lie nilpotent.*

PROOF. We use induction on the order of G. Pick a central element z of order p in G. Let $\overline{G} = G/\langle z\rangle$. By induction

$$[F\overline{G}, \underbrace{F\overline{G}, \ldots, F\overline{G}}_{n}] = 0.$$

This implies that

$$[FG, \underbrace{FG, \ldots, FG}_{n}]$$

is contained in $\Delta(G, \langle z\rangle)$, the kernel of the natural projection $FG \to F\overline{G}$. We conclude that $[FG, FG, \ldots, FG] \subseteq (1-z)FG$. Thus

$$[FG, \underbrace{FG, \ldots, FG}_{2n}] \subseteq (1-z)^2 FG$$

and consequently

$$[FG, \underbrace{FG, \ldots, FG}_{pn}] \subseteq (1-z)^p FG = 0.$$

□

By applying Theorem 10.2 we obtain the

THEOREM 10.3 (Passi, Passman and Sehgal [PPS]). *FG is Lie nilpotent if and only if G is nilpotent and p-Abelian where $p \geqslant 0$ is the characteristic of F.*

Let us write $\delta^{[1]}(FG) = [FG, FG]$, $\delta^{[i+1]}(FG) = [\delta^{[i]}(FG), \delta^{[i]}(FG)]$. We say that FG is Lie solvable if $\delta^{[n]}(FG) = 0$ for some n.

THEOREM 10.4 (Passi, Passman and Sehgal [PPS]). *Necessary and sufficient conditions for Lie solvability of FG,* $\operatorname{char} F = p \geqslant 0$ *are*

(i) *G is p-Abelian when $p \neq 2$,*

(ii) *G has a 2-Abelian subgroup of index at most 2 when $p = 2$.*

Continuing in the same vein we say that FG is Lie n-Engel if we have

$$[x, \underbrace{y, \ldots, y}_{n}] = 0$$

for all $x, y \in FG$. A classification of Lie n-Engel group rings is given by

THEOREM 10.5 [Se10, p. 155]. *Let* char $F = p \geqslant 0$. *Then necessary and sufficient conditions for FG to be Lie n-Engel are*

(i) *G is nilpotent and contains a normal p-Abelian subgroup A with G/A a finite p-group if $p > 0$,*
(ii) *G is Abelian if $p = 0$.*

Let us now turn to symmetric (and skew symmetric) elements. We say that $(FG)^+$ (respectively $(FG)^-$) is Lie nilpotent if we have $[x_1, \ldots, x_n] = 0$ for all $x_i \in (FG)^+$ (respectively $(FG)^-$). We similarly define when $(FG)^+$ is Lie n-Engel, etc. In this regard we have the following result.

THEOREM 10.6 (Giambruno and Sehgal [GS3]). *Suppose* char $F \neq 2$ *and that G has no 2-elements. Then*

$$(FG)^+ \text{ or } (FG)^- \text{Lie nilpotent} \quad \Rightarrow \quad FG \text{ Lie nilpotent.}$$

A related result is

THEOREM 10.7 (Giambruno and Sehgal [GS3]). *Let A be an additive subgroup of FG. Suppose that $[A, FG, FG, \ldots, FG] = 0$. Then $[A, FG]\,FG$ is an (associative) nilpotent ideal.*

By taking $A = FG$ in the above we obtain:

$$\begin{aligned} FG \text{ Lie nilpotent} \quad &\Rightarrow \quad [FG, FG]FG = \Delta(G, G') \text{ nilpotent} \\ &\Rightarrow \quad G' \text{ is a finite } p\text{-group.} \end{aligned}$$

Similar to Theorem 10.6 we have

THEOREM 10.8 [Lee2]. *If $(FG)^+$ or $(FG)^-$ is Lie n-Engel for some n, G has no 2-elements and* char $F \neq 2$, *then FG is Lie m-Engel for some m.*

There is a general result due to Zalesskii and Smirnov [ZS] about certain rings with involution. Again, denote by R^- the subset of skew symmetric elements.

THEOREM 10.9 (Zalesskii and Smirnov [ZS]). *Suppose that $R = \langle R^-, 1\rangle$ and that* char $R \neq 2$. *Then*

$$R^- \text{ Lie nilpotent} \quad \Rightarrow \quad R \text{ Lie nilpotent.}$$

REMARK 10.10. In general $\langle (FG)^-, 1\rangle \neq FG$. For example, let $G = \langle a, b \colon b^2 = 1,\ a^b = a^{-1}\rangle$ be the infinite dihedral group. Then $(FG)^- = \{\sum c_i(a^i - a^{-i}) \mid c_i \in F\} \subset F\langle a\rangle$. Thus $\langle (FG)^-, 1\rangle \subseteq F\langle a\rangle \neq FG$. However, from the identity

$$2g^2 = 2 + \left(g^2 - g^{-2}\right) + \left(g - g^{-1}\right)^2$$

it follows that $g^2 \in \langle (FG)^-, 1\rangle$ for $g \in G$. Thus if G is a finite group of odd order and $\operatorname{char} F \neq 2$ then $\langle (FG)^-, 1\rangle = FG$.

In connection with Theorem 10.6, possibilities of 2-elements and/or $p = 2$ remain to be discussed. Let us see what may happen if we allow G to have 2-elements. Let

$$G = K_8 = \left\langle a, b \colon a^4 = 1 = b^4,\ a^b = a^{-1},\ ba = abz\right\rangle$$

be the quaternion group of order 8. Then

$$(FG)^+ = F\left(a + a^{-1}\right) \oplus F\left(b + b^{-1}\right) \oplus F\left(ab + (ab)^{-1}\right) \oplus F \oplus Fz$$

is clearly commutative. Moreover, for $p \neq 2$, G' is not a p-group. Thus FG is not Lie nilpotent if $\operatorname{char} F \neq 2$.

We have seen that $(FK_8)^+$ is Lie nilpotent. Further, if E is an elementary Abelian 2-group and $G = K_8 \times E$, then the calculation above gives us that $(FG)^+$ is Lie nilpotent. Moreover, if $G = K_8 \times E \times P$ where P is a finite p-group and $\operatorname{char} F = p$, it is easily seen that $(FG)^+$ is Lie nilpotent. In fact this is all that can happen as seen in the following result of G. Lee.

THEOREM 10.11 [Lee1]. *Suppose that $K_8 \nsubseteq G$ and* $\operatorname{char} F = p > 2$. *Then $(FG)^+$ is Lie nilpotent $\Leftrightarrow$ FG is Lie nilpotent.*

COROLLARY 10.12. *Suppose that* $\operatorname{char} F = 0$ *and $K_8 \nsubseteq G$. Then $(FG)^+$ is Lie nilpotent if and only is G is Abelian.*

THEOREM 10.13 [Lee1]. *Suppose that $K_8 \subseteq G$ and* $\operatorname{char} F = p > 2$. *Then $(FG)^+$ is Lie nilpotent if and only if $G = K_8 \times E \times P$ where $E^2 = 1$ and P is a finite p-group.*

THEOREM 10.14. *Suppose that* $\operatorname{char} F = 0$ *and $K_8 \subseteq G$. Then $(FG)^+$ is Lie nilpotent if and only if $G = K_8 \times E$ where $E^2 = 1$.*

Recall from Theorem 10.6 that in the absence of 2-elements and provided $p \neq 2$, $(FG)^-$ is Lie nilpotent if and only if FG is Lie nilpotent. This is no more true if we allow 2-elements as seen by the following example. Let $G = D_8 = \langle a, b \colon a^4 = 1 = b^2,\ a^b = a^{-1}\rangle$ be the dihedral group of order 8. Then $(FG)^- \subseteq F\langle a\rangle$ is commutative. However, FG is not Lie nilpotent if $p \neq 2$.

10.2. *Properties of the unit group*

In this section we classify groups G so that $\mathcal{U}(RG)$ has some well known property where R is the ring of rational integers $\boldsymbol{Z}$ or a field F.

DEFINITION 10.15. We say that the group $\mathcal{U} = \mathcal{U}(RG)$ satisfies a group identity if there exists a nontrivial word $w(x_1, \ldots, x_n)$ in the free group generated by $x_1, \ldots, x_n$ such that $w(u_1, \ldots, u_n) = 1$ for all $u_i \in \mathcal{U}$. We write $\mathcal{U} \in \boldsymbol{GI}$ or $\mathcal{U}$ satisfies a $\boldsymbol{GI}$.

EXAMPLES 10.16.
(1) If $\mathcal{U}$ is Abelian then it satisfies (x_1, x_2).
(2) If $\mathcal{U}$ is nilpotent then it satisfies $(x_1, \ldots, x_n)$ for some n.
(3) If $\mathcal{U}$ is metabelian then it satisfies $((x_1, x_2), (x_3, x_4))$.
(4) If $\mathcal{U}$ is torsion of bounded exponent e then it satisfies x^e.

At first, we shall restrict our attention to the case of $\boldsymbol{Z}G$ and $\boldsymbol{Q}G$.

Let G be finite and let $\mathcal{U} = \mathcal{U}(\boldsymbol{Z}G)$. Suppose $\mathcal{U} \in \boldsymbol{GI}$. We recall (7.5), that if u is a nontrivial bicyclic unit then $\langle u, u^* \rangle$ is a free group, contained in $\mathcal{U}$, and it cannot satisfy a $\boldsymbol{GI}$. Thus all bicyclic units are trivial. Consequently all subgroups of G are normal. So G is Abelian or Hamiltonian. In the latter case, $G = K_8 \times E \times O$ where K_8 is the quaternion group of order 8, $E^2 = 1$ and O is an odd order Abelian group. We also know [Se13, p. 21] that the group $\mathcal{U}(\boldsymbol{Z}(K_8 \times C_p))$ contains a free group if p is an odd prime. It follows that $O = 1$ and $G = K_8 \times E$. We have proved that

$$\mathcal{U} \in GI \quad \Rightarrow \quad G \text{ is Abelian or } G = K_8 \times E.$$

Conversely, if G is Abelian then, of course, $(\mathcal{U}, \mathcal{U}) = 1$. If $G = K_8 \times E$ then by Theorem 3.5 we have that $\mathcal{U} = K_8 \times E$ and $(\mathcal{U}, \mathcal{U}, \mathcal{U}) = 1$. We have proved

THEOREM 10.17. *Let G be a finite group and $\mathcal{U} = \mathcal{U}(\boldsymbol{Z}G)$. Then the following conditions are equivalent*:
(i) $\mathcal{U} \in \boldsymbol{GI}$.
(ii) *G is Abelian or a Hamiltonian* 2*-group.*
(iii) $(\mathcal{U}, \mathcal{U}, \mathcal{U}) = 1$.
(iv) *$\mathcal{U}$ is nilpotent.*
(v) *$\mathcal{U}$ is solvable.*

An easy consequence is

THEOREM 10.18. *Let G be a finite group and $\mathcal{U} = \mathcal{U}(\boldsymbol{Q}G)$. Then $\mathcal{U} \in \boldsymbol{GI}$ if and only if G is Abelian.*

Now, we look at properties of nilpotence and solvability for infinite groups. In the case of solvability we have to remember that for torsion-free groups, the units of $\boldsymbol{Z}G$ are conjectured to be trivial. In the absence of an answer to this conjecture it is improbable that one can classify G so that $\mathcal{U}(\boldsymbol{Z}G)$ is solvable. Considering this we do have a, more or less, satisfactory result:

THEOREM 10.19 [Se10, p. 203]. *Suppose that $\mathcal{U}(\boldsymbol{Z}G)$ is solvable. Then*

($*$) *The torsion elements, $T = T(G)$, of G form a group which is Abelian or a Hamiltonian 2-group with every subgroup of T normal in G.*

Conversely, if G is a solvable group satisfying ($*$) *and*

($**$) *$G/T(G)$ is nilpotent,*

then $\mathcal{U}(\boldsymbol{Z}G)$ is solvable.

PROOF. Suppose $\mathcal{U} = \mathcal{U}(\boldsymbol{Z}G)$ is solvable. Then the argument in the paragraph before (10.17), using the fact that $\langle u, u^*\rangle$ is free if u is a nontrivial bicyclic unit, gives ($*$). To prove the solvability of $\mathcal{U}$ under the assumptions ($*$) and ($**$) it is enough to show

($***$) $\mathcal{U}(\boldsymbol{Z}G) = \mathcal{U}(\boldsymbol{Z}T(G)) \cdot G$.

This is proved in [Se10, p. 203]. □

For rational group algebras we have

THEOREM 10.20. *Suppose that $\mathcal{U} = \mathcal{U}(\boldsymbol{Q}G)$ is solvable. Then*

($*$) *The torsion elements of G form an Abelian subgroup $T(G)$ with every subgroup normal in G.*

Conversely, if G is a solvable group satisfying ($*$) *and*

($**$) *$G/T(G)$ is nilpotent,*

then $\mathcal{U}$ is solvable.

THEOREM 10.21. *Let G be any group. Then $\mathcal{U} = \mathcal{U}(\boldsymbol{Q}G)$ is nilpotent if and only if G is nilpotent with all torsion elements central.*

The next theorem was proved for finite groups by Polcino Milies [Po3] and for arbitrary groups by Sehgal and Zassenhaus [SZ2].

THEOREM 10.22. *$\mathcal{U}(\boldsymbol{Z}G)$ is nilpotent if and only if G is nilpotent and the torsion subgroup T of G satisfies one of the following*:

(i) *T is central in G,*

(ii) *T is an Abelian 2-group and for $x \in G$, $t \in T$*

$$x^{-1}tx = t^{\delta(x)}, \quad \delta(x) = \pm 1,$$

(iii) *$T = E \times K_8$ where $E^2 = 1$ and K_8 is the quaternion group of order* 8. *Moreover, E is central in G and conjugation by $x \in G$ induces on K_8 one of the four inner automorphisms.*

Now we discuss when $\mathcal{U}(FG)$, the group of units of the group algebra FG with char $F = p$, is nilpotent or solvable. We begin with two easy but very useful observations.

LEMMA 10.23. *Let I be a two-sided nil ideal in a ring R. Then the natural epimorphism $R \to R/I$ induces an epimorphism of multiplicative groups $\mathcal{U}(R) \to \mathcal{U}(R/I)$.*

LEMMA 10.24. *Suppose that I is a two-sided ideal in a ring R and u, v are units in R which are $\equiv 1 \bmod I$. Then the multiplicative commutator $(u, v) \equiv 1 \bmod I^2$.*

For finite groups G, necessary and sufficient conditions for $\mathcal{U}(FG)$ to be solvable have been given by Bateman [Bat], Bovdi and Khripta [BK3] and Passman [P20]. We avoid the difficulties with characteristic 2 or 3 and give

LEMMA 10.25 (Bateman [Bat]). *If F is a field of positive characteristic $p \neq 2, 3$ and G is a finite group then $\mathcal{U}(FG)$ is solvable if and only if G' is a p-group.*

PROOF. Let us first suppose that $\mathcal{U}(FG)$ is solvable and prove that G' is a p-group. Without loss of generality we may assume that $F = \mathbf{Z}/p\mathbf{Z}$. We use induction on $|G|$. If G contains a normal p-subgroup $P \neq \{1\}$ then by Lemma 10.23, $\mathcal{U}(F(G/P))$ is solvable. By induction $(G/P)'$ is a p-group and G' is also a p-group. Thus we may assume that G has no nontrivial normal p-subgroup. Let J be the Jacobson radical of FG. Then

$$FG/J \stackrel{\psi}{=} \bigoplus_i (F_i)_{n_i},$$

a direct sum of $n_i \times n_i$ matrix rings over finite fields F_i. Lemma 10.23 implies that $\mathcal{U}(FG/J)$ is solvable. It follows that each $n_i = 1$ and

$$FG/J \stackrel{\psi}{=} \bigoplus_i F_i.$$

Since $G \cap (1 + J)$ is a normal p-subgroup we may conclude that $G \cap (1 + J) = \{1\}$. Therefore,

$$\lambda : G \to \bigoplus_i F_i, \quad \lambda(g) = \psi(\bar{g}) \in \bigoplus_i F_i$$

is a monomorphism. Thus G is Abelian.

Now, we prove the converse. Recall that $\Delta(G, G')$ is nilpotent because G' is a finite p-group, where the characteristic of F is p and $\Delta(G, G') = (FG)(\Delta G')$. Suppose $(\Delta G')^s = 0$. Then $(\Delta(G, G'))^s = 0$ as well.

Let $u, v \in KG$. Then $(u, v) \equiv 1 \bmod \Delta(G, G')$. By the last lemma, $((u, v), (w, t)) \equiv 1 \bmod (\Delta(G, G'))^2$. Repeating this we find that $\delta_{s+1}(\mathcal{U}(KG)) = \{1\}$ and consequently, $\mathcal{U}(FG)$ is solvable. □

For nilpotence we have the next result of Khripta [Kh2].

THEOREM 10.26. *Let G be a group having an element of order p. If F is a field of characteristic p then $\mathcal{U}(FG)$ is nilpotent $\Leftrightarrow$ G is nilpotent with G' a finite p-group.*

To complete the classification of groups G, for which the group $\mathcal{U}(FG)$ is nilpotent when F has characteristic $p > 0$, it remains to consider the case when G has no p-element. This was done by Fisher, Parmenter and Sehgal [FPS] and also by Khripta as announced by Zalesskii and Mikhalev [ZM]. We state the result below. We refer the reader for the proof to [Se10, p. 182].

THEOREM 10.27. *Suppose FG is a group algebra over a field F of characteristic $p > 0$. Suppose G has no element of order p. Then $\mathcal{U}(FG)$ is nilpotent if and only if G is nilpotent and one of the following holds*:

(a) *The torsion elements of G form a central group.*

(b) $|F| = 2^\beta - 1 = p$, *a Mersenne prime*; *the torsion elements form an Abelian group T of exponent $(p^2 - 1)$ and for all $x \in G$, $t \in T$ we have $x^{-1}tx = t$ or t^p.*

10.3. *Hartley's conjecture*

In order to connect the multiplicative structure to the additive structure, Brian Hartley made the

CONJECTURE 10.28. *Let G be a torsion group and F be an infinite field. Let $\mathcal{U} = \mathcal{U}(FG)$. Then*

$$\mathcal{U} \in \boldsymbol{GI} \quad \Rightarrow \quad FG \in \boldsymbol{PI}.$$

The first results on this conjecture were obtained by Warhurst [Warh] who studied some special cases in 1981. Also, Menal [Me] suggested a possible solution for some finite p-groups. Gonçalves and Mandel [GM] classified group algebras FG of torsion groups over infinite fields whose group of units satisfies a semigroup identity ($w_1 = w_2$ where w_1 and w_2 are distinct semigroup words) proving in this way Hartley's conjecture for semigroup identities. Dokuchaev and Gonçalves [DG] dealt with this question for integral group rings. Giambruno, Jespers and Valenti [GJV] settled in the positive the conjecture when G has no p-elements. By using the construction suggested by Menal the authors proved

THEOREM 10.29 (Giambruno, Sehgal and Valenti [GSV3]). *If G is a torsion group and F is infinite, then $\mathcal{U} \in \boldsymbol{GI} \;\Rightarrow\; FG \in \boldsymbol{PI}$.*

However, more is true.

THEOREM 10.30 (Passman [P25]). *Let G be torsion and let F be infinite. Then*

(1) *if* $\operatorname{char} F = 0$, *$\mathcal{U}(FG)$ satisfies a group identity if and only if G is Abelian,*

(2) *if* $\operatorname{char} F = p > 0$, *$\mathcal{U} \in \boldsymbol{GI} \;\Leftrightarrow\; FG \in \boldsymbol{PI}$ and G' is of bounded exponent p^m.*

It turns out that Hartley's conjecture is also true for finite fields (see Liu [Liu]). Moreover, the last theorem can be extended as well.

THEOREM 10.31 (Liu and Passman [LiP]). *Let F be a field of characteristic $p > 0$ and G a torsion group. If G' is a p-group, then the characterization of the last theorem holds. If G' is not a p-group, then $\mathcal{U} \in \boldsymbol{GI} \Leftrightarrow FG \in \boldsymbol{PI}$, G has bounded period and F is finite.*

This result was further extended to more general groups by Giambruno, Sehgal and Valenti [GSV5].

THEOREM 10.32. *Suppose that F is infinite or G has an element of infinite order. Let the characteristic of F be $p \geqslant 0$. Let T be the set of all torsion elements of G. Let P be the set of p-elements. If $p = 0$, we set $P = 1$. We have the following*:

(a) *If $\mathcal{U}(FG)$ satisfies a group identity then P is a subgroup.*

(b) *If P is of unbounded exponent and $\mathcal{U}(FG)$ satisfies a group identity then*

  (i) *G contains a p-Abelian subgroup of finite index.*

  (ii) *G' is of bounded p-power exponent.*

Conversely, if P is a subgroup and G satisfies (i) *and* (ii) *then $\mathcal{U}(FG)$ satisfies a group identity.*

(c) *If P is of bounded exponent and $\mathcal{U}(FG)$ satisfies a group identity then*:

  (0) *P is finite or G has a p-Abelian subgroup of finite index.*

  (1) *$T(G/P)$ is an Abelian p'-subgroup and so T is a group.*

  (2) *Every idempotent of $F(G/P)$ is central.*

Conversely, if P is a subgroup, G satisfies (0), (1), (2) *and G/T is nilpotent then $\mathcal{U}(FG)$ satisfies a group identity.*

We can ask if the identities at the unit level are also controlled by symmetric elements. We use the notation: $\mathcal{U}^+ = \{u \in \mathcal{U}(FG)\colon\ u^* = u\}$. We say $\mathcal{U}^+$ satisfies a group identity, $\mathcal{U}^+ \in \boldsymbol{GI}$, if there is a nontrivial word $w(x_1, \ldots, x_m)$ such that $w(u_1, \ldots, u_m) = 1$ for all $u_i \in \mathcal{U}^+$. We should keep in mind the example, of the last section, of the quaternion group where elements of $\mathcal{U}^+(FK_8)$ commute. We have an analogue of Hartley's conjecture:

THEOREM 10.33 (Giambruno, Sehgal and Valenti [GSV4]). *Let F be an infinite field of characteristic $\neq 2$ and let G be a torsion group. Then*

$$\mathcal{U}^+ \in \boldsymbol{GI} \quad \Rightarrow \quad FG \in \boldsymbol{PI}.$$

In fact, it is possible to say more.

THEOREM 10.34 (Giambruno, Sehgal and Valenti [GSV4]). *Let F be an infinite field and G a torsion group. If* $\operatorname{char} F = 0$, *$\mathcal{U}^+ \in \boldsymbol{GI}$ if and only if G is Abelian or a Hamiltonian 2-group. If* $\operatorname{char} F = p > 2$, *then $\mathcal{U}^+ \in \boldsymbol{GI}$ if and only if $FG \in \boldsymbol{PI}$ and either $K_8 \nsubseteq G$ and G' is of bounded exponent p^k or $K_8 \subseteq G$ and*

(1) *the p-elements of G form a subgroup P and G/P is a Hamiltonian 2-group,*

(2) *G is of bounded exponent $4p^s$.*

For groups for which $\mathcal{U}^+$ is nilpotent we have the two results of Lee [Lee3], dependent on whether or not K_8 is contained in G.

THEOREM 10.35. *Let F be a field of characteristic $p \neq 2$ and G a torsion group not containing K_8. Then $\mathcal{U}^+(FG)$ is nilpotent if and only if $\mathcal{U}(FG)$ is nilpotent.*

THEOREM 10.36. *Let F be a field of characteristic $p \neq 2$ and G a torsion group containing K_8. Then $\mathcal{U}^+(FG)$ is nilpotent if and only if either*
(1) *$p > 2$ and $G \simeq K_8 \times E \times P$ where $E^2 = 1$ and P is a finite p-group, or*
(2) *$p = 0$ and $G \simeq K_8 \times E$ with $E^2 = 1$.*

As consequence we have

COROLLARY 10.37. *Let F be a field of characteristic $p \neq 2$ and G a torsion group. Then $\mathcal{U}^+(FG)$ is nilpotent if and only if $(FG)^+$ is Lie nilpotent.*

11. Zero divisors

We have already seen that if g is an element of finite order n in a group G then in the group ring KG we have the identity $(g-1)(1+g+\cdots+g^{n-1}) = 0$, namely, $(g-1)$ is a zero divisor. We have the

ZERO DIVISOR CONJECTURE 11.1. *If G is a torsion free group and K is a field then KG has no zero divisors.*

We give a brief summary of the fantastic results obtained on this problem. First recall that a ring R is said to be semiprime if R has no nilpotent ideals whereas it is said to be prime if a product of nonzero ideals is always nonzero. It is a classical result, going back to P. Jordan, that if K is a field of characteristic zero then KG is semiprime. We have the following classification which is an analogue of Maschke's theorem for group rings of finite groups.

THEOREM 11.2 (Passman [P1]). *Let K be a field of characteristic $p > 0$. Then KG is semiprime if and only if G has no finite normal subgroup of order divisible by p.*

For prime group rings we have

THEOREM 11.3 (Connell [Con]). *The group ring KG of a group G over a field K is prime if and only if G has no nontrivial finite normal subgroups.*

Crucial in the proof of these theorems is the Passman map $\pi : KG \to K\phi$, the projection of KG to the group ring of the FC-subgroup ϕ, given by

$$\pi\left(\sum_{g\in G} a_g g\right) = \sum_{g\in\phi} a_g g,$$

which has the following pleasant properties.

THEOREM 11.4 (Passman [P1]). *Let A, B be ideals in KG. Then*
(i) *$\pi(A)$ is an ideal in $K\phi$.*
(ii) *$A \neq 0$ if and only if $\pi(A) \neq 0$.*
(iii) *$AB = 0$ implies that $\pi(A)\pi(B) = 0$.*

PROOF. See [Pas2, p. 90]. □

We discuss when the augmentation ideal $\Delta_K(G)$ is nilpotent or residually nilpotent (i.e., $\Delta^\omega = \bigcap_n \Delta^n = 0$). We have the well known and easy to prove result of Coleman, Connell and Losey.

THEOREM 11.5. *Let K be a commutative ring with identity. Then the augmentation ideal $\Delta(G)$ of the group ring KG of an arbitrary group G is nilpotent if and only if G is a finite p-group and K is of p-power characteristic.*

PROOF. See [Se10, p. 27]. □

The classification of groups with residually nilpotent augmentation ideals is deep and difficult. For a comprehensive treatment see [Pas2]. The results are as follows. The integral case was handled by Lichtman [Lic]. To state the result we need a

DEFINITION. A group G is said to be discriminated by a class $\mathcal{C}$ of groups if for every finite subset $g_1, \dots, g_n$ of distinct elements of G, there exists a group $H \in \mathcal{C}$ and a homomorphism $\varphi : G \to H$ such that $\varphi(g_i) \neq \varphi(g_j)$ for all $i \neq j$.

Then we have

THEOREM 11.6 (Lichtman). *Let G be a group. Then $\Delta_{\mathbf{Z}}(G)$ is residually nilpotent if and only if one of the following holds*:
(i) *G is residually 'torsion-free nilpotent'.*
(ii) *G is discriminated by the class of nilpotent p_i-groups, $i \in I$, of bounded exponents, where $\{p_i\colon\ i \in I\}$ is some set of primes.*

PROOF. See [Pas2, p. 92] or [Lic]. □

This result was generalized to the residual nilpotence of the Lie-powers of $\Delta_{\mathbf{Z}}(G)$ (namely $\bigcap_n \Delta^{(n)} = 0$) by Musson and Weiss [MW]. A related result of Parmenter and Passi giving necessary and sufficient conditions for the existence of an element $i \in \mathbb{Z}G$ such that $(\bigcap_n \Delta^n)(1 - i) = 0$ can be found in [Pas2, p. 103].

THEOREM 11.7 (Jennings [Jen3], Hartley [Har1]). *Let G be a group and K a field. Then $\Delta_K(G)$ is residually nilpotent if and only if one of the following holds*:
(i) *G is residually 'torsion-free nilpotent' and* char $K = 0$.
(ii) *G is residually 'nilpotent p-group of bounded exponent' and* char $K = p > 0$.

A consequence of Connell's theorem is

THEOREM 11.8 (Connell [Con]). *A group ring KG of a group G over a field K is Artinian if and only if G is finite.*

It has not been possible to classify Noetherian group rings. We have a partial result due to P. Hall. Remember that for group rings right Noetherian is the same as left Noetherian.

THEOREM 11.9. *Suppose that G has a series*

$$G = G_0 \rhd G_1 \rhd \cdots \rhd G_k = 1$$

with G_i/G_{i+1} either finite or infinite-cyclic for $0 \leqslant i \leqslant k-1$. Then KG is Noetherian.

The inductive step in the proof requires an analogue of the Hilbert basis theorem, namely if R is a Noetherian ring then the twisted group ring $R^t(G)$ is also Noetherian.

It is elementary and well known that a commutative integral domain has a field of fractions. This can also be done for noncommutative rings R satisfying

ORE'S CONDITION. Given $a, b \in R$ with b a nonzero divisor then there exist $a_1, b_1 \in R$ with b_1 a nonzero divisor such that $ab_1 = ba_1$.

DEFINITION. A ring $Q(R) \supset R$ is said to be a right quotient ring for R if:

(1) every element of R which is not a zero divisor is invertible in $Q(R)$,

(2) every $x \in Q(R)$ is of the form $x = ab^{-1}$, $a, b \in R$, where b is not a zero divisor.

If $Q(R)$ is a right quotient ring of R we say that R is a right order in $Q(R)$ and $Q(R)$ is the classical ring of right quotients.

THEOREM 11.10. *A necessary and sufficient condition that R has a right quotient ring is that R satisfies the* (*right*) *Ore's condition.*

Suppose KG is a Noetherian domain. Then by Goldie's Theorem [Gol1,Gol2] it is an order in a full matrix ring $M_n(D)$ over a division ring. Moreover, by a theorem of Faith and Utumi [FU], KG contains $M_n(C)$ when C is an order in D. Thus

$$M_n(C) \subseteq KG \subseteq M_n(D).$$

Since KG is a domain it follows that $n = 1$. We have proved

PROPOSITION 11.11. *If a Noetherian group ring KG is a domain then its classical ring of quotients, $Q(KG)$, is a division ring.*

We already saw in Section 3 that if G is ordered (or is more generally a unique product group) then KG has no zero divisors. However, it was proved by Promislow [Pr] that not all torsion free groups are unique product groups. The following example was given by Rips and Segev [RSe].

EXAMPLE. Let G be the group

$$G=\langle x, y\colon x^{-1}y^2x=y^{-2}, y^{-1}x^2y=x^{-2}\rangle.$$

Then G is a torsion free Abelian-by-finite group which does not have the unique product property.

In the 1970s J. Lewin [Lew] used Cohn's theory of n-firs to prove

THEOREM 11.12. *Let $G=A\underset{N}{*}B$ be a free product with amalgamation and K a field. Suppose*
(i) $N\lhd A$, $N\lhd B$;
(ii) *KA and KB have no zero divisors*;
(iii) *KN satisfies the Ore condition.*
Then KG has no zero divisors.

By using this theorem E. Formanek [For3] proved that if G is a torsion free supersolvable group and K a field then KG has no zero divisors. Recall that G is said to be supersolvable if it has a finite normal series

$$1=G_0\leqslant G_1\leqslant\cdots\leqslant G_n=G$$

with each G_{i+1}/G_i cyclic. He observed the next lemma and applied the Lewin result.

LEMMA 11.13. *If G is infinite supersolvable then G has a quotient G/H which is infinite cyclic or infinite dihedral.*

We give an elementary proof due to Zalesskii (see [Bov2]).

THEOREM 11.14 (Formanek [For3]). *If G is torsion free supersolvable and K a field then KG has no zero divisors.*

PROOF. Let H be the subgroup of the last lemma. Then by induction on the Hirsch number we may assume that KH has no zero divisors.

If G/H is infinite cyclic then KG, being the crossed product of the ring KH without zero divisors and an infinite cyclic group, has no zero divisors. Therefore we may assume that

$$G/H=\langle\bar{a},\bar{b}\colon \bar{a}^2=1,\ \bar{a},\bar{b}a^{-}=\bar{b}^{-1}\rangle,$$

where $\bar{a}=aH$, $\bar{b}=bH$. Then for $G_1=\langle a, H\rangle$ we have that KG_1 has no zero divisors.

By Proposition 11.8, the group algebra KH has a division ring of right quotients $Q(KH)$ and the crossed product $(KH)(\langle\bar{a}\rangle,\rho,\sigma)\simeq KG_1$, induces the crossed product $Q(KH)(\langle\bar{a}\rangle,\rho,\sigma)$. The latter ring is a two-dimensional vector space over $Q(KH)$; hence it is Artinian. Moreover, it has no zero divisors and so it coincides with the division ring of quotients of KG_1.

Represent the elements of KG in the form

$$y = \sum_i \lambda_i b^i = \sum_i (k_1 + k_i' a) b^i, \quad \lambda_i \in kG,\ k_i, k_i' \in KH.$$

Let $yu = 0$. Multiplying y, if necessary, on the right by b^s for suitable s, we can assume that y has the form $\sum_{i=0}^m \lambda_i b^i$ and $0 \neq \lambda_m \in KG_1$. We can also assume that $\lambda_m \in KH$: in the division ring of quotients $Q(KH) + aQ(KH)$ of KG_1 there is an inverse $x^{-1}(x_1 + x_1' a)$ for λ_m and $(x_1 + x_1' a)\lambda_m \in KH$. Then instead of y we could take $(x_1 + x_1' a) y$.

As shown before the division ring of quotients for the group algebra $KG_2 = K\langle b, H\rangle$ exists. The crossed product of the cyclic group $\langle aG_2\rangle$ and the division ring $Q(KG_2)$ induced by the crossed product $KG_2(\langle aG_2\rangle, \rho, \sigma) = KG$ is a two-dimensional space over a division ring and is Artinian, and in such a ring every left zero divisor is a right zero divisor. Therefore, for a suitable $z \in KG$, $zy = 0$ and z has an expression of the form $\sum_{i=-n}^n \mu_i b^i$, where μ_n and μ_{-n} are not simultaneously zero. Moreover, it is convenient to assume that z has been chosen so that n is minimal.

If $\mu_j = e_j + e_j' a$ $(e_j, e_j' \in KH)$ and $h_i = a^{-1} b^i a b^i \in H$ then

$$\begin{aligned}\mu_j b^j \cdot \lambda_i b^i = {} & \bigl[e_j b^j k_i b^{-j} + e_j'\bigl(a b^j k_i b^{-j} a^{-1}\bigr)a\bigr] b^{i+j} \\ & + \bigl[e_j'\bigl(a b^j k_i' b^{-j} a^{-1}\bigr)a^2 h_j + e_j b^j k_i' b^{-j} a h_j\bigr] b^{i-j}. \qquad (*)\end{aligned}$$

Let $zy = \sum_i \nu_i b^i$. Then from $(*)$ it follows that $\nu_i = 0$ for $i > m + n$ and $\nu_{n+m} b^{n+m} = \mu_n b^n \lambda_m b^m$. Therefore the equation $zy = 0$ implies $\mu_n = 0$, and by our choice of z, $\mu_{-n} \neq 0$.

As shown above, for suitable $z_1 \in KG_1$ we have $z_1 \mu_{-n} \in KH$ and choosing instead of z the element $z_1 z$ we can again assume that $\mu_{-n} \in KH$. If $\bar{z} = bz = \sum_{i=-n}^n \bar{\mu}_i b^i$ then $\bar{\mu}_{-n} = b e_{-n+1}' a b^{-n+1}$. Therefore $\tilde{\mu}_{-n} \in KH$ for $\tilde{z} = abz = \sum_{i=-n}^n \tilde{\mu}_i b^i$ and $\alpha \mu_{-n} + \beta \tilde{\mu}_{-n} = 0$ for some $\alpha, \beta \in KH$ because of the existence of common left multiples for μ_{-n} and $\tilde{\mu}_{-n}$. Then

$$z_2 = \alpha z + \beta \tilde{z} = \sum_{i=-n+1}^n \rho_i b^i \neq 0 \quad \text{and} \quad z_2 y = 0.$$

By what has been shown, $\rho_n = 0$, but this contradicts the minimality of n. □

Extending this work Farkas and Snider [FS] and Cliff [Cl] handled torsion free polycyclic-by-finite groups in characteristic 0 and p respectively by a clever application of a K-theory argument. The characteristic zero case of torsion free Abelian-by-finite groups had already been solved by Brown [Bro1].

THEOREM 11.15. *Let K be a field and let G be a torsion free polycyclic-by-finite group. Then KG is a domain whose ring of quotients is a division ring.*

For the next result see Brown [Bro2].

THEOREM 11.16. *Let R be a ring and G a solvable by finite group. Then RG has a right Artinian right quotient ring if and only if R has a right Artinian quotient ring and the finite subgroups of G have bounded order.*

The latest and most far reaching result on the zero divisor problem is based on a marvelous induction theorem of John Moody [Moo1,Moo2]. For a Noetherian ring S, let $G_0(S)$ denote the Grothendieck group associated with the category of all finitely generated right S-modules. If R is right Noetherian and G is polycyclic-by-finite then, a crossed product $R * G$ is also right Noetherian. Moreover, if $H \subseteq G$ then the functor

$$M \to M \underset{R*H}{\otimes} R * G$$

preserves exact sequences and hence induces a natural group homomorphism $G_0(R * H) \to G_0(R * G)$ which is called the induction map. We have

MOODY'S INDUCTION THEOREM 11.17. *Let R be a right Noetherian ring and let G be a polycyclic-by-finite group. Let $\mathcal{F}$ be the set of finite subgroups of G. Then the natural induction map*

$$\bigoplus_{H \in \mathcal{F}} G_0(R * H) \to G_0(R * G)$$

is surjective.

PROOF. See [Moo1,Moo2] and Passman [P24]. □

This has the following important consequence as proved by Kropholler, Linnell and Moody [KLM]. They prove the results for a class of groups which contains solvable by finite groups.

THEOREM 11.18. *Let K be a division ring and G a solvable-by-finite group. Suppose that the FC-subgroup ϕ of G is torsion free and that the finite subgroups of G have bounded order. Let ℓ denote the ℓ.c.m. of the orders of the finite subgroups. Then KG has a right quotient ring which is an $\ell \times \ell$ matrix ring over a division ring.*

A special case of the last theorem is

THEOREM 11.19. *Let K be a division ring and G a torsion free solvable-by-finite group. Then KG is a domain with a right quotient ring which is a division ring.*

References

[Ak1] Akasaki, T., *Idempotent ideals in integral group rings*, J. Algebra **23** (1972), 343–346.
[Ak2] Akasaki, T., *Idempotent ideals in integral group rings, II*, Arch. Math. **24** (1973), 126–128.

[Al] Aleev, R.Z., *Higman's central unit theory, units of integral group rings of finite cyclic groups and Fibonacci numbers*, Internat. J. Algebra Comput. **4** (1994), 309–358.

[AH1] Allen, P.J. and Hobby, C., *A characterization of units of* $\mathbf{Z}[A_4]$, J. Algebra **66** (1980), 534–543.

[AH2] Allen, P.J. and Hobby, C., *A note on the unit group of* $\mathbf{Z}S_3$, Proc. Amer. Math. Soc. **99** (1987), 9–14.

[AH3] Allen, P.J. and Hobby, C., *Units in integral group rings of some metacyclic groups*, Canad. Math. Bull. **30** (1987), 231–240.

[AH4] Allen, P.J. and Hobby, C., *A characterization of units in* $\mathbf{Z}S_4$, Comm. Algebra **16** (1988), 1479–1505.

[AH5] Allen, P.J. and Hobby, C., *Elements of finite order in* $V(\mathbf{Z}A_4)$, Pacific J. Math. **138** (1989), 1–8.

[Am1] Amitsur, S.A., *The identities of* PI*-rings*, Proc. Amer. Math. Soc. **4** (1953), 27–34.

[Am2] Amitsur, S.A., *Radicals of polynomial rings*, Canad. J. Math. **8** (1956), 355–361.

[Am3] Amitsur, S.A., *Groups with representation of bounded degree II*, Illinois J. Math. **5** (1961), 198–205.

[AB] Artamonov, V.A. and Bovdi, A.A., *Integral group rings: Groups of units and classical K-theory*, J. Soviet Math. **57** (1991), 2931–2958.

[AP] Arora, S.R. and Passi, I.B.S., *Central height of the unit group of an integral group ring*, Comm. Algebra **21** (1993), 3673–3683.

[AHP] Arora, S.R., Hales, A.W. and Passi, I.B.S., *Jordan decomposition and hypercentral units in integral group rings*, Comm. Algebra **21** (1993), 25–35.

[AA] Ayoub, R.G. and Ayoub, C., *On the group ring of a finite Abelian group*, Bull. Austral. Math. Soc. (1969), 245–261.

[Ba] Bagiński, C., *The isomorphism question for modular group algebras of metacyclic* p*-groups*, Proc. Amer. Math. Soc. **104** (1988), 39–42.

[BaC] Bagiński, C. and Caranti, A., *The modular group algebras of* p*-groups of maximal class*, Canad. J. Math. **40** (1988), 1422–1435.

[Bas1] Bass, H., *The Dirichlet unit theorem, induced characters, and Whitehead groups of finite groups*, Topology **4** (1965), 391–410.

[Bas2] Bass, H., *Algebraic* K*-Theory*, Benjamin, New York (1968).

[Bas3] Bass, H., *Euler characteristics and characters of discrete groups*, Invent. Math. **35** (1976), 155–196.

[BMS] Bass, H., Milnor, J. and Serre, J.-P., *Solution of the congruence subgroup problem for* SL_n $(n \geqslant 3)$ *and* Sp_{2n} $(n \geqslant 2)$, Publ. Math. I.H.E.S. **33** (1967), 59–137.

[Bat] Bateman, J.M., *On the solvability of unit groups of group algebras*, Trans. Amer. Math. Soc. **157** (1971), 73–86.

[BC] Bateman, J.M. and Coleman, D.B., *Group algebras with nilpotent unit groups*, Proc. Amer. Math. Soc. **19** (1968), 448–449.

[BD] Bergman, G.M. and Dicks, W., *On universal derivations*, J. Algebra **36** (1975), 193–211.

[Be1] Berman, S.D., *On certain properties of the integral group rings*, Dokl. Akad. Nauk SSSR (N.S.) **91** (1953), 7–9.

[Be2] Berman, S.D., *On a necessary condition for the isomorphism of integral group rings*, Dopovidi Akad. Nauk USSR (1953), 313–316.

[Be3] Berman, S.D., *On the equation* $x^n = 1$ *in an integral group ring*, Ukrainsk. Mat. Zh. **7** (1955), 253–261.

[Be4] Berman, S.D., *Group algebras of countable Abelian* p*-groups*, Soviet Math. Dokl. **8** (1967), 871–873.

[BM] Berman, S.D. and Mollov, T.Zh., *On group rings of Abelian* p*-groups of any cardinality*, Matem. Zametki **6** (1969), 381–392.

[Bh] Bhandari, A.K., *Some remarks on the unit groups of integral group rings*, Arch. Math. **44** (1985), 319–322.

[BL1] Bhandari, A.K. and Luthar, I.S., *Torsion units of integral group rings of metacyclic groups*, J. Number Theory **17** (1983), 270–283.

[BL2] Bhandari, A.K. and Luthar, I.S., *Conjugacy classes of torsion units of the integral group ring of* D_p, Comm. Algebra **11** (1983), 1607–1627.

[BP] Bhandari, A.K. and Passi, I.B.S., *Residually Lie nilpotent group rings*, Arch. Math. **58** (1992), 1–6.

[BR] Bhandari, A.K. and Ritter, J., *Large subgroups in the unit groups of arithmetic orders*, J. Algebra **178** (1995), 512–529.

[BW] Bhandari, A.K. and Weiss, A., *Residual solvability of local group rings*, Comm. Algebra **17** (1989), 2635–2662.

[BSe] Bhandari, A.K. and Sehgal, S.K., *An induction theorem for units of p-adic group rings*, Canad. Math. Bull. **34** (1991), 31–35.

[BJ] Bhattacharya, P.B. and Jain, S.K., *A note on the adjoint group of a ring*, Arch. Math. **21** (1970), 366–368.

[BRT] Billig, Y., Riley, D. and Tasic, V., *Nonmatrix varieties and nil-generated algebras whose units satisfy a group identity*, J. Algebra **190** (1997), 241–252.

[Bl1] Bleher, F.M., *Zassenhaus Vermutung end einfache Gruppen*, Diplomarbeit Universität Stuttgart (1993).

[Bl2] Bleher, F.M., *Automoprhismen von Gruppenringen und Blocktheorie*, Dissertation, Stuttgart (1995).

[Bl3] Bleher, F.M., *Tensor products and a conjecture of Zassenhaus*, Arch. Math. **64** (1995), 289–298.

[Bl4] Bleher, F.M., *Finite groups of Lie type of small rank*, Pacific J. Math. **187** (1999), 215–239.

[BGK] Bleher, F.M., Geck, M. and Kimmerle, W., *Automorphisms of generic Iwahori–Hecke algebras and integral group rings of finite Coxeter groups*, J. Algebra **197** (1997), 615–655.

[BHK] Bleher, F.M., Hiss, G. and Kimmerle, W., *Autoequivalences of blocks and a conjecture of Zassenhaus*, J. Pure Appl. Algebra **103** (1995), 23–43.

[BK] Bleher, F.M. and Kimmerle, W., *On the structure of integral group rings of sporadic groups*, LMS J. Comp. Math. **3** (2000), 274–306.

[BKRW] Bleher, F.M., Kimmerle, W., Roggenkamp, K. and Wursthorn, M., *Computational aspects of the isomorphism problem*, Algorithmic Algebra and Number Theory, Heidelberg (1997), 313–329.

[BH] Borel, E. and Harish-Chandra, *Arithmetic subgroups of algebraic groups*, Ann. Math. **75** (1962), 485–535.

[Bou] Bourbaki, N., *Eléments d'Histoire des Mathématiques*, Hermann, Paris (1969).

[Bov1] Bovdi, A.A., *Group rings of torsion free groups*, Sibirsk. Mat. Zh. **1** (1960), 555–558 (in Russian).

[Bov2] Bovdi, A.A., *Group Rings*, Uzhgorod (1974).

[BK1] Bovdi, A.A. and Khripta, I.I., *Normal subgroups of a multiplicative group of a ring*, Mat. Sb. **87** (1972), 338–350; Math. USSR Sb. **16** (1972), 349–362.

[BK2] Bovdi, A.A. and Khripta, I.I., *Finite dimensional group algebras having solvable unit groups*, Trans. Science Conference Uzhgorod State University (1974), 227–233.

[BK3] Bovdi, A.A. and Khripta, I.I., *Group algebras of periodic groups with solvable multiplicative group*, Mat. Zametki **22** (1977), 421–432.

[BK4] Bovdi, A.A. and Khripta, I.I., *Group algebras with polycyclic multiplicative group*, Ukrainian Math. J. **38** (1986), 373–375.

[BMi1] Bovdi, A.A. and Mihovski, S.V., *Idempotents in crossed products*, Dokl. Akad. Nauk SSSR **195**; Soviet Math. Dokl. **11** (1970), 1439–1442.

[BMS] Bovdi, A., Marciniak, Z. and Sehgal, S.K., *Torsion units in infinite group rings*, J. Number Theory **47** (1994), 284–299.

[BKS] Bovdi, V., Kovács, L.G. and Sehgal, S.K., *Symmetric units in modular group algebras*, Comm. Algebra **24** (1996), 803–808.

[Br1] Brauer, R., *Über Systeme hypercomplexer Zahlen*, Math. Z. **30** (1929), 79–107.

[Br2] Brauer, R., *Representations of finite groups*, Lectures in Modern Mathematics, Vol. I, Wiley, New York (1963), 133–175.

[BN] Brauer, R. and Noether, E., *Über minimale Zerfällungskörper irreduzibler Darstellungen*, Sitz. Preuss. Akad. Wiss. Berlin (1927), 221–228.

[Bro1] Brown, K.A., *On zero divisors in group rings*, Bull. London Math. Soc. **8** (1976), 251–256.

[Bro2] Brown, K.A., *Artinian quotient rings of group rings*, J. Algebra **49** (1977), 63–80.

[Bro3] Brown, K.A., *The structure of modules over polycyclic groups*, Math. Proc. Cambridge Philos. Soc. **89** (1981), 257–283.

[Bu] Burns, R.G., *Central idempotents in group rings*, Canad. Math. Bull. **13** (1970), 527–528.

[BH] Burns, R.G. and Hale, V.W.D., *A note on group rings of certain torsion-free groups*, Canad. Math. Bull. **15** (1972), 441–445.

[Car] Carlson, J.F., *Periodic modules over modular group algebras*, J. London Math. Soc. **15** (1977), 431–436.

[Cau] Cauchy, A.L., *Oeuvres Complètes de A.L. Cauchy*, 26 Vols., Gauthier-Villars, Paris (1882–1938).

[Cay] Cayley, A., *On the theory of groups as depending on the symbolic equation* $\theta^n = 1$, Phil. Mag. **7** (1854), 40–47.

[CP1] Chadha, G.S. and Passi, I.B.S., *Idempotents in matrix group rings*, J. Pure Appl. Algebra **94** (1994), 283–284.
[CP2] Chadha, G.S. and Passi, I.B.S., *Bass conjecture and the group trace property*, Comm. Algebra **26** (1998), 627–639.
[Cl] Cliff, G.H., *Zero divisors and idempotents in group rings*, Canad. J. Math. **32** (1980), 596–602.
[CS1] Cliff, G.H. and Sehgal, S.K., *On the trace of an idempotent in a group ring*, Proc. Amer. Math. Soc. **62** (1977), 11–14.
[CS2] Cliff, G.H. and Sehgal, S.K., *Group rings whose units form an FC-group*, Math. Z. **161** (1978), 163–168.
[CS3] Cliff, G.H. and Sehgal, S.K., *Groups which are normal in the unit groups of their group rings*, Arch. Math. **33** (1979), 529–537.
[CS4] Cliff, G.H. and Sehgal, S.K., *Group rings with units torsion over the center*, Manuscripta Math. **33** (1980), 145–158.
[CSW] Cliff, G.H., Sehgal, S.K. and Weiss, A., *Units of integral group rings of metabelian groups*, J. Algebra **73** (1981), 167–185.
[CW1] Cliff, G.H. and Weiss, A., *Moody's induction theorem*, Illinois J. Math. **32** (1988), 489–500.
[CW2] Cliff, G.H. and Weiss, A., *Finite groups of matrices over group rings*, Trans. Amer. Math. Soc. **350** (2000), 457–475.
[Coe1] Coelho, S.P., *Group rings with units of bounded exponent over the center*, Canad. J. Math. **34** (1982), 1349–1364.
[CoP] Coelho, S.P. and Polcino Milies, C., *A note on central idempotents in group rings*, Proc. Edinburgh Math. Soc. **30** (1987), 69–72.
[CL1] Cohn, J.A. and Livingstone, D., *On groups of order p^3*, Canad. J. Math. **15** (1963), 622–624.
[CL2] Cohn, J.A. and Livingstone, D., *On the structure of group algebras*, Canad. J. Math. **17** (1965), 583–593.
[Coh1] Cohn, P.M., *Generalization of a theorem of Magnus*, Proc. London Math. Soc. **2** (1952), 297–310.
[Coh2] Cohn, P.M., *On the free product of associative rings III*, J. Algebra **8** (1968), 376–383.
[Coh3] Cohn, P.M., *Free Rings and Their Relations*, 2nd edn, Academic Press, London (1985).
[Col1] Coleman, D.B., *Finite groups with isomorphic group algebras*, Trans. Amer. Math. Soc. **105** (1962), 1–8.
[Col2] Coleman, D.B., *On the modular group ring of p-group*, Proc. Amer. Math. Soc. **15** (1964), 511–514.
[Col3] Coleman, D.B., *Idempotents in group rings*, Proc. Amer. Math. Soc. **17** (1966), 962.
[Col4] Coleman, D.B., *On group rings*, Canad. J. Math. **22** (1970), 249–254.
[CP] Coleman, D.B. and Passman, D.S., *Units in modular group rings*, Proc. Amer. Math. Soc. **25** (1970), 510–512.
[Con] Connell, I.G., *On the group ring*, Canad. J. Math. **15** (1963), 650–685.
[CR] Curtis, C.W. and Reiner, I., *Representation Theory of Finite Groups and Associative Algebras*, Wiley, New York (1962).
[Da] Dade, E.C., *Deux groups finis distincts ayant la même algebre de group sur tout corps*, Math. Z. **119** (1971), 345–348.
[Deh] Dehn, M., *Über die Grundlagen der Projectiven Geometrie und allgemeine Zahlsysteme*, Math. Ann. **85** (1922), 184–194.
[Den1] Dennis, R.K., *Units of group rings*, J. Algebra **43** (1976), 655–665.
[Den2] Dennis, R.K., *The Structure of the Unit Group of Group Rings*, Lecture Notes Pure Appl. Math., Vol. 26, Marcel Dekker, New York (1977).
[Des] Deskins, E., *Finite Abelian groups with isomorphic group algebras*, Duke Math. J. **23** (1956), 35–40.
[Di] Dieckman, E.M., *Isomorphism of group algebras of p-groups*, Thesis, Washington Univ. (1967).
[Do] Dokuchaev, M., *Finite subgroups of units in integral group rings*, Resenhaus **2** (1996), 293–302.
[DG] Dokuchaev, M. and Gonçalves, J.Z., *Semigroup identities on units of integral group rings*, Glasgow Math. J. **39** (1997), 1–6.
[DJ] Dokuchaev, M.A. and Juriaans, S.O., *Finite subgroups in integral group rings*, Canad. J. Math. **48** (1996), 1170–1179.
[DJP] Dokuchaev, M.A., Juriaans, S.O. and Polcino Milies, C., *Integral group rings of Frobenius groups and the conjectures of H.J. Zassenhaus*, Comm. Algebra **25** (1997), 2311–2325.

[DS] Dokuchaev, M.A. and Sehgal, S.K., *Torsion units in integral group rings of solvable groups*, Comm. Algebra **22** (1994), 5005–5020.

[DSS] Dokuchaev, M.A. and Sobral Singer, M.L., *Units in group rings of free products of prime cyclic groups*, Canad. J. Math. **50** (1998), 312–322.

[DuS] Dubois, P.F. and Sehgal, S.K., *Another proof of the invariance of Ulm's functions in commutative modular group rings*, Math. J. Okayama Univ. **15** (1972), 137–139.

[FU] Faith, C. and Utumi, Y., *On Noetherian prime rings*, Trans. Amer. Math. Soc. **114** (1965), 53–60.

[F] Farkas, D., *Group rings: An annotated questionaire*, Comm. Algebra **8** (1980), 585–602.

[FL] Farkas, D.R. and Linnell, P., *Zero divisors in group rings: something old, something new*, Contemporary Math. **93** (1989), 155–166.

[FS] Farkas, D.R. and Snider, R.L., *K_0 and Noetherian group rings*, J. Algebra **42** (1976), 192–198.

[Fe1] Fernandes, N.A., *Torsion units in the integral group ring of S_4*, Bol. Soc. Brasileira de Mat. **18** (1987), 1–10.

[Fe2] Fernandes, N.A., *A characterization of units in $\mathbf{Z}D_m$*, J. Indian Math. Soc. **55** (1990), 89–110.

[FPS] Fisher, J.L., Parmenter, M.M. and Sehgal, S.K., *Group rings with solvable n-Engel unit groups*, Proc. Amer. Math. Soc. **59** (1976), 195–200.

[For1] Formanek, E., *A short proof of a theorem of Jennings*, Proc. Amer. Math. Soc. **26** (1970), 405–407.

[For2] Formanek, E., *Idempotents in Noetherian group rings*, Canad. J. Math. **25** (1973), 366–369.

[For3] Formanek, E., *The zero divisor question for supersolvable groups*, Bull. Austral. Math. Soc. **9** (1973), 67–71.

[Fox] Fox, R.H., *Free differential calculus I*, Ann. Math. **57** (1953), 547–560.

[Fr] Frobenius, G., *Neuer Beweis des Sylowschen Satzes*, J. Reine Angew. Math. **100** (1887), 179–181.

[Fu1] Fuchs, L., *Infinite Abelian Groups, Vol. 1*, Academic Press, New York (1970).

[Fu2] Fuchs, L., *Infinite Abelian Groups, Vol. 2*, Academic Press, New York (1973).

[Fur1] Furukawa, T., *A note on the isomorphism invariants of a modular group algebra*, Math. J. Okayama Univ. **23** (1981), 1–5.

[Fur2] Furukawa, T., *On isomorphism invariants of integral group rings*, Math. J. Okayama Univ. **23** (1981), 125–130.

[Fur3] Furukawa, T., *Note on groups with isomorphic group algebras*, Math. J. Okayama Univ. **24** (1982), 1–6.

[Fur4] Furukawa, T., *The group of normalized units of a group ring*, Osaka J. Math. **23** (1986), 217–221.

[Fur5] Furukawa, T., *Isomorphism of group rings of infinite nilpotent groups*, Osaka J. Math. **24** (1987), 95–105.

[GJ] Giambruno, A. and Jespers, E., *Central idempotents and units in rational group algebras of alternating groups*, Internat. J. Algebra Comput. **8** (1998), 467–477.

[GJV] Giambruno, A., Jespers, E. and Valenti, A., *Group identities on units of rings*, Arch. Math. **63** (1994), 291–296.

[GS1] Giambruno, A. and Sehgal, S.K., *A Lie property in group rings*, Proc. Amer. Math. Soc. **105** (1989), 287–292.

[GS2] Giambruno, A. and Sehgal, S.K., *Automorphisms of the integral group ring of the wreath product of p-group with S_n*, Rocky Mountain J. Math. **22** (1992), 1303–1316.

[GS3] Giambruno, A. and Sehgal, S.K., *Lie nilpotence of group rings*, Comm. Algebra **21** (1993), 4253–4261.

[GS4] Giambruno, A. and Sehgal, S.K., *Generators of large subgroups of units of integral group rings of nilpotent groups*, J. Algebra **174** (1995), 150–156.

[GSV1] Giambruno, A., Sehgal, S.K. and Valenti, A. *Automorphisms of the integral group ring of the hyperoctahedral group*, Comm. Algebra **18** (1990), 1989–2005.

[GSV2] Giambruno, A., Sehgal, S.K. and Valenti, A., *Automorphisms of the integral group rings of some wreath products*, Comm. Algebra **19** (1991), 519–534.

[GSV3] Giambruno, A., Sehgal, S.K. and Valenti, A., *Group algebras whose units satisfy a group identity*, Proc. Amer. Math. Soc. **125** (1997), 629–634.

[GSV4] Giambruno, A., Sehgal, S.K. and Valenti, A., *Symmetric units and group identities*, Manuscripta Math. **96** (1998), 443–461.

[GSV5] Giambruno, A., Sehgal, S.K. and Valenti, A., *Group identities on units of group algebras*, J. Algebra **226** (2000), 488–504.

[Gol1] Goldie, A.W., *The structure of prime rings under ascending chain conditions*, Proc. London Math. Soc. **8** (1958), 589–608.

[Gol2] Goldie, A.W., *Semiprime rings with maximum condition*, Proc. London Math. Soc. **10** (1960), 201–220.

[Go1] Gonçalves, J.Z., *Free subgroups of units in group rings*, Canad. Math. Bull. **27** (1984), 309–312.

[Go2] Gonçalves, J.Z., *Free groups in subnormal subgroups and the residual nilpotence of the group of units of group rings*, Canad. Math. Bull. **27** (1984), 365–370.

[Go3] Gonçalves, J.Z., *Free subgroups in the group of units of group rings II*, J. Number Theory **21** (1985), 121–127.

[Go4] Gonçalves, J.Z., *Integral group rings whose group of units is solvable – an elementary proof*, Bol. Soc. Bras. Mat. **16** (1985), 1–9.

[Go5] Gonçalves, J.Z., *Normal and subnormal subgroups in the group of units of groups rings*, Bull. Austral. Math. Soc. **31** (1985), 355–363.

[Go6] Gonçalves, J.Z., *Group rings with solvable unit groups*, Comm. Algebra **14** (1986), 1–20.

[GM] Gonçalves, J.Z. and Mandel, A., *Semigroup identities on units of group algebras*, Arch. Math. **57** (1991), 539–545.

[GMS] Gonçalves, J.Z., Mandel, A. and Shirvani, M., *Free products of units in algebras I. Quaternion algebras*, J. Algebra **214** (1999), 301–316.

[GP1] Gonçalves, J.Z. and Passman, D.S., *Construction of free subgroups in the group of units of modular group algebras*, Comm. Algebra **24** (1996), 4211–4215.

[GP2] Gonçalves, J.Z. and Passman, D.S., *Free unit groups in group algebras*, J. Algebra **246** (2001), 226–252.

[GRS] Gonçalves, J.Z., Ritter, J. and Sehgal, S.K., *Subnormal subgroups in* $U(\mathbf{Z}G)$, Proc. Amer. Math. Soc. **103** (1988), 375–382.

[GJP] Goodaire, E.G., Jespers, E. and Parmenter, M.M., *Determining units in some integral group rings*, Canad. Math. Bull. **33** (1990), 242–246.

[Gr1] Gruenberg, K.W., *Residual properties of infinite soluble groups*, Proc. London Math. Soc. **7** (1957), 29–62.

[Gr2] Gruenberg, K.W., *Cohomological Topics in Group Theory*, Lecture Notes in Math., Vol. 143, Springer-Verlag, New York (1970).

[GRV] Gulliksen, T., Ribenboim, P. and Viswanathan, T., *An elementary note on group rings*, J. Reine Angew. Math. **242** (1970), 148–162.

[Gu1] Gupta, N.D., *A problem of R.H. Fox*, Canad. Math. Bull. **24** (1981), 129–136.

[Gu2] Gupta, N.D., *Free Group Rings*, Contemporary Math. Series, Vol. 66, AMS (1987).

[Gu3] Gupta, N.D., *The dimension subgroup conjecture*, Bull. London Math. Soc. **22** (1990), 453–456.

[GL] Gupta, N.D. and Levin, F., *On the Lie ideals of a ring*, J. Algebra **81** (1983), 225–231.

[GT] Gupta, N.D. and Tahara, K., *The seventh and eighth Lie dimension subgroups*, J. Pure Appl. Algebra **88** (1993), 107–117.

[Ham] Hall, M., *The Theory of Groups*, MacMillan, New York (1959).

[Hap1] Hall, P., *Finiteness conditions for soluble groups*, Proc. London Math. Soc. **4** (1954), 419–436.

[Hap2] Hall, P., *The Edmonton Notes on Nilpotent Groups*, Queen Mary College Math. Notes (1969).

[Har1] Hartley, B., *The residual nilpotence of wreath products*, Proc. London Math. Soc. **20** (1970), 365–392.

[Har2] Hartley, B., *Free groups in normal subgroups of unit groups and arithmetic groups*, Contemporary Math. **93** (1989), 173–178.

[HP] Hartley, B. and Pickel, P.F., *Free subgroups in the unit groups of integral group rings*, Canad. J. Math. **32** (1980), 1342–1352.

[Haw1] Hawkins, T., *The origins of the theory of group characters*, Archive Hist. Exact Sci. **7** (1970–1971), 142–170.

[Haw2] Hawkins, T., *Hypercomplex numbers, Lie groups and the creation of group representation theory*, Archive Hist. Exact Sci. **8** (1972), 243–287.

[Haw3] Hawkins, T., *New light on Frobenius' creation of the theory of group characters*, Archive Hist. Exact Sci. **12** (1974), 217–243.

[He1] Herstein, I.N., *Topics in Ring Theory*, Univ. Chicago Press (1965).

[He2] Herstein, I.N., *Noncommutative Rings*, Carus Mathematical Monographs, No. 15, Math. Assoc. Amer. (1968).

[Hew] Hertweck, M., *A counterexample to the isomorphism problem for integral group rings*, Ann. of Math. **154** (2001), 115–138.

[Hi1] Higman, G., *Units in group rings*, D.Phil. Thesis, University of Oxford, Oxford (1940).

[Hi2] Higman, G., *The units of group rings*, Proc. London Math. Soc. **46** (1940), 231–248.

[Hil] Hill, E.T., *Ideals in the modular group ring of a p-group*, Proc. Amer. Math. Soc. **28** (1971), 389–390.

[Ho1] Hoechsmann, K., *Functors on finite vector spaces and units in Abelian group rings*, Canad. Math. Bull. **29** (1986), 79–83.

[Ho2] Hoechsmann, K., *Units and class-groups in integral elementary Abelian group rings*, J. Pure Appl. Algebra **47** (1987), 253–264.

[Ho3] Hoechsmann, K., *On the Bass–Milnor index of Abelian p-groups*, Contemporary Math. **93** (1989), 179–195.

[Ho4] Hoechsmann, K., *Local units and circular index in Abelian p-group rings*, J. Pure Appl. Algebra **82** (1992), 253–272.

[Ho5] Hoechsmann, K., *Constructing units in commutative group rings*, Manuscripta Math. **75** (1992), 5–23.

[Ho6] Hoechsmann, K., *On the arithmetic of commutative group rings*, Group Theory, Algebra and Number Theory (Saarbrücken, 1993), de Gruyter, Berlin (1996), 145–201.

[HR1] Hoechsmann, K. and Ritter, J., *Logarithms and units in p-adic Abelian group rings*, Arch. Math. **49** (1987), 23–28.

[HR2] Hoechsmann, K. and Ritter, J., *Constructible units in Abelian p-group rings*, J. Pure Appl. Algebra **6** (1990), 325–339.

[HoS1] Hoechsmann, K. and Sehgal, S.K., *Units in regular elementary Abelian group rings*, Arch. Math. **47** (1986), 413–417.

[HoS2] Hoechsmann, K. and Sehgal, S.K., *Integral group rings without proper units*, Canad. Math. Bull. **30** (1987), 36–42.

[HoS3] Hoechsmann, K. and Sehgal, S.K., *Units in regular Abelian p-group rings*, J. Number Theory **30** (1988), 375–381.

[HSW] Hoechsmann, K., Sehgal, S.K. and Weiss, A., *Cyclotomic units and the unit group of an elementary Abelian group ring*, Arch. Math. **45** (1985), 5–7.

[Hol1] Holvoet, R., *Sur les Z_2-algb̀res du groupe di'edral d'ordre 8 et du groupe quaternionique*, C. R. Acad. Sci. Paris, Sér. A **262** (1966), 209–210.

[Hol2] Holvoet, R., *Sur les algebres modulaires des groupes non commutatifs d'ordre 12*, C. R. Acad. Sci. Paris, Sér. A **262** (1966), 690–692.

[Hol3] Holvoet, R., *Sur l'isomorophie d'algebres des groupes*, Bull. Soc. Math. Belg. **20** (1968), 264–282.

[Hol4] Holvoet, R., *De groepalgebra van een eindige p-groep over een veld met karakteristiek p*, Simon Stevin **42** (1969), 157–170.

[HP] Hughes, I. and Pearson, K.R., *The group of units of the integral group ring* $\mathbf{Z}S_3$, Canad. Math. Bull. **15** (1972), 529–534.

[Hu] Hurley, T.C., *Identifications in a free group*, J. Pure Appl. Algebra **48** (1987), 249–261.

[HS1] Hurley, T.C. and Sehgal, S.K., *The Lie dimension subgroup conjecture*, J. Algebra **143** (1991), 46–56.

[HS2] Hurley, T.C. and Sehgal, S.K., *The modular Fox subgroups*, Comm. Algebra **24** (1996), 4563–4580.

[HS3] Hurley, T.C. and Sehgal, S.K., *Groups related to Fox subgroups*, Comm. Algebra **28** (2000), 1051–1059.

[IP] Isaacs, I.M. and Passman, D.S., *Groups with representations of bounded degree*, Canad. J. Math. **16** (1964), 299–309.

[JM] Jackowski, S. and Marciniak, Z., *Group automorphisms inducing the identity map on cohomology*, J. Pure Appl. Algebra **44** (1987), 241–250.

[Jak] Jackson, D.A., *The groups of units of the integral group rings of finite metabelian and finite nilpotent groups*, Quart. J. Math. Oxford (2) **20** (1969), 319–331.

[Jac1] Jacobson, N., *Structure of Rings*, Amer. Math. Soc. Colloquium, Vol. 37, Providence, RI (1956).

[Jac2] Jacobson, N., *PI-Algebras: An Introduction*, Lecture Notes in Math., Vol. 441, Springer, Berlin (1975).

[JKMS] Jain, S.K., Kanwar, P., Malik, S. and Srivastva, J.B., *KD_∞ is a CS algebra*, Proc. Amer. Math. Soc. **28** (1999), 397–400.

[Jen1] Jennings, S.A., *The structure of the group ring of a p-group over a modular field*, Trans. Amer. Math. Soc. **50** (1941), 175–185.

[Jen2] Jennings, S.A., *On rings whose associated Lie rings are nilpotent*, Bull. Amer. Math. Soc. **53** (1947), 593–597.

[Jen3] Jennings, S.A., *The group ring of a class of infinite nilpotent groups*, Canad. J. Math. **7** (1955), 169–187.

[Jen4] Jennings, S.A., *Radical rings with nilpotent associated groups*, Trans. Roy. Soc. Canad. **49** (1955), 31–38.

[Jes1] Jespers, E., *Bicyclic units in some integral group rings*, Canad. Math. Bull. **38** (1995), 80–86.

[Jes2] Jespers, E., *Free normal complements and the unit group in integral group rings*, Proc. Amer. Math. Soc. **122** (1994), 59–66.

[Jes3] Jespers, E., *Units in integral group rings: A survey*, Methods in Ring Theory, Lecture Notes in Pure and Applied Math., Vol. 198, Dekker, New York (1998), 141–169.

[JD] Jespers, E. and Del Rio, A., *A structure theorem for the unit group of the integral group ring of some finite groups*, J. Reine Angew. Math. **521** (2000), 99–117.

[JL1] Jespers, E. and Leal, G., *Describing units in integral group rings of some* 2*-groups*, Comm. Algebra **19** (1991), 1808–1827.

[JL2] Jespers, E. and Leal, G., *Generators of large subgroups of the unit group of integral group rings*, Manuscripta Math. **78** (1993), 303–315.

[JL3] Jespers, E. and Leal, G., *Degree 1 and 2 representations of nilpotent groups and applications to units of group rings*, Manuscripta Math. **86** (1995), 479–498.

[JL4] Jespers, E. and Leal, G., *Units of integral group rings of Hamiltonian groups*, Comm. Algebra **23** (1995), 623–628.

[JL5] Jespers, E. and Leal, G., *Free products of Abelian groups in the unit group of integral group rings*, Proc. Amer. Math. Soc. **126** (1998), 1257–1265.

[JL6] Jespers, E. and Leal, G., *Congruence subgroups in group rings*, Comm. Algebra **28** (2000), 3155–3168.

[JLD] Jespers, E., Leal, G. and Del Rio, A., *Products of free groups in the unit group of integral group rings*, J. Algebra **180** (1996), 22–40.

[JLP] Jespers, E., Leal, G. and Parmenter, M.M., *Bicyclic and Bass cyclic units in group rings*, Canad. Math. Bull. **36** (1993), 178–182.

[JLPo] Jespers, E., Leal, G. and Polcino Milies, C., *Units of integral group rings of some metacyclic groups*, Canad. Math. Bull. **37** (1994), 228–277.

[JP1] Jespers, E. and Parmenter, M., *Bicyclic units in $\mathbf{Z}S_3$*, Bull. Soc. Math. Belg. Ser. B **44** (1992), 141–146.

[JP2] Jespers, E. and Parmenter, M., *Units of group rings of groups of order 16*, Glasgow Math. J. **35** (1993), 367–379.

[JPS] Jespers, E., Parmenter, M. and Sehgal, S.K., *Central units of integral group rings of nilpotent groups*, Proc. Amer. Math. Soc. **124** (1996), 1007–1012.

[JPSm] Jespers, E., Parmenter, M. and Smith, P., *Revisiting a theorem of Higman*, Proceedings, Groups ’93, C.M. Campbell, T.C. Hurley, E.F. Robertson, S.J. Tobin and J.J. Ward, eds, London Math. Soc. Lecture Notes, Vol. 211, London (1995), 269–273.

[JPo] Jespers, E. and Polcino Milies, C., *Units of group rings*, J. Pure Appl. Algebra **107** (1996), 233–251.

[Kap1] Kaplansky, I., *Groups with representations of bounded degree*, Canad. J. Math. **1** (1949), 105–112.

[Kap2] Kaplansky, I., *Problems in the theory of rings*, NAS-NRC Publ. 502, Washington (1957), 1–3.

[Kap3] Kaplansky, I., *Fields and Rings*, Univ. Chicago Press, Chicago (1969).

[Kap4] Kaplansky, I., *Problems in the theory of rings (revisited)*, Amer. Math. Monthly **77** (1970), 445–454.

[Kap5] Kaplansky, I., *Encyclopaedia Brittanica*, Yearbook (1974).

[KM] Kargapolov, M. and Merjlakov, Yu., *Fundamentals of the Theory of Groups*, Springer-Verlag, New York (1979).

[Kar] Karpilovsky, G., *Unit Groups of Group Rings*, Longman Scientific & Technical, New York (1989).

[Kh1] Khripta, I.I., *On the multiplicative group of a group ring*, Thesis, Uzhgorod (1971).

[Kh2] Khripta, I.I., *The nilpotence of the multiplicative group of a group ring*, Mat. Zametki **11** (1972), 191–200; Math. Notes **11** (1972), 119–124.

[KLST] Kimmerle, W., Lyons, R., Sandling, R. and Teague, D., *Composition factors from the group ring and Artin's Theorem on orders of simple groups*, Proc. London Math. Soc. **60** (1990), 89–122.
[KiS1] Kimmerle, W. and Sandling, R., *The determination of Abelian Hall subgroups by a conjugacy class structure*, Publications Mat. **36** (1992), 685–691.
[KiS2] Kimmerle, W. and Sandling, R., *Group theoretic and group ring theoretic determination of certain Sylow and Hall subgroups and the resolution of a question of R. Brauer*, J. Algebra **171** (1995), 329–346.
[Kl1] Kleinert, E., *Einheiten in* $\mathbf{Z}[D_{2m}]$, J. Number Theory **13** (1981), 541–561.
[Kl2] Kleinert, E., *On the class number of some group rings*, Arch. Math. **49** (1987), 200–207.
[Kl3] Kleinert, E., *Units in* $\mathbf{Z}[Q_p]$, J. Number Theory **26** (1987), 227–236.
[Kl4] Kleinert, E., *A theorem of units of integral group rings*, J. Pure Appl. Algebra **49** (1987), 161–171.
[Kl5] Kleinert, E., *Units of classical orders: a survey*, Enseign. Math. **40** (3–4) (1994), 205–248.
[Kl6] Kleinert, E., *Handling integral* p*-group rings*, Comm. Algebra **24** (1996), 3193–3210.
[Kl] Kline, M., *Mathematical Thought from Ancient to Modern Times*, Oxford Univ. Press, New York (1972).
[Kli] Klingler, L., *Construction of a counterexample to a conjecture of Zassenhaus*, Comm. Algebra **19** (1991), 2303–2330.
[KS] Kmet, S. and Sehgal, S.K., *Some isomorphism invariants of integral group rings*, Rocky Mountain J. Math. **15** (1985), 451–458.
[KLM] Kropholler, P.H., Linnell, P.A. and Moody, J.A., *Applications of a new* K*-theoretic theorem to soluble group rings*, Proc. Amer. Math. Soc. **104** (1988), 675–684.
[L] Lam, T.Y., *A First Course in Noncommutative Rings*, Springer-Verlag, New York (1991).
[Lam] Lambek, J., *Lectures on Rings and Modules*, Blaisdell, Toronto (1966).
[Lan1] Lang, S., *Cyclotomic Fields*, Springer, New York (1978).
[Lan2] Lang, S., *Algebra*, Addison-Wesley, Redwood (1984).
[Laz] Lazard, M., *Sur les groupes nilpotents et les anneaux de Lie*, Ann. Sci. Ecole Norm Sup. **71** (1954), 101–190.
[LD] Leal, G. and Del Río, A., *Products of free groups in the unit group of integral group rings, II*, J. Algebra **191** (1997), 240–251.
[Lee1] Lee, G., *Group rings whose symmetric elements are Lie nilpotent*, Proc. Amer. Math. Soc. **127** (1999), 3153–3159.
[Lee2] Lee, G., *The Lie* n*-Engel property in group rings*, Comm. Algebra **28** (2000), 867–881.
[Lee3] Lee, G., *Nilpotent symmetric units in group rings*, Comm. Algebra (to appear).
[LS1] Lee, G.T. and Sehgal, S.K., *Generators of subgroups of finite index in* $GL_m(\boldsymbol{Z}G)$, Advances in Ring Theory, Birkhäuser, Boston (1997), 211–219.
[LS2] Lee, G.T. and Sehgal, S.K., *Torsion matrices over commutative integral group rings*, Publ. Mat. **44** (2000), 359–367.
[LeS1] Levin, F. and Sehgal, S.K., *On Lie nilpotent group rings*, J. Pure Appl. Algebra **37** (1985), 33–39.
[LeS2] Levin, F. and Sehgal, S.K., *Units of the integral group ring of the infinite dihedral group*, Infinite Groups and Group Rings (Tuscaloosa, AL, 1992), Ser. Algebra, Vol. 1, World Sci. Publishing, River Edge, NJ (1993), 57–68.
[Lew] Lewin, J., *A note on zero divisors in group rings*, Proc. Amer. Math. Soc. **31** (1972), 357–359.
[Li] Li, Y., *The hypercenter and the* n*-centre of the unit group of an integral group ring*, Canad. J. Math. **50** (1998), 401–411.
[LiPa1] Li, Y. and Parmenter, M.M., *Central units of the integral group ring* $\boldsymbol{Z}A_5$, Proc. Amer. Math. Soc. **125** (1997), 61–65.
[LiPa2] Li, Y. and Parmenter, M.M., *Hypercentral units in integral group rings*, Proc. Amer. Math. Soc. **129** (2001), 2235–2238.
[LPS] Li, Y., Parmenter, M.M. and Sehgal, S.K., *On the normalizer property for integral group rings*, Comm. Algebra **27** (1999), 4217–4223.
[Lic] Lichtman, A., *The residual nilpotency of the augmentation ideal and the residual nilpotency of some classes of groups*, Israel J. Math. **26** (1977), 276–293.
[LiS] Lichtman, A.I. and Sehgal, S.K., *The elements of finite order in the group of units of group rings of free products of groups*, Comm. Algebra **17** (1989), 2223–2253.

[Liu] Liu, Chia-Hsin, *Group algebras with units satisfying a group identity*, Proc. Amer. Math. Soc. **127** (1999), 327–336.

[LiP] Liu, Chia-Hsin and Passman, D.S., *Group algebras with units satisfying a group identity II*, Proc. Amer. Math. Soc. **127** (1999), 337–341.

[Lor1] Lorenz, M., *Finite normalizing extensions of rings*, Math. Z. **176** (1981), 447–484.

[Lor2] Lorenz, M., *The Goldie rank of prime supersolvable group algebras*, Mitt. Math. Sem. Giessen **149** (1981), 115–129.

[Lor3] Lorenz, M., *K_0 of skew group rings and simple Noetherian rings without idempotents*, J. London Math. Soc. **32** (1985), 41–50.

[Lor4] Lorenz, M., *Goldie ranks of prime polycyclic crossed products*, Perspectives in Ring Theory, Kluwer, Dordrecht (1988), 215–219.

[Los1] Losey, G., *On dimension subgroups*, Trans. Amer. Math. Soc. **97** (1960), 474–486.

[Los2] Losey, G., *On the structure of $Q_2(G)$ for finitely generated groups*, Canad. J. Math. **25** (1973), 353–359.

[Los3] Losey, G., *A remark on the units of finite order in the group ring of a finite group*, Canad. Math. Bull. **17** (1974), 129–130.

[Los4] Losey, G., *N-series and filtrations of the augmentation ideal*, Canad. J. Math. **26** (1974), 962–977.

[LP1] Luthar, I.S. and Passi, I.B.S., *Zassenhaus conjecture for A_5*, Proc. Indian Acad. Sci. **99** (1989), 1–5.

[LP2] Luthar, I.S. and Passi, I.B.S., *Torsion units in matrix group rings*, Comm. Algebra **20** (1992), 1223–1228.

[LT1] Luthar, I.S. and Trama, P., *Zassenhaus conjecture for certain integral group rings*, J. Indian Math. Soc. **55** (1990), 199–212.

[LT2] Luthar, I.S. and Trama, P., *Zassenhaus conjecture for S_5*, Comm. Algebra **19** (1991), 2353–2362.

[M] Magnus, W., *Über Beziehungen zwichen höhren Kommutatorn*, J. Reine Angew. Math. **177** (1937), 105–115.

[Mak] Makasikis, A., *Sur l'isomorphie d'algèbres de groupes sur un champ modulaire*, Bull. Soc. Math. Belg. **28** (1976), 91–109.

[Mal1] Malcev, A.I., *On embedding of group algebras in a division algebra*, Dokl. Akad. Nauk SSSR **60** (1948), 1499–1501 (in Russian).

[Mal2] Malcev, A.I., *On certain classes of infinite solvable groups*, Mat. Sb. **28** (1951), 567–588 (in Russian).

[MRSW] Marciniak, Z., Ritter, J., Sehgal, S.K. and Weiss, A., *Torsion units in integral group rings of some metabelian groups, II*, J. Number Theory **25** (1987), 340–352.

[MS1] Marciniak, Z. and Sehgal, S.K., *Hirsch induction and torsion units in group rings*, Arch. Math. **64** (1995), 374–384.

[MS2] Marciniak, Z. and Sehgal, S.K., *Finite matrix groups over nilpotent group rings*, J. Algebra **181** (1996), 565–583.

[MS3] Marciniak, Z. and Sehgal, S.K., *Zassenhaus conjecture and infinite nilpotent groups*, J. Algebra **184** (1996), 207–212.

[MS4] Marciniak, Z. and Sehgal, S.K., *Constructing free subgroups of integral group ring units*, Proc. Amer. Math. Soc. **125** (1997), 1005–1009.

[MS5] Marciniak, Z. and Sehgal, S.K., *Zassenhaus conjectures for infinite groups*, Canad. Math. Soc. Proc. **22** (1998), 99–111.

[MS6] Marciniak, Z. and Sehgal, S.K., *Subnormal subgroups of group ring units*, Proc. Amer. Math. Soc. **126** (1998), 343–348.

[MS7] Marciniak, Z. and Sehgal, S.K., *Units in group rings and geometry*, Methods in Ring Theory, Lecture Notes Pure Appl. Math., Vol. 198, Dekker, New York (1998), 185–198.

[MS8] Marciniak, Z. and Sehgal, S.K., *Torsion matrices over Abelian group rings*, J. Group Theory **3** (2000), 67–75.

[Maz] Mazur, M., *On the isomorphism problem for infinite group rings*, Expo. Math. **13** (1995), 433–445.

[May] May, W., *Commutative group algebras*, Trans. Amer. Math. Soc. **136** (1969), 139–149.

[Me] Menal, P., *Private letter to B. Hartley* (April 6, 1981).

[MNO] Michler, G, Newman, M. and O'Brien, E., *Modular group algebras* (unpublished).

[Mol1] Molien, T., *Über Systeme höherer complexer Zahlen*, Math. Ann. **XLI** (1893), 83–156.

[Mol2] Molien, T., *Über die Invarianten der linearen Substitutionsgruppen*, S'ber Akad. d. Wiss. Berlin (1897), 1152–1156.

[Moo1] Moody, J.A., *Induction theorems for infinite groups*, Bull. Amer. Math. Soc. **17** (1987), 113–116.

[Moo2] Moody, J.A., *Brauer induction theorem for G_0 of certain infinite groups*, J. Algebra **122** (1989), 1–14.

[Mor] Moran, S., *Dimension subgroups* mod n, Proc. Cambridge Philos. Soc. **68** (1970), 579–582.

[Mos] Moser, C., *Representation de -1 comme somme de carres dans un corps cyclotomique quelconque*, J. Number Theory **5** (1973), 139–141.

[MN] Motose, K. and Ninomiya, Y., *On the solvability of unit groups of group rings*, Math. J. Okayama Univ. **15** (1972), 209–214.

[MT1] Motose, K. and Tominaga, H., *Group rings with nilpotent unit groups*, Math. J. Okayama Univ. **14** (1969), 43–46.

[MT2] Motose, K. and Tominaga, H., *Group rings with solvable unit groups*, Math. J. Okayama Univ. **15** (1971), 37–40.

[MW] Musson, I. and Weiss, A., *Integral group rings with residually nilpotent unit groups*, Arch. Math. **38** (1982), 514–530.

[NT] Nakajima, A. and Tominaga, H., *A note on group rings of p-groups*, Math. J. Okayama Univ. **13** (1968), 107–109.

[NS1] Neisse, O. and Sehgal, S.K., *Units in integral group rings over Solomon fields*, Comm. Algebra **26** (1998), 3985–3991.

[NS2] Neisse, O. and Sehgal, S.K., *Gauss units in integral group rings*, J. Algebra **204** (1998), 588–596.

[Ne1] Neumann, B.H., *On ordered division rings*, Trans. Amer. Math. Soc. **66** (1949), 202–252.

[Ne2] Neumann, B.H., *Groups with finite classes of conjugate elements*, Proc. London Math. Soc. **1** (1951), 178–187.

[No1] Noether, E., *Hypercomplexe Grössen und Darstellungtheorie*, Math. Z. **30** (1929), 641–692.

[No2] Noether, E., *Nichtkommutative Algebra*, Math. Z. **37** (1933), 513–541.

[Ob1] Obayashi, T., *Solvable groups with isomorphic group algebras*, J. Math. Soc. Japan **18** (1966), 394–397.

[Ob2] Obayashi, T., *Integral group rings of finite groups*, Osaka J. Math. **7** (1970), 253–266.

[Or] Ore, O., *Linear equations in non-commutative fields*, Ann. Math. **32** (1931), 463–477.

[Oy] Oyama, T., *On the groups with the same table of characters as the alternating groups*, Osaka J. Math. **1** (1964), 91–101.

[Par1] Parmenter, M.M., *Dimension theory, isomorphism and related problems in group rings*, Thesis, University of Alberta (1972).

[Par2] Parmenter, M.M., *Isomorphic group rings*, Canad. Math. Bull. **18** (1975), 567–576.

[Par3] Parmenter, M.M., *Conjugates of units in integral group rings*, Comm. Algebra **23** (1995), 5503–5507.

[Par4] Parmenter, M.M., *Central units in integral group rings, Algebra some recent advances*, Indian Nat. Sci. Acad. (1999), 110–116.

[Par5] Parmenter, M.M., *Units and isomorphism in group rings*, Quest. Math. **8** (1985), 9–14.

[Par6] Parmenter, M.M., *Torsion free normal complements in unit groups of integral group rings*, C. R. Math. Rep. Acad. Sci. Canada **12** (1990), 113–118.

[Par7] Parmenter, M.M., *Free torsion-free normal complements in integral group rings*, Comm. Algebra **21** (1993), 3611–3617.

[PPSe] Parmenter, M.M., Passi, I.B.S. and Sehgal, S.K., *Polynomial ideals in group rings*, Canad. J. Math. **25** (1973), 1174–1182.

[PP] Parmenter, M.M. and Polcino Milies, C., *Group rings whose units form a nilpotent or FC group*, Proc. Amer. Math. Soc. **68** (1978), 247–248.

[PS1] Parmenter, M.M. and Sehgal, S.K., *Idempotent elements and ideals in group rings and the intersection theorem*, Arch. Math. **24** (1973), 586–600.

[PS2] Parmenter, M.M. and Sehgal, S.K., *Automorphisms of the integral group ring of the wreath product of p-group with S_n, the case $n = 2$*, Rocky Mountain J. Math. **23** (1993), 1379–1383.

[Pa] Pascaud, J., *Anneaux de groupes réduits*, C. R. Acad. Sci., Paris Sér. A **277** (1973), 719–722.

[Pas1] Passi, I.B.S., *Dimension subgroups*, J. Algebra **9** (1968), 152–182.

[Pas2] Passi, I.B.S., *Group Rings and Their Augmentation Ideals*, Lecture Notes in Math., Vol. 715, Springer-Verlag, Berlin (1979).

[Pas3] Passi, I.B.S., *Trace function methods in group rings*, Research Bull. Panjab Univ. **45** (1995), 149–174.

[PP] Passi, I.B.S. and Passman, D.S., *Algebraic elements in group rings*, Proc. Amer. Math. Soc. **108** (1990), 871–877.

[PPS] Passi, I.B.S., Passman, D.S. and Sehgal, S.K., *Lie solvable group rings*, Canad. J. Math. **25** (1973), 748–757.

[PS1] Passi, I.B.S. and Sehgal, S.K., *Isomorphism of modular group algebras*, Math. Z. **129** (1972), 65–73.

[PS2] Passi, I.B.S. and Sehgal, S.K., *Lie dimension subgroups*, Comm. Algebra **3** (1975), 59–73.

[P1] Passman, D.S., *Nil ideals in group rings*, Michigan Math. J. **9** (1962), 375–384.

[P2] Passman, D.S., *The group algebras of groups of order p^4 over a modular field*, Michigan Math. J. **12** (1965), 405–415.

[P3] Passman, D.S., *Isomorphic groups and group rings*, Pacific J. Math. **15** (1965), 561–583.

[P4] Passman, D.S., *Group algebras of groups of order p^4 over a modular field*, Michigan Math. J. **12** (1965), 405–415.

[P5] Passman, D.S., *Infinite Group Rings*, Marcel Dekker, New York (1971).

[P6] Passman, D.S., *Idempotents in group rings*, Proc. Amer. Math. Soc. **28** (1971), 371–374.

[P7] Passman, D.S., *Linear identities in group rings*, Pacific J. Math. **36** (1971), 457–484.

[P8] Passman, D.S., *Group rings satisfying a polynomial identity II*, Pacific J. Math. **39** (1971), 425–438.

[P9] Passman, D.S., *Group rings satisfying a polynomial identity III*, Proc. Amer. Math. Soc. **31** (1972), 87–90.

[P10] Passman, D.S., *On a ring of quotients of a group ring*, Proc. Amer. Math. Soc. **33** (1972), 221–225.

[P11] Passman, D.S., *Group rings satisfying a polynomial identity*, J. Algebra **20** (1972), 103–117.

[P12] Passman, D.S., *Primitive group rings*, Pacific J. Math. **47** (1973), 499–506.

[P13] Passman, D.S., *On the semisimplicity of group rings of linear groups*, Pacific J. Math. **47** (1973), 221–228.

[P14] Passman, D.S., *On the semisimplicity of group rings of linear groups II*, Pacific J. Math. **48** (1973), 215–234.

[P15] Passman, D.S., *A new radical for group rings*?, J. Algebra **28** (1974), 556–572.

[P16] Passman, D.S., *The semisimplicity problem for group rings*, Symp. Math. **8** (1974), 333–342.

[P17] Passman, D.S., *Advances in group rings*, Israel J. Math. **19** (1974), 67–107.

[P18] Passman, D.S., *On the semisimplicity of group rings of some locally finite groups*, Pacific J. Math. **58** (1975), 179–207.

[P19] Passman, D.S., *What is a group ring?*, Amer. Math. Monthly **83** (1976), 173–185.

[P20] Passman, D.S., *Observations on group rings*, Comm. Algebra **5** (1977), 1119–1162.

[P21] Passman, D.S., *The Algebraic Structure of Group Rings*, Wiley-Interscience, New York (1977).

[P22] Passman, D.S., *Group rings of polycyclic groups*, Group Theory: Essays for Philip Hall, Academic Press, London (1984), 207–256.

[P23] Passman, D.S., *On the Goldie rank of group algebras*, Comm. Algebra **13** (1985), 1305–1311.

[P24] Passman, D.S., *Infinite Crossed Products*, Academic Press, New York (1989).

[P25] Passman, D.S., *Group Algebras whose units satisfy a group identity II*, Proc. Amer. Math. Soc. **125** (1997), 657–662.

[P26] Passman, D.S., *Semiprimitivity of Group Algebras: Past Results and Recent Progress*, Trends in Ring Theory, Conference Proc. Canadian Math. Soc., Vol. 22, Amer. Math. Soc., Providence, RI (1998).

[PSm] Passman, D.S. and Smith, P.F., *Units in integral group rings*, J. Algebra **69** (1981), 213–239.

[Pe1] Pearson, K.R., *On the units of a modular group ring*, Bull. Austral. Math. Soc. **7** (1972), 169–182.

[Pe2] Pearson, K.R., *On the units of a modular group ring II*, Bull. Austral. Math. Soc. **8** (1973), 435–442.

[Pei] Peirce, B., *Linear associative algebras*, Amer. J. Math. **4** (1881), 97–215.

[PW] Perlis, S. and Walker, G.L., *Abelian group algebras of finite order*, Trans. Amer. Math. Soc. **68** (1950), 420–426.

[Pet1] Peterson, G., *On automorphisms of the integral group ring of S_n*, Proc. Amer. Math. Soc. **59** (1976), 14–18.

[Pet2] Peterson, G., *On automorphisms of integral group I*, Arch. Math. **28** (1977), 577–583.

[Pet3] Peterson, G., *Automorphisms of integral group rings II*, Illinois J. Math. **21** (1977), 836–844.

[Po1] Polcino Milies, C., *The units of the integral group ring $\mathbf{Z}D_4$*, Bol. Soc. Brasil. Mat. **4** (1973), 85–92.

[Po2] Polcino Milies, C., *Anéis de Grupo*, IV Escola de Álgebra, São Paulo (1976).

[Po3] Polcino Milies, C., *Integral group rings with nilpotent unit groups*, Canad. J. Math. **28** (1976), 954–960.

[Po4] Polcino Milies, C., *A glance at the early history of group rings*, Groups – St. Andrews 1981, C.M. Campbell and E.F. Robinson, eds, London Math. Soc. Lecture Notes Series, Vol. 71, Cambridge Univ. Press, London (1982), 270–280.

[Po5] Polcino Milies, C., *Units of group rings: a short survey*, Groups – St. Andrews 1981, C.M. Campbell and E.F. Robinson, eds, London Math. Soc. Lecture Notes Series, Vol. 71, Cambridge Univ. Press, London (1982), 281–297.

[Po6] Polcino Milies, C., *Torsion units in group rings and a conjecture of H.J. Zassenhaus*, Group and Semigroup Rings, North-Holland Mathematics Studies, Vol. 126 (1986), 179–192.

[PRS] Polcino Milies, C., Ritter, J. and Sehgal, S.K., *On a conjecture of Zassenhaus on torsion units in integral group rings II*, Proc. Amer. Math. Soc. **97** (1986), 201–206.

[PSe1] Polcino Milies, C. and Sehgal, S.K., *FC-elements in a group ring*, Comm. Algebra **9** (1981), 1285–1293.

[PSe2] Polcino Milies, C. and Sehgal, S.K., *Torsion units in integral group rings of metacyclic groups*, J. Number Theory **19** (1984), 103–114.

[PSe3] Polcino Milies, C. and Sehgal, S.K., *Central units of integral group rings*, Comm. Algebra **27** (1999), 6233–6241.

[PSe4] Polcino Milies, C. and Sehgal, S.K., *An Introduction to Group Rings*, Kluwer, Dordrecht (2002).

[Pr] Promislow, S.D., *A simple example of a torsion-free non unique product group*, Bull. London Math. Soc. **20** (1988), 302–304.

[Rac1] Raggi Cárdenas, F.F., *Units in group rings I*, An. Inst. Mat. Univ. Nac. Autonoma Mexico **7** (1967), 27–35.

[Rac2] Raggi Cárdenas, F.F., *Units in group rings II*, An. Inst. Mat. Univ. Nac. Autonoma Mexico **8** (1968), 91–103.

[Rei] Reiner, I., *Maximal Orders*, Academic Press, New York (1975).

[Re] Renault, G., *Sur les anneaux de groupes*, C. R. Acad. Sci. Paris **273** (1971), 84–87.

[Rib] Ribenboim, P., *Rings and Modules*, Interscience, New York (1969).

[Ril] Riley, D., *Restricted Lie dimension subgroups*, Comm. Algebra **19** (1991), 1493–1499.

[Rip] Rips, E., *On the fourth integer dimension subgroup*, Israel J. Math. **12** (1972), 342–346.

[RSe] Rips, E. and Segev, Y., *Torsion-free groups without unique product property*, J. Algebra **108** (1987), 116–126.

[Rit] Ritter, J., *Large subgroups in the unit group of group rings (a survey)*, Bayreuther Math. Schriften **33** (1990), 153–171.

[RS1] Ritter, J. and Sehgal, S.K., *Integral group rings of some p-groups*, Canad. J. Math. **34** (1982), 233–246.

[RS2] Ritter, J. and Sehgal, S.K., *Isomorphism of group rings*, Arch. Math. **40** (1983), 32–39.

[RS3] Ritter, J. and Sehgal, S.K., *On a conjecture of Zassenhaus on torsion units in integral group rings*, Math. Ann. **264** (1983), 257–270.

[RS4] Ritter, J. and Sehgal, S.K., *Certain normal subgroups of units in group rings*, J. Reine Angew. Math. **381** (1987), 214–220.

[RS5] Ritter, J. and Sehgal, S.K., *Generators of subgroups of $\mathcal{U}(\mathbf{Z}G)$*, Contemp. Math. **93** (1989), 331–347.

[RS6] Ritter, J. and Sehgal, S.K., *Construction of units in integral group rings of finite nilpotent groups*, Bull. Amer. Math. Soc. (N.S.) **20** (1989), 165–168.

[RS7] Ritter, J. and Sehgal, S.K., *Integral group rings with trivial central units*, Proc. Amer. Math. Soc. **108** (1990), 327–329.

[RS8] Ritter, J. and Sehgal, S.K., *Construction of units in integral group rings of finite nilpotent groups*, Trans. Amer. Math. Soc. **324** (1991), 602–621.

[RS9] Ritter, J. and Sehgal, S.K., *Construction of units in group rings of monomial and symmetric groups*, J. Algebra **142** (1991), 511–526.

[RS10] Ritter, J. and Sehgal, S.K., *Units of group rings of solvable and Frobenius groups over large rings of cyclotomic integers*, J. Algebra **158** (1993), 116–129.

[Ro1] Robinson, D.J.S., *Finiteness Conditions and Generalized Solvable Groups*, Ergebnisse der Math., Vols. 62, 63, Springer-Verlag, New York (1972).

[Ro2] Robinson, D.J.S., *A Course in the Theory of Groups*, Springer-Verlag, New York (1993).

[Rog] Roggenkamp, K.W., *Integral group rings of solvable finite groups have no idempotent ideals*, Arch. Math. **25** (1974), 125–128.

[RoS1] Roggenkamp, K.W. and Scott, L.L., *Nonsplitting examples for normalized units in integral group rings of metacyclic Frobenius groups*, C. R. Math. Rep. Acad. Sci. Canad. **3** (1981), 29–32.

[RoS2] Roggenkamp, K.W. and Scott, L.L., *Units in metabelian group rings: Nonsplitting examples for normalized units*, J. Pure Appl. Algebra **27** (1983), 299–314.

[RoS3] Roggenkamp, K.W. and Scott, L.L., *Units in group rings: Splittings and the isomorphism problem*, J. Algebra **96** (1985), 397–417.

[RoS4] Roggenkamp, K.W. and Scott, L.L., *The isomorphism problem for integral group rings of finite nilpotent groups*, Proc. Int. Conf. St. Andrews, 1985, London Math. Soc. Lect. Notes Ser., Vol. 121 (1986), 291–299.

[RoS5] Roggenkamp, K.W. and Scott, L.L., *Isomorphisms for p-adic group rings*, Ann. Math. **126** (1987), 593–647.

[RT] Roggenkamp, K.W. and Taylor, M.J., *Group Rings and Class Groups*, Birkhäuser, Basel, Boston (1992).

[Rö1] Röhl, F., *On the isomorphism problem for integral group rings of circle groups*, Math. Z. **180** (1982), 419–422.

[Rö2] Röhl, F., *On induced isomorphism of group rings*, Proc. Groups – Korea, Lecture Notes in Math., Vol. 1098, Springer-Verlag, Berlin (1983), 136–141.

[Ros1] Roseblade, J.E., *Group rings of polycyclic groups*, J. Pure Appl. Algebra **3** (1973), 307–328.

[Ros2] Roseblade, J.E., *Applications of the Artin–Rees lemma to group rings*, Symp. Math. **17** (1976), 471–478.

[Ros3] Roseblade, J.E. and Smith, P.F., *A note on hypercentral group rings*, J. London Math. Soc. **13** (1976), 183–190.

[Ross1] Rosset, S., *The Goldie rank of virtually polycyclic group rings*, The Brauer Group, Lecture Notes in Math., Vol. 844, Springer, Berlin (1981), 35–45.

[Rot] Rotman, J., *The Theory of Groups*, Allyn and Bacon, Boston (1965).

[Row1] Rowen, L.H., *Polynomial Identities in Ring Theory*, Academic Press, New York (1988).

[Row2] Rowen, L.H., *Ring Theory, Vol. II*, Academic Press, New York (1988).

[Sak1] Saksanov, A.I., *On group rings of finite groups I*, Publ. Math. Debrecen **18** (1971), 187–209.

[Sak2] Saksanov, A.I., *On an integral ring of characters of a finite group*, Vesti Akad. Nauk USSR, Ser. Fiz.-Mat. (1966), 69–76.

[San1] Sandling, R., *The modular group rings of p-groups*, Thesis, Univ. of Chicago (1969).

[San2] Sandling, R., *Fixed point free actions on p-groups*, Arch. Math. **22** (1971), 231–236.

[San3] Sandling, R., *Dimension subgroups over arbitrary coefficient rings*, J. Algebra **21** (1972), 250–265.

[San4] Sandling, R., *The dimension subgroup problem*, J. Algebra **21** (1972), 216–231.

[San5] Sandling, R., *Note on the integral group ring problem*, Math. Z. **124** (1972), 255–258.

[San6] Sandling, R., *Group rings of circle and unit groups*, Math. Z. **140** (1974), 195–202.

[San7] Sandling, R., *Graham Higman's thesis "Units in Group Rings"*, Integral Representations and Applications, Lecture Notes in Math., Vol. 882, Springer-Verlag, Berlin (1981), 93–116.

[San8] Sandling, R., *Units in the modular group algebra of a finite Abelian p-group*, J. Pure Appl. Algebra **33** (1984), 337–346.

[San9] Sandling, R., *The isomorphism problem for group rings: a survey*, Orders and their Applications (Oberwolfach 1984), Lecture Notes in Math., Vol. 1142, Springer-Verlag, Berlin (1985), 256–288.

[San10] Sandling, R., *The modular group algebra of a central-elementary-by-Abelian p-group*, Arch. Math. **52** (1989), 22–27.

[San11] Sandling, R., *The modular group algebra problem for metacyclic p-groups*, Proc. Amer. Math. Soc. **124** (1996), 1347–1350.

[Se1] Sehgal, S.K., *On the isomorphism of group algebras*, Math. Z. **95** (1967), 71–75.

[Se2] Sehgal, S.K., *On the isomorphism of integral group rings I*, Canad. J. Math. **21** (1969), 410–413.

[Se3] Sehgal, S.K., *On the isomorphism of integral group rings II*, Canad. J. Math. **21** (1969), 1182–1188.

[Se4] Sehgal, S.K., *Units in commutative integral group rings*, Math. J. Okayama Univ. **14** (1970), 135–138.

[Se5] Sehgal, S.K., *Isomorphism of p-adic group rings*, J. Number Theory **2** (1970), 500–508.

[Se6] Sehgal, S.K., *On class sums in p-adic group rings*, Canad. J. Math. **23** (1971), 541–543.

[Se7] Sehgal, S.K., *Lie properties in modular group algebra*, Orders, Group Rings and Related Topics (Proc. Conf. Ohio State Univ., Columbus, OH), Lecture Notes in Math., Vol. 353, Springer-Verlag, New York (1973), 152–160.

[Se8] Sehgal, S.K., *Certain algebraic elements in group rings*, Arch. Math. **26** (1975), 139–143.

[Se9] Sehgal, S.K., *Nilpotent elements in group rings*, Manuscripta Math. **15** (1975), 65–80.

[Se10] Sehgal, S.K., *Topics in Group Rings*, Marcel Dekker, New York (1978).

[Se11] Sehgal, S.K., *Torsion units in group rings*, Methods in Ring Theory (Antwerp, 1983), NATO Adv. Sci. Inst. Ser. C: Math. Phys. Sci., Vol. 129, Reidel, Dordrecht (1984), 497–504.

[Se12] Sehgal, S.K., *Units of integral group rings – a survey*, Algebraic Structures and Number Theory, World Sci. Publishing, Teaneck, NJ (1990), 255–268.

[Se13] Sehgal, S.K., *Units in Integral Group Rings*, Longman, Essex (1993).

[Se14] Sehgal, S.K., *Large groups of units of integral group rings of finite nilpotent groups*, Resenhaus **1** (1993), 116–124.

[Se15] Sehgal, S.K., *Symmetric elements and identities in group algebras*, Algebra, Some Recent Advances, Indian Nat. Sci. Acad. (1999), 207–213.

[SSZ] Sehgal, Sudarshan K., Sehgal, Surinder K. and Zassenhaus, H.J., *Isomorphism of integral group rings of Abelian by nilpotent class two group*, Comm. Algebra **12** (1984), 2401–2407.

[SW] Sehgal, S.K. and Weiss, A., *Torsion units in integral group rings of some metabelian groups*, J. Algebra **103** (1986), 490–499.

[SZ1] Sehgal, S.K. and Zassenhaus, H.J., *Group rings whose units form an FC-group*, Math. Z. **153** (1977), 29–35.

[SZ2] Sehgal, S.K. and Zassenhaus, H.J., *Integral group rings with nilpotent unit groups*, Comm. Algebra **5** (1977), 101–111.

[SZ3] Sehgal, S.K. and Zassenhaus, H.J., *Group rings without nontrivial idempotents*, Arch. Math. **28** (1977), 378–380.

[SL] Shin, H. and Lyou, I., *On the structure of $V(\mathbf{Z}D_4)$*, Bull. Korean Math. Soc. **34** (1997), 645–651.

[SLD] Shin, H., Lyou, I. and Dixon, M.R., *On free product in $V(\mathbf{Z}S_3)$*, Semigroups (Kunming, 1995), Springer, Singapore (1998), 290–295.

[Si] Siegel, C.L., *Discontinuous groups*, Ann. Math. **44** (1943), 674–689.

[Sj] Sjögren, J.A., *Dimension and lower central subgroups*, J. Pure Appl. Algebra **14** (1979), 175–194.

[Smm] Smith, M.K., *Group algebras*, J. Algebra **18** (1971), 477–499.

[Sm1] Smith, P.F., *On the intersection theorem*, Proc. London Math. Soc. **21** (1970), 385–398.

[Sm2] Smith, P.F., *Localization in group rings*, Proc. London Math. Soc. **22** (1971), 69–90.

[Sm3] Smith, P.F., *Quotient rings of group rings*, J. London Math. Soc. **3** (1971), 645–660.

[Sm4] Smith, P.F., *On the dimension of group rings*, Proc. London Math. Soc. **25** (1972), 288–302.

[Sm5] Smith, P.F., *A note on idempotent ideals in group rings*, Arch. Math. **27** (1976), 22–27.

[Sn] Snider, R.L., *The zero divisor conjecture for some solvable groups*, Pacific J. Math. **90** (1980), 191–196.

[St] Strojnowski, A., *A note on u.p. groups*, Comm. Algebra **8** (1980), 231–234.

[Sy] Sylow, L., *Théorèmes sur les groupes de substitutions*, Math. Ann. **5** (1872), 584–594.

[Ta1] Tahara, K., *The fourth dimension subgroups and polynomial maps*, J. Algebra **45** (1977), 102–131.

[Ta2] Tahara, K., *The fourth dimension subgroups and polynomial maps, II*, Nagoya Math. J. **69** (1978), 1–7.

[Ti] Tits, J., *Free subgroups in linear groups*, J. Algebra **20** (1972), 250–270.

[Ut] Utumi, Y., *On quotient rings*, Osaka Math. J. **8** (1956), 1–18.

[V1] Valenti, A., *On the automorphism group of the integral group ring of $S_k wr S_n$*, J. Pure Appl. Algebra **78** (1992), 203–211.

[V2] Valenti, A., *Torsion units in integral group rings*, Proc. Amer. Math. Soc. **120** (1994), 1–4.

[Val] Valitskas, A.I., *Rational identities of radical algebras*, Soviet Math. (Izv. VUZ) **29** (1985), 88–99.

[Vas1] Vaserstein, L.N., *On the subgroup SL_2 over Dedekind rings of arithmetic type*, Math. USSR Sbornik **18** (1972), 321–332.

[Vas2] Vaserstein, L.N., *The structure of classic arithmetic group of rank greater than one*, Math. USSR Sbornik **20** (1973), 465–492.

[Wa] Walker, R., *Local rings and normalizing sets of elements*, Proc. London Math. Soc. **24** (1972), 27–45.

[Wal1] Wallace, D.A.R., *Group algebras with radicals of square zero*, Proc. Glasgow Math. Assoc. **5** (1962), 158–159.

[Wal2] Wallace, D.A.R., *The Jacobson radicals of the group algebra of a group and of certain normal subgroups*, Math. Z. **100** (1967), 282–294.

[Wal3] Wallace, D.A.R., *Some applications of subnormality in groups in the study of group algebras*, Math. Z. **108** (1968), 53–62.

[Wal4] Wallace, D.A.R., *On commutative and central conditions on the Jacobson radical of the group algebra of a group*, Proc. London Math. Soc. **19** (1969), 385–402.

[Wal5] Wallace, D.A.R., *The radical of the group algebra of a subgroup, of a polycyclic group and of a restricted SN-group*, Proc. Edinburgh Math. Soc. **17** (1970), 165–171.

[War] Ward, H.N., *Some results on the group algebra of a group over a prime field*, Harvard Seminar on Finite Groups (1960–1961).

[Warh] Warhurst, D.S., *Topics in group rings*, Thesis, Manchester (1981).

[We1] Weiss, A., *Idempotents in group rings*, J. Pure Appl. Algebra **16** (1980), 207–213.

[We2] Weiss, A., *Rigidity of p-adic torsion*, Ann. Math. **127** (1988), 317–332.

[We3] Weiss, A., *Torsion units in integral group rings*, J. Reine Angew. Math. **415** (1991), 175–187.

[Wh] Whitcomb, A., *The group ring problem*, Ph.D. Thesis, University of Chicago (1968).

[Wi] Wiechecki, L., *Finitely generated group ring units*, Proc. Amer. Math. Soc. **127** (1999), 51–55.

[Wo1] Woods, S.M., *On perfect group rings*, Proc. Amer. Math. Soc. **27** (1971), 49–52.

[Wo2] Woods, S.M., *Some results on semi-perfect group rings*, Canad. J. Math. **26** (1974), 121–129.

[Yu] Yunus, I.A., *A problem of Fox*, Dokl. Akad. Nauk SSSR **278** (1984), 53–56 (in Russian).

[Za1] Zalesskii, A.E., *On the semisimplicity of a crossed product*, Sibirsk. Mat. Zh. **6** (1965), 1009–1013 (in Russian).

[Za2] Zalesskii, A.E., *On group rings of solvable groups*, Izv. Akad. Nauk BSSR, Ser. Fiz.-Mat. (1970), 13–21 (in Russian).

[Za3] Zalesskii, A.E., *On group rings of linear groups*, Sib. Math. J. **12** (1971), 346–352.

[Za4] Zalesskii, A.E., *On a problem of Kaplansky*, Dokl. Akad. Nauk SSSR **203** (1972), 749–751 (in Russian); Soviet Math. Dokl. **13** (1972), 449–452.

[Za5] Zalesskii, A.E., *On the semisimplicity of a modular group algebra of a solvable group*, Soviet Math. **14** (1973), 101–105.

[Za6] Zalesskii, A.E., *The Jacobson radical of the group algebra of a solvable group is locally nilpotent*, Izv. Akad. Nauk SSSR, Ser. Mat. **38** (1974), 983–994 (in Russian).

[Za7] Zalesskii, A.E., *Intersection theorems in group rings*, No. 395-74, VINITI (1974) (in Russian). (This manuscript is kept in the depot of VINITI, Moscow, and is available upon request.)

[ZM] Zalesskii, A.E. and Mikhalev, A.V., *Group rings*, Itogi Nauki i Tekhniki (Sovremennye Problemy Matematiki), Vol. 2 (1973), 5–118; J. Soviet Math. **4** (1975), 1–78.

[ZS] Zalesskii, A.E. and Smirnov, M.B., *Lie algebra associated with linear group*, Comm. Algebra **9** (1981), 2075–2100.

[Zas1] Zassenhaus, H., *Ein Verfahren jeder endlichen p-Gruppe einen Lie-Ring mit der Charackteristik p zuzuordnen*, Abh. Math. Sem. Univ. Hamburg **13** (1940), 200–207.

[Zas2] Zassenhaus, H.J., *On units of orders*, J. Algebra **20** (1972), 368–395.

[Zas3] Zassenhaus, H., *On the torsion units of finite group rings*, Studies in Mathematics (in honor of A. Almeida Costa), Instituo de Alta Cultura, Lisbon (1974), 119–126.

[Zas4] Zassenhaus, H., *When is the unit group of a Dedekind order solvable?*, Comm. Algebra **6** (1978), 1621–1627.

This Page Intentionally Left Blank

Semiregular, Weakly Regular, and π-regular Rings

Askar Tuganbaev

Department of Mechanics and Mathematics, Moscow State University, Moscow, 119899 GSP, Russia
E-mail: askar@tuganbaev.mccme.ru

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

Preface

All rings are assumed to be associative and (except for nil-rings and for some stipulated cases) to have nonzero identity elements.

A ring A is said to be *regular* if for every element a of A, there is an element $b \in A$ such that $a = aba$. Regular rings are well investigated. For example, the books [53] and [94] are devoted to regular rings.

A ring A is called a *π-regular* ring if for every element $a \in A$, there is an element $b \in A$ such that $a^n = a^n b a^n$ for some positive integer n. A ring A is said to be *strongly π-regular* if for every $x \in A$, there is a positive integer m with $x^m \in x^{m+1}A$. It is proved in [43] that a ring A is strongly π-regular $\Longleftrightarrow$ for every $x \in A$, there is a positive integer n such that $x^n \in Ax^{n+1}$.

Every strongly π-regular ring is π-regular [7]. If F is a skew field and M is a right vector F-space with an infinite basis $\{e_i\}_{i=1}^{\infty}$, then the ring $\mathrm{End}(M_F)$ is a regular (and π-regular) ring which is not strongly π-regular. The factor ring of the integers with respect to the ideal generated by the integer 4 is a strongly π-regular ring which is not a regular ring.

We highlight [3,69,74,7,104] among the papers concerning π-regular and strongly π-regular rings. Also, see [1,5,10,11,21,31,33,43,54,61,62,64–66,75–78,85–87,96,101,107–109,115,119].

For a ring A, we denote by $J(A)$ the *Jacobson radical* of A which is equal to the intersection of all maximal right ideals of A by definition. A ring A is said to be *semiregular* if $A/J(A)$ is a regular ring, and all idempotents of $A/J(A)$ are images of idempotents of A for the natural epimorphism $A \to A/J(A)$. The classes of semiregular and π-regular rings are quite large. Every regular ring is a semiregular π-regular ring. In addition, all right or left Artinian rings are semiregular π-regular rings, and the endomorphism ring of any injective module is a semiregular ring.

Semiregular rings were considered in many sources. Here, we just highlight [8,82,95,109,114,116].

A ring A is said to be *right weakly regular* if $B^2 = B$ for every right ideal B of A. A ring A is called a *right weakly π-regular* ring if for every element $a \in A$, there is a positive integer $n = n(a)$ such that $a^n A = (a^n A)^2$. A ring A is said to be *simple* if every nonzero ideal of A is equal to A. A ring A is called a *biregular* ring if every principal ideal of A can be generated by a central idempotent of A. All simple or biregular rings are (right and left) weakly regular. Weakly regular rings were considered, for example, in [9,16–19,28,38,59,63,90,105].

If M and N are two modules, then $\mathrm{Hom}(M, N)$ denotes the Abelian group formed by all homomorphisms $M \to N$. For a module M, the direct product and the direct sum of I isomorphic copies of M are denoted by M^I and $M^{(I)}$, respectively. If X and Y are subsets of a right A-module M, then $(X : Y)$ denotes the subset $\{a \in A \mid xa \in Y \ \forall x \in X\}$ of A. If N is a submodule of a module M and $f : X \to M$ is a module homomorphism, then $f^{-1}(N)$ denotes the submodule $\{x \in X \mid f(x) \in N\}$ of X.

A submodule N of a module M is called a *maximal* submodule (in M) if M/N is a simple module (i.e. if N is the kernel of a nonzero homomorphism of M into a simple module). The set of all maximal submodules of a module M is denoted by $\max(M)$.

A submodule N of a module M is said to be *superfluous* (in the module M) if $N + M' \neq M$ for every proper submodule M' of M. We denote by $\mathrm{gs}(M)$ the set of all endomorphisms f of a module M such that $f(M)$ is a superfluous submodule in M.

A submodule N of a module M is *essential* (in M) if N has nonzero intersection with any nonzero submodule of M. In this case, we say that the module M is an *essential extension* of N.

A module M is said to be *simple* if M does not have nonzero proper submodules. A nonzero module M is said to be *semisimple* if M is a direct sum of simple modules.

The intersection of kernels of all homomorphisms from a module M into simple modules is called the *Jacobson radical* of M and is denoted by $J(M)$. The radical $J(M)$ coincides with the sum of all superfluous submodules of M. Note that either $J(M) = M$ (if $\max(M) = \emptyset$) or $J(M)$ coincides with the intersection of all maximal submodules of M (if $\max(M) \neq \emptyset$). A module M is said to be *semiprimitive* if $J(M) = 0$.

A ring is called a *prime* ring if the product of any two of its nonzero ideals is not equal to zero. A ring is said to be a *domain* if the product of any two its nonzero elements is not equal to zero.

An ideal P of a ring A is a *prime* ideal if the factor ring A/P is a prime ring (i.e. if $BC \not\subseteq P$ for any two of ideals B and C of the ring A which are not contained in P). A proper ideal B of a ring A is a *completely prime* ideal if A/B is a domain.

A prime (resp. completely prime) ideal P of a ring A which does not contain properly other prime ideals (resp. completely prime ideals) of A is called a *minimal prime* ideal (*minimal completely prime* ideal) of A.

If A is a ring and M is a subgroup of the additive group of A, then we say that idempotents of A *can be lifted modulo* M if for every element x of A with $x - x^2 \in M$, there exists an idempotent e of A with $e - x \in M$.

Let B be an ideal of a ring A, and let $\overline{A} = A/B$. The natural homomorphic image of an element a of A in the ring $\overline{A}$ is denoted by $\overline{a}$. We say that *idempotents of* $\overline{A}$ *can be lifted to idempotents of* A if the following three equivalent conditions hold. A ring A is *right primitive* if A has a maximal right ideal which does not contain any nonzero ideal of A. An ideal P of a ring A is a *right primitive* ideal if the ring A/P is right primitive.

A ring is *normal* if all its idempotents are central. A ring without nonzero nilpotent ideals is called a *semiprime* ring. A ring without nonzero nilpotent elements is called a *reduced* ring. (All reduced rings are semiprime and normal.)

1. Semiregular modules and rings

1.1. *Regular modules and rings.* A module M is said to be *regular* if every cyclic submodule of M is a direct summand of M.

A module M is regular $\Longleftrightarrow$

every finitely generated submodule of M is a direct summand of M $\Longleftrightarrow$

for every finitely generated submodule N of M, there is an idempotent endomorphism f of M such that $f(M) = N$.

A ring A is a *regular* ring if the following seven equivalent conditions hold.

(1) For every element $a \in A$, there is an element $b \in A$ with $a = aba$.

(2) Every principal right ideal of A is generated by an idempotent.
(3) Every principal left ideal of A is generated by an idempotent.
(4) Every finitely generated right ideal of A is generated by an idempotent.
(5) Every finitely generated left ideal of A is generated by an idempotent.
(6) A_A is a regular module.
(7) $_AA$ is a regular module.

1.2. *Abelian regular rings.* A ring A is called an *Abelian regular* ring if the following equivalent conditions hold.
(1) For every element $a \in A$, there is an element $b \in A$ with $a = a^2b$.
(2) For every element $a \in A$, there is an element $b \in A$ with $a = ba^2$.
(3) Every element of A is a product of a central idempotent and an invertible element.
(4) A is a regular reduced ring.
(5) A is a regular normal ring.
(6) A is a regular ring, and all maximal right ideals of A are ideals in A.
(7) A is a regular ring, and all right or left ideals of A are ideals in A.

1.3. Every cyclic right module over an Abelian regular ring is a regular module.

1.4. *Semiregular modules.* We say that a submodule N of a module M *lies above a direct summand* of M if there is a direct decomposition $M = P \oplus Q$ such that $P \subseteq N$ and $Q \cap N$ is a superfluous submodule of Q. In this case $Q \cap N$ is a superfluous submodule in M and $Q \cap N \subseteq J(M)$.

A module M is said to be *semiregular* if the following four equivalent conditions hold.
(1) Every cyclic submodule of M lies above a direct summand of M.
(2) Every finitely generated submodule of M lies above a direct summand of M.
(3) For every cyclic submodule N of M, there is an idempotent endomorphism f of M such that $f(M) \subseteq N$ and $(1 - f)(N)$ is a superfluous submodule of M.
(4) For every finitely generated submodule N of M, there is an idempotent endomorphism f of M such that $f(M) \subseteq N$, and $(1 - f)(N)$ is a superfluous submodule of M.

1.5. *Relatively projective and quasi-projective modules.* A module M is said to be *projective with respect to* a module N (or N*-projective*) if for every epimorphism $h : N \to \overline{N}$ and every homomorphism $\overline{f} : M \to \overline{N}$, there is a homomorphism $f : M \to N$ such that $\overline{f} = hf$.

A module which is projective with respect to itself is called a *quasi-projective* or *self-projective* module.

1.6. *Finitely supplemented and amply f-supplemented modules.* Let M be a module, and let U be a submodule of M. A submodule V of M is called a *supplement* of U in M if $U + V = M$ and $U \cap V$ is a superfluous submodule in V.

A module M is said to be *finitely supplemented* or f*-supplemented* if every submodule of M has a supplement in M.

A finitely supplemented module M is said to be *amply f-supplemented* if for any two submodules U, V of M such that $U+V=M$ and U is a finitely generated module, there is a supplement V' of U in M with $V' \subseteq V$.

(1) Every semiregular module is a finitely supplemented module.
(2) Every semiregular quasi-projective module is an amply f-supplemented module.
(3) Let M be a finitely generated quasi-projective module. Then
M is a semiregular module $\Longleftrightarrow$
M is an amply f-supplemented module.

1.7. Let M be a module such that $J(M)$ is a superfluous submodule in M. Then
M is a semiregular module $\Longleftrightarrow$
for every cyclic submodule N of M, there is an idempotent endomorphism f of M such that $f(M) \subseteq N$ and $(1-f)(N) \subseteq J(M) \Longleftrightarrow$
for every finitely generated submodule N of M, there is an idempotent endomorphism f of M such that $f(M) \subseteq N$ and $(1-f)(N) \subseteq J(M)$.

1.8. *Semiregular rings.* A ring A is *semiregular* if the following five equivalent conditions hold.

(1) $A/J(A)$ is a regular ring, and all idempotents of $A/J(A)$ can be lifted to idempotents of A.
(2) A_A is a semiregular module.
(3) ${}_AA$ is a semiregular module.
(4) A is a right amply f-supplemented ring.
(5) A is a left amply f-supplemented ring.

1.9. If M is a semiregular module, then $M/J(M)$ is a regular module.

1.10. For a module M, the following conditions are equivalent.

(1) M is a semiprimitive semiregular module.
(2) M is a regular module.
(3) Every submodule of M is a semiprimitive regular module.

1.11. Let M_A be a semiregular countably generated module such that $J(M)$ is a superfluous submodule in M.

Then M is a direct sum of a countable set of cyclic modules.

Consequently, every finitely generated semiregular module is a finite direct sum of cyclic modules.

1.12. Every countably generated regular module is a countable direct sum of cyclic modules.

1.13. *Free and projective modules.* A right A-module M is called a *free cyclic* module if $M_A \cong A_A$ (i.e. there is an element $m \in M$ such that $M = mA$ and $r(m)=0$). In this case, m is called a *free generator* of M.

A module M is said to be *free* if M is a direct sum of free cyclic modules M_i with free generators m_i $(i \in I)$. The set $\{m_i\}_{i \in I}$ is called a *basis* of M.

A right A-module M is said to be *projective* if the following equivalent conditions hold.

(1) M is projective with respect to every right A-module N.

(2) Every epimorphism $N \to M$ of right A-modules is split.

(3) M is a direct summand of a free module.

1.14. Let M be a quasi-projective module such that $J(M)$ is a superfluous submodule in M, and let $h : M \to M/J(M)$ be the natural epimorphism. Then

M is a semiregular module $\Longleftrightarrow$

$M/J(M)$ is a regular module, and for every direct decomposition $h(M) = \overline{X} \oplus \overline{Y}$ such that $\overline{X}$ is a finitely generated module, there is a direct decomposition $M = X \oplus Y$ such that $h(X) = \overline{X}$ and $h(Y) = \overline{Y}$.

1.15. Let M be a quasi-projective finitely generated module. Then

M is a semiregular module $\Longleftrightarrow$

$M/J(M)$ is a regular module, and for every direct decomposition $M/J(M) = \overline{X} \oplus \overline{Y}$, there is a direct decomposition $M = X \oplus Y$ such that the modules $\overline{X}$ and $\overline{Y}$ are natural images of X and Y, respectively.

1.16. Let M be a regular quasi-projective module, Q be a direct summand of M, and let N be a finitely generated submodule in M.

Then $Q + N$ is a direct summand of M, and there is a submodule U of N such that $Q + N = Q \oplus U$.

1.17. Let M be a right module over a ring A which is projective with respect to the module A_A, and suppose that all idempotents of $A/J(A)$ can be lifted to idempotents of A. Then

M is a semiregular module $\Longleftrightarrow$

$M/J(M)$ is a regular module.

1.18. (1) Let M be a regular module which is a direct summand of a direct sum of countably generated modules.

Then M is a direct sum of cyclic regular modules.

(2) Let M be a regular projective right module over a ring A.

Then M is isomorphic to a direct sum of cyclic regular direct summands of A_A, and every countably generated submodule K of M is projective.

1.19. (1) Let M be a direct summand of a direct sum of projective regular right A-modules M_i $(i \in I)$.

Then M is a projective regular module which is isomorphic to a direct sum of cyclic regular direct summands of A_A.

(2) Let M be a projective right module over a regular ring A.

Then M is a regular module which is isomorphic to a direct sum of cyclic regular direct summands of A_A.

In addition, every countably generated submodule of M is projective.

1.20. Let M_A be a semiregular module such that $J(M)$ is a superfluous submodule in M.

(1) If M is a direct summand of a direct sum of countably generated modules, then M is a direct sum of cyclic modules.

(2) If M is a projective module, then M is a direct sum of cyclic modules.

1.21. Let A be a ring. Assume that for every set $\{a_i\}_{i\in I}$ of elements of A and each set $\{U_i\}_{i\in I}$ of finitely generated left ideals of A such that the intersection of every finite subset of the set $\{a_i+U_i\}_{i\in I}$ is non-empty, the intersection of all sets a_i+U_i is non-empty.

Then A is a semiregular ring.

1.22. *Pure-injective modules.* A submodule V of a right module U over a ring S is called a *pure* submodule in the module U_S if for every left S-module M, a natural group homomorphism $V\otimes_S M\to U\otimes_S M$ is a monomorphism. (In this case, we can consider $V\otimes_S M$ as a subgroup in $U\otimes_S M$.)

A right module X over a ring A is said to be *pure-injective* if for every module M_A and each pure submodule N of M, all homomorphisms $N\to X$ can be extended to homomorphisms $M\to X$. (This means that the natural homomorphism $\mathrm{Hom}(M_A, X_A)\to \mathrm{Hom}(N_A, X_A)$ is an epimorphism.)

1.23 [122]. If M is a pure-injective module, then the ring $\mathrm{End}(M)$ is semiregular.

1.24. *Continuous modules.* A submodule H of a module M is a *closed* submodule (in M) if H has no proper essential extensions in the module M.

A module M is said to be *continuous* if every submodule of M which is isomorphic to a closed submodule of M is a direct summand of M.

For a right A-module M, the set of all endomorphisms φ of M_A such that $\mathrm{Ker}(\varphi)$ is an essential submodule in M is denoted by $\mathrm{sg}(M)$. The set $\mathrm{sg}(M)$ is an ideal of the ring $\mathrm{End}(M)$.

1.25. For a continuous module M, the following assertions hold.

(1) $\mathrm{End}(M)/\mathrm{sg}(M)$ is a regular ring, and $J(\mathrm{End}(M))=\mathrm{sg}(M)$.

(2) Every countable set $\{\bar{e}_i\}_{i=1}^{\infty}$ of orthogonal idempotents of $\mathrm{End}(M)/\mathrm{sg}(M)$ can be lifted to a countable set $\{e_i\}_{i=1}^{\infty}$ of orthogonal idempotents of $\mathrm{End}(M)$.

(3) $\mathrm{End}(M)$ is a semiregular ring.

1.26. *Skew power series rings.* Let φ be an injective endomorphism of a ring A. We denote by $A_\ell[\![x,\varphi]\!]$ the *left skew* (*power*) *series ring* consisting of formal series $\sum_{i=0}^{\infty} a_i x^i$ of an indeterminate x with canonical coefficients $a_i\in A$, where addition is defined naturally (i.e. componentwise) and multiplication is defined by the rule $x^i a=\varphi^i(a)x^i$. The *right skew* (*power*) *series ring* $A_r[\![x,\varphi]\!]$ consists of series $\sum_{i=0}^{\infty} x^i a_i$, and multiplication is defined with the use of the rule $ax^i=x^i\varphi^i(a)$. For any two series $f\in A_\ell[\![x,\varphi]\!]$ and

$g \in A_r[\![x,\varphi]\!]$, we denote by f_i and g_i the coefficient $a_i \in A$ of x^i in the canonical forms of f and g, respectively.

Let φ be an injective endomorphism of the ring A.

(1) $A_\ell[\![x,\varphi]\!]$ is a domain $\Longleftrightarrow$ A is a domain.

(2) All idempotents of $A/J(A)$ can be lifted to idempotents of A $\Longleftrightarrow$ all idempotents of $A_\ell[\![x,\varphi]\!]/J(A_\ell[\![x,\varphi]\!])$ can be lifted to idempotents of $A_\ell[\![x,\varphi]\!]$.

(3) $A_\ell[\![x,\varphi]\!]$ is a local ring $\Longleftrightarrow$ A is a local ring.

(4) $A_\ell[\![x,\varphi]\!]$ is a semiregular ring $\Longleftrightarrow$ A is a semiregular ring.

1.27. *Exchange rings and I_0-rings.* A ring A is called an *exchange ring* if the following two equivalent conditions hold.

(i) For any two elements $a, b \in A$ with $a+b=1$, there are idempotents $e \in aA$ and $f \in bA$ with $e+f=1$.

(ii) For any two elements $a, b \in A$ with $a+b=1$, there are idempotents $e' \in Aa$ and $f' \in Aa$ with $e'+f'=1$.

A ring A is called an *I_0-ring* if the following three equivalent conditions hold.

(1) Every right ideal of A which is not contained in $J(A)$ contains a nonzero idempotent.

(2) Every left ideal of A which is not contained in $J(A)$ contains a nonzero idempotent.

(3) For every element $a \in A \setminus J(A)$, there is a nonzero element $x \in A$ with $x = xax$.

1.28. Every semiregular ring is an exchange ring, and every exchange ring is an I_0-ring.

1.29. Let B be a unitary subring of a ring A, $\{A_i\}_{i=1}^{\infty}$ be a countable set of copies of A, D be the direct product of all rings A_i, and let R be the subring in D generated by the ideal $\bigoplus_{i=1}^{\infty} A_i$ and by the subring $B' \equiv \{(b,b,b,\dots) \mid b \in B\}$.

(1) R is a regular ring $\Longleftrightarrow$ A and B are regular rings.

(2) R is an exchange ring $\Longleftrightarrow$ A and B are exchange rings.

(3) R is an I_0-ring $\Longleftrightarrow$ A is an I_0-ring.

(4) If A is the field of rational numbers and B is the ring of all rational numbers with odd denominators, then R is a commutative reduced semiprimitive exchange ring which is not a semiregular ring.

(5) If A is the field of rational numbers and B be the ring of integers, then the ring R is a commutative reduced semiprimitive I_0-ring which is not an exchange ring.

1.30. *Semiperfect rings.* A ring A is said to be *semiperfect* if $A/J(A)$ is a semisimple ring, and all idempotents of $A/J(A)$ can be lifted to idempotents of A.

A ring A is said to be *semilocal* if $A/J(A)$ is a semisimple ring. (Consequently, every semiperfect ring is semilocal.)

A ring A is said to be *orthogonally finite* if A does not contain infinite sets of nonzero orthogonal idempotents.

For a ring A, the following conditions are equivalent.

(1) A is a semiperfect ring.

(2) A is an orthogonally finite semiregular ring.

(3) A is an orthogonally finite exchange ring.

(4) A is an orthogonally finite I_0-ring.
(5) For every right ideal L of A, there are an idempotent $e \in L$ and a right ideal M of A such that $M \subseteq J(A)$ and $L = eA + M$.
(6) A is a semilocal I_0-ring.
(7) All factor rings of A are orthogonally finite semiperfect exchange rings.

2. Weakly π-regular and weakly regular rings

2.1. *Weakly π-regular rings.* A ring A is said to be *right weakly π-regular* if the following equivalent conditions hold.

(i) For every element $a \in A$, there is a positive integer $n = n(a)$ such that $a^nA = (a^nA)^2$.
(ii) For every element $a \in A$, there is a positive integer n such that $a^n = a^nb$, where $b \in Aa^nA$.
(iii) For every element $a \in A$, there is a positive integer n such that $a^n(1-b) = 0$, where $b \in Aa^nA$.

2.2. If M is a right (resp. left) A-module, then the set of all elements $m \in M$ such that $r(m)$ (resp. $\ell(m)$) is an essential right (resp. left) ideal of A is denoted by $\mathrm{Sing}(M)$.

A module M over a ring A is said to be *nonsingular* if $\mathrm{Sing}(M) = 0$.

The set $\mathrm{Sing}(M)$ is a fully invariant submodule in M. Therefore $\mathrm{Sing}(A_A)$ and $\mathrm{Sing}({}_AA)$ are ideals in A for every ring A.

2.3. For a right weakly π-regular ring A, the following assertions hold.

(1) If a is a right regular element of A, then $AaA = A$.
(2) If A is a domain, then A is a simple ring.
(3) If A is a reduced ring, then A is a left weakly π-regular ring.
(4) $J(A)$ is a nil-ideal of A.
(5) $\mathrm{Sing}({}_AA)$ is a nil-ideal in A.

2.4. *Rings of bounded index.* A positive integer n is called the *index* of a nilpotent element a of a ring A if $a^n = 0$ and $a^{n-1} \neq 0$.

A ring A is said to be a *ring of index at most n* if there is a positive integer n such that $a^n = 0$ for every nilpotent element $a \in A$. The least among these numbers n is the *index* of A and is denoted by $i(A)$.

A ring A is called a ring *of bounded index* (*of nilpotence*) if A is a ring of index at most n for some positive integer n.

2.5. If A is a right weakly π-regular ring of index at most n, then every prime factor ring A/P of A is a simple ring of index at most n.

Therefore, all prime ideals of A are maximal ideals and minimal prime ideals.

2.6. *Right weakly regular rings.* A ring A is a *right weakly regular* ring if the following nine equivalent conditions hold.

(1) $B^2 = B$ for every right ideal B of A.
(2) $B^2 = B$ for every principal right ideal B of A.
(3) For every element $a \in A$, there are elements $x_1, \dots, x_n, y_1, \dots, y_n$ of A such that $a = \sum_{i=1}^n ax_iay_i$.
(4) For every element $a \in A$, there are elements $x_1, \dots, x_n, y_1, \dots, y_n$ of A such that $a(1 - \sum_{i=1}^n x_iay_i) = 0$.
(5) $BC = B$ for any two right ideals B and C of A such that $B \subseteq C$.
(6) $B \cap D \subseteq BD$ for any two right ideals B and D of A.
(7) $B \cap D = BD$ for every right ideal B of A and for every ideal D of A.
(8) $aA \cap D \subseteq aD$ for every element $a \in A$ and for every ideal D of A.
(9) $d \in dD$ for every ideal D of A and for every element d of the ideal D.

2.7. For a right weakly regular ring A, the following assertions hold.
(1) A is a semiprimitive ring. In particular, A is a semiprime ring.
(2) The center C of A is a commutative regular ring.
(3) A is a left nonsingular ring.
(4) If a is a right regular element of A, then $AaA = A$.
(5) Assume that A is an indecomposable ring and either every essential left ideal of A contains a regular element or every essential right ideal of A contains a regular element.
Then A is a simple ring.

2.8. A ring A is regular $\Longleftrightarrow$
A is right weakly regular, and all prime quotient rings of A are regular $\Longleftrightarrow$
A is a left weakly regular ring, and all prime quotient rings of A are regular.

2.9. *Biregular rings.* A ring is said to be *biregular* if every 1-generated two-sided ideal of it is generated by a central idempotent.
(1) Every simple ring A is a weakly regular ring.
(2) If every right ideal of a ring A is generated by idempotents, then A is a weakly regular ring.
(3) Every biregular ring A is a weakly regular ring.
(4) A is a biregular ring without nontrivial central idempotents $\Longleftrightarrow$ A is a simple ring.
(5) A is a biregular ring without infinite sets of nontrivial central orthogonal idempotents $\Longleftrightarrow$ A is a finite direct product of simple rings.

2.10. For a reduced ring A, the following conditions are equivalent.
(1) A is a right weakly π-regular ring.
(2) A is a left weakly π-regular ring.
(3) Every prime ideal of A is a maximal ideal.
(4) Every prime factor ring of A is a simple domain.
(5) A is a biregular ring.
(6) A is a right weakly regular ring.
(7) A is a left weakly regular ring.

2.11. (1) A is a reduced indecomposable right weakly π-regular ring $\Longleftrightarrow$
A is a reduced indecomposable left weakly π-regular ring $\Longleftrightarrow$
A is a simple domain.

(2) A is a reduced right weakly π-regular ring which does not contain infinite sets of nontrivial central orthogonal idempotents $\Longleftrightarrow$
A is a reduced left weakly π-regular ring which does not contain infinite sets of nontrivial central orthogonal idempotents $\Longleftrightarrow$
A is a finite direct product of simple domains.

2.12. *The prime radical.* For a ring A, the intersection $P(A)$ of all prime ideals of A is called the *prime radical* of A.

The prime radical $P(A)$ contains all nilpotent ideals of A and is the least semiprime ideal of A.

In addition, $P(A)$ coincides with the intersection of all semiprime ideals of A and is equal to the intersection of all minimal prime ideals of A.

2.13. Let A be a ring, and let $P(A)$ be the prime radical of A. Then
$A/P(A)$ is a reduced right weakly regular ring $\Longleftrightarrow$
$A/P(A)$ is a reduced right weakly π-regular ring $\Longleftrightarrow$
all prime quotient rings of A are simple domains.

2.14. A module M is said to be *finite-dimensional* (in the sense of Goldie) if M does not contain infinite direct sums of nonzero submodules.

(1) A is a right weakly regular left finite-dimensional ring $\Longleftrightarrow$
A is a finite direct product of simple left finite-dimensional rings.

(2) A is a right weakly regular right finite-dimensional right nonsingular ring $\Longleftrightarrow$
A is a finite direct product of simple right finite-dimensional rings.

2.15. Assume that A is a ring such that $(X + Z) \cap (Y + Z) = (X \cap Y) + Z$ for every right ideal X, for each left ideal Y, and for every ideal Z of A. Then
A is a weakly regular ring $\Longleftrightarrow$
A is a semiprime ring, every prime factor ring of A is a weakly regular ring, and the union of every chain of semiprime ideals of A is a semiprime ideal.

2.16. For a ring A, the following conditions are equivalent.

(1) A is an Abelian regular ring.
(2) A is a semiprime right weakly π-regular right ring of bounded index, and all maximal right ideals of A are ideals in A.
(3) A is a semiprime right weakly π-regular ring of bounded index, and all maximal left ideals of A are ideals in A.

2.17. *Injective and quasi-injective modules.* Let N be a module. A module M is said to be *injective with respect to* N (or N-*injective*) if for every submodule $\overline{N}$ of N, all homomorphisms $\overline{N} \to M$ can be extended to homomorphisms $N \to M$.

An A-module M is called an *injective* module if for every right A-module N, the module M is N-injective.

A module M is said to be *quasi-injective* if M is an M-injective module.

For every quasi-injective module M, the ring $\mathrm{End}(M)$ is semiregular.

2.18. If A is a right weakly regular ring, then every injective right module over any factor ring A/B of A is an injective A-module.

2.19. A module M is said to be *uniform* if any two of nonzero submodules of M have nonzero intersection.

Every uniform module over an Abelian regular ring is an injective simple module.

2.20. *Pierce stalks.* Let A be a ring, and let $S(A)$ be the non-empty set of all proper ideals of A generated by central idempotents.

An ideal $P \in S(A)$ is a *Pierce ideal* of A if P is a maximal (with respect to inclusion) element of the set $S(A)$.

If P is a Pierce ideal of A, then the factor ring A/P is called a *Pierce stalk* of A.

2.21. (1) A is a biregular ring $\Longleftrightarrow$
all Pierce stalks of A are simple rings.

(2) A is a regular ring $\Longleftrightarrow$
all Pierce stalks of A are regular rings $\Longleftrightarrow$
all indecomposable factor rings of A are regular rings.

(3) A is an Abelian regular ring $\Longleftrightarrow$
all Pierce stalks of A are Abelian regular rings $\Longleftrightarrow$
all Pierce stalks of A are skew fields.

3. Strongly π-regular and π-regular rings

3.1. *Strongly π-regular rings.* A ring A is said to be *strongly π-regular* if the following five equivalent conditions hold.

(1) For every element $x \in A$, there is an element y of A such that $x^m = x^{m+1}y$ for some positive integer m.

(2) For every element $x \in A$, there is an element z of A such that $x^n = zx^{n+1}$ for some positive integer n.

(3) For every element $x \in A$, the descending chain of principal right ideals $xA \supseteq x^2A \supseteq \cdots \supseteq x^nA \supseteq \cdots$ of A eventually terminates.

(4) For every element $x \in A$, the descending chain of principal left ideals $Ax \supseteq Ax^2 \supseteq \cdots \supseteq Ax^n \supseteq \cdots$ of A eventually terminates.

(5) For every element $x \in A$, there is a positive integer t such that $x^t \in x^{t+j}A \cap Ax^{t+j}$ for all positive integers j.

3.2. *Zorn rings and π-regular rings.* A ring A is called a *Zorn* ring if A is an I_0-ring, and $J(A)$ is a nil-ideal.

A ring A is said to be π-*regular* if the following three equivalent conditions hold.
(1) For every element $a \in A$, there is an element b of A such that $a^n = a^n b a^n$ for some positive integer n.
(2) For every element $a \in A$, there is a positive integer n such that the principal right ideal $a^n A$ is generated by an idempotent.
(3) For every element $a \in A$, there is a positive integer n such that the principal left ideal Aa^n is generated by an idempotent.

3.3. (1) Every regular or strongly π-regular rings is π-regular.
(2) Every π-regular ring is an exchange ring.
(3) A is a Zorn ring $\Longleftrightarrow$ A is an I_0-ring, and $J(A)$ is a nil-ideal.
(4) Every π-regular ring A is a Zorn ring.
In particular, $J(A)$ is a nil-ideal.

3.4. Let B be a unitary subring of a ring A, $\{A_i\}_{i=1}^{\infty}$ be a countable set of copies of A, D be the direct product of all rings A_i, and let R be the subring in D generated by the ideal $\bigoplus_{i=1}^{\infty} A_i$ and by the subring $\{(b, b, b, \ldots) \mid b \in B\}$. Then
(1) R is a strongly π-regular (resp. π-regular) ring $\Longleftrightarrow$
A and B are strongly π-regular (resp. π-regular) rings.
(2) Let F be a field, A be the ring of all 2×2 matrices over F, and let B be the ring of upper triangular 2×2 matrices over F, $\{A_i\}_{i=1}^{\infty}$ be a countable set of copies of A.
Then R is a strongly π-regular semiprimitive ring which is not a right or left weakly regular ring.
Consequently, R is not a semiregular ring.
In addition, R has a non-semiprime quotient ring and satisfies the polynomial identity $[[X_1, X_2]^2, X_3] = 0$ (where $[x, y]$ denotes the commutator $xy - yx$).

3.5. For a module M, the following assertions hold.
(1) $\mathrm{End}(M)$ is a strongly π-regular ring $\Longleftrightarrow$
for every endomorphism f of M, there is a positive integer n such that $M = \mathrm{Ker}(f^n) \oplus f^n(M)$.
(2) If $\mathrm{End}(M)$ is a strongly π-regular ring, then every injective or surjective endomorphism of M is an automorphism.

3.6. For a ring A, the following eight conditions are equivalent.
(1) A is a strongly π-regular ring.
(2) The quotient ring A/N with respect to the prime radical N of A is a strongly π-regular ring.
(3) Every prime quotient ring of A is a strongly π-regular ring.
(4) Every injective endomorphism of every cyclic right A-module is an automorphism.
(5) For every element $a \in A$, there is a positive integer n such that $A_A = r(a^n) \oplus a^n A$.
(6) For every element $a \in A$, there is a positive integer n such that ${}_A A = Aa^n \oplus \ell(a^n)$.
(7) For every element $a \in A$, there is an element $x \in A$ such that $ax = xa$ and $a^n = a^{n+1}x = xa^{n+1}$.

(8) For every element $a \in A$, there is a positive integer n such that the system of equations $\{ay = ya, a^n = a^{2n}y, y^2a^n = y\}$ has exactly one solution.

3.7. For a ring A, the following conditions are equivalent.

(1) All matrix rings A_n are strongly π-regular rings.

(2) For every positive integer n, all injective endomorphisms of every cyclic right A_n-module are automorphisms.

(3) All injective endomorphisms of every finitely generated right A-module are automorphisms.

(4) All injective endomorphisms of every finitely generated left A-module are automorphisms.

3.8. Let R be a ring, $\{A_i\}_{i \in I}$ be a set of strongly π-regular unitary subrings of R, and let $A \equiv \bigcap_{i \in I} A_i$.

Then A is a strongly π-regular ring.

3.9. For a ring A, the following five conditions are equivalent.

(1) A is an Abelian regular ring.

(2) A is a reduced ring, and for every completely prime ideal Q of A, the quotient ring A/Q is a π-regular ring.

(3) A is a strongly π-regular reduced ring.

(4) A is a π-regular reduced ring.

(5) A is a reduced ring, every regular element of A is an invertible element, and for every element $a \in A$, there is an idempotent $e \in A$ with $r(a) = eA$.

3.10. *Properties of the center.* Let A be a ring, and let C be the center of A.

(1) If A is a π-regular prime ring, then C is a field.

(2) Assume that for every element $a \in A$, there is a polynomial $f \in C[x]$ such that $f(a) = 0$ and the coefficient of the lowest term of the polynomial $f(x)$ is invertible in the ring A.

Then A is a strongly π-regular ring.

(3) Assume that A is a semiprime ring and for every element $c \in C$, there is a positive integer $m = m(c)$ such that $c^m \in c^{m+1}A$.

Then C is a commutative regular ring.

(7) If A is a semiprime ring and the center of every prime factor ring of A is a field, then C is a commutative regular ring.

(8) If A is a π-regular semiprime ring, then C is a commutative regular ring.

(9) If A is a right weakly π-regular ring, then C is a strongly π-regular commutative ring, and C/P is a field for every prime ideal P of C.

3.11. Let A be a strongly π-regular ring, and let n be a positive integer. Then the following five conditions are equivalent.

(1) $r(a^n) = r(a^{n+1})$ for every element $a \in A$.

(2) $\ell(a^n) = \ell(a^{n+1})$ for every element $a \in A$.

(3) $a^nA = a^{n+1}A$ for every element $a \in A$.

(4) $Aa^n = Aa^{n+1}$ for every element $a \in A$.
(5) A is a ring of index at most n.

3.12. Let A be a π-regular ring of index at most n, and let N be the prime radical of A.
(1) $J(A) = N$, A/N is a semiprimitive ring of index at most n, and all idempotents of the factor ring $A/J(A)$ can be lifted to idempotents of A.
(2) If A is a semiprime ring, then A is a semiprimitive ring.
(3) If A is a semiprime orthogonally finite ring, then A is a semisimple ring.
(4) If Q is a prime ideal of A, then A/Q is a simple Artinian ring of index at most n.
(5) If $A/J(A)$ is an orthogonally finite ring, then A is a semiperfect ring.

3.13. Let A be a ring, N be the prime radical of A, n be a positive integer, and let A/N be a ring of index at most n. Then the following six conditions are equivalent.
(1) A is a π-regular ring.
(2) A is a strongly π-regular ring.
(3) All prime factor rings of A are π-regular rings of bounded index.
(4) All prime factor rings of A are simple Artinian rings of index at most n.
(5) $a^nA + N = a^{n+1}A + N$ for every element $a \in A$.
(6) $Aa^n + N = Aa^{n+1} + N$ for every element $a \in A$.

3.14. Let A be a ring of index at most n. Then the following conditions are equivalent.
(1) A is a π-regular ring.
(2) A is a strongly π-regular ring.
(3) Every prime factor ring of A is a π-regular ring of bounded index.
(4) All prime factor rings of A are simple Artinian rings of index at most n.
(5) $a^nA + N = a^{n+1}A + N$ for every element $a \in A$.
(6) $Aa^n + N = Aa^{n+1} + N$ for every element $a \in A$.
(7) $a^nA = a^{n+1}A$ for every element $a \in A$.
(8) $Aa^n = Aa^{n+1}$ for every element $a \in A$.

3.15. A ring A is said to be a *ring of stable range 1* if for any four elements f, v, x, and y of A with $fx + vy = 1$, there is an element h of A such that $f + vh$ is an invertible element of A.

Every strongly π-regular ring is a ring of stable range 1.

3.16. A is a π-regular ring, and every right primitive factor ring of A is a ring of bounded index $\Longleftrightarrow$

A is a strongly π-regular ring, and every indecomposable semiprimitive factor ring $\overline{A}$ of A is a simple Artinian ring.

3.17. Let A be a π-regular ring. For every element $a \in A$, denote by $h(a)$ the least positive integer n such that $a^n \in a^nAa^n$. The following three conditions are equivalent.
(1) All right primitive factor rings of A are rings of bounded index.

(2) For every countable set $\{e_i\}_{i=1}^{\infty}$ of orthogonal idempotents of A and for every countable set $\{x_i\}_{i=1}^{\infty}$ of elements of A, there is a positive integer n such that $h(\cdots h(h(e_1x_1)e_2x_2)\cdots e_nx_n)=0$.

(3) For every π-regular subring S of A, all indecomposable semiprimitive quotient rings of S are Artinian rings.

3.18. *Distributive, distributively generated and, semidistributive modules.* A module M is called a *distributive* module if $F\cap(G+H)=F\cap G+F\cap H$ for any three submodules F, G, and H of M.

A module is said to be *distributively generated* if it is generated by distributive submodules.

A direct sum of distributive modules is called a *semidistributive* module.

(1) If A is a right distributively generated Zorn ring, then $J(A)$ coincides with the prime radical of A.

(2) If A is a semiprime right distributively generated Zorn ring, then A is a semiprimitive ring.

(3) Let n be a positive integer, and let A be a Zorn prime ring which is a sum of n distributive right ideals.

Then A is a simple Artinian ring of index at most n which is isomorphic to a ring of all $k\times k$ matrices over a skew field with $k\leqslant n$.

3.19. Let A be a π-regular semiprime ring which is a sum of n distributive right ideals.

Then A is a strongly π-regular semiprimitive ring of index at most n, every prime factor ring of A is isomorphic to a ring of all $k\times k$ matrices over a skew field with $k\leqslant n$, and $a^nA=a^{n+1}A$, $Aa^n=Aa^{n+1}$ for every element $a\in A$.

3.20. Let A be a π-regular right distributively generated ring, and let N be the prime radical of A.

Then A is a strongly π-regular ring, and A/N is a semiprimitive ring of bounded index.

A set $\{e_{ij}\}_{i,j=1}^{n}$ of nonzero elements of a ring A is called a *system of n^2 matrix units* if $e_{ij}e_{st}=\delta_{js}e_{it}$, where $\delta_{jj}=1$ and $\delta_{js}=0$ for $j\neq s$.

3.21 [76]. A is a π-regular ring, and all right primitive quotient rings of A are Artinian rings $\Longleftrightarrow$

there does not exist an infinite sequence of systems of matrix units in A $\{e_{ij}^{(k)}\}_{i,j=1}^{n(k)}$ $(k=1,2,\dots)$ such that $n(1)<n(2)<\cdots$ and $e_{11}^{(k+1)}\in e_{11}^{(k)}Ae_{11}^{(k)}$.

3.22 [21]. For a strongly π-regular ring A, the following assertions hold.

(1) If B is a commutative ring and $f:B\to A$ is a ring homomorphism, then there is a commutative strongly π-regular subring of A containing $f(B)$.

(2) Let $A=\sum_{i=1}^{n}a_iA$, where all elements a_i commute with each other. Then $A=\sum_{i=1}^{n}Aa_i$, and there are elements $b_1,\dots,b_n\in A$ which commute with each other and with the a_i such that $\sum_{i=1}^{n}a_ib_i=1$.

(3) If 2 is invertible in the ring A, then every element $a\in A$ can be expressed as a sum of two units which commute with each other and with a.

(4) Every element $a \in A$ can be expressed as a sum of a unit and an idempotent which commute with each other and with a.

3.23. Let $A = \begin{pmatrix} F & 0 \\ F(x) & F(x) \end{pmatrix}$, where F is a field, x is an indeterminate, and $F(x)$ is the field of rational functions.

(i) All matrix rings A_n are strongly π-regular, and hence all injective endomorphisms of finitely generated right or left A-modules are automorphisms.

(ii) All surjective endomorphisms of finitely generated left A-modules are automorphisms.

(iii) Not all surjective endomorphisms of cyclic right A-modules are automorphisms.

3.24. *Denominator sets and rings of quotients.* Let T be a multiplicative subset of a ring A such that $1 \in T$, $0 \notin T$, and the product of any two elements of the set T belongs to T. The set T is a *right denominator set* in the ring A if there are a ring A_T and a ring homomorphism $f_T \equiv f : A \to A_T$ such that $f(T) \subseteq U(A_T)$, $A_T = \{f(a)f(t)^{-1} \mid a \in A,\ t \in T\}$, and $\mathrm{Ker}(f)$ coincides with the set $\{a \in A \mid at = 0$ for some element $t \in T\}$. In this case, the ring A_T is called the *right ring of quotients* for A with respect to T, and f_T is said to be the *canonical homomorphism* for A_T.

Let T be a right denominator set in a ring A.

If A is a π-regular (resp. regular, strongly π-regular) ring, then A_T is a π-regular (resp. regular, strongly π-regular) ring, and the canonical homomorphism $f : A \to A_T$ is surjective.

3.25. *Rings of quotients with respect to maximal ideals of a central subring.* Let R be a unitary subring of the center of A, and let $\max(R)$ be the set of all maximal ideals of R. For every $M \in \max(R)$, denote by A_M the ring of quotients of A with respect to the central multiplicative subset $R \setminus M$ of A.

(1) A is a regular ring $\Longleftrightarrow$
$A_{T(M)}$ is a regular ring for every maximal ideal M of R.

(2) A is a regular ring, and $A_{T(M)}$ is an orthogonally finite ring for every maximal ideal M of R $\Longleftrightarrow$
$A_{T(M)}$ is a semisimple ring for every maximal ideal M of R.

(3) Let A be a prime ring. Then
A is a strongly π-regular ring $\Longleftrightarrow$
$A_{T(M)}$ is a strongly π-regular ring for every maximal ideal M of R.

3.26. Let A be a ring, and let R be a unitary subring of the center of A.

(1) A is a right weakly regular ring $\Longleftrightarrow$
$A_{T(M)}$ is a right weakly regular ring for every maximal ideal M of R.

(2) A is a strongly π-regular ring $\Longleftrightarrow$
$A_{T(M)}$ is a strongly π-regular ring for every maximal ideal M of R.

3.27 [24]. A is a normal strongly π-regular ring $\Longleftrightarrow$
A is a normal ring, and all factor rings of A are Zorn rings $\Longleftrightarrow$
all Pierce stalks of A are local strongly π-regular rings.

References

[1] P. Ara, *Strongly π-regular rings have stable range one*, Proc. Amer. Math. Soc. **124** (11) (1996), 3293–3298.

[2] P. Ara, K.C. O'Meara and D.V. Tjukavkin, *Cancellation of projective modules over regular rings with comparability*, J. Pure Appl. Algebra **107** (1) (1996), 19–38.

[3] R. Arens and I. Kaplansky, *Topological representation of algebras*, Trans. Amer. Math. Soc. **63** (1948), 457–481.

[4] E.P. Armendariz and J.W. Fisher, *Regular P.I.-rings*, Proc. Amer. Math. Soc. **39** (2) (1973), 247–251.

[5] E.P. Armendariz, J.W. Fisher and R.L. Snider, *On injective and surjective endomorphisms of finitely generated modules*, Comm. Algebra **6** (7) (1978), 659–672.

[6] E.P. Armendariz, J.W. Fisher and S.A. Steinberg, *Central localization of regular rings*, Proc. Amer. Math. Soc. **46** (3) (1974), 315–321.

[7] G. Azumaya, *Strongly π-regular rings*, J. Fac. Sci. Hokkaido Univ. **13** (1954), 659–672.

[8] G. Azumaya, *F-semi-perfect modules*, J. Algebra **136** (1991), 73–85.

[9] G. Baccella, *On flat factor rings and fully right idempotent rings*, Ann. Univ. Ferrara Sez. VII **26** (1980), 125–141.

[10] A. Badawi, *On semicommutative π-regular rings*, Comm. Algebra **22** (1) (1993), 151–157.

[11] A. Badawi, *On Abelian π-regular rings*, Comm. Algebra **25** (4) (1997), 1009–1021.

[12] K.I. Beidar and A.V. Mikhalev, *Semiprime rings with bounded index of nilpotent elements*, Trudy Sem. Petrovsk., No. 13 (1988), 237–249.

[13] K.I. Beidar and R. Wisbauer, *Strongly and properly semiprime rings and modules*, Ring Theory, Proc. Ohio State – Denison Conf., 1992, World Sci. Pub., Singapore (1993).

[14] K.I. Beidar and R. Wisbauer, *Properly semiprime self-pp-modules*, Comm. Algebra **23** (3) (1995), 841–861.

[15] L.P. Belluce, *Spectral spaces and non-commutative rings*, Comm. Algebra **19** (1991), 1855–1865.

[16] G.F. Birkenmeier, *A survey of regularity conditions and the simplicity of prime factor rings*, Vietnam J. Math. **23** (1) (1995), 29–38.

[17] G.F. Birkenmeier, J.Y. Kim and J.K. Park, *A connection between weak regularity and the simplicity of prime factor rings*, Proc. Amer. Math. Soc. **122** (1) (1994), 53–58.

[18] G.F. Birkenmeier, J.Y. Kim and J.K. Park, *Regularity conditions and the simplicity of prime factor rings*, J. Pure Appl. Algebra **115** (1997), 213–230.

[19] W.D. Blair and H. Tsutsui, *Fully prime rings*, Comm. Algebra **22** (13) (1994), 5389–5400.

[20] W.D. Burgess, *Confirmation of a conjecture of Dauns and Hofmann on biregular rings*, Abelian Groups and Modules, A. Facchini and C. Menini, eds, Kluwer Acad. Publishers, Dordrecht (1995), 67–72.

[21] W.D. Burgess and P. Menal, *On strongly π-regular rings and homomorphisms into them*, Comm. Algebra **16** (1988) 1701–1725.

[22] W.D. Burgess and W. Stephenson, *Pierce sheaves of non-commutative rings*, Comm. Algebra **4** (1) (1976), 51–75.

[23] W.D. Burgess and W. Stephenson, *An analogue of the Pierce sheaf for non-commutative rings*, Comm. Algebra **6** (9) (1978), 863–886.

[24] W.D. Burgess and W. Stephenson, *Rings all of whose Pierce stalks are local*, Canad. Math. Bull. **22** (2) (1979), 159–164.

[25] D.G. Burkholder, *Azumaya rings with locally perfect centers*, J. Algebra **103** (1986), 606–618.

[26] D.G. Burkholder, *Azumaya rings, Pierce stalks and central ideal algebras*, Comm. Algebra **17** (1) (1989), 103–113.

[27] D.G. Burkholder, *Products of Azumaya rings and Kochen's map*, Comm. Algebra **17** (1) (1989), 115–134.

[28] V. Camillo and Y. Xiao, *Weakly regular rings*, Comm. Algebra **22** (10) (1994), 4095–4112.

[29] V. Camillo and H.-P. Yu, *Exchange rings, units and idempotents*, Comm. Algebra **22** (12) (1994), 4737–4749.

[30] V. Camillo and H.-P. Yu, *Stable range one for rings with many idempotents*, Trans. Amer. Math. Soc. **347** (8) (1995), 3141–3147.

[31] R. Camps and P. Menal, *Power cancellation for Artinian modules*, Comm. Algebra **19** (7) (1991), 2081–2095.
[32] R. Camps and P. Menal, *The power substitution property for rings of continuous functions*, J. Algebra **161** (2) (1993), 455–466.
[33] M. Chacron, *On algebraic rings*, Bull. Austral. Math. Soc. **1** (1969), 385–389.
[34] V.R. Chandran, *On two analogues of Cohen's theorem*, Indian J. Pure Appl. Math. **8** (1977), 54–59.
[35] M. Contessa, *On PM-rings*, Comm. Algebra **10** (1982), 93–108.
[36] M. Contessa, *On certain classes of PM-rings*, Comm. Algebra **12** (12) (1984), 1447–1469.
[37] M. Contessa, *Ultraproducts of PM-rings and MP-rings*, J. Pure Appl. Algebra **32** (1984), 11–20.
[38] R.C. Courter, *Fully idempotent rings have regular centroids*, Proc. Amer. Math. Soc. **43** (1974), 293–296.
[39] J. Dauns and K.H. Hofmann, *The representation of biregular rings by sheaves*, Math. Z. **91** (1966), 103–123.
[40] J. Dauns and K.H. Hofmann, *Representation of Rings by Sections*, Memoires of the Amer. Math. Soc., Vol. 83, Providence, RI (1968).
[41] G. De Marco and A. Orsatti, *Commutative rings in which every prime ideal is contained in a unique maximal ideal*, Proc. Amer. Math. Soc. **30** (1971), 459–466.
[42] M. Desrochers and R. Raphael, *On finite extensions of hereditary regular rings*, Comm. Algebra **21** (8) (1993), 2631–2636.
[43] F. Dischinger, *Sur les anneaux fortement π-réguliers*, C. R. Acad. Sci. **283** (8) (1976), A571–A573.
[44] A.N. Dranishnikov, *The power substitution property for rings of complex and real functions on compact metric spaces*, Proc. Amer. Math. Soc. **123** (9) (1995), 2887–2893.
[45] D.R. Farkas and R.L. Snider, *Locally finite-dimensional algebras*, Proc. Amer. Math. Soc. **81** (3) (1981), 369–372.
[46] S. Feigelstock, *An embedding theorem for weakly regular rings and fully idempotent rings*, Comment. Math. Univ. St. Pauli **27** (2) (1978/79), 101–103.
[47] J.W. Fisher and R.L. Snider, *On the Von Neumann regularity of rings with regular prime factor rings*, Pacific J. Math. **54** (1) (1974), 135–144.
[48] J.W. Fisher and R.L. Snider, *Rings generated by their units*, J. Algebra **42** (1976), 363–368.
[49] J.A. Fraser and W.K. Nicholson, *Reduced $p.p.$-rings*, Math. Japon. **34** (5) (1989), 715–725.
[50] L. Fuchs and K.M. Rangaswamy, *On generalized regular rings*, Math. Z. **107** (1968), 71–81.
[51] R. Gilmer, *Commutative rings with periodic multiplicative semigroup*, Comm. Algebra **21** (11) (1993), 4025–4028.
[52] R. Gilmer, *Zero-dimensional subrings of commutative rings*, Abelian Groups and Modules, A. Facchini and C. Menini, eds, Kluwer Acad. Publishers, Dordrecht (1995), 209–219.
[53] K.R. Goodearl, *Von Neumann Regular Rings*, Pitman, London (1979).
[54] K.R. Goodearl, *Surjective endomorphisms of finitely generated modules*, Comm. Algebra **15** (3) (1987), 589–609.
[55] K.R. Goodearl and D.E. Handelman, *Homogenization of regular rings of bounded index*, Pacific J. Math. **84** (1) (1979), 63–78.
[56] K.R. Goodearl and P. Menal, *Stable range one for rings with many units*, J. Pure Appl. Algebra **54** (1988), 261–287.
[57] K.R. Goodearl and J. Moncasi, *Cancellation of finitely generated modules over regular rings*, Osaka J. Math. **26** (1989), 679–685.
[58] K.R. Goodearl and R.R. Warfield, *Algebras over zero-dimensional rings*, Math. Ann. **223** (2) (1976), 157–168.
[59] V. Gupta, *Weakly π-regular rings and group rings*, Math. J. Okayama Univ. **19** (2) (1977), 123–127.
[60] V. Gupta, *A generalization of strongly regular rings*, Acta Math. Hung. **43** (1984), 57–61.
[61] R.E. Hartwig, *On a theorem of Flanders*, Proc. Amer. Math. Soc. **85** (3)(1982), 613–616.
[62] Y. Hirano, *Some studies on strongly π-regular rings*, Math. J. Okayama Univ. **20** (2) (1978), 141–149.
[63] Y. Hirano, *On fully right idempotent rings and direct sums of simple rings*, Math. J. Okayama Univ. **22** (1980), 43–49.
[64] Y. Hirano, *On π-regular rings with involution*, Math. J. Okayama Univ. **27** (1985), 45–47.
[65] Y. Hirano, *Some characterizations of π-regular rings of bounded index*, Math. J. Okayama Univ. **32** (1990), 97–101.

[66] Y. Hirano and J.P. Park, *On self-injective strongly π-regular rings*, Comm. Algebra **21** (1) (1993), 85–91.
[67] C.Y. Hong, H.K. Kim and J.K. Park, *Rings with restricted descending chain conditions*, Comm. Algebra **25** (8) (1997), 2579–2584.
[68] N. Jacobson, *Structure of Rings*, Amer. Math. Soc., Providence, RI (1968).
[69] I. Kaplansky, *Topological representation of algebras*, Trans. Amer. Math. Soc. **68** (1950), 62–75.
[70] K. Koh, *On one-sided ideals of a prime type*, Proc. Amer. Math. Soc. **28** (2) (1971), 321–329.
[71] K. Koh, *On a representation of a strongly harmonic ring by sheaves*, Pacific J. Math. **41** (1972), 459–468.
[72] N. Lang, *On centroids of generalized regular rings*, Ann. Scuola Norm. Sup. Pisa (3) **26** (1972), 573–585.
[73] L. Lesieur, *Sur quelques classes d'anneaux non commutatifs*, Comm. Algebra **14** (8) (1986), 1481–1488.
[74] J. Levitzki, *On the structure of algebraic algebras and related rings*, Trans. Amer. Math. Soc. **74** (1953), 384–409.
[75] V.T. Markov, *On B-rings with polynomial identity*, Trudy Sem. Petrovsk., No. 7 (1981), 232–238.
[76] P. Menal, *On π-regular rings whose primitive factor rings are Artinian*, J. Pure Appl. Algebra **20** (1) (1981), 71–78.
[77] P. Menal, *Spectral Banach algebras of bounded index*, J. Algebra **154** (1) (1993), 27–66.
[78] K. Meyberg and B. Zimmermann-Huisgen, *Rings with descending chain condition on certain principal ideals*, Indag. Math. **39** (3) (1977), 225–229.
[79] C. Năstăsescu and N. Popescu, *Anneaux semi-Artiniens*, Bull. Soc. Math. France **96** (1968), 357–368.
[80] E. Nauwelaerts and F. van Oystaeyen, *Zariski extensions and biregular rings*, Israel J. Math. **37** (4) (1980), 315–326.
[81] W.K. Nicholson, *I-rings*, Trans. Amer. Math. Soc. **207** (1975), 361–373.
[82] W.K. Nicholson, *Semiregular modules and rings*, Canad. J. Math. **28** (5) (1976), 1105–1120.
[83] W.K. Nicholson, *Lifting idempotents and exchange rings*, Trans. Amer. Math. Soc. **229** (1977), 269–278.
[84] W.K. Nicholson, *On exchange rings*, Comm. Algebra **25** (6) (1997), 1917–1918.
[85] C. Niță, *Anneaux N-réguliers*, C. R. Acad. Sci. **268** (1969), A1241–A1243.
[86] C. Niță, *Remarques sur les anneaux N-réguliers*, C. R. Acad. Sci. **271** (1970), A345–A348.
[87] M. Ohori, *On strongly π-regular rings and periodic rings*, Math. J. Okayama Univ. **19** (2) (1977), 123–127.
[88] F. van Oystaeyen and J. van Geel, *Local-global results for regular rings*, Comm. Algebra **4** (9) (1976), 811–821.
[89] R.S. Pierce, *Modules over Commutative Regular Rings*, Memoires of the Amer. Math. Soc., Vol. 70, Providence, RI (1967).
[90] V.S. Ramamurthi, *Weakly regular rings*, Canad. Math. Bull. **16** (3) (1973), 317–321.
[91] K.M. Rangaswamy, *Endomorphism representations of Zorn rings and subclasses of rings with enough idempotents*, J. Reine Angew. Math. **276** (1975), 44–55.
[92] R. Raphael, *Algebraic closures for some non-commutative rings*, Comm. Algebra **17** (7) (1989), 1687–1708.
[93] G. Shin, *Prime ideals and sheaf representation of a pseudo symmetric ring*, Trans. Amer. Math. Soc. **184** (1973), 43–60.
[94] L.A. Skornyakov, *Complemented Modular Lattices and Regular Rings*, Oliver and Boyd, Edinburgh and London (1964).
[95] J.B. Srivastava and S.K. Shah, *Semilocal and semiregular group rings*, Nederl. Akad. Wetensch. Indag. Math. **42** (3) (1980), 347–352.
[96] A. Steger, *Elementary factorization in π-regular rings*, Canad. J. Math. **18** (1966), 307–313.
[97] S.-H. Sun, *On biregular rings and their duality*, J. Pure Appl. Algebra **89** (1993), 329–337.
[98] G. Szeto, *The localization of zero-dimensional rings*, Bull. Soc. Math. Belg. **28** (1976), 193–196.
[99] G. Szeto, *On weakly biregular algebras*, Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys. **26** (9–10) (1977) , 775–780.
[100] K. Tanabe, *On rings whose Artinian modules are precisely Noetherian modules*, Comm. Algebra **22** (10) (1994), 4023–4032.
[101] K. Tanabe, *Some conditions for nil ideals and π-regular ideals*, Comm. Algebra **27** (6) (1999), 2573–2581.
[102] D.V. Tjukavkin, *Rings all of whose one-sided ideals are generated by idempotents*, Comm. Algebra **17** (5) (1989), 1193–1198.

[103] D.V. Tjukavkin, *On regular rings of bounded index*, Mat. Zametki **54** (1) (1993), 87–93.
[104] H. Tominaga, *Some results on π-regular rings of bounded index*, Math. J. Okayama Univ. **4** (1955), 135–141.
[105] H. Tsutsui, *Fully prime rings. II*, Comm. Algebra **24** (9) (1996), 2981–2989.
[106] A.A. Tuganbaev, *Distributive semiprime rings*, Mat. Zametki **58** (5) (1995), 736–761.
[107] A.A. Tuganbaev, *Bass rings and perfect rings*, Uspekhi Mat. Nauk **51** (1) (1996), 173–174.
[108] A.A. Tuganbaev, *Maximal submodules and locally perfect rings*, Mat. Zametki **64** (1) (1998), 136–142.
[109] A.A. Tuganbaev, *Semidistributive Modules and Rings*, Kluwer Academic Publishers, Dordrecht/Boston/London (1998), 362 pp.
[110] A.A. Tuganbaev, *Distributive Modules and Related Topics*, Gordon and Breach, Amsterdam (1999).
[111] J. Viola-Prioli, *On absolutely torsion-free rings*, Pacific J. Math. **56** (1975), 275–283.
[112] R.B. Warfield, *Exchange rings and decompositions of modules*, Math. Ann. **199** (1972), 31–36.
[113] R. Wisbauer, *F-semiperfekte und perfekte Moduln in $\sigma[M]$*, Math. Z. **173** (1980), 229–234.
[114] R. Wisbauer, *Foundations of Module and Ring Theory*, Gordon and Breach, Philadelphia (1991).
[115] W. Xue, *Three questions on strongly π-regular rings and regular rings*, Comm. Algebra **21** (2) (1993), 699–704.
[116] W. Xue, *Semiregular modules and F-semiperfect modules*, Comm. Algebra **23** (3) (1995), 1035–1046.
[117] X. Yao, *Weakly right duo rings*, Pure Appl. Math. Sci. **21** (1985), 19–24.
[118] H.-P. Yu, *On quasiduo rings*, Glasgow Math. J. **37** (1995), 21–31.
[119] H.-P. Yu, *On strongly π-regular rings of stable range one*, Bull. Austral. Math. Soc. **51** (3) (1995), 433–437.
[120] H.-P. Yu, *On the structure of exchange rings*, Comm. Algebra **25** (2) (1997), 661–670.
[121] J. Zelmanowitz, *Regular modules*, Trans. Amer. Math. Soc. **163** (1972), 341–355.
[122] B. Zimmermann-Huisgen and W. Zimmermann, *Algebraically compact rings and modules*, Math. Z. **161** (1978), 81–93.

Max Rings and V-rings

Askar Tuganbaev

Department of Mechanics and Mathematics, Moscow State University, Moscow, 119899 GSP, Russia
E-mail: askar@tuganbaev.mccme.ru

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1. Max rings

All rings are assumed to be associative and (except for nil-rings and some stipulated cases) to have nonzero identity elements. A module is said to be *simple* if it does not have nonzero proper submodules. A direct sum of simple modules is called a *semisimple* module. A submodule N of a module M is called a *maximal* submodule (in M) if the module M/N is simple. A ring A is called a *right max* ring if every nonzero right A-module has a maximal submodule.

It is directly verified that every semisimple ring is a right and left max ring, and every factor ring of a right max ring is a right max ring.

For a module M, the set of all the maximal submodules of M is denoted by $\max(M)$. For a module M, the intersection of kernels of all the homomorphisms from M into simple modules is called the *Jacobson radical* of the module M and is denoted by $J(M)$. Note that either $J(M) = M$ (if $\max(M) = \emptyset$) or $J(M)$ coincides with the intersection of all the maximal submodules of M (if $\max(M) \neq \emptyset$). A module M is said to be *semiprimitive* if $J(M) = 0$.

A ring J without identity element is said to be *right* (resp. *left*) T-*nilpotent* if for any sequence $a_1, a_2 \dots$ of elements in J, there exists a subscript k such that $a_k a_{k-1} \cdots a_1 = 0$ (resp. $a_1 a_2 \cdots a_k = 0$).

1.1 [11,45]. A ring A is a right max ring $\Longleftrightarrow$
the factor ring $A/J(A)$ is a right max ring, and the ideal $J(A)$ is right T-nilpotent.

A ring A is called a *semilocal* ring if the factor ring $A/J(A)$ is Artinian. A ring A is said to be *right perfect* if the ideal $J(A)$ is right T-nilpotent, and the factor ring $A/J(A)$ is Artinian.

1.2 [11]. A semilocal ring A is a right max ring $\Longleftrightarrow$
A is a right perfect ring.

A ring A is said to be *regular* if for every element a of A, there exists an element b of A such that $a = aba$. A ring is said to be *orthogonally finite* if it does not contain an infinite set of nonzero orthogonal idempotents.

1.3 [45]. A commutative ring A is a max ring $\Longleftrightarrow$
the factor ring $A/J(A)$ is regular, and the ideal $J(A)$ is right T-nilpotent.

1.4 [45]. A commutative ring A is a perfect ring $\Longleftrightarrow$
A is an orthogonally finite max ring.

A ring A is said to be *simple* if every nonzero ideal of A coincides with A. A ring is called a *domain* if the product of any two of its nonzero elements is not equal to zero.

1.5 [25,60]. There exists a right max ring A such that A is a simple orthogonally finite domain, A is not a division ring, $A/J(A)$ is not a regular ring, all simple right A-modules are isomorphic, and all the one-sided ideals of A are principal.

Let A be a ring. The center of A is denoted by $C(A)$. If N is a subset of a right (resp. left) A-module M, then the annihilator of N in A is denoted by $r(N)$ (resp. $\ell(N)$).

An element a of a ring A is said to be *right regular* (resp. *left regular*) in A if $r(a) = 0$ (resp. $\ell(a) = 0$). Right and left regular elements are called *regular* elements.

1.6 [84]. Let A be a right max ring. Then every regular central element of a A is invertible in A, and the center of every semiprime factor ring of A is a commutative regular ring.

Let A be a ring, R be a subring in $C(A)$, M be a maximal ideal of R, and let $T \equiv R \setminus M$. Then there exist a ring A_M and a ring homomorphism f such that all the elements of $f(T)$ are invertible in A_M, $A_M = \{f(a)f(t)^{-1} \mid a \in A,\ t \in T\}$, and $\mathrm{Ker}(f)$ coincides with the set $\{a \in A \mid at = 0$ for some element $t \in T\}$. The ring A_M is called the *localization* of A with respect to M, and f is called the *canonical homomorphism*.

1.7 [35]. A commutative ring A is a max ring $\Longleftrightarrow$
A_M is a perfect ring for every maximal ideal M of A.

1.8 [85]. Let A be a ring, and let R be a unitary subring of the center $C(A)$ of the ring A. Then the following conditions are equivalent.

(1) A is a right max ring.
(2) For every maximal ideal M of R and every factor ring B of A, the localization A_M is a right max ring, and all regular central elements of the ring B are invertible in B.
(3) For every maximal ideal M of R, the localization A_M is a right max ring, and the canonical homomorphism $A \to A_M$ is surjective.

A ring A is *strongly π-regular* if for every $x \in A$, there exists a positive integer m such that $x^m \in x^{m+1}A$. It is proved in [29] that a ring A is strongly π-regular $\Longleftrightarrow$ for every $x \in A$, there exists a positive integer n such that $x^n \in Ax^{n+1}$.

1.9 [85]. Let A be a ring with the center C, and let R be a unitary subring in C. Assume that A_M is a semilocal ring for every maximal ideal M of the ring R. Then
A is a right max ring $\Longleftrightarrow$
A_M is a right max ring for every maximal ideal M of R $\Longleftrightarrow$
A_M is a right perfect ring for every maximal ideal M of R $\Longleftrightarrow$
A is a strongly π-regular right max ring.

1.10 [85]. Let A be a ring with the center C, and let the ring A_M be a finitely generated module over its center for every maximal ideal M of the ring C. Then
A is a right max ring $\Longleftrightarrow$
A is a left max ring $\Longleftrightarrow$
A_M is a right max ring for every maximal ideal M of C $\Longleftrightarrow$
A_M is a left max ring for every maximal ideal M of C $\Longleftrightarrow$
A_M is a right perfect ring for every maximal ideal M of C $\Longleftrightarrow$
A_M is a left perfect ring for every maximal ideal M of C $\Longleftrightarrow$
C_M is a perfect ring for every maximal ideal M of C $\Longleftrightarrow$
$C/J(C)$ is a regular ring, and $J(C)$ is a T-nilpotent ideal $\Longleftrightarrow$

A is a strongly π-regular ring, and $J(A)$ is a right T-nilpotent ideal $\Longleftrightarrow$
A is a strongly π-regular ring, and $J(A)$ is a left T-nilpotent ideal.

A ring A is said to be *π-regular* if for every element a of A, there exists an element b of A such that $a^n = a^n b a^n$ for some positive integer n.

1.11 [87]. Let A be a ring, and let R be a unitary central subring of A.
(1) If the ring A is π-regular, then
A is a right max ring $\Longleftrightarrow$
A_M is a right max ring for every maximal ideal M of R.
(2) A is a strongly π-regular right max ring $\Longleftrightarrow$
A_M is a strongly π-regular right max ring for every maximal ideal M of R.

A ring A is called a *PI-ring* if A satisfies the polynomial identity $f(x_1, \dots, x_n) = 0$, where $f(x_1, \dots, x_n)$ is a polynomial of noncommuting variables with coefficients in the ring of integers $\mathbf{Z}$, and $\mathbf{Z}$ coincides with the ideal generated by the coefficients of $f(x_1, \dots, x_n)$.

1.12 [5]. Let A be a *PI*-ring.
(1) A is a right perfect ring $\Longleftrightarrow$
A is an orthogonally finite right max ring.
(2) If A is right max ring, then the factor ring $A/J(A)$ is π-regular.
(3) If $J(A)$ is right T-nilpotent and $B^2 = B$ for every ideal B of the factor ring $A/J(A)$, then A is a right max ring.

1.13 [65]. Let A be a *PI*-ring. Then
A is a right max ring $\Longleftrightarrow$
A is a left max ring $\Longleftrightarrow$
the factor ring $A/J(A)$ is π-regular, and $J(A)$ is right or left T-nilpotent.

A ring is called a *prime* ring if the product of any two of its nonzero ideals is not equal to zero. A ring A is said to be *semiprime* if A does not have a nonzero nilpotent ideal. An ideal P of a ring A is said to be *prime* (resp. *semiprime*) if the factor ring A/P is a prime (resp. semiprime) ring.

1.14 [65]. Let A be a semiprime *PI*-ring. Then
A is a right max ring $\Longleftrightarrow$
A is a left max ring $\Longleftrightarrow$
A is a π-regular ring
every prime ideal of A is a maximal ideal $\Longleftrightarrow$
for any semiprime ideal P of A, the center of the factor ring A/P is a regular ring $\Longleftrightarrow$
there exists a positive integer n such that $a^n \in a^{n+1}A$ for every element a of A $\Longleftrightarrow$
there exists a positive integer n such that $a^n \in Aa^{n+1}$ for every element a of A $\Longleftrightarrow$
there exists a positive integer n such that $B^n = B^{n+1}$ for every ideal B of A $\Longleftrightarrow$
for every ideal B of A, there exists a positive integer $n = n(B)$ such that $B^n = B^{n+1}$
$\Longleftrightarrow$

for every element a of A, there exists a positive integer $n = n(a)$ such that $a^n \in (AaA)^{n+1}$.

1.15 [65]. Let F be a field, A be the ring of all 2×2 matrices over F, B be the ring of upper triangular 2×2 matrices over F, $\{A_i\}_{i=1}^{\infty}$ be a countable set of copies of A, D be the direct product of all rings A_i, and let R be the subring in D generated by the ideal $\bigoplus_{i=1}^{\infty} A_i$ and by the subring $\{(b, b, b, \ldots) \mid b \in B\}$.
Then R is a strongly π-regular semiprime max PI-ring which is not a regular ring.

Let A be a ring, C be the center of A, and let a be an element of A. If a is a root of a polynomial $f(x)$ with coefficients in C and the leading coefficient of $f(x)$ is invertible (resp. regular) in A, then the element a is said to be *integral* (resp. *algebraic*) over $C(A)$. A ring A is *integral* (resp. *algebraic*) over its center if all its elements are integral (resp. algebraic) over C.

1.16 [84]. Let A be a ring. Assume that every prime factor ring of A is algebraic over its center.
(1) A is a right perfect ring $\Longleftrightarrow$
A is an orthogonally finite right max ring.
(2) If A is a right max ring, then A is strongly π-regular.

1.17 [84]. Let A be a ring. Assume that either A is integral over its center or all the prime factor rings of A are PI-rings. Then
A is an orthogonally finite right max ring $\Longleftrightarrow$
A is a right perfect ring.

A ring A is called a *ring of index at most n* if there exists a positive integer n such that $a^n = 0$ for every nilpotent element a of the ring A. A ring A is called a *ring of bounded index* if A is a ring of index at most n for some positive integer n. A ring A is said to be *right primitive* if A has a maximal right ideal which does not contain any nonzero ideal of A.

1.18 [84]. A is a right perfect ring $\Longleftrightarrow$
A is an orthogonally finite right max ring, $A/J(A)$ is a ring of bounded index, and every right primitive factor ring of A is strongly π-regular.

A ring without nonzero nilpotent elements is called a *reduced* ring. A ring is said to be *normal* if all of its idempotents are central.

1.19 [84]. Let A be a right max ring such that all right primitive factor rings of A are strongly π-regular rings. Then every right regular element a of A is invertible in A. In addition, if A is a reduced ring, then A is an normal regular ring.

A ring A is said to be *right quasi-invariant* if all maximal right ideals of A are ideals in A. A ring A is said to be *right invariant* if all right ideals of A are ideals in A.

1.20 [84]. Let A be a ring. Then

A is a right max ring, $A/J(A)$ is a reduced ring, and all right primitive factor rings of A are strongly π-regular rings $\Longleftrightarrow$

A is a right quasi-invariant right max ring $\Longleftrightarrow$

$A/J(A)$ is an normal regular ring, and $J(A)$ is right T-nilpotent.

1.21 [87]. Assume that all right primitive factor rings of a ring A are right max rings, $A/J(A)$ is a regular ring, and the ideal $J(A)$ is right T-nilpotent.

Then A is a right max ring.

A module M is said to be *semi-Artinian* if every nonzero factor module of M contains a simple submodule.

1.22 [20]. Let A be a left semi-Artinian ring with the maximum condition on (right and left) primitive ideals.

Then A is a right max ring.

1.23 [20]. There exists a right and left semi-Artinian ring which is not a right max ring.

1.24 [19]. Let A be a right max ring which is a right principal ideal domain.

Then A is a simple ring.

1.25 [19]. Let A be a right max ring which is a right Ore domain, and let a be an element of A such that aA is a maximal right ideal in A.

(1) A/a^nA is a semisimple module for every positive integer N.

(2) If $b \in A$ and b generates a maximal right ideal in A, then either A/abA or A/baA is semisimple.

Let N be a module. A module M is said to be *injective with respect to N* or *N-injective* if for every submodule $\overline{N}$ of N, each homomorphism $\overline{N} \to M$ can be extended to a homomorphism $N \to M$. A module M is said to be *quasi-injective* if M is injective with respect to M. Let A be a ring. A module M_A is said to be *injective* if M is injective with respect to every right A-module. A ring A is said to be *right self-injective* if the module A_A is injective.

A module M is called an *essential extension* of its submodule N if N has nonzero intersection with any nonzero submodule of M. In this case, N is called an *essential* submodule in M. If M is an injective module and N is an essential submodule in M, then M is called the *injective hull* of N. The injective hull is unique up to isomorphism.

A module T_A is called a *cogenerator* if for every nonzero module M_A, there exists a nonzero homomorphism $M \to T$. If A is a ring, then the injective hull of the direct sum of representatives of all isomorphism classes of simple right A-modules is called the *minimal cogenerator*.

1.26 [34]. A is a right max ring $\Longleftrightarrow$

there is a cogenerator T_A such that every nonzero submodule of T has a maximal submodule $\Longleftrightarrow$

every nonzero quasi-injective right A-module has a maximal submodule.

1.27 [34]. Let A be a ring. If A has an injective cogenerator E_A such that $\mathrm{End}(E_A)$ is a left semi-Artinian ring, then A is a right max ring, and the ring $\mathrm{End}(E_A)$ is right perfect; moreover, there are only finitely many isomorphism classes of simple right A-modules.

1.28. Let A be a ring, B be a unitary subring in A, $\{A_i\}_{i=1}^{\infty}$ be a countable set of copies of A, D be the direct product of all rings A_i, and let R be the subring in D generated by the ideal $\bigoplus_{i=1}^{\infty} A_i$ and by the subring $\{(b, b, b, \ldots) \mid b \in B\}$. Then
R is a right max ring $\Longleftrightarrow$ A and B are right max rings.

Let A be a ring, and let $S(A)$ be the non-empty set of all proper ideals of A generated by central idempotents. If P is a maximal element in the set $S(A)$, then the factor ring A/P is called a *Pierce stalk* of A.

Let A be a ring, γ be an ordinal, and let $I = \{I_\alpha \mid 0 \leqslant \alpha < \gamma\}$ be a sequence of ideals of A. The set I is called a *Pierce chain* if (i) $I_0 = 0$, (ii) $I_\alpha \subseteq I_\beta$ for $\alpha \leqslant \beta < \gamma$, (iii) A/I_β is a Pierce stalk of $R/I(\beta - 1)$ if $\beta < \gamma$ and β is not a limit ordinal, and (iv) $I_\beta = \bigcup_{\alpha<\beta} I_\alpha$ if $0 < \beta < \gamma$ and β is a limit ordinal. A factor ring A/B of A is called a *maximal indecomposable factor* of A if the ring A/B is indecomposable and B is a part of a Pierce chain.

1.29. A is a right max ring $\Longleftrightarrow$
all Pierce stalks of A are right max rings $\Longleftrightarrow$
every maximal indecomposable factor of A is a right max ring.

A ring A is called a *right quasi-max ring* if every nonzero Artinian right A-module has a maximal submodule.

1.30 [97].

(1) A is a right quasi-max ring $\Longleftrightarrow$
the category of all right A-modules has a cogenerator C such that every nonzero Artinian submodule of C has a maximal submodule.

(2) If A is a commutative semilocal ring such that $J(A)$ is a nil-ideal, then A is a quasi-max ring.

2. V-rings

A ring A is called a *right V-ring* if every simple right A-module is an injective module.

2.1 [66]. Let A be a ring. Then
A is a right V-ring $\Longleftrightarrow$
every right A-module M is a semiprimitive module $\Longleftrightarrow$
every cyclic right A-module M is a semiprimitive module $\Longleftrightarrow$ every right ideal of A is the intersection of some set of maximal right ideals of A.

It follows from 2.1 that *every right V-ring is a right max ring*.

2.2 [69]. A is a right V-ring $\Longleftrightarrow$
the category of right A-modules has a cogenerator which is a direct sum of simple modules.

2.3 (*Kaplansky*). Let A be a commutative ring. Then
A is a V-ring $\Longleftrightarrow$ A is a regular ring.

2.4 [25,60]. There exists a simple domain A such that A is not a regular ring, A is a right V-ring, all simple right A-modules are isomorphic, and all one-sided ideals of A are principal.

A ring A is said to be *right weakly regular* if $B^2 = B$ for every right ideal B of A.

2.5 [66]. Every right V-ring is right weakly regular.

2.6 [42]. If A is a regular ring whose right primitive factor rings are Artinian, then all direct sums of isomorphic simple right A-modules are injective.

2.7 [9]. Let A be a ring all of whose right primitive factor rings are Artinian. Then
A is a right V-ring $\Longleftrightarrow$
all direct sums of isomorphic simple right A-modules are injective $\Longleftrightarrow$
A is a right weakly regular ring $\Longleftrightarrow$
A is a regular ring.

2.8 [32]. A is a right V-ring $\Longleftrightarrow$
A is a right weakly regular ring, and all right primitive factor rings of A are right V-rings.

2.9 [6]. A is a right V-ring $\Longleftrightarrow$
A_M is a right V-ring for every maximal ideal M of $C(A)$.

2.10 [1]. If A is a left semi-Artinian left or right V-ring, then A is a regular ring.

2.11 [66]. Let A be a ring, G be a group, and let $A[G]$ be the group ring.
(1) If $A[G]$ is a right V-ring, then A is a right V-ring, G is locally finite, and the order of each element in G is invertible in A.
(2) If G is a finite group, then
$A[G]$ is a right V-ring $\Longleftrightarrow$
A is a right V-ring, and the order of G is invertible in A.
(3) If A is a regular ring and $A[G]$ is a right V-ring, then $A[G]$ is a regular ring.

2.12 [37]. Let A be a ring, G be a group, and let the group ring $A[G]$ be a right V-ring. If A is either a *PI*-ring or a left semi-Artinian ring, then $A[G]$ is a regular ring.

2.13 [38]. Let A be a field, G be a countable group, and let $A[G]$ be the group ring.
(1) If A contains all the roots of unity, then

$A[G]$ is a right V-ring $\Longleftrightarrow$
G is locally finite, the order of each element in G is invertible in A, and G has an Abelian subgroup of finite index.

(2) If $A[G]$ is a regular ring, then
$A[G]$ is a right V-ring $\Longleftrightarrow$
every primitive factor ring of $A[G]$ is Artinian.

Let U be a right module over a ring A. A submodule V of U is called a *pure* submodule in U_A if for every left A-module M, the natural group homomorphism $V \otimes_A M \to U \otimes_A M$ is a monomorphism. A module M_A is said to be *pure Baer injective* if for each pure right ideal B of A, any homomorphism $f : B_A \to M$ can be extended to a homomorphism $A \to M$.

2.14 [2]. Let A be a right V-ring. If every semisimple right A-module is pure Baer injective, then A satisfies the ascending chain condition on pure right ideals.

2.15 [34]. Let A be a ring, E_A be the minimal injective cogenerator, and let $J(\mathrm{End}(E)) = 0$. If every proper submodule of E is contained in some maximal submodule, then R is a right V-ring, and there is only a finite number of isomorphism classes of simple right R-modules.

2.16 [36]. A ring R is a right V-ring $\Longleftrightarrow$
there exists a semisimple right R-module W such that for every right ideal I of R, $I = \mathrm{ann}_R\, \mathrm{ann}_W\, I$, where ann denotes the respective annihilator.

A submodule N of a module M is a *superfluous* submodule (in M) if $N + M' \neq M$ for every proper submodule M' of M. An epimorphism $f : M \to N$ of R-modules is called a superfluous cover if $\mathrm{Ker}(f)$ is superfluous in M.

2.17 [57]. A is a right V-ring $\Longleftrightarrow$
all cyclic right A-modules have no proper superfluous covers.

A ring A is said to be *unit-regular* if for every element a of A, there exists an invertible element b of A such that $a = aba$. A ring A is called a *directly finite* ring if $xy = 1$ for all $x, y \in A$ such that $yx = 1$.

A module M is said to be *projective with respect to* a module N if for every epimorphism $h : N \to \overline{N}$ and each homomorphism $\bar{f} : M \to \overline{N}$, there exists a homomorphism $f : M \to N$ such that $\bar{f} = hf$. A module M over a ring A is called a *projective module* if M is projective with respect to every A-module N. A module M is said to be *quasi-projective* if M is projective with respect to M. A module M is said to be *hereditary* (resp. *semihereditary*) if all submodules (resp. all finitely generated submodules) of M are projective modules.

2.18 [10]. (1) A is a right semi-Artinian right V-ring $\Longleftrightarrow$
every right A-module contains a nonzero injective module.

(2) Every right semi-Artinian right V-ring is regular.

(3) If A is a right and left semi-Artinian right and left V-ring, then every right primitive factor ring of A is Artinian.
(4) Let A be a ring such that every right primitive factor ring of A is Artinian. Then
A is a right semi-Artinian right V-ring $\Longleftrightarrow$
A is a left semi-Artinian left V-ring $\Longleftrightarrow$
A is a regular right and left semi-Artinian ring.
(5) Let A be a right semi-Artinian right V-ring whose proper ideals are prime. Then
R is a unit-regular regular ring $\Longleftrightarrow$
R is a directly finite ring.
(6) There exists a directly finite right semi-Artinian right V-ring which is not unit-regular.
(7) There exists a hereditary unit-regular right semi-Artinian right V-ring which is not a left V-ring.

A module M is said to be *finite-dimensional* (in the sense of Goldie) if M does not contain infinite direct sums of nonzero right ideals. A ring A is called a *right Goldie* ring if A is a right finite-dimensional ring with the maximum condition on right annihilators.

2.19 [61]. Let A be a right Noetherian right V-ring, and let E be the injective hull of A_A. Then
(1) E/A_A is a semisimple module $\Longleftrightarrow$
for every essential right ideal B of A, the module A_A/B is Artinian.
(2) If A is a simple ring and the module A_A/B is Artinian for every essential right ideal B of A, then A is a left Noetherian left V-ring if and only if A is left Goldie.

2.20 [61]. Let A be a right and left Noetherian right and left V-ring, and let E be the injective hull of A_A. Then
E/A_A is a semisimple module $\Longleftrightarrow$
A is a right and left hereditary ring.

2.21 [88]. Let A be a right self-injective regular V-ring such that the dimension of every simple right A-module M over the division ring $\mathrm{End}(M)$ is less than $2^{2^{\aleph_0}}$.
Then A is a ring of bounded index.

If M is a right module over a ring A, then $\mathrm{Sing}(M)$ denotes the set of all the elements m of M such that $r(m)$ is an essential right ideal of A. A module M is said to be *singular* (resp. *nonsingular*) if $\mathrm{Sing}(M) = M$ (resp. $\mathrm{Sing}(M) = 0$).
For a module M, the sum of all simple submodules of M is called the *socle* of M. It is denoted by $\mathrm{Soc}(M)$.

2.22 [54]. Let A be a ring which is not right Noetherian. Then
A is a right V-ring such that $\mathrm{Soc}(A_A)$ is not finitely generated and $A/\mathrm{Soc}(A_A)$ is a division ring $\Longleftrightarrow$
each cyclic right A-module is either nonsingular or injective.

2.23 [16]. A is a right max ring $\Longleftrightarrow$
every maximal indecomposable factor of A is a right max ring.

2.24 [46]. Assume A is a right Goldie right V-ring such that the right classical ring of quotients Q of A is two-sided, and the right A-module A/B is semi-Artinian for every essential right ideal B of A.

If S is a subring in Q and $R \subseteq Q$, then S is a right Goldie right V-ring, and the right S-module S/T is semi-Artinian for every essential right ideal T of S.

2.25 [47]. Every right V-ring with the maximum condition on left annihilators is a finite direct product of simple right V-rings.

2.26 [4]. Let A be a regular right V-ring, M be a right A-module, and let M/MP be an Artinian module for each prime ideal P of A.
Then $\operatorname{End}(M)$ is a strongly π-regular ring.

Let L be a partially ordered set (poset). If $L_1, L_2 \in L$ and $L_2 \leqslant L_1$, then L_1/L_2 denotes the poset in L such that $N \in L_1/L_2 \Longleftrightarrow L_2 \leqslant N \leqslant L_1$. The *deviation* $\operatorname{dev}(L)$ of L can be defined by transfinite induction as follows.

(1) $\operatorname{dev}(L) = -1 \Longleftrightarrow L$ contains exactly one element.
(2) Consider an ordinal $\alpha \geqslant 0$; also, assume that we have already defined which posets have deviation β for ordinals $\beta < \alpha$. Then $\operatorname{dev}(L) = \alpha \Longleftrightarrow$
  (i) we have not already defined $\operatorname{dev}(L) = \beta$ for some $\beta < \alpha$,
  (ii) for every (countable) descending chain $L_1 > L_2 > \cdots$ of elements of L, we have $\operatorname{dev}(L_i/L_{i+1}) < \alpha$ for all but finitely many subscripts i (this means that, for all but finitely many i, the deviation of L_i/L_{i+1} has previously been defined, and therefore, is an ordinal less than α).

If the lattice $\operatorname{Lat}(M)$ of submodules of a module M has a deviation, then this deviation is called the *Krull dimension* of M.

2.27 [14]. A right V-ring with right Krull dimension is right Noetherian.

A right A-module M is called a *rational extension* of its submodule N if for all $x, y \in M$ with $y \neq 0$, there exists $a \in A$ such that $xa \in N$ and $ya \neq 0$. A module is said to be *rationally complete* if it does not have proper rational extensions. Every injective module is rationally complete.

2.28 [15]. (1) Let A be a ring. All simple right A-modules are rationally complete provided that A has the following property $(*)$: for each maximal right ideal M, any right ideal I and any homomorphism $f : A/M \to A/I$ so that A/I is an essential extension of $f(A/M)$, there exists a right ideal B so that $I \subseteq B$ and $A/M \cong A/B$.

(2) If A is a right Noetherian ring, then
all simple right A-modules are rationally complete $\Longleftrightarrow$
A has the property $(*)$.

2.29 [110]. A ring A is a right V-ring $\Longleftrightarrow$
for any right ideal I of A and any maximal right subideal K of I, there exists a maximal right ideal M of R such that $M \cap I = K$.

Some results about V-rings are presented in [7,15,23,30,39,43,41,51,53,56,59,62,70,71, 77,80,79,82,93,95,96,99,101–111,115–117].

3. Rings and modules related to V-rings

A ring A is called a *right GV-ring* if all simple singular right A-modules are injective. It can be directly verified that A is a right GV-ring $\Longleftrightarrow$ every simple right A-module is either injective or projective.

3.1 [46]. Every semiprime right GV-ring with the maximum condition on left annihilators is a finite direct product of simple right V-rings.

3.2 [8]. Let A be a ring, and let S be the right socle of A.
(1) A is a right GV-ring $\Longleftrightarrow$
S_A is projective and A/S is a right V-ring.
(2) If A is a semiprime right GV-ring, then
A is right (resp. left) finite-dimensional $\Longleftrightarrow$
A is a right (resp. left) Goldie ring $\Longleftrightarrow$
A is a finite direct product of simple right V-rings which are right (resp. left) Goldie rings.
(3) If all maximal essential right ideals of A are two-sided, then
A is regular and A/S is normal regular $\Longleftrightarrow$
A/S^2 is a normal regular ring and $S^2 \cap J(A) = 0 \Longleftrightarrow$
A is a right GV-ring and $S^2 \cap J(A) = 0 \Longleftrightarrow$
A is a regular ring $\Longleftrightarrow$
A is a right weakly regular ring.
(4) A is a self-injective regular ring and A/S is a normal regular ring $\Longleftrightarrow$
$A = A_1 \times A_2 \times A_3$, where A_1 is a semisimple Artinian ring, A_2 is a direct product of division rings and A_3 is a normal regular self-injective ring with zero socle.

3.3 [54]. For a ring A, the following conditions are equivalent.
(1) Each cyclic right A-module is either nonsingular or injective.
(2) Either A is a direct product of a semisimple Artinian ring and a domain over which every singular right module is injective,
or A is a right V-ring such that $\mathrm{Soc}(A_A)$ is not finitely generated and $A/\mathrm{Soc}(A_A)$ is a division ring.

A right A-module M is said to be *p-injective* if every homomorphism from a principal right ideal of A to M can be extended to a homomorphism from A itself to M. A left A-module E is *flat* if for any monomorphism of right A-modules $u : M_1 \to M_2$, the group homomorphism $M_1 \otimes E \to M_2 \otimes E$ is a monomorphism.

3.4 [72]. Let B be an ideal of a ring A, and let all simple right A-modules be p-injective.
(1) The left A-module A/B is flat.
(2) If the ideal B is a maximal right ideal, then the right A-module A/B is injective.

3.5 [110]. If A is a ring, then
every simple right A-module is p-injective $\Longleftrightarrow$
for any principal right ideal I of A and any maximal right subideal K of I, there exists a maximal right ideal M of A such that $M \cap I = K$.

A right A-module M is said to be *GP*-injective if for any nonzero element $a \in A$, there exists a positive integer n such that $a^n \neq 0$ and each homomorphism $a^n A \to M_A$ can be extended to a homomorphism $A_A \to M$.

In [68], the authors characterize von Neumann regular rings whose simple right modules are *GP*-injective.

3.6 [28]. Let A be a ring, and let all simple singular right A-modules be *GP*-injective.
(1) If A is a prime ring, then either $J(A) = 0$ or A is a domain.
(2) If A is right invariant, then A is strongly regular.
(3) If every maximal right ideal of A is a right annihilator ideal, then A is semisimple Artinian.

3.7 [22]. Let A be a ring, and let every simple right A-module be either *GP*-injective or flat.
(1) If A is semiprime, then the center of A is a regular ring.
(2) If A is a reduced ring then A is a left and right weakly regular ring.

A ring A is called a *weakly right invariant* if for any nonzero element $a \in A$, there exists a positive integer n such that $a^n A$ is an ideal in A.

3.8 [58]. Let A be a ring, and let all simple singular right A-modules be *GP*-injective.
(1) If $r(a)$ is an ideal in A for any $a \in A$, then A is a reduced right and left weakly regular ring.
(2) If every nonzero right ideal of A contains a nonzero ideal of A, then A is a reduced right and left weakly regular ring.
(3) If A is a normal right or left quasi-invariant ring, then A is a normal regular ring.
(4) If A is a weakly right invariant ring, then A is a normal regular ring.

A module M_A is called a *V-module* if every submodule of M is the intersection of maximal submodules of M. This is equivalent to having all simple right A-modules M-injective.

A right module M over a ring A is said to be *regular in the sense of Zelmanowitz* if, for each $m \in M$, there is $f \in \mathrm{Hom}(M, A_A)$ with $m = mf(m)$. A module U is said to be *locally projective* if given a monomorphism $0 \to V \to U$ with V finitely generated and an epimorphism $f : X \to Y$ and a homomorphism $g : U \to Y$, there exists a homomorphism $h : U \to X$ such that $g|_V = fh|_V$.

3.9 [48,49]. Let M be a module over a PI-ring. Then
M is regular in the sense of Zelmanowitz $\Longleftrightarrow$
M is a locally projective V-module.

3.10 [50]. Let M_A be a finitely generated projective module with trace ideal $T = \sum(\mathrm{Im}(f)\colon f \in \mathrm{Hom}(M, A_A))$. Then
M is a V-module $\Longleftrightarrow$
$\mathrm{End}(M)$ is a right V-ring and $m \in mT$ for each $m \in M$.

3.11 [31]. Let M be a quasi-projective module. Then
M is a semi-Artinian V-module $\Longleftrightarrow$
each nonzero cyclic subquotient of M has a nonzero M-injective submodule.

A module M is called a *semi-V-module* if every nonzero homomorphic image of M has a nonzero V-submodule.

3.12 [21]. (1) A ring A is a right semi-V-ring $\Longleftrightarrow$
$J(A)$ is left T-nilpotent and $A/J(A)$ is a right semi-V-ring.
(2) If A is a commutative ring, then
A is a semi-V-ring $\Longleftrightarrow$
$J(A)$ is T-nilpotent and $A/J(A)$ is a regular ring.

3.13 [100]. A V-module with Krull dimension is Noetherian.

A right A-module M is called a GV-module if every singular simple right A-module is M-injective.

3.14 [55]. Let M be a GV-module.
(1) M has a maximal submodule.
(2) $M/\mathrm{Soc}(M)$ is a Noetherian module $\Longleftrightarrow$
$M/\mathrm{Soc}(M)$ has Krull dimension.

A module M_A is said to be *generalized co-semisimple* if every singular simple A-module is M-injective or M-projective. There exists a generalized co-semisimple module which is not a GV-module [94].

3.15 [94]. If M is a generalized co-semisimple module, then
M is Noetherian $\Longleftrightarrow$
M has Krull dimension $\Longleftrightarrow$
every factor module of M is finite-dimensional.

For a module M, the category of all submodules of modules which are homomorphic images of direct sums of isomorphic copies of M is denoted by $\sigma[M]$. A module N is said to be *M-singular* if $N = L/K$ for L and K in $\sigma[M]$ and K essential in L. The class of M-singular modules is closed under submodules, homomorphic images and direct sums, so for each N in $\sigma[M]$ there is a largest submodule $Z_M(N)$ contained in N which is M-singular. A module is said to be *locally Noetherian* if all finitely generated submodules are Noetherian.

3.16 [94]. (1) M is a generalized co-semisimple module $\Longleftrightarrow$
$Z_M(N) \cap \mathrm{Rad}(N) = 0$ for every module $N \in \sigma[M] \Longleftrightarrow$
$M/\operatorname{Soc}(M)$ is a V-module, and every finitely generated submodule of $Z_M(M) \cap \operatorname{Soc}(M)$ is a direct summand in $M \Longleftrightarrow$
every module in $\sigma[M]$ is a generalized co-semisimple module.

(2) $M/\operatorname{Soc}(M)$ is a locally Noetherian V-module $\Longleftrightarrow$
every M-singular module in $\sigma[M]$ is a locally Noetherian V-module $\Longleftrightarrow$
every M-singular semisimple module in $\sigma[M]$ is $M/\operatorname{Soc}(M)$-injective.

(3) M is a generalized co-semisimple module with the maximum condition on essential submodules $\Longleftrightarrow$
M is generalized co-semisimple and M/K is finite-dimensional for every essential submodule $K \Longleftrightarrow$
$M/\operatorname{Soc}(M)$ is a V-module and $Z_M(M) \cap J(N) = 0$.

(4) Let M be a module such that $M/\operatorname{Soc}(M)$ is finitely generated. Then
M is a generalized co-semisimple module with the minimum condition on essential submodules $\Longleftrightarrow$
$M/\operatorname{Soc}(M)$ is semisimple and $Z_M(M) \cap J(N) = 0$.

3.17 [94]. Let M be a quasi-projective module over a ring A.

(1) M is generalized co-semisimple $\Longleftrightarrow$
$M/\operatorname{Soc}(M)$ is a V-module and $\operatorname{Soc}(M)$ is M-projective $\Longleftrightarrow$
$M/\operatorname{Soc}(M)$ is a V-module and $Z_M(M) \cap \operatorname{Soc}(M) = 0$.

(2) If the factor module $M/J(M)$ is semisimple, then
M is a generalized co-semisimple $\Longleftrightarrow$
every (cyclic) M-singular module is M-injective.

(3) If the ring A is commutative and M is finitely generated, then
M is generalized co-semisimple $\Longleftrightarrow$
the endomorphism ring of M is regular $\Longleftrightarrow$
the factor ring $A/r(M)$ is regular.

A ring A is called a right *QI*-ring if every quasi-injective right A-module is injective. A ring A is called a right *PCI*-ring if every proper cyclic right A-module is injective.

3.18 [12]. (1) Let A be a right and left hereditary left Noetherian ring. Then
A is a right *QI*-ring $\Longleftrightarrow$
A is a right Noetherian right V-ring.

(2) If A is a right and left Noetherian domain, then
A is a right *PCI*-ring $\Longleftrightarrow$
A is a right hereditary right V-ring $\Longleftrightarrow$
A is a left hereditary left V-ring.

3.19 [26]. (1) A right *PCI*-ring is either semisimple or a simple right hereditary right Ore V-domain.

(2) A left Ore right Noetherian right *PCI*-domain is left Noetherian.

A ring A is said to be *right coherent* if every finitely generated right ideal of A is finitely presented.

3.20 [14]. (1) If A is a right and left Noetherian ring, then
A is a right *PCI*-ring $\Longleftrightarrow$
A is a left *PCI*-ring.
(2) If A is a semiprime right Goldie ring, then
A is a right *QI*-ring $\Longleftrightarrow$
all singular quasi-injective right A-modules are injective.

3.21 [27]. (1) A right *PCI*-ring is right Noetherian.
(2) A is a right and left *PCI*-ring $\Longleftrightarrow$
A is semisimple or a two-sided simple Noetherian hereditary Ore V-domain.
(3) A right *PCI*-ring A is a left *PCI*-ring $\Longleftrightarrow$
A is left coherent.

3.22 [40]. Let A be a ring. Then
every cyclic right A-module is injective or projective $\Longleftrightarrow$
$A = B \times C$, where B is semisimple, and C is a simple right semihereditary right Ore domain which is a right *PCI*-ring.

3.23 [33]. Let A be a right *QI*-ring such that $\mathrm{Soc}(A/B) \neq 0$ for every essential right ideal B.

Then A is a right hereditary ring.

References

[1] J.S. Alin and E.P. Armendariz, *A class of rings having all singular simple modules injective*, Math. Scand. **23** (1968), 233–240.

[2] N.M. Al-Thani, *Pure Baer injective modules*, Internat. J. Math. Math. Sci. **3** (1997), 529–538.

[3] E.P. Armendariz, *On semiprime PI algebras over commutative regular rings*, Pacific J. Math. **66** (1) (1976), 23–38.

[4] E.P. Armendariz, *Modules with Artinian prime factors*, Proc. Amer. Math. Soc. **78** (3) (1980), 311–314.

[5] E.P. Armendariz and J.W. Fisher, *Regular PI-rings*, Proc. Amer. Math. Soc. **39** (1973), 247–251.

[6] E.P. Armendariz, J.W. Fisher and S.A. Steinberg, *Central localization of regular rings*, Proc. Amer. Math. Soc. **46** (3) (1974), 315–321.

[7] A.G. Athanassiadis, *A note on V-rings*, Bull. Soc. Math. Greece (N.S.) **12** (1) (1971), 91–95.

[8] G. Baccella, *Generalized V-rings and von Neumann regular rings*, Rend. Sem. Mat. Univ. Padova **72** (1984), 117–133.

[9] G. Baccella, *Von Neumann regularity of V-rings with Artinian primitive factor rings*, Proc. Amer. Math. Soc. **103** (3) (1988), 747–749.

[10] G. Baccella, *Semi-Artinian V-rings and semi-Artinian von Neumann regular rings*, J. Algebra **173** (3) (1995), 587–612.

[11] H. Bass, *Finistic dimension and a homological generalization of semiprimary rings*, Trans. Amer. Math. Soc. **95** (3) (1960), 466–488.

[12] A.K. Boyle, *Hereditary QI-rings*, Trans. Amer. Math. Soc. **192** (1974), 115–120.

[13] A.K. Boyle, *Injectives containing no proper quasi-injective submodules*, Comm. Algebra **4** (8) (1976), 775–785.

[14] A.K. Boyle and K.R. Goodearl, *Rings over which certain modules are injective*, Pacific J. Math. **58** (1) (1975), 43–53.

[15] S.H. Brown, *Rings over which every simple module is rationally complete*, Canad. J. Math. **25** (1973), 693–701.

[16] W.D. Burgess and W. Stephenson, *An analogue of the Pierce sheaf for noncommutative rings*, Comm. Algebra **6** (9) (1978), 863–886.

[17] K.A. Byrd, *Rings whose quasi-injective modules are injective*, Proc. Amer. Math. Soc. **33** (1972), 235–240.

[18] K.A. Byrd, *When are quasi-injectives injective?*, Canad. Math. Bull. **15** (1972), 599–600.

[19] V.P. Camillo, *On some rings whose modules have maximal submodules*, Proc. Amer. Math. Soc. **50** (1975), 97–100.

[20] V.P. Camillo and K.R. Fuller, *A note on Loewy rings and chain conditions on primitive ideals*, Lecture Notes in Math., Vol. 700, Springer, Berlin (1979).

[21] V. Camillo and M.F. Yousif, *Semi-V-modules*, Comm. Algebra **17** (1) (1989), 165–177.

[22] J.L. Chen and N.Q. Ding, *On a generalization of V-rings and SF-rings*, Kobe J. Math. **11** (1) (1994), 101–105.

[23] D.P. Choudhury and K. Tewari, *Tensor purities, cyclic quasiprojectives and cocyclic copurity*, Comm. Algebra **7** (15) (1979), 1559–1572.

[24] F. Couchot, *Classes d'anneaux contenant les V-anneaux et les anneaux absolument plats*, Lecture Notes in Math., Vol. 740, Springer, Berlin (1979), 170–183.

[25] J.H. Cozzens, *Homological properties of the ring of differential polynomials*, Bull. Amer. Math. Soc. **76** (1970), 75–79.

[26] J.H. Cozzens and C. Faith, *Simple Noetherian Rings*, Cambridge University Press, Cambridge (1975).

[27] R.F. Damiano, *A right PCI ring is right Noetherian*, Proc. Amer. Math. Soc. **77** (1) (1979), 11–14.

[28] N.Q. Ding and J.L. Chen, *Rings whose simple singular modules are YJ-injective*, Math. Japon. **40** (1) (1994), 191–195.

[29] F. Dischinger, *Sur les anneaux fortement π-réguliers*, C. R. Acad. Sci. **283** (8) (1976), A571–A573.

[30] X. Du and J.L. Zhang, *MELT right V-rings are von Neumann regular*, Chinese Sci. Bull. **40** (11) (1995), 967–968.

[31] N.V. Dung and P.F. Smith, *On semi-Artinian V-modules*, J. Pure Appl. Algebra **82** (1) (1992), 27–37.

[32] C. Faith, *Lectures on Injective Modules and Quotient Rings*, Lecture Notes in Math., Vol. 49, Springer, Berlin (1967), 1–140.

[33] C. Faith, *On hereditary rings and Boyle's conjecture*, Arch. Math. **27** (2) (1976), 113–119.

[34] C. Faith, *Rings whose modules have maximal submodules*, Publ. Mat. **39** (1) (1995), 201–214.

[35] C. Faith, *Locally perfect commutative rings are those whose modules have maximal submodules*, Comm. Algebra **23** (13) (1995), 4885–4886.

[36] C. Faith and P. Menal, *A new duality theorem for semisimple modules and characterization of Villamayor rings*, Proc. Amer. Math. Soc. **123** (6) (1995), 1635–1637.

[37] J.W. Fisher, *Von Neumann regular rings versus V-rings*, Ring Theory (Proc. Conf., Univ. Oklahoma, Norman, OK, 1973), Lecture Notes in Pure and Appl. Math., Vol. 7, Dekker, New York (1974), 101–119.

[38] D.R. Farkas and R.L. Snider, *Group algebras whose simple modules are injective*, Trans. Amer. Math. Soc. **194** (1974), 241–248.

[39] J.L. Garsia Hernandez and J.L. Gomez Pardo, *V-rings relative to Gabriel topologies*, Comm. Algebra **13** (1) (1985), 59–83.

[40] S.C. Goel, S.K. Jain and S. Singh, *Rings whose cyclic modules are injective or projective*, Proc. Amer. Math. Soc. **53** (1) (1975), 16–18.

[41] J.M. Goursaud, *Sur les V-anneaux reguliers*, Seminaire d'Algebre Non Commutative (annee 1975–1976), Exp. No. 5, Publ. Math. Orsay, Nos. 186-7655, U. E. R. Math., Univ. Paris XI, Orsay (1976).

[42] J.M. Goursaud and J. Valette, *Sur l'enveloppe injective des anneaux de groupes reguliers*, Bull. Soc. Math. France **103** (1) (1975), 91–102.

[43] J.S. Golan and Z. Papp, *Cocritically nice rings and Boyle's conjecture*, Comm. Algebra **8** (18) (1980), 1775–1798.

[44] R.M. Hamsher, *Commutative, Noetherian rings over which every module has a maximal submodule*, Proc. Amer. Math. Soc. **17** (1966), 1471–1472.

[45] R.M. Hamsher, *Commutative rings over which every module has a maximal submodule*, Proc. Amer. Math. Soc. **18** (1967), 1133–1137.
[46] F. Hansen, *Certain overrings of right hereditary, right Noetherian rings are V-rings*, Proc. Amer. Math. Soc. **52** (1975), 85–90.
[47] F. Hansen, *On one-sided prime ideals*, Pacific J. Math. **58** (1) (1975), 79–85.
[48] Y. Hirano, *Regular modules and V-modules*, Proceedings of the 13th Symposium on Ring Theory (Okayama Univ., Okayama, 1980), Okayama Univ., Okayama (1981), 24–43.
[49] Y. Hirano, *Regular modules and V-modules*, Hiroshima Math. J. **11** (1) (1981), 125–142.
[50] Y. Hirano, *Regular modules and V-modules. II*, Math. J. Okayama Univ. **23** (2) (1981), 131–135.
[51] Y. Hirano and H. Tominaga, *Regular rings, V-rings and their generalizations*, Hiroshima Math. J. **9** (1) (1979), 137–149.
[52] V.A. Hiremath, *Coflat modules*, Indian J. Pure Appl. Math. **17** (2) (1986), 223–230.
[53] X.H. Hu, *Quasiregular rings and regular rings*, J. Math. (Wuhan) **14** (4) (1994), 519–522.
[54] D.V. Huynh, S.K. Jain and S.R. Lopez-Permouth, *On a class of non-Noetherian V-rings*, Comm. Algebra **24** (9) (1996), 2839–2850.
[55] D.V. Huynh, P.F. Smith and R. Wisbauer, *A note on GV-modules with Krull dimension*, Glasgow Math. J. **32** (3) (1990), 389–390.
[56] A. Idelhadj and E.A. Kaidi, *The dual of the socle-fine notion and applications*, Commutative Ring Theory (Fes, 1995), Lecture Notes in Pure and Appl. Math., Vol. 185, Dekker, New York (1997), 359–367.
[57] S.K. Jain, S.R. Lopez-Permouth and L.H. Rowen, *Superfluous covers*, Comm. Algebra **23** (5) (1995), 1663–1677.
[58] N.K. Kim, S.B. Nam and J.Y. Kim, *On simple singular GP-injective modules*, Comm. Algebra **27** (5) (1999), 2087–2096.
[59] M.H. Klin, *V-rings and their connection with the automorphism groups of binary relations*, Mat. Issled. **7** (1) (1972), 204–205.
[60] L.A. Koifman, *Rings over which each module has a maximal submodule*, Mat. Zametki **7** (1970), 359–367.
[61] K.A. Kosler, *On hereditary rings and Noetherian V-rings*, Pacific J. Math. **103** (2) (1982), 467–473.
[62] Z.K. Liu, *Characterization for two kinds of rings by using the quasi-injectivity of some modules*, Northeast. Math. J. **7** (2) (1991), 141–149.
[63] Z.K. Liu, *A characterization of F-V-rings by generalized injectivity*, Acta Math. Sinica **38** (2) (1995), 200–206.
[64] Z.K. Liu and J. Ahsan, *Characterizations of F-V-rings by quasi-continuous modules*, Northeast. Math. J. **11** (4) (1995), 417–424.
[65] V.T. Markov, *On B-rings with a polynomial identity*, Trudy Sem. Petrovsk., No. 7 (1981), 232–238.
[66] G.O. Michler and O.E. Villamayor, *On rings whose simple modules are injective*, J. Algebra **25** (1973), 185–201.
[67] S. Mohamed, *On PCI rings*, J. Univ. Kuwait Sci. **2** (1975), 21–23.
[68] S.B. Nam, N.K. Kim and J.Y. Kim, *On simple GP-injective modules*, Comm. Algebra **23** (14) (1995), 5437–5444.
[69] C. Nastasescu, *Quelques remarques sur la dimension homologique des anneaux*, J. Algebra **19** (1971), 470–485.
[70] K. Ohtake, *Commutative rings of which all radicals are left exact*, Comm. Algebra **8** (16) (1980), 1505–1512.
[71] S. Page, *Semihereditary and fully idempotent FPF rings*, Comm. Algebra **11** (3) (1983), 227–242.
[72] V.S. Ramamurthi, *On the injectivity and flatness of certain cyclic modules*, Proc. Amer. Math. Soc. **48** (1975), 21–25.
[73] V.S. Ramamurthi and K.M. Rangaswamy, *Generalized V-rings*, Math. Scand. **31** (1972), 69–77.
[74] G. Renault, *Sur les anneaux A tels que tout A-module a gauche non nul contient un sous-module maximal*, C. R. Acad. Sci. Paris Ser. A–B **264** (1967), A622–A624.
[75] G. Renault, *Sur les anneaux A tels que tout A-module a gauche non nul contient un sous-module maximal*, C. R. Acad. Sci. Paris Ser. A–B **267** (1968), A792–A794.
[76] S.H. Rim, *Semisimple Artinian localizations related with V-rings*, Commun. Korean Math. Soc. **10** (4) (1995), 839–847.
[77] B. Sarath, *Krull dimension and Noetherianness*, Illinois J. Math. **20** (2) (1976), 329–353.

[78] B. Sarath and K. Varadarajan, *Injectivity of certain classes of modules*, J. Pure Appl. Algebra **5** (1974), 293–305.
[79] J. Shapiro and M. Teply, *Semisimple localizations and V-rings*, Comm. Algebra **16** (8) (1988), 1673–1688.
[80] M.S. Shrikhande, *On hereditary and cohereditary modules*, Canad. J. Math. **25** (1973), 892–896.
[81] Y. Takehana, *V-rings relative to hereditary torsion theories*, Tsukuba J. Math. **6** (2) (1982), 293–298.
[82] M.L. Teply, *On the idempotence and stability of kernel functors*, Glasgow Math. J. **37** (1) (1995), 37–43.
[83] A.A. Tuganbaev, *Max rings and perfect rings*, Uspekhi Mat. Nauk **51** (1) (1996), 173–174.
[84] A.A. Tuganbaev, *Rings over which each module has a maximal submodule*, Mat. Zametki **61** (3) (1997), 407–415.
[85] A.A. Tuganbaev, *Maximal submodules and locally perfect rings*, Mat. Zametki **64** (1998), 136–142.
[86] A.A. Tuganbaev, *Semidistributive Modules and Rings*, Mathematics and its Applications, Vol. 449, Kluwer Academic Publishers, Dordrecht (1998).
[87] A.A. Tuganbaev, *Max rings*, Preprint, Moscow (2000).
[88] D.V. Tyukavkin, *Regular self-injective rings and V-rings*, Algebra i Logika **33** (5) (1994), 564–575.
[89] K. Varadarajan, *Generalised V-rings and torsion theories*, Comm. Algebra **14** (3) (1986), 455–467.
[90] K. Varadarajan and K. Wehrhahn, *p-injectivity of simple pre-torsion modules*, Glasgow Math. J. **28** (2) (1986), 223–225.
[91] D.G. Wang, *Generalized V-modules*, J. Qufu Normal Univ. (Natural Science Edition) **19** (1993), 30–34.
[92] D.G. Wang, *On V-modules relative to torsion theory*, Rings and Radicals (Shijiazhuang, 1994), Pitman Res. Notes Math. Ser., Vol. 346, Longman, Harlow (1996), 249–252.
[93] D.G. Wang, *On some rings and self-injectivity*, Vietnam J. Math. **24** (2) (1993), 215–222.
[94] R. Wisbauer, *Generalized co-semisimple modules*, Comm. Algebra **18** (12) (1990), 4235–4253.
[95] P. Wu and X. Hu, *Some new characterizations of V-rings*, Proceedings of the Second Japan–China International Symposium on Ring Theory and the 28th Symposium on Ring Theory (Okayama, 1995), Okayama Univ., Okayama (1996), 147–149.
[96] T. Wurfel, *Sur les V-anneaux semi-Artiniens*, Rev. Roumaine Math. Pures Appl. **20** (4) (1975), 503–507.
[97] W. Xue, *Quasi-Hamsher modules and quasi-max rings*, Math. J. Okayama Univ. **39** (1997), 71–79.
[98] W. Xue, *Characterizations of hereditary modules and V-modules*, Math. J. Okayama Univ. **39** (1997), 7–16.
[99] W. Xue, *A note on YJ-injectivity*, Riv. Mat. Univ. Parma (6) **1** (1998), 31–37.
[100] M.F. Yousif, *V-modules with Krull dimension*, Bull. Austral. Math. Soc. **37** (2) (1988), 237–240.
[101] R. Yue Chi Ming, *On generalizations of V-rings and regular rings*, Math. J. Okayama Univ. **20** (2) (1978), 123–129.
[102] R. Yue Chi Ming, *On regular rings and V-rings*, Monatsh. Math. **88** (4) (1979), 335–344.
[103] R. Yue Chi Ming, *On V-rings and prime rings*, J. Algebra **62** (1) (1980), 13–20.
[104] R. Yue Chi Ming, *On von Neumann regular rings IV*, Riv. Mat. Univ. Parma (4) **6** (1980), 47–54.
[105] R. Yue Chi Ming, *On von Neumann regular rings V*, Math. J. Okayama Univ. **22** (2) (1980), 151–160.
[106] R. Yue Chi Ming, *On regular and continuous rings II*, Kyungpook Math. J. **21** (2) (1981), 171–178.
[107] R. Yue Chi Ming, *On V-rings and unit-regular rings*, Rend. Sem. Mat. Univ. Padova **64** (1981), 127–140.
[108] R. Yue Chi Ming, *Remarks on regular rings and V-rings*, Ricerche Mat. **30** (1) (1981), 3–14.
[109] R. Yue Chi Ming, *On injective modules and annihilators*, Ricerche Mat. **33** (2) (1984), 147–158.
[110] R. Yue Chi Ming, *On V-rings, p-V-rings and injectivity*, Kyungpook Math. J. **32** (2) (1992), 219–228.
[111] R. Yue Chi Ming, *On self-injectivity and regularity*, Rend. Sem. Fac. Sci. Univ. Cagliari **1** (1994), 9–24.
[112] R. Yue Chi Ming, *A note on YJ-injectivity*, Demonstratio Math. **30** (3) (1997), 551–556.
[113] J.L. Zhang, *Characterizations of normal regular rings*, Northeast. Math. J. **10** (3) (1994), 359–364.
[114] J.L. Zhang, *SF-rings whose maximal essential left ideals are ideals*, Adv. in Math. (China) **23** (3) (1994), 257–262.
[115] J.L. Zhang and X.N. Du, *Von Neumann regularity of SF-rings*, Comm. Algebra **21** (7) (1993), 2445–2451.
[116] S.H. Zhang, *T-regular rings and TV-rings*, J. Math. Res. Exposition **11** (1) (1991), 31–36.
[117] S.H. Zhang, *V-modules and V-additive categories*, Northeast Math. J. **10** (3) (1994), 346–350.
[118] M. Zhou, *The relation between commutative generalized regular rings and generalized V-rings*, Acta Math. Sinica **35** (1) (1992), 140–143.
[119] M. Zhou, *S-injective modules and S-V rings*, Northeast Math. J. **8** (3) (1992), 303–310.
[120] Y. Zhou, *Modules arising from some relative injectives*, Bull. Austral. Math. Soc. **53** (2) (1996), 249–260.

Section 3C
Coalgebras

This Page Intentionally Left Blank

Coassociative Coalgebras

Walter Michaelis

Department of Mathematics, University of New Orleans, New Orleans, LA 70148, USA

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

1. Introduction

We begin this introduction by saying a few words about coalgebras, our intent in writing this article, and the background we assume on the part of the reader. In short, we begin with an introduction to this introduction.

Coalgebras have been around for some time, most notably in connection with Hopf algebras arising in algebraic topology in the 1940's; algebraic geometry and Lie theory in the 1950's; combinatorics, functional analysis, category theory, multilinear algebra, and field theory in the 1960's; number theory in the 1970's; quantum groups in connection with inverse scattering problems, the Ising model, statistical mechanics, Hamiltonian mechanics and Poisson Lie groups in the 1980's; non-commutative ring theory, knot theory, topological quantum field theory and quantum stochastic processes in the 1990's; and most recently in connection with Feynman diagrams at the beginning of the 21st century. A selection of papers on all this is included in the references. Because coalgebras appear in so many places (most often as part of the more elaborate structure of a *bialgebra* or *Hopf algebra*, but sometimes because a familiar algebra turns out to be understood best by virtue of the fact that it is the *dual algebra* of a coalgebra), and because the topic of coalgebras is not generally covered in detail in introductory mathematics or physics courses, it is of interest to have an article devoted to a consideration of the elementary part of the theory of coalgebras, directed to people from a variety of areas who wish to learn about coalgebras, and that is our aim here. More specifically, this article is intended as a primer on *ungraded, coassociative, counitary* coalgebras over a *field* for working mathematicians and physicists. Some familiarity with *ungraded, associative, unitary* algebras over a *field* can, therefore, be assumed. Moreover, it is natural to consider coalgebras with reference to algebras because coalgebras are defined *dually* to algebras in a sense to be made precise shortly, and because, at the *finite-dimensional* level, the *categories* of *finite-dimensional coalgebras* and *finite-dimensional algebras* are *dual* to one another, as stated at the beginning of Section 4 (cf. 2.86).

Before proceeding further, we stipulate that throughout this introduction, save explicit mention to the contrary, K will denote an *arbitrary* field; the term *vector space* will mean a *vector space over* K; and the term *linear map* will mean a *morphism of vector spaces*. We recall next that an *ungraded algebra over a field* K (or *algebra*, for short) is a *vector space* V together with a *binary operation* $V \times V \to V$ *linear* in each variable; i.e., it is a *vector space* V together with a *linear* map $\varphi : V \otimes V \to V$ called the *product* or *multiplication* of V. If the algebra has a *two-sided identity element* (or *unit element*), the algebra is said to be *unitary* (or *unital*). If the *multiplication* of the algebra is *associative*, the algebra is said to be *associative*. If the multiplication of the algebra satisfies the *Jacobi identity* and is *strongly anti-commutative*, the algebra is said to be a *Lie algebra*. Unfamiliar terms appearing in this introduction and needed later will be defined in due course. On occasion we shall direct the reader's attention to material collected in the appendices.

Coalgebras are defined "dually" to *algebras* as follows: Whereas, as above, an *algebra* is a *vector space* V together with a *linear* map $\varphi : V \otimes V \to V$, called the *multiplication* or *product* of V, a *coalgebra* is a *vector space* V together with a *linear* map $\Delta : V \to V \otimes V$, in the *opposite* direction, called the *comultiplication* or *coproduct* of V. The *counterpart* of a *unitary* algebra is a *unitary* coalgebra. The *counterpart* of an *associative* algebra is an *associative* coalgebra. The *counterpart* of a *Lie* algebra is a *Lie* coalgebra. (*Nota Bene*. In much of the literature, *unitary* coalgebras are referred to as *counitary* coalgebras; and *associative* coalgebras are referred to as *coassociative* coalgebras.) Later in this article, the *underlying vector space, V,* of an *algebra* (V, $\varphi : V \otimes V \to V$) will often be denoted by A; the *underlying vector space*, V, of a *coalgebra* (V, $\Delta : V \to V \otimes V$) will often be denoted by C; the *underlying vector space* of a *Lie algebra* will often be denoted by L; and the *underlying vector space* of a *Lie coalgebra* will often be denoted by M.

To obtain the *detailed* definition of an *associative* coalgebra, given in Section 2, in contradistinction to the general definition given above, we *first* need to express the *associativity* of the *multiplication* of an algebra in an *element-free* way by requiring the *commutativity* of a certain diagram (Figure 2.2(a)). Then the (*definition* of the) *associativity* of the *comultiplication* of a coalgebra is obtained by requiring the *commutativity* of the (so-called) *dual* diagram (Figure 2.6(a)) that one gets from the diagram expressing the associativity of the multiplication of an algebra by *reversing all the arrows* in that diagram, and relabeling the arrows appropriately (e.g., replacing each "φ" appearing as part of a label by "Δ"). Similarly, the defining property of a *unitary* algebra can be given diagrammatically (as in Figure 2.2(b)), once one realizes that if an algebra (V, $\varphi : V \otimes V \to V$) has a two-sided *identity element* 1_V then that element may be obtained as the image $\eta(1_K) =: 1_V$ (in the *center* of V) of the *identity element* 1_K of the *ground field* K under a *linear* map $\eta : K \to V$ making a certain diagram (Figure 2.2(b)) commute. Accordingly, the *counterpart* for a *coalgebra* (V, $\Delta : V \to V \otimes V$) of an *algebra* (V, $\varphi : V \otimes V \to V$) having a *two-sided unit*, meaning a *K-linear* map $\eta : K \to V$ making a certain diagram (Figure 2.2(b)) commute, is for the *coalgebra* (V, $\Delta : V \to V \otimes V$) to have a *two-sided counit*, meaning a *linear* map $\varepsilon : V \to K$, in the *opposite* direction, making the *dual* diagram (Figure 2.6(b)) commute. Summarizing: Having defined an *ungraded, associative, unitary algebra* (V, $\varphi : V \otimes V \to V$, $\eta : K \to V$) *diagrammatically*, one obtains the *definition* of *an ungraded, associative, unitary coalgebra* (V, $\Delta : V \to V \otimes V$, $\varepsilon : V \to K$) from the *dual* diagrams obtained by *reversing all the arrows of the corresponding diagrams* for (V, $\varphi : V \otimes V \to V$, $\eta : K \to V$), and *relabeling the arrows* appropriately. Details will be given in Section 2. From a categorical perspective (once one defines the relevant properties of an algebra or of a coalgebra in terms of its structure maps rather than its elements), then, either the definition of an *algebra* or the definition of a *coalgebra* may be taken as primary, a fact which inspired Sweedler [N-S, p. 57] to quip: " 'An algebra is defined by taking the defining diagrams for a coalgebra and reversing arrows.' – old coalgebraist's joke." In fact, just such an approach is taken by Jonah [Jon, p. 8] as we review briefly now and in greater detail below (in Remark 2.96) for the benefit of the categorically minded. The formal, technical device that must be added in order to display the precise way in which algebras and coalgebras are dual to one another is that of a *monoidal category* discussed in greater detail following 2.97, below (for those desiring such a discussion, and forbidden to the eyes of others). The general idea is this: First one defines an algebra to be an algebra over a certain monoidal category, and one likewise defines a coalgebra to be a coalgebra over the same monoidal category. Next, having defined the notion of an *algebra over a monoidal category*, one then, to paraphrase Jonah [Jon, p. 8], observes two things, namely, (1) that a *coalgebra over a monoidal category* is simply *an algebra over the opposite monoidal category*, and (2) that the *corresponding categories are dual to one another*. In the present case, for the benefit of the categorically minded, having defined an *ungraded algebra over a field K* to be an *algebra over the monoidal category* (**Vect**, $\otimes$), where **Vect** is the *category of K-vector spaces* and where $\otimes$: **Vect** $\times$ **Vect** $\to$ **Vect** is the *functor* which assigns to any pair of *K-vector spaces* their *tensor product over K*, one then has (in Jonah's words:) "by definition or observation" that *an ungraded coalgebra over a field K* is simply an *algebra* over the *opposite monoidal category* ($\mathbf{Vect}^{\mathrm{op}}$, $\otimes^{\mathrm{op}}$). Of course Jonah [Jon, p. 8], turns this situation on its head by, symmetrically, first defining a coalgebra over a monoidal category and then remarking that "(by definition or observation)"

(1) *an algebra over a monoidal category* is a *coalgebra over the opposite monoidal category*, and
(2) "the corresponding categories are dual."

At this juncture, to paraphrase a remark made by Jim Stasheff [Sta-2, p. 13] in connection with *Hopf algebras*, we note that "there are two main theologies regarding" coalgebras, "namely, *graded* and *ungraded*." The precise definitions will be given in Section 2, but here we note that categorically speaking, a *graded* algebra (respectively, a *graded* coalgebra) is an *algebra* (respectively, a *coalgebra*) *over the monoidal category* (**Graded Vect**, $\otimes_{\mathbf{g}}$), where **Graded Vect** is the *category of graded K-vector spaces* and where

$$\otimes_{\mathbf{g}} : \mathbf{Graded\,Vect} \times \mathbf{Graded\,Vect} \to \mathbf{Graded\,Vect}$$

is the *functor* which assigns to any pair of *graded K-vector spaces* their *graded tensor product* over K.

Historically, *graded coalgebras* arose in connection with *graded Hopf algebras* in the area of *algebraic topology* through the work of Hopf [Hopf-1] on the homology of Lie groups, while *ungraded coalgebras* arose in connection with *ungraded Hopf algebras* in the area of *algebraic geometry*

(1) through Cartier's analysis of Dieudonné's work on *hyperalgebras*, and
(2) through the work of Hochschild and Mostow [Hoc-Mos-1,Hoc-Mos-3] on the *representation theory* of *algebraic groups* and *algebraic Lie algebras*.

It was Pierre Cartier who (according to a private communication[1] from Yvette Kosmann-Schwarzbach) first observed that the notion of a *coalgebra* underlay both the work of Hopf in the area of *algebraic topology* and the work of Dieudonné in the area of *formal groups*.

In the remainder of this introduction, we shall list some other situations in which coalgebras and Hopf algebras have arisen. We shall go into details in a few instances. The reader unfamiliar with the topics mentioned should not despair, for beginning with Section 2, the rest of this paper will be fairly self-contained, requiring for the most part only a knowledge of linear and multi-linear algebra, especially, familiarity with the tensor product of vector spaces over a field. Background in category theory will be minimal, requiring just a familiarity with the terms *category, functor*, and *natural transformation* such as one would pick up by osmosis from one's general reading. In this connection, the book [Mac-Bir] by Mac Lane and Birkhoff should be more than adequate for the algebra as well as the category theory. Another especially gentle introduction to category theory (for non specialists) is the book [Gol] by Goldblatt. But our mentioning these books is *not* at all to suggest that

[1]Letter from Yvette Kosmann-Schwarzbach at the Centre de Mathématiques, Ecole Polytechnique, Palaiseu Cedex, France, dated September 22, 1994.

a reader of the present article should feel it necessary to consult either one before reading further. The same holds true for other references throughout. They are given to denote priority or for further optional reading.

For the rest of this introduction, we adopt an approach similar to that of Heinrich Tietze in his wonderful book [Tie], in that we shall try to provide a brief excursion through the exotic land (terrain) of coalgebras similar to a sight-seeing expedition on an alpine cable car ride for those who wish to enjoy the mountains without undergoing the ardure of having to hike in and out, either because they have not the time or are not in shape to do so. This concludes our introduction to the Introduction. We now continue with the introduction proper.

Historically, as noted briefly above, (graded) coalgebras arose out of the work [Hopf-1] of the Swiss topologist Heinz Hopf on the homology of compact Lie groups though the word *coalgebra* was coined much later by P. Cartier (cf. 2.15, below). In the situation considered by Hopf (described more fully below), the coalgebra structure on the homology was compatible with the algebra structure in the sense that the coalgebra structure maps were algebra morphisms. When a vector space is endowed simultaneously with the structure of an algebra and a coalgebra in such a way that the coalgebra structure maps are algebra morphisms, then the resulting object is called a *bialgebra* (or, sometimes, a *Hopf algebra*). In the case considered by Hopf, the homology bialgebra had additionally what is commonly called a "conjugation" or "antipode," to be defined in Section 2. Nowadays, it is customary to distinguish between a bialgebra and a Hopf algebra (= a bialgebra with antipode). But, in many respects, the word "bialgebra" seems preferable to the words "Hopf algebra" since the former suggests *two* structures whereas the latter suggest a special type of algebra as opposed to an algebra with additional structure.

In this spirit, it would seem appropriate to speak of *bialgebras* and of *Hopf bialgebras* (= those types of bialgebras that are special – like the one arising in Hopf's context). At all events, historically, coalgebras arose as part of a more elaborate structure called a Hopf algebra, and – to a large extent – individual coalgebras continue to be important in that context and because of that connection.

Of course the more one understands about coalgebras in their own right, the more one can exploit such knowledge in the Hopf context.

For example, for the benefit of those familiar with the basics of Lie theory, in the case of the universal enveloping algebra UL of a finite-dimensional Lie algebra L over a field of characteristic zero, a result of Harish-Chandra (cf. [H-C, p. 905, Theorem 1]) asserts that UL has "sufficiently many (finite-dimensional) representations" (to separate points), or, equivalently (cf. [Mi-5, p. 20, Lemma 14]), that the intersection $\mathcal{R}(UL)$ of all cofinite two-sided ideals of UL is zero. Harish-Chandra's original proof is long and involved, but one can give a short, non-computational, Hopf-theoretic proof of this result (cf. [Mi-5]) once one observes that $\mathcal{R}(UL)$ is a "coideal" of UL having zero intersection with the "primitives" of UL (cf. Appendix 5.1). This example illustrates why it is useful to study coalgebras as such. For if one knows a lot about coalgebras, one can sometimes exploit those facts to establish results about the algebra part of a Hopf algebra – results that at first blush simply involve the algebra structure with no apparent reference to (or awareness of) the hidden yet present underlying coalgebra structure.

In other instances, where one is conscious of the algebra and coalgebra portions of a Hopf algebra, it is the compatibility of those structures that is of primary importance. Historically this was so in Hopf's case though it must be mentioned that Hopf's paper [Hopf-1] in the *Annals of Mathematics* in 1941 was not written using modern terminology. If you like, it is with prophetic hindsight that we can look back at Hopf's original paper and understand that it was precisely this compatibility that was the crux of Hopf's result. In 1941, cohomology was not as developed as it is today so homology was used and even there the argument given is a *tour-de-force* in that it doesn't make use of the notion of tensor product, etc. But though Hopf worked with homology rather than cohomology (using the so-called "Umkehrhomomorphismus" introduced earlier by Freudenthal), the spirit of his argument is well captured by a cohomological explanation since the (rational) homology and cohomology Hopf algebras of a compact connected Lie group are dual to one another. For an excellent presentation of Hopf's work in a modern setting (from a cohomological perspective) the reader is referred to Dieudonné [D-3, pp. 234–242]. Here, however, we follow Samelson's illustration [Sa, pp. 12–13, §13] of Hopf's methods in sketching a much shorter cohomological argument used to prove that even-dimensional spheres can't be Lie groups.

By way of background, recall that a goal of algebraic topology is to establish topological results by algebraic means. Toward this end, one considers functors (covariant and contravariant) from the category **Top** of topological spaces and continuous functions to various categories of algebraic objects (e.g., groups, rings, vector spaces, ...) and their associated maps. For example, if X is a topological space, let $H^n(X)$ denote the nth *cohomology group of* X *with integer coefficients* and set

$$H^*(X) = \left\{H^n(X)\right\}_{n=0}^{\infty}.$$

This definition turns the *integral cohomology* of X into a *graded* group according to the convention that a *graded group* A is a family $A = \{A_n\}_{n=0}^{\infty}$ of *groups* A_n, one for each integer $n \geqslant 0$. A *morphism* $f : A \to B$ *of graded groups* $A = \{A_n\}_{n=0}^{\infty}$ and $B = \{B_n\}_{n=0}^{\infty}$ is a family $f = \{f_n\}_{n=0}^{\infty}$ of *group homomorphisms* $f_n : A_n \to B_n$, one for each integer $n \geqslant 0$. If A and B are graded groups, so is their *tensor product* $A \otimes B$ under the definition

$$(A \otimes B)_n = \bigoplus_{\substack{p+q=n,\\ p,q\geqslant 0}} A_p \otimes B_q.$$

It turns out that if $f : X \to Y$ is any continuous function, then f gives rise to a morphism

$$H^*(f) : H^*(Y) \to H^*(X)$$

of graded groups in such a way that the assignment

$$X \mapsto H^*(X) \quad \text{and} \quad f \mapsto H^*(f)$$

defines a *contravariant* functor

$$H^* : \mathbf{Top} \to \mathbf{Graded\ Groups}.$$

By the Künneth theorem together with the Eilenberg–Zilber theorem (cf. [Mac-1, pp. 166, 239]), if X is a topological space with the property that $H^n(X)$ is a finitely generated free Abelian group for all $n \geqslant 0$, then, for each integer $n \geqslant 0$, there exists a natural isomorphism

$$H^n(X \times X) \cong \bigoplus_{\substack{p+q=n,\\ p,q\geqslant 0}} H^p(X) \otimes H^q(X)$$

which may be written more compactly as

$$H^*(X \times X) \cong H^*(X) \otimes H^*(X)$$

according to the above conventions.

In this case, the *diagonal map*

$$d : X \to X \times X$$

given by

$$x \mapsto (x, x)$$

gives rise, by functoriality, to a *multiplication* (a morphism of graded groups)

$$\varphi := d^* : H^*(X) \otimes H^*(X) \xrightarrow{\cong} H^*(X \times X) \xrightarrow{H^*(d)} H^*(X)$$

called the *cup product* equipping $H^*(X)$ with the structure of a *graded ring* or *graded algebra over* $\mathbb{Z}$, the *ring of integers*. This means, in particular, that if $a \in H^p(X)$ and $b \in H^q(X)$, then

$$\varphi(a \otimes b) := a \cdot b := a \cup b \in H^{p+q}(X)$$

where $a \cup b$ denotes the *cup product* of a with b. With this enrichment, H^* is a functor from **Top** to **Graded Rings**. Now, if X is in addition a topological group, then the *multiplication* $m : X \times X \to X$ on X induces a *comultiplication*

$$\Delta := m^* : H^*(X) \xrightarrow{H^*(m)} H^*(X \times X) \xrightarrow{\cong} H^*(X) \otimes H^*(X)$$

on $H^*(X)$ equipping $H^*(X)$ with the structure of a *graded* "*co-ring*" or *graded coalgebra over* $\mathbb{Z}$ (in fact, the structure of a *graded* "*bi-ring*" or *graded bialgebra over* $\mathbb{Z}$). The isomorphism

$$H^*(X \times X) \cong H^*(X) \otimes H^*(X)$$

is that of graded rings (in fact, of graded bi-rings), where $H^*(X) \otimes H^*(X)$ has the structure of a graded ring (as the *tensor product of graded rings*) according to the definition

$$(a \otimes b) \cdot (c \otimes d) = (-1)^{pq}\, a \cdot c \otimes b \cdot d$$

for $a \in H^m(X)$, $b \in H^p(X)$, $c \in H^q(X)$, and $d \in H^r(X)$.

The fact that the comultiplication

$$\Delta : H^*(X) \to H^*(X) \otimes H^*(X)$$

is a morphism of graded rings (one of the defining conditions for a bi-ring or a Hopf ring) can be used to prove that even-dimensional spheres cannot be Lie groups. The argument goes as follows:

Let $H^* := H^*(S^n) := \{H^p(S^n)\}_{p=0}^{\infty}$ denote the *cohomology of the n-sphere S^n with integer coefficients*, i.e., the *integral cohomology of S^n*. It is known [Mun, p. 281, Corollary 47.2] that

$$H^0 := H^0(S^n) = \mathbb{Z} \quad \text{and} \quad H^n := H^n(S^n) = \mathbb{Z}$$

while

$$H^p := H^p(S^n) = 0 \quad \text{if } p \neq 0, n.$$

Let $e \in H^0 = \mathbb{Z}$ be a generator of H^0. One can choose e so that e serves as a *two-sided identity element* for H^*. If $x \in H^n = \mathbb{Z}$ is a *generator* of H^n, then $x^2 = 0$ for dimensional reasons:

$$x \in H^n \quad \Rightarrow \quad x^2 = x \cdot x = x \cup x \in H^{n+n} = H^{2n} = 0.$$

On the other hand

$$(H^* \otimes H^*)^n := \bigoplus_{\substack{p+q=n,\\ p,q \geqslant 0}} H^p \otimes H^q = H^0 \otimes H^n + H^n \otimes H^0$$

so using the fact that S^n has a two-sided "unit" – which in turn furnishes $H^*(S^n)$ with a two-sided "counit" – one can readily show that

$$\Delta(x) = e \otimes x + x \otimes e.$$

Because Δ is a morphism of graded rings,

$$x^2 = 0 \quad \Rightarrow \quad 0 = \Delta(x^2) = (\Delta x)^2.$$

But, by the sign conventions for multiplication in the graded ring $H^* \otimes H^*$, one finds (since n is even!) that

$$\begin{aligned}
(\Delta x)^2 &= (e \otimes x + x \otimes e)^2 \\
&= (e \otimes x)^2 + (e \otimes x) \cdot (x \otimes e) + (x \otimes e) \cdot (e \otimes x) + (x \otimes e)^2 \\
&= e^2 \otimes x^2 + 2(x \otimes x) + x^2 \otimes e^2 \\
&= 2(x \otimes x) \neq 0
\end{aligned}$$

so one has a contradiction.

NOTE. $x \in H^n \Rightarrow (e \otimes x) \cdot (x \otimes e) = (-1)^{n^2}(e \cdot x \otimes x \cdot e) = x \otimes x$ since n is even.

This contradiction shows that the assumption that S^n is a Lie group is untenable whenever n is an even integer > 0. In fact, the above proof (as well as Hopf's original proof) only uses the fact that S^n is a pointed topological space having a multiplication and a two-sided homotopy unit – a fact realized by Hopf. Such spaces are called *H-spaces* or *Hopf spaces* in honor of Hopf, and the above proof shows that S^n is not an H-space in case n is even. A celebrated theorem of J. Frank Adams shows that S^n is an H-space only if $n = 0, 1, 3$ or 7 – cf. Adams ([Ad-1] and [Ad-2]). Actually, S^0, S^1, and S^3 are Lie groups while S^7 is not even homotopy associative (cf. Whitehead [Wh-2, p. 119, Example 2]). Here

$S^0 =$ $\mathbb{Z}/2\mathbb{Z} =$ two points,
$S^1 =$ the unit sphere (elements of modulus 1) in the field of *complex numbers*,
$S^3 =$ the unit sphere in the division algebra of *quaternions*, and
$S^7 =$ the unit sphere in the division algebra of *octonions* or *Cayley numbers*.

The interested reader is also referred to [S-S, pp. 30–31, especially to Propositon 3.3.3].

In contrast to the two cases considered so far, coalgebras can be important even if they are not the underlying coalgebra of a Hopf algebra (or bialgebra).

Our reason is that, as we shall see later, the vector space dual of a coalgebra always carries the structure of an algebra. Whenever an algebra arises, in this way, as the *dual algebra* of a coalgebra, it is helpful to realize that. For it is frequently the case that certain properties of a coalgebra (for instance, associativity) imply the corresponding property for the dual algebra yet are practically self-evident at the coalgebra level while not so immediately transparent at the algebra level.

A second, often related, reason is that, in the words of Joni and Rota [J-R, p. 2], coalgebras seem to provide "a valuable formal framework for the study of combinatorial problems". By way of illustration, in combinatorics, the *umbral calculus*, developed by Gian-Carlo Rota and his school (cf. [R-K-O, p. 74] and [R-R]), has its foundation in the observation by Cartier [Ca-1, pp. 3–13, Proposition 6; pp. 3–14, Proposition 7] that the *algebra of formal power series* is the *dual algebra* of the *divided power coalgebra* (of which we shall have more to say in Section 2). It turns out that the coalgebra context leads to a better understanding of "the classical technique of treating indices as exponents, from which the umbral calculus derives its name" (from [J-R, p. 26]).

Because the connection between formal power series and generating functions is so well known (cf. [D-R-S]), the existence of a *coalgebra–combinatorics* interplay should not be too surprising. But there are other points of contact between these two disciplines. The fact is that coalgebras have appeared in a variety of ways in recent work in combinatorics. To convince oneself of this, one need just look, for example, at the article of Joni and Rota [J-R] or at the more recent works of Schmitt [Schmi], Haiman and Schmitt [Hai-Sch], Hirschhorn and Raphael [Hirs-Raph], or Spiegel and O'Donnell [S-O]. These circumstances have prompted some to suggest that coalgebras may be more suited to combinatorics than algebras. After all, algebras result from putting things together, so coalgebras should furnish a natural setting for taking things apart. From that standpoint, bialgebras ought to provide a natural framework for assembling and disassembling various prescribed entities – a task with which combinatorics is frequently concerned.

One way of trying to give form to these perhaps speculative sounding remarks is to begin by recalling that the algebra structure on the cohomology of a Lie group was furnished by a map

$$H^* \otimes H^* \to H^*$$

whereas the coalgebra structure was provided by a map

$$H^* \to H^* \otimes H^*$$

in the opposite direction. If we interpret this turning around of arrows somewhat loosely as implying a reversal of order, and if we think of the tensor product sign, $\otimes$, as a sort of place holder, then we arrive at the following informal contrast of the *algebra-coalgebra* situation.

In algebra, you put two things together – you combine them – to make a third thing. In coalgebra, you take something apart – you decompose it into its constituent parts. You cut it up into the "original" atoms – basic parts – from which it was assembled. This being the case, coalgebras ought to arise quite naturally in the study of number theory, combinatorics, and "puzzles" of various kinds (cf. [J-R, pp. 11–17]).

One way to get a coalgebra is to "undo" an algebra. If the algebra that is undone has an associative multiplication, then you get what is usually called a coalgebra, namely, an *associative coalgebra*. If the algebra that is undone has a Lie structure, then you get a *Lie coalgebra*. Breaking something up into its constituent parts is frequently not unique. That is to say: an object appearing in an algebra can frequently be obtained in several ways. So as not to slight any of these ways, we take them all together – we "add" them up. Thus, for example, for any integer $n \geqslant 0$, we have

$$x^n = x^k \cdot x^{n-k} \quad \text{for all integers } k \text{ with } 0 \leqslant k \leqslant n.$$

Therefore, we expect that

$$\Delta\left(x^n\right) = \sum_{k=0}^{n} (\ )x^k \otimes x^{n-k}$$

where the coefficient ( ) *counts* the number of ways of picking k objects from n objects. We are thus led to *define*

$$\Delta\left(x^n\right) = \sum_{k=0}^{n} \binom{n}{k} x^k \otimes x^{n-k}$$

where $\binom{n}{k}$ denotes the binomial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Alternatively, we can think of x as being *primitive* (or prime, or basic) because

$$x = 1 \cdot x = x \cdot 1$$

and this primitivity of x is then reflected in the fact that (or translates to the requirement that)

$$\Delta(x) = 1 \otimes x + x \otimes 1$$

i.e., that

$$\Delta\left(x^1\right) = x^0 \otimes x^1 + x^1 \otimes x^0.$$

Of course, if one wants to construct a bialgebra structure on $K[x]$, then the formula for $\Delta(x^n)$ is a consequence of the requirement (one of the conditions defining a bialgebra) that Δ be an algebra map: for if A is an *ungraded associative* algebra, then so is $A \otimes A$ with the definition that

$$(a \otimes b) \cdot (c \otimes d) = a \cdot c \otimes b \cdot d.$$

Since

$$(1 \otimes x) \cdot (x \otimes 1) = x \otimes x = (x \otimes 1) \cdot (1 \otimes x),$$

it follows that $1 \otimes x$ and $x \otimes 1$ commute in the associative algebra $K[x] \otimes K[x]$. Thus, the binomial formula can be applied to expand

$$(1 \otimes x + x \otimes 1)^n$$

to obtain the desired result for $\Delta(x^n) = (\Delta x)^n$.

One might well ask: "What is the combinatorial significance of the requirement that Δ be an algebra map?" The *suspicion* is that this is a type of combinatorial compatibility: "Decomposition should respect composition. Cutting-up should be compatible with gluing-together. Unglue what was glued, or glue what was cut-up."

As a second example of obtaining a coalgebra by "undoing" an algebra, we consider an example from number theory. Following Graham, Knuth, and Patashnik (cf. [G-K-P, p. 102]), we write $p \backslash n$ to denote the fact that p *divides* n (throughout this example, p and n shall be positive integers). Since

$$n = p \cdot \frac{n}{p} \quad \text{whenever } p \backslash n,$$

we would like to have

$$\Delta(n) = \sum_{\substack{p \backslash n, \\ p \geqslant 1}} p \otimes \frac{n}{p} = \sum_{\substack{p \cdot q = n, \\ p, q \geqslant 1}} p \otimes q.$$

But, our algebras and coalgebras should be built on vector spaces over a field K, so to give meaning to the above we are led (following Joni and Rota [J-R, pp. 16–17]) to *define* D to be the vector space over K having as basis the set

$$\{n^x \mid n = 1, 2, 3, \dots\}$$

of variables n^x, one for each positive integer n. Then $\Delta : D \to D \otimes D$ is taken to be the vector space map defined by the requirement that

$$\Delta(n^x) = \sum_{\substack{p \cdot q = n, \\ p, q \geqslant 1}} p^x \otimes q^x.$$

(The *counit* $\varepsilon : D \to K$, to be introduced in the next section, is defined by the requirement that $\varepsilon(1^x) = 1$ while $\varepsilon(n^x) = 0$ if $n > 1$.) As Joni and Rota point out [J-R, p. 17], the coalgebra D has additionally a natural algebra structure given by

$$m^x \cdot n^x = (m \cdot n)^x.$$

But, D is *not* a bialgebra though if m and n are *relatively prime*, then

$$\Delta(m^x \cdot n^x) = \Delta(m^x) \cdot \Delta(n^x).$$

It turns out that if D^* denotes the *dual algebra* of the coalgebra D, then

$$D^* \cong \text{the algebra of formal Dirichlet series}$$

with isomorphism given by

$$f \mapsto \varphi(s) = \sum_n \frac{f(n^x)}{n^s}$$

(cf. Apostol [Ap, pp. 29–30, §2.6]). D is called the *Dirichlet coalgebra*.

REMARK. D^* is a commutative, associative algebra under the so-called *Dirichlet product* (or *Dirichlet convolution*) precisely because D is a commutative, associative coalgebra. At the coalgebra level, these properties are truly self-evident (cf. 2.100, 2.101).

We conclude our examples (of coalgebras obtained by "undoing" algebras) by sampling a Lie theoretic flavor of the above.

In the case of the Lie algebra $(\mathbb{E}^3, \times)$ consisting of Euclidean 3-space together with *Lie bracket* given by the *vector cross product*, we have – for *basis vectors* $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ – that

$$\begin{aligned} &[\hat{\imath}, \hat{\jmath}] = \hat{k} = -[\hat{\jmath}, \hat{\imath}], \\ &[\hat{\jmath}, \hat{k}] = \hat{\imath} = -[\hat{k}, \hat{\jmath}], \quad \text{and} \\ &[\hat{k}, \hat{\imath}] = \hat{\jmath} = -[\hat{\imath}, \hat{k}]. \end{aligned}$$

Since our "rule" for obtaining a coalgebra from an algebra is to *add up* all the basic tensors whose left and right factors multiply together to give the coalgebra generator one is trying to decompose, we get, upon setting $e_1 := \hat{i}$, $e_2 := \hat{j}$, and $e_3 := \hat{k}$, that

$$\begin{aligned} \Delta(e_1) &= e_2 \otimes e_3 - e_3 \otimes e_2, \\ \Delta(e_2) &= e_3 \otimes e_1 - e_1 \otimes e_3, \quad \text{and} \\ \Delta(e_3) &= e_1 \otimes e_2 - e_2 \otimes e_1. \end{aligned}$$

This Δ equips Euclidean 3-space with the structure of a *Lie coalgebra* (cf. [Mi-2, p. 5, Example 1]). However, it turns out that this Δ does not equip the Lie algebra $(\mathbb{E}^3, \times)$ with the structure of a *Lie bialgebra* (cf. [Mi-9]).

As a further illustration of how coalgebras have appeared historically, we describe next a general method for producing coalgebras, namely, we outline a way of associating to a given group G (finite, topological, Lie, or affine algebraic) the underlying coalgebra of a commutative Hopf algebra. This construction has its basis in the observation that if G is a group and K is a field and if K^G denotes the set of all functions $f : G \to K$, then K^G is a commutative algebra when addition, multiplication, and scalar multiplication are defined pointwise. Accordingly, if one can define a subalgebra $\mathrm{Fun}(G)$ of K^G in such a way that

$$\mathrm{Fun}(G \times G) = \mathrm{Fun}(G) \otimes \mathrm{Fun}(G)$$

with the assignment

$$G \mapsto \mathrm{Fun}(G)$$

giving rise to a *contravariant* functor from that category of groups of which G is a member to that category of algebras of which $\mathrm{Fun}(G)$ is a member, then the *multiplication*

$$G \times G \xrightarrow{m} G$$

on G will induce a *comultiplication* on $\mathrm{Fun}(G)$ via

$$\mathrm{Fun}(G) \xrightarrow{\mathrm{Fun}(m)} \mathrm{Fun}(G \times G) = \mathrm{Fun}(G) \otimes \mathrm{Fun}(G)$$

with $\mathrm{Fun}(G)$ the underlying coalgebra of a *commutative* Hopf algebra. [The *counit* $\varepsilon : \mathrm{Fun}(G) \to K$ of $\mathrm{Fun}(G)$ will be the *algebra map* $\mathrm{Fun}(G) \to \mathrm{Fun}(\{e\}) \cong K$ induced by the inclusion $\{e\} \hookrightarrow G$ where $e =$ the *identity element* of G, while the *antipode* of $\mathrm{Fun}(G)$ will be the *algebra map* $S := \mathrm{Fun}(i) : \mathrm{Fun}(G) \to \mathrm{Fun}(G)$ induced by the *inverse map* $i : G \to G$. Thus, $\varepsilon(f) := f(e)$ while $S(f)(x) := f(x^{-1})$.]

To implement the procedure just outlined, we make use of the fact (cf. [H-5, p. 2, Proposition 1.2]) that, for any field K and any non-empty sets X and Y, there is a canonical morphism

$$\pi : K^X \otimes K^Y \to K^{X \times Y}$$

of K-algebras given by

$$\pi(\Sigma f \otimes g)(x, y) = \Sigma f(x) \cdot g(y)$$

with the property that π is *injective* and that the image of π consists of all functions h such that the K-space spanned by the partial functions h_y, where y ranges over Y and $h_y(x) := h(x, y)$, is *finite-dimensional*.

We consider four examples:

(1) When G is a *finite group*, one can take $\mathrm{Fun}(G)$ equal to all of K^G since in this case the canonical algebra injection $\pi : K^G \otimes K^G \to K^{G\times G}$ is an *isomorphism* for dimensional reasons. For G *finite*, the *Hopf algebra* K^G turns out (cf. 3.27 and 3.29, below) to be *the dual Hopf algebra* of *the group algebra* $K[G]$ mentioned as the third of the next set of three examples and considered more fully in Section 3.

(2) When G is a *topological group* and when $K = \mathbb{R} :=$ the field of *real* numbers or $K := \mathbb{C} :=$ the field of *complex* numbers, one can take

$$\mathrm{Fun}(G) = \mathcal{R}(G) := \mathcal{R}_K(G) := \text{the } K\text{-algebra of all } K\text{-valued } \textit{continuous} \\ \textit{representative functions on } G$$

where, by definition, a function $f : G \to K$ is called a *representative function* in case the subspace of K^G spanned by all translates $x \cdot f \cdot y$ of f as x and y range over G is *finite-dimensional* (by definition, $(x \cdot f \cdot y)(z) = f(y \cdot z \cdot x) \in K$) since in this case $\pi : K^G \otimes K^G \to K^{G\times G}$ induces an *algebra isomorphism*

$$\mathcal{R}(G) \otimes \mathcal{R}(G) \xrightarrow{\cong} \mathcal{R}(G \times G),$$

cf. Hochschild [H-2, p. 26, Lemma 3.1].

(3) When G is a *Lie group* and $K = \mathbb{R}$ or $\mathbb{C}$, one can take

$$\mathrm{Fun}(G) = \mathcal{R}(G) \cap C^{\omega}(G)$$

where $C^{\omega}(G)$ denotes the space of all K-valued *analytic* functions defined on G, cf. Abe [Ab, p. 74, Remark].

(4) When G is an *affine algebraic group*, one can take

$$\mathrm{Fun}(G) = \mathcal{P}(G) := \text{the } K\text{-algebra of all } \textit{polynomial functions} \text{ defined on } G,$$

cf. Hochschild [H-5, p. 10] or Manin [Man-1, p. 13].

In this context we note (cf. Bröcker and tom Dieck [B-t.D, pp. 151–156]) that when G is a *compact Lie group* and $K = \mathbb{C}$, then $\mathcal{R}_{\mathbb{C}}(G)$ is a *finitely generated commutative* Hopf algebra over $\mathbb{C}$. Set

$$G_{\mathbb{C}} := \mathrm{Hom}_{\mathbb{C}-\mathbf{alg}}\big[\mathcal{R}_{\mathbb{C}}(G), \mathbb{C}\big] := \mathbf{Alg}_{\mathbb{C}}\big[\mathcal{R}_{\mathbb{C}}(G), \mathbb{C}\big] \\ := \text{the collection of all } \mathbb{C}\textit{-algebra homomorphisms from } \mathcal{R}_{\mathbb{C}}(G) \textit{ to } \mathbb{C}.$$

Then $G_{\mathbb{C}}$ has the structure of an *affine algebraic group* called the *complex algebraic hull* of G.

NOTE. It turns out (cf. [Ab, p. 159]) that the *contravariant functor*

$$\Phi : \textbf{Groups} \to \textbf{Comm Hopf Alg},$$

from the *category of groups* to the *category of commutative Hopf algebras*, which assigns to any *group* its *Hopf algebra* of *representative functions* (cf. [Ab, p. 72]), is *adjoint on the right* to the *contravariant functor*

$$\Psi : \textbf{Comm Hopf Alg} \to \textbf{Groups}$$

which assigns to any *commutative Hopf algebra* H the *group* $\mathrm{Hom}_{\mathbf{Alg}}[H, K]$ of all *algebra maps* $f : H \to K$, a *group under convolution* (cf. 2.58(a), below, or [Hey-Sw-1, p. 204, Proposition 1.5.2 (1)] or [Ab, p. 65, Theorem 2.1.5]). Additionally, Φ induces on each of four subcategories of the category of groups (finite, compact topological, compact Lie, or affine algebraic) a *contravariant functor* having a *contravariant inverse* and hence yielding an *anti-equivalence* between each of the four subcategories of **Groups** mentioned above and a corresponding subcategory of the category **Comm Hopf Alg** that can be completely characterized. Applied to the category **Cpt Top Groups** of *compact topological groups*, the resulting theorem giving the *duality* is *Hochschild's version* of the *Tannaka Duality Theorem* (cf. [H-2, p. 30, Theorem 3.5]), hereafter referred to as the *Tannaka Hochschild Duality Theorem*. It states that there is a *categorical duality* (*anti-equivalence*) between the *category of compact topological groups* and a certain subcategory of the category of co-commutative $\mathbb{R}$-Hopf algebras which we chose to call *Tannaka Hochschild Duality Hopf Algebras*. These are *commutative* $\mathbb{R}$*-Hopf algebras* satisfying the following two conditions:

(1) The set $\mathrm{Hom}_{\mathbf{Alg}}[H, \mathbb{R}]$ of all $\mathbb{R}$-algebra morphisms $f : H \to \mathbb{R}$ separates points, meaning that if $x, y \in H$ with $x \neq y$, then there exists an $f \in \mathrm{Hom}_{\mathbf{Alg}}[H, \mathbb{R}]$ such that $f(x) \neq f(y)$

and

(2) H supports a *positive definite left (invariant) integral* meaning an element $J \in H^*$ such that

$$J\left(f^2\right) > 0 \quad \text{whenever } f \neq 0$$

and such that

$$\rho_H \circ (1_H \otimes J) \circ \Delta = \eta \circ J \in \mathrm{Hom}_{\mathbb{R}}[H, H]$$

where $\rho_H : H \otimes_{\mathbb{R}} \mathbb{R} \to H$ is the canonical natural isomorphism, where $\Delta : H \to H \otimes H$ is the *comultiplication* of H (defined in 2.6), and where $\eta : \mathbb{R} \to H$ is the *unit* of H (defined in 2.2).

The *contravariant functors*

$$\textbf{Cpt Top Gps} \underset{\Psi}{\overset{\Phi}{\rightleftarrows}} \textbf{THD Hopf Alg}$$

establishing the claimed *duality* are given by

$$\Phi(G) := \mathcal{R}_{\mathbb{R}}(G) \cap C_{\mathbb{R}}(G) =: \mathcal{R}(G)$$
$$:= \text{all real-valued continuous representative functions on } G$$

and

$$\Psi(H) := \mathrm{Hom}_{\mathbf{Alg}}(H, \mathbb{R}) \equiv G\bigl(H^0\bigr)$$

where H^0 is the *dual Hopf algebra* of H (defined in 3.48) and $G(H^0)$ is the *group* of *group-like elements* of H^* (cf. 2.65, 2.67, and 3.26, below). When G is a *compact topological group*, $J : \mathcal{R}(G) \to \mathbb{R}$ is given by the *Haar integral* (cf. [H-2, p. 9], [Ab, p. 144], [Mon-1, p. 26, Example 2.4.5], or [Char-Pres, p. 115, Example 4.1.18; p. 454]). For further explicit details, see, for example [Tannaka], [Chevalley-1, p. 211], [H-2, p. 30, Theorem 3.4.3], [Mon-1, pp. 159–160, especially item 9.3.3], and [Char-Pres, pp. 106, 115–116, 135, 147].

The *Tannaka Duality Theorem* as presented in [Tannaka] and [Chevalley-1] is not stated in terms of Hopf algebras. Rather, the Hopf algebra version of that theorem is due to Hochschild who proved (cf. [H-2, p. 30, Theorem 3.5]) that certain types of $\mathbb{R}$-Hopf algebras appear precisely as the $\mathbb{R}$-Hopf algebra of real-valued representative functions on a certain compact topological group obtained from the Hopf algebra. These are, in Hochschild's terminology, "reduced Hopf algebras having a symmetry and a gauge." Above, we recast Hochschild's result using the by now more standard terminology as found in [Ab]. Montgomery [Mon-1] states a version of Hochschild's result over $\mathbb{C}$ and mentions Tannaka's original result as well as more recent categorical formulations regarding "recovering a group or Hopf algebra from a suitable category of its representations" (quoted from [Mon-1, p. 160]). Chari and Pressley (cf. [Char-Pres, p. 106, Example 4.1.7; p. 116, Example 4.1.19]) present an alternate way of looking at representative functions on a group, state (what we choose to call) the *Tannaka Hochschild Duality Theorem* over $\mathbb{R}$, and note that if, additionally, the Hopf algebra is *finitely generated* as an algebra, then the group G of algebra morphisms $f : G \to \mathbb{R}$ (which becomes a compact topological group when equipped with the so-called *finite-open topology*) "can be given the structure of a compact Lie group and conversely" (quoted from Example 4.1.19 on p. 116). [Char-Pres, pp. 135–149] also consider more generally what conditions on a category "will guarantee that it is equivalent to a category of representations (or corespresentations) of a Hopf algebra of some special type" (quoted from the Remark on p. 149). Abe [Ab, pp. 161–162] explains more completely why the category of compact Lie groups and the category of finitely generated commutative $\mathbb{R}$-Hopf algebras satisfying conditions (1) and (2), above, are anti-equivalent; and then proves, as a consequence of the Fundamental Theorem of Coalgebras (cf. 4.12 and 5.3, below), that every *compact topological group* is a *projective limit* of *compact Lie groups*. Hofmann [Hof] presents a beautiful treatment of an analog of Hochschild's version of the Tannaka Duality Theorem for compact semigroups in that he establishes an anti-equivalence between the category of *compact Hausdorff topological semigroups* and the category of *commutative C^*-bialgebras* thereby generalizing

(1) the *duality* (cf. [Hof, p. IV, Introduction; p. 135, Theorem 15.8]) between *compact abelian groups* and *discrete abelian groups* (discovered by Pontryagin in the early 1930s);

(2) the *duality* (cf. [Hof, p. IV, Introduction, p. 33, Theorem 6.6]) between *compact spaces* and *commutative C^*-algebras* (due to Gelfand and Naimark in the late 1930s);

and

(3) The Tannaka Duality Theorem as described by Tannaka in the late 1930s and by Hochschild in the late 1950s and early 1960s.

Schikhof (cf. [Schik, p. 231, Theorem 6.1]) proves a *non-archimedean* analog of Hochschild's version of the Tannaka Duality Theorem in that he establishes a *category anti-equivalence* between the *category* of 0-*dimensional Hausdorff semigroups with identity* (i.e., monoids) and the *category* of *commutative K-Hopf algebras generated by idempotents*, K being a *field* with a *non-archimedean valuation*. Using that result together with the Fundamental Theorem of Coalgebras, Schikhof then establishes (cf. [Schik, p. 231, Corollary 6.2]) the following result well known to people working in the area of semigroups (cf. [Hof-Mos-1, p. 52, Proposition 8.10]): Every 0-dimensional Hausdorff semigroup with identity is a projective limit of finite semigroups.

For a further consideration of the *non-archimedean* case, the reader is referred to [Diarra-1] and [Diarra-2].

For a nice treatment of various other duality theorems (e.g., those of Kreĭn, Kac–Paljutkin, Eymard, Tatsuuma, Ernest, and Takesaki), the reader is referred to [Enoc-Schw].

Each of the four examples considered above before the intervening NOTE illustrates the general principle according to which if you have a reasonable collection of scalar-valued functions defined functorially on a category of spaces closed under finite products, then morally speaking the function space on a finite product of those spaces ought to be the tensor product of the corresponding function spaces. In purely algebraic contexts the $\otimes$-product sign will be the usual algebraic tensor product sign – as above. But in certain topological contexts – such as the case of the collection $C^\infty(G)$ of all *smooth* functions on a Lie group G – *we may need to complete the usual tensor product in a nice way* since, in the case just cited (cf. Kirillov [Kir, p. 148, §10.4]),

$$C^\infty(G \times G) \cong C^\infty(G) \mathbin{\widehat{\otimes}} C^\infty(G).$$

If we permit ourselves this type of flexibility in the requirement that

$$\mathrm{Fun}(G \times G) = \mathrm{Fun}(G) \otimes \mathrm{Fun}(G),$$

then, depending on the specific context (i.e., the specific nature of G, $\mathrm{Fun}(G)$, and $\otimes$), we may encounter a variety of kindred notions, for instance, *graded coalgebras*, *complete coalgebras*, *Banach coalgebras*, *C^*-coalgebras*, etc. The common feature of all these examples is that each coalgebra so obtained is the underlying coalgebra of a Hopf algebra whose *multiplication is commutative*, for we are dealing in each case with a function ring and to multiply two functions you just multiply their values – which are numbers (i.e., scalars). On the other hand, the comultiplication may be non-commutative though it will be commutative for a commutative group (in fact, precisely for a commutative group).

We can generalize the class of examples just considered in yet another way:

In the case of the K-algebra $\mathcal{R}(G) = \mathcal{R}_K(G)$ of all K-valued *continuous representative functions* on a topological group G (where $K = \mathbb{R}$ or $\mathbb{C}$), the property of the *contravariant functor* $\mathcal{R}$ that made that example work was that $\mathcal{R}$ was *multiplicative* in the sense that

$$\mathcal{R}(G \times G) \cong \mathcal{R}(G) \otimes \mathcal{R}(G).$$

But there are also *covariant multiplicative functors* which furnish us with examples of coalgebras that appear as the underlying coalgebra of a Hopf algebra though in this case the Hopf algebra so obtained is one whose comultiplication is commutative – for reasons discussed below.

We consider three examples:

(1) Let H_* be the (covariant) functor which assigns to each connected topological space X having the homotopy type of a finite CW complex its homology groups with rational coefficients. For each such space X, the *diagonal map*

$$d : X \to X \times X$$

sending x to (x, x) induces a map

$$H_*(X) \xrightarrow{H_*(d)} H_*(X \times X) \xrightarrow{\cong} H_*(X) \otimes H_*(X)$$

of connected graded algebras that equips $H_*(X)$ with the structure of a connected graded commutative coalgebra. If, additionally, X has the structure of a Lie group G (or, even, of a homotopy associative H-space), then $H_*(G)$ will have the structure of a connected graded Hopf algebra whose *comultiplication* will be *commutative* (in the graded sense); in this case the multiplication $G \times G \xrightarrow{m} G$ on G will give rise to a multiplication

$$H_*(G) \otimes H_*(G) \xrightarrow{\cong} H_*(G \times G) \xrightarrow{H_*(m)} H_*(G)$$

on $H_*(G)$, cf. Milnor and Moore [M-M-2, p. 211].

NOTE. By definition, a graded algebra or coalgebra $A := \{A_n\}_{n=0}^{\infty}$ is *connected* if its component, A_0, in degree 0 is isomorphic to the ground field. Here, the fact that the 0th homology group $H_0(X)$ is isomorphic to the ground field $\mathbb{Q}$ is a consequence (and reflection) of the *connectivity* of the topological space X.

(2) Let $U : \mathbf{Lie\,Alg} \to \mathbf{Alg}$ be the (covariant) functor from the category of Lie algebras to the category of associative unitary algebras which assigns to each *Lie algebra* L its *universal enveloping algebra* UL. The *diagonal map* $d : L \to L \times L$ given by $x \mapsto (x, x)$ is a map of Lie algebras (when the Lie bracket of $L \times L$ is defined coordinate-wise), so d induces an algebra map

$$\Delta : U(L) \xrightarrow{U(d)} U(L \times L) \xrightarrow{\cong} U(L) \otimes U(L)$$

equipping UL with the structure of a Hopf algebra whose *comultiplication* is *commutative*. Under the map $\Delta : UL \to UL \otimes UL$

$$ix \mapsto ix \otimes 1_{UL} + 1_{UL} \otimes ix$$

for each $x \in L$, where $i : L \to \mathcal{L}(UL)$ is the canonical injective morphism of the Lie algebra L into the Lie algebra $\mathcal{L}(UL)$ associated to the associative algebra UL and where 1_{UL} denotes the identity element of UL.

(3) Let $K : \mathbf{Groups} \to \mathbf{Alg}$ be the (covariant) functor which assigns to each *group* G its *group algebra* $K[G]$ – cf. Section 3, below. The *diagonal map* $d : G \to G \times G$ given by $x \mapsto (x, x)$ is a group homomorphism and so induces an algebra map

$$\Delta : K[G] \xrightarrow{K(d)} K[G \times G] \xrightarrow{\cong} K[G] \otimes K[G]$$

equipping $K[G]$ with the structure of a Hopf algebra whose *comultiplication* is *commutative*. Under the map $\Delta : K[G] \to K[G] \otimes K[G]$,

$$g \mapsto g \otimes g \quad \text{for all } g \in G.$$

When the group G is finite, $K[G]$ is a finite-dimensional Hopf algebra whose dual Hopf algebra is K^G (cf. 3).

The common feature of all these examples is that each coalgebra so obtained is the underlying coalgebra of a Hopf algebra whose *comultiplication* is *commutative*; for the comultiplication is induced by the *diagonal map* d which sends x to (x, x), and thus the commutativity of the comultiplication is a consequence of the fact that

$$d = \tau \circ d$$

where τ is the map sending (x, y) to (y, x). By contrast, the *algebra structure* on UL will not be *commutative* unless the *bracket* of L is *trivial*; the algebra structure on $K[G]$ will not be *commutative* unless the *multiplication* of G is *commutative*; and the algebra structure on $H_*(G)$ will not be *commutative* unless the multiplication of G is *chain homotopy commutative*. Yet, the *homology ring* of an H-space can be *associative* and *commutative* even though the *multiplication* of that space is neither *homotopy associative* nor *homotopy commutative*. An example of an H-space for which this is so is given by the 7-sphere S^7 (cf. [Sta-1, pp. 69–70] and also [Lin, p. 1102, Paragraph 1]).

It turns out that there is a categorical framework to explain just why these and related examples work. That setting is provided by the notion (due to Bénabou – cf. Bénabou [Bén] or Kleisli [Kleis]) of a *multiplicative category* and *multiplicative functors* (covariant and contravariant) between two such – though these days *multiplicative categories* are usually called *monoidal categories* (cf. [Eil-Kel], [Mac-4] or [Kleis]) or *tensor categories* (cf. [J-S-1], [J-S-3] or [Kass]). When one views algebras from this somewhat broader (though, we feel, quite natural) perspective as *algebras over a monoidal category*, then one encounters such diverse examples of algebras as ungraded algebras, graded algebras, differential graded algebras, complete algebras, profinite (or linearly compact) algebras, C^*-algebras, von Neumann algebras, triples, etc. From this perspective, a coalgebra is – in the same spirit – a *coalgebra over a monoidal category*, where, by definition, *a coalgebra over a monoidal category* is an *algebra over the opposite monoidal category*. (This idea is explored more fully at the end of Section 2.) By this device one obtains "co-notions" and "co-examples" to those listed above, i.e., one encounters such diverse examples of coalgebras as ungraded coalgebras, graded coalgebras, differential graded coalgebras, complete coalgebras, profinite (or linearly compact) coalgebras, C^*-coalgebras, von Neumann coalgebras, cotriples, etc.

We close this section with an example that illustrates, once again, the ubiquity of coalgebras and the surprising (perhaps, mysterious) way in which they arise where one would have been hard-pressed to expect them.

In a conversation [Sw-3] that I had with Moss Sweedler, Sweedler mentioned to me that at one time he turned his attention to the *Jacobian conjecture* (a result in the area of commutative algebra dealing with the invertibility of a certain map, cf. [Ba-Co-Wr]). He came to realize that the *surjectivity* of the map in question was equivalent to a certain result about topological coalgebras. There is a topological coalgebra with a continuous coalgebra isomorphism to itself, where the Jacobian conjecture comes down to the inverse coalgebra isomorphism being continuous. See Appendix 5.2 for details.

2. Definitions and preliminary results

Henceforth, K shall denote a fixed but arbitrary ground field; all vector spaces and tensor products shall be over K; $char(K)$ shall denote the characteristic of K; and **Vect** shall denote the *category of vector spaces over* K.

As indicated in the previous section, algebras and coalgebras are defined "dually" to one another. To obtain the definition of a coalgebra – by duality – from that of an algebra, it will be convenient to write down the conditions that define an algebra via diagrams, more precisely, *commutative diagrams*.

REMARK 2.1. We recall (from Arbib and Manes [Arb-Man, p. 2]) that "... *commutare* is the Latin for *exchange*, and we say that a diagram *commutes* if we can exchange paths between two given points with impunity."

We now turn to the definition of an associative unitary algebra in diagram form – a form that is at once simple, suggestive, and in essence fundamental to mastering the notion of a coalgebra, Hopf algebra, or quantum group.

DEFINITION 2.2. An *associative unitary algebra over* K, or a K*-algebra* (or, more simply, an *algebra*), is a triple (A, φ_A, η_A) where A is a vector space and where

$$\varphi = \varphi_A : A \otimes A \to A \quad \text{and} \quad \eta = \eta_A : K \to A$$

are vector space maps for which the diagrams

$$\begin{tikzcd} A\otimes A\otimes A \arrow[r, "1\otimes\varphi"] \arrow[d, "\varphi\otimes 1"'] & A\otimes A \arrow[d, "\varphi"] \\ A\otimes A \arrow[r, "\varphi"] & A \end{tikzcd} \quad \text{and} \quad \begin{tikzcd} K\otimes A \arrow[d, "\eta\otimes 1"'] & A \arrow[l, "\cong", "\lambda^{-1}"'] \arrow[r, "\cong", "\rho^{-1}"'] \arrow[d, equal] & A\otimes K \arrow[d, "1\otimes\eta"] \\ A\otimes A \arrow[r, "\varphi"] & A & A\otimes A \arrow[l, "\varphi"'] \end{tikzcd}$$

(Figure 2.2(a)) (Figure 2.2(b))

commute. Here, the maps

$$\lambda^{-1} = \lambda_A^{-1} : A \xrightarrow{\cong} K \otimes A \quad \text{and} \quad \rho^{-1} = \rho_A^{-1} : A \xrightarrow{\cong} A \otimes K$$

are the natural K-linear isomorphisms given, respectively, for $a \in A$, by

$$a \mapsto 1_K \otimes a \quad \text{and} \quad a \mapsto a \otimes 1_K$$

where 1_K denotes the *unit* (or *identity*) *element* of the ground field K. Sometimes we shall abbreviate (A, φ_A, η_A) by A. The map φ_A is called the *multiplication* of A while η_A is called the *unit* of A. The ground field K will be considered trivially as the K-algebra (K, φ_K, η_K) where $\varphi_K : K \otimes K \to K$ is the *multiplication* of K and where $\eta_K : K \to K$ is the *identity map* from K to itself.

NOTATION AND TERMINOLOGY 2.3. If (A, φ_A, η_A) is an algebra and if $x, y \in A$, then we set

$$xy := x \cdot y := x \cdot_A y := \varphi_A(x \otimes y)$$

and call $x \cdot_A y$ the *product* of x and y in A; further we set

$$1_A := \eta_A(1_K)$$

and call 1_A the *unit* (or *identity*) *element* of A. The Greek letter *eta*, η, is to remind us of the German word *Einheit* meaning *unity*.

NOTE. By 2.50(f), a unit of an algebra is unique, hence is called *the unit* of the algebra. In the setting of a ring having a 2-sided (multiplicative) identity element, elements of the ring having 2-sided multiplicative inverses are called *units* (cf. [J-3, pp. 28–29], [Mac-Bir, p. 13] or [Art, p. 347]). As M. Artin notes (loc. cit.): “The identity element 1 of a ring is always a unit, and any reference to ‘the’ unit element in R refers to the identity. This is ambiguous terminology, but it is too late to change it.”

REMARKS 2.4 (a). If (A, φ, η) is an algebra, then, for all $x, y, z \in A$, we have that

$$\begin{tikzcd} x \otimes y \otimes z \arrow[r, mapsto, "1\otimes\varphi"] \arrow[d, mapsto, "\varphi\otimes 1"'] & x \otimes (y \cdot z) \arrow[d, mapsto, "\varphi"] \\ (x \cdot y) \otimes z \arrow[r, mapsto, "\varphi"] & (x \cdot y) \cdot z = x \cdot (y \cdot z). \end{tikzcd}$$

Thus, the commutativity of Figure 2.2(a) is equivalent to the usual way of writing down the *associativity axiom* for φ.

(b) If (A, φ, η) is an algebra, then, for any $a \in A$,

$$\begin{tikzcd} 1_K \otimes a \arrow[d, "\eta\otimes 1"'] & a \arrow[l, mapsto, "\cong", "\lambda^{-1}"'] \arrow[d, equal] \\ 1_A \otimes a \arrow[r, mapsto, "\varphi"] & 1_A \cdot a \end{tikzcd} \quad \text{and} \quad \begin{tikzcd} a \arrow[r, mapsto, "\cong", "\rho^{-1}"'] \arrow[d, equal] & a \otimes 1_K \arrow[d, "1\otimes\eta"] \\ a \cdot 1_A & a \otimes 1_A. \arrow[l, mapsto, "\varphi"'] \end{tikzcd}$$

Thus, the commutativity of Figure 2.2(b) is equivalent to the usual way of requiring that 1_A be a *left* as well as a *right unit* (or *identity*) *element* for the multiplication on A, and so expresses the *axiom* that η_A is a *two-sided unit* for φ_A.

REMARK 2.5. Defining an algebra via maps and commutative diagrams has the virtue that we can define a new object called a "coalgebra" simply by formally inverting all the arrows in the definition of an algebra and adding a "co" as a prefix to the description of each arrow thus formally inverted: In this way corresponding to the *multiplication* map $\varphi : A \otimes A \to A$ we have a *comultiplication* map (usually denoted Δ) in the reverse direction, and corresponding to the *unit* map $\eta : K \to A$ we require a *counit* map (denoted ε) in the reverse direction.

We are thus led to the following:

DEFINITION 2.6. An *associative unitary coalgebra over K*, also called a *coassociative counitary coalgebra over K* or a *K-coalgebra* (or, more simply, a *coalgebra*), is a triple $(C, \Delta_C, \varepsilon_C)$ consisting of a vector space C and vector space maps

$$\Delta = \Delta_C : C \to C \otimes C \quad \text{and} \quad \varepsilon = \varepsilon_C : C \to K$$

for which the diagrams

$$\begin{tikzcd} C \arrow[r, "\Delta"] \arrow[d, "\Delta"'] & C \otimes C \arrow[d, "1\otimes\Delta"] \\ C \otimes C \arrow[r, "\Delta\otimes 1"] & C \otimes C \otimes C \end{tikzcd} \quad \text{and} \quad \begin{tikzcd} C \otimes C \arrow[d, "\varepsilon\otimes 1"'] & C \arrow[l, "\Delta"'] \arrow[r, "\Delta"] \arrow[d, equal] & C \otimes C \arrow[d, "1\otimes\varepsilon"] \\ K \otimes C \arrow[r, "\cong", "\lambda"'] & C & C \otimes K \arrow[l, "\cong"', "\rho"] \end{tikzcd}$$

(Figure 2.6(a)) (Figure 2.6(b))

commute. Here the maps

$$\lambda = \lambda_C : K \otimes C \xrightarrow{\cong} C \quad \text{and} \quad \rho = \rho_C : C \otimes K \xrightarrow{\cong} C$$

are the natural K-linear isomorphisms given, respectively, for $k \in K$ and $c \in C$, by

$$k \otimes c \mapsto k \cdot c \quad \text{and} \quad c \otimes k \mapsto k \cdot c.$$

Sometimes we shall abbreviate $(C, \Delta_C, \varepsilon_C)$ by C. The map Δ_C is variously called the *comultiplication*, *coproduct*, or *diagonal* of C while the map ε_C is called the *counit* of C. The ground field K will be considered trivially as the K-coalgebra $(K, \Delta_K, \varepsilon_K)$ where $\Delta_K : K \to K \otimes K$ is the K-linear map defined by $\Delta_K(1) = 1 \otimes 1$, where 1 is the unit (or identity) element of K, and where $\varepsilon_K : K \to K$ is the *identity map* from K to itself.

TERMINOLOGY, NOTATION AND CONSEQUENCES 2.7. (a) The Greek letter *delta*, Δ, which begins with the English letter d, stands for the word *diagonal* which is to remind us of the origin of this map in algebraic topology (cf. Section 1). The Greek letter *epsilon*, ε, which begins with the English letter e is to remind us of the German word *Einheit* (meaning *unity*) which also begins with the English letter e. Milnor and Moore [M-M-2, pp. 217–218, Definitions 2.1] call Δ the *comultiplication* of the coalgebra C and ε the *unit* of C. This usage seems a bit inconsistent but the choice of *unit* as opposed to *counit* when referring to ε makes sense if one thinks (cf. 2.96, below) of a coalgebra over a monoidal category as an algebra over the opposite monoidal category. Such a categorical perspective also explains why we speak of an *associative unitary coalgebra* rather than a *coassociative counitary coalgebra* as is quite common.

(b) If $c \in C$, then

$$\Delta(c) = \sum_{i=1}^{n} c_{1i} \otimes c_{2i}$$

for some $\{c_{1i}, c_{2i}\}_{i=1}^{n} \subseteq C$. For simplicity, we suppress the index "i" and instead write

$$\Delta(c) = \sum_{c} c_1 \otimes c_2.$$

Here the "c" under the summation sign serves to remind us that the terms $c_1 \otimes c_2$ following $\sum_c$ come from diagonalizing (i.e., applying Δ to) c. By this convention

$$(\Delta \otimes 1) \circ \Delta(c) = \sum_{c} \sum_{c_1} c_{11} \otimes c_{12} \otimes c_2$$

while

$$(1 \otimes \Delta) \circ \Delta(c) = \sum_{c} \sum_{c_2} c_1 \otimes c_{21} \otimes c_{22}.$$

Of course, by the *associativity* of $\varDelta$,

$$\sum_c \sum_{c_1} c_{11} \otimes c_{12} \otimes c_2 = \sum_c \sum_{c_2} c_1 \otimes c_{21} \otimes c_{22}$$

for each $c \in C$. In like manner, the *unitary* property of $\varepsilon : C \to K$ guarantees that

$$c = \sum_c \varepsilon(c_1) \cdot c_2 = \sum_c \varepsilon(c_2) \cdot c_1$$

for each $c \in C$. These last two equations are sometimes called the *associativity* and *identity equations* (cf. Winter [Win, p. 173]). Occasionally, it simplifies things further to suppress the summation symbol(s). Then the associativity of $\varDelta : C \to C \otimes C$ and the unitary property of $\varepsilon : C \to K$ read, respectively, as follows: For each $c \in C$,

$$c_{11} \otimes c_{12} \otimes c_2 = c_1 \otimes c_{21} \otimes c_{22} \quad \text{while } c = \varepsilon(c_1) \cdot c_2 = \varepsilon(c_2) \cdot c_1.$$

For additional information about the sigma notation see [Sw-1, pp. 10–12] or [Dăs-Năs-Rai, pp. 4–8].

ALTERNATE CONVENTIONS 2.8. (a) Other commonly used conventions are

$$\varDelta(c) = \sum_{(c)} c_{(1)} \otimes c_{(2)} \tag{1}$$

due to Heyneman and Sweedler [Hey-Sw-1, p. 197], and

$$\varDelta(x) = \sum_i {}_i x \otimes x_i \tag{2}$$

due to Winter [Win, p. 173]. According to Winter: "...(the i range over some index set and ${}_i x$ or x_i are 0 for all but finitely many i). The ${}_i x$, x_i are certainly not unique. However, at times we shall arbitrarily choose the x_i, say, to be the basis for C, so that the ${}_i x$ for which $\varDelta(x) = \sum_i {}_i x \otimes x_i$ are then uniquely determined by x. The ${}_i x$ and x_i are referred to informally as the *left* and *right cofactors* of x."

(b) Joseph, in [Jos, p. 12], states with reference to a coalgebra A that: "There comes a moment when the geometric elegance of diagrams must be forsaken for the banal realities of algebraic computation. Here coalgebras cause a particular difficulty. Thus, for example $\varDelta(a)$ is actually a sum and moreover the terms are only determined up to linear transformations. To Sweedler's well-known solution to this problem we have added the logical simplification of omitting both Σ and the parenthesis, writing this sum simply as $\varDelta(a) = a_1 \otimes a_2$."

(c) Lambe and Radford, in [L-R, p. 267], write: "The Heyneman–Sweedler notation for $\varDelta(c)$ is $\varDelta(c) = \Sigma c_{(1)} \otimes c_{(2)}$. We will follow the physicists' lead and drop the summation symbol and write $\varDelta(c) = c_{(1)} \otimes c_{(2)}$."

REMARKS 2.9. (a) In the case of an associative algebra A, one can define the three-fold product of elements a, b and c of A, denoted $a \cdot b \cdot c$, to be the common value of the (iterated) two-fold products $(a \cdot b) \cdot c$ and $a \cdot (b \cdot c)$; and one may then inductively define the $(n+1)$-fold product of elements of A via

$$a_1 \cdots a_n \cdot a_{n+1} := (a_1 \cdots a_n) \cdot a_{n+1}.$$

By analogy, in the case of an associative coalgebra C, Heyneman and Sweedler adopt the notation

$$\sum_{(c)} c_{(1)} \otimes c_{(2)} \otimes c_{(3)}$$

for the common value of

$$(\Delta \otimes 1) \circ \Delta(c) = (1 \otimes \Delta) \circ \Delta(c).$$

Then, having defined

$$\Delta := \Delta_1 : C \to C \otimes C$$

via

$$\Delta(c) := \Delta_1(c) = \sum_{(c)} c_{(1)} \otimes c_{(2)}.$$

Heyneman and Sweedler [Hey-Sw-1, p. 197] define

$$\Delta_n : C \to \bigotimes^{n+1} C$$

inductively, for any integer $n \geqslant 2$, via

$$\Delta_n := (\Delta \otimes 1 \otimes \cdots \otimes 1) \circ \Delta_{n-1} := \left(\Delta \otimes 1^{n-1}\right) \circ \Delta_{n-1}$$

and write

$$\Delta_{n-1}(c) := \sum_{(c)} c_{(1)} \otimes c_{(2)} \otimes \cdots \otimes c_{(n)}.$$

We advocate streamlining their notation by dropping the parentheses as indicated above.

(b) Using Joseph's notation (of 2.8(b)), the counterpart of setting

$$a \cdot b \cdot c := (a \cdot b) \cdot c = a \cdot (b \cdot c)$$

to denote the common value of

$$\varphi \circ (\varphi \otimes 1)(a \otimes b \otimes c) = \varphi \circ (1 \otimes \varphi)(a \otimes b \otimes c),$$

for $a, b, c \in A$, is to set

$$c_1 \otimes c_2 \otimes c_3 := c_{11} \otimes c_{12} \otimes c_2 = c_1 \otimes c_{21} \otimes c_{22}$$

to denote the common value of

$$(\Delta \otimes 1) \circ \Delta(c) = (1 \otimes \Delta) \circ \Delta(c)$$

for $c \in C$.

(c) The obvious analog of "generalized associativity" for the multiplication of an associative algebra holds for the comultiplication of an associative coalgebra (cf. [Dăs-Năs-Rai, p. 5, Proposition 1.1.7]).

(d) Now that we have the definition of a coalgebra, it is time to define a morphism between two such. To motivate that definition, we first recall the dual situation.

DEFINITION 2.10. If (A, φ_A, η_A) and (B, φ_B, η_B) are algebras, then an *algebra map* f from (A, φ_A, η_A) to (B, φ_B, η_B) is a K-vector space map $f : A \to B$ for which the diagrams

$$\begin{array}{ccc} A \otimes A & \xrightarrow{\varphi_A} & A \\ {\scriptstyle f\otimes f}\downarrow & & \downarrow{\scriptstyle f} \\ B \otimes B & \xrightarrow{\varphi_B} & B \end{array} \qquad \text{and} \qquad \begin{array}{ccc} K & \xrightarrow{\eta_A} & A \\ \| & & \downarrow{\scriptstyle f} \\ K & \xrightarrow{\eta_B} & B \end{array}$$

(Figure 2.10(a)) (Figure 2.10(b))

commute.

REMARKS 2.11 (a). The commutativity of Figure 2.10(a) guarantees that $f(x \cdot_A y) = f(x) \cdot_B f(y)$ for all $x, y \in A$ and therefore is equivalent to the usual way of writing down the *definition* that $f : A \to B$ is *compatible with the multiplications* φ_A and φ_B on A and on B, respectively.

(b) Similarly the commutativity of Figure 2.10(b) is equivalent to the usual way of specifying that f send the unit (or identity) element of A to the unit (or identity) element of B, and so expresses the *condition* that $f : A \to B$ *preserve the units of* A *and of* B.

(c) It follows from (a) and (b), above, that the commutativity of Figures 2.10(a) and 2.10(b) is precisely equivalent to the usual way of writing down the defining properties of a *morphism of algebras*. We therefore define a *morphism of coalgebras* as follows:

DEFINITION 2.12. If $(C, \Delta_C, \varepsilon_C)$ and $(D, \Delta_D, \varepsilon_D)$ are coalgebras, then a *coalgebra map* f from $(C, \Delta_C, \varepsilon_C)$ to $(D, \Delta_D, \varepsilon_D)$ is a K-vector space map $f : C \to D$ for which the diagrams

$$\begin{tikzcd} C \arrow[r, "\Delta_C"] \arrow[d, "f"'] & C \otimes C \arrow[d, "f \otimes f"] \\ D \arrow[r, "\Delta_D"] & D \otimes D \end{tikzcd} \qquad \text{and} \qquad \begin{tikzcd} C \arrow[r, "\varepsilon_C"] \arrow[d, "f"'] & K \arrow[d, equal] \\ D \arrow[r, "\varepsilon_D"] & K \end{tikzcd}$$

(Figure 2.12(a)) (Figure 2.12(b))

commute.

REMARKS 2.13. (a) The commutativity of Figure 2.12(a) defines what it means to say that f *is compatible with the comultiplications* Δ_C and Δ_D on C and on D, respectively. In terms of the notation introduced in 2.7 above, this condition takes the form

$$\Delta\big[f(c)\big] = \sum_c f(c_1) \otimes f(c_2) \quad \text{for all } c \in C,$$

i.e.,

$$\sum_{f(c)} \big(f(c)\big)_1 \otimes \big(f(c)\big)_2 = \sum_c f(c_1) \otimes f(c_2) \quad \text{for all } c \in C.$$

(b) The commutativity of Figure 2.12(b) defines what it means to say that f *respects the counits* of C and of D. In terms of elements, this condition takes the form

$$\varepsilon_D\big[f(c)\big] = \varepsilon_C(c), \quad \text{for all } c \in C.$$

NOTATION 2.14. With the above definitions, we obtain the *category* **Alg** *of algebras and the category* **Coalg** *of coalgebras*.

REMARKS 2.15. (a) In one of the (several extant) preprint versions [M-M-1, p. 4, Footnote 3] of their seminal paper “On the structure of Hopf algebras,” Milnor and Moore state: “The term coalgebra was introduced by P. Cartier.” The relevant reference is Pierre Cartier’s famous “Hyperalgèbres et groupes de Lie formels” [Ca-1, especially, Exposé n°4, p. 4-01, §2, and Exposé n°3, p. 3–12, Lemma 2].

(b) *The Fundamental Theorem of Coalgebras* (4.12) is also due to Cartier as noted in (4.11) and (4.13).

(c) Since a subalgebra of an algebra is a subspace on which there is an algebra structure for which the inclusion is a morphism of algebras, we define a subcoalgebra of a coalgebra to be a subspace on which there is a coalgebra structure for which the inclusion is a morphism of coalgebras. Accordingly, we have the following:

DEFINITION 2.16. If $(C, \Delta_C, \varepsilon_C)$ and $(D, \Delta_D, \varepsilon_D)$ are coalgebras and if D is a subspace of C, then $(D, \Delta_D, \varepsilon_D)$ is said to be a *subcoalgebra* of $(C, \Delta_C, \varepsilon_C)$ in case the natural inclusion $i : D \hookrightarrow C$ is a morphism of coalgebras. In this case, we say, more briefly, that D is a *subcoalgebra* of C. A *subcoalgebra* D of C is said to be *proper* or *non-trivial* in case $D \neq 0$ and $D \neq C$.

REMARK. Any coalgebra (C, Δ, ε) always has two *trivial* subcoalgebras, namely, 0 and C itself. This is in contrast to the case of unitary algebras where 0 is not a subalgebra of a non-zero algebra since a non-zero algebra always contains an identity element distinct from 0. (Of course an algebra (A, φ, η) always has two *trivial* subalgebras, namely, $K \cdot 1_A$ and A itself.)

NOTE. In the above, 0 denotes the one-element coalgebra whose underlying vector space, $\{0\}$, considered as a set, has just one element, its origin, also denoted 0.

As an easy consequence, we have the following:

PROPOSITION 2.17. (a) *If $(D, \Delta_D, \varepsilon_D)$ is a subcoalgebra of $(C, \Delta_C, \varepsilon_C)$, then $\Delta_C(D) \subseteq D \otimes D$.*

(b) *Conversely, if $(C, \Delta_C, \varepsilon_C)$ is a coalgebra and if D is a subspace of C for which $\Delta_C(D) \subseteq D \otimes D$, then with $\varepsilon_D : D \to K$ defined via $\varepsilon_D := \varepsilon_C \circ i$, where $i : D \hookrightarrow C$ is the inclusion, and with $\Delta_D : D \to D \otimes D$ defined to be the unique K-linear map induced by the restriction of Δ_C to D, i.e., the unique K-linear map making the diagram*

$$\begin{tikzcd} D \arrow[r, dashed, "\Delta_D"] \arrow[d, "i"'] & D \otimes D \arrow[d, "i \otimes i"] \\ C \arrow[r, "\Delta_C"] & C \otimes C \end{tikzcd}$$

commute, $(D, \Delta_D, \varepsilon_D)$ is a subcoalgebra of $(C, \Delta_C, \varepsilon_C)$.

DEFINITION 2.18. A coalgebra is *locally finite* in case each of its elements lies in some finite-dimensional subcoalgebra.

REMARK 2.19. It is a basic result from the theory of associative unitary coalgebras (cf. [Sw-1, p. 46, Theorem 2.2.1], [Gr-1, p. 65, Lemma III.1.8], or [Lar-1, p. 351, Proposition 2.5]) that over a field any such coalgebra is *locally finite*. We shall prove this in 4.12 and again in 5.3, below. The *associativity* of Δ is the crucial part of the hypothesis since (cf. 4.14, below) over a field (1) *associative coalgebras* (without counit) are likewise *locally finite*, and (cf. [Mi-1, p. 4, Example (I.1.3.d)] or [Mi-2, pp. 9–10]) (2) there exist *Lie coalgebras* (special types of non-associative coalgebras) which are *not* locally finite. (In a more general approach to writing about coalgebras than the one adopted here, one would first define a *coalgebra* to be an ordered pair (C, Δ) where C is a K-vector space and $\Delta : C \to C \otimes C$ is a K-linear map. An *associative coalgebra* would then be a coalgebra (C, Δ) whose diagonal $\Delta : C \to C \otimes C$ satisfies the *associativity axiom* given by

Figure 2.6(a), above. A *unitary coalgebra* would then be an ordered triple (C, Δ, ε) consisting of a coalgebra (C, Δ) together with a K-linear map $\varepsilon : C \to K$ satisfying the *unitary property* specified by Figure 2.6(b), above. Finally, an *associative unitary coalgebra* would be a unitary coalgebra (C, Δ, ε) for which (C, Δ) is an associative coalgebra. A *Lie coalgebra* would then be a coalgebra (M, Δ) in which Δ satisfies certain axioms (cf. [Mi-1] or [Mi-2]) that are the counterpart (dual) of those satisfied by the *bracket* of a *Lie algebra*.) Local finiteness is an important property for coalgebras because it enables one to obtain results about locally finite coalgebras by dualizing results about finite-dimensional algebras.

REMARKS 2.20. (a) As Sweedler notes [Sw-1, p. 16], and we quote almost verbatim, "... when one defines a subalgebra of an algebra, a condition on the unit (= identity element) must be added (since) a subspace closed under multiplication may have a different unit or no unit at all... ." Indeed, (cf. [J-1, pp. 84–86] or [B-J-N, pp. 193–194, Example 2.7(c)]) any ring R is a $\mathbb{Z}$-algebra. Given a ring R, embed it in a ring with unit, A, via setting

$$\begin{aligned}
&A := \{(r, n) \mid r \in R,\ n \in \mathbb{Z}\},\\
&(r, n) + (s, m) := (r + s, n + m), \quad \text{and}\\
&(r, n) + (s, m) := (rs + mr + ns, nm).
\end{aligned}$$

Then A is a ring with unit (0, 1). Let B be the subring defined via

$$B := \{(r, 0) \mid r \in R\}.$$

Then A and B are $\mathbb{Z}$-algebras. If R does not have a unit neither will B have a unit. But even if R does have a unit, the units of A and B will differ. On the other hand, Sweedler (loc. cit) continues: "In the coalgebra case, the counit takes care of itself." Analogously, the unit of a quotient algebra takes care of itself.

(b) Quotient algebras are important and so too are quotient coalgebras. Now, (two-sided) *ideals* are what one factors algebras by to get *quotient algebras*. Correspondingly, (two-sided) *coideals* are what one factors coalgebras by to get *quotient coalgebras*. By this we mean that *ideals* should be *kernels of algebra morphisms* while *coideals* should be *kernels of coalgebra morphisms*. This means, in the algebra context, that ideals should be defined in such a way that the following theorem holds: If I is a subspace of an algebra A, then I is a (two-sided) *ideal* of A if and only if there exists a necessarily unique algebra structure on the quotient vector space A/I for which the natural vector space projection $\pi : A \to A/I$ is a morphism of algebras. The guiding principle here is Edwin Hewitt's adage: Old theorems never die; they just become definitions [Herr-Stre, p. 100]. Now our algebras have (two-sided) units which algebra morphisms are required to preserve while coalgebras have (two-sided) counits which coalgebra morphisms are correspondingly required to respect.

Because coalgebras are obtained from algebras by "turning the arrows around" and because a (two-sided) ideal of an algebra A can be viewed as a subspace I of A for which

$$\varphi_A(I \otimes A + A \otimes I) \subseteq I$$

it is natural to propose that a (two-sided) coideal of a coalgebra C should be a subspace I of C for which

$$\Delta_C(I) \subseteq I \otimes C + C \otimes I.$$

But, should we take this to be the definition, we would miss a significant extra ingredient (namely, compatibility between π and ε_C and $\varepsilon_{C/I}$), and therefore not obtain the correct definition of a coideal. In the case of algebras, the corresponding extra condition (namely, compatibility between π and η_A and $\eta_{A/I}$) takes care of itself – as explained below.

To obtain the correct definition, we make use of two facts. In what follows, as usual, 0 may on occasion denote the one-element vector space whose only element is its origin, also denoted 0.

LEMMA 2.21. *Let A be a vector space, let I be a subspace of A, and let $\pi : A \to A/I$ denote the natural vector space projection. Then*

$$\operatorname{Ker}(\pi \otimes \pi) = \operatorname{Ker}(\pi) \otimes A + A \otimes \operatorname{Ker}(\pi) = I \otimes A + A \otimes I.$$

LEMMA 2.22. *Let $\pi_1 : A_1 \to B_1$ and $\pi_2 : A_2 \to B_2$ be vector space surjections, and let $f : A_1 \to A_2$ be a vector space map. Then $f(\operatorname{Ker}\pi_1) \subseteq \operatorname{Ker}\pi_2$ if and only if there is a (necessarily unique) vector space map $\tilde{f} : B_1 \to B_2$ making the following diagram of short exact sequences commute*:

$$\begin{tikzcd}
0 \arrow[r] & \operatorname{Ker}\pi_1 \arrow[r, hook] \arrow[d] & A_1 \arrow[r, two heads, "\pi_1"] \arrow[d, "f"] & B_1 \arrow[r] \arrow[d, "\tilde{f}"] & 0 \\
0 \arrow[r] & \operatorname{Ker}\pi_2 \arrow[r, hook] & A_2 \arrow[r, two heads, "\pi_2"] & B_2 \arrow[r] & 0
\end{tikzcd}$$

For a proof of Lemma 2.21, the reader is referred to [Bly, p. 204, Theorem 5.13] or to [Fog, p. 26, Lemma 1.27]. Lemma 2.22 is an immediate consequence of the following more general result:

PROPOSITION 2.23. *Consider the diagram*

$$\begin{tikzcd}
0 \arrow[r] & A \arrow[r, "\alpha"] \arrow[d, "g"] & B \arrow[r, "\beta"] \arrow[d, "f"] & C \arrow[r] \arrow[d, "h"] & 0 \\
0 \arrow[r] & A' \arrow[r, "\alpha'"] & B' \arrow[r, "\beta'"] & C' \arrow[r] & 0
\end{tikzcd}$$

of vector spaces and linear transformations in which the rows are exact sequences but g and h are not yet specified. Then a (necessarily unique) linear transformation $g : A \to A'$ exists making the left square commute if and only if a (necessarily unique) linear transformation $h : C \to C'$ exists making the right square commute.

REMARK 2.24. The above result says the following (cf. [Rotm, p. 27, Exercise 2.7]): Given the commutative square

$$\begin{tikzcd} V \arrow[r, "\gamma"] \arrow[d, "S"'] & W \arrow[d, "T"] \\ V' \arrow[r, "\sigma"] & W' \end{tikzcd}$$

in **Vect**, the linear transformation S *restricts* to a linear transformation $\widetilde{S}\colon \operatorname{Ker}\gamma \to \operatorname{Ker}\sigma$ while the linear transformation T *induces* a linear transformation $\widetilde{T}\colon \operatorname{Coker}\gamma \to \operatorname{Coker}\sigma$.

REMARK 2.25. Shortly, we shall apply Lemmas 2.21 and 2.22 to obtain the definition of a (two-sided) *coideal* of a coalgebra. But first, we analyze the method by which we obtain the definition of a (two-sided) *ideal* of an algebra.

As noted earlier, we want a subspace I of an algebra A to be a (two-sided) ideal of A precisely when the quotient vector space A/I has an algebra structure for which the natural projection $\pi : A \to A/I$ is a morphism of algebras. But for π to be a morphism the following two diagrams must commute:

$$\begin{tikzcd} A \otimes A \arrow[r, "\varphi_A"] \arrow[d, "\pi\otimes\pi"'] & A \arrow[d, "\pi"] \\ A/I \otimes A/I \arrow[r, "\varphi_{A/I}"'] & A/I \end{tikzcd} \quad \text{and} \quad \begin{tikzcd} K \arrow[r, "\eta_A"] \arrow[d, equal] & A \arrow[d, "\pi"] \\ K \arrow[r, "\eta_{A/I}"'] & A/I. \end{tikzcd}$$

Now, the commutativity of the diagram on the right forces $\eta_{A/I} := \pi \circ \eta_A$ but does not otherwise make any demands on I. By contrast, the commutativity of the diagram on the left forces $\varphi_A(I \otimes A + A \otimes I) \subseteq I$ because that commutative diagram embeds in the following diagram of short exact sequences

$$\begin{tikzcd} 0 \arrow[r] & I \otimes A + A \otimes I \arrow[r, hook] \arrow[d] & A \otimes A \arrow[r, "\pi\otimes\pi"] \arrow[d, "\varphi_A"] & A/I \otimes A/I \arrow[r] \arrow[d, "\varphi_{A/I}"] & 0 \\ 0 \arrow[r] & I \arrow[r, hook] & A \arrow[r, "\pi"] & A/I \arrow[r] & 0 \end{tikzcd}$$

and by Lemmas 2.21 and 2.22, a (necessarily) unique linear map

$$\varphi_{A/I}\colon A/I \otimes A/I \to A/I$$

exists for which the rectangle on the right commutes if and only if $\varphi_A(I \otimes A + A \otimes A) \subseteq I$.

Accordingly, we are led to the following:

DEFINITION 2.26. A (two-sided) *ideal* of an algebra (A, φ_A, η_A) is a subspace I of A for which

$$\varphi_A(I \otimes A + A \otimes I) \subseteq I.$$

An ideal I of A is said to be a *non-trivial ideal* in case $I \neq 0$ and $I \neq A$.

REMARK 2.27. With ideals defined as above, it is straightforward to show that I is an ideal of A if and only if the quotient vector space A/I has a (necessarily) unique algebra structure for which the natural projection $\pi : A \to A/I$ is a morphism of algebras. (One need only check that $\varphi_{A/I}$ is associative and that $\eta_{A/I}$ is a 2-sided unit for $\varphi_{A/I}$.)

REMARK 2.28. "Dualizing" the above approach, we see that a (two-sided) *coideal* of a coalgebra C should be a subspace I of C for which the quotient vector space C/I has a necessarily unique coalgebra structure for which the natural projection $\pi : C \to C/I$ is a morphism of coalgebras. Now for π to be a morphism the following two diagrams must commute:

$$\begin{tikzcd} C \arrow[r, "\Delta_C"] \arrow[d, "\pi"'] & C \otimes C \arrow[d, "\pi\otimes\pi"] \\ C/I \arrow[r, "\Delta_{C/I}"'] & C/I \otimes C/I \end{tikzcd} \quad \text{and} \quad \begin{tikzcd} C \arrow[r, "\varepsilon_C"] \arrow[d, "\pi"'] & K \arrow[d, equal] \\ C/I \arrow[r, "\varepsilon_{C/I}"'] & K. \end{tikzcd}$$

The commutativity of the diagram at the right tells us that $I \subseteq \operatorname{Ker} \varepsilon_C$ since $I = \operatorname{Ker} \pi$. It also tells us that there can be at most one linear map $\varepsilon_{C/I} : C/I \to K$ such that $\varepsilon_{C/I} \circ \pi = \varepsilon_C$ since $\pi : C \to C/I$ is surjective.

NOTE. In the algebra case, the (corresponding) requirement that $\eta_{A/I} = \pi \circ \eta_A$ forced the definition of $\eta_{A/I}$ but otherwise made no demands on I.

The commutativity of the diagram at the left forces $\Delta_C(I) \subseteq I \otimes C + C \otimes I$ because that commutative diagram embeds in the following diagram of short exact sequences

$$\begin{tikzcd} 0 \arrow[r] & I \arrow[r, hook] \arrow[d] & C \arrow[r, "\pi"] \arrow[d, "\Delta_C"] & C/I \arrow[r] \arrow[d, "\Delta_{C/I}"] & 0 \\ 0 \arrow[r] & I \otimes C + C \otimes I \arrow[r, hook] & C \otimes C \arrow[r, "\pi\otimes\pi"'] & C/I \otimes C/I \arrow[r] & 0 \end{tikzcd}$$

and by Lemmas 2.21 and 2.22, a (necessarily) unique linear map

$$\Delta_{C/I} : C/I \to C/I \otimes C/I$$

exists for which the rectangle on the right commutes if and only if $\Delta_C(I) \subseteq I \otimes C + C \otimes I$. Accordingly, we are led to the following:

DEFINITION 2.29. A (two-sided) *coideal* of a coalgebra $(C, \Delta_C, \varepsilon_C)$ is a subspace I of C for which

$$\Delta_C(I) \subseteq I \otimes C + C \otimes I$$

and

$$I \subseteq \operatorname{Ker} \varepsilon_C.$$

REMARKS 2.30. (a) Notice the presence of a second condition (viz., $I \subseteq \operatorname{Ker} \varepsilon$) – one whose analogue in the case of algebras takes care of itself. This condition is necessary to insure that the natural projection $\pi : C \to C/I$, where I is a coideal of C, is a morphism of **Coalg**.

(b) Another way to see the necessity of the condition $I \subseteq \operatorname{Ker} \varepsilon$ required of a coideal I of a coalgebra C (in addition to the requirement that $\Delta(I) \subseteq I \otimes C + C \otimes I$) is to observe that, over a field, any subcoalgebra $(D, \Delta_D, \varepsilon_D)$ of a coalgebra (C, Δ, ε) satisfies the condition

$$\Delta(D) \subseteq D \otimes C + C \otimes D,$$

so that if we wish to *distinguish* between *subcoalgebras* and *coideals*, even as in the case of algebras we *distinguish* between *ideals* and *subalgebras*, then we shall need an extra condition for a coideal. Now, if $(D, \Delta_D, \varepsilon_D)$ is a *non-zero* subcoalgebra of $(C, \Delta, \varepsilon) := (C, \Delta_C, \varepsilon_C)$, then $D \nsubseteq \operatorname{Ker} \varepsilon$ because, for each $d \in D$,

$$d = \sum_d \varepsilon(d_1) \cdot d_2.$$

Thus, a *nonzero subcoalgebra* is *never* a *coideal*, even as a *subalgebra* is *never* a *proper ideal*. Indeed, but for the requirement that a subalgebra contain the unit (= identity) element of the algebra, an ideal would be a subalgebra. Since, however, a subalgebra is required to contain the unit (= identity) element of the algebra, it follows that a *proper* (i.e., *non-trivial*) *ideal* can *never* be a *subalgebra*, and that a *subalgebra* can *never* be a *proper ideal*. (Recall (cf. [Art, p. 357]) that an *ideal* of an algebra is *proper* (or *non-trivial*) if it is not the zero ideal (= the principal ideal (0) generated by 0) or the entire algebra (= the unit ideal = the principal ideal (1) generated by the unit (or identity element) 1 of the algebra). Accordingly, 0 and A are considered *trivial ideals* of the algebra (A, φ_A, η_A).)

REMARK 2.31. With coideals defined as above, we can now establish the following theorem (cf. [Sw-1, p. 22, Theorem 1.4.7]):

THEOREM 2.32 (The Fundamental Isomorphism Theorem for Coalgebras). *Let I be a coideal of a coalgebra C, and let $\pi : C \to C/I$ denote the natural vector space projection. Then*

(a) *C/I has a unique coalgebra structure for which $\pi : C \to C/I$ is a morphism of coalgebras.*
(b) *If D is a coalgebra and if $f : C \to D$ is any morphism of coalgebras, then* $\operatorname{Ker} f$ *is a coideal of C.*
(c) *If $f : C \to D$ is a morphism of coalgebras and if I is a coideal of C such that $I \subseteq \operatorname{Ker} f$, then there exists a unique morphism $\tilde{f} : C/I \to D$ of coalgebras such that $f = \tilde{f} \circ \pi$. In case $I = \operatorname{Ker} f$, $\tilde{f} : C/\operatorname{Ker} f \to \operatorname{Im} f$ is an isomorphism of coalgebras.*

PROOF. (a) By virtue of the foregoing observations, we need just show that $\Delta_{C/I}$ and $\varepsilon_{C/I}$ equip C/I with the structure of an *associative unitary* coalgebra (i.e., that $\Delta_{C/I}$ is associative and that $\varepsilon_{C/I}$ is a two-sided counit for $\Delta_{C/I}$), for the fact that π is a morphism of coalgebras is then immediate. For convenience, let $\overline{C} := C/I$, $\Delta := \Delta_C$, $\varepsilon := \varepsilon_C$, $\overline{\Delta} := \Delta_{C/I} := \Delta_{\overline{C}}$, and $\bar{\varepsilon} := \varepsilon_{C/I} := \varepsilon_{\overline{C}}$. We claim that because the natural projection $\pi : C \to \overline{C}$ is surjective, the *associativity* of $\overline{\Delta}$ follows from that of Δ by a consideration of the following diagram

$$\begin{tikzcd}
C \arrow[rr, "\Delta"] \arrow[dr, "\pi"] \arrow[dd, equal] & & C\otimes C \arrow[rr, "1\otimes\Delta"] \arrow[dr, "\pi\otimes\pi"] & & C\otimes(C\otimes C) \arrow[dr, "\pi\otimes(\pi\otimes\pi)"] \arrow[dd, dashed, "\cong"] & \\
& \overline{C} \arrow[rr, "\overline{\Delta}"] \arrow[dd, equal] & & \overline{C}\otimes\overline{C} \arrow[rr, "1\otimes\overline{\Delta}"] & & \overline{C}\otimes(\overline{C}\otimes\overline{C}) \arrow[dd, "\cong"] \\
C \arrow[rr, dashed, "\Delta"] \arrow[dr, "\pi"] & & C\otimes C \arrow[rr, dashed, "\Delta\otimes 1"] \arrow[dr, dashed, "\pi\otimes\pi"] & & (C\otimes C)\otimes C \arrow[dr, dashed, "(\pi\otimes\pi)\otimes\pi"] & \\
& \overline{C} \arrow[rr, "\overline{\Delta}"'] & & \overline{C}\otimes\overline{C} \arrow[rr, "\overline{\Delta}\otimes 1"'] & & (\overline{C}\otimes\overline{C})\otimes\overline{C}
\end{tikzcd}$$

in which all the faces commute with the possible exception of the front face. For the associativity of $\overline{\Delta}$ is expressed by the commutativity of the front face while the associativity of Δ is expressed by the commutativity of the back face. But, since $\pi : C \to \overline{C}$ is surjective, the commutativity of the front face follows at once from the commutativity of all the remaining faces. To establish the fact that $\bar{\varepsilon}$ is a two-sided counit one applies a similar argument. For instance, we claim that because π is surjective, the fact that $\bar{\varepsilon}$ is a right counit follows from the fact that ε is a right counit by a consideration of the following diagram:

$$\begin{tikzcd}
C \arrow[rr, "\Delta"] \arrow[dr, "\pi"] \arrow[dd, equal] & & C\otimes C \arrow[dr, "\pi\otimes\pi"] \arrow[dd, dashed, "1\otimes\varepsilon"] & \\
& \overline{C} \arrow[rr, "\overline{\Delta}"] \arrow[dd, equal] & & \overline{C}\otimes\overline{C} \arrow[dd, "1\otimes\bar{\varepsilon}"] \\
C \arrow[rr, dashed, "\cong"] \arrow[dr, "\pi"] & & C\otimes K \arrow[dr, dashed, "\pi\otimes 1"] & \\
& \overline{C} \arrow[rr, "\cong"] & & \overline{C}\otimes K
\end{tikzcd}$$

By a similar argument applied to the appropriate mirror image of the above diagram one can show that, since π is surjective, $\bar{\varepsilon}$ is a left counit because ε is a left counit. This concludes the proof of part (a). Proofs for the remaining parts are routine – cf. [Sw-1, pp. 22–26, Theorem 1.4.7]. □

DEFINITION/NOTATION 2.33. (a) For any vector space V, the natural *twisting map* is the linear *isomorphism*

$$\tau : V \otimes V \to V \otimes V$$

which sends $x \otimes y$ to $y \otimes x$ for all $x, y \in V$. Since τ *switches* the 1st and 2nd tensorands, some authors, e.g., [Man-1, p. 12], use the notation $S_{(12)}$ in place of τ. More generally, for any integer $n \geqslant 2$ and any element $\sigma \in S_n$ (:= the symmetric group on the n symbols $1, 2, \ldots, n$), $S_\sigma := S_\sigma^{(n)}$ will denote the obvious action of σ on the n-fold tensor product $V^{\otimes n}$. In this spirit, $S_{(23)} := S_{(23)}^{(4)} : V^{\otimes 4} \to V^{\otimes 4}$ would denote the linear map

$$1 \otimes \tau \otimes 1 : V \otimes V \otimes V \otimes V \to V \otimes V \otimes V \otimes V$$

which *switches* the 2nd and 3rd tensorands of the 4-fold tensor product while leaving the 1st and 4th tensorands intact, thus sending $a \otimes b \otimes c \otimes d$ to $a \otimes c \otimes b \otimes d$ whenever $a, b, c, d \in V$.

(b) If V and W are vector spaces, we let

$$\tau_{V,W} : V \otimes W \to W \otimes V$$

denote the isomorphism of vector spaces which sends $v \otimes w$ to $w \otimes v$ for all $(v, w) \in V \times W$. If $f : V \to V'$ and $g : W \to W'$ are vector space maps, then the following diagram commutes:

$$\begin{tikzcd}
V \otimes W \arrow[r, "\tau_{V,W}"'] \arrow[d, "f \otimes g"'] & W \otimes V \arrow[d, "g \otimes f"] \\
V' \otimes W' \arrow[r, "\tau_{V',W'}"'] & W' \otimes V'
\end{tikzcd}$$

Hence τ is what in [K-R-T, p. 18, item 3.1] is called a *commutativity constraint* in the monoidal category (**Vect**, $\otimes$, K).

REMARK 2.34. An algebra A is *commutative* in case $x \cdot y = y \cdot x$, for all $x, y \in A$. We may express this condition in an element-free way, via a commutative diagram, as follows:

DEFINITION 2.35. An *algebra* (A, φ_A, η_A) is *commutative* in case the diagram

$$\begin{tikzcd}
A \otimes A \arrow[dd, "\tau"'] \arrow[dr, "\varphi_A"] & \\
& A \\
A \otimes A \arrow[ur, "\varphi_A"'] &
\end{tikzcd}$$

commutes.

"Dualizing" the above, we obtain the following:

DEFINITION 2.36. A *coalgebra* $(C, \Delta_C, \varepsilon_C)$ is *commutative* in case the diagram

$$\begin{tikzcd}
& C \otimes C \arrow[dd, "\tau"] \\
C \arrow[ur, "\Delta_C"] \arrow[dr, "\Delta_C"'] & \\
& C \otimes C
\end{tikzcd}$$

commutes. In terms of elements, a coalgebra C is *commutative* if and only if $\Delta(c) = \sum_c c_1 \otimes c_2 = \sum_c c_2 \otimes c_1$, for all $c \in C$.

REMARKS 2.37. (a) Many authors use the adjective "cocommutative" rather than "commutative" when referring to a coalgebra. This usage does have the advantage that, for example, when dealing with a bialgebra or a Hopf algebra (to be defined below), one can speak of a commutative, cocommutative bialgebra or Hopf algebra. Nonetheless, from a categorical perspective, our terminology makes more sense. But it does force one to speak, for example, of a bialgebra or a Hopf algebra having a commutative multiplication and a commutative comultiplication.

(b) If (A, φ, η) is an algebra, set $\varphi^{op} := \varphi \circ \tau$. Then (A, φ^{op}, η) is also an algebra because the following diagrams commute:

$$\begin{tikzcd}
& & A \arrow[dll, "\lambda^{-1}"'] \arrow[drr, "\rho^{-1}"] \arrow[d, equal] & & \\
K \otimes A \arrow[d, "\eta \otimes 1"'] & A \otimes K \arrow[l, "\tau"'] \arrow[d, "1 \otimes \eta"] & A \arrow[l, "\rho^{-1}"'] \arrow[r, "\lambda^{-1}"] \arrow[d, equal] & K \otimes A \arrow[r, "\tau"] \arrow[d, "\eta \otimes 1"] & A \otimes K \arrow[d, "1 \otimes \eta"] \\
A \otimes A \arrow[r, "\tau"'] & A \otimes A \arrow[r, "\varphi"'] & A & A \otimes A \arrow[l, "\varphi"] & A \otimes A \arrow[l, "\tau"]
\end{tikzcd}$$

and

$$\begin{tikzcd}
A \otimes A \otimes A \arrow[r, "\tau\otimes 1 = S_{(12)}"] \arrow[d, "1\otimes\tau = S_{(23)}"] & A \otimes A \otimes A \arrow[r, "\varphi\otimes 1"] \arrow[d, "\tau_{A\otimes A, A} = S_{(123)}"] & A \otimes A \arrow[d, "\tau = \tau_{A,A}"] \\
A \otimes A \otimes A \arrow[r, "\tau_{A, A\otimes A} = S_{(132)}"'] \arrow[d, "1\otimes\varphi"'] & A \otimes A \otimes A \arrow[r, "1\otimes\varphi"'] \arrow[d, "\varphi\otimes 1"] & A \otimes A \arrow[d, "\varphi"] \\
A \otimes A \arrow[r, "\tau = \tau_{A,A}"'] & A \otimes A \arrow[r, "\varphi"'] & A.
\end{tikzcd}$$

NOTE. In S_3, the *symmetric group on* 3*-letters*, $(123)(12) = (13)$ and $(132)(23) = (13)$. *Terminology*: φ^{op} is called the *opposite multiplication* of φ; (A, φ^{op}, η) is called the *opposite algebra* of (A, φ, η); and if we abbreviate (A, φ, η) by A, then we abbreviate (A, φ^{op}, η) by A^{op}. Thus an algebra A is *commutative* if and only if $A = A^{op}$.

(c) If (C, Δ, ε) is a coalgebra, set $\Delta^{op} := \tau \circ \Delta$. Then $(C, \Delta^{op}, \varepsilon)$ is also a coalgebra because the following diagrams commute:

$$\begin{tikzcd}
C \otimes C \arrow[d, "\varepsilon\otimes 1"'] & C \otimes C \arrow[l, "\tau"'] \arrow[d, "1\otimes\varepsilon"] & C \arrow[l, "\Delta"'] \arrow[r, "\Delta"] \arrow[d, equal] & C \otimes C \arrow[r, "\tau"] \arrow[d, "\varepsilon\otimes 1"] & C \otimes C \arrow[d, "1\otimes\varepsilon"] \\
K \otimes C \arrow[r, "\tau"'] \arrow[drr, "\lambda"'] & C \otimes K \arrow[r, "\rho"'] & C \arrow[d, equal] & K \otimes C \arrow[l, "\lambda"] & C \otimes K \arrow[l, "\tau"] \arrow[dll, "\rho"] \\
 & & C & &
\end{tikzcd}$$

and

$$\begin{tikzcd}
C \arrow[r, "\Delta"] \arrow[d, "\Delta"'] & C \otimes C \arrow[r, "\tau = \tau_{C,C}"] \arrow[d, "1\otimes\Delta"] & C \otimes C \arrow[d, "\Delta\otimes 1"] \\
C \otimes C \arrow[r, "\Delta\otimes 1"] \arrow[d, "\tau = \tau_{C,C}"'] & C \otimes C \otimes C \arrow[r, "\tau_{C, C\otimes C} = S_{(132)}"'] \arrow[d, "\tau_{C\otimes C, C} = S_{(123)}"] & C \otimes C \otimes C \arrow[d, "\tau\otimes 1 = S_{(12)}"] \\
C \otimes C \arrow[r, "1\otimes\Delta"'] & C \otimes C \otimes C \arrow[r, "1\otimes\tau = S_{(23)}"'] & C \otimes C \otimes C.
\end{tikzcd}$$

NOTE. In S_3, $(23)(123) = (13)$ and $(12)(132) = (13)$. *Terminology*: Δ^{op} is called the *opposite comultiplication* of Δ; $(C, \Delta^{op}, \varepsilon)$ is called the *opposite coalgebra* of (C, Δ, ε); and if we abbreviate (C, Δ, ε) by C, then we abbreviate $(C, \Delta^{op}, \varepsilon)$ by C^{op}. Thus a coalgebra C is *commutative* if and only if $C = C^{op}$. (N.B.: Both the notation and the terminology vary with, for example, some authors writing Δ^{cop} for $\tau \circ \Delta$ and calling $(C, \Delta^{cop}, \varepsilon)$ the *coopposite* (or *co-opposite*) *coalgebra* of (C, Δ, ε).) Things get interesting in the case of a bialgebra or Hopf algebra. For a bialgebra $H := (H, \varphi, \eta, \Delta, \varepsilon)$, one may consider $H^{op} := (H, \varphi^{op}, \eta, \Delta, \varepsilon)$, $H^{cop} := (H, \varphi, \eta, \Delta^{op}, \varepsilon)$ and $H^{op,cop} := (H, \varphi^{op}, \eta, \Delta^{op}, \varepsilon)$. For further discussion, see, for example, [Mon-1, pp. 2, 6, 9], [Char-Pres, pp. 103–104], [Kass, pp. 41, 46, 52–55], [Maj-1, pp. 9–10], [Kli-Sch, pp. 9, 12–15], [L-R, pp. 3, 43–45], and [Dăs-Năs-Rai, pp. 32–33, 149, 156].

(d) In the graded context the definition of commutativity for algebras and coalgebras requires a tiny bit of care. For when dealing with graded vector spaces, in place of the *ungraded twisting map*

$$\tau : V \otimes V \to V \otimes V$$

(defined in 2.33 and used in Definitions 2.35 and 2.36), one uses the *graded twisting map*, also denoted

$$\tau : V \otimes V \to V \otimes V,$$

defined for all $n \geqslant 0$ (the context will make clear how τ is to be understood) by

$$\tau_n(x_p \otimes y_q) := (-1)^{p \cdot q} y_q \otimes x_p$$

where $p \geqslant 0$, $q \geqslant 0$, $p + q = n$, $x_p \in V_p$ and $y_q \in V_q$. Recall that – in the simplest case – a *graded K-vector space V* (or, more simply, a graded vector space) is a *family* $V := \{V_n\}_{n=0}^{\infty}$ of K-vector spaces V_n, one for each integer $n \geqslant 0$, while a *morphism $f : V \to W$ of graded K-vector spaces* from $V := \{V_n\}_{n=0}^{\infty}$ to $W := \{W_n\}_{n=0}^{\infty}$ (or, more simply, a *morphism of graded vector spaces*) is a *family* $f := \{f_n\}_{n=0}^{\infty}$ of *K-vector space morphisms* $f_n : V_n \to W_n$, one for each integer $n \geqslant 0$. If V and W are *graded vector spaces*, so is their *tensor product $V \otimes W$ defined* by

$$(V \otimes W)_n := \bigoplus_{\substack{p+q=n \\ p,q \geqslant 0}} V_p \otimes W_q.$$

A *graded K-vector space V* is *connected* in case $V_0 \cong K$ as K-vector spaces.

(e) The *ground field K* may (and, on occasion, will) be considered as *the graded K-vector space* $K := \{K_n\}_{n=0}^{\infty}$ concentrated in degree 0 defined by setting $K_0 := K$ and $K_n := 0$ for each integer $n \geqslant 1$. Then K is a (trivially) *graded algebra* as well as a (trivially) *graded coalgebra* according to the definitions given in (f) and (g), below.

(f) *A graded algebra over K* (or *a graded K-algebra*, or, more simply, *a graded algebra*) is a triple (A, φ, η) consisting of a graded K-vector space $A := \{A_n\}_{n=0}^{\infty}$ and morphisms $\varphi : A \otimes A \to A$ and $\eta : K \to A$ of graded K-vector spaces for which the diagrams 2.2(a) and 2.2(b), above, considered as diagrams in the category of graded K-vector spaces, commute.

NOTE. If $V := \{V_n\}_{n=0}^{\infty}$ is any graded K-vector space, then there exist natural isomorphisms

$$K \otimes V \cong V \quad \text{and} \quad V \otimes K \cong V$$

of graded K-vector spaces because, for every integer $n \geqslant 0$,

$$(K \otimes V)_n = K \otimes V_n \cong V_n \quad \text{and} \quad (V \otimes K)_n = V_n \otimes K \cong V_n.$$

A graded algebra (A, φ, η) *is commutative* (or, for emphasis, *graded commutative*) in case $\varphi = \varphi \circ \tau$ where $\tau : A \otimes A \to A \otimes A$ is the *graded twisting map* defined in (d), above. Classically, such an algebra was called *anti-commutative* or *skew-commutative*. *A graded K-algebra* (A, φ, η) *is connected* in case $\eta_0 : K \to A_0$ is an *isomorphism* of K-vector spaces.

(g) *A graded coalgebra over K* (or *a graded K-coalgebra*, or, more simply, *a graded coalgebra*) is a triple (C, Δ, ε) consisting of a graded K-vector space $C := \{C_n\}_{n=0}^{\infty}$ and morphisms $\Delta : C \to C \otimes C$ and $\varepsilon : C \to K$ of graded K-vector spaces for which the diagrams 2.6(a) and 2.6(b), above, considered as diagrams in the category of graded K-vector spaces, commute. *A graded coalgebra* (C, Δ, ε) *is commutative* in case $\Delta = \tau \circ \Delta$ where $\tau : C \otimes C \to C \otimes C$ is the *graded twisting map* defined in (d), above. *A graded K-coalgebra is connected* in case $\varepsilon_0 : C_0 \to K$ is an *isomorphism* of K-vector spaces.

DEFINITION 2.38. The *tensor product of* the *algebras* (A, φ_A, η_A) and (B, φ_B, η_B), in that order, is the *algebra*

$$(A \otimes B, \varphi_{A\otimes B}, \eta_{A\otimes B})$$

whose *multiplication* $\varphi_{A\otimes B}$ and *unit* $\eta_{A\otimes B}$ are the K-linear maps defined, respectively, by

$$\varphi_{A\otimes B} := (\varphi_A \otimes \varphi_B) \circ S_{(23)} : A \otimes B \otimes A \otimes B \xrightarrow{S_{(23)}} A \otimes A \otimes B \otimes B \xrightarrow{\varphi_A \otimes \varphi_B} A \otimes B$$

and by

$$\eta_{A\otimes B} := (\eta_A \otimes \eta_B) \circ \Delta_K : K \xrightarrow{\Delta_K} K \otimes K \xrightarrow{\eta_A \otimes \eta_B} A \otimes B,$$

$\Delta_K : K \xrightarrow{\cong} K \otimes K$ being the natural K-vector space isomorphism given by $\Delta(1) := 1 \otimes 1$ where $1 := 1_K$ is the identity element of the ground field K.

REMARK 2.39. (a) According to the above definition,

$$(a_1 \otimes b_1) \cdot (a_2 \otimes b_2) := (a_1 \cdot a_2) \otimes (b_1 \cdot b_2)$$

while

$$1_{A\otimes B} := 1_A \otimes 1_B.$$

(b) In the graded case, one uses the *graded twisting map* of 2.37(d). Thus, for example, if $a_m \in A_m$ and $b_p \in B_p$ while $a_q \in A_q$ and $b_n \in B_n$, then

$$\begin{aligned}(a_m \otimes b_p) \cdot (a_q \otimes b_n) &= (-1)^{p \cdot q} a_m \cdot a_q \otimes b_p \cdot b_n \\ &\in A_{m+q} \otimes B_{p+n} \\ &\subseteq (A \otimes B)_{m+q+p+n}.\end{aligned}$$

DEFINITION 2.40. The *tensor product of* the *coalgebras* $(C, \Delta_C, \varepsilon_C)$ and $(D, \Delta_D, \varepsilon_D)$, in that order, is the *coalgebra*

$$(C \otimes D, \Delta_{C \otimes D}, \varepsilon_{C \otimes D})$$

whose *comultiplication* $\Delta_{C \otimes D}$ and *counit* $\varepsilon_{C \otimes D}$ are the K-linear maps defined, respectively, by

$$\begin{aligned}\Delta_{C \otimes D} := S_{(23)} \circ (\Delta_C \otimes \Delta_D) : C \otimes D &\xrightarrow{\Delta_C \otimes \Delta_D} C \otimes C \otimes D \otimes D \\ &\xrightarrow{S_{(23)}} C \otimes D \otimes C \otimes D\end{aligned}$$

and by

$$\varepsilon_{C \otimes D} := \varphi_K \circ (\varepsilon_C \otimes \varepsilon_D) : C \otimes D \xrightarrow{\varepsilon_C \otimes \varepsilon_D} K \otimes K \xrightarrow{\varphi_K} K,$$

$\varphi_K : K \otimes K \xrightarrow{\cong} K$ being the natural K-vector space isomorphism given by the *multiplication of the ground field* K.

REMARKS 2.41. (a) According to the above definitions, for all $c \in C$ and for all $d \in D$,

$$\Delta_{C \otimes D}(c \otimes d) := \sum_c \sum_d (c_1 \otimes d_1) \otimes (c_2 \otimes d_2)$$

whenever

$$\Delta_C(c) := \sum_c c_1 \otimes c_2 \quad \text{and} \quad \Delta_D(d) := \sum_d d_1 \otimes d_2$$

while

$$\varepsilon_{C \otimes D}(c \otimes d) := \varepsilon_C(c) \cdot \varepsilon_D(d) \in K.$$

(b) Chari and Pressley, in [Char-Pres, p. 103], adopt the following suggestive notation (or mnemonic device), reminiscent of that used in defining the *classical Yang–Baxter equation* (cf. [Dr-2, p. 804] or [Mi-9, pp. 370–371]): If $\Delta_C(c) = \sum_c c_1 \otimes c_2$, if $\Delta_D(d) = \sum_d d_1 \otimes$

d_2, and if $1 := 1_K :=$ the identity element of the ground field K, set

$$\Delta_C^{13}(c) = \sum_c c_1 \otimes 1 \otimes c_2 \otimes 1 \in C \otimes K \otimes C \otimes K \subseteq C \otimes D \otimes C \otimes D,$$
$$\Delta_D^{24}(d) = \sum_d 1 \otimes d_1 \otimes 1 \otimes d_2 \in K \otimes D \otimes K \otimes D \subseteq C \otimes D \otimes C \otimes D,$$

and define

$$\begin{aligned}&(c_1 \otimes 1 \otimes c_2 \otimes 1) \cdot (1 \otimes d_1 \otimes 1 \otimes d_2)\\ &\quad = (c_1 \cdot 1) \otimes (1 \cdot d_1) \otimes (c_2 \cdot 1) \otimes (1 \cdot d_2) = c_1 \otimes d_1 \otimes c_2 \otimes d_2.\end{aligned}$$

With these definitions,

$$\Delta_{C\otimes D}(c \otimes d) := \Delta_C^{13}(c) \cdot \Delta_D^{24}(d).$$

(c) In the graded case, one uses the graded twisting map. Because the notation in that situation can get more involved, we postpone the details till later. For now, till further notice, all vector spaces, algebras and coalgebras will be ungraded.

DEFINITION 2.42. A *K-bialgebra*, or, more simply, a *bialgebra* is a 5-tuple $(H, \varphi, \eta, \Delta, \varepsilon)$ consisting of a K-vector space H and K-linear maps $\varphi : H \otimes H \to H$, $\eta : K \to H$, $\Delta : H \to H \otimes H$ and $\varepsilon : H \to K$ such that

(1) (H, φ, η) is an associative unitary algebra, i.e., a K-algebra;
(2) (H, Δ, ε) is an associative unitary coalgebra, i.e., a K-coalgebra; and
(3) Any one of the following *four* equivalent conditions holds:
  (a) Δ and ε are algebra morphisms.
  (b) φ and η are coalgebra morphisms.
  (c) The following *four* diagrams commute:

$$\begin{tikzcd} H \otimes H \arrow[rr, "\varphi"] \arrow[d, "\Delta\otimes\Delta"'] & & H \arrow[d, "\Delta"] \\ H \otimes H \otimes H \otimes H \arrow[r, "1\otimes\tau\otimes 1"] & H \otimes H \otimes H \otimes H \arrow[r, "\varphi\otimes\varphi"] & H \otimes H, \end{tikzcd}$$

(Figure 1)

$$\begin{tikzcd} K \arrow[rr, "\eta"] \arrow[d, equal] & & H \arrow[d, "\Delta"] \\ K \arrow[r, "\Delta_K"] & K \otimes K \arrow[r, "\eta\otimes\eta"] & H \otimes H \end{tikzcd}$$

(Figure 2)

$$\begin{array}{ccc} H\otimes H & \xrightarrow{\varphi} & H \\ \downarrow{\scriptstyle \varepsilon\otimes\varepsilon} & & \downarrow{\scriptstyle \varepsilon} \\ K\otimes K & \xrightarrow{\varphi_K} & K, \end{array} \qquad \begin{array}{ccc} K & \xrightarrow{\eta} & H \\ \| & & \downarrow{\scriptstyle \varepsilon} \\ K & \xrightarrow{1_K} & K. \end{array}$$

(Figure 3) (Figure 4)

(d) The following *four* identities hold:
(1) $\Delta(1_H) = 1_H \otimes 1_H$.
(2) $\Delta(a \cdot b) = \sum_a \sum_b (a_1 \cdot b_1) \otimes (a_2 \cdot b_2)$, for all $a, b \in H$.
(3) $\varepsilon(1_H) = 1_K$.
(4) $\varepsilon(a \cdot b) = \varepsilon(a) \cdot \varepsilon(b)$, for all $a, b \in H$.

REMARKS 2.43. (a) Conditions (a) and (d) of Definition 2.42(3) are clearly equivalent. To establish the equivalence of conditions (a), (b) and (c) of Definition 2.42(3), observe that

$$\Delta \text{ is an algebra map} \quad \Leftrightarrow \quad \text{Figures 1 and 2 commute}$$

and

$$\varepsilon \text{ is an algebra map} \quad \Leftrightarrow \quad \text{Figures 3 and 4 commute,}$$

whereas

$$\varphi \text{ is an coalgebra map} \quad \Leftrightarrow \quad \text{Figures 1 and 3 commute}$$

and

$$\eta \text{ is an coalgebra map} \quad \Leftrightarrow \quad \text{Figures 2 and 4 commute.}$$

(b) Sometimes we shall abbreviate $(H, \varphi, \eta, \Delta, \varepsilon)$ by H.

(c) We previously noted, respectively, in Definitions 2.2 and 2.6, above, that the ground field K may be considered trivially as the K-algebra (K, φ_K, η_K) (where $\varphi_K : K \otimes K \to K$ is the multiplication of K and where $\eta_K : K \to K$ is the identity map on K) and as the K-coalgebra $(K, \Delta_K, \varepsilon_K)$ (where $\varepsilon_K : K \to K$ is the identity map on K and where $\Delta_K : K \to K \otimes K$ is the K-linear map defined by $\Delta_K(1) = 1 \otimes 1$ where 1 is the identity element of K). It is trivial to verify that ε_K and Δ_K are K-algebra maps. Thus, $(K, \varphi_K, \eta_K, \Delta_K, \varepsilon_K)$ is a K-bialgebra.

DEFINITION 2.44. An *antipode for a bialgebra* $(H, \varphi, \eta, \Delta, \varepsilon)$ is a K-linear map $S : H \to H$ for which the following diagram

$$\begin{tikzcd}
H \otimes H \arrow[dd, "S\otimes 1"'] & H \arrow[l, "\Delta"'] \arrow[r, "\Delta"] \arrow[d, "\varepsilon"] & H \otimes H \arrow[dd, "1\otimes S"] \\
 & K \arrow[d, "\eta"] & \\
H \otimes H \arrow[r, "\varphi"'] & H & H \otimes H \arrow[l, "\varphi"]
\end{tikzcd}$$

commutes.

DEFINITION 2.45. A *Hopf algebra* is a 6-tuple $(H, \varphi, \eta, \Delta, \varepsilon, S)$ consisting of a *bialgebra* $(H, \varphi, \eta, \Delta, \varepsilon)$ together with an *antipode* S. Sometimes we shall abbreviate $(H, \varphi, \eta, \Delta, \varepsilon, S)$ by H. The *ground field* K may be considered trivially as the *Hopf algebra* $(K, \varphi_K, \eta_K, \Delta_K, \varepsilon_K, S_K)$ whose *bialgebra structure* $(K, \varphi_K, \eta_K, \Delta_K, \varepsilon_K)$ is specified by Remark 2.43(c) and whose *antipode* $S_K : K \to K$ is taken to be the *identity map* from K to itself.

REMARKS 2.46. (a) It may be shown (cf. [Kass, p. 50, Proposition III. 3.1(a)] or [Dăs-Năs-Rai, p. 151]) that whenever $H := (H, \varphi, \eta, \Delta, \varepsilon)$ is a *bialgebra* over K, then the set

$$M(H, H) := \mathrm{Hom}(H, H) := \mathbf{Vect}(H, H) =: \mathrm{End}(H)$$

of all K-vector space *endomorphisms* of H has the structure of a *monoid* (in fact, of an algebra, called the *convolution algebra* of H) whose *binary operation* $*$, called the *convolution product* is defined, for all $f, g \in \mathrm{End}(H)$, by

$$f * g := \varphi \circ (f \otimes g) \circ \Delta : H \xrightarrow{\Delta} H \otimes H \xrightarrow{f\otimes g} H \otimes H \xrightarrow{\varphi} H,$$

and whose *two-sided identity element* is given as the composite

$$\eta \circ \varepsilon : H \xrightarrow{\varepsilon} K \xrightarrow{\eta} H.$$

Accordingly, for any bialgebra H, an *antipode*, should it exist, is simply a *two-sided inverse in the convolution algebra of H* of the *identity map* $1_H : H \to H$; i.e., an *antipode* for H is a K-linear map $S : H \to H$ such that

$$S * 1_H = \eta \circ \varepsilon = 1_H * S \in \mathrm{Hom}(H, H).$$

The left-half of the diagram in 2.44 says (according to [Kap, p. 31]) that $S : H \to H$ is a *left antipode* (i.e., a *left inverse* of 1_H in the monoid $(\mathrm{Hom}(H, H), *, \eta \circ \varepsilon)$) while the right-half of that diagram says (loc. cit.) that $S : H \to H$ is a *right antipode* (i.e., a *right inverse* of 1_H in $(\mathrm{Hom}(H, H), *, \eta \circ \varepsilon)$). Because a *two-sided inverse* of any element of a monoid (should such an inverse exist) is *unique* (because any *left inverse* must necessarily coincide with any *right inverse*) it follows that the *antipode* of a *Hopf algebra* is always *unique*!

NOTE. More generally (cf. [Sw-1, p. 69]), whenever (C, Δ, ε) is a *coalgebra* and (A, φ, η) is an *algebra*, let $\mathrm{Hom}(C, A)$ denote the K-vector space of all K-vector space maps $f : C \to A$; let

$$\varphi_{\mathrm{Hom}(C,A)} : \mathrm{Hom}(C, A) \otimes \mathrm{Hom}(C, A) \to \mathrm{Hom}(C, A)$$

denote the composite

$$\mathrm{Hom}(C, A) \otimes \mathrm{Hom}(C, A) \hookrightarrow \mathrm{Hom}(C \otimes C, A \otimes A) \xrightarrow{\mathrm{Hom}(\Delta,\varphi)} \mathrm{Hom}(C, A)$$

of the *inclusion* followed by the K-linear map

$$\mathrm{Hom}(\Delta, \varphi) : \mathrm{Hom}(C \otimes C, A \otimes A) \to \mathrm{Hom}(C, A)$$

induced by the K-linear maps

$$\Delta : C \to C \otimes C \quad \text{and} \quad \varphi : A \otimes A \to A;$$

and let $\eta_{\mathrm{Hom}(C,A)} : K \to \mathrm{Hom}(C, A)$ denote the composite

$$K \xrightarrow[\cong]{\lambda} \mathrm{Hom}(K, K) \xrightarrow{\mathrm{Hom}(\varepsilon,\eta)} \mathrm{Hom}(C, A)$$

of the *natural isomorphism* $K \cong K^*$ given by the *left regular representation* $\lambda : K \xrightarrow{\cong} \mathrm{Hom}(K, K)$ *of* K *on itself* (given, for all $k, k' \in K$, by $\lambda(k)(k') := k \cdot k'$) followed by the K-linear map

$$\mathrm{Hom}(\varepsilon, \eta) : \mathrm{Hom}(K, K) \to \mathrm{Hom}(C, A)$$

induced by the K-linear maps $\varepsilon : C \to K$ and $\eta : K \to A$. Then, for all $f, g \in \mathrm{Hom}(C, A)$,

$$\mathrm{Hom}(\Delta, \varphi)(f \otimes g) : C \to A$$

is given as the composite $\varphi \circ (f \otimes g) \circ \Delta : C \to A$ defined by the commutative diagram

$$\begin{tikzcd}
C \otimes C \arrow[r, "f\otimes g"] & A \otimes A \arrow[d, "\varphi"] \\
C \arrow[u, "\Delta"] \arrow[r, "\mathrm{Hom}(\Delta,\varphi)(f\otimes g)"'] & A,
\end{tikzcd}$$

while, for all $k \in K$,

$$\operatorname{Hom}(\varepsilon, \eta)\big[\lambda(k)\big] : C \to A$$

is given as the composite $\eta \circ \lambda(k) \circ \varepsilon : C \to A$ defined by the commutative diagram

$$\begin{tikzcd} K \arrow[r, "\lambda(k)"] & K \arrow[d, "\eta"] \\ C \arrow[u, "\varepsilon"] \arrow[r, "{\operatorname{Hom}(\varepsilon,\eta)[\lambda(k)]}"'] & A. \end{tikzcd}$$

As maps, $\eta \circ \lambda(k) \circ \varepsilon = k \cdot (\eta \circ \varepsilon) \in \operatorname{Hom}(C, A)$ because, for all $c \in C$,

$$\begin{aligned} \eta \circ \lambda(k) \circ \varepsilon(c) &= \eta\big[\lambda(k)[\varepsilon(c)]\big] := \eta\big[k \cdot [\varepsilon(c)]\big] \\ &= k \cdot \eta\big[\varepsilon(c)\big] =: k \cdot (\eta \circ \varepsilon)(c). \end{aligned}$$

Accordingly, for all $f, g \in \operatorname{Hom}(C, A)$

$$\begin{aligned} f * g &:= \varphi_{\operatorname{Hom}(C,A)}(f \otimes g) := \operatorname{Hom}(\Delta, \varphi)(f \otimes g) \\ &:= \varphi \circ (f \otimes g) \circ \Delta \in \operatorname{Hom}(C, A) \end{aligned}$$

while

$$1_{\operatorname{Hom}(C,A)} := \eta_{\operatorname{Hom}(C,A)}(1_K) := 1_K \cdot (\eta \circ \varepsilon) = \eta \circ \varepsilon \in \operatorname{Hom}(C, A).$$

A proof of the fact that $(\operatorname{Hom}(C, A), \varphi_{\operatorname{Hom}(C,A)}, \eta_{\operatorname{Hom}(C,A)})$ is an algebra, called the *convolution* algebra of all linear maps from the coalgebra C to the algebra A, may be found in [Kass, p. 50, Proposition III.3.1(a)], [Kli-Sch, pp. 10–11, Proposition 1], or [Dăs-Năs-Rai, p. 151]. For every coalgebra (C, Δ, ε) and every algebra (A, φ, η), set

$$M(C, A) := \operatorname{Hom}(C, A).$$

From the above, it follows that the ordered triple $(M(C, A), *, \eta \circ \varepsilon)$ is a *monoid* whose binary operation, $* : M(C, A) \times M(C, A) \to M(C, A)$, is given, for all $f, g \in M(C, A)$ by

$$f * g := \varphi \circ (f \otimes g) \circ \Delta$$

and whose *identity element* is the K-linear map $\eta \circ \varepsilon : C \xrightarrow{\varepsilon} K \xrightarrow{\eta} A$.

NOTE. The *graded* counterpart of the above is as follows: If (C, Δ, ε) is a *connected graded coalgebra* (as defined in Remark 2.37(g), above) while (A, φ, η) is a *connected graded algebra* (as defined in Remark 2.37(f), above) and if $G(C, A)$ denotes the set of all morphisms $f : C \to A$ of graded K-vector spaces (as defined in Remark 2.37(d), above) for which the component, f_0, in degree 0 is the identity morphism of K, then it is shown in [M-M-1, p. 57, Corollary 6.8], [M-M-2, p. 259, Proposition 8.2], or [S-S, p. 33, Proposition 3.4.1] that the set $G(C, A)$ together with the binary operation

$$*: G(C, A) \times G(C, A) \to G(C, A)$$

defined, for all $f, g \in G(C, A)$ by

$$f * g := \varphi \circ (f \otimes g) \circ \Delta$$

is a *group* whose *identity element* is the composite

$$\eta \circ \varepsilon : C \xrightarrow{\varepsilon} K \xrightarrow{\eta} A.$$

Returning to the ungraded context, it is shown in [Kass, p. 50, Proposition III3.1(b)] that the K-vector space map

$$\iota_{C,A} : A \otimes C^* \to \mathrm{Hom}(C, A)$$

defined, for all $(a, f, c) \in A \times C^* \times C$, by

$$\iota_{C,A}(a \otimes f)(c) := f(c) \cdot a$$

is in fact a morphism of algebras where C^* is an algebra, viz., the *dual algebra* (defined in 2.58(a), below) of the coalgebra C, and $A \otimes C^*$ is then an algebra as the *tensor product of algebras* (as defined in Definition 2.38). The map $\iota_{C,A}$ is always *injective* (cf. [L-R, p. 2, Exercise 1.3.10]) and is bijective if either C or A is finite-dimensional (cf. [Kass, p. 28, Corollary II.2.3]). For the analogous result, with the roles of finite-dimensional A and C reversed, see [Sw-1, pp. 70–71, Exercise 3] where it is described how $\mathrm{Hom}(A, C)$ may be endowed with the structure of a coalgebra in such a way that the natural K-vector space isomorphism $\mathrm{Hom}(A, C) \cong C \otimes A^*$ is in fact a *coalgebra isomorphism*.

(b) It may be shown (cf. [M-M-2, p. 259, Proposition 8.6; p. 260, Proposition 8.7], [Hey-Sw-1, p. 204, Proposition 1.5.2] or [Gr-1, p. 19, Satz I.1.9]) that the *antipode* S of a Hopf algebra $(H, \varphi, \eta, \Delta, \varepsilon, S)$ is always a *bialgebra anti-endomorphism*; i.e., it is a *bialgebra morphism* from $H^{op,cop}$ to H (or, equivalently, a bialgebra morphism from H to $H^{op,cop}$) meaning

(1) that $S \circ \eta = \eta$ and that

$$\varphi \circ (S \otimes S) = S \circ \varphi^{op} := S \circ \varphi \circ \tau,$$

or, equivalently, by 2.33(b) since $\tau^2 = 1$, that

$$S \circ \varphi = \varphi \circ (S \otimes S) \circ \tau = \varphi \circ \tau \circ (S \otimes S) =: \varphi^{op} \circ (S \otimes S)$$

and

(2) that $\varepsilon \circ S = \varepsilon$ and that

$$(S \otimes S) \circ \Delta = \Delta^{op} \circ S := \tau \circ \Delta \circ S,$$

or, equivalently, by 2.33(b) since $\tau^2 = 1$, that

$$\Delta \circ S = \tau \circ (S \otimes S) \circ \Delta = (S \otimes S) \circ \tau \circ \Delta =: (S \otimes S) \circ \Delta^{op}.$$

Proofs of the above *using elements* provides (in the words of [Kli-Sch, p. 14] a "nice exercise" in the use of the *Heyneman–Sweedler notation*, and may be found in [Hey-Sw-1, pp. 204–205, Proposition 1.5.2], [Sw-1, pp. 74–78, Proposition 4.0.1], [Ab, pp. 62–65, Theorem 2.1.4], [Kass, pp. 52–54, Theorem III.3.4(a)], [L-R, pp. 42–43, Proposition 1.6.1], [Kli-Sch, pp. 13–14, Proposition 5], or [Dăs-Năs-Rai, pp. 153–155, Proposition 4.2.6]. A (more elegant) proof just using maps may be found (in part) in [M-M-1, pp. 60 and 61, Propositions 6.14 and 6.16], [M-M-2, pp. 259–260, Propositions 8.6 and 8.7], and [Gr-1, pp. 19–20, Satz I.1.9].

(c) If $H := (H, \varphi, \eta, \Delta, \varepsilon)$ is a *bialgebra*, the set

$$\mathrm{End}(H) := \mathrm{Hom}(H, H) := \mathbf{Vect}(H, H)$$

of all K-vector space maps from H to itself has *two distinct* monoid structures, namely,

$$\big(\mathrm{End}(H), *, \eta \circ \varepsilon\big),$$

the *convolution* algebra of H as noted above in 2.46(a), and

$$\big(\mathrm{End}(H), \circ, 1_H\big),$$

the usual *endomorphism algebra* of H. In other words, the set $\mathrm{End}(H)$ of vector space endomorphisms of H is a monoid under *convolution* (with *identity element* the composite $\eta \circ \varepsilon : H \xrightarrow{\varepsilon} K \xrightarrow{\eta} H$) as well as under *composition of maps* (with *identity element* the identity map $1_H : H \to H$). By definition, an *antipode* of a bialgebra H is an *inverse* S of 1_H in the *monoid* $(\mathrm{End}(H), *, \eta \circ \varepsilon)$, while the *order of the antipode* S of a Hopf algebra $(H, \varphi, \eta, \Delta, \varepsilon, S)$ is the *order* of the element S in the *monoid* $(\mathrm{End}(H), \circ, 1_H)$. In other words, if S is the antipode of a Hopf algebra H and if n is a positive integer, let

$$\begin{aligned} S^n &:= \text{the } n\text{-fold composite of } S \text{ with itself} \\ &:= S \circ \cdots \circ S \end{aligned}$$

(so that there are n copies of S in $S \circ \cdots \circ S$). If $S^n = 1_H$ for some positive integer n, then S is said to have *finite order*, and in this case the *order* of S is defined to be the *least positive integer n* for which the n-fold composite $S^n = 1_H \in \mathrm{Hom}(H, H) =: \mathrm{End}(H)$. On the other hand, if

$$S^n \neq 1_H \in \mathrm{End}(H)$$

for any positive integer n, then S is said to have *infinite order.* As noted in [Sw-1, p. 89], the antipode of the ground field K, considered trivially as a Hopf algebra according to 2.45 above has order 1, though one may also have $S = 1_H$ in less trivial cases. This is the case if $H = KG =$ the group algebra of a group G all of whose elements have order 2 (for instance, if $G = C_2 =$ the cyclic group of order 2, or if $G = C_2 \times C_2 =$ the Klein 4-group, etc.): For we shall see in 2.66(b), below, that if $\Delta(g) = g \otimes g$ for an element g of a Hopf algebra H, then necessarily g has an inverse in H, and moreover, $S(g) = g^{-1}$. Because KG has G as its *basis*, it follows that if the linear map $S : H \to H$ is the identity on a *basis*, then it is the identity map all over. Regarding the order of the antipode S of a Hopf algebra H, as noted in [Sw-1, p. 89], since $S : H \to H$ is a bialgebra anti-endomorphism, $S^n : H \to H$ will be a *bialgebra endomorphism* if n is *even* and a *bialgebra anti-endomorphism* if n is *odd.* It follows that if the antipode of a Hopf algebra H has *finite odd order* then the identity map $1_H : H \to H$ is a *bialgebra anti-endomorphism* whence $\varphi = \varphi \circ \tau$ and also $\Delta = \tau \circ \Delta$. Accordingly, as noted in [Larson-2, p. 127], if a Hopf algebra is neither commutative nor cocommutative, then the order of the antipode, if finite, must be even. Now in [Sw-1, pp. 89–90, Example-Exercise], Sweedler shows that the antipode of a Hopf algebra can have any even order or can have infinite order. See also [Mon-1, p. 8, Example 1.5.8]. In addition, Sweedler gave an example (unpublished, but described in [Lar-4], [Ta-1, p. 2632], [Mon-1, p. 8, Example 1.5.6], [Kass, p. 68, item 7; p. 174, Example 1; p. 196, Exercise 2; p. 198, Notes; pp. 363–364], [Kli-Sch, pp. 19–20, Example 9, p. 244, Example 2; p. 337, Example 2] or [Dăs-Năs-Rai, pp. 165–166, item 6]) of a 4-dimensional Hopf algebra having an *antipode* of order 4 provided that the characteristic of the ground field is *not* 2. When the ground field has characteristic 2, the antipode of Sweedler's example has order 2. Radford (in both [R-1, pp. 95–96] and [R-2, pp. 56–57]) gives (an example of) a 4-dimensional Hopf algebra H having an antipode S of order 4 independent of the characteristic of the ground field. (In fact, Radford gives an example of a free rank 4 Hopf algebra over an arbitrary commutative ring.) When the characteristic of the ground field is different from 2, Radford's example is isomorphic to Sweedler's. An explicit isomorphism, recounted below, is given in [Ta-1, p. 2633] where Taft, after giving a nice account of previous work, also constructs, for any prescribed integer $q \geqslant 2$ and any integer $n \geqslant 1$, a Hopf algebra of dimension q^{n+1} (over a field containing a primitive qth root of 1) having an antipode of order $2q$. When $q = 2$ and $n = 1$, Taft's example reduces to Sweedler's. As Taft notes, the condition on the ground field holds, in particular, if its characteristic is zero or relatively prime to q. As described in [Ta-1, p. 2633], Sweedler's 4-dimensional Hopf algebra, H, is described in terms of generators and relations as follows: As a vector space over a field K of characteristic $\neq 2$,

$$H = (K \cdot 1) \oplus (K \cdot x) \oplus (K \cdot y) \oplus (K \cdot w).$$

The multiplication of H is determined by the Cayley table

	1	x	y	w
1	1	x	y	w
x	x	1	w	y
y	y	$-w$	0	0
w	w	$-y$	0	0

while the diagonal of H is determined by setting

$$\Delta(1) = 1 \otimes 1,$$
$$\Delta(x) = x \otimes x,$$
$$\Delta(y) = y \otimes x + 1 \otimes y,$$

and

$$\Delta(w) = w \otimes 1 + x \otimes w.$$

Finally, the counit and antipode of H are defined by setting $\varepsilon(1) = \varepsilon(x) = 1$, $\varepsilon(y) = \varepsilon(w) = 0$, $S(1) = 1$, $S(x) = x$, $S(y) = w$, and $S(w) = -y$. By contrast, Radford's 4-dimensional Hopf algebra, H, has

$$H_1 = (K \cdot 1) \oplus (K \cdot z) \oplus (K \cdot a) \oplus (K \cdot b)$$

subject to

$$\begin{array}{c|cccc} & 1 & z & a & b \\ \hline 1 & 1 & z & a & b \\ z & z & 1 & b & a \\ a & a & -1+z-b & a & -a \\ b & b & 1-z-a & b & -b, \end{array}$$

$$\Delta(1) = 1 \otimes 1,$$
$$\Delta(z) = z \otimes z,$$
$$\Delta(a) = a \otimes z + 1 \otimes a,$$

and

$$\Delta(b) = b \otimes 1 + z \otimes b.$$

$\varepsilon(1) = \varepsilon(z) = 1$, $\varepsilon(a) = \varepsilon(b) = 0$, $S(1) = 1$, $S(z) = z$, $S(a) = 1 - z + b$, and $S(b) = -a$. The Hopf algebra isomorphism $F : H \to H_1$, given in [Ta-1, p. 2633] is defined (in case the characteristic of the ground field K is not equal to 2) by setting

$$F(1) := 1,$$
$$F(x) := z,$$
$$F(y) := -1 + z + 2a,$$

and

$$F(w) := 1 - z + 2b.$$

A proof that $H_1^* \cong H_1$ as Hopf algebras is given in [R-1, pp. 96–97, item 4.5.2] as well as in [R-2, p. 58, item 2.2]; it is also shown (cf. [R-1, p. 96, item 4.5.1] as well as [R-2, pp. 57–58, item 2.1.0]) that $H \cong H^{cop}$ and $H \cong H^{op}$ as Hopf algebras. Similarly, in [Char-Pres, pp. 131–132, item F(a)], it is shown that $H \cong H^*$ as Hopf algebras. This is also sketched in [Kli-Sch, p. 337, Example 2]. In all, seven interesting facts about H are established in [Char-Pres, pp. 131–133, Sweedler's example] including, among others, the fact that the only (non-zero) *group-like elements* of H are 1 and x, meaning (by Definition 2.65, below) that if $g \in H$ satisfies $\Delta(g) = g \otimes g$, then necessarily $g = 1$ or $g = x$. As noted by Kaplansky [Kap, p. 44, item c]: "Over an arbitrary field there is exactly one four-dimensional Hopf algebra which is neither commutative nor cocommutative. It is the one given by Radford ([R-2])." Montgomery (in [Mon-1, p. 8, Example 1.5.6]) states that the "smallest non-commutative, non-cocommutative Hopf algebra has dimension 4, and is unique (and so self-dual) for a given K of characteristic $\neq 2\ldots.$ This example was first described by Sweedler." Chari and Pressley [Char-Pres, p. 131, Sweedler's example] remark that "before the advent of quantum groups, very few examples of Hopf algebras which are neither commutative nor cocommutative were known (except for those obtained by taking tensor products of the standard commutative and cocommutative examples). However, Sweedler (1969) constructed an interesting four-dimensional example, which we shall use to illustrate some... concepts..." Other papers dealing with the construction of antipodes of any even order n are [Ta-W-1,Ta-W-2,R-T-W,Wa-1]. In this connection we call the reader's attention to the paper [Par-1] of Pareigis with the title "A non-commutative non-cocommutative Hopf algebra in 'nature'." It is an interesting but curious fact that the three-dimensional subcoalgebra (of Sweedler's four-dimensional Hopf algebra) determined by setting

$$\begin{aligned}
\Delta(1) &= 1 \otimes 1,\\
\Delta(x) &= x \otimes x,\\
\Delta(y) &= y \otimes x + 1 \otimes y,
\end{aligned}$$

arises (cf. [Me, p. 189, formula (2.4); and p. 198, Example 3]) in connection with *Azéma martingales*. (For information about Azéma martingales, see [Schür, pp. 81, 93, 94, 137]. As noted in [Schür, p. 81, §4, White noise on Bose Fock space], "Azéma martingales are closely related to additive q-white noise and to the interpolations between Bose and Fermi quantum Brownian motion... .")

(d) If H is a finite-dimensional Hopf algebra, then it is shown in [Lar-Sw, p. 83, Proposition 2] and in [Sw-1, p. 101, Corollary 5.1.6] that the antipode S of H is *bijective*. A proof using quite a different argument may be found in [L-R, p. 45, Theorem 1.6.2]. In point of fact, the bijectivity of the *antipode* S of a *finite-dimensional* Hopf algebra H follows at once from the fact established in [R-4] that the *order of the antipode of a finite-dimensional Hopf algebra is always finite*; for if $S^n = 1_H$ with $n \geqslant 2$, then $S^{n-1} \circ S = 1_H = S \circ S^{n-1}$, so that $S^{-1} = S^{n-1}$. (Of course if $S = 1_H$, then S is trivially *bijective*!) Above we remarked that if a Hopf algebra H is either commutative or cocommutative, then the antipode S is an involution, i.e., $S^2 = 1_H$. In [Kap, p. 45, Appendix 2, Conjecture 5], Kaplansky conjectured that if H is a finite-dimensional Hopf algebra over an algebraically closed field, and if either H or H^* is semisimple as an algebra, then $S^2 = 1$. In practice, Kaplansky's fifth conjecture is taken to be (cf. [Lar-Rad-1,Lar-Rad-2], or [Mon-1, p. 27]) the assertion that the antipode of a finite-dimensional cosemisimple Hopf algebra is an involution. This conjecture was solved [Lar-Rad-1, p. 195, Theorem 4] in characteristic 0 via two papers of Larson and Radford in that they proved that a finite-dimensional cosemisimple Hopf algebra over a field K of characteristic 0 is semisimple [Lar-Rad-2, p. 276, Theorem 3.3] and that if H is a finite-dimensional semisimple and cosemisimple Hopf algebra with antipode S over a field K of characteristic 0 or of characteristic $p > (\dim H)^2$, then $S^2 = 1$ [Lar-Rad-1, p. 194, Theorem 3]. As noted in [Mon-1, p. 27], "Kaplansky's conjecture remains open in characteristic p." But since the publication of [Mon-1], Yorck Sommerhäuser has settled Kaplansky's fifth conjecture over fields of large positive characteristic (cf. [Som, p. 223, Theorem]) by proving that if H is a finite-dimensional semisimple Hopf algebra, and if the characteristic p of K is 0 or satisfies $p > m^{m-4}$ where $m = 2(\dim H)^2$, then H is cosemisimple and the antipode of H is an involution.

Conditions under which the antipode of an infinite-dimensional Hopf algebra is injective are given in [Sw-1, p. 102, Cor. 5.1.7]. In general, the antipode of a Hopf algebra need not be bijective. An example is given by Takeuchi in [Tak-1].

At this juncture we note that for some authors (e.g., [Char-Pres, p. 103, Definition 4.1.3(iv)] or [Etin-Schi, p. 68, Definition 8.1.3]), the antipode of a Hopf algebra is bijective by definition, so the reader should beware (though Chari and Pressley [Char-Pres, p. 104, Remark 5] do remark that their assumption that S is bijective is not customary).

(e) These days there is a great interest in quantum groups, another name given to certain special Hopf algebras having neither a commutative multiplication φ nor a commutative comultiplication Δ. The fact that $S^2 = 1$ if either $\varphi = \varphi \circ \tau$ or $\Delta = \tau \circ \Delta$ together with the existence of Hopf algebras for which $S^2 \neq 1$ indicates, from a Hopf theoretic point of view, one reason for the current interest in and importance of quantum groups. In this context it is nonetheless worthwhile to realize that there are conditions (as described above) other than $\varphi = \varphi \circ \tau$ or $\Delta = \tau \circ \Delta$ that will insure that $S^2 = 1$.

REMARKS 2.47. (a) In the first section, we sketched the way in which functions on a group can give rise to Hopf algebras – the antipode being induced by the inverse map $x \mapsto x^{-1}$. If one starts with a semi-group with identity (i.e., with a monoid rather than with a group) one winds up with a bialgebra rather than a Hopf algebra. Another connection between groups and Hopf algebras is that certain kinds of Hopf algebras are *group objects* in an appropriate category of coalgebras (cf. [Gr-1, p. 20, Korollar I.1.11]).

(b) In Lemma 2.66(b), below, we shall prove that if H is a Hopf algebra with antipode S and if g is an element of H for which $\Delta(g) = g \otimes g$ (such an element will be called a *group-like element* of H), then g has an inverse in H and $S(g) = g^{-1}$.

(c) In Lemma 2.74(b), below, we shall prove that if H is a Hopf algebra with antipode S and if p is an element of H for which $\Delta(p) = p \otimes 1 + 1 \otimes p$ (such an element will be called a *primitive element* of H), then $S(p) = -p$.

(d) In their seminal paper "On the structure of Hopf algebras," Milnor and Moore [M-M-2] did *not* include an antipode as part of their definition of a Hopf algebra. However, the Hopf algebras that they were interested in were *graded* and *connected* [*connected* (compare with Remarks 2.37(f) and (g), above, and with part (e), below) meaning that the *component* $H_0 \cong K$]; and it turns out ([M-M-2, p. 259, Proposition 8.2] or [S-S, p. 34, Proposition 3.4.2]) that a connected graded bialgebra always has an antipode! It is, however, to be noted that Milnor and Moore used the term *conjugation* [M-M-2, p. 259, Definition 8.4] for what we call an *antipode*. In using the word "antipode," we follow the terminology used by Sweedler [Sw-1, p. 71], Abe [Ab, p. 61], and Hochschild [H-4, p. 18]; [H-5, p. 5]. Other expressions used for an *antipode* are *antipodism* used by Cartan–Eilenberg [C-E, pp. 222, 269, 351] and Ditters [Dit, p. 2, Definition 1.1.6], *inversion* used by Ditters

[Dit, p. 2, Definition 1.1.6], *symmetry* used by Hochschild [H-2, p. 27], and *involution* used by Grünenfelder [Gr-1, p. 19, Definition I.1.11] and by Michaelis [Mi-2, p. 34]. If H has either a commutative multiplication φ or a commutative comultiplication Δ, then $S^2 = 1$ (cf. [Sw-1, p. 74, Proposition 4.0.1(6)] or [Gr-1, p. 19, Satz I.1.9]). On the other hand, for any even integer n one can construct a (finite-dimensional) Hopf algebra whose antipode has order n – cf. Taft and Wilson [Ta-W-2] as well as Taft [Ta-1], Radford [R-2,R-3,R-4], Taft and Wilson [Ta-W-1], Radford, Taft, and Wilson [R-T-W], and Waterhouse [Wa-1].

(e) *A graded bialgebra over K* (or *a graded K-bialgebra*, or, more simply, *a graded bialgebra*) is defined, *mutatis mutandis*, as a 5-tuple $(H, \varphi, \eta, \Delta, \varepsilon)$ in precise analogy to an ungraded bialgebra as defined in Definition 2.42 (with condition (3)(d) left out or else suitably modified) with the obvious qualification that $H := \{H_n\}_{n=0}^{\infty}$ is a graded K-vector space; that all tensor products appearing in the definitions of $\varphi : H \otimes H \to H$ and of $\Delta : H \to H \otimes H$, as well as in any diagrams, are tensor products of graded K-vector spaces; that all maps are maps of graded K-vector spaces; that the twisting map $\tau : H \otimes H \to H \otimes H$ when it appears (e.g., so as to modify 2.42 (3)(d)(2)) is the *graded twisting map* defined in 2.37(d); and finally that all the diagrams that are required to commute are to be considered as diagrams in the category of graded K-vector spaces. *A graded Hopf algebra over K* (or *a graded K-Hopf algebra*, or, more simply, *a graded Hopf algebra*) is defined as a 6-tuple $(H, \varphi, \eta, \Delta, \varepsilon, S)$ analogously. A *graded K-bialgebra* $(H, \varphi, \eta, \Delta, \varepsilon)$, or a *graded K-Hopf algebra* $(H, \varphi, \eta, \Delta, \varepsilon, S)$, is *connected* in case $\eta_0 : K \to H_0$ is a K-vector space *isomorphism* with inverse $\varepsilon_0 : H_0 \to K$. As noted in (d), above, any connected graded bialgebra over K is a graded Hopf algebra over K.

NOTE. The term "connected" comes from the fact that the *rational cohomology* of a connected Lie group having the homotopy type of a finite CW complex is a *connected* graded Hopf algebra over $\mathbb{Q}$. In fact, more generally, the 0th homology group or cohomology group of a connected topological space is always the coefficient group.

DEFINITION/NOTATION 2.48. (a) If $(A, \varphi_A, \eta_A, \Delta_A, \varepsilon_A)$ and $(B, \varphi_B, \eta_B, \Delta_B, \varepsilon_B)$ are bialgebras, then a *bialgebra map f* from $(A, \varphi_A, \eta_A, \Delta_A, \varepsilon_A)$ to $(B, \varphi_B, \eta_B, \Delta_B, \varepsilon_B)$ is a K-vector space map $f : A \to B$ which is simultaneously an *algebra map* from (A, φ_A, η_A) to (B, φ_B, η_B) and a *coalgebra map* from $(A, \Delta_A, \varepsilon_A)$ to $(B, \Delta_B, \varepsilon_B)$. The *category of all bialgebras and bialgebra maps* will be denoted **Bialg**.

(b) If $(A, \varphi_A, \eta_A, \Delta_A, \varepsilon_A, S_A)$ and $(B, \varphi_B, \eta_B, \Delta_B, \varepsilon_B, S_B)$ are Hopf algebras with antipodes S_A and S_B, respectively, then a *Hopf algebra map* from $(A, \varphi_A, \eta_A, \Delta_A, \varepsilon_A, S_A)$ to $(B, \varphi_B, \eta_B, \Delta_B, \varepsilon_B, S_B)$ is a *bialgebra map f* from $(A, \varphi_A, \eta_A, \Delta_A, \varepsilon_A)$ to $(B, \varphi_B, \eta_B, \Delta_B, \varepsilon_B)$ for which $f \circ S_A = S_B \circ f$.

NOTE. By [Sw-1, pp. 81–82, Lemma 4.0.4], or [Dăs-Năs-Rai, p. 152, Proposition 4.2.5], *any bialgebra map between Hopf algebras is*, in fact, *a Hopf algebra map*. The *category of all Hopf algebras and Hopf algebra maps* will be denoted **Hopf Alg**.

REMARK 2.49. One of the defining conditions (cf. Definition 2.42(3)) for a bialgebra or a Hopf algebra is that the coalgebra structure maps

$$\Delta : H \to H \otimes H \quad \text{and} \quad \varepsilon : H \to K$$

should be algebra maps or, equivalently, that the algebra structure maps

$$\varphi : H \otimes H \to H \quad \text{and} \quad \eta : K \to H$$

should be coalgebra maps. It is therefore natural to wonder under what – if any – conditions the algebra structure maps are algebra maps or the coalgebra structure maps are coalgebra maps.

PROPOSITION 2.50. (a) *If (A, φ, η) is an algebra over K and if $(K, \varphi_K, 1_K)$ is considered trivially as a K-algebra according to the definition given in Definition* 2.2*, then $\eta : K \to A$ is an algebra map.*

(b) *If (C, Δ, ε) is a coalgebra over K and if $(K, \Delta_K, 1_K)$ is considered trivially as a K-coalgebra according to the definition given in Definition* 2.6*, then $\varepsilon : C \to K$ is a coalgebra map. Hence* $\operatorname{Ker}\varepsilon$ *is a coideal of C.*

(c) *If $\gamma : K \to A$ is a morphism of algebras, then $\gamma = \eta$.*

(d) *If $\gamma : C \to K$ is a morphism of coalgebras, then $\gamma = \varepsilon$.*

(e) *K is a zero object in the category of bialgebras and in the category of Hopf algebras.*

(f) *The unit map $\eta : K \to A$ of an algebra (A, φ, η) is unique, hence so is the unit* (*or identity*) *element $1_A := \eta(1_K)$ of A.*

(g) *The counit map $\varepsilon : C \to K$ of a coalgebra (C, Δ, ε) is unique.*

PROOF. (a) We must show that the following two diagrams commute:

$$\begin{array}{ccc} K \otimes K & \xrightarrow{\varphi_K} & K \\ {\scriptstyle \eta\otimes\eta}\downarrow & & \downarrow{\scriptstyle \eta} \\ A \otimes A & \xrightarrow[\varphi_A]{} & A \end{array} \quad \text{and} \quad \begin{array}{ccc} K & \xrightarrow{1_K} & K \\ \| & & \downarrow{\scriptstyle \eta} \\ K & \xrightarrow[\eta]{} & A. \end{array}$$

Now the diagram at the right obviously commutes while the diagram at the left commutes since η is K-linear and $\eta(1_K) =: 1_A :=$ the *identity element* of A so that

$$\begin{aligned} \eta(k_1 \cdot k_2) &= \eta(k_1 \cdot k_2 \cdot 1_K) = k_1 \cdot k_2 \cdot 1_A = k_1 \cdot 1_A \cdot k_2 \cdot 1_A \\ &= \eta(k_1 \cdot 1_K) \cdot \eta(k_2 \cdot 1_K) = \eta(k_1) \cdot \eta(k_2). \end{aligned}$$

(b) “Dually”, with regard to the diagrams

$$\begin{array}{ccc} C & \xrightarrow{\Delta_C} & C \otimes C \\ {\scriptstyle \varepsilon}\downarrow & & \downarrow{\scriptstyle \varepsilon\otimes\varepsilon} \\ K & \xrightarrow{\Delta_K} & K \otimes K \end{array} \quad \text{and} \quad \begin{array}{ccc} C & \xrightarrow{\varepsilon} & K \\ {\scriptstyle \varepsilon}\downarrow & & \| \\ K & \xrightarrow{1_K} & K \end{array}$$

the diagram at the right obviously commutes while the diagram at the left commutes since if $c \in C$ with $\Delta(c) = \sum_c c_1 \otimes c_2$, then $c = \sum_c \varepsilon(c_1) \cdot c_2$ so $\varepsilon(c) = \sum_c \varepsilon(c_1) \cdot \varepsilon(c_2)$ and therefore (because Δ is K-linear)

$$\begin{aligned} \Delta\big[\varepsilon(c)\big] &= \varepsilon(c) \cdot \Delta(1) = \varepsilon(c) \cdot (1 \otimes 1) \\ &= \Bigg[\sum_c \varepsilon(c_1) \cdot \varepsilon(c_2)\Bigg] \cdot (1 \otimes 1) \end{aligned}$$

$$\begin{aligned} &= \sum_c \varepsilon(c_1) \otimes \varepsilon(c_2) \\ &= (\varepsilon \otimes \varepsilon)\big[\Delta(c)\big]. \end{aligned}$$

(c) This is immediate from the commutativity of the diagram

$$\begin{tikzcd} K \arrow[r, "1"] \arrow[d, equal] & K \arrow[d, "\gamma"] \\ K \arrow[r, "\eta"'] & A. \end{tikzcd}$$

(d) This is immediate from the commutativity of the diagram

$$\begin{tikzcd} C \arrow[r, "\varepsilon"] \arrow[d, "\gamma"'] & K \arrow[d, equal] \\ K \arrow[r, "1"] & K. \end{tikzcd}$$

(e) This is immediate from Definition 2.42 together with (a) and (c), respectively, (b) and (d), above.

(f) This is immediate from (a) and (c), above.

(g) This is immediate from (b) and (d), above. □

The graded version of the following result is stated without proof in [M-M-2, pp. 213 and 218] and in [Gr-1, p. 11, Satz I.1.1].

PROPOSITION 2.51. (a) *If* (A, φ, η) *is a unitary algebra with a not necessarily associative multiplication* φ*, then*

$$\varphi \text{ is an algebra map} \quad \Leftrightarrow \quad \varphi \text{ is associative and commutative.}$$

(b) *If* (C, Δ, ε) *is a unitary coalgebra with a not necessarily associative comultiplication* Δ*, then*

$$\Delta \text{ is a coalgebra map} \quad \Leftrightarrow \quad \Delta \text{ is associative and commutative.}$$

PROOF. (a) ($\Rightarrow$) First suppose that $\varphi : A \otimes A \to A$ is a morphism of K-algebras. Then, in particular, the following diagram commutes:

$$\begin{tikzcd} A \otimes A \otimes A \otimes A \arrow[r, "1\otimes\tau\otimes 1"] \arrow[d, "\varphi\otimes\varphi"'] & A \otimes A \otimes A \otimes A \arrow[r, "\varphi\otimes\varphi"] & A \otimes A \arrow[d, "\varphi"] \\ A \otimes A \arrow[rr, "\varphi"] & & A. \end{tikzcd}$$

If $a, b, c, d \in A$, then chasing the element $a \otimes b \otimes c \otimes d \in A \otimes A \otimes A \otimes A$ around the diagram, beginning in the northwest corner, we find that

$$(a \cdot b) \cdot (c \cdot d) = (a \cdot c) \cdot (b \cdot d) \tag{$*$}$$

for all $a, b, c, d \in A$. By taking $c = 1_A = \eta(1_K)$ in $(*)$ we find that

$$(a \cdot b) \cdot d = a \cdot (b \cdot d), \quad \text{for all } a, b, d \in A,$$

so φ is associative. By taking $a = 1_A$ and $d = 1_A$ in $(*)$, we find that

$$b \cdot c = c \cdot b, \quad \text{for all } b, c \in A,$$

so φ is commutative. We now establish the above result in an element-free way by translating the elementwise proofs into diagrammatic proofs by means of the following commutative diagrams, the first of which establishes the associativity of φ and the second of which establishes the commutativity of φ.

$$\begin{tikzcd}
A\otimes A\otimes A \arrow[r, equal] \arrow[dddd, "\varphi\otimes 1"'] & A\otimes A\otimes A \arrow[r, equal] \arrow[d, "1\otimes 1\otimes\lambda^{-1}"'] & A\otimes A\otimes A \arrow[r, equal] \arrow[d, "\rho^{-1}\otimes 1\otimes 1"] & A\otimes A\otimes A \arrow[ddd, "1\otimes\varphi"] \\
& A\otimes A\otimes K\otimes A \arrow[r, "1\otimes\tau\otimes 1"] \arrow[d, "1\otimes 1\otimes\eta\otimes 1"'] & A\otimes K\otimes A\otimes A \arrow[d, "1\otimes\eta\otimes 1\otimes 1"] & \\
& A\otimes A\otimes A\otimes A \arrow[r, "1\otimes\tau\otimes 1"] \arrow[dd, "\varphi\otimes\varphi"'] & A\otimes A\otimes A\otimes A \arrow[d, "\varphi\otimes\varphi"] & \\
& & A\otimes A \arrow[r, equal] & A\otimes A \arrow[d, "\varphi"] \\
A\otimes A \arrow[r, equal] & A\otimes A \arrow[rr, "\varphi"] & & A
\end{tikzcd}$$

(For the associativity of φ)

$$\begin{tikzcd}
A\otimes A \arrow[r, equal] \arrow[dddd, "1\otimes 1"'] & A\otimes A \arrow[r, "\tau"] \arrow[d, "\lambda^{-1}\otimes\rho^{-1}"'] & A\otimes A \arrow[r, equal] \arrow[d, "\lambda^{-1}\otimes\rho^{-1}"] & A\otimes A \arrow[ddd, "1\otimes 1"] \\
& K\otimes A\otimes A\otimes K \arrow[r, "1\otimes\tau\otimes 1"] \arrow[d, "\eta\otimes 1\otimes 1\otimes\eta"'] & K\otimes A\otimes A\otimes K \arrow[d, "\eta\otimes 1\otimes 1\otimes\eta"] & \\
& A\otimes A\otimes A\otimes A \arrow[r, "1\otimes\tau\otimes 1"] \arrow[dd, "\varphi\otimes\varphi"'] & A\otimes A\otimes A\otimes A \arrow[d, "\varphi\otimes\varphi"] & \\
& & A\otimes A \arrow[r, equal] & A\otimes A \arrow[d, "\varphi"] \\
A\otimes A \arrow[r, equal] & A\otimes A \arrow[rr, "\varphi"] & & A
\end{tikzcd}$$

(For the commutativity of φ)

($\Leftarrow$) Conversely, suppose that φ is associative and commutative. We must show that the following two diagrams commute:

$$\begin{tikzcd}
A \otimes A \otimes A \otimes A \arrow[r, "1\otimes\tau\otimes 1"] \arrow[d, "\varphi\otimes\varphi"'] & A \otimes A \otimes A \otimes A \arrow[r, "\varphi\otimes\varphi"] & A \otimes A \arrow[d, "\varphi"] \\
A \otimes A \arrow[rr, "\varphi"] & & A
\end{tikzcd}$$

and

$$\begin{tikzcd}
K \arrow[r, "\Delta_K"] \arrow[d, equal] & K \otimes K \arrow[r, "\eta\otimes\eta"] & A \otimes A \arrow[d, "\varphi"] \\
K \arrow[rr, "\eta"] & & A .
\end{tikzcd}$$

The second diagram commutes because (by 2.50(a)) $\eta : K \to A$ is an algebra morphism and $\Delta_K = \varphi_K^{-1}$. The first diagram commutes because it can be decomposed into commutative sub-diagrams as follows:

$$\begin{tikzcd}[column sep=small]
 & & & A^{\otimes 4} \arrow[rrrr, "1\otimes\tau\otimes 1"] \arrow[ddlll, "\varphi\otimes\varphi"'] \arrow[d, "1\otimes 1\otimes\varphi"] \arrow[drr, "1\otimes\varphi\otimes 1"] & & & & A^{\otimes 4} \arrow[dll, "1\otimes\varphi\otimes 1"'] \arrow[d, "1\otimes 1\otimes\varphi"] \arrow[ddrrr, "\varphi\otimes\varphi"] & & & \\
 & & & A^{\otimes 3} \arrow[dl, "\varphi\otimes 1"'] \arrow[dr, "1\otimes\varphi"] & & A^{\otimes 3} \arrow[dl, "1\otimes\varphi"'] \arrow[dr, "1\otimes\varphi"] & & A^{\otimes 3} \arrow[dl, "1\otimes\varphi"'] \arrow[dr, "\varphi\otimes 1"] & & & \\
A^{\otimes 2} \arrow[drrrrr, "\varphi"'] & & A^{\otimes 2} \arrow[drrr, "\varphi"] & & A^{\otimes 2} \arrow[dr, "\varphi"] & & A^{\otimes 2} \arrow[dl, "\varphi"'] & & A^{\otimes 2} \arrow[dlll, "\varphi"'] & & A^{\otimes 2} \arrow[dlllll, "\varphi"] \\
 & & & & & A & & & & &
\end{tikzcd}$$

This diagram commutes because reading it from left to right across the top (= outer) layer (consisting of 2 quadrilaterals followed by 1 triangle followed by 2 quadrilaterals) and then from left to right across the inner band (of 3 quadrilaterals) one finds that the 5 subdiagrams of the top (= outer) layer commute successively by definition (since $\varphi \circ (1 \otimes 1) = \varphi$), by the associativity of φ, by the commutativity of φ, by the associativity of φ, and by definition (since $\varphi \circ (1 \otimes 1) = \varphi$) while the 3 quadrilaterals of the inner band successively commute by the associativity of φ, by definition, and by the associativity of φ.

NOTE. The commutativity of the above diagram records the fact that, for all $a, b, c, d \in A$,

$$
\begin{aligned}
(a \cdot b) \cdot (c \cdot d) &= a \cdot \big[b \cdot (c \cdot d)\big] \\
&= a \cdot \big[(b \cdot c) \cdot d\big] \\
&= a \cdot \big[(c \cdot b) \cdot d\big] \\
&= a \cdot \big[c \cdot (b \cdot d)\big] \\
&= (a \cdot c) \cdot (b \cdot d)
\end{aligned}
$$

since "·" is associative and commutative. Now, even though establishing the commutativity of the immediately preceding diagram amounts to showing (as we have just done, above) that

$$(a \cdot b) \cdot (c \cdot d) = (a \cdot c) \cdot (b \cdot d), \quad \text{for all } a, b, c, d \in A,$$

an "element-free" proof is preferred because it can so readily be "dualized", as we do next.

(b) ($\Rightarrow$) First suppose that $\Delta : C \to C \otimes C$ is a morphism of K-coalgebras. Then, in particular, the following diagram commutes:

$$
\begin{tikzcd}
C \arrow[rr, "\Delta"] \arrow[d, "\Delta"'] & & C \otimes C \arrow[d, "\Delta\otimes\Delta"] \\
C \otimes C \arrow[r, "\Delta\otimes\Delta"] & C \otimes C \otimes C \otimes C \arrow[r, "1\otimes\tau\otimes1"] & C \otimes C \otimes C \otimes C.
\end{tikzcd}
$$

The commutativity of this diagram implies the commutativity of the two diagrams that follow. The first of these establishes the *associativity* of Δ; the second establishes the *commutativity* of Δ.

$$
\begin{tikzcd}
C \arrow[rr, "\Delta"] \arrow[d, "\Delta"'] & & C \otimes C \arrow[r, equal] \arrow[dd, "\Delta\otimes\Delta"] & C \otimes C \arrow[dddd, "1\otimes\Delta"] \\
C \otimes C \arrow[r, equal] \arrow[ddd, "\Delta\otimes1"'] & C \otimes C \arrow[d, "\Delta\otimes\Delta"'] & & \\
& C \otimes C \otimes C \otimes C \arrow[r, "1\otimes\tau\otimes1"] \arrow[d, "1\otimes1\otimes\varepsilon\otimes1"'] & C \otimes C \otimes C \otimes C \arrow[d, "1\otimes\varepsilon\otimes1\otimes1"] & \\
& C \otimes C \otimes K \otimes C \arrow[r, "1\otimes\tau\otimes1"] \arrow[d, "1\otimes1\otimes\lambda"'] & C \otimes K \otimes C \otimes C \arrow[d, "\rho\otimes1\otimes1"] & \\
C \otimes C \otimes C \arrow[r, equal] & C \otimes C \otimes C \arrow[r, equal] & C \otimes C \otimes C \arrow[r, equal] & C \otimes C \otimes C
\end{tikzcd}
$$

(For the associativity of Δ)

$$\begin{tikzcd}
C \arrow[rr, "\Delta"] \arrow[d, "\Delta"'] & & C\otimes C \arrow[r, equal] \arrow[dd, "\Delta\otimes\Delta"] & C\otimes C \arrow[dddd, "1\otimes 1"] \\
C\otimes C \arrow[r, equal] \arrow[ddd, "1\otimes 1"'] & C\otimes C \arrow[d, "\Delta\otimes\Delta"'] & & \\
& C\otimes C\otimes C\otimes C \arrow[r, "1\otimes\tau\otimes 1"] \arrow[d, "\varepsilon\otimes 1\otimes 1\otimes\varepsilon"'] & C\otimes C\otimes C\otimes C \arrow[d, "\varepsilon\otimes 1\otimes 1\otimes\varepsilon"] & \\
& K\otimes C\otimes C\otimes K \arrow[r, "1\otimes\tau\otimes 1"] \arrow[d, "\lambda\otimes\rho"'] & K\otimes C\otimes C\otimes K \arrow[d, "\lambda\otimes\rho"] & \\
C\otimes C \arrow[r, equal] & C\otimes C \arrow[r, "\tau"] & C\otimes C \arrow[r, equal] & C\otimes C
\end{tikzcd}$$

(For the commutativity of Δ)

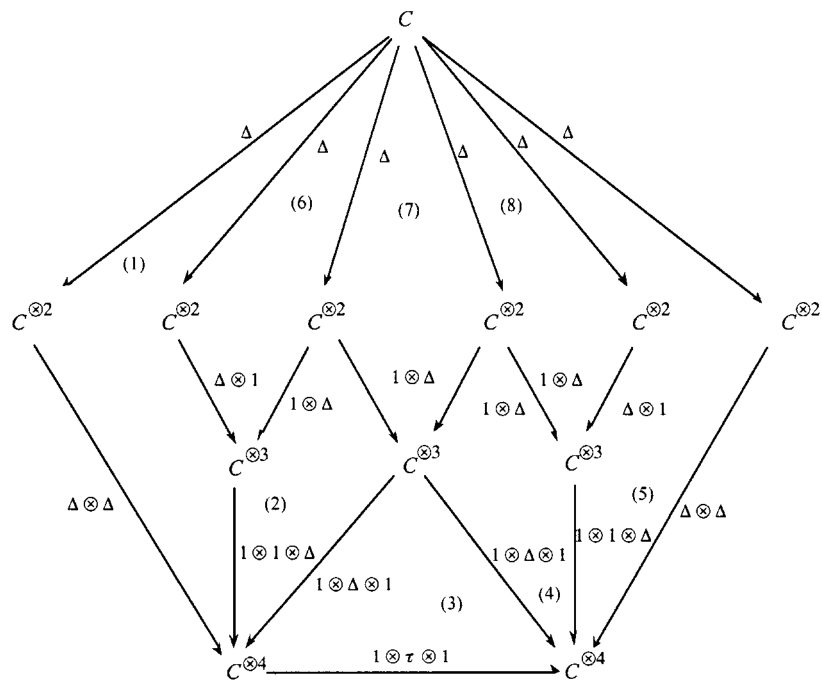
($\Leftarrow$) Conversely, suppose that Δ is associative and commutative. We must show that the following two diagrams commute:

$$\begin{tikzcd}
C \arrow[rr, "\Delta"] \arrow[d, "\Delta"'] & & C\otimes C \arrow[d, "\Delta\otimes\Delta"] \\
C\otimes C \arrow[r, "\Delta\otimes\Delta"] & C\otimes C\otimes C\otimes C\otimes C \arrow[r, "1\otimes\tau\otimes 1"] & C\otimes C\otimes C\otimes C
\end{tikzcd}$$

and

$$\begin{tikzcd}
C \arrow[rr, "\varepsilon"] \arrow[d, "\Delta"'] & & K \arrow[d, equal] \\
C\otimes C \arrow[r, "\varepsilon\otimes\varepsilon"] & K\otimes K \arrow[r, "\varphi_K"] & K.
\end{tikzcd}$$

The second diagram commutes because (by 2.50(b)) $\varepsilon : C \to K$ is a coalgebra morphism and $\varphi_K = \Delta_K^{-1}$. The first diagram commutes because it can be decomposed into commutative sub-diagrams as follows:



In the above diagram, subdiagrams (2), (4), (6), and (8) commute by the associativity of Δ; (3) commutes by the commutativity of Δ; (1) and (5) commute because $\Delta \circ 1 = \Delta$ and $(1 \otimes 1) \circ \Delta = \Delta$; and finally (7) obviously commutes. $\square$

REMARK 2.52. Before proceeding further, we pause to consider some examples of coalgebras. For us the most transparent example of a coalgebra is the *dual coalgebra* of a *finite-dimensional algebra*. Here *dual* refers to *vector space dual*. The importance of the vector space dual for what follows lies in the fact that the *vector space dual of a coalgebra* inherits naturally the structure of an *algebra* (referred to as the *dual algebra* of the given *coalgebra*) while the *vector space dual of a finite-dimensional algebra* inherits naturally the structure of a *finite-dimensional coalgebra* (referred to as the *dual coalgebra* of the given *finite-dimensional algebra*). In Section 3 (respectively, Section 4), we shall consider in detail just how algebras give rise to coalgebras and vice-versa. For the time being, however, we shall content ourselves with writing down careful statements of those results needed to present the examples we choose to showcase at this time. We begin with some notation.

NOTATION 2.53. For any ordered pair (V, W) of K-vector spaces let

$$\rho = \rho_{V,W} : V^* \otimes W^* \to (V \otimes W)^*$$

denote the natural K-linear map given by

$$\rho(f \otimes g)(x \otimes y) := f(x) \cdot g(y) \in K$$

for all $f \in V^*$, $g \in W^*$, $x \in V$, and $y \in W$, or, equivalently, by

$$\rho(f \otimes g) = \varphi_K \circ (f \otimes g)$$

where $\varphi_K : K \otimes K \to K$ denotes the multiplication of the ground field K. When $W = V$, we write

$$\rho_V := \rho_{V,V} : V^* \otimes V^* \to (V \otimes V)^*.$$

NOTE. This ρ should not be confused with the ρ of Definition 2.6.

REMARK 2.54. It is well-known (cf. [Sw-1, p. 317, Proposition A.2]) that $\rho_{V,W}$ is always *injective* and (cf. [Mac-1, p. 147, Proposition 4.3] or [Bly, p. 227, Corollary]) that $\rho_{V,W}$ is *surjective* if either V or W is *finite-dimensional*.

DEFINITION 2.55. The *left regular representation of* the ground field K *on itself* is the K-linear map $\lambda : K \to K^*$ defined by

$$\lambda(k)(k') := k \cdot k', \quad \text{for all } k, k' \in K,$$

i.e., by $\lambda(k) := f_k : K \to K$ where

$$f_k(k') := k \cdot k', \quad \text{for all } k, k' \in K.$$

NOTE. This λ should not be confused with the λ of Definition 2.6.

REMARK 2.56. It is easy to see that $\lambda : K \to K^*$ is a K-vector space *isomorphism* with *inverse* $\lambda^{-1} : K^* \to K$ given, for all $f \in K^*$, by

$$\lambda^{-1}(f) := f(1) \in K$$

where $1 := 1_K :=$ the *identity element* of K.

REMARK 2.57. If (A, φ, η) is an algebra and A is *finite-dimensional*, then the *multiplication* $\varphi : A \otimes A \to A$ on A gives rise to a linear map

$$\Delta : A^* \xrightarrow{\varphi^*} (A \otimes A)^* \xrightarrow{\rho^{-1}} A^* \otimes A^*$$

on A^* which furnishes A^* with a *comultiplication*, while the *unit* $\eta : K \to A$ on A gives rise to a linear map

$$\varepsilon : A^* \xrightarrow{\eta^*} K^* \xrightarrow{\lambda^{-1}} K$$

on A^* which furnishes A^* with a *counit*. If $f \in A^*$, then

$$\begin{aligned}
&\rho^{-1} \circ \varphi^*(f) =: \Delta(f) := \sum_f f_1 \otimes f_2 \\
&\quad \Leftrightarrow \quad f \circ \varphi =: \varphi^*(f) = \rho\Bigl(\sum_f f_1 \otimes f_2\Bigr) \\
&\quad \Leftrightarrow \quad f(x \cdot_A y) = \sum_f f_1(x) \cdot f_2(y), \quad \text{for all } x, y \in A;
\end{aligned}$$

while if $f \in A^*$, then

$$\begin{aligned}
\varepsilon(f) &= \lambda^{-1} \circ \eta^*(f) = \lambda^{-1}\bigl[\eta^*(f)\bigr] \\
&= \lambda^{-1}[f \circ \eta] = f \circ \eta[1_K] \\
&= f\bigl(\eta[1_K]\bigr) = f(1_A) \in K.
\end{aligned}$$

According to 4.4(b), if A is a *finite-dimensional* vector space and if

$$\varphi : A \otimes A \to A \quad \text{and} \quad \eta : K \to A$$

are K-linear maps, then

$$(A, \varphi, \eta) \text{ is an } \textit{algebra} \quad \Leftrightarrow \quad \bigl(A^*, \rho^{-1} \circ \varphi^*, \lambda^{-1} \circ \eta^*\bigr) \text{ is a } \textit{coalgebra}.$$

In this way, finite-dimensional algebras give rise to finite-dimensional coalgebras. Conversely, coalgebras give rise to algebras; however, in this direction there is no restriction on the dimension of the underlying vector space. For, according to 4.4(a), if (C, Δ, ε) is a coalgebra, then $(C^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda)$ is always an algebra (and conversely). Thus, the vector space dual of a coalgebra always has the structure of an algebra, whereas, in general, only the vector space dual of a *finite-dimensional* algebra has the structure of a coalgebra.

DEFINITION/TERMINOLOGY 2.58. (a) If (C, Δ, ε) is a *coalgebra*, then the *algebra* $(C^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda)$ is called the *dual algebra* of (C, Δ, ε). The *multiplication* $\Delta^* \circ \rho$ of C^* is called the *convolution product*, and one writes $f * g$ in place of $\Delta^* \circ \rho(f \otimes g)$ for $f, g \in C^*$. Then

$$(f * g)(c) = \sum_c f(c_1) \cdot g(c_2)$$

whenever $f, g \in C^*$ and $c \in C$ with $\Delta(c) = \sum_c c_1 \otimes c_2$. The *identity element* of the algebra $(C^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda)$ is the K-linear map $\varepsilon : C \to K$, i.e., $1_{C^*} = \varepsilon_C$.

(b) If (A, φ, η) is a *finite-dimensional algebra*, then the *coalgebra* $(A^*, \rho^{-1} \circ \varphi^*, \lambda^{-1} \circ \eta^*)$ is called the *dual coalgebra* of (A, φ, η). The *comultiplication* Δ_{A^*} of A^* is related to the multiplication "$\cdot_A$" of A by the defining property that, for all $f \in A^*$,

$$\Delta_{A^*}(f) := \sum_f f_1 \otimes f_2 \quad \Leftrightarrow \quad f(x \cdot_A y) = \sum_f f_1(x) \cdot f_2(y) \in K, \quad \text{for all } x, y \in A.$$

The *counit* ε_{A^*} of A^* is defined to be "evaluation at the identity element of A":

$$\varepsilon_{A^*}(f) := f(1_A) \in K.$$

REMARK 2.59. The assignment $C \mapsto C^*$ is functorial. It turns out that the *linear dual functor*

$$(-)^* : (\mathbf{Coalg})^{op} \to \mathbf{Alg}$$

so obtained from the opposite category of the category of coalgebras to the category of algebras has a right adjoint

$$(-)^0 : (\mathbf{Alg})^{op} \to \mathbf{Coalg}$$

called the *upper-zero* or the *continuous linear dual functor*. The latter name is appropriate because if one endows A with the *cofinite ideal topology* (having as neighborhood basis of the origin of A the collection of all *cofinite two-sided ideals* of A), and if one endows K with the *discrete topology*, then the subspace A^0 of A^* consists precisely of those elements of the linear dual space A^* which are continuous with respect to the given topologies, namely,

$$A^0 = \{f \in A^* \mid \operatorname{Ker} f \text{ contains a } \textit{cofinite } 2\textit{-sided ideal} \text{ of } A\}.$$

Observe that $A^0 = A^*$ whenever A is *finite-dimensional*.

NOTE. By definition, a subspace W of a vector space V is *cofinite* in case V/W is *finite-dimensional*, in which case the *codimension* of W, denoted $\operatorname{codim}(W)$, is defined to be the dimension of V/W, denoted $\dim(V/W)$.

REMARK 2.60. Having described the process by which the vector space dual of any finite-dimensional algebra can be endowed with the structure of a coalgebra, we now consider a few specific low-dimensional examples. Along the way we shall introduce some important concepts, for instance the notion of a *group-like element* of a coalgebra and of a *primitive element* of a coalgebra. In what follows, as usual, 0 shall denote the one-element vector space whose only element, its *origin*, is also denoted 0.

REMARK 2.61 (0-*dimensional coalgebras*). There is exactly one coalgebra of dimension 0, with $\Delta : 0 \to 0 \otimes 0$ and $\varepsilon : 0 \to K$ the obvious (unique) K-linear maps.

REMARK 2.62. In the category **Coalg** of associative unitary coalgebras, one can have a 0-dimensional subcoalgebra of a non-zero coalgebra. By contrast, in the category **Alg** of associative unitary algebras, one cannot have a 0-dimensional subalgebra of a non-zero algebra because a subalgebra is required to have an identity element, and if $1_A = 0$ then $A = 0 := \{0\}$. It is natural to ask for the coalgebra counterpart of this. The result is that, in contradistinction to the case of **Alg** where one can have a 0-dimensional quotient algebra A/I of a non-zero algebra A (in this case $1_{A/I} = 0$), in **Coalg** one cannot have a 0-dimensional quotient coalgebra of a non-zero coalgebra because from the commutativity of the diagram

$$\begin{tikzcd} C \arrow[r, "\varepsilon_C"] \arrow[d] & K \arrow[d, equal] \\ 0 \arrow[r] & K \end{tikzcd}$$

one concludes that $\varepsilon := e_C = 0$ and hence that $C = 0$ since for every $x \in C$, $x = \sum_x \varepsilon(x_1) \cdot x_2$. We may summarize these observations by saying that in the category of (associative) unitary coalgebras *one can have* a 0-dimensional sub coalgebra of a non-zero coalgebra, but *one cannot have* a 0-dimensional quotient coalgebra of a non-zero coalgebra. “Dually,” in the category of (associative) unitary algebras *one can have* a 0-dimensional quotient algebra of a non-zero algebra, but *one cannot have* a 0-dimensional subalgebra of a non-zero algebra.

REMARK 2.63. In the course of discussing 1-dimensional coalgebras, next, we shall make use of the fact (established below) that the *comultiplication* of a unitary coalgebra is always *injective*. We shall also encounter what classically is called a *group-like element* of a coalgebra. To avoid burying the above-mentioned fact and definition in an example we take care of those preliminaries first.

LEMMA 2.64. *If (C, Δ, ε) is a (unitary) coalgebra, then $\Delta : C \to C \otimes C$ is always injective. “Dually,” if (A, φ, η) is an (unitary) algebra, then $\varphi : A \otimes A \to A$ is always surjective.*

PROOF. Δ is injective because $\rho \circ (1 \otimes \varepsilon) \circ \Delta = 1_C :=$ the *identity map* on C. “Dually,” φ is *surjective* because $\varphi \circ (1 \otimes \eta) \circ \rho^{-1} = 1_A :=$ the *identity map* on A. [Here, ρ and ρ^{-1} are, respectively, as in Definitions 2.6 and 2.2. Further, whenever $g \circ f = 1_X$, then $f : X \to Y$ is *injective* and $g : Y \to X$ is *surjective*.] $\square$

DEFINITION 2.65. Let (C, Δ, ε) be a coalgebra. A *non-zero* element $g \in C$ is called a *group-like element* of C in case $\Delta(g) = g \otimes g$. The *set of all group-like elements* of C is denoted $G(C)$.

LEMMA 2.66. (a) *Let (C, Δ, ε) be a coalgebra and let $g \in G(C)$, i.e., let g be a group-like element of C. Then $\varepsilon(g) = 1_K :=$ the identity element of K.*

(b) *Let* $(H, \varphi, \eta, \Delta, \varepsilon, S)$ *be a Hopf algebra, and let* $g \in G(H)$. *Then* g *has an inverse in* H *and*

$$g^{-1} = S(g).$$

PROOF. (a) If $g \in G(C)$, then from the commutativity of Figure 2.6(b),

$$\begin{tikzcd} g \otimes g \arrow[d, mapsto, "\varepsilon\otimes 1"'] & g = 1 \cdot g \arrow[l, mapsto, "\Delta"'] \arrow[d, equal] \\ \varepsilon(g) \otimes g \arrow[r, mapsto, "\cong", "\lambda"'] & \varepsilon(g) \cdot g \end{tikzcd}$$

so $[\varepsilon(g) - 1] \cdot g = 0$ and therefore $\varepsilon(g) = 1$ since $g \neq 0$ because g is *group-like*.

(b) Chasing g through the diagram that defines S (in Definition 2.44), we find that

$$\begin{tikzcd} g \otimes g \arrow[dd, "S\otimes 1"'] & g \arrow[l, mapsto, "\Delta"'] \arrow[r, mapsto, "\Delta"] \arrow[d, mapsto, "\varepsilon"] & g \otimes g \arrow[dd, mapsto, "1\otimes S"] \\ & \varepsilon(g) = 1_K \arrow[d, mapsto, "\eta"] & \\ S(g) \otimes g \arrow[r, mapsto, "\varphi"'] & S(g) \cdot g = 1_H = g \cdot S(g) & g \otimes S(g) \arrow[l, mapsto, "\varphi"] \end{tikzcd}$$

whence $S(g) \cdot g = 1_H = g \cdot S(g)$, and therefore g^{-1} exists in H and $g^{-1} = S(g)$. □

REMARKS 2.67. (a) Some authors, for example Abe [Ab, p. 59] or Winter [Win, p. 121, Definition 5.3.13], define

$$G(C) := \big\{ g \in G \mid \Delta(g) = g \otimes g \text{ and } \varepsilon(g) = 1 \big\}.$$

We have chosen to follow Sweedler [Sw-1, pp. 55 and 57] in this matter. Because $\varepsilon : C \to K$ is K-linear, $\varepsilon(0) = 0$; so these definitions are equivalent because if $\varepsilon(g) = 1$, then necessarily $g \neq 0$. Sweedler's formulation, however, has the distinct virtue of underscoring the fact that *group-like elements* of a coalgebra *cannot be zero*; and that fact is critical in the proof that, for any algebra A, $G(A^0) = \mathbf{Alg}(A, K) := \mathrm{Hom}_{\mathbf{Alg}}(A, K) :=$ the set of algebra maps (= homomorphisms) from A to K.

(b) Theoretically, it is possible for an associative unitary coalgebra (C, Δ, ε) to have one or more group-like elements since if $x \in C$ and $\Delta(x) = x \otimes x$, then $(\Delta \otimes 1) \circ \Delta(x) = x \otimes x \otimes x = (1 \otimes \Delta) \circ \Delta(x)$; and, using the notation of 2.6,

$$\begin{aligned} \lambda \circ (\varepsilon \otimes 1) \circ \Delta(x) &= \lambda\big[\varepsilon(x) \otimes x\big] = \lambda(1 \otimes x) \\ &= 1 \cdot x = x = x \cdot 1 = \rho(x \otimes 1) \\ &= \rho\big[x \otimes \varepsilon(x)\big] \\ &= \rho \circ (1 \otimes \varepsilon) \circ \Delta(x). \end{aligned}$$

On the other hand, any given associative unitary coalgebra may or may not have group-like elements. For example (cf. 2.78(b), below), if $C = (K \cdot c) \oplus (K \cdot s)$ with $\Delta : C \to C \otimes C$ and $\varepsilon : C \to K$ defined by setting

$$\begin{aligned}
\Delta(c) &:= c \otimes c - s \otimes s, \\
\Delta(s) &:= s \otimes c + c \otimes s, \\
\varepsilon(c) &:= 1 \quad \text{and} \quad \varepsilon(s) := 0,
\end{aligned}$$

then one can readily verify that (C, Δ, ε) is a coalgebra. Suppose that $x = k_1 \cdot c + k_2 \cdot s \in G(C)$ for some $k_1, k_2 \in K$. Then, by comparing coefficients of the *basis elements* $c \otimes c$, $c \otimes s$, $s \otimes c$, and $s \otimes s$ of $C \otimes C$ in the left-hand side and right-hand side expansions of $k_1 \cdot \Delta(c) + k_2 \cdot \Delta(s) =: \Delta(x) = x \otimes x := (k_1 \cdot c + k_2 \cdot s) \otimes (k_1 \cdot c + k_2 \cdot s)$, one finds that $k_1^2 = k_1$, $k_2^2 = -k_1$, and $k_2 \cdot (k_1 - 1) = 0$. If $K = \mathbb{C}$, one has $k_1 = 1$ and $k_2 = \pm i$; but if $K = \mathbb{R}$, then there are no such k_1 and k_2. A *second example* of a *coalgebra* having *no group-like elements* is furnished by the matrix coalgebra $\mathcal{M}_n^c(K)$ when $n > 1$ ($\mathcal{M}_n^c(K)$ will be defined in 2.82, below). For if $C = \mathcal{M}_n^c(K)$, then, by 2.87, $C = [\mathcal{M}_n(K)]^* =$ the dual coalgebra of the algebra of all n-by-n matrices with entries in K. By 2.58(b), for each $f \in C$, $\Delta(f) = f \otimes f \Leftrightarrow f(x \cdot y) = f(x) \cdot f(y)$ for all $x, y \in \mathcal{M}_n(K)$. Hence,

$$\begin{aligned}
G(C) = \mathbf{Alg}\big[\mathcal{M}_n(K), K\big] &:= \mathrm{Hom}_{\mathbf{Alg}}\big[\mathcal{M}_n(K), K\big] \\
&:= \text{the collection of all } K\text{-algebra morphisms } f : \mathcal{M}_n(K) \to K.
\end{aligned}$$

But, for $n > 1$, there are no algebra homomorphisms $f : \mathcal{M}_n(K) \to K$ because (cf. [Mac-Bir, p. 414, Theorem 10] and 2.84(a), below), the matrix algebra $\mathcal{M}_n(K)$ is *simple*; so, if $f : \mathcal{M}_n(K) \to K$ were a *K-algebra map*, then $\operatorname{Ker} f$ would be a 2-*sided ideal* of $\mathcal{M}_n(K)$, hence either 0 or all of $\mathcal{M}_n(K)$. Now, if $n > 1$, $\operatorname{Ker} f$ cannot be 0 for otherwise f would be *injective* which would lead to the contradiction that

$$1 < n^2 = \dim_K\big[\mathcal{M}_n(K)\big] \leqslant \dim_K(K) = 1.$$

On the other hand, for any integer $n \geqslant 1$, $\operatorname{Ker} f$ cannot be all of $\mathcal{M}_n(K)$ because otherwise one would have the contradiction that $0_K = f(\mathrm{I}_n) = 1_K$ (where $\mathrm{I}_n =$ the n-by-n identity matrix). It follows that $G(C) = \emptyset =$ the empty set.

(c) One can show (cf. [R-1, p. 24, Corollary 1.2.3.(1)]) that if C and D are coalgebras, then

$$G(C \otimes D) = \big\{c \otimes d \mid c \in G(C) \text{ and } d \in G(D)\big\}.$$

NOTATION 2.68. If V is a K-vector space and v is a non-zero element of V, then

$$K \cdot v := \{k \cdot v \mid k \in K\} := \text{the 1-dimensional subspace of } V \text{ generated by } v.$$

EXAMPLE 2.69 (1-*dimensional coalgebras*). The situation in dimension 1 is characterized by the following result.

THEOREM.

(a) *If* (C, Δ, ε) *is a* 1*-dimensional* K*-coalgebra, then* $C = K \cdot g$ *for a unique group-like element* g *of* C.

(b) *If* (A, φ, η) *is a* 1*-dimensional* K*-algebra, then* $A = K \cdot a$ *for a unique idempotent element* a *of* A.

(c) *If* (C, Δ, ε) *is a* 1*-dimensional* K*-coalgebra and if* g *is the unique group-like element of* $C(= K \cdot g)$*, let* $g^* \in C^*$ *be defined by* $g^*(k \cdot g) := k$ *for all* $k \in K$ (*so* g^* *is the basis element of* C^* *dual to* g). *Then* g^* *is the unique idempotent element of* $C^*(= K \cdot g^*)$.

(d) *If* (A, φ, η) *is a* 1*-dimensional* K*-algebra and* a *is the unique idempotent element of* $A(= K \cdot a)$*, let* $a^* \in A^*$ *be defined by* $a^*(k \cdot a) := k$ *for all* $k \in K$ (*so* a^* *is the basis element of* A^* *dual to* a). *Then* a^* *is the unique group-like element of* $A^*(= K \cdot a^*)$.

PROOF. (a) Suppose $C = K \cdot c$ for some (necessarily non-zero) *basis element* c of C. Then $c \otimes c$ is a *basis element* of $C \otimes C$, so

$$\Delta(c) = k \cdot (c \otimes c), \quad \text{for some } k \in K.$$

Because Δ is *injective* (by 2.64, above),

$$c \neq 0 \quad \Rightarrow \quad k \neq 0.$$

Then

$$\Delta(k \cdot c) = k \cdot c \otimes k \cdot c, \quad \text{with } k \cdot c \neq 0.$$

Set $g := k \cdot c$. Then g is a *group-like* element of C and also a *basis* element of C, so $K \cdot g = C$. If g' is also a *group-like* element of C, then

$$g' = \lambda \cdot g \quad \text{for some } \textit{non-zero } \lambda \in K.$$

Then

$$\begin{aligned} \lambda^2 \cdot (g \otimes g) &= (\lambda \cdot g) \otimes (\lambda \cdot g) = g' \otimes g' = \Delta(g') \\ &= \Delta(\lambda \cdot g) = \lambda \cdot \Delta(g) = \lambda \cdot (g \otimes g) \end{aligned}$$

so

$$\left(\lambda^2 - \lambda\right) \cdot (g \otimes g) = 0.$$

But $g \otimes g \neq 0$, so

$$0 = \lambda^2 - \lambda = \lambda \cdot (\lambda - 1)$$

so $\lambda = 0$ or $\lambda = 1$. But $\lambda \neq 0$ because $\lambda \cdot g = g' \in G(C)$ and by definition group-likes are non-zero. Therefore $\lambda = 1$, so $g' = 1 \cdot g = g$, so

$$C = K \cdot g$$

for a *unique group-like* element $g \in C$.

(b) Since A is 1-dimensional and has an identity element, 1_A, it must be that

$$1_A \neq 0.$$

Otherwise, for all $a \in A$, $a = 1_A \cdot a = 0 \cdot a = \varphi(0 \otimes a) = 0$ contrary to A being 1-dimensional. Hence $K \cdot 1_A$ is a 1-dimensional subspace of A and therefore

$$A = K \cdot 1_A.$$

Obviously, 1_A is an *idempotent* element of A. To establish uniqueness, suppose that $A = K \cdot a$ with *a idempotent*. Then $0 \neq a \in A = K \cdot 1_A$, so $a = \lambda \cdot 1_A$ for some non-zero $\lambda \in K$, and therefore

$$1_K \cdot a = a = a^2 = \lambda \cdot 1_A \cdot \lambda \cdot 1_A = \lambda \cdot \lambda \cdot 1_A \cdot 1_A = \lambda \cdot \lambda \cdot 1_A = \lambda \cdot a.$$

Then $(\lambda - 1_K) \cdot a = 0$ with $a \neq 0$ so $\lambda = 1_K$ and therefore $a = 1_A$.

(c) Because C is a 1-dimensional coalgebra, C^* is a 1-dimensional algebra and so, by (b), above, C^* has a *unique idempotent* element. If g and g^* are as in the hypotheses, then, for all $k \in K$,

$$\begin{aligned} g^* * g^*(k \cdot g) &= k \cdot \big[(g^* * g^*)(g)\big] \\ &= k \cdot \big[g^*(g) \cdot g^*(g)\big] \\ &= k \cdot 1 \cdot 1 = k \\ &= g^*(k \cdot g), \end{aligned}$$

and therefore, as an element of C^*,

$$g^* * g^* = g^*$$

so g^* is an *idempotent* element of C^* and hence is the *unique idempotent* element of C^*.

(d) A^* is a 1-dimensional coalgebra and so, by (a), above, has a *unique group-like* element. If a and a^* are as in the hypotheses, then $\{a\}$ is a *basis* for A while $\{a^*\}$ is a *basis* for A^*. Consequently, $\{a \otimes a\}$ is a *basis* for $A \otimes A$ while $\{a^* \otimes a^*\}$ is a *basis* for $A^* \otimes A^*$. Then

$$\rho^{-1} \circ \varphi^*(a^*) =: \Delta(a^*) = k \cdot (a^* \otimes a^*) \quad \text{for some } k \in K,$$

so, upon applying ρ to each side, we find that

$$a^* \circ \varphi = \varphi^*(a^*) = \rho\big[k \cdot (a^* \otimes a^*)\big] = k \cdot \rho(a^* \otimes a^*).$$

Consequently,

$$\begin{aligned}1_K &= a^*(a) = a^*(a \cdot a) = a^* \circ \varphi(a \otimes a)\\ &= k \cdot \rho(a^* \otimes a^*)(a \otimes a)\\ &= k \cdot a^*(a) \cdot a^*(a) = k \cdot 1_K \cdot 1_K = k.\end{aligned}$$

So

$$\Delta(a^*) = a^* \otimes a^*,$$

and therefore a^* is a *group-like* element of A^* and hence is *the unique group-like* element of A^*. □

REMARKS 2.70. (a) In the course of discussing 2-dimensional coalgebras, below, we shall make use of the fundamental fact (established next) that the (*distinct*) *group-like elements* of a coalgebra *are always linearly independent*. We shall also encounter what classically is called a *primitive* element of a coalgebra having a unique *group-like* element. To avoid burying the above fact and definition in an example, we dispense with those preliminaries first. The linear independence of the group-like elements of a coalgebra is a basic result (cf. Lazard [Laz, p. 495, Lemma I.I.8] or Larson [Lar-1, p. 353, Lemma 3.1] or Sweedler [Sw-1, pp. 54–56, Proposition 3.2.1(b)]); and it will be established in Lemma 2.71, below. Thereafter, in Definition 2.72, we'll consider *variations on a primitive theme*.

(b) *Apropos* the fact that the (*distinct*) *group-likes* are *linearly independent*, Kaplansky [Kap, p. 24, Proof of Theorem 15] observes that "More generally, any sum of distinct simple subcoalgebras is direct." – cf. Remarks 2.88 and 3.42, below.

(c) The fact that the (*distinct*) *group-like elements* of a coalgebra are *linearly independent* has as a consequence the fact that if G is any group, and if $K[G]$ denotes the *group algebra* on G, then $G(K[G]) = G$ – cf. Lemma 3.7, below.

NOTE. In Sweedler's notation and terminology, $K[G] = KG$ and is called the *group-like coalgebra* on G. More generally, for any non-empty set S, Sweedler [Sw-1, p. 6, Examples of Coalgebras (1)] lets KS denote the K-vector space having S as its *basis*, and defines K-linear maps

$$\Delta : KS \to KS \otimes KS \quad \text{and} \quad \varepsilon : KS \to K$$

on the *basis*, S, of KS via

$$\Delta(s) = s \otimes s \quad \text{and} \quad \varepsilon(s) = 1, \quad \text{for all } s \in S.$$

He then asserts, *loc. cit.*: "Then $(KS, \Delta, \varepsilon)$ is a coalgebra and it is sometimes called the *group-like coalgebra* on the set S." We shall consider this coalgebra in detail at the beginning of Section 3, below. A proof of the fact that $G[KS] = S$ for any non-empty set S may be found, for example, in [Gr-1, p. 31, Theorem I.3.6.a] and in [Dăs-Năs-Rai, pp. 61–62, Solution to Exercise 1.4.16].

LEMMA 2.71. *Let (C, Δ, ε) be a coalgebra. Then the set*

$$G(C) = \{x \in C \mid x \neq 0 \text{ and } \Delta(x) = x \otimes x\}$$

of group-like elements of C is linearly independent over K.

PROOF. Suppose, by way of contradiction, that $G(C)$ is linearly dependent over K. Then there is a subset $\{x_1, \dots, x_s\} \subseteq G(C)$ of s distinct group-like elements of C which is linearly dependent over K, so there exist $k_1, \dots, k_s \in K$ *not all zero* such that

$$\sum_{i=1}^{s} k_i x_i = 0.$$

In fact, there exist integers i, j with $1 \leqslant i < j \leqslant s$ such that $k_i \neq 0$ *and* $k_j \neq 0$. For otherwise there would exist exactly one integer r with $1 \leqslant r \leqslant s$ for which $k_r \neq 0$. But then

$$0 = \sum_{i=1}^{s} k_i x_i = k_r x_r$$

and so $k_r \neq 0 \Rightarrow x_r = 0$ contrary to the fact that $g \in G(C) \Rightarrow g \neq 0$. Let

$$m := \max\{i \mid 1 \leqslant i \leqslant s \text{ with } k_i \neq 0\}.$$

By the above observations, $m \geqslant 2$, so

$$0 = \sum_{i=1}^{s} k_i x_i = \sum_{i=1}^{m} k_i x_i,$$

and therefore $k_m x_m = \sum_{i=1}^{m-1} (-k_i) \cdot x_i$. Consequently $x_m = \sum_{i=1}^{m-1} (\frac{-k_i}{k_m}) \cdot x_i$. By finding a maximal linearly independent subset of $x_1, \dots, x_{m-1}$ having the same span as $x_1, \dots, x_{m-1}$, and by relabeling the subscripts and the scalars $\frac{-k_i}{k_m}$, we may – without loss of generality – suppose that there are $n+1$ distinct elements $g, g_1, g_2, \dots, g_n$ of $G(C)$ such that

$$g = \sum_{i=1}^{n} \lambda_i \cdot g_i$$

with $\{g_1, \dots, g_n\}$ linearly independent over K and each $\lambda_i \neq 0 \in K$. If we set

$$V := \mathrm{Span}_K\{g_1, \dots, g_n\} := K g_1 \oplus \cdots \oplus K g_n$$

then $\{g_i \otimes g_j\}_{i,j=1}^n$ is a *basis* for $V \otimes V$ over K since $\{g_i\}_{i=1}^n$ is a *basis* for V over K. Then, from the fact that

$$g = \sum_{i=1}^n \lambda_i g_i$$

we conclude that

$$\begin{aligned}\sum_{i=1}^n \lambda_i \cdot (g_i \otimes g_i) &= \sum_{i=1}^n \lambda_i \cdot \Delta(g_i) \\ &= \Delta\left[\sum_{i=1}^n \lambda_i g_i\right] = \Delta(g) \\ &= g \otimes g = \left(\sum_{i=1}^n \lambda_i g_i\right) \otimes \left(\sum_{j=1}^n \lambda_j g_j\right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j (g_i \otimes g_j).\end{aligned}$$

By the linear independence of $\{g_i \otimes g_j\}_{i,j=1}^n$ over K we find that $\lambda_i^2 = \lambda_i$ for all $i \in \{1, \ldots, n\}$ while $\lambda_i \lambda_j = 0$ whenever $i \neq j$. Because each $\lambda_i \lambda_j \neq 0$, this is impossible unless $n = 1$ and $\lambda_1 = 1$, in which case $g = g_1$, contrary to the fact that g is distinct from g_1. This contradiction shows that $G(C)$ is linearly independent over K. □

DEFINITION 2.72. (a) If (C, Δ, ε) is a coalgebra with *group-like* element g, then an element $x \in C$ is called a *g-primitive element of C* if

$$\Delta(x) = g \otimes x + x \otimes g.$$

The *set of all g-primitive elements of C* is denoted $P_g(C)$. In fact, $P_g(C)$ is a subspace as is easy to check.

(b) If (C, Δ, ε) is a coalgebra, and if g and h are *group-like* elements of C, then an element $x \in C$ is called a (g, h)-*primitive* element of C (or, more precisely, a (g, h)-*skew primitive* element of C) in case

$$\Delta(x) = g \otimes x + x \otimes h.$$

The set of all (g, h)-*primitive* elements of C is denoted $P_{g,h}(C)$, i.e.,

$$P_{g,h}(C) := \{x \in C \mid \Delta(x) = g \otimes x + x \otimes h\}.$$

NOTE. $P_g(C) = P_{g,g}(C)$.

(c) If $(H, \varphi, \eta, \Delta, \varepsilon)$ is a bialgebra or if $(H, \varphi, \eta, \Delta, \varepsilon, S)$ is a Hopf algebra, then an element $x \in H$ is called a *primitive* element of H in case

$$\Delta(x) = 1 \otimes x + x \otimes 1$$

where $1 = 1_H =$ the identity element of H. The set of all *primitive* elements of H is denoted $P(H)$.

NOTE. $P(H) := P_1(H) := P_{1,1}(H) := \{x \in H \mid \Delta(x) = 1 \otimes x + x \otimes 1\}$.

REMARKS 2.73. (a) It may happen that a Hopf algebra has many primitive elements and many group-like elements. An example is given by $K[G] \otimes U(L)$ where $K[G]$ is the group-algebra on a group G and $U(L)$ is the universal enveloping algebra of a Lie algebra L. On the other hand, *a finite-dimensional Hopf algebra over a field of characteristic zero can have no non-zero primitive elements* (cf. [Dăs-Năs-Rai, p. 158, Exercise 4.2.16; and p. 174, Solution to Exercise 4.2.16]).

(b) One may generalize Definition 2.72(b) to the case of augmented coalgebras as follows. By definition, an *augmented* (or *supplemented*) *coalgebra* is a coalgebra (C, Δ, ε) together with a coalgebra map $\eta := \eta_C : K \to C$. If $(C, \Delta, \varepsilon, \eta)$ is an augmented coalgebra, then $\varepsilon \circ \eta = 1_K : K \to K$ where $1_K : K \to K$ denotes the identity map on K. It follows that

$$C = \operatorname{Im} \eta \oplus \operatorname{Ker} \varepsilon.$$

Set $1 := 1_C := \eta_C(1_K)$ where $1_K =$ the identity element of K. In this situation, Grünenfelder [Gr-1, p. 27, Definition I.3.1.b] defines the set $P(C)$ of *primitive* elements of C via

$$P(C) := \{p \in C \mid \Delta(p) = p \otimes 1 + 1 \otimes p\}.$$

In the graded context, Milnor and Moore [M-M-2, p. 224, Definition 3.7] give an equivalent definition (cf. [Mi-3, p. 130]).

(c) One can show (cf. [R-1, p. 24, Corollary 1.2.3.(2)]) that if C and D are coalgebras and if $g \in G(C)$ and $h \in G(D)$, then

$$P_{g \otimes h}(C \otimes D) = P_g(C) \otimes Kh + Kg \otimes P_h(D).$$

(d) Definition 2.72(a) follows the convention of [L-R, p. 9, Exercise 1.1.3] and [L-R, p. 9, Definition 1.1.18], though Lambe and Radford call (the) elements of $P_{g,h}(C)$ $g:h$*-skew primitive*. By contrast, Montgomery in [Mon-1, p. 4, Definition 1.3.4(b)] calls an element $c \in C$ g, h*-primitive* if $\Delta(c) = c \otimes g + h \otimes c$ and denotes the set of all g, h*-primitives* by $P_{g,h}(C)$.

NOTE. *Augmented coalgebras* are the counterpart of *augmented algebras* (cf. [M-M-2, p. 214] or [Gr-1, p. 14, Definition I.1.7]) while *augmented algebras* are what Cartan and Eilenberg [C-E, Chapter I, p. 182 ff.] called *supplemented algebras.* (An *augmented* (or *supplemented*) *algebra* is an algebra (A, φ, η) together with an algebra map $\varepsilon := \varepsilon_A : A \to K$.)

LEMMA 2.74. (a) *If x is a (g, h)-primitive element of a coalgebra C, then $\varepsilon(x) = 0$. In particular, if x is a g-primitive element of a coalgebra that has a group-like element g (or if x is a primitive element of an augmented coalgebra, or of a bialgebra, or of a Hopf algebra), then $\varepsilon(x) = 0$.*

(b) *If x is a (g, h)-primitive element of a Hopf algebra H, then*

$$S(x) = -g^{-1}xh^{-1} := -g^{-1} \cdot x \cdot h^{-1}$$

where S is the antipode of H. In particular, if x is a primitive element of H, then

$$S(x) = -x.$$

PROOF. (a) Recall that if g is a *group-like* element of a coalgebra, then $\varepsilon(g) = 1 := 1_K$. Likewise, if $1 := 1_C := \eta_C(1_K)$ for an augmented coalgebra C, or if $1 := 1_H := \eta_H(1_K)$ for a bialgebra or a Hopf algebra H, then $\varepsilon(1) = 1_K$. Then, in the first case, by the commutativity of the left half of Figure 2.6(b) (defining the *counit* of a coalgebra) we have that

$$\begin{tikzcd}
g \otimes x + x \otimes h \arrow[d, mapsto, "\varepsilon\otimes 1"'] & x \arrow[l, mapsto, "\Delta"'] \arrow[d, equal] \\
1_k \otimes x + \varepsilon(x) \otimes h \arrow[r, mapsto, "\cong", "\lambda"'] & 1_k \cdot x + \varepsilon(x) \cdot h
\end{tikzcd}$$

and therefore that $x = \varepsilon(x) \cdot h + x$. Thus $\varepsilon(x) \cdot h = 0$ with $\varepsilon(x) \in K$ and $h \neq 0$ (because $h \in G(C)$), so $\varepsilon(x) = 0$. In the remaining cases, one obtains the equation $x = x + \varepsilon(x) \cdot 1$ with $1 \neq 0$ from which we likewise conclude that $\varepsilon(x) = 0$.

(b) Chasing x through the commutative diagram that defines S (in Definition 2.44), and making use of the fact that $\varepsilon(x) = 0$ (by part (a), above) while g and h are *invertible* with $S(g) = g^{-1}$ (by Lemma 2.66(b), above) we find that

$$\begin{tikzcd}
g \otimes x + x \otimes h \arrow[dd, mapsto, "S\otimes 1"'] & x \arrow[l, mapsto, "\Delta"'] \arrow[d, mapsto, "\varepsilon"] \\
 & 0 \arrow[d, mapsto, "\eta"] \\
S(g) \otimes x + S(x) \otimes h \arrow[r, mapsto, "\varphi"] & S(g) \cdot x + S(x) \cdot h = 0
\end{tikzcd}$$

so that

$$0 = S(g) \cdot x + S(x) \cdot h$$

and therefore $S(x) \cdot h = -S(g) \cdot x$, whence

$$S(x) = -S(g) \cdot x \cdot h^{-1} = -g^{-1} \cdot x \cdot h^{-1}. \qquad \square$$

NOTE. In the case where x is a primitive element of a Hopf algebra H, then by Definition 2.72(c), $\Delta x = 1 \otimes x + x \otimes 1$ so upon chasing elements as above but with g and h replaced by 1, we find that $S(x) = -x$ since $1 = 1_H =$ the identity of the algebra H and since $S(1) = 1$ (because by chasing 1 through the diagram defining S (cf. Definition 2.44) we find (as in the proof of Lemma 2.66(b)) that $S(1) = 1$).

EXAMPLE 2.75 (*of a* 2-*dimensional coalgebra*). The situation in dimension 2, being more involved than in dimensions 0 or 1, is best approached by starting with a (particularly elementary) 2-dimensional algebra and seeing what 2-dimensional coalgebra it gives rise to by duality. If A is the underlying vector space of a 2-dimensional associative unitary algebra (A, φ, η), then

$$A = K \cdot a \oplus K \cdot b$$

for linearly independent elements a and b of A. Without loss of generality, we may take a to be the (unique) 2-sided identity element, 1_A, of A. Then the only option, in specifying the multiplication on A, is the choice of scalars λ, μ in the equation

$$b^2 = \lambda \cdot a + \mu \cdot b.$$

As our first example, we consider the particularly simple situation where

$$\lambda = \mu = 0.$$

In this case, the multiplication table for A may be written as follows:

$\cdot_A$	0	a	b
0	0	0	0
a	0	a	b
b	0	b	0.

This table tells us that

$$a = 1_A = \text{the (unique) 2-sided identity element of } A$$

while

$$b = \text{a nilpotent element of } A \text{ of order 2.}$$

REMARK. One may realize A concretely as the subalgebra of the algebra $\mathcal{M}_2(\mathbb{K})$ of all 2×2 matrices with entries in a field $\mathbb{K}$ generated by the linearly independent elements

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

When $\mathbb{K} = \mathbb{R} =$ the field of real numbers, one may also realize A as the quotient algebra $\mathbb{R}[x]/(x^2)$ of the algebra $\mathbb{R}[x]$ of all real polynomials in the indeterminate x by the ideal (x^2) generated by x^2. (This algebra is sometimes referred to as the *algebra of functions on the vanishing point-pair* – cf. Arnold, Gusein-Zade, and Varchenko [A-GZ-V, pp. 72–73] or as the *algebra* $\mathbb{R}[\varepsilon]$ *of dual numbers*, as defined in Anders Kock [Koc, p. 3 of the Preface] or Armand Borel [Bo, p. 67]. Here $\varepsilon^2 = 0$.)

We now wish to see what coalgebra arises from A by duality. This requires some calculations – to be carried out momentarily. But first we announce the result of those computations thereby revealing how (modulo evident identifications) the table giving the algebra structure on A encodes the structure of the coalgebra to which it gives rise by duality. Such an interpretation rests on the fact, described in Section 1, that we shall have obtained our coalgebra by "undoing an algebra." In the computations that follow, "$\cdot$" shall either denote *scalar multiplication* or the *multiplication*, $\cdot_A$, *of the algebra* A; the context will make clear what meaning is intended.

In the present case, the 2-dimensional algebra

$$A = (K \cdot a) \oplus (K \cdot b)$$

with multiplication table

$\cdot_A$	a	b
a	a	b
b	b	0

gives rise to a 2-dimensional coalgebra

$$C = (K \cdot a) \oplus (K \cdot b)$$

with diagonal $\Delta : C \to C \otimes C$ given by

$$\Delta(a) = a \otimes a$$

and by

$$\Delta(b) = a \otimes b + b \otimes a$$

and then, of course, $\varepsilon(a) = 1$ while $\varepsilon(b) = 0$, so that a is *the unique group-like* element of C while b is a *primitive* element of C. Here the fact that

$$a \cdot a = a$$

while

$$b = a \cdot b \quad \textit{and} \quad b = b \cdot a$$

is reflected in the fact that

$$\Delta(a) = a \otimes a$$

while

$$\Delta(b) = a \otimes b + b \otimes a.$$

Put differently, here, since

$$a = a \cdot a$$

while

$$b = a \cdot b \quad \textit{and also} \quad b = b \cdot a,$$

we *define*

$$\Delta(a) := a \otimes a$$

and

$$\Delta(b) := a \otimes b + b \otimes a.$$

Then the *algebra* A and the *coalgebra* C are *dual* to one another. This is the content of the following:

THEOREM 2.76. *The* 2*-dimensional algebra having a basis* $\{a, b\}$ *consisting of a* 2*-sided identity element* a *together with a nilpotent element* b *of order* 2 *gives rise by duality to the* 2*-dimensional coalgebra having a basis* $\{a^*, b^*\}$ *consisting of a unique group-like element* a^* *and a primitive element* b^*. *Conversely, the* 2*-dimensional coalgebra having a basis* $\{g, p\}$ *consisting of a unique group-like element* g *together with a primitive element* p *gives rise by duality to the* 2*-dimensional algebra having a basis* $\{g^*, p^*\}$ *consisting of a* 2*-sided identity element* g^* *and a nilpotent element* p^* *of order* 2.

The proof is straightforward (and long and tedious if all details are written down). A more general version of the second half of the theorem is proved below in 2.77(c).

REMARK 2.77. (a) Since in the *dual algebra* $(C^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda)$, defined in 2.58(a), $1_{C^*} = \varepsilon : C \to K$ while, by the definition of g^* and ε, $g^* = \varepsilon \in C^*$ (since both agree on the basis $\{g, p\}$ of C), it is immediate from the general theory (i.e., from the fact that

$1_{C^*} = \varepsilon$) that $g^* = 1_{C^*}$. Thus $g^* * g^* = g^*$ and $g^* * p^* = p^* = p^* * g^*$ giving part of the proof for free.

(b) The same sort of computations as those needed for the proof of 2.76 show that if (C, Δ, ε) is the coalgebra having *basis* elements $g, p_1, p_2, \ldots, p_n$ with g the unique group-like element and each p_i primitive (i.e., $\Delta(g) = g \otimes g$, $\Delta(p_i) = g \otimes p_i + p_i \otimes g$, $\varepsilon(g) = 1$, and $\varepsilon(p_i) = 0$), then the dual algebra $(A, \varphi, \eta) = (C^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda)$ has *dual basis* $\{g^*, p_i^*\}_{i=1}^n$ with $g^* = 1_{C^*}$ and $p_i^* * p_j^* = 0$ for all (i, j) with $1 \leqslant i, j \leqslant n$.

(c) More generally, for *any* K-vector space V, no matter what its dimension, (cf., [L-R, p. 14, Exercise 1.2.4] and [Hey-Rad, p. 225, Example 2.7]) if (C, Δ, ε) is the coalgebra with $C = K \cdot g \oplus V$ as a vector space and $\Delta : C \to C \otimes C$ and $\varepsilon : C \to K$ defined, for each (basis element) $v \in V$, by

$$\Delta(g) = g \otimes g, \qquad \Delta(v) = g \otimes v + v \otimes g$$

and

$$\varepsilon(g) = 1 \quad \text{while } \varepsilon(v) = 0,$$

then the dual algebra $(A, \varphi, \eta) = (C^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda)$ has $A = K \cdot \varepsilon \oplus I$ as a vector space with I a two-sided ideal of A satisfying $I^2 = 0$ where by our conventions, as per usual, $0 := \{0\}$.

PROOF. Because $C = (K \cdot g) \oplus V$, the short exact sequence

$$0 \longrightarrow K \cdot g \overset{i}{\hookrightarrow} C \overset{\pi}{\longrightarrow} C/(K \cdot g) \to 0$$

gives rise to the *split* short exact sequence

$$0 \longrightarrow [C/(K \cdot g)]^* \overset{\pi^*}{\longrightarrow} C^* \overset{i^*}{\longrightarrow} (K \cdot g)^* \longrightarrow 0$$

with splitting

$$\begin{tikzcd}
(K \cdot g)^* \arrow[r, "f"] \arrow[dr, "1_{(K\cdot g)^*}"'] & C^* \arrow[d, "i^*"] \\
& (K \cdot g)^*
\end{tikzcd}$$

whence

$$\begin{aligned} A := C^* &= \operatorname{Im}(f) \oplus \operatorname{Ker}(i^*) \\ &:= f\big[(K \cdot g)^*\big] \oplus (K \cdot g)^{\perp}. \end{aligned}$$

Set $I := (K \cdot g)^{\perp}$. Then I has codimension 1 because

$$A/I := C^*/\operatorname{Ker}(i^*) \cong \operatorname{Im}(i^*) = (K \cdot g)^*.$$

Further, I is an ideal of C^* because $I = (K \cdot g)^{\perp}$ and $K \cdot g$ is a subcoalgebra of C. This is a consequence of 4.5(a) but may also easily be proved directly. If $f_1, f_2 \in I = (K \cdot g)^{\perp} := \{\beta \in C^* \mid \beta(K \cdot g) = 0\}$, then (because $\Delta(g) = g \otimes g$)

$$(f_1 * f_2)(g) := f_1(g) \cdot f_2(g) = 0 \cdot 0 = 0;$$

while (because, for all $v \in V$, $\Delta(v) := g \otimes v + v \otimes g$) one has, for all $v \in V$, that

$$\begin{aligned}(f_1 * f_2)(v) &:= f_1(g) \cdot f_2(v) + f_1(v) \cdot f_2(g) \\ &= 0 \cdot f_2(v) + f_1(v) \cdot 0 = 0.\end{aligned}$$

Consequently, because $C = K \cdot g \oplus V$, one has, for all $c \in C$, that $(f_1 * f_2)(c) = 0$, so $f_1 * f_2 \equiv 0 \in C^*$. □

CLAIMS. (1) $(K \cdot g)^* = K \cdot \hat{g}$ *where* $\hat{g} \in (K \cdot g)^*$ *is defined, for all* $k \in K$, *by*

$$\hat{g}(k \cdot g) := k \cdot \hat{g}(g) := k \cdot 1 = k.$$

(2) *A splitting* $f : (K \cdot g)^* \to C^*$ *of the short exact sequence*

$$0 \longrightarrow \big[C/(K \cdot g)\big]^* \xrightarrow{\pi^*} C^* \xrightarrow{i^*} (K \cdot g)^* \longrightarrow 0$$

is defined on the basis $\{\hat{g}\}$ *of* $(K \cdot g)^*$ $(= K \cdot \hat{g})$ *by setting*

$$f(\hat{g}) := g^* : C \to K$$

where, by definition,

$$g^*(g) := 1$$

while

$$g^*(v) = 0 \quad \textit{for each } v \in V.$$

PROOF OF (1). By definition,

$$(K \cdot g)^* := \{\alpha : K \cdot g \to K \mid \alpha \text{ is } K\text{-linear}\}.$$

Consequently, if $\alpha \in (K \cdot g)^*$, then $\alpha : K \cdot g \to K$ is completely determined by its value at g. If, for instance, $\alpha(g) := k_\alpha \in K$, then $\alpha = k_\alpha \cdot \hat{g}$ because, for each $k \in K$,

$$\alpha(k \cdot g) = k \cdot \alpha(g) := k \cdot k_\alpha$$

while

$$k_\alpha \cdot \hat{g}(k \cdot g) = k_\alpha \cdot k = k \cdot k_\alpha.$$ □

PROOF OF (2). By definition, the K-linear maps $\varepsilon : C \to K$ and $g^* : C \to K$ both satisfy $\varepsilon(g) = g^*(g) := 1$ and $\varepsilon(v) = g^*(v) = 0$ for each $v \in V$. Consequently $g^* \equiv \varepsilon : C \to K$. Moreover, $i^*(g^*) := g^* \circ i =: g^* |_{K \cdot g} = \hat{g}$ because, for every $\in K$,

$$g^*(k \cdot g) := k \cdot g^*(g) = k \cdot 1 = k =: \hat{g}(k \cdot g).$$

Hence $f : (K \cdot g)^* \to C^*$ is a splitting of

$$0 \longrightarrow [C/(K \cdot g)]^* \xrightarrow{\pi^*} C^* \xrightarrow{i^*} (K \cdot g)^* = K \cdot \hat{g} \longrightarrow 0$$

because, for each $k \in K$,

$$\begin{aligned} i^* \circ f(k \cdot \hat{g}) &= i^*\big[f(k \cdot \hat{g})\big] \\ &= k \cdot i^*\big[f(\hat{g})\big] \\ &= k \cdot i^*(g^*) = k \cdot \hat{g}. \end{aligned}$$

Consequently,

$$\begin{aligned} A = C^* &= \mathrm{Im}(f) \oplus \mathrm{Ker}(i^*) \\ &= f\big[(K \cdot g)^*\big] \oplus (K \cdot g)^{\perp} \\ &= f\big[K \cdot \hat{g}\big] \oplus I \\ &= K \cdot f(\hat{g}) \oplus I \\ &= K \cdot g^* \oplus I \\ &= K \cdot \varepsilon \oplus I. \end{aligned}$$

□

REMARK 2.78. The coalgebra with basis $\{c, s\}$ and with Δ and ε defined by

$$\begin{aligned} &\Delta(c) = c \otimes c - s \otimes s, \\ &\Delta(s) = s \otimes c + c \otimes s, \\ &\varepsilon(c) = 1, \quad \text{and} \quad \varepsilon(s) = 0 \end{aligned}$$

is what Nichols and Sweedler [N-S, p. 54, Example 1] called the *trigonometric coalgebra* τ. They urge the reader to think of c and s as the trigonometric functions *sine* and *cosine* in which case applying ε corresponds to evaluation at 0. On pp. 50–51 [loc. cit.], Nichols and Sweedler even suggest a context in which such diagonalization formulas arise "from generic addition formulas." Their idea (which we paraphrase) is this: Let $K[x]$ denote the algebra of polynomials in the indeterminate x with coefficients in the field K. One can *define* an *algebra map*

$$\Delta : K[x] \to K[x] \otimes K[x]$$

via

$$\Delta(1) = 1 \otimes 1$$

and

$$\Delta(x) = 1 \otimes x + x \otimes 1.$$

Then, provided one *defines* $x^0 := 1$ and $(\Delta x)^0 := 1 \otimes 1$, one has, for any integer $n \geqslant 0$, that

$$\Delta\big(x^n\big) = (\Delta x)^n;$$

and, for any polynomial $f(x) \in K[x]$, one has that

$$\begin{aligned}\Delta\big[f(x)\big] &= \Delta\left[\sum_{i=0}^{n} a_i x^i\right] = \sum_{i=0}^{n} a_i \Delta\big(x^i\big)\\ &= \sum_{i=0}^{n} a_i (\Delta x)^i\\ &=: f(\Delta x).\end{aligned}$$

Further, provided one defines $(1 \otimes x)^0 := 1 \otimes x^0 := 1 \otimes 1$, one has that

$$\begin{aligned}1 \otimes f(x) &= 1 \otimes \left[\sum_{i=0}^{n} a_i x^i\right] = \sum_{i=0}^{n} \big[1 \otimes a_i x^i\big]\\ &= \sum_{i=0}^{n} a_i \big[1 \otimes x^i\big] = \sum_{i=0}^{n} a_i [1 \otimes x]^i\\ &=: f(1 \otimes x);\end{aligned}$$

and similarly one has that

$$f(x) \otimes 1 = f(x \otimes 1)$$

provided one *defines* $(x \otimes 1)^0 := x^0 \otimes 1 := 1 \otimes 1$. Then, within the appropriate topological framework (see [N-S, pp. 50–51] for details), $\Delta : K[x] \to K[x] \otimes K[x]$ gives rise to the map

$$\widehat{\Delta} : K\big[[x]\big] = \widehat{K[x]} \to \widehat{K[x] \otimes K}[x] = K\big[[1 \otimes x, x \otimes 1]\big]$$

extending Δ and subsequently also denoted Δ. In characteristic 0, one can use the power series expansions for $\sin(x)$ and $\cos(x)$ to obtain the following:

$$\begin{aligned}\Delta(\sin x) &= \sin(\Delta x) = \sin(1 \otimes x + x \otimes 1)\\ &= \sin(1 \otimes x) \cdot \cos(x \otimes 1) + \cos(1 \otimes x) \cdot \sin(x \otimes 1)\\ &= (1 \otimes \sin x) \cdot (\cos x \otimes 1) + (1 \otimes \cos x) \cdot (\sin x \otimes 1)\\ &= \cos x \otimes \sin x + \sin x \otimes \cos x\\ &= \sin x \otimes \cos x + \cos x \otimes \sin x,\end{aligned}$$

and similarly

$$\begin{aligned}\Delta(\cos x) &= \cos(\Delta x) = \cos(1 \otimes x + x \otimes 1)\\ &= \cos(1 \otimes x) \cdot \cos(x \otimes 1) - \sin(1 \otimes x) \cdot \sin(x \otimes 1)\\ &= (1 \otimes \cos x) \cdot (\cos x \otimes 1) - (1 \otimes \sin x) \cdot (\sin x \otimes 1)\\ &= \cos x \otimes \cos x - \sin x \otimes \sin x.\end{aligned}$$

(For a related discussion, especially on generic "addition laws", see [Ste, pp.1–7, Coalgebras and addition laws] and [S-S, Chapter 1].) In like manner,

$$\begin{aligned}\Delta\big(\mathrm{e}^x\big) &= \mathrm{e}^{\Delta x} = \mathrm{e}^{1\otimes x + x\otimes 1}\\ &= \mathrm{e}^{1\otimes x} \cdot \mathrm{e}^{x\otimes 1} = \big(1 \otimes \mathrm{e}^x\big) \cdot \big(\mathrm{e}^x \otimes 1\big)\\ &= \mathrm{e}^x \otimes \mathrm{e}^x,\end{aligned}$$

and, more generally,

$$\Delta\big(\mathrm{e}^{\lambda x}\big) = \mathrm{e}^{\lambda x} \otimes \mathrm{e}^{\lambda x}, \quad \text{for all } \lambda \in K.$$

The computation given for $\Delta(\mathrm{e}^x)$ illustrates how the exponential function, exp, carries *primitive elements* into *group-like elements*. This fact has been observed by Serre [Se-1, p. LA 4.14, Corollary 7.3]; [Se-2, p. 27, Corollary 7.3] in his treatment of the Campbell–Hausdorff formula. (For a related treatment, see [Q, p. 270, Proposition 2.6; p. 271, formula (2.7)].)

NOTE. One may treat exp over $\mathbb{C}$ and $\mathbb{C} \subset \mathcal{M}_2(\mathbb{R})$, defined in 2.79(b), below, as usual!

We look next at the dual of one of the most basic examples of a finite-dimensional algebra. But first we set the notation.

NOTATION AND TERMINOLOGY 2.79. (a) For any integers m and n, δ_{mn} shall denote the *Kronecker delta* defined, as usual, by

$$\delta_{mn} = \begin{cases} 1, & \text{if } m = n,\\ 0, & \text{if } m \neq n.\end{cases}$$

(b) For any positive integer n, $\mathcal{M}_n(K)$ shall denote the underlying *vector space* of the algebra of all *n-by-n matrices* with entries in the *field K*.

(c) Given integers i and j with $1 \leqslant i, j \leqslant n$,

$$E_{ij} := E_{ij}^{(n)} \in \mathcal{M}_n(K)$$

shall denote the *n-by-n matrix* defined by

$$(E_{ij})_{rs} := \delta_{ri} \cdot \delta_{sj}.$$

In other words, E_{ij} has a 1 in the (i, j)th spot and 0's elsewhere. Finally, for i and j as above,

$$E^{ij} \in \left[\mathcal{M}_n(K)\right]^*$$

shall denote the *linear transformation* defined, for any $B \in \mathcal{M}_n(K)$, by $E^{ij}(B) := B_{ij}$. Then, for any integers i, j, r, s with $1 \leqslant i, j, r, s \leqslant n$, one has that

$$\begin{aligned} E^{ij}(E_{rs}) &:= \delta_{ir} \cdot \delta_{js} \\ &=: (E_{rs})_{ij} = \begin{cases} 1, & \text{if } i = r \ \& \ j = s, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Then $\{E_{ij}\}_{i,j=1}^n$ is the *standard basis* for $\mathcal{M}_n(K)$, and $\{E^{ij}\}_{i,j=1}^n$ is the *dual basis* for $[\mathcal{M}_n(K)]^*$.

THEOREM 2.80. (a) *If $(A, \varphi, \eta) := (\mathcal{M}_n(K), \varphi, \eta)$ is the K-algebra of all n-by-n matrices with entries in K, and if*

$$(A^*, \Delta, \varepsilon) := \left(A^*, \rho^{-1} \circ \varphi^*, \lambda^{-1} \circ \eta^*\right)$$

is the dual coalgebra of (A, φ, η), then A^* has a basis

$$\left\{X^{ij} \mid 1 \leqslant i, j \leqslant n\right\}$$

on which Δ and ε are given, respectively, by

$$\Delta\left(X^{ij}\right) := \sum_{k=1}^n X^{ik} \otimes X^{kj}$$

and by

$$\varepsilon\left(X^{ij}\right) := \delta_{ij}.$$

(b) *Conversely, if (C, Δ, ε) is a K-coalgebra with basis*

$$\{e_{ij} \mid 1 \leqslant i, j \leqslant n\}$$

for C on which Δ and ε are given, respectively, by

$$\Delta(e_{ij}) := \sum_{k=1}^n e_{ik} \otimes e_{kj}$$

and by

$$\varepsilon(e_{ij}) = \delta_{ij},$$

then the dual algebra

$$(C^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda)$$

of the coalgebra (C, Δ, ε) *is isomorphic to the algebra of all* n*-by-*n *matrices with entries in* K.

The proof is straightforward.

REMARK 2.81. Using the obvious vector space isomorphism between $\mathcal{M}_n(K)$ and $[\mathcal{M}_n(K)]^*$ we are led, by the previous result, to define the following coalgebra structure on the space $\mathcal{M}_n(K)$ of all n-by-n matrices with entries in K.

DEFINITION 2.82. The n-by-n *matrix coalgebra* is the coalgebra $(\mathcal{M}_n^c(K), \Delta, \varepsilon)$ where $\mathcal{M}_n^c(K)$ and $\mathcal{M}_n(K)$ are identical as vector spaces (i.e., $\mathcal{M}_n^c(K) := \mathcal{M}_n(K)$) and where $\Delta : \mathcal{M}_n^c(K) \to \mathcal{M}_n^c(K) \otimes \mathcal{M}_n^c(K)$ and $\varepsilon : \mathcal{M}_n^c(K) \to K$ are the linear transformations defined, respectively, for all integers i, j with $1 \leqslant i, j \leqslant n$, by

$$\Delta(E_{ij}) := \sum_{k=1}^{n} E_{ik} \otimes E_{kj}$$

and by

$$\varepsilon(E_{ij}) := \delta_{ij}.$$

REMARK 2.83. Nichols and Sweedler [N-S, p. 54, Example 5] call this coalgebra the *comatric coalgebra* and they point out that it is *not* commutative if $n \geqslant 2$. We applaud their use of the "upper c" to denote the coalgebra $\mathcal{M}_n^c(K)$ – in fact they use the trivial variant $M^c(n, k)$ – but we prefer the name *matrix coalgebra* just as in the algebra context we prefer the name *matrix algebra* to the name *matric algebra* popular in the 1930's and 1940's (cf. MacDuffee [Mac D, p. 4], for an amusingly involved discussion of the history of this terminology, and also Albert [Al, pp. 55, 56, 317] and Weyl [Weyl, Chapter III]).

REMARKS 2.84. (a) The alert reader may be troubled, momentarily, by the fleeting spector of a matrix bialgebra – the case where the algebra and coalgebra structures on $\mathcal{M}_n(K)$ are compatible in the sense that Δ and ε are algebra maps. In such a case (i.e., were this possible) one might prefer a terminology that distinguishes the algebra structure from the coalgebra structure even as in another context one might wish to speak, for emphasis and/or ease of recognition, of an associative, coassociative bialgebra (or Hopf algebra) or of a commutative, cocommutative bialgebra or Hopf algebra. Fortunately, we can banish such a pang of discomfort by observing that because $\mathcal{M}_n(K)$ is a *simple algebra* (cf. Birkhoff and Mac Lane [Bir-Mac, p. 414, Theorem 10]), $\mathcal{M}_n(K)$ can *never* have a bialgebra (or a Hopf algebra) structure if $n > 1$. For otherwise, the *counit*, $\varepsilon_{\mathcal{M}_n(K)} : \mathcal{M}_n(K) \to K$, of $\mathcal{M}_n(K)$ would be a K-algebra morphism contrary to the fact, already established in 2.67(b), that there are no K-algebra morphisms $f : \mathcal{M}_n(K) \to K$ when $n > 1$. Of course when $n = 1$, then $\mathcal{M}_n(K) \equiv K$, and (by 2.50(c)) the *identity map* $1_K : K \to K$ is the unique K-algebra morphism $f : K \to K$.

(b) M. Takeuchi [Tak-3, p. 232] calls a bialgebra M over a field K a *matric bialgebra* if it has a set of generators x_{ij}, with $1 \leqslant i, j \leqslant n$, such that

$$\Delta(x_{ij}) = \sum_{k=1}^{n} x_{ik} \otimes x_{kj} \quad \text{and} \quad \varepsilon(x_{ij}) = \delta_{ij}.$$

He then asserts that as a consequence of the *Fundamental Theorem of Coalgebras* [Sw-1, p. 46, Theorem 2.2.1] every finitely generated bialgebra is a matric bialgebra. We shall prove this in 4.16.

REMARKS 2.85. With the terminology just introduced, we can restate Theorem 2.80 as follows:

(a) The dual coalgebra of a matrix algebra is (isomorphic to) a matrix coalgebra.

(b) The dual algebra of a matrix coalgebra is (isomorphic to) a matrix algebra.

REMARK 2.86. It may be shown (cf. [Mi-1, p. 22, Theorem (I.2.4)] or [Dăs-Năs-Rai, p. 22, Proposition 1.3.14]) that if one starts with a finite-dimensional algebra A, then – modulo the natural vector space identification of A with its double dual A^{**} – dualizing twice returns the algebra A to itself:

$$A = \text{f.d. alg.} \mapsto A^* = \text{f.d. coalg.} \mapsto A^{**} = \text{f.d. alg.} \cong A.$$

Similarly, if one starts with a finite-dimensional coalgebra C, then – modulo the natural vector space identification of C with its double dual C^{**} – dualizing twice returns the coalgebra C to itself:

$$C = \text{f.d. coalg.} \mapsto C^* = \text{f.d. alg.} \mapsto C^{**} = \text{f.d. coalg.} \cong C.$$

In consequence, we have the following:

COROLLARY 2.87.

(a) [$A \cong$ *a matrix algebra*] $\Leftrightarrow$ [$A^* \cong$ *a matrix coalgebra*].

(b) [$C \cong$ *a matrix coalgebra*] $\Leftrightarrow$ [$C^* \cong$ *a matrix algebra*].

REMARK 2.88. In view of the above it should not come as a complete surprise that once one introduces the notion of a *simple coalgebra* [This is *a non-zero coalgebra with no subcoalgebras other than* 0 *and itself*.] then one can prove that

(a) a finite-dimensional coalgebra C is simple $\Leftrightarrow$ the dual algebra C^* is simple, and

(b) a finite-dimensional algebra A is simple $\Leftrightarrow$ the dual coalgebra A^* is simple.

Thus, the simplicity of the full matrix algebra $\mathcal{M}_n(K)$ is equivalent to the simplicity of the full matrix coalgebra $\mathcal{M}_n^c(K)$.

NOTE. By 4.12, *any simple associative coalgebra over a field must be finite-dimensional*.

REMARK 2.89. The technique used above shows how one can obtain finite-dimensional coalgebras by dualizing finite-dimensional algebras. Since, however, there is a duality between the category of finite-dimensional algebras and the category of finite-dimensional coalgebras, there is not too much point in expending energy on such a project *except* in so far as it helps to build a collection of coalgebra examples and a resulting intuition about them. With this in mind, and because of its relevance to quantum groups (mentioned in 2.46(e)), we now take a closer look at $\mathcal{M}_n^c(K)$ when $n = 2$.

EXAMPLES 2.90 (*The matrix coalgebra* $\mathcal{M}_2^c(K)$). To avoid the clutter of subscripts in the formulas

$$\Delta(E_{ij}) := \sum_{k=1}^{2} E_{ik} \otimes E_{kj}$$

and

$$\varepsilon(E_{ij}) := \delta_{ij},$$

we write a, b, c, and d, respectively, in place of the *basis* elements

$$E_{11} := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad E_{12} := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad E_{21} := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \text{and}$$
$$E_{22} := \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

of $\mathcal{M}_2^c(K)$. We then obtain the 4-dimensional (non-commutative) coalgebra $\mathcal{M}_2^c(K) := (K \cdot a) \oplus (K \cdot b) \oplus (K \cdot c) \oplus (K \cdot d)$ with basis $\{a, b, c, d\}$ on which Δ and ε are given, respectively, by

$$\begin{aligned} \Delta(a) &:= a \otimes a + b \otimes c, \\ \Delta(b) &:= a \otimes b + b \otimes d, \\ \Delta(c) &:= c \otimes a + d \otimes c, \quad \text{and} \\ \Delta(d) &:= c \otimes b + d \otimes d, \end{aligned}$$

while

$$\varepsilon(a) := \varepsilon(d) := 1 \quad \text{and} \quad \varepsilon(b) := \varepsilon(c) := 0.$$

In his second lecture (on June 22, 1988) at the Université de Montréal, Yuri Manin presented a nice way to remember these formulas. Namely, one first thinks of a, b, c and d as being elements of some non-commutative ring R, for example a ring of operators on a Hilbert space. One next forms the matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ composed of the individual elements a, b, c, d; and one then applies Δ to the matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. However, in doing so, one employs two rather unconventional ways of writing, namely:

(1) One first applies Δ to each entry of the matrix separately in that one *defines*

$$\Delta\begin{pmatrix} a & b \\ c & d \end{pmatrix} := \begin{pmatrix} \Delta a & \Delta b \\ \Delta c & \Delta d \end{pmatrix}.$$

(2) Next, one *defines*

$$\Delta\begin{pmatrix} a & b \\ c & d \end{pmatrix} := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \otimes \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

but in so doing, one again employs a slightly unconventional form of writing in that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \otimes \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is to mean, by definition, that one multiplies the 2-by-2 matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ by itself in the usual way with the exception that where one would ordinarily multiply the entries one does so "in a tensor way" (i.e., one replaces "$\cdot$" by "$\otimes$"). Putting these two rather unconventional ways of writing together, we obtain

$$\begin{aligned} \begin{pmatrix} \Delta a & \Delta b \\ \Delta c & \Delta d \end{pmatrix} &=: \Delta\begin{pmatrix} a & b \\ c & d \end{pmatrix} \\ &:= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \otimes \begin{pmatrix} a & b \\ c & d \end{pmatrix} \\ &:= \begin{pmatrix} a \otimes a + b \otimes c & a \otimes b + b \otimes d \\ c \otimes a + d \otimes c & c \otimes b + d \otimes d \end{pmatrix} \end{aligned}$$

whence

$$\begin{aligned} &\Delta(a) := a \otimes a + b \otimes c, \\ &\Delta(b) := a \otimes b + b \otimes d, \\ &\Delta(c) := c \otimes a + d \otimes c, \end{aligned}$$

and

$$\Delta(d) := c \otimes b + d \otimes d.$$

By the same token, one *defines*

$$\begin{pmatrix} \varepsilon a & \varepsilon b \\ \varepsilon c & \varepsilon d \end{pmatrix} =: \varepsilon\begin{pmatrix} a & b \\ c & d \end{pmatrix} := I_2 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

whence

$$\varepsilon(a) := \varepsilon(d) := 1 \quad \text{and} \quad \varepsilon(b) := \varepsilon(c) := 0.$$

NOTE. The above shorthand has the added virtue of being consistent with the fact that if g is an element of an associative unitary coalgebra (C, Δ, ε) and if $\Delta(g) = g \otimes g$ (i.e., if g is *group-like*), then necessarily $\varepsilon(g) = 1$. In this context, we identify the identity element, $1 = 1_K$, of the ground field with the identity scalar matrix I_2. *N.B.* The reader should be aware of the unorthodox conventions employed here, and not be misled into believing that $\mathcal{M}_2^c(K)$ is a bialgebra, which (by 3.27, 2.87(b) and 2.84(a)) is *not* the case!.

REMARK 2.91. The matrix coalgebra $\mathcal{M}_2^c(K) := (K \cdot a) \oplus (K \cdot b) \oplus (K \cdot c) \oplus (K \cdot d)$ has a nice 3-dimensional non-commutative *quotient*, namely, the *quotient coalgebra* $T := (K \cdot a) \oplus (K \cdot b) \oplus (K \cdot d)$ of *upper-triangular* matrices $\left(\begin{smallmatrix} a & b \\ 0 & d \end{smallmatrix}\right)$ with *basis* $\{a, b, d\}$ on which Δ and ε are given, respectively, by

$$\begin{aligned} &\Delta(a) := a \otimes a, \\ &\Delta(b) := a \otimes b + b \otimes d, \quad \text{and} \\ &\Delta(d) := d \otimes d, \end{aligned}$$

and by

$$\varepsilon(a) := \varepsilon(d) := 1 \quad \text{while } \varepsilon(b) := 0.$$

(Here, for a, b, c, $d \in \mathcal{M}_2^c(K)$, T is a *quotient* of $\mathcal{M}_2^c(K)$ under $a \mapsto a$, $b \mapsto b$, $c \mapsto 0$, and $d \mapsto d$.) The coalgebra T is sometimes defined to be the 3-dimensional coalgebra with *basis* consisting of two *group-like* elements, a and d, and a third *basis element*, b, satisfying

$$\Delta(b) = a \otimes b + b \otimes d \quad \text{while } \varepsilon(b) = 0.$$

But, by thinking of Δ and ε as being given, respectively, by

$$\begin{aligned} \begin{pmatrix} \Delta a & \Delta b \\ 0 & \Delta d \end{pmatrix} &=: \Delta \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \\ &:= \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \otimes \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \\ &:= \begin{pmatrix} a \otimes a & a \otimes b + b \otimes d \\ 0 & d \otimes d \end{pmatrix} \end{aligned}$$

and by

$$\begin{pmatrix} \varepsilon a & \varepsilon b \\ 0 & \varepsilon d \end{pmatrix} =: \varepsilon \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} := I_2 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

we have a delightfully easy way to recall this example! This, apparently, is what was intended by [N-S, p. 54, Example 6] when they wrote: "This coalgebra lives inside of" $\mathcal{M}_2^c(K)$. It must, however, be noted that T is *not* a *subcoalgebra* of $\mathcal{M}_2^c(K)$ because

$\mathcal{M}_2^c(K)$ is a *simple* coalgebra (cf. Corollary 2.87(b) and Remark 2.88, above) and so cannot have any subcoalgebras other than 0 and itself. On the other hand, $K \cdot a$ and $K \cdot d$ are two 1-dimensional subcoalgebras of T. Also,

$$\Delta_T(a) \neq \Delta_{\mathcal{M}_2^c(K)}(a) \quad \text{and} \quad \Delta_T(d) \neq \Delta_{\mathcal{M}_2^c(K)}(d).$$

NOTE. This example furnishes an example for the claim made in Proposition 4.8(d), below (i.e., it provides a counterexample which shows that a certain result is not true in general).

EXAMPLES 2.92 (*The matrix coalgebra $\mathcal{M}_n^c(K)$, revisited*). Having experienced the elegance and power of Manin's way of retrieving (or encoding) the Δ and ε of $(\mathcal{M}_n^c(K), \Delta, \varepsilon)$ when $n = 2$, we now verify that his approach works just as nicely for any integer $n > 2$. To implement this approach, one first adopts the convention that, for any n-by-n matrices X and Y with entries in the non-commutative K-algebra $R := \mathcal{M}_n(K)$, $X \otimes Y$ shall denote the usual matrix product $X \cdot Y$ with the singular exception that "$\cdot$" is to be replaced throughout by "$\otimes$". The logic behind this is that once we identify X with $X \otimes I$ and Y with $I \otimes Y$ via

$$\begin{array}{ccccc}
\mathcal{M}_n(R) & \xrightarrow{\cong} & \mathcal{M}_n(R) \otimes K & \hookrightarrow & \mathcal{M}_n(R) \otimes \mathcal{M}_n(R) \\
\Downarrow & & \Downarrow & & \Downarrow \\
X & \longmapsto & X \otimes 1_K & \longmapsto & X \otimes I
\end{array}$$

and

$$\begin{array}{ccccc}
\mathcal{M}_n(R) & \longrightarrow & K \otimes \mathcal{M}_n(R) & \hookrightarrow & \mathcal{M}_n(R) \otimes \mathcal{M}_n(R) \\
\Downarrow & & \Downarrow & & \Downarrow \\
Y & \longmapsto & 1_K \otimes Y & \longmapsto & I \otimes Y
\end{array}$$

(where I is the identity element of the matrix algebra $\mathcal{M}_n(R)$) so that, by definition

$$(X \otimes I)_{ij} := X_{ij} \otimes 1 := X_{ij} \otimes 1_R$$

while

$$(I \otimes Y)_{ij} := 1 \otimes Y_{ij} := 1_R \otimes Y_{ij},$$

then

$$\begin{aligned}
(X \otimes I) \cdot (I \otimes Y) &:= (X \cdot I) \otimes (I \cdot Y) \\
&:= X \otimes Y
\end{aligned}$$

where the "·" is just the usual product for n-by-n matrices with entries in a ring. Consequently,

$$\begin{aligned}(X \otimes Y)_{ij} &:= \big[(X \otimes I) \cdot (I \otimes Y)\big]_{ij} \\ &:= \sum_{k=1}^{n} (X \otimes I)_{ik} \cdot (I \otimes Y)_{kj} \\ &:= \sum_{k=1}^{n} (X_{ik} \otimes 1) \cdot (1 \otimes Y_{kj}) \\ &:= \sum_{k=1}^{n} X_{ik} \otimes Y_{kj}.\end{aligned}$$

(Of course, the initially defined products $X \otimes I$ and $I \otimes Y$ are just special cases of this more general formula. The reader may find it instructive to write out the details of both the special and general case when $n = 2$ and $X = Y = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$.) With this as background, the usual formulas for the Δ and ε of $(\mathcal{M}_n^c(K), \Delta, \varepsilon)$ are now neat, immediate and easy consequences of the definitions

$$\Delta(E) := E \otimes E \quad \text{and} \quad \varepsilon(E) := I_n$$

and

$$\Delta(E_{ij}) := (\Delta E)_{ij} \quad \text{and} \quad \varepsilon(E_{ij}) := (\varepsilon E)_{ij}$$

where E is *defined* to be the n-by-n matrix whose (i, j)th entry is the *basis element* E_{ij} of $\mathcal{M}_n(K)$. In sum, Δ and ε are retrieved *coordinatewise* once the *basis matrix* E is made *group-like*. Surely one cannot capture the definitions of Δ and ε more elegantly or succinctly than this! We therefore trust that the reader now has an enhanced appreciation for, and understanding of, the formulas

$$\begin{aligned}\Delta(E_{ij}) &:= (\Delta E)_{ij} := (E \otimes E)_{ij} \\ &:= \sum_{k=1}^{n} E_{ik} \otimes E_{kj}\end{aligned}$$

and

$$\varepsilon(E_{ij}) := (\varepsilon E)_{ij} := (I_n)_{ij} = \delta_{ij}$$

that equip $\mathcal{M}_n^c(K) \equiv \mathcal{M}_n(K)$ with the structure of a coalgebra. In a sense, both the algebra and coalgebra structures on $\mathcal{M}_n(K)$ are the same (to the extent that such statement makes any sense at all) in that each structure is rooted in the definition of the product of matrices. This observation provides a clue as to why *these particular* algebra and coalgebra structures on $\mathcal{M}_n(K)$ do *not* equip $\mathcal{M}_n(K)$ with the structure of a bialgebra. (Of course we previously noted that $\mathcal{M}_n(K)$ can never have the structure of a bialgebra while retaining its usual algebra structure.) For a bialgebra, we require *two* structures, one algebra and one coalgebra, plus a compatibility between the two. Here, in essence, we have just one structure! (What we are saying is this: given a finite-dimensional algebra A, the dual vector space A^* has the structure of a finite-dimensional coalgebra. By using the fact that A and A^* are isomorphic as finite-dimensional vector spaces, we can transport the coalgebra structure of A^* back to A. Then A will have both an algebra and a coalgebra structure. But, in general, since the coalgebra structure on A comes essentially from the algebra structure of A, it is too much to expect that these 2 structures will be "compatible" in the precise sense required of a bialgebra. And in the present case where $A = \mathcal{M}_n(K)$ they are *not compatible* because A is *simple*.) And so it is that we are left with the paradox: *Identical is not* the same as *compatible*. It turns out that in the Lie case, an analogous thing occurs: When one tries to "undo" the *cross product* on Euclidean 3-space $\mathbb{E}^3$, one does indeed obtain a *Lie coalgebra* structure on $\mathbb{E}^3$. But because we have obtained it *by duality*, it is *too compatible*, hence incompatible, with the usual Lie algebra structure on $\mathbb{E}^3$ to yield a *Lie bialgebra*. In the case of the (simple) Lie algebra ($\mathbb{E}^3$, $\times$) one can endow $\mathbb{E}^3$ with the structure of a Lie coalgebra so as to equip $\mathbb{E}^3$ with the structure of a Lie bialgebra. (For details, see Michaelis [Mi-9, p. 371, Example 2.14].)

NOTE. Above we are suggesting that *identifying* L with L^* or A with A^* (for finite-dimensional L or A) rarely produces a bialgebra!

REMARK 2.93. What makes the matrix coalgebra $(\mathcal{M}_2^c(K), \Delta, \varepsilon)$ so interesting is that it serves as a *guide* in defining the underlying coalgebra structure of the *quantum matrix bialgebra* $M_q(2)$, defined, for arbitrary $q \in \mathbb{C}$, below. When the complex parameter $q = 1$, the underlying algebra of $M_q(2)$ is just the *commutative polynomial algebra* $\mathbb{C}[a, b, c, d]$ – known as the *coordinate ring* $\mathcal{O}(G)$ *of the semi-group* $G = \mathcal{M}_2(\mathbb{C})$ – while the *comultiplication*

$$\Delta : M_1(2) \to M_1(2) \otimes M_1(2)$$

and *counit*

$$\varepsilon : M_1(2) \to K = \mathbb{C}$$

of $M_1(2)$ are determined (according to the conventions noted above) by the equations

$$\begin{aligned}\begin{pmatrix} \Delta a & \Delta b \\ \Delta c & \Delta d \end{pmatrix} &=: \Delta \begin{pmatrix} a & b \\ c & d \end{pmatrix} \\ &:= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \otimes \begin{pmatrix} a & b \\ c & d \end{pmatrix} \\ &:= \begin{pmatrix} a \otimes a + b \otimes c & a \otimes b + b \otimes d \\ c \otimes a + d \otimes c & c \otimes b + d \otimes d \end{pmatrix}\end{aligned}$$

and

$$\begin{pmatrix} \varepsilon a & \varepsilon b \\ \varepsilon c & \varepsilon d \end{pmatrix} =: \varepsilon \begin{pmatrix} a & b \\ c & d \end{pmatrix} := I_2 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

NOTE. More generally, following K. Schmüdgen [Schmü, p. 3, Example] or S.P. Smith [Smit, p. 134, Example] one can define the *coordinate ring* $\mathcal{O}(G)$ *of any multiplicative submonoid* G *of* $\mathcal{M}_n(\mathbb{C})$ *as follows*: *First, for any integers* i *and* j *such that* $1 \leqslant i, j \leqslant n$, *one defines the* (i, j)*th coordinate function*

$$X_{ij} : G \to \mathbb{C}$$

on G *by setting*

$$X_{ij}(B) := B_{ij}, \quad \text{for all } B := (B_{ij}) \in G.$$

Next, one lets

$$A := \mathcal{O}(G) := \mathbb{C}[X_{ij} \mid 1 \leqslant i, j \leqslant n\}$$

be the *free commutative, associative unitary algebra* generated by the coordinate functions $X_{ij} : G \to \mathbb{C}$ and the function $1 : G \to \mathbb{C}$ (taking the constant value $1 \in \mathbb{C}$). In other words, $A := \mathbb{C}[X_{ij}]$ is the *polynomial algebra* in the *commuting variables* X_{ij}. Matrix multiplication furnishes us with the formula

$$\begin{aligned} X_{ij}(B \cdot C) := (B \cdot C)_{ij} &:= \sum_{k=1}^{n} B_{ik} \cdot C_{kj} \\ &:= \sum_{k=1}^{n} X_{ik}(B) \cdot X_{kj}(C) \in \mathbb{C} \end{aligned}$$

for all $B, C \in G \subseteq \mathcal{M}_n(\mathbb{C})$ which tells us that the (i, j)th coordinate of a product of two matrices is a *polynomial function* of the entries of the factors. We now define *algebra homomorphisms*

$$\varDelta : A \to A \otimes A \quad \text{and} \quad \varepsilon : A \to \mathbb{C}$$

on the generators X_{ij} of $A := \mathbb{C}[X_{ij}]$ via

$$\varDelta(X_{ij}) := \sum_{k=1}^{n} X_{ik} \otimes X_{kj}$$

and

$$\varepsilon(X_{ij}) := \delta_{ij}, \quad \text{for all } i, j \in \{1, \dots, n\}.$$

On the basis of our discussion of the matrix coalgebra $\mathcal{M}_n^c(K)$, we can be sure that Δ and ε defined in this way equip $A := \mathbb{C}[X_{ij}]$ with the structure of an associative unitary coalgebra. Alternatively, the associativity of the multiplication of G gives rise to the function identity

$$(\Delta \otimes 1) \circ \Delta = (1 \otimes \Delta) \circ \Delta : A \to A \otimes A \otimes A$$

even as the identity

$$g \cdot I_n = I_n \cdot g, \quad \text{for all } g \in G$$

(where $I_n :=$ the n-by-n identity matrix) yields the function identity

$$(\varepsilon \otimes 1) \circ \Delta = (1 \otimes \varepsilon) \circ \Delta = 1 : A \to A$$

on $\mathbb{C} \otimes A \cong A \otimes \mathbb{C} \cong A$. The fact that Δ and ε are algebra morphisms says that the coordinate ring $\mathcal{O}(G)$ of the submonoid G of $\mathcal{M}_n(\mathbb{C})$ is a *bialgebra*. It turns out that if G is a *subgroup* of the *special linear group*

$$SL_n(\mathbb{C}) := \{B \in \mathcal{M}_n(\mathbb{C}) \mid \det(B) = 1\}$$

then $\mathcal{O}(G)$ may be endowed with an *antipode* so that $\mathcal{O}(G)$ becomes a *Hopf algebra*.

REMARK 2.94. The next bit of this section is addressed to the categorically minded. Those not so inclined may safely proceed to the next section (in fact, they are *forbidden* to read what follows).

REMARK 2.95. The categorical duality

$$\mathbf{Alg_{f.d.}} \cong \mathbf{Coalg_{f.d.}}$$

between *finite-dimensional algebras* and *finite-dimensional coalgebras* can be extended to categorical dualities

$$\mathbf{Alg} \cong \mathbf{Profinite\ Coalg}$$

and

$$\mathbf{Coalg} \cong \mathbf{Profinite\ Alg}$$

between *algebras* and *profinite coalgebras* and between *coalgebras* and *profinite algebras*. These ideas go back at least to J. Dieudonné [D-1] and P. Gabriel [Ga, pp. 476–530]. Of more recent vintage are the works of Larson [Lar-1], J. Dieudonné [D-2, pp. 7–22], R.G. Heyneman and D.E. Radford [Hey-Rad, pp. 219–222], G.M. Bergman [Ber, pp. 41–43,

§§17–18], M. Takeuchi [Tak-2], [Tak-4, Section 1] and P. Bonneau, M. Flato, M. Gerstenhaber and G. Pinczon [B-F-G-P]. The basic idea is this: The usual *linear dual functor*

$$(-)^* : \mathbf{Vect} \to \mathbf{Vect}$$

on the *category of vector spaces* is *adjoint to itself on the right* (cf. Mac Lane [Mac-3, p. 86] or Michaelis [Mi-8, p. 140, Lemma 3.1]), but it does not furnish a categorical equivalence (though it does when restricted to the full subcategory $\mathbf{Vect_{f.d.}}$ of *finite-dimensional vector spaces*). To obtain a category equivalent to **Vect** one considers $(-)^*$ to take values in the category **Profinite Vect** of *profinite vector spaces* (these are *projective*, i.e., inverse, *limits* of finite-dimensional vector spaces having the discrete topology). Then the functor in the reverse direction furnishing the categorical equivalence is none other than the *continuous linear dual functor*

$$(-)' : \mathbf{Profinite\ Vect} \to \mathbf{Vect}.$$

(If $\mathcal{V}$ is a profinite vector space, then the *continuous linear dual of* $\mathcal{V}$ is the vector space $\mathcal{V}'$ defined by

$$\begin{aligned} \mathcal{V}' &:= \left\{ f : \mathcal{V} \to K \;\middle|\; \begin{array}{l} f \text{ is } \textit{linear} \text{ and } \textit{continuous} \\ \text{where } K \text{ has the } \textit{discrete topology} \end{array} \right\} \\ &:= \{ f \in \mathcal{V}^* \mid \operatorname{Ker} f \text{ is open} \}.) \end{aligned}$$

It turns out that the categorical equivalence between **Vect** and **Profinite Vect** can be extended to a categorical equivalence between **Alg** and **Profinite Coalg** and to one between **Coalg** and **Profinite Alg** (as well as to ones between the Lie counterparts thereof). In this way, theorems about coalgebras can be viewed as theorems about *linearly compact* algebras (cf. Lefschetz [Lef, p. 78, Definition 27.1] or Köthe [Köt, pp. 95–113]), so some (e.g., Bergman [Ber, pp. 42, §18]) are tempted to say: Why bother? For us, the answer is this: The topology under consideration is essentially algebraic yet in many cases messy so it frequently is advantageous to do away with the topology and focus on the coalgebra. At the same time we must acknowledge that at times a sound topological grounding serves as an invaluable guide to asking the pertinent algebraic question. For example, innocently asking whether a certain topology on the universal enveloping algebra UL of a Lie algebra L is Hausdorff translates to the question of whether UL is *proper* (= *residually finite*), a condition that in turn is equivalent to UL having "sufficiently many finite-dimensional representations to separate points" – a question requiring more insight (cf. Appendix 5.1), than its topological precursor and counterpart! Incidentally, in one of the early articles on coalgebras, Larson [Lar-1, p. 351, Proposition 2.5] gave a proof of the *Fundamental Theorem of Coalgebras* (to the effect that each element of an associative unitary coalgebra lies in a finite-dimensional subcoalgebra) within a topological framework. In subsequent treatments, others exorcised the topology. One last point: A *profinite algebra* $(A, \varphi : A \widehat{\otimes} A \to A, \eta : K \to A)$ can be viewed as an ordinary algebra because there always is an embedding of the *usual tensor product* $A \otimes A$ into the *completed tensor prod-*

uct $A \mathbin{\widehat{\otimes}} A$. However, a profinite coalgebra $(C, \Delta : C \to C \mathbin{\widehat{\otimes}} C, \varepsilon : C \to K)$ will not usually give rise to an ordinary coalgebra because the maps go the wrong way:

$$C \to C \mathbin{\widehat{\otimes}} C \hookleftarrow C \otimes C.$$

REMARK 2.96. At the beginning of Sections 1 and 2 we remarked that algebras and coalgebras are defined "dually" to one another and that the diagrammatic definition of an algebra had the virtue that one could obtain the diagrammatic definition of a coalgebra simply by formally "turning the arrows around." Here we point out that the proper formal framework for obtaining the definition of a coalgebra from that of an algebra by "turning the arrows around" is that of the *opposite category* of a category. But a subtlety is involved. In the case of an algebra or a coalgebra, the tensor product, $\otimes$, appearing in $A \otimes A \xrightarrow{\varphi} A$ or in $A \xrightarrow{\Delta} A \otimes A$ is to be considered as an *additional formal piece of data* – not something of internal categorical significance which would then also have to be dualized (much as an *H-space multiplication* $X \times X \to X$ is "dual" to a *co-H-space comultiplication* $X \to X \vee X$). Thus, instead of defining an algebra or a coalgebra as we did above, technically speaking those very definitions can be thought of as defining an *algebra* or a *coalgebra over the monoidal category* $(\mathbf{V}, \otimes, K)$ where

$$\mathbf{V} := \text{the category } \mathbf{Vect} \text{ of vector spaces over the field } K$$

and

$$\otimes : \mathbf{V} \times \mathbf{V} \to \mathbf{V}$$

is the functor assigning to each ordered pair (V, W) of K-vector spaces their tensor product over K.

NOTE 2.97. In [Jon, p. 4], Jonah credits Peter Hilton with bringing to his attention that he should be working over a category with multiplication, nowadays known as a monoidal category.

For the precise definition of a monoidal category, the reader is referred to Mac Lane [Mac-4, pp. 157–158] or to Eilenberg and Kelly [Eil-Kel, pp. 471–475]. It turns out that if $\mathbf{V}^{op}$ denotes the opposite category of $\mathbf{V}$ (i.e., the category whose objects are the same as those of $\mathbf{V}$ but whose morphisms are in 1-1 correspondence with those of $\mathbf{V}$ but with the direction of f and f^{op} reversed so that $f : A \to B$ in $\mathbf{V}$ if and only if $f^{op} : B^{op} := B \to A^{op} := A$ in $\mathbf{V}^{op}$) and if $\otimes^{op} : \mathbf{V}^{op} \times \mathbf{V}^{op} \to \mathbf{V}^{op}$ denotes the functor induced in the obvious way by $\otimes : \mathbf{V} \times \mathbf{V} \to \mathbf{V}$ so that, for $A, B \in \mathrm{Obj}(\mathbf{V})$ and $f, g \in \mathrm{Morph}(\mathbf{V})$,

$$A^{op} \otimes^{op} B^{op} := (A \otimes B)^{op} \quad \text{and} \quad (f^{op}) \otimes^{op} (g^{op}) = (f \otimes g)^{op},$$

then $(\mathbf{V}^{op}, \otimes^{op}, K^{op})$ is likewise a monoidal category. In this context and depending on one's point of view (regarding which notion ought to be considered as primary), a coalgebra over the monoidal category $(\mathbf{V}, \otimes, K)$ is defined to be an algebra over the (opposite) monoidal category $(\mathbf{V}^{op}, \otimes^{op}, K^{op})$ – cf. [Mac-4, p. 167]; and – in like manner – an algebra over the monoidal category $(V, \otimes, K)$ is defined to be a coalgebra over the (opposite) monoidal category $(\mathbf{V}^{op}, \otimes^{op}, K^{op})$. And now the richness of this viewpoint becomes apparent: By replacing the monoidal category $(\mathbf{V}, \otimes, K)$ with an arbitrary monoidal category $(\mathcal{C}, \Box, e)$ one can now define an algebra (respectively, a coalgebra) over $(\mathcal{C}, \Box, e)$. In this way, one encounters the varied algebras and coalgebras mentioned at the end of Section 1. In the remainder of this article, however, unless explicit mention is made to the contrary, by algebra, coalgebra, bialgebra or Hopf algebra, we shall mean those respective entities over the monoidal category $(\mathbf{V}, \otimes, K)$.

REMARK 2.98. In Section 1 and again in items 2.75, 2.76 and 2.92, we gave examples of associative unitary coalgebras obtained by "undoing" or "taking apart" in all possible ways the elements of certain associative unitary algebras. We now discuss the theory behind the construction of those examples as well as behind two examples to be considered in Section 3, namely, the *cut coalgebra* (3.53) and the *divided power coalgebra* (3.34(f) and 3.35).

DEFINITION 2.99. A monoid $(M, \Box, e)$ is said to have the *finite factorization property* or the *finite product property* in case each $m \in M$ can be written as a product

$$m = p \Box q$$

for at most finitely many $(p, q) \in M \times M$.

EXAMPLES 2.100. Any time one has a monoid $(M, \Box, e)$ having the *finite factorization property*, one may (cf. [Kap, p. 4] and [L-R, p. 5, Example 2]) obtain the structure of an *associative unitary K-coalgebra on $K[M]$*, the K-vector space having M as its *basis*, by setting

$$\Delta(m) := \sum_{\substack{(p,q)\in M\times M\\ m=p\Box q}} p \otimes q \quad \text{and} \quad \varepsilon(m) := \delta_{m,e} \cdot 1_K$$

for each $m \in M$ (see also [J-R, p. 21, the example dealing with the *incidence coalgebra of a locally finite category*], [N-S, pp. 55–56, Examples 6, 7], or [S-O, p. 138, Definition 3.2.30]). The fact that $\Delta : K[M] \to K[M] \otimes K[M]$ and $\varepsilon : K[M] \to K$ do indeed equip $K[M]$ with the structure of an *associative unitary K-coalgebra*, called the *monoid coalgebra of M*, follows from the observations that, upon setting $M' := M - \{e\}$, one has

$$\Delta(e) := e \otimes e + \sum_{\substack{(p,q)\in M'\times M'\\ e=p\Box q}} p \otimes q,$$

$$\Delta(m) := e \otimes m + m \otimes e + \sum_{\substack{(p,q)\in M'\times M'\\ m=p\Box q}} p \otimes q,$$

for each $m \in M'$, and

$$\begin{aligned}(1 \otimes \Delta) \circ \Delta(m) &:= \sum_{\substack{(p,u,v) \in M \times M \times M \\ p \,\square\, u \,\square\, v = m}} p \otimes u \otimes v \\ &= \sum_{\substack{(r,s,q) \in M \times M \times M \\ r \,\square\, s \,\square\, q = m}} r \otimes s \otimes q \\ &:= (\Delta \otimes 1) \circ \Delta(m),\end{aligned}$$

for each $m \in M$.

REMARK 2.101. Some authors (e.g., [Kap, p. 4]) alternatively take $K[M] =$ the K-vector space having the set $\{x_m \mid m \in M\}$ of variables x_m, one for each element m of M, as its *basis*, and then define

$$\Delta(m) := \sum_{\substack{(p,q) \in M \times M \\ m = p \,\square\, q}} x_p \otimes x_q \quad \text{and} \quad \varepsilon(x_m) := \delta_{m,e} \cdot 1_K$$

while others (e.g., [J-R, pp. 16–17] and [S-O, pp. 135–136]) alternatively take $K[M]$ to be the K-vector space having the set $\{m^x \mid m \in M\}$ of formal variables m^x, one for each element m of M, as its *basis*, and then define

$$\Delta\big(m^x\big) := \sum_{\substack{(p,q) \in M \times M \\ m = p \,\square\, q}} p^x \otimes q^x \quad \text{and} \quad \varepsilon\big(m^x\big) := \delta_{m,e} \cdot 1_K.$$

By taking $(M, \square, e)$ to be the monoid $(\mathbb{N}, \cdot, 1)$ of *positive integers under multiplication* and using the latter notation one obtains the *Dirichlet coalgebra* mentioned in Section 1 (cf. Joni and Rota (*loc. cit.*) as well as Spiegel and O'Donnell (*loc. cit.*)).

REMARK 2.102. For any *multiplicatively written* monoid $(M, \square, e)$, the K-vector space $K[M]$, having M as its *basis*, has the structure of an *associative unitary algebra* called the *monoid algebra of M*. This shall be considered in greater detail in Section 3. Here, suffice it to say that the binary operation, $\square =: \cdot_M$, of M induces by linear extension a multiplication, $\cdot_{K[M]}$, on $K[M]$ via the stipulation that, for all $(m, n) \in M \times M$,

$$m \cdot_{K[M]} n := m \,\square\, n := m \cdot_M n.$$

Using the notation of [Kap], this condition translates to the condition that

$$x_m \cdot_{K[M]} x_n := x_{m \,\square\, n} =: x_{m \cdot_M n},$$

whereas using the notation of [J-R] or [S-O], this condition translates to the condition that

$$m^x \cdot_{K[M]} n^x := (m \,\square\, n)^x =: (m \cdot_M n)^x.$$

REMARK 2.103. Suppose that $(M, \square, e)$ is a (possibly infinite) monoid satisfying the *finite factorization property*. For each $m \in M$, let $m^* \in (K[M])^*$ be defined by setting

$$m^*(n) = \delta_{m,n} \cdot 1_K$$

for each $n \in M$. Set

$$M^* := \{m^* \mid m \in M\}.$$

If $m^*, n^* \in M^*$, then the product $m^* * n^*$ in the *dual algebra* $(K[M])^*$ of the *monoid coalgebra* $K[M]$ is given by

$$m^* * n^* = (m \square n)^*,$$

as can be verified directly from the definitions. Hence, as monoids,

$$(M^*, *, e^*) \cong (M, \square, e).$$

THEOREM 2.104. *The monoid coalgebra of a finite monoid and the monoid algebra of that same monoid are mutual duals. If M is an infinite monoid satisfying the finite factorization property, then the* dual algebra $(K[M])^*$ *of the* monoid coalgebra $K[M]$ *contains* $K[M^*]$ *as a* subalgebra isomorphic *to the* monoid algebra $K[M]$*, i.e.,*

$$A := K[M] \cong K[M^*] \subseteq \big(K[M]\big)^* =: C^*.$$

REMARK 2.105. For infinite monoids satisfying the *finite factorization property* the containment is strict already at the vector space level (cf. [J-4, pp. 238, 245; p. 247, Theorem 2]): For if V is a vector space over a division ring D, then $\dim_D(V^*) > \dim_D(V)$. [In fact, if $d =$ the *cardinality* of D, and if $b =$ the *dimension* of V, then the $\dim_D(V^*) = d^b$, the *cardinality* of the set of all mappings from a set of *cardinality* b into a set of *cardinality* d.] Furthermore, the vector space dual A^* of an *infinite-dimensional algebra* A does not (as we shall see in 3.44, below) generally have the structure of a coalgebra arising from the algebra structure of A; hence C and A cannot be mutual duals.

REMARK 2.106. Even for a *finite* monoid M, the *monoid coalgebra* and *monoid algebra* structures on $K[M]$ do not generally endow $K[M]$ with the structure of a *bialgebra*. Of course, for $M = \{e\}$, $K[M]$ even has the structure of a *Hopf algebra*. On the other hand, the failure of the *monoid coalgebra* and *monoid algebra* structures on $K[M]$ to endow it with the structure of a *bialgebra* can easily be checked in Example 2.107. For a *finite group* G of order $n > 1$ the *group algebra* and *group coalgebra* structures on $K[G]$ *never* endow $K[G]$ with the structure of a *bialgebra*. The reason is that, for any $g \in G - \{e\}$, one has that

$$\varepsilon\big(g \square g^{-1}\big) := \varepsilon(e) := 1 \neq 0 = 0 \cdot 0 =: \varepsilon(g) \cdot \varepsilon\big(g^{-1}\big).$$

More generally, if $(M, \square, e)$ is a *monoid* having the *finite factorization property*, and if there exist elements $a, b \in M - \{e\} =: M'$ such that $a \square b = e$ (possibly, with $a = b$), then the *monoid algebra* and *monoid coalgebra* structures on $K[M]$ *never* endow $K[M]$ with the structure of a *bialgebra* because

$$\varepsilon(a \square b) := \varepsilon(e) := 1 \neq 0 = 0 \cdot 0 =: \varepsilon(a) \cdot \varepsilon(b).$$

Additionally, were $K[M]$ to be a *bialgebra*, then e, the *identity element* of the *monoid algebra* $K[M]$, would have to be *group-like*. But,

$$\Delta(e) := e \otimes e + \sum_{\substack{(p,q) \in M' \times M' \\ e = p \square q}} p \otimes q,$$

so $\Delta(e) - e \otimes e \neq 0$ because (a, b) is one of the $(p, q) \in M' \times M'$ for which $p \square q = e$. On the other hand, if $(M, \square, e)$ is the *two-element monoid* with underlying set $M = \{e, m\}$ whose *non-identity element* m is *idempotent* (meaning that $m \square m = m$), and if the *characteristic* of K is *equal* to 2 or 3, then the *monoid algebra* and *monoid coalgebra* structures on $K[M]$ *do* equip $K[M]$ with the structure of a *bialgebra* because, in this case, as one may readily check, Δ and ε *are algebras maps*.

EXAMPLES 2.107. With the above as background, we now describe the *cut coalgebra*. Towards this end, recall that given a non-empty set X, one defines $= W_X :=$ the set of all *words* (strings) $x_1 x_2 \cdots x_n$ of *finite length* n in the *alphabet* X, including the *empty word* ( ) of length 0 denoted by the symbol 1. Words are multiplied by *concatenation* (or *juxtaposition*) meaning that one sets

$$x_1 x_2 \cdots x_n \cdot y_1 y_2 \cdots y_m := x_1 x_2 \cdots x_n y_1 y_2 \cdots y_m.$$

One then has, for example, for any $x_1, x_2 \in X$, that

$$x_1 \cdot x_2 := x_1 x_2.$$

Accordingly, (1) the *empty sequence* denoted by 1 is the *two-sided identity element* of W_X, (2) W_X is the *free monoid on the non-empty set* X, and (3) $K[W_X]$, the K-vector space having W_X as its *basis*, is just $K\langle x_\alpha \rangle$, the *algebra of all polynomials in the non-commuting variables* $x_\alpha \in X$. Because the *monoid* W_X clearly has the *finite product property*, one obtains, by the procedure outlined above, the *monoid coalgebra structure on* $K[W_X]$ by setting

$$\begin{aligned} \Delta(x_1 \cdots x_n) := {} & 1 \otimes (x_1 \cdots x_n) + x_1 \otimes (x_2 \cdots x_n) \\ & + \sum_{p=2}^{n-1} (x_1 \cdots x_p) \otimes (x_{p+1} \cdots x_n) \\ & + (x_1 \cdots x_{n-1}) \otimes x_n + (x_1 \cdots x_n) \otimes 1 \end{aligned}$$

and

$$\varepsilon(1) := 1 := 1_K \quad \text{while } \varepsilon(x_1 \cdots x_n) = 0 \ \forall n \geqslant 1.$$

The *monoid coalgebra* $K[W_X]$, because its *comultiplication* is "cut", is sometimes referred to as the *cut coalgebra on* X, other times as the *tensor coalgebra* on $K[X]$, the K-vector space having X as its *basis* (cf. 3.53, below). The only *group-like element of the monoid coalgebra* $K[W_X]$ is 1 while the *space of* 1*-primitive elements* of $K[W_X]$ is just $K[X]$. There is also an algebra structure on $K[W_X]$ coming from the monoid structure on W_X (concatenation). But here, the *monoid algebra* and *monoid coalgebra* structures on $K[W_X]$ are each derived from the monoid structure of W_X and *do not* combine to define a bialgebra structure on $K[W_X]$ (cf. the Note just prior to Remark 3.54, below). However, the *shuffle product* (defined in the penultimate Note of Remark 3.53, below, as well as in the chapter on Lie algebras in this volume) does combine with the *cut coproduct* to equip $K[W_X]$ with the structure of a bialgebra and even a Hopf algebra with a suitably defined antipode.

EXAMPLES 2.108. When $X = \{x\}$, one clearly has that

$$W_X = W_{\{x\}} := \{1, x, x^2, \ldots, x^n, \ldots\} = \{x^n\}_{n=0}^{\infty},$$

so that $K[W_X]$, the *monoid algebra of the free monoid on the singleton set* $\{x\}$, is just $K[x]$, the K-algebra of all polynomials in the variable x with coefficients in K, whereas $K[W_X]$, the *monoid coalgebra* of $W_{\{x\}}$, has $\{x^n\}_{n=0}^{\infty}$ as its *basis* and has its *comultiplication* and its *counit* given, respectively, by setting

$$\Delta(x^n) := \sum_{k=0}^{n} x^k \otimes x^{n-k} \quad \text{and} \quad \varepsilon(x^n) := \delta_{n,0} \cdot 1_K.$$

The resulting coalgebra $(K[W_{\{x\}}], \Delta, \varepsilon)$, the *cut coalgebra on* $\{x\}$, is also variously called the *divided power coalgebra*, the *divided powers coalgebra*, or *the standard reduced incidence coalgebra of the set of non-negative integers under the usual ordering* (cf. [S-O, p. 135, Definition 3.2.23 and Example 3.2.24] as well as 3.35, below). If $(\mathbb{N}, +, 0)$ denotes the *monoid of non-negative integers under addition*, and if we form the *isomorphic monoid* whose elements are written as x_n rather than $n \in \mathbb{N}$, then the construction of 2.101 applied to the *monoid* $(\mathbb{N}, +, 0)$ yields the *monoid coalgebra* $K[\mathbb{N}]$, whose *comultiplication* and *counit* are given, respectively, by setting

$$\Delta(x_n) := \sum_{\substack{(p,q)\in\mathbb{N}\times\mathbb{N} \\ p+q=n}} x_p \otimes x_q := \sum_{k=0}^{n} x_k \otimes x_{n-k}$$

and

$$\varepsilon(x_n) := \delta_{n,0} \cdot 1_K.$$

This (cf. [L-R, pp. 5–6, Example 2], [Dăs-Năs-Rai, pp. 164–165, Example 5], and 3.34(f) and 3.35, below) is the more conventional description of the *divided power coalgebra.*

3. A guide to selected examples

When organizing an excursion into the land of coalgebras, one is faced with a choice: What scenic spots should be included and in what context and with what emphasis?

For us, the scenic spots are provided by examples, and we feel that the most coherent way to present them is within a theoretical context.

One of the most basic examples of a coalgebra is the *group-like coalgebra on a non-empty set* X, defined next.

DEFINITION 3.1. Given a non-empty set X, let $K[X]$, or just KX, denote the K-vector space having X as its *basis* and let

$$\Delta : K[X] \to K[X] \otimes K[X]$$

and

$$\varepsilon : K[X] \to K$$

be the K-vector space maps defined (on the *basis* X of $K[X]$) by setting

$$\Delta(x) := x \otimes x \quad \text{and} \quad \varepsilon(x) := 1$$

for each $x \in X$. Then the *coalgebra* $(K[X], \Delta, \varepsilon)$ so defined is called the *group-like coalgebra on* X.

REMARKS 3.2. (a) The verification that $(K[X], \Delta, \varepsilon)$ is, in fact, a *coalgebra* is completely routine.

(b) The name *group-like coalgebra on* X reflects the fact that $(K[X], \Delta, \varepsilon)$ is the smallest coalgebra containing X in such a way that each element of X is *group-like*. [Indeed, if C is any coalgebra for which $X \subseteq G(C) =$ *the set of group-like elements* of X, then, for each $x \in X$, $K \cdot x$ is a 1-dimensional subcoalgebra of C, and therefore by 4.7(a), $\sum_{x \in X} K \cdot x$ is a subcoalgebra of C.] But by 2.71 the (distinct) group-like elements of a K-coalgebra are linearly independent over K so

$$\mathrm{Span}_K(X) := \sum_{x \in X} K \cdot x = \bigoplus_{x \in X} K \cdot x =: K[X]$$

and therefore $K[X]$ is a subcoalgebra of any coalgebra C which contains X among its group-like elements. Accordingly, $K[X] \subseteq \bigcap_{\gamma \in \Lambda} C_\gamma$ where Λ indexes the collection of all coalgebras C_γ for which $X \subseteq G(C_\gamma)$. But $K[X]$ itself is one such C_γ and by 4.7(c) the intersection of all subcoalgebras of a coalgebra is again a subcoalgebra thereby proving that

$$K[X] = \bigcap_{\gamma \in \Lambda} C_\gamma = \text{the smallest coalgebra } C \text{ for which } X \subseteq G(C).$$

(c) If $\{(C_i, \Delta_i, \varepsilon_i)\}_{i \in I}$ is any collection of coalgebras, then (cf. [Win, p. 174]) the direct sum, $\bigoplus_{i \in I} C_i$, is a coalgebra with $\Delta(x) := \sum_{i \in I} \Delta_i(x_i)$ and $\varepsilon(x) := \sum_{i \in I} \varepsilon_i(x_i)$ for each $x := \sum_{i \in I} x_i \in \bigoplus_{i \in I} C_i$. This is another way to view the coalgebra structure of $K[X]$ $(= \bigoplus_{x \in X} K \cdot x)$, namely, to consider $K[X]$ as the *direct sum*, $\bigoplus_{x \in X} K \cdot x$, of the 1-dimensional coalgebras $K \cdot x$, one for each $x \in X$.

(d) When X is the underlying set of a *multiplicatively written monoid* (respectively, a *multiplicatively written group*), then $K[X]$ is the underlying coalgebra of a *bialgebra* (respectively, a *Hopf algebra*), as we'll show below.

(e) In connection with what is to follow, it is convenient to recall the following definition.

DEFINITION 3.3. The *covariant functor* $F : \mathcal{A} \to \mathcal{B}$ is *left adjoint* to the *covariant functor* $G : \mathcal{B} \to \mathcal{A}$ and the *covariant functor* $G : \mathcal{B} \to \mathcal{A}$ is *right adjoint* to the *covariant functor* $F : \mathcal{A} \to \mathcal{B}$ in case there exists a *natural set bijection*

$$\Phi_{(A,B)} : \mathrm{Hom}_{\mathcal{B}}\big[F(A), B\big] \xrightarrow{\cong} \mathrm{Hom}_{\mathcal{A}}\big[A, G(B)\big]$$

for each object (A, B) of $\mathcal{A}^{op} \times \mathcal{B}$.

REMARKS 3.4. (a) *Naturality* of $\Phi_{(-,-)} : \mathrm{Hom}_{\mathcal{B}}[F(-), -] \xrightarrow{\cong} \mathrm{Hom}_{\mathcal{A}}[-, G(-)]$ means that, for any morphism $f : (A, B) \to (A_1, B_1)$ of $\mathcal{A}^{op} \times \mathcal{B}$, the diagram

$$\begin{tikzcd}
\mathrm{Hom}_{\mathcal{B}}[F(A), B] \arrow[r, "\cong", "\Phi_{(A,B)}"'] \arrow[d, "{\mathrm{Hom}_{\mathcal{B}}[F(f),g]}"'] & \mathrm{Hom}_{\mathcal{A}}[A, G(B)] \arrow[d, "{\mathrm{Hom}_{\mathcal{A}}[f,G(g)]}"] \\
\mathrm{Hom}_{\mathcal{B}}[F(A_1), B_1] \arrow[r, "\cong", "\Phi_{(A_1,B_1)}"'] & \mathrm{Hom}_{\mathcal{A}}[A_1, G(B_1)]
\end{tikzcd}$$

(in **Sets**) commutes where if $\alpha \in \mathrm{Hom}_{\mathcal{B}}[F(A), B]$ and $\beta := \Phi_{(A,B)}(\alpha) \in \mathrm{Hom}_{\mathcal{A}}[A, G(B)]$, then, by definition,

$$\mathrm{Hom}_{\mathcal{B}}\big[F(f), g\big](\alpha) : F(A_1) \to B_1$$

and

$$\mathrm{Hom}_{\mathcal{A}}\big[f, G(g)\big](\beta) : A_1 \to G(B_1)$$

are defined, respectively, by the commutative diagrams

$$\begin{tikzcd} F(A) \arrow[r, "\alpha"] & B \arrow[d, "g"] \\ F(A_1) \arrow[u, "F(f)"] \arrow[r, dashed, "\mathrm{Hom}_{\mathcal{B}}[F(f),g](\alpha)"'] & B_1 \end{tikzcd} \quad \text{and} \quad \begin{tikzcd} A \arrow[r, "\beta"] & G(B) \arrow[d, "G(g)"] \\ A_1 \arrow[u, "f"] \arrow[r, dashed, "\mathrm{Hom}_{\mathcal{A}}[F,G(g)](\beta)"'] & G(B_1) \end{tikzcd}$$

in $\mathcal{B}$ (respectively, in $\mathcal{A}$). In other words,

$$\mathrm{Hom}_{\mathcal{B}}\big[F(f), g\big](\alpha) := g \circ \alpha \circ F(f) : F(A_1) \xrightarrow{F(f)} F(A) \xrightarrow{\alpha} B \xrightarrow{g} B$$

while

$$\mathrm{Hom}_{\mathcal{A}}\big[f, G(g)\big](\beta) := G(g) \circ \beta \circ f : A_1 \xrightarrow{f} A \xrightarrow{\beta} G(B) \xrightarrow{G(g)} G(B_1).$$

Accordingly, the equality

$$\Phi_{(A_1,B_1)} \circ \mathrm{Hom}_{\mathcal{B}}\big[F(f), g\big] = \mathrm{Hom}_{\mathcal{A}}\big[f, G(g)\big] \circ \Phi_{(A,B)},$$

required by the commutativity of the foregoing diagram, entails that, for each $\alpha \in \mathrm{Hom}_{\mathcal{B}}[F(A), B]$,

$$\Phi_{(A_1,B_1)}\big[g \circ \alpha \circ F(A)\big] = G(g) \circ \Phi_{(A,B)}(\alpha) \circ f \tag{$*$}$$

as elements of $\mathrm{Hom}_{\mathcal{A}}[A_1, G(B_1)]$. Similarly, the equality

$$\Phi^{-1}_{(A_1,B_1)} \circ \mathrm{Hom}_{\mathcal{A}}\big[f, G(g)\big] = \mathrm{Hom}_{\mathcal{B}}\big[F(f), g\big] \circ \Phi^{-1}_{(A,B)}$$

resulting from the commutativity of the foregoing diagram, entails that, for each $\beta \in \mathrm{Hom}_{\mathcal{A}}[A, G(B)]$,

$$\Phi^{-1}_{(A_1,B_1)}\big[G(g) \circ \beta \circ f\big] = g \circ \Phi^{-1}_{(A,B)}(\beta) \circ F(f) \tag{$**$}$$

as elements of $\mathrm{Hom}_{\mathcal{B}}[F(A_1), B_1]$.

(b) The identities $(*)$ and $(**)$, just given (above) can be used to show that if

$$\left(\prod_{i \in I} B_i,\ \left\{\pi_i : \prod_{j \in I} B_j \to B_i\right\}_{i \in I}\right)$$

is the *categorical product* of objects B_i of $\mathcal{B}$, then

$$\left(G\left(\prod_{i \in I} B_i\right),\ \left\{G(\pi_i) : G\left(\prod_{j \in I} B_j\right) \to G(B_i)\right\}_{i \in I}\right)$$

is the *categorical product* of the objects $G(B_i)$ of $\mathcal{A}$ whenever the covariant *functor* $G:\mathcal{B}\to\mathcal{A}$ is *right adjoint* to the covariant *functor* $F:\mathcal{A}\to\mathcal{B}$.

RECOLLECTIONS/NOTATIONS 3.5. (a) For any non-empty set X, let $i_X : X \to K[X]$ denote the function sending x to $1_K \cdot x$ $(= x)$, for each $x \in X$. Because any linear transformation is completely determined by its values on a *basis* of its domain, it follows that the K-vector space $K[X]$ is *free* on X and that the ordered pair $(K[X], i_X)$ is a *free K-module* ($=K$-vector space) on X in the sense that, for any given K-vector space V and any function $f : X \to V$, there is a unique K-vector space map $F : K[X] \to V$ for which the diagram

$$\begin{tikzcd} X \arrow[r, "i_X"] \arrow[d, equal] & K[X] \arrow[d, "F"] \\ X \arrow[r, "f"'] & V \end{tikzcd}$$

(in **Sets**) commutes. It follows that, for all K-vector space maps $F, G : K[X] \to V$,

$$F, G \in \mathrm{Hom}_{\mathbf{Vect}}\big[K[X], V\big] \quad \Leftrightarrow \quad F \circ i_X = G \circ i_X \in \mathrm{Hom}_{\mathbf{Sets}}[X, V].$$

In the interest of clarity, we shall usually distinguish between the linear transformation $F : K[X] \to V$ and its *restriction*

$$F|_X := F \circ i_X : X \to V$$

to the *basis*, X, of $K[X]$, though on occasion, by abuse of notation, we might not do so (as in the case of $\Delta : K[X] \to K[X] \otimes K[X]$ and $\varepsilon : K[X] \to K$, above).

(b) If $f : X \to Y$ is any function between non-empty sets X and Y, let $K[f] : K[X] \to K[Y]$ denote the unique K-linear map making the following diagram commute:

$$\begin{tikzcd} X \arrow[r, "i_X"] \arrow[d, "f"'] & K[X] \arrow[d, "{K[f]}"] \\ Y \arrow[r, "i_Y"'] & K[Y]. \end{tikzcd}$$

Then, the assignment

$$X \mapsto K[X] \quad \text{and} \quad f \mapsto K[f]$$

defines a functor $K : \mathbf{Sets} \to \mathbf{Vect}$, from the category of *sets* to the category of *K-vector spaces*, which is *left adjoint* to the forgetful functor $J : \mathbf{Vect} \to \mathbf{Sets}$.

NOTE. In the present context, the commutative diagram

$$\begin{tikzcd} X \arrow[r, "i_X"] \arrow[d, equal] & K[X] \arrow[d, "F"] \\ X \arrow[r, "f"'] & V \end{tikzcd}$$

expressing the *universal mapping property* (= U.M.P.) satisfied by the *free K-module* $(K[X], i_X)$ *on* X should actually be written

$$\begin{tikzcd} X \arrow[r, "i_X"] \arrow[d, equal] & J(K[X]) \arrow[d, "J(F)"] \\ X \arrow[r, "f"'] & J(V) \end{tikzcd}$$

in which case, for any object $(X, V) \in (\mathbf{Sets})^{op} \times \mathbf{Vect}$, the *natural set bijection*

$$\Phi : \mathrm{Hom}_{\mathbf{Vect}}\big[K[X], V\big] \xrightarrow{\cong} \mathrm{Hom}_{\mathbf{Sets}}\big[X, J(V)\big]$$

is defined, with reference to the preceding diagram, by

$$\Phi(F) := J(F) \circ i_X.$$

We shall, however, usually avoid the clutter of including the symbol J (and thus forget to mention the forgetful functor!).

PROPOSITION 3.6. *The functor* $K : \mathbf{Sets} \to \mathbf{Vect}$ (*which, at the object level, assigns, to any set X, the K-vector space,* $K[X]$*, having X as its basis*) *takes values in* **Coalg** *when, at the object level, for any set X,* $K[X] =$ *the group-like coalgebra on X in case* $X \neq \emptyset$ (= *the empty set*) *while* $K[\emptyset] := \{0\} :=$ *the vector space,* $\{0\}$*, whose only element is* 0*, considered trivially as a K-coalgebra.*

PROOF. We need only verify that, for any morphism $f : X \to Y$ of **Sets**, the induced K-vector space map $K[f] : K[X] \to K[Y]$ is a map of coalgebras. If $X = \emptyset =$ the *empty set*, this is trivial since, for any coalgebra C, the *unique* K-linear map, $\{0\} \to C$, from $\{0\}$ to C is always a morphism of **Coalg**. It remains to show that, for any function $f : X \to Y$ between non-empty sets X and Y, the unique linear extension $K[f] : K[X] \to K[Y]$ of $i_Y \circ f : X \xrightarrow{f} Y \to K[Y]$ to $K[X]$ is a map of **Coalg**. But this is so because a linear map is completely determined by what it does on a *basis* of its domain; and, therefore, more generally, any linear map from one coalgebra to another is always a morphism of **Coalg** in case it maps a *basis* of its *domain* consisting of *group-like elements* to a basis of its *codomain* consisting of *group-like elements*. □

PROPOSITION 3.7. *For any set X, the set of group-like elements of X is just the set X, i.e.,*

$$G\big(K[X]\big) = X.$$

LEMMA 3.8. *If $(C, \Delta_C, \varepsilon_C)$ and $(D, \Delta_D, \varepsilon_D)$ are coalgebras, and if $F:(C, \Delta_C, \varepsilon_C) \to (D, \Delta_D, \varepsilon_D)$ is a morphism of* **Coalg**, *then $F[G(C)] \subseteq G(D)$, so $F: C \to D$ induces a function $G(F): G(C) \to G(D)$, the unique map of* **Sets** *for which the diagram*

$$\begin{tikzcd} G(C) \arrow[r, "i_{G(C)}"] \arrow[d, dashed, "G(F)"'] & C \arrow[d, "F"] \\ G(D) \arrow[r, "i_{G(D)}"'] & D \end{tikzcd}$$

(*in which the horizontal maps are the inclusions*) *commutes.*

COROLLARY 3.9. *With reference to the preceding lemma, the assignment $C \mapsto G(C)$ and $F \mapsto G(F)$ defines a (covariant) functor $G:$* **Coalg** $\to$ **Sets**.

COROLLARY 3.10. $G \circ K = 1_{\mathbf{Sets}} : \mathbf{Sets} \to \mathbf{Sets}$.

PROPOSITION 3.11. *For any non-empty set X and any coalgebra (C, Δ, ε), there is a natural bijection*

$$\Phi := \Phi_{X,C} : \mathrm{Hom}_{\mathbf{Coalg}}\big[K[X], C\big] \xrightarrow{\cong} \mathrm{Hom}_{\mathbf{Sets}}\big[X, G(C)\big]$$

given by $\Phi(F) := G(F)$.

PROOF. If $F \in \mathrm{Hom}_{\mathbf{Coalg}}[K[X], C]$ then, by definition,

$$G(F): X = G\big(K[X]\big) \to G(C)$$

is the unique map of **Sets** for which

$$F|_X := F \circ i_X = F \circ i_{G(K[X])} = i_{G(C)} \circ G(F).$$

For the assignment in the reverse direction, observe that if $f : X \to G(C)$ is any map of **Sets**, then the unique linear extension, $F : K[X] \to C$, to $K[X]$, of

$$i_{G(C)} \circ f : X \xrightarrow{f} G(C) \xhookrightarrow{i_{G(C)}} C$$

is a *coalgebra map* because it maps a *basis* of $K[X]$ consisting of *group-like elements* of $K[X]$ (consisting, in fact, of *the* group-like elements, X, of $K[X]$) to a subset of C consisting of group-like elements. To see that the functions

$$\mathrm{Hom}_{\mathbf{Coalg}}\big[K[X], C\big] \underset{\Psi}{\overset{\Phi}{\rightleftarrows}} \mathrm{Hom}_{\mathbf{Sets}}\big[X, G(C)\big]$$

given by

$$\Phi(F) := G(F)$$

and

$$\Psi(F) := \text{the unique extension to } K[X] \text{ of } i_{G(C)} \circ f$$

are *inverse* to one another observe that, for any given $f \in \mathrm{Hom}_{\mathbf{Sets}}[X, G(C)]$, the commutativity of the diagram

$$\begin{tikzcd} X \arrow[r, "i_X"] \arrow[d, "f"'] & K[X] \arrow[d, "F =: \Psi(f)"] \\ G(C) \arrow[r, hook, "i_{G(C)}"'] & C \end{tikzcd}$$

(with $F := \Psi(f) : K[X] \to C$ a map of **Coalg**) implies that of the diagram

$$\begin{tikzcd} X = G(K[X]) \arrow[r, hook, "i_X"] \arrow[d, "G(F)"'] & K[X] \arrow[d, "F"] \\ G(C) \arrow[r, hook, "i_{G(C)}"'] & C \end{tikzcd}$$

so the equality

$$i_{G(C)} \circ G(F) = F \circ i_X = i_{G(C)} \circ f$$

implies that $G(F) = f$, from which it follows that, for each $f \in \mathrm{Hom}_{\mathbf{Sets}}[X, G(C)]$,

$$\begin{aligned} 1_{\mathrm{Hom}_{\mathbf{Sets}}[X,G(C)]}(f) := f = G(F) &=: \Phi(F) \\ &=: \Phi\big[\Psi(f)\big] \\ &= \Phi \circ \Psi(f), \end{aligned}$$

whence

$$\Phi \circ \Psi = 1_{\mathrm{Hom}_{\mathbf{Sets}}[X,G(C)]}.$$

Conversely, for any given $F \in \mathrm{Hom}_{\mathbf{Coalg}}[K[X], C]$, we have the commutative diagram

$$\begin{tikzcd}
X = G(K[X]) \arrow[r, hook, "i_X"] \arrow[d, "\Phi(F):=G(F)"'] & K[X] \arrow[d, "F"] \\
G(C) \arrow[r, hook, "i_{G(C)}"'] & C
\end{tikzcd}$$

from which it is obvious that $F : K[X] \to C$, itself, is the *unique linear extension* to $K[X]$ of $\Phi(F) : X \to G(C)$. Accordingly, $F =: \Psi[\Phi(F)]$; and hence, for any given $F \in \mathrm{Hom}_{\mathbf{Coalg}}[K[X], C]$,

$$\begin{aligned} 1_{\mathrm{Hom}_{\mathbf{Coalg}}[K[X],C]}(F) := F &= \Psi[\Phi(F)] \\ &= \Psi \circ \Phi(F), \end{aligned}$$

whence

$$\Psi \circ \Phi = 1_{\mathrm{Hom}_{\mathbf{Coalg}}[K[X],C]}.$$

Finally, the *naturality* of

$$\Phi : \mathrm{Hom}_{\mathbf{Coalg}}\big[K[X], C\big] \to \mathrm{Hom}_{\mathbf{Sets}}[X, G(C)]$$

follows from the fact that if $(f, g) \in \mathrm{Hom}_{(\mathbf{Sets})^{op} \times \mathbf{Coalg}}[(X_1, C_1), (X_2, C_2)]$ so that

$$f : X_2 \to X_1 \quad \text{in } \mathbf{Sets}$$

while

$$g : C_1 \to C_2 \quad \text{in } \mathbf{Coalg}$$

then the diagram

$$\begin{tikzcd}
\mathrm{Hom}_{\mathbf{Coalg}}\big[K[X_1], C_1\big] \arrow[r, "\Phi_{X_1,C_1}"] \arrow[d, "{\mathrm{Hom}[K[f],g]}"'] & \mathrm{Hom}_{\mathbf{Sets}}[X_1, G(C_1)] \arrow[d, "{\mathrm{Hom}[f,G(g)]}"] \\
\mathrm{Hom}_{\mathbf{Coalg}}[K[X_2], C_2] \arrow[r, "\Phi_{X_2,C_2}"] & \mathrm{Hom}_{\mathbf{Sets}}[X_2, G(C_2)]
\end{tikzcd}$$

commutes because, for any given $\alpha \in \mathrm{Hom}_{\mathbf{Coalg}}[K[X_1], C_1]$,

$$\begin{aligned} \Phi_{X_2,C_2} \circ \mathrm{Hom}\big[K[f], g\big](\alpha) &= \Phi_{X_2,C_2}\big[g \circ \alpha \circ K[f]\big] \\ &= G\big[g \circ \alpha \circ K[f]\big] \\ &= G(g) \circ G(\alpha) \circ G\big(K[f]\big) \end{aligned}$$

$$\begin{aligned}
&= G(g) \circ G(\alpha) \circ f \\
&= \mathrm{Hom}\big[f, G(g)\big]\big[G(\alpha)\big] \\
&= \mathrm{Hom}\big[f, G(g)\big] \circ G(\alpha) \\
&= \mathrm{Hom}\big[f, G(g)\big] \circ \Phi_{X_1, C_1}(\alpha). \qquad \square
\end{aligned}$$

REMARK. Observe that whenever $F \in \mathrm{Hom}_{\mathbf{Coalg}}[K[X], C]$, then the diagram

$$\begin{tikzcd}
\emptyset \neq X = G[K[X]] \arrow[r, hook, "i_X"] \arrow[d, "\Phi(F):=G(F)"'] & K[X] \arrow[d, "F"] \\
G(C) \arrow[r, hook, "i_{G(C)}"'] & C
\end{tikzcd}$$

(in **Sets**) commutes. Hence, if $G(C)$ has *no* elements (i.e., is the *empty set*, $\emptyset$) neither does *either* $\mathrm{Hom}_{\mathbf{Sets}}[X, G(C)]$ or $\mathrm{Hom}_{\mathbf{Coalg}}[K[X], C]$.

COROLLARY 3.12. *The functor* $G : \mathbf{Coalg} \to \mathbf{Sets}$ *is right adjoint to the functor* $K : \mathbf{Sets} \to \mathbf{Coalg}$.

REMARK 3.13. When X is a non-empty set, then, trivial as the verification is that $(K[X], \Delta, \varepsilon)$ is a coalgebra, there is a deep reason for this which also sheds light on the *bialgebra structure* (respectively, the *Hopf algebra structure*) on $(K[X], \Delta, \varepsilon)$ when X is the underlying set of a *multiplicatively written monoid* (respectively, a *multiplicatively written group*); and it is to a discussion of this that we now turn following the introduction of some notation.

NOTATION 3.14. For non-empty sets X and Y, *define*

$$X \otimes Y := \big\{x \otimes y \in K[X] \otimes K[Y] \mid (x, y) \in X \times Y\big\}.$$

PROPOSITION 3.15. *For non-empty sets X and Y, $X \otimes Y$ is a basis for $K[X] \otimes K[Y]$.*

PROOF. This follows (cf. [Spin, p. 232, Proposition 11.6], for example) from the fact that X is a *basis* of $K[X]$ and Y is a *basis* of $K[Y]$. $\square$

NOTATION 3.16. For non-empty sets X and Y, let

$$\phi := \phi_{X,Y} : X \times Y \xrightarrow{\cong} X \otimes Y$$

denote the obvious *set bijection* from the *basis*, $X \times Y$, of $K[X \times Y]$ to the *basis*, $X \otimes Y$, of $K[X] \otimes K[Y]$, and let

$$\Phi := \Phi_{X,Y} : K[X \times Y] \to K[X] \otimes K[Y]$$

denote the *linear extension* to $K[X \times Y]$ of the composite set map

$$X \times Y \xrightarrow{\phi_{X,Y}} X \otimes Y \hookrightarrow K[X] \otimes K[Y],$$

i.e., let

$$\Phi := \Phi_{X,Y} : K[X \times Y] \xrightarrow{\cong} K[X] \otimes K[Y]$$

denote the *unique K-vector space isomorphism* for which the diagram

$$\begin{tikzcd} X \times Y \arrow[r, hook, "i_{X\times Y}"] \arrow[d, "\cong"', "\phi_{X,Y}"] & K[X \times Y] \arrow[d, "\Phi_{X,Y}"] \\ X \otimes Y \arrow[r, hook] & K[X] \otimes K[Y] \end{tikzcd}$$

(in **Sets**) commutes; and let

$$\Psi := \Psi_{X,Y} : K[X] \otimes K[Y] \xrightarrow{\cong} K[X \times Y]$$

denote the *inverse* of $\Phi_{X,Y}$.

It is easy to verify that Φ (and hence Ψ) are *natural transformations*.

PROPOSITION 3.17. *For non-empty sets X and Y, the inverse K-vector space isomorphisms*

$$\Phi := \Phi_{X,Y} : K[X \times Y] \xrightarrow{\cong} K[X] \otimes K[Y]$$

and

$$\Phi^{-1} := \Psi := \Psi_{X,Y} : K[X] \otimes K[Y] \xrightarrow{\cong} K[X \times Y]$$

are isomorphisms of coalgebras where $K[X \times Y]$ is the group-like coalgebra on $X \times Y$ and where $K[X] \otimes K[Y]$ is a coalgebra as the tensor product of the group-like coalgebras $K[X]$ and $K[Y]$.

REMARK 3.18. We now expand on the idea touched on in 3.12 in that we present an alternate but revealing way to think of (or understand) the coalgebra structure on $K[X]$ whenever X is a *non-empty* set. In essence, the coalgebra structure on $K[X]$ results from the two facts (elaborated on and then established below) that, if $d : X \to X \times X$ denotes the *set diagonal* defined, for each $x \in X$, by $d(x) := (x, x)$, while $c : X \to \{1_K\}$ denotes the *set* *counit* defined as the unique function from X to a terminal set (They are all isomorphic!) which, with malice aforethought, we choose to be the singleton set $\{1_K\}$ where $1_K =$ the identity element of the ground field K, then

(1) the ordered triple (X, d, c) is a *comonoid* in the *symmetric monoidal category* $(\mathbf{Sets}, \times, \{1_K\})$, and

(2) the *functor* $K : \mathbf{Sets} \to \mathbf{Vect}$ is *multiplicative* (cf. Section 1 and also below at the end of this section) and hence takes *comonoids* in the *symmetric monoidal category* $(\mathbf{Sets}, \times, \{1_K\})$ to *comonoids* in the *symmetric monoidal category* $(\mathbf{Vect}, \otimes, K)$. For the general definition of

(1) a *symmetric monoidal category* (also known as a *braided tensor category*),

(2) a *comonoid* (respectively, a *monoid*) in a *symmetric monoidal category*, and

(3) a *multiplicative functor between symmetric monoidal categories* (also known, respectively, as a *monoidal functor* or as a *tensor functor*), we refer the reader, for example, to [Mac-4, pp. 157–171, 180], [Man-2, pp. 26–29], [Char-Pres, pp. 138–160], [Kass, pp. 281–293], or [Hof-Mor, pp. 729–737].

NOTE. What [Mac-4] or [Hof-Mor] call a *monoid* (respectively, a *comonoid*) *in a monoidal category*, [Jon, pp. 5–8] calls an *algebra* (respectively, an *associative coalgebra with counit*) *over a category with multiplication* and [Hof, pp. 11–27] calls an *algebra with identity* (respectively, a *cogebra with coidentity*) *over a multiplicative category*.

Here, suffice it to say that what we have previously defined in 2.6 (respectively, in 2.2) as a *coalgebra* (respectively, as an *algebra*) is more properly thought of as a *comonoid* (respectively, as a *monoid*) in the *symmetric monoidal category* $(\mathbf{Vect}, \otimes, K)$, as indicated at the end of Section 2; while, for any non-empty set X, the ordered triple (X, d, c) is a *comonoid* in the *symmetric monoidal category* $(\mathbf{Sets}, x, \{1_K\})$, meaning that the diagrams

$$\begin{tikzcd} X \arrow[rr, "d"] \arrow[d, "d"'] & & X \times X \arrow[d, "1\times d"] \\ X \times X \arrow[r, "d\times 1"'] & (X \times X) \times X \arrow[r, "\cong", "\beta"'] & X \times (X \times X) \end{tikzcd}$$

and

$$\begin{tikzcd} X \times X \arrow[d, "c\times 1"'] & X \arrow[l, "d"'] \arrow[r, "d"] \arrow[d, equal] & X \times X \arrow[d, "1\times c"] \\ \{1_K\} \times X \arrow[r, "\cong", "\lambda"'] & X & X \times \{1_K\} \arrow[l, "\cong"', "\rho"] \end{tikzcd}$$

in **Sets** (in which $(X \times X) \times X \xrightarrow[\beta]{\cong} X \times (X \times X)$, $\{1_K\} \times X \xrightarrow[\lambda]{\cong} X$, and $X \times \{1_K\} \xrightarrow[\rho]{\cong} X$, are the obvious, natural, set bijections) commute.

A moment's reflection reveals the fact that the *diagonal* and the *counit* of the *group-like coalgebra*, $(K[X], \Delta_{K[X]}, \varepsilon_{K[X]})$, on the non-empty set X are nothing other than, respectively, the composite

$$\begin{tikzcd} K[X] \arrow[r, "K[d]"] \arrow[d, equal] & K[X \times X] \arrow[d, "\Phi_{X,X}", "\cong"'] \\ K[X] \arrow[r, "\Delta_{K[X]}"'] & K[X] \otimes K[X] \end{tikzcd}$$

and the map

$$\begin{tikzcd} K[X] \arrow[r, "K[c]"] \arrow[d, equal] & K[\{1_K\}] \arrow[d, equal] \\ K[X] \arrow[r, "\varepsilon_{K[X]}"'] & K \end{tikzcd}$$

induced, respectively, by the *diagonal* and the *counit* of the *comonoid* (X, d, c). Then, since $K : \mathbf{Sets} \to \mathbf{Vect}$ is a (covariant) *functor*, the commutativity of the diagram

$$\begin{tikzcd} X \arrow[rr, "d"] \arrow[d, "d"'] & & X \times X \arrow[d, "1\times d"] \\ X \times X \arrow[r, "d\times 1"'] & (X \times X) \times X \arrow[r, "\cong", "\beta"'] & X \times (X \times X) \end{tikzcd}$$

(in **Sets**) yields the commutativity of the diagram

$$\begin{tikzcd} K[X] \arrow[rr, "K[d]"] \arrow[d, "K[d]"'] & & K[X \times X] \arrow[d, "K[1\times d]"] \\ K[X \times X] \arrow[r, "K[d\times 1]"'] & K[(X \times X) \times X] \arrow[r, "\cong", "K[\beta]"'] & K[X \times (X \times X)] \end{tikzcd}$$

(in **Vect**), while the *naturality* of the K*-vector space isomorphism*

$$\Phi_{X,Y} : K[X \times Y] \xrightarrow{\cong} K[X] \otimes K[Y]$$

yields the commutativity (in **Vect**) of the diagrams

$$\begin{tikzcd} K[X \times X] \arrow[r, "\Phi_{X,X}"] \arrow[d, "K[1_X\times d]"'] & K[X] \otimes K[X] \arrow[d, "\cong"', "K[1_X]\otimes K[d]=1_{K[X]}\otimes K[d]"] \\ K[X \times (X \times X)] \arrow[r, "\Phi_{X,X\otimes X}"'] & K[X] \otimes K[X \times X] \end{tikzcd}$$

and

$$\begin{array}{ccc} K[X \times X] & \xrightarrow{K[d \times 1_X]} & K[(X \times X) \times X] \\ {\scriptstyle \Phi_{X,X}}\downarrow\cong & & \cong\downarrow{\scriptstyle \Phi_{X \times X, X}} \\ K[X] \otimes K[X] & \xrightarrow[K[d] \otimes K[1_X] = K[d] \otimes 1_{K[X]}]{} & K[X \times X] \otimes K[Y]. \end{array}$$

Finally, the diagram

$$\begin{array}{ccccc} K[(X \times X) \times X] & \xrightarrow{K[\beta]} & K[X \times (X \times X)] & \xrightarrow[\cong]{\Phi_{X, X \times X}} & K[X] \otimes K[X \times X] \\ {\scriptstyle \Phi_{X \times X, X}}\downarrow\cong & & & & \cong\downarrow{\scriptstyle 1_{K[X]} \otimes \Phi_{X,X}} \\ K[X \times X] \otimes K[X] & \xrightarrow[\Phi_{X,X} \otimes 1_{K[X]}]{\cong} & (K[X] \otimes K[X]) \otimes K[X] & \xrightarrow{\cong} & K[X] \otimes (K[X] \otimes K[X]) \end{array}$$

(in **Vect**) commutes because the linear composites giving the two ways of moving along the edges of this diagram from the *northwest corner* to the *southeast* corner agree on each *basis element*, $((x, y), z)$, of $(X \times X) \times X$ as is verified by the following diagram chase:

$$\begin{array}{ccccc} ((x, y), z) & \xmapsto{K[\beta]} & (x, (y, z)) & \xmapsto{\Phi_{X, X \times X}} & x \otimes (y, z) \\ {\scriptstyle \Phi_{X \times X, X}}\downarrow & & & & \downarrow{\scriptstyle 1_{K[X]} \otimes \Phi_{X,X}} \\ (x, y) \otimes z & \xmapsto[\Phi_{X,X} \otimes 1_{K[X]}]{} & (x \otimes y) \otimes z & \longmapsto & x \otimes (y \otimes z). \end{array}$$

Because $\Delta_{K[X]} \equiv \Phi_{X,X} \circ K[d]$, we find, upon putting the above four diagrams together in the obvious way, that the diagram

$$\begin{array}{ccccc} K[X] & & \xrightarrow{\Delta_{K[X]}} & & K[X] \otimes K[X] \\ {\scriptstyle \Delta_{K[X]}}\downarrow & & & & \downarrow{\scriptstyle 1_{K[X]} \otimes \Delta_{K[X]}} \\ K[X] \otimes K[X] & \xrightarrow[\Delta_{K[X]} \otimes 1_{K[X]}]{} & (K[X] \otimes K[X]) \otimes K[X] & \xrightarrow{\cong} & K[X] \otimes (K[X] \otimes K[X]) \end{array}$$

(in **Vect**) commutes. Similarly, the obvious commutativity of the diagram

$$\begin{array}{ccc} X & \xrightarrow{d} & X \times X \\ \Vert & & \downarrow{\scriptstyle 1 \times c} \\ X & \xleftarrow[\rho_X]{} & X \times \{1_K\} \end{array}$$

(in **Sets**), together with the naturality of

$$\Phi_{X,Y} : K[X \times Y] \xrightarrow{\cong} K[X] \otimes K[Y]$$

and the fact that $K:\mathbf{Sets}\to\mathbf{Vect}$ is a *functor*, guarantees the commutativity of the diagram

$$\begin{tikzcd}
K[X] \arrow[r, "K[d]"] \arrow[d, equal] & K[X\times X] \arrow[rr, "\Phi_{X,X}", "\cong"'] \arrow[d, "K[1_X\times c]"] & & K[X]\otimes K[Y] \arrow[d, "K[1_X]\otimes K[c]=1_{K[X]}\otimes K[c]"] \\
K[X] \arrow[r, "K[\rho_X]"'] & K[X\times\{1_K\}] & & K[X]\otimes K[\{1_K\}]=K[X]\otimes K, \arrow[ll, "\cong"', "\Phi^{-1}_{X,\{1_K\}}"]
\end{tikzcd}$$

hence of the diagram

$$\begin{tikzcd}
K[X] \arrow[r, "\Delta_{K[X]}"] \arrow[d, equal] & K[X]\otimes K[X] \arrow[d, "1_{K[X]}\otimes\varepsilon_{K[X]}"] \\
K[X] & K[X]\otimes K \arrow[l, "\rho_{K[X]}"]
\end{tikzcd}$$

(in **Vect**). In like manner, the commutativity of the diagram

$$\begin{tikzcd}
K[X]\otimes K[X] \arrow[d, "\varepsilon_{K[X]}\otimes 1_{K[X]}"'] & K[X] \arrow[l, "\Delta_{K[X]}"'] \arrow[d, equal] \\
K\otimes K[X] \arrow[r, "\lambda_{K[X]}"'] & K[X]
\end{tikzcd}$$

(in **Vect**) is due, at the most basic level, to the obvious commutativity of the diagram

$$\begin{tikzcd}
X\times X \arrow[d, "c\times 1"'] & K[X] \arrow[l, "d"'] \arrow[d, equal] \\
\{1_K\}\times X \arrow[r, "\lambda_X"'] & X
\end{tikzcd}$$

(in **Sets**).

REMARK 3.19. When the non-empty set X carries *additionally* the structure of a *multiplicatively written monoid* (respectively, a *multiplicatively written group*), then the *multiplicativity* of the functor $K:\mathbf{Sets}\to\mathbf{Vect}$ guarantees that the *group-like coalgebra* on X carries *additionally* the structure of a *bialgebra* (respectively, a *Hopf algebra*) as we shall show below. First, however, it will be convenient to establish the following (two) preliminary results.

Proposition 3.20. *If* (M, φ_M, e) *is a* multiplicatively written monoid *with* binary operation $\varphi_M : M \times M \to M$ *and* two-sided identity element $e \in M$*, let* $u = u_M : \{1_K\} \to M$ *be defined by* $u(1_K) := e$*, i.e., by the commutative diagram*

$$\begin{tikzcd} \{1_K\} \arrow[r, "u"] \arrow[d, "\cong"'] & M \arrow[d, equal] \\ \{e\} \arrow[r, hook, "incl."'] & M. \end{tikzcd}$$

Then the following results obtain:

(1) (M, φ_M, u_M) *is a* monoid *in the* symmetric monoidal category (**Sets**, $\times$, $\{1_K\}$) *and*

(2) *the* K*-vector space* $K[M]$*, having* M *as its basis, has the structure of an* algebra *whose* multiplication*,* $\varphi_{K[M]}$*, is the composite*

$$\begin{tikzcd} K[M] \otimes K[M] \arrow[r, "\Psi_{M,M} := \Phi^{-1}_{M,M} =: \Phi^{-1}"] \arrow[d, equal] & K[M \times M] \arrow[d, "K[\varphi_M]"] \\ K[M] \otimes K[M] \arrow[r, "\varphi_{K[M]}"'] & K[M] \end{tikzcd}$$

(*where, for all* $(x, y) \in M \times M$*,* $\Phi(x, y) := x \otimes y$) *and whose* unit*,* $\eta_{K[M]}$*, is the map*

$$\begin{tikzcd} K \arrow[r, "\eta_{K[M]}"] \arrow[d, equal] & K[M] \arrow[d, equal] \\ K[\{1_K\}] \arrow[r, "K[u_M]"'] & K[M] \end{tikzcd}$$

i.e., $(K[M], K[\varphi_M] \circ \Psi_{M,M}, K[u_M])$ *is a* comonoid *in the* symmetric monoidal category (**Vect**, $\otimes$, K) *because*

(a) (M, φ_M, u_M) *is a* comonoid *in the* symmetric monoidal category (**Sets**, $\times$, $\{1_K\}$) *and*

(b) *the* functor $K :$ **Sets** $\to$ **Vect** *is* multiplicative.

Proof. (1) The conclusion simply asserts the trivial to verify commutativity of the diagrams

$$\begin{tikzcd} M \times (M \times M) \arrow[r, "1 \times \varphi_M"] \arrow[d, "\alpha"', "\cong"] & M \times M \arrow[dd, "\varphi_M"] \\ (M \times M) \times M \arrow[d, "\varphi_M \times 1"'] & \\ M \times M \arrow[r, "\varphi_M"'] & M \times M \end{tikzcd}$$

and

$$\begin{tikzcd}
\{1_K\} \times M \arrow[d, "u\times 1"'] & M \arrow[l, "\cong"'] \arrow[r, "\cong"] \arrow[d, equal] & M \times \{1_K\} \arrow[d, "1\times u"] \\
M \times M \arrow[r, "\varphi_M"'] & M & M \times M \arrow[l, "\varphi_M"]
\end{tikzcd}$$

(2) By turning around the arrows in the diagrams appearing in the proof of 3.18 (after replacing X by M) and then relabeling the arrows appropriately, one finds that the associativity of $\varphi_{K[M]}$, expressed diagrammatically, follows readily from the associativity of φ_M, expressed diagrammatically, even as, "dually," as shown in the proof of 3.18, the associativity of $\Delta_{K[X]}$, expressed diagrammatically, follows readily from the associativity of $d : X \to X \times X$, expressed diagrammatically. Similarly, the fact that $\eta_{K[M]}$ is a two-sided unit for $(K[M], \varphi_{K[M]})$, expressed diagrammatically, follows readily from the fact that $u : \{1_K\} \xrightarrow{\cong} \{e\} \hookrightarrow M$ is a two-sided unit for (M, φ_M), expressed diagrammatically, even as "dually," as shown in the proof of 3.18, the fact that $\varepsilon_{K[X]}$ is a two-sided counit for $(K[X], \Delta_{K[X]})$ follows readily from the fact that $c : X \to \{1_K\}$ is a two-sided counit for (X, d).

NOTE. With regard to the proof of (2), observe, alternatively, that, since $M \otimes M \otimes M := \{x \otimes y \otimes z \in K[M] \otimes K[M] \otimes K[M] \mid (x, y, z) \in M \times M \times M\}$ is a *basis* of $K[M] \otimes K[M] \otimes K[M]$ (since M is a basis of $K[M]$), and since $\varphi_{K[M]} : K[M] \otimes K[M] \to K[M]$ is K-linear, the associativity of $\varphi_{K[M]}$ follows at once from the associativity of φ_M because, for all $x, y \in M$, one finds, by chasing the element $x \otimes y \in M \otimes M$ around the commutative diagram

$$\begin{tikzcd}
M \otimes M \arrow[r, hook] \arrow[d, "\cong"'] & K[M] \otimes K[M] \arrow[d, "\Psi_{M,M}"] \\
M \times M \arrow[r, hook, "i_{M\times M}"'] \arrow[d, "\varphi_M"'] & K[M \times M] \arrow[d, "K[\varphi_M]"] \\
M \arrow[r, hook, "i_M"'] & K[M]
\end{tikzcd}$$

that

$$\varphi_{K[M]}(x \otimes y) = x \cdot_M y := \varphi_M(x, y)$$

and, therefore, on a typical basis element, $x \otimes y \otimes z$, of $K[M] \otimes K[M] \otimes K[M]$, the composites $\varphi_{K[M]} \circ (\varphi_{K[M]} \otimes 1)$ and $\varphi_{K[M]} \circ \big(1 \otimes \varphi_{K[M]}\big)$ agree since they yield, respectively, $(x \cdot y) \cdot z$ and $x \cdot (y \cdot z)$, and these are equal since φ_M is associative.

Further, since $1_{K[M]} := \eta_{K[M]}(1_K) := e$ and since $\varphi_{K[M]}$ and $\eta_{K[M]}$ are K-linear, the fact that e is a *two-sided identity element* for the *algebra* $K[M]$ follows at once from the fact that e is a *two-sided identity element* for the *monoid* M. □

PROPOSITION 3.21. *Given any associative unitary algebra* (A, φ, η), *let* $\mu : A \times A \to A$ *be a function defined by the commutative diagram*

$$\begin{tikzcd} K[A \times A] \arrow[r, "\pi"] & A \otimes A \arrow[d, "\varphi"] \\ A \times A \arrow[u, "i_{A\times A}"] \arrow[r, "\mu"'] & A \end{tikzcd}$$

(*in* **Sets**) *in which* $i_{A\times A} : A \times A \to K[A \times A]$ *is the inclusion of* $A \times A$ *into* $K[A \times A]$ *as its basis, and in which* $\pi : K[A \times A] \to A \otimes A$ *is the canonical natural projection arising in the construction of the tensor product* $A \otimes A$. *Then the following results obtain*:

(a) *The ordered triple* $(A, \mu, 1_A)$ *is a multiplicatively written monoid with multiplication*

$$\mu := \varphi \circ \pi \circ i_{A\times A} : A \times A \xrightarrow{i_{A\times A}} K[A \times A] \xrightarrow{\pi} A \otimes A \xrightarrow{\varphi} A$$

and two-sided identity element $1_A \in A$.

(b) *If* (M, φ, e) *is any multiplicatively written monoid with multiplication* $\varphi_M : M \times M \to M$ *and two-sided identity element* $e \in M$, *and if* $f : (M, \varphi_M, e) \to (A, \mu, 1_A)$ *is any morphism of monoids, then the unique* K*-linear map* $F : K[M] \to A$ *for which the diagram*

$$\begin{tikzcd} M \arrow[r, "i_M"] \arrow[d, equal] & K[M] \arrow[d, "F"] \\ M \arrow[r, "f"'] & A \end{tikzcd}$$

(*in* **Sets**) *commutes is a morphism of associative unitary algebras,*

(c) *If* $f : (M, \varphi_M, e_M) \to (N, \varphi_N, e_N)$ *is a morphism of multiplicatively written monoids, and if* $K[f] : K[M] \to K[N]$ *denotes the unique* K*-vector space map filling in the diagram*

$$\begin{tikzcd} M \arrow[r, "i_M"] \arrow[d, "f"'] & K[M] \arrow[d, "K[f]"] \\ N \arrow[r, "i_N"'] & K[N] \end{tikzcd}$$

(*in* **Sets**), *then* $K[f]$ *is a map of associative unitary* K*-algebras*; *and hence the assignment*

$$M \mapsto K[M] \quad \textit{and} \quad f \mapsto K[f]$$

defines a functor (*also denoted*) $K : \mathbf{Mon} \to \mathbf{Alg}$, *from the category of* monoids *to the category of* associative unitary K-algebras, *which is* left adjoint *to the* forgetful functor $J : \mathbf{Alg} \to \mathbf{Mon}$ *defined on objects by*

$$J(A, \varphi, \eta) := (A, \mu, 1_A) := \big(A, \varphi \circ \pi \circ i_{A\times A}, \eta(1_K)\big).$$

PROOF. Routine. □

DEFINITION/PROPOSITION 3.22. If (M, φ_M, e_M) and (N, φ_N, e_N) are multiplicatively written monoids, then so is the direct product $(M \times N, \varphi_{M\times N}, e_{M\times N})$ with *binary operation*, $\varphi_{M\times N}$, and *two-sided identity element*, $e_{M\times N}$, defined *coordinatewise* via

$$\begin{aligned}\varphi_{M\times N}\big((m_1, n_1), (m_2, n_2)\big) &:= \big(\varphi_M(m_1, m_2), \varphi_N(n_1, n_2)\big)\\ &:= (m_1 \cdot_M m_2, n_1 \cdot_N n_2)\end{aligned}$$

and

$$e_{M\times N} := (e_M, e_N),$$

respectively.

PROOF. Straightforward. □

PROPOSITION 3.23. *If* (M, φ_M, e_M) *and* (N, φ_N, e_N) *are multiplicatively written monoids and if*

$$\Phi_{M,N} : K[M \times N] \xrightarrow{\cong} K[M] \otimes K[N]$$

is the unique natural K*-linear isomorphism extending the natural set bijection* $\phi_{M,N} : M \times N \xrightarrow{\cong} M \otimes N$ *from the* K-basis *of* $K[M \times N]$ *to the* K-basis *of* $K[M] \otimes K[N]$*, then* $\Phi_{M,N}$ *is a map of* K-algebras *where* $K[M \times N]$ *is an* algebra *because its* basis*,* $M \times N$*, is a* monoid (*as the* direct product *of the* monoids M *and* N) *and where* $K[M] \otimes K[N]$ *is an* algebra *as the* tensor product *of the* monoid algebras $K[M]$ *and* $K[N]$.

PROPOSITION 3.24. *If* (M, φ_M, e_M) *is a* multiplicatively written monoid*, then*

$$\big(K[M], \varphi_{K[M]}, \eta_{K[M]}, \Delta_{K[M]}, \varepsilon_{K[M]}\big)$$

is a bialgebra *whose* underlying algebra *is the* monoid algebra $(K[M], \varphi_{K[M]}, \eta_{K[M]})$ *of* M (*as given in* 3.20(2)) *and whose* underlying coalgebra *is the* group-like coalgebra *on* (*the non-empty set*) M (*as given in* 3.1*, for* $X = M \ni e$).

COROLLARY 3.25. *If* $f : (M, \cdot_M, e_M) \to (N, \cdot_N, e_N)$ *is a* morphism of multiplicatively written monoids*, then* $K[f] : K[M] \to K[N]$ *is a* morphism of bialgebras *and the assignment*

$$M \mapsto K[M] \quad \textit{and} \quad f \mapsto K[f]$$

defines a functor (*also denoted*) K : **Monoids** → **Bialgebras***, from the* category of multiplicatively written monoids *to the category of* K-bialgebras*, which is* left adjoint *to the functor* G : **Bialgebras** → **Monoids** *which assigns to any* bialgebra H *its* monoid*,* $G(H)$*, of* group-like elements.

REMARKS 3.26. (a) If $(H, \varphi, \eta, \Delta, \varepsilon)$ is a bialgebra, then the set of group-like elements of H is easily seen to form (the underlying set of) a multiplicatively written monoid whose binary operation is induced by the multiplication of H and whose identity element is just the identity element 1_H of H. If $(H, \varphi, \eta, \Delta, \varepsilon, S)$ is a Hopf algebra, then the just described monoid of group-like elements of H is, in fact, the underlying monoid of the group in which $g^{-1} = S(g)$ for each $g \in G(H)$. We express the above more succinctly by saying that if H is a *bialgebra*, then $G(H)$, the set of *group-like elements of* H, is a multiplicatively written *monoid* in the obvious way whereas if H is a *Hopf algebra*, then $G(H)$ is a multiplicatively written *group* in the obvious way with $g^{-1} = S(g)$ for each $g \in G(H)$.

(b) One then has the following easy to establish consequences of the adjointness

$$K : \mathbf{Sets} \rightleftarrows \mathbf{Coalg} : G \quad \text{(Corollary 3.12)},$$
$$K : \mathbf{Mon} \rightleftarrows \mathbf{Alg} : J \quad \text{(Proposition 3.21(c))}$$

namely,

$$K : \mathbf{Mon} \rightleftarrows \mathbf{Bialg} : G,$$
$$K : \mathbf{Groups} \rightleftarrows \mathbf{Hopf\ Alg} : G.$$

Since G, *being a right adjoint, preserves products* (cf. [Hil-Sta, p. 68, Theorem 7.7]), and since the *product in the category* of *cocommutative coalgebras* (respectively, the *category* of *cocommutative bialgebras*, or *cocommutative Hopf algebras*) is the *tensor product* (cf. [Gr-1, p. 18, Korollar I.1.7] or [Dăs-Năs-Rai, p. 31, Proposition 1.4.21]) whereas the *product in the category* of *monoids* (respectively, the *category* of *groups*) is the *direct product* (cf. [Mac-4, p. 69] or [Arb-Man, p. 54, *Products in* **Mon** and **Grp**]]), it follows that for all cocommutative bialgebras (respectively, all cocommutative Hopf algebras) H_1 and H_2

$$G(H_1 \otimes H_2) \cong G(H_1) \times G(H_2).$$

In fact, one can prove this directly without assuming that the bialgebra or Hopf algebra is cocommutative (cf. [Cae, p. 179, Proposition 7.1.8]).

(c) *Nonisomorphic groups can have isomorphic group algebras*. For example, the *dihedral group of order* 8 and the *quaternion group* have *isomorphic group algebras*. See, for example, [Palm, p. 129].

REMARK 3.27. It turns out that the K*-vector space dual*, H^*, of any *finite-dimensional Hopf algebra* H has again the *structure of a Hopf algebra*, called the *dual Hopf algebra*. In greater detail (see, for example, [Dăs-Năs-Rai, p. 149, Proposition 4.1.6; p. 156, Proposition 4.2.11]), if $(H, \varphi, \eta, \Delta, \varepsilon, S)$ is a *finite-dimensional* Hopf algebra, so is $(H^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda, \rho^{-1} \circ \varphi^*, \lambda^{-1} \circ \eta^*, S^*)$. Since the *group-algebra*, $K[G]$, of any *finite* group, G, is a *finite-dimensional Hopf algebra*, it follows that, for any *finite* group G, $(K[G])^*$ has the structure of a *finite-dimensional Hopf algebra*. Because any K-linear function $f : G \to K$ is completely determined by its values on (i.e., its restriction to) the *basis* G of $K[G]$, one clearly has (first as underlying sets, then) as K-vector spaces, that

$$(K[G])^* \equiv K^G := \text{the } K\text{-vector space of all functions } f : G \to K,$$

with addition (of functions) and (the) multiplication (of a function) by a scalar ($k \in K$) defined pointwise. The fact that $K^G \equiv (K[G])^*$ is one (albeit, indirect) way to see that, for any *finite group* G, K^G has the structure of a *finite-dimensional Hopf algebra*. One may also see this directly. To do so, first note that whenever S is a *non-empty set*, the *set diagonal* $d : S \to S \times S$ induces the K-linear map (the *product*)

$$K^S \otimes K^S \xrightarrow{\pi} K^{S\times S} \xrightarrow{d^*} K^S$$

on K^S, where $\pi : K^S \otimes K^S \to K^S$ is the K-vector space *injection* of Section 1 given, for all $(f, g) \in K^S \times K^S$ and all $(x, y) \in S \times S$, by

$$\pi\Big(\sum f \otimes g\Big)(x, y) := \sum f(x) \cdot g(y).$$

When, additionally, the non-empty set S is *finite*, then the canonical K-vector space *injection*

$$\pi : K^S \otimes K^S \to K^{S\times S}$$

is an *isomorphism* of K-algebras. Therefore, for a finite group G, the *group multiplication* $m : G \times G \to G$ *induces* a K-linear map

$$K^G \xrightarrow{m^*} K^{G\times G}$$

and hence a *comultiplication*

$$K^G \xrightarrow{m^*} K^{G\times G} \xrightarrow{\pi^{-1}} K^G \otimes K^G.$$

The other structure maps of the *Hopf algebra* K^G, where G is a *finite* group, are given, respectively, as the composites

$$\begin{aligned}
&\eta_{K^G} := p^* \circ \alpha : K \xrightarrow{\alpha} K^{\{e\}} \xrightarrow{p^*} K^G,\\
&\varepsilon_{K^G} := \beta \circ \iota^* : K^G \xrightarrow{\iota^*} K^{\{e\}} \xrightarrow{\beta} K, \quad \text{and}\\
&S_{K^G} := i^* : K^G \xrightarrow{i^*} K^G
\end{aligned}$$

where $e :=$ the *identity* element of G, $\iota : \{e\} \hookrightarrow G$ is the *inclusion* map, $p : G \to \{e\}$ is the *projection* map, $i : G \to G$ is the *inverse* map assigning to any $g \in G$ its *inverse*, g^{-1}, and where $\beta : K^{\{e\}} \xrightarrow{\cong} K$ and $\alpha : K \xrightarrow{\cong} K^{\{e\}}$ are the *set bijections*, *inverse to one another*, given, respectively, by $\beta(f) := f(e)$ and by $\alpha(k) := f_k : \{e\} \to K$ where $f_k(e) := k$, for each $k \in K$.

REMARK 3.28. The direct proof that the *contravariant functor* $K^{(-)} : \mathbf{Sets} \to \mathbf{Vect}$ induces a *contravariant functor* (also denoted) $K^{(-)} : \mathbf{Groups} \to \mathbf{Hopf\,Alg}$ proceeds in the same way as the proof that the *covariant functor* $K := K[-] : \mathbf{Sets} \to \mathbf{Vect}$ induces a *covariant functor* (also denoted) $K := K[-] : \mathbf{Groups} \to \mathbf{Hopf\,Alg}$ except that the arrows in the diagrams expressing the defining properties of a group (in **Sets**) have their sense reversed when the *contravariant functor* $K^{(-)} : \mathbf{Sets} \to \mathbf{Vect}$ is applied to any one of the three commutative diagrams in **Sets** expressing the defining properties of a group, but in each case the relevant multiplicative functor [$K^{(-)}$ or $K[-]$] takes a *finite bimonoid* (respectively, *group*) in the *monoidal category* (**Sets**, $\times$, 1), $1 := \{*\}$ denoting any one of the isomorphic terminal objects of **Sets**, to a *finite-dimensional bimonoid* (respectively, *group*) in the *monoidal category* (**Vect**, $\otimes$, K).

NOTE. The finiteness condition on G is *not* necessary in the covariant case of $K[G]$. By contrast, it is *crucial* to get

$$K^{G \times G} \cong K^G \otimes K^G.$$

REMARK 3.29. It is a *nice exercise* to verify, for a *finite group* G, that the Hopf algebra structure that one gets on K^G, directly, is precisely the same as the Hopf algebra structure that $K^G (= (K[G])^*)$ has as the *dual Hopf algebra* of the *finite-dimensional* Hopf algebra $K[G]$.

REMARK 3.30. Returning now to the case of the monoid algebra, we recall (cf. [Gril, pp. 69 and 146–148]) that the *monoid algebra* $K[FCM(X)]$ of the *free commutative monoid* $FCM(X)$ on a set X is the *algebra of polynomials in the commuting variables* of X. As we know, $K[FCM(X)]$ is a bialgebra in case the coalgebra structure results from considering each element of $FCM(X)$, and in particular each element of X, to be *group-like*. In fact, one obtains a distinct coalgebra structure on $K[FCM(X)]$, the algebra of polynomials in the commmuting variables of X, one making it a bialgebra, even a Hopf algebra, by declaring each element of X to be *primitive* (with respect to 1_H, which is *group-like*). When this is done for $X = \{x\}$, a 1-element set, one obtains the familiar Hopf algebra structure on the polynomial algebra $K[x]$, given by declaring 1 ($= 1_{K[X]} = 1_K$, the *identity element* of the ground field considered as a constant polynomial) *group-like* and x *primitive* and correspondingly (by, respectively, Lemmas 2.66 and 2.74) $S(1) = 1$ while $S(x) = -x$. This construction can be generalized to obtain a Hopf algebra structure on the polynomial algebra $K[x_1, x_2, \ldots, x_n]$ in the *n commuting variables* $x_1, x_2, \ldots, x_n$. Again, 1 is *group-like*, each x_i is *primitive*, and, therefore, $S(1) = 1$ while $S(x_i) = -x_i$, for each i.

As a Hopf algebra, the *algebra of polynomials* (in the – however many – *commuting variables* $x_\beta \in \{x_\alpha\}_{\alpha \in I}$, I an index set) is a special case of the *symmetric algebra* $S(V)$ *on the* K*-vector space* V (having $\{x_\alpha\}_{\alpha \in I}$ as its *basis*). The *symmetric algebra* $S(V)$, in turn, is a special case of the *universal enveloping algebra* $U(L)$ *of a Lie algebra* L (in this case the K-vector space V considered trivially as the *Lie algebra* $\mathrm{Triv}(V)$ all of whose *Lie brackets* are *zero*). Now, for any Lie algebra L, $U(L)$ is a *quotient algebra* of the *tensor algebra*

$T(L)$, while, for any K-vector space V, $S(V)$ is a *quotient algebra* of the *tensor algebra* $T(V)$. For any K-vector space V, $T(V) := \bigoplus_{n=0}^{\infty} T_n(V)$ where $T_0(V) := K$, $T_n(V) := V^{\otimes n}$ for any integer $n \geqslant 1$, and where the *product* (= multiplication) in $T(V)$ is given by *juxtaposition*. In other words, if $\{x_\alpha\}_{\alpha \in I}$ is a K*-basis* for V, then $T(V) \equiv K\langle x_\alpha \rangle :=$ the K*-algebra of all polynomials in the non-commuting variables* $x_\beta \in \{x_\alpha\}_{\alpha \in I}$. For any *associative unitary* K*-algebra* A*, any* K*-linear map* $f : V \to A$ *extends uniquely* to an *algebra map* $F : T(V) \to A$. In other words, for any K-vector space V, $T(V)$ together with the *inclusion* $i : V =: T_1(V) \hookrightarrow T(V)$, is the *free associative unitary* K*-algebra on* V in the sense that the obviously defined *functor* $T : \mathbf{Vect} \to \mathbf{Alg}$ is *left adjoint* to the *forgetful functor*.

PROPOSITION 3.31. *For any* K*-vector space* V*, the* tensor algebra $T(V)$ *has a* unique Hopf algebra structure *determined by setting* $\Delta(1) = 0$*,* $\varepsilon(1) = 1$*,* $S(1) = 1$*, and*

$$\Delta(v) := 1 \otimes v + v \otimes 1, \quad \varepsilon(v) := 0, \quad \textit{and} \quad S(v) := -v,$$

for each $v \in V$*. These definitions entail that, for all* $v_1, v_2, \ldots, v_n \in V$*,*

$$\varepsilon(v_1 v_2 \ldots v_n) = 0, \quad S(v_1 v_2 \ldots v_{n-1} v_n) = (-1)^n (v_n v_{n-1} \ldots v_2 v_1),$$

and

$$\begin{aligned}
&\Delta(v_1 v_2 \ldots v_n) \\
&\quad = \prod_{i=1}^{n} \Delta(v_i) = \prod_{i=1}^{n} (1 \otimes v_i + v_i \otimes 1) \\
&\quad = 1 \otimes (v_1 \ldots v_n) + \sum_{p=1}^{n-1} (v_{i_1} \ldots v_{i_p}) \otimes (v_{j_1} \ldots v_{j_{n-p}}) + (v_1 \ldots v_n) \otimes 1
\end{aligned}$$

where the sum ranges over all pairs of strictly increasing sequences

$$i_1 < i_2 < \cdots < i_p \quad \textit{and} \quad j_1 < j_2 < \cdots < j_{n-p}$$

which partition *the set* $\{1, 2, \ldots, n\}$ *into* disjoint *subsets* $\{i_r\}_{r=1}^{p}$ *and* $\{j_s\}_{s=1}^{n-p}$.

SKETCH OF PROOF. The formulae for $\varepsilon(v_1 \ldots v_n)$ and for $S(v_1 \ldots v_n)$ are immediate consequences, respectively, of the definitions and the fact that $\varepsilon : TV \to K$ and $S : TV \to (TV)^{op}$ are morphisms of **Alg**. The formula for $\Delta(v_1 v_2 \ldots v_n)$ follows upon expanding $\prod_{i=1}^{n} (1 \otimes v_i + v_i \otimes 1)$. A proof (by induction) may be found, for example, in [Kass, pp. 47–48, Theorem III.2.4]. □

REMARK 3.32. One may describe $\Delta(v_1 \ldots v_n)$ in terms of *shuffles* as follows (cf. [Kli-Sch, p. 19, Example 8]):

$$\Delta(v_1 \ldots v_n) = \sum_{k=0}^{n} \sum_{\sigma \in \mathcal{P}_{n,k}} (v_{\sigma(1)} \ldots v_{\sigma(k)}) \otimes (v_{\sigma(k+1)} \ldots v_{\sigma(n)})$$

where $\mathcal{P}_{n,k}$ denotes the set of all *permutations* σ of $\{1, 2, \dots, n\}$ such that $\sigma(1) < \cdots < \sigma(k)$ and $\sigma(k+1) < \cdots < \sigma(n)$. The elements of $\mathcal{P}_{n,k}$ are called $(k, n-k)$-*shuffles*. In the above formula, it is to be understood that the summand corresponding to $k = 0$ is $1 \otimes (v_1 \dots v_n)$ whereas the summand corresponding to $k = n$ is $(v_1 \dots v_n) \otimes 1$ (cf. [K-R-T, p. 12, Example 1.5]).

REMARK 3.33. The formulae in 3.31 and 3.32 express $\Delta(v_1 \dots v_n)$ as a sum of 2^n summands. These may be obtained in an orderly way according to the following decision-making process: Start with the expression $1 \otimes (v_1 v_2 \dots v_n)$. Then, for each $i \in \{1, 2, \dots, n\}$, beginning with 1 and proceeding in order to n, decide whether v_i is to *stay* (S) where it is on the right-hand side of the $\otimes$-product sign, or, instead, *move* (M) to the spot immediately to the left-hand side of the $\otimes$-product sign. This pair of choices (S or M), one for each of the n *productands* v_i of $v_1 \dots v_n$, results altogether in 2^n outcomes which may be listed in an orderly way from a tableau of the resulting decision tree diagram. The reader encountering this for the first time will find it instructive to work out the details, say, for $n \in \{2, 3, 4\}$.

REMARK 3.34 (a). By the famous *Poincaré–Birkhoff–Witt Theorem* (cf. [J-2, p. 159, Theorem 3] or [M-P, p. 41, Theorem 1]), if $\{x_\alpha\}$ is a *well-ordered basis* of a *Lie algebra L*, then, upon identifying each $x_\alpha \in L$ with its *image* $i_{UL}(x_\alpha)$ in $U(L)$ under the *canonical, natural map* $i_{UL}: L \to U(L)$ (necessarily an *injection* by the P-B-W Theorem) the set

$$\{x_{\alpha_1} x_{\alpha_2} \dots x_{\alpha_n}\}_{\alpha_1 \leqslant \alpha_2 \leqslant \cdots \leqslant \alpha_n}$$

of so-called *standard monomials*, together with $1 \in K$, forms a K-vector space *basis* of $U(L)$. It turns out that $U(L)$ and $S(V)$ are *quotient Hopf algebras*, respectively, of $T(L)$ and of $T(V)$ *by respective Hopf ideals*. From this fact it easily follows that the *coalgebra structures* on $U(L)$ and on $S(L)$ are *likewise given by shuffles* (as in 3.32). Accordingly, the obvious K-vector space *isomorphism* between $U(L)$ and $S(L)$ obtained by sending a P-B-W *basis* of $U(L)$ to one of $S(L)$ provides an *isomorphism* of $U(L)$ and $S(L)$ *as augmented coalgebras*. As a consequence, by the functoriality of $(-)^*$: **Coalg** $\to$ **Alg**, one obtains an *isomorphism* $[U(L)]^* \cong [S(L)]^*$ of the *dual augmented* algebras.

(b) In *characteristic* 0, the so-called *canonical symmetrization* map $\omega: S(L) \to U(L)$ given (cf. [Dix, p. 78]) by

$$\omega(x_1 x_2 \dots x_n) := \frac{1}{n!}\left[\sum_{\sigma \in S_n} x_{\sigma(1)} x_{\sigma(2)} \dots x_{\sigma(n)}\right]$$

is an *isomorphism of augmented coalgebras* (cf. [Smo, p. 467]).

(c) Quillen [Q, p. 281, Theorem 2.3] calls the analogous result for *DG* Lie algebras the *Poincaré–Birkhoff–Witt Theorem*.

(d) By the functoriality of $(-)^*$: **Coalg** $\to$ **Alg**, one obtains, from (b), as from (a), an *isomorphism* $[U(L)]^* \cong [S(L)]^*$ of the *dual augmented algebras*.

(e) Some authors, e.g., [Smo, p. 467] call the *dual algebra*, $[U(L)]^*$, of the *coalgebra* $U(L)$, the *jet algebra* of the *Lie algebra L*.

(f) When L is 1-dimensional, $U(L) = S(L) = K[x] =$ the familiar algebra of polynomials in x with coefficients in K. In this case, the diagonal $\Delta : K[x] \to K[x] \otimes K[x]$ takes the simpler form

$$\Delta(x^n) = \sum_{k=0}^{n} \binom{n}{k} x^k \otimes x^{n-k}.$$

In *characteristic* 0, if a transformation of the *basis* of $K[x]$ is made (cf. [Ca-1, p. 3-13]) via

$$x^n \leftrightarrow \frac{x^n}{n!} =: c_n,$$

then the resulting coalgebra structure on $K[x]$ with this new *basis* is known as the *divided power* (or *divided powers*) coalgebra. Here $\varepsilon(c_n) = \delta_{0,n} \cdot 1_K$ and

$$\Delta(c_n) = \sum_{k=0}^{n} c_k \otimes c_{n-k},$$

a formula that defines the *divided power coalgebra* on the K-vector space

$$\mathcal{D}^\infty := \bigoplus_{i=0}^{\infty} K \cdot c_i$$

in any characteristic.

(g) In *characteristic* 0, with $\dim L = n < \infty$, by *changing the basis of* $U(L)$ *in the analogous fashion to changing the basis of* $K[x]$ as above, one finds by a generalization of the argument in 3.36, below, that $[U(L)]^*$ identifies with the *algebra* $K[[t_1, \dots, t_n]]$ *of formal power series in* $t_1, \dots, t_n$ (cf. [Ca-1, p. 3-13, Proposition 6] or [Dix, p. 90, Proposition 2.7.5] or [Mon-1, p. 75, Lemma 5.5.4]).

DEFINITION 3.35. The *divided power coalgebra* is the *coalgebra* $(\mathcal{D}^\infty, \Delta, \varepsilon)$ where $\mathcal{D}^\infty$ is the K-vector space with *basis* $\{d_n\}_{n=0}^{\infty}$ and with Δ and ε defined by setting

$$\Delta(d_n) := \sum_{k=0}^{n} d_k \otimes d_{n-k}$$

and

$$\varepsilon(d_n) := \delta_{0,n} \cdot 1_K.$$

For any integer $N > 1$, $\mathcal{D}^\infty$ has the *finite-dimensional analog* $\mathcal{D}^N$ where $\mathcal{D}^N$ is the K-vector space with *basis* $\{d_n\}_{n=0}^{N}$ and where Δ and ε are given as above. Observe, here, that d_0 is *group-like* while d_1 is d_0*-primitive*.

REMARK 3.36. We now identify $(\mathcal{D}^\infty)^*$, the *dual algebra* of the *divided power coalgebra* $\mathcal{D}^\infty$ (cf. [Sw-1, p. 43, Example-Exercise], [N-S, p. 58] or [Dăs-Năs-Rai, pp. 18–19, Example 1.3.8.2]). To do so, for each $m \in \mathbb{N}$, define $t^m \in (\mathcal{D}^\infty)^*$ by setting

$$t^m(d_n) := \delta_{m,n} \cdot 1_K.$$

Then $(\mathcal{D}^\infty)^* \cong K[[t]]$ as K-vector spaces. By definition, $(\mathcal{D}^\infty)^*$ has its *product*, the *convolution product*, given, for all $f, g \in (\mathcal{D}^\infty)^*$, by

$$f * g(d_n) := \sum_{k=0}^{n} f(d_k) \cdot g(d_{n-k}).$$

In particular, then, for all $r, s \in \mathbb{N}$,

$$\begin{aligned} t^r * t^s(d_n) &:= \sum_{k=0}^{n} t^r(d_k) \cdot t^s(d_{n-k}) \\ &= \sum_{k=0}^{n} \delta_{r,k} \cdot 1_K \cdot \delta_{s,n-k} \cdot 1_K \\ &= \delta_{r+s,n} \cdot 1_K \end{aligned}$$

whence

$$t^r * t^s = t^{r+s}.$$

The *unit* $\eta : K \to (\mathcal{D}^\infty)^*$ of $(\mathcal{D}^\infty)^*$ is given by setting

$$\eta(k)(d_n) := k \cdot \delta_{0,n}.$$

Accordingly, the *dual algebra* $(\mathcal{D}^\infty)^*$ *of the divided power coalgebra* $\mathcal{D}^\infty$ is *canonically isomorphic* to *the* K*-algebra* $K[[t]]$ of *formal power series in* t under the map

$$\phi : (\mathcal{D}^\infty)^* \to K[[t]]$$

given by setting

$$\phi(f) := \sum_{n=0}^{\infty} f(d_n) \cdot t^n.$$

REMARK 3.37. The *divided power coalgebra* $\mathcal{D}^\infty$ may be given the structure of a *bialgebra* (cf. [N-S, p. 71]), even a Hopf algebra (cf. [Dăs-Năs-Rai, pp. 164–165, Example 5]).

TAKING STOCK 3.38. Interestingly, the group algebra of a group and the universal enveloping algebra of a Lie algebra (which, at first blush, at least for many people, have just the structure of an algebra) each have an additional, compatible, coalgebra structure making each a Hopf algebra. This fact leads to several questions:

(1) What kind of algebra can be the underlying algebra of a bialgebra or Hopf algebra? Dually, what kind of coalgebra can be the underlying coalgebra of a bialgebra or Hopf algebra?

(2) What kind of Hopf algebra can be the group algebra of a group?

(3) What kind of Hopf algebra can be the universal enveloping algebra of a Lie algebra?

(4) What kind of Hopf algebra can be a type of *semi-direct product* of a group algebra and an enveloping algebra?

(5) Are there some algebras that cannot be given the structure of a bialgebra or a Hopf algebra? Likewise, are there some coalgebras that cannot be given the structure of a bialgebra or a Hopf algebra?

Because these questions pertain to bialgebras or Hopf algebras, to be considered elsewhere in this *Handbook*, here, next, in Remarks 3.39, 3.40, 3.41, 3.42, and 3.43, we content ourselves with brief answers, respectively, to questions 1, 2, 3, 4, and 5, above.

REMARK 3.39. The question of what kind of algebra can serve as the underlying algebra of a Hopf algebra was partially answered in the *graded context* by the famous *Hopf–Leray Theorem* (cf. [M-M-1, p. 27, Theorem 4.6] and [M-M-2, p. 252, Theorem 7.5]) which asserts that if B is a connected graded Hopf algebra over a field K of characteristic 0 [*Connected* means, by definition, that $B_0 = K$, where B_0 is the component of B of degree 0, cf. 2.37(d)], then as an algebra B is isomorphic to the (graded) symmetric algebra $S(Q(B))$ on the space $Q(B) := \overline{B}/\overline{B}^2$ of *indecomposables* of B. [Here, by definition, $\overline{B} := \operatorname{Ker} \varepsilon$ where $\varepsilon : B \to K$ is the *counit* (or *augmentation*) of B. In short, $\overline{B} :=$ the *augmentation ideal* of the *augmented algebra* $(B, \varphi, \eta, \varepsilon)$, cf. the Note following Remark 2.73(b).] In his 1975 University of Chicago Ph.D. thesis, [N-1], Warren Nichols established an ungraded counterpart of both the *Hopf–Leray Theorem* [N-1, p. 57, Theorem 7(i)] and a *Dual Hopf–Leray Theorem* [N-1, p. 58, Theorem 7(ii)]; see also [N-2, p. 70, Theorem 8; p. 71, Remark; and p. 64, Introduction] and [Block, pp. 314–315]. Because Nichols is dealing with ungraded bialgebras, his definition of "connected" is a technical one but one which corresponds to that used by algebraic topologists in the graded case – where the 0th homology group of a connected space is simply the coefficient group.

NOTE. For a discussion of what Nichols felt was known about the above topic at the time of his paper, see [N-2, p. 64, Introduction].

REMARKS 3.40 (a). In [Zhu, p. 57, Theorem 2], Yongchang Zhu, extending a result [Kac-3] of G.I. Kac (sometimes spelled Kats or Katz), proved that any Hopf algebra of prime dimension p over an algebraically closed field K of characteristic 0 is a group algebra, specifically, the group algebra, $K[C_p]$, of the cyclic group, C_p, of order p (cf. [Dăs-Năs-Rai, p. 329, Theorem 7.6.4; p. 358, Bibliographical Notes; and pp. 168–169 and 175–176, Exercises 4.3.6–4.3.8]). In so doing, Zhu established a conjecture, the 8th out of 10, made by Irving Kaplansky in his 1975 University of Chicago Notes *Bialgebras* [Kap, pp. 45–46]. A different proof of [Zhu, p. 57, Theorem 2] was subsequently given by Pavel Etingof and Shlomo Gelaki [E-G-1, p. 194, Theorem 1.7]. Montgomery [Mon-2, p. 267, Item (0.2)] remarks that Kac established a result analogous to that of Zhu for *Kac algebras*. [A *Kac algebra* (cf. [Străt, p. 278, Item 18.15] or [Enoc-Schw, p. 56, Definition]) is a certain kind of *Hopf–von Neumann algebra* (cf. [Străt, p. 256, Item 18.2] or [Enoc-Schw, p. 8, Definition 1.1.1]). A *Hopf–von Neumann algebra* (cf. [Er, p. 203, Definition 3.1]) is a *von Neumann algebra* admitting a compatible coalgebra structure making it a Hopf algebra. In [Zhu, pp. 56–57], Zhu states that the original theorem of G.I. Kats was given "in the context of ring groups. A ring group is a Hopf algebra over $\mathbb{C}$ with a C^*-algebra structure compatible with the Hopf algebra ..., and it is in particular semisimple." Zhu goes on [*loc. cit.*, p. 57] to state that results of [Larson-3] and [Lar-Rad-1] "allow the generalization of Kats's argument to semisimple Hopf algebras over an algebraically closed field of characteristic 0." For an especially nice, fairly expository treatment of Kaplansky's *Ten Conjectures on Hopf algebras* [Kap, pp. 45–46], see [Sommer-2].

(b) Other results giving conditions on a bialgebra (respectively, Hopf algebra) that will guarantee it is the monoid algebra of a monoid (respectively, the group algebra of a group) may be found in [Lar-1], [Taka], [Larson-2], [Larson-3], [Gray, pp. 197–200, especially, p. 199, Theorem 14], [Dem-Gab-1, p. 509, Theorem 3.6], [Sweed-4, pp. 69–70, Theorem 4.1], [Ab, p. 223, Theorem 4.6.2], [Mon-1, p. 22, Theorem 2.3.1; p. 83, Theorem 5.7.1; p. 85, Theorem 5.7.4], [Mon-2, p. 267, items (1.1), (1.3), and (1.4)], and [Etin-Schi, p. 72, Proposition 8.2]. See also [Chil, p. 85, Theorem 15.2; p. 86, Remark].

REMARK 3.41. In [Sw-1, p. 274, Theorem 13.0.1; p. 193, Theorem 9.2.2; and p. 224, Exercise 2 and Proposition 11.0.9] or [Mon-1, p. 73, Proposition 5.5.3.1; and p. 79, Proposition 5.6.5], it is shown that over a field of characteristic 0, a *cocommutative Hopf algebra* H is the *universal enveloping algebra of a Lie algebra* (specifically, that $H = U(L)$ where $L = P(H)$, the *Lie algebra* of *primitive elements* of H) if and only if H is *connected*, meaning it has a *unique group-like* element. Sweedler remarks on page 2 of the Preface to [Sw-1] that this (unpublished) result is due to Bertram Kostant.

NOTE. To any *associative algebra* A, one can associate a *Lie algebra* $\mathcal{L}(A)$ called the *associated Lie algebra* of A. As a K-vector space, $\mathcal{L}(A) \equiv A$. The *Lie bracket* of $\mathcal{L}(A)$ is given by setting

$$[x, y] := [x, y]_{\mathcal{L}(A)} := (1 - \tau) \circ \varphi_A(x \otimes y) := x \cdot y - y \cdot x$$

where "$\cdot$" denotes the *product* of A. It is an easy exercise to verify, for any bialgebra or Hopf algebra H, that $[x, y] \in P(H)$ whenever $x, y \in P(H)$ (cf. [Ca-1, pp. 3-14–3-15, Lemma 3]). This proves that $P(H)$ is a *sub Lie algebra* of $\mathcal{L}(H)$ for any *bialgebra* or *Hopf algebra* H.

REMARK 3.42. An unpublished result due to Bertram Kostant gives conditions under which a *cocommutative Hopf algebra* can be written as the *semidirect product* of an *enveloping algebra* and a *group algebra*. Various people have written this result down in various ways. See, for example, [Lar-1, p. 350, Introduction; p. 357, Theorem 5.3; p. 358, Remark], [Hey-Sw-1, p. 194, Introduction; p. 238, Theorem 3.5.8], [Sw-1, Preface (p. 2);

§8.1, §13.1], [Gr-1, p. 83, Theorem III.3.10], [Gr-2, p. 181, Theorem 3.13], [Win, pp. 186–190] and [Mon-1, p. 78, Corollary 5.6.4(3)]. Cartier and Gabriel established *Kostant's result*, in characteristic 0, *in dual form* (cf. [D-2, p. 42]) in the context of *formal groups* (cf. [Mon-1, p. 79]). Here we give the statement of Kostant's theorem in the form found in Luzius Grünenfelder's 1969 E.T.H. Ph.D. thesis [Gr-1]. There, [Gr-1, p. 83, Theorem III.3.10], it is shown that if H is a Hopf algebra over a field of characteristic 0, then H is the *semi-direct* (or *smash*) product $U[P(H)]\#K[G(H)]$ of $U[P(H)]$ and $K[G(H)]$ if and only if H is *cocommutative* and the *coradical* $\operatorname{Cor}(H)$ of H is *generated by group-like elements*, i.e., if and only if

$$\operatorname{Cor}(H) = \bigoplus_{g \in G(H)} K \cdot g = K\bigl[G(H)\bigr].$$

An analogous, but slightly different, result (stated in [Gr-1, p. 83, Theorem III.3.10]) holds in characteristic p. In this connection we note that:

(a) The *semi-direct product of Hopf algebras* is a generalization of the *semi-direct product of groups* and the *semi-direct product of Lie algebras* and it is compatible with each in the obvious sense. See, for example, [Hey-Sw-1, pp. 209–210, Examples 1.8.2 and 1.8.3] or [Molnar-2, p. 29, Introduction; p. 47, Examples 4.3; p. 47, Theorem 5.1].

(b) The *coradical* $\operatorname{Cor}(C)$ of a coalgebra C is (cf. [Sw-1, p. 181, Definition]) the *sum* (necessarily *direct*, it turns out) of all the *simple subcoalgebras* of C, a *coalgebra* being *simple* (cf. [Sw-1, p. 157, Definition]) if it has no subcoalgebras other than 0 and itself. As expected, for any coalgebra C, $[\operatorname{Cor}(C)]^{\perp} = \operatorname{Jac}(C^*)$, the *Jacobson radical* of C^* (cf. [Gr-1, p. 74, Theorem III.2.12], [Gr-2, p. 174, Theorem 2.5(a)], [Sw-1, p. 183], [Ab, p. 84, Theorem 2.3.9(i)], [Mon-1, p. 58, Remark 5.1.7; p. 63, Proposition 5.2.9], or [Dăs-Năs-Rai, p. 120, Proposition 3.1.8(ii)]).

NOTE. By 4.12, *any simple associative unitary coalgebra over a field K necessarily is finite-dimensional*, hence (by 2.86 and 2.88(a)) the dual coalgebra of a finite-dimensional simple associative unitary algebra which by a celebrated theorem of Wedderburn (cf. [Ni, p. 586, Theorem 11.5] or [He, pp. 385–386]) is just $\mathcal{M}_n(D)$, the algebra of all n-by-n matrices with entries in D where D is a *division algebra, finite-dimensional over K*. According to Frobenius's Theorem (cf. [Gril, p. 497, Theorem 16.6.8]), a division $\mathbb{R}$-algebra which is finite-dimensional over $\mathbb{R}$ is isomorphic to $\mathbb{R}$, $\mathbb{C}$, or the *quaternion algebra* $\mathbb{H}$ (so denoted after its discoverer Hamilton) – cf. [Gril, pp. 618–620].

(c) A coalgebra is *pointed* (cf. [Sw-1, p. 157, Definition]) if and only if *all the simple subcoalgebras* of C are 1-*dimensional*.

(d) The *sum* of *distinct simple subcoalgebras* of a *coalgebra* is *direct* (cf. [Sw-1, p. 166, Corollary 8.0.6(a)]).

(e) In a *pointed coalgebra C*,

$$\operatorname{Cor}(C) = \text{the sum of all the 1-dimensional subcoalgebras of } C.$$

(f) By [Sw-1, p. 158, Lemma 8.0.1(e)], $G(C)$ is in 1–1 correspondence with the set of 1-*dimensional subcoalgebras* of C via $g \leftrightarrow K \cdot g$.

(g) Hence, in a *pointed coalgebra* C, the *coradical* $\mathrm{Cor}(C)$ of C has a *basis* of *group-like* elements, i.e.,

$$\mathrm{Cor}(C) = \bigoplus_{g\in G(C)} K \cdot g = K\big[G(C)\big].$$

(h) The version of [Gr-1, p. 83, Theorem III.3.10] found in [Sw-1, p. 279, Exercise] reads as follows: If H is a *cocommutative, pointed Hopf algebra* over a *field of characteristic* 0, then,

$$H \cong U\big[P(H)\big]\#K\big[G(H)\big].$$

as Hopf algebras while $H \cong U[P(H)] \otimes K[G(H)]$ *as coalgebras*, and hence, *as coalgebras*

$$H \cong \bigoplus_{g\in G(H)} U\big[P(H)\big],$$

"the direct sum of copies $U[P(H)]$, indexed by $G(H)$."

REMARK 3.43. As previously noted, neither the matrix algebra $\mathcal{M}_n(K)$ nor the matrix coalgebra $\mathcal{M}_n^c(K)$ can have the structure of a bialgebra if $n > 1$.

REMARK 3.44. We now turn to the construction of a rather broad class of coalgebras. The idea is this: Whereas the *vector space dual* C^* of an *arbitrary* coalgebra C can always be endowed with the structure of an *algebra* coming from the coalgebra structure of C – this because the *comultiplication* $C \xrightarrow{\Delta} C \otimes C$ of C gives rise to the *multiplication* $C^* \xleftarrow{\Delta^*} (C \otimes C)^* \overset{\rho}{\hookleftarrow} C^* \otimes C^*$ on C^* [where ρ is the *natural* K-vector space *injection*], by contrast, the *vector space dual* A^* of an *arbitrary infinite-dimensional* algebra A does not in general support a coalgebra structure induced by the multiplication $\varphi: A \otimes A \to A$ of A because in this case we have the situation $A^* \xrightarrow{\varphi^*} (A \otimes A)^* \overset{\rho}{\hookleftarrow} A^* \otimes A^*$ and, unfortunately, the map ρ "goes the wrong way," and $\rho: A^* \otimes A^* \to (A \otimes A)^*$ cannot be inverted unless A is *finite-dimensional*. There is, however, a functorial way of associating to any algebra A (whatever its dimension) a coalgebra denoted A^0. We shall refer to A^0 simply as "A-upper zero" or "the upper zero of A." Other common ways of referring to A^0 include (1) *the coalgebra dual to* A [N-S, p. 80], and (2) *the dual coalgebra of* A (cf. [L-R, p. 19, Definition 1.3.1]), (3) *the dual* K*-coalgebra of the* K*-algebra* A (cf. [Win, p. 179] and [Ab, p. 74]), (4) *the finite dual to* A (cf. [Mon-1, p. 3, Definition 1.2.3] and [Dăs-Năs-Rai, pp. 33–49, §1.5]), and (5) *the restricted dual to* A (cf. [Kor-Soi, p. 13, Proposition 2.2.6]). Other notations for A^0 include $A^\star$ (cf. [Jos, pp. 26–35, §1.4]), $A^\bullet$ (cf. [Kor-Soi, p. 13]), and A' (cf. [H-5, p. 229], [Bour-1, p. A III.202, Exercise 27(b)] and [Bour-2, p. 654, Exercise 27(b)]). **Nota Bene:** In [H-5], A^o is used to denote $A^* = \mathrm{Hom}_{\mathbf{Vect}}[A, K]$). One can think of A^0 as the *continuous linear dual of* A when the ground field K has the *discrete topology* and when A has as its *neighborhood basis of the origin* the *collection* $C.I.(A)$ *of all cofinite two-sided ideals of* A. Then, with $I^\perp := \{f \in A^* \mid f(I) =$

$\{0\}\} =$ the *annihilator* of I in A^* as in 4.1, below, one *defines* $A^0 := \bigcup_{I \in C.I.(A)} I^{\perp}$, i.e., one *defines*

$$A^0 = \{f \in A^* \mid \operatorname{Ker} f \text{ contains a } \textit{cofinite two-sided ideal} \text{ of } A\}.$$

From the description of A^0 it is easy to see just why A^0 has the structure of a coalgebra. The insight comes from the following observation: If $i : I \hookrightarrow A$ denotes the *inclusion*, then

$$I^{\perp} := \operatorname{Ker}(i^* : A^* \to I^*)$$

so the exact sequence $0 \to I \overset{i}{\hookrightarrow} A \xrightarrow{\pi} A/I \to 0$, in which i is the *inclusion* and π the *projection*, gives rise to the exact sequence $0 \to (A/I)^* \xrightarrow{\pi^*} A^* \xrightarrow{i^*} I^* \to 0$. It follows that

$$I^{\perp} = \operatorname{Ker}(i^*) = \operatorname{Im}(\pi^*) \cong (A/I)^*.$$

But, if I is a *cofinite two-sided ideal* of A, then A/I is a *finite-dimensional algebra*, so $(A/I)^*$ is a *finite-dimensional coalgebra*. Thus

$$A^0 = \bigcup_{I \in C.I.(A)} I^{\perp} \cong \bigcup_{I \in C.I.(A)} (A/I)^*$$

has the *structure of a coalgebra* as the *direct limit* of *finite-dimensional coalgebras*. Of course, if A is finite-dimensional, then $A^0 = A^*$.

REMARK 3.45. Though, for an infinite-dimensional algebra A, A^* will, in general, not be a coalgebra, it may be in case $A^* = A^0$. Heyneman and Radford [Hey-Rad, p. 231, Example 3.1.3(a)] show that, for an infinite-dimensional A, one may, at one extreme, have A^0 as large as possible (namely, $A^0 = A^*$) while, at the other extreme, one may have A^0 as small as possible (namely, $A^0 = 0 := \{0\}$).

For the one extreme, one takes an A with lots and lots of cofinite two-sided ideals; whereas for the other extreme, one takes an A with as few cofinite two-sided ideals as possible. Specifically, to construct an example of an infinite-dimensional algebra A such that $A^0 = A^*$, one takes A to be any infinite-dimensional algebra *having a two-sided ideal* I of *codimension* 1 *whose square is* 0. It is easy to see that, for such an algebra A (with $A = K \oplus I$ and $I^2 = 0$), the *multiplication is necessarily commutative* and that *any subspace of* I *is necessarily a two-sided ideal of* A since, for example, if $S \subseteq I$, then

$$A \cdot S = (K + I) \cdot S \subseteq K \cdot S + I \cdot S \subseteq S + I \cdot I = S + I^2 = S + 0 = S.$$

In [Hey-Rad, p. 231, Example 3.1.3(b)] the authors construct such an A as the dual algebra C^* of a special coalgebra C. In 2.77(c), above, we likewise obtain such an A as C^* for some C. One may, however, easily construct such an A directly as follows: Take V to be any *infinite-dimensional* K-vector space. Endow V with the structure of an *associative algebra without* unit, by declaring all products of elements of V to be 0. Then *adjoin a unit* (as in 4.14) to obtain $A = K \oplus V$ with $(\beta, w) \cdot (\alpha, v) := (\beta\alpha, \beta v + \alpha w)$ for all $\beta, \alpha \in K$ and all $w, v \in V$. One readily checks that (for this A)

(1) $I := 0 \oplus V$ is an *ideal* of A of *codimension* 1 whose *square* is $0_A := (0_K, 0_V)$;
(2) $(1, 0) := (1_K, 0_V) =$ the *two-sided identity element* of A;
(3) the *multiplication* of A is *commutative*; and
(4) *any subspace* of I is an *ideal* of A.

To see that $A^0 = A^*$ *for any algebra A having a two-sided ideal I of codimension* 1 *whose square is* 0, it suffices to show that $A^* \subseteq A^0$ (since, by definition, $A^0 \subseteq A^*$). If $f \in A^*$, then $(\operatorname{Ker} f) \cap I$ being a *subspace* of I is an *ideal* of A. Moreover, $A/(\operatorname{Ker} f) \cong \operatorname{Im} f \subseteq K$, so $\operatorname{codim}(\operatorname{Ker} f) \in \{0, 1\}$. By [Mac-Bir, p. 206, Exercise 10] and [A-M-V-W, p. 44, Exercise 10],

$$\begin{aligned}\operatorname{codim}\big[(\operatorname{Ker} f) \cap I\big] &= \operatorname{codim}(\operatorname{Ker} f) + \operatorname{codim}(I) - \operatorname{codim}(\operatorname{Ker} f + I)\\ &\leqslant \operatorname{codim}(\operatorname{Ker} f) + \operatorname{codim}(I)\\ &\leqslant 1 + 1 = 2,\end{aligned}$$

thereby proving that $f \in A^0$. For the other extreme, take A to be a *field extension* of K which is *infinite-dimensional* as a K-vector space (e.g., $\mathbb{R}$ as a $\mathbb{Q}$-vector space). Since a field has no ideals other than 0 and itself, $A^0 = 0$ since if I is a *cofinite two-sided ideal* of A contained in the kernel of f, for $f \in A^*$, then, necessarily, $I = A$, so $f \equiv 0$.

REMARKS 3.46. (a) It is easy to see (cf. [Sw-1, p. 110, Lemma 6.0.1] or [Dăs-Năs-Rai, p. 34, Lemma 1.5.2]) that if $f: A \to B$ is a *morphism* of **Alg** and if $f^*: B^* \to A^*$ denotes the *transpose* of $f: A \to B$, then $f^*(B^0) \subseteq A^0$ and hence $f^*: B^* \to A^*$ *induces* a map $f^0: B^0 \to A^0$, *the unique K-vector space map* for which the diagram

$$\begin{tikzcd} B^* \arrow[r, "f^*"] & A^* \\ B^0 \arrow[u, "\mathrm{incl}"] \arrow[r, dashed, "f^0"'] & A^0 \arrow[u, "\mathrm{incl}"'] \end{tikzcd}$$

commutes. It is then straightforward (cf. [Dăs-Năs-Rai, p. 37, Proposition 1.5.4]) to show that the induced map $f^0: B^0 \to A^0$ is a *morphism* of **Coalg**. Accordingly, the *contravariant linear dual functor* $(-)^*: \mathbf{Vect} \to \mathbf{Vect}$ induces a *contravariant functor* $(-)^0: \mathbf{Alg} \to \mathbf{Coalg}$.

(b) Using the fact (cf. [Mac-4, p. 86] or [Mi-8, p. 140, Lemma 3.1]) that the *contravariant functor* $(-)^*: \mathbf{Vect} \to \mathbf{Vect}$ is *adjoint to itself on the right*, meaning that, for any object $(V, W) \in (\mathbf{Vect})^{op} \times \mathbf{Vect}$, there is a natural set bijection

$$\operatorname{Hom}_{\mathbf{Vect}}[V, W^*] \cong \operatorname{Hom}_{\mathbf{Vect}}[W, V^*],$$

one can prove that the *contravariant functors*

$$(-)^*: \mathbf{Coalg} \to \mathbf{Alg} \quad \text{and} \quad (-)^0: \mathbf{Alg} \to \mathbf{Coalg}$$

are *adjoint to one another on the right*, meaning that, for any object

$$(A, C) \in (\mathbf{Alg})^{op} \times \mathbf{Coalg},$$

there is a natural set bijection

$$\mathrm{Hom}_{\mathbf{Alg}}[A, C^*] \cong \mathrm{Hom}_{\mathbf{Coalg}}[C, A^0].$$

A sketch of a proof may be found in Heyneman and Sweedler [Hey-Sw-1, p.200, Remark], Sweedler [Sw-1, p. 118, Theorem 6.0.5], Abe [Ab, p. 87, Theorem 2.3.14], or [Dăs-Năs-Rai, p. 44, Theorem 1.5.22].

REMARK 3.47. It turns out that the *contravariant functor* $(-)^0 : \mathbf{Alg} \to \mathbf{Coalg}$ gives rise to a *contravariant functor*, also denoted $(-)^0$, from the *category* **Hopf Alg** of *Hopf algebras* to itself because H^0 happens to be a *Hopf algebra* whenever H is. The idea is this: If H is a *Hopf algebra*, then H is of course an algebra, so H^0 is a coalgebra. Now even as the algebra structure on H gives rise to a coalgebra structure on H^0, we can expect "dually" that the coalgebra structure on H will give rise to an algebra structure on H^0. How? Well, H is a coalgebra, so H^* is an algebra. Furthermore, H^0 sits inside of H^*, so if there is any justice in the world (and of that one certainly has ample opportunity to despair) then H^0 will turn out to be a subalgebra of H^*, and with these two structures H^0 will be a bialgebra and even a Hopf algebra – with the antipode $S : H \to H$ giving rise to an antipode $S^0 : H^0 \to H^0$ induced by the restriction to H^0 of $S^* : H^* \to H^*$. This is indeed the case, and it turns out that the identity element 1_{H^0} of the algebra H^0 is simply the counit $\varepsilon : H \to K$ of the coalgebra H, and that, additionally,

$$1_{H^0} = 1_{H^*} = \varepsilon : H \to K.$$

(cf. [Hey-Sw-1, pp. 205–206, Remark 1], [Sw-1, pp. 122–123, Section 6.2], [N-S, p. 81], [H-5, pp. 228–229], [Mon-1, p. 151, Theorem 9.1.3], [Jos, p. 27, Corollary 1.4.3], [L-R, p. 44, Theorem 1.6.1], [Kli-Sch, p. 22, Theorem 12] or [Dăs-Năs-Rai, pp. 176–177, Solution to Exercise 4.3.9]).

DEFINITION 3.48. If H is a *Hopf algebra*, then the *Hopf algebra* H^0 (denoted $H^\star$ by [Jos, p. 26], $H^\bullet$ by [Kor-Soi, p. 13], or H' by [H-5, p. 229], [Bour-1, p. A III.202, Exercise 27(b)], and [Bour-2, p. 654, Exercise 27(b)]) is variously called *the dual Hopf algebra of H* (cf. [Sw-1, p. 122], [Ab, p. 87], [Mi-8, p. 127, Definition 1.3], and [Kli-Sch, p. 23]), *the dual Hopf algebra to H* (cf. [Kor-Soi, p. 13, Proposition 2.2.6]), *the finite dual* of H (cf. [Mon-1, p. 3, Definition 1.2.3, p. 149] and [Dăs-Năs-Rai, pp. 169, 176, 177, Exercise 4.3.9 and its Solution]), *the Hopf dual of H* (cf. [Char-Pres, p. 13] and [Jos, p. 26]), *the restricted dual of H* (cf. [Kass, pp. 71, 120, 163]), *the restricted dual to H* (cf. [Kor-Soi, p. 13, Proposition 2.2.6]), *the Hopf algebra dual to H* (cf. [H-5, p. 229]), and *the continuous linear dual of H* with respect to the cofinite ideal topology on H and the discrete topology on the ground field K (cf. [H-5, p. 228] and [Mi-8, p. 127, Definition 1.3]).

REMARK 3.49. An example of H^0 was considered by Hochschild in the case where $H = U(L)$, the *universal enveloping algebra* of a *finite-dimensional Lie algebra L. Hochschild's Hopf algebra $R(L)$ of representative functions on $U(L)$*, as defined on p. 514 of [H-1] is precisely $(UL)^0$ in present day notation (cf. [H-1, p. 500], [Hoch, pp. 40–41], [H-5, pp. 229–231], [Dixmier-2, pp. 99–100, Commentaires et compléments 2.8.16, 2.8.17], [Dix, pp. 99–100, Supplementary Remarks 2.8.16, 2.8.17], [Mi-8, p. 152, Remark (4)], [Char-Pres, p. 114, Example 4.1.17], and [Kor-Soi, p. 14, Example 2.2.8; pp. 17–18, §3.1]).

REMARK 3.50. The following result is stated (without proof) on the bottom of p. 87 of Abe [Ab].

THEOREM 3.51. *The* contravariant functor $(-)^0 : \mathbf{Hopf\,Alg} \to \mathbf{Hopf\,Alg}$ *is* adjoint to itself on the right; *i.e., for all Hopf algebras H_1 and H_2, there is a* natural set bijection

$$\mathrm{Hom}_{\mathbf{Hopf\,Alg}}\big[H_1, H_2^0\big] \cong \mathrm{Hom}_{\mathbf{Hopf\,Alg}}\big[H_2, H_1^0\big].$$

[*The* covariant version *of this result reads as follows*: *The functor* $(-)^0 : (\mathbf{Hopf\,Alg})^{op} \to \mathbf{Hopf\,Alg}$ *is* right adjoint *to the* functor $(-)^{0^{op}} : \mathbf{Hopf\,Alg} \to (\mathbf{Hopf\,Alg})^{op}$.]

PROOF. For a proof containing all the gory details see Michaelis [Mi-8, p. 148, Theorem 3.10]. □

REMARK 3.52. As it happens, coalgebras not only have associative incarnations, they also have Lie incarnations. The notion of a *Lie coalgebra* was discovered independently by André [An] and by Michaelis [Mi-1] in the Spring of 1969. By definition a *Lie coalgebra over the monoidal category* $(\mathbf{Vect}, \otimes, K)$ is a *Lie algebra over the opposite monoidal category* $(\mathbf{Vect}^{op}, \otimes^{op}, K^{op})$. Now, just as Lie algebras have Lie coalgebra counterparts, so, too, the *universal enveloping algebra* $U(L)$ of a *Lie algebra* L has its counterpart in the *universal coenveloping coalgebra* $U^c(M)$ of a *Lie coalgebra* M. Further, the *upper-zero* A^0 of an *associative unitary algebra* A has its counterpart in the *upper-zero* L^0 of a *Lie algebra* L.

NOTE. The upper-zero construction for Lie algebras requires care owing to subtleties arising from the fact that *Lie coalgebras, unlike associative coalgebras, need not be locally finite* (cf. [Mi-1, p. 4, Example (I.1.3(d)); p. 31, Corollary (I.2.11)], [Mi-2, pp. 9–10] and [Mi-7, p. 343, Proposition 3; p. 345, Corollary 11; p. 346, Theorem 15]).

It turns out (cf. [Mi-1, pp. 75–77, Theorem (II.2.5)], [Mi-2, pp. 31–32], and [Mi-8, p. 169, Theorem 3.12]) that, for any Lie algebra L,

$$U^c\big(L^0\big) = (UL)^0.$$

This fact ties Lie coalgebras in with *Hochschild's algebra of representative functions* on $U(L)$. More generally, the *contravariant upper zero* functor $(-)^0 : \mathbf{Alg} \to \mathbf{Coalg}$ can be used to construct the coalgebra counterparts T^c, S^c, and U^c of T, S, and U. The fact is that $T(V)$, $S(V)$, and $U(L)$ satisfy *universal mapping properties* which may be *dualized* to obtain $T^c(V)$, $S^c(V)$ and $U^c(M)$ with *corresponding dual* UMP's. Details will be given in [Mi-11], but, in the meantime, the reader may look at [Mi-1], [Mi-2] and [Blo]. At this, juncture, however, we comment briefly on the construction of $T^c(V)$ since that construction is a prototype for the construction of $S^c(V)$ and $U^c(M)$. The construction of $T^c(V)$

that we give follows Sweedler [Sw-1, p. 125, Theorem 6.4.1; p. 128, Lemma 6.4.2] though the notation and terminology that follows is ours and not his. By definition, a *cofree, associative, unitary coalgebra on a K-vector space V* consists of an object $T^c(V)$ of **Coalg** together with a morphism $\pi_{T^c(V)}: F[T^c(V)] \to V$ of **Vect** (where F: **Coalg** $\to$ **Vect** is the *forgetful functor*) such that if C is any object of **Coalg** and if $g: F(C) \to V$ is any morphism of **Vect**, then there is a *unique morphism* $G: C \to T^c(V)$ of **Coalg** making the diagram

$$\begin{tikzcd} V & F[T^c(V)] \arrow[l, "\pi_{T^c(V)}"'] \\ V \arrow[u, equal] & F(C) \arrow[l, "g"] \arrow[u, "F(G)"'] \end{tikzcd}$$

commute. The construction of $T^c(V)$ for an arbitrary K-vector space V as given in [Sweed-2] or [Sw-1] is a two-step process and goes as follows. One first shows that if (TV, i_{TV}) denotes the *tensor algebra on V*, then $(TV)^0$ together with the linear map $(TV)^0 \hookrightarrow (TV)^* \xrightarrow{(i_{TV})^*} V^*$ satisfies the UMP required of a *cofree, associative, unitary coalgebra on the vector space V^**. We abbreviate this fact by writing, simply, $T^c(V^*) = (TV)^0$. Thus, as a first step in the general construction of $T^c(V)$, we see how to construct $T^c(W)$ in case W is the vector space dual V^* of a vector space V. One then shows how to construct $T^c(W)$ whenever W is a subspace of a vector space V for which $T^c(V)$ is known. To do this, one simply observes that the diagram

$$\begin{tikzcd} V & F[T^c(V)] \arrow[l, "\pi_{T^c(V)}"'] \\ W \arrow[u, "\text{incl}"] & \end{tikzcd}$$

may always be "filled in" to yield the *commutative diagram*

$$\begin{tikzcd} V & F[T^c(V)] \arrow[l, "\pi_{T^c(V)}"'] \\ W \arrow[u, "\text{incl}"] & F[T^c(W)] \arrow[l, "\pi_{T^c(W)}"] \arrow[u, "\text{incl}"'] \end{tikzcd}$$

giving a construction of $T^c(W)$ for any subspace W of V. In other words, if $T^c(V)$ is known and if W is a subspace of V, then $T^c(W)$ may be constructed as a certain subspace [indeed *subcoalgebra*] of $T^c(V)$. Specifically, $T^c(W) := \sum D$ where the sum is taken over all *subcoalgebras D* of $T^c(V)$ such that $\pi_{T^c(V)}(D) \subseteq W \subseteq V$. Then $\pi_{T^c(W)}: T^c(W) \to W$ is *defined* to be the K-linear map *induced by the restriction* to $T^c(W)$ of $\pi_{T^c(V)}: T^c(V) \to V$ (cf. [Sweed-2, p. 265, Theorem 1.2], [Sw-1, p. 128, Lemma 6.4.2], or [Dăs-Năs-Rai, p. 51, Lemma 1.6.4; p. 52, Lemma 1.6.5, Theorem 1.6.6]). Since any vector space V *embeds* in its *double dual* V^{**} via the canonical natural map $\chi_V: V \to V^{**}$, given, for all $(v, f) \in V \times V^*$, by $\chi_V(v)(f) := f(v)$, the above two-step procedure yields a construction of $T^c(V)$ for an arbitrary vector space V. Now, even as $T^c(V^*) = (TV)^0$ and $U^c(L^0) = (UL)^0$, so, too, $S^c(V^*) = (SV)^0$.

NOTE. Just as to any *associative algebra* A there is an *associated Lie algebra* $\mathcal{L}(A)$ having the same underlying vector space as A, so, too, to any *associative coalgebra* C there is an *associated Lie coalgebra* $\mathcal{L}^c(C)$ having the same underlying vector space as C. Then, even as the *universal enveloping algebra functor* U is *defined* as a *left adjoint* of $\mathcal{L}$, so too the *universal coenveloping coalgebra functor* U^c is defined as a *right adjoint* of $\mathcal{L}^c$. Similarly, just as the *tensor algebra functor* T is *defined* as the *left adjoint* of the *forgetful functor* $F : \mathbf{Alg} \to \mathbf{Vect}$, and the *symmetric algebra functor* S is *defined* as the *left adjoint* of the *forgetful functor* $F : \mathbf{Comm\,Alg} \to \mathbf{Vect}$ from the *category* of *commutative, associative, unitary algebras* to the *category* of *vector spaces*, so, too, by analogy, we *define* the *tensor coalgebra functor* T^c to be the *right adjoint* of the *forgetful functor* $F : \mathbf{Coalg} \to \mathbf{Vect}$, and we *define* the *symmetric coalgebra functor* S^c to be the *right adjoint* of the *forgetful functor* $F : \mathbf{Comm\,Coalg} \to \mathbf{Vect}$ from the *category* of *commutative, associative, unitary coalgebras* to the *category* of *vector spaces*. Finally, just as the *free Lie algebra functor* L is *defined* as the *left adjoint* of the *forgetful functor* $F : \mathbf{Lie\,Alg} \to \mathbf{Vect}$ from the *category* of *Lie algebras* to the *category* of *vector spaces*, so, too, we *define* the *cofree Lie coalgebra functor* L^c as the *right adjoint* of the *forgetful functor* $F : \mathbf{Lie\,Coalg} \to \mathbf{Vect}$ from the *category* of *Lie coalgebras* to the *category* of *vector spaces*. Moreover both $T^c(V)$ and $S^c(V)$ are special cases of $U^c(M)$; specifically, $S^c(V) = U^c[\mathrm{Triv}(V)]$ where $\mathrm{Triv}(V)$ denotes the vector space V considered trivially as a Lie *coalgebra* [i.e., $\Delta : V \to V \otimes V$ sends all of V to 0], and $T^c(V) = U^c[L^c(V)]$. If we write $Q \dashv R$ to denote the fact that the *covariant functor* $Q : \mathcal{D} \to \mathcal{E}$ is *left adjoint* to the *covariant functor* $R : \mathcal{E} \to \mathcal{D}$ and that the *covariant functor* $R : \mathcal{E} \to \mathcal{D}$ is *right adjoint* to the *covariant functor* $Q : \mathcal{D} \to \mathcal{E}$, i.e., that for each object (D, E) of $\mathcal{D}^{op} \times \mathcal{E}$ there is a *natural set bijection*

$$\mathrm{Hom}_{\mathcal{E}}\big[Q(D), E\big] \cong \mathrm{Hom}_{\mathcal{D}}\big[D, R(E)\big],$$

then the relationships noted above are summarized in the following table:

$$\begin{array}{ccc} T \dashv F & & F \dashv T^c \\ S \dashv F & \text{versus} & F \dashv S^c \\ U \dashv \mathcal{L} & & \mathcal{L}^c \dashv U^c \\ L \dashv F & & F \dashv L^c \\ S(V) = U[\mathrm{Triv}(V)] & & S^c(V) = U^c[\mathrm{Triv}(V)] \\ T(V) = U[L(V)] & & T^c(V) = U^c[L^c(V)]. \end{array}$$

As one would expect, the *universal coenveloping coalgebra* of a *Lie coalgebra* provides us with an important additional example of a *commutative* Hopf algebra. For example, in the case where the *Lie algebra* L is *one-dimensional*, $UL = K[x]$, and one can then show that, as a vector space, $(K[x])^0 \cong$ the K-vector space of all *linearly recursive sequences* (cf. [Fl], [P-T] and [Mon-1, p. 152, Example 9.1.7]). This illustrates yet another way to obtain some interesting examples. Pursuing this a bit further, suppose that L is a 2-*dimensional abelian Lie algebra*. In this case $UL = SL = K[x, y] \cong K[x] \otimes K[y]$, so $(UL)^0 \cong (K[x] \otimes K[y])^0$. Now it turns out (cf. [Hey-Sw-1, p. 197, Lemma 1.3.1(b)], [Sw-1, p. 110, Lemma 6.0.1(b)], [Dăs-Năs-Rai, p. 34, Lemma 1.5.2 (ii)] or [Jos, p. 34, Item 1.4.17]) that, for any *associative, unitary algebras* A and B,

$$(A \otimes B)^0 \cong A^0 \otimes B^0$$

as objects of **Alg**. Thus

$$\big(K[x, y]\big)^0 \cong \big(K[x]\big)^0 \otimes \big(K[y]\big)^0;$$

however, an explicit description of $(K[x, y])^0$ and more generally of $(K[x_1, \dots, x_n])^0$ can get rather messy. For a nice treatment, see Fliess [Fl], Peterson and Taft [P-T], and Taft [Ta-2].

Earlier, we *defined* $A^0 = \bigcup_{I \in C.I.(A)} I^{\perp}$. In [Mi-1, p. 40], [Mi-2, p. 19], [Mi-5, p. 20, Proposition 13], and [Mi-8, p. 126], A^0 is *defined* as the *maximal good subspace* of A^*, a *subspace* V of A^* being *good* in case

$$\varphi^*(V) \subseteq \rho(V \otimes V)$$

where $\varphi : A \otimes A \to A$ is the *multiplication* of A. Equivalently, $V \subseteq A^*$ is a *good subspace* of A^* in case one can *define* a map $\Delta : V \to V \otimes V$ *filling in the diagram*

$$\begin{tikzcd}
A^* \arrow[r, "\varphi^*"] & (A \otimes A)^* \\
 & A^* \otimes A^* \arrow[u, "\rho"'] \\
V \arrow[uu, "\mathrm{incl}"] \arrow[r, dashed, "\Delta"] & V \otimes V. \arrow[u, "\mathrm{incl}"']
\end{tikzcd}$$

It can be shown that V is a *good subspace* of A^* if and only if (V, Δ) is a *coalgebra* (cf. [Mi-1, pp. 32–35] for a proof in the *Lie* case which (proof) is strictly analogous to the proof in the *associative* case). This fact shows that A^0 is the *largest subspace* of A^* carrying a coalgebra structure induced by the restriction to A^0 of the transpose $\varphi^* : A^* \to (A \otimes A)^*$ of the multiplication $\varphi : A \otimes A \to A$ of A, and it is for this reason that it is natural to call A^0 the *dual coalgebra of* (the algebra) A.

NOTE. Consistent with *defining* a (not necessarily associative) *coalgebra* C to be *locally finite* in case each element of C lies in a *finite-dimensional subcoalgebra* of C, we *define* $\mathrm{Loc}(C)$ for *any* (not necessarily associative) *coalgebra* C via setting

$$\mathrm{Loc}(C) := \sum_{\substack{D = \text{fin.dim'l} \\ \text{subcoalgebra of } C}} D.$$

$\mathrm{Loc}(C)$ is obviously also *the sum of all locally finite subcoalgebras of* C and hence is *the largest locally finite subcoalgebra of* C. Clearly, a *coalgebra* C is *locally finite* $\Longleftrightarrow C = \mathrm{Loc}(C)$. It may be shown that for any (associative or Lie) *algebra* A,

$$\mathrm{Loc}\big(A^0\big) = \{f \in A^* \mid \mathrm{Ker}\, f \text{ contains a \textit{cofinite ideal} of } A\}$$

where A^0 is *defined* as the *maximal good subspace* of A and where *ideal* means a *two-sided ideal* if A is an *associative algebra* while *ideal* means a *Lie ideal* if A is a *Lie algebra*. A *proof* of this fact in the case where A is a *Lie algebra* is given in [Mi-1, p. 39, Theorem (I.3.7)]. The proof for the case where A is *associative* is strictly analogous. Of course, if A is an *associative algebra*, then A^0 is an *associative coalgebra*, so by 4.12 and 4.14, below, A^0 is *locally finite* and hence

$$A^0 = \{f \in A^* \mid \mathrm{Ker}\, f \text{ contains a \textit{cofinite two-sided} ideal of } A\}.$$

It may be shown (cf. [Mi-1, p. 41, Proposition (I.3.8)]) that, for any *associative algebra* A, $\mathcal{L}^c(A^0)$ is a *sub Lie coalgebra* of $\mathrm{Loc}[(\mathcal{L}A)^0]$. In general, however, $\mathcal{L}^c(A^0) \subsetneqq \mathrm{Loc}[(\mathcal{L}A)^0]$ as may be seen by taking A to be an infinite-dimensional, commutative, simple algebra (for example, an infinite-dimensional, commutative field extension of the ground field K) and observing that, for any (associative or Lie) *algebra* A, the *elements of* $\mathrm{Loc}(A^0)$ *are in one-to-one correspondence with the finite-dimensional representations* of A since (cf. [Mi-5, p. 20, Lemma 14]) the *cofinite ideals* of A (*two-sided* or *Lie*, as the case may be) are precisely the *kernels* of *finite-dimensional representations* of A.

Returning, now, to the case of A^0, for A an associative unitary algebra, there is another way to introduce A^0 and that is via *representative functions*. This is the approach taken in Abe [Ab, p. 74]; see also [Mon-1, p. 151, Example 9.1.5]. In [Blo-Ler], Block and Leroux generalize the notion of a representative function following which they generalize the construction of A^0 in that given a graded K-algebra A (associative with 1) and a K-vector space V, Block and Leroux construct a coalgebra A^0_V which reduces to A^0 if either $V = K$ or if A is trivially graded (meaning that A is an ungraded K-algebra considered as a graded K-algebra $\{A_n\}_{n=0}^{\infty}$ by concentrating A in degree 0: $A_0 := A$ while $A_n := 0$ for $n > 0$). In [Blo, pp. 277–280], Block gives an alternate construction of $T^c(V)$ based on one in [Blo-Ler] and using the generalization of a representative function given there. By analogy to constructing the underlying vector space of a *free algebra* on V as the *direct sum*, for all integers $n \geqslant 0$, of *n-fold tensor products* of V, Block (*loc.cit.*) constructs, as outlined below in 3.53, a realization of the underlying vector space of a *cofree algebra* on V as a *space embedded in the direct product* (instead of sum), for all integers $n \geqslant 0$, of *n-fold tensor products* of V.

REMARK 3.53. As is well-known (and easy to prove, as a consequence of the UMP's they satisfy), *free* and *cofree objects*, if they exist, are unique up to isomorphism. Now, for any K-vector space V, the tensor algebra, $T(V)$, on V has its name because the product of generators of $T(V)$ is given by the tensor product, $\otimes$. Of course, once one, for example, writes $x_\alpha \cdot x_\beta$ in place of $x_\alpha \otimes x_\beta$, for $x_\alpha, x_\beta \in V$, then $T(V)$ may be identified with $K\langle x_\alpha \rangle$, the K-algebra of all polynomials in the *non-commuting* variables $x_\beta \in \{x_\alpha\}_{\alpha \in I}$

where (as in 3.50), $\{x_\alpha\}_{\alpha \in I}$ is a *K-basis* of V. In this case, one does not, as a rule, refer to $K\langle x_\alpha \rangle$ as the tensor algebra on $\bigoplus_{\alpha \in I} K \cdot x_\alpha$, but rather as the *algebra of all polynomials in the non-commuting variables* $x_\beta \in \{x_\alpha\}_{\alpha \in I}$. Nonetheless, because the tensor algebra on V is one incarnation of the *free, associative, unitary K-algebra on a K-vector space V*, it is natural, by abuse of terminology and notation, to refer to any one of the naturally isomorphic incarnations of the *free, associative, unitary K-algebra on a K-vector space V* simply as the *tensor algebra on V*. Then, by analogy, it is natural to call any one of the naturally isomorphic incarnations of the *cofree, associative, unitary K-coalgebra on a K-vector space V* the *tensor coalgebra on V* and to use the notation $T^c(V)$ for it by analogy to using the notation $T(V)$ for the *tensor algebra on V*. And this is what we have done in the past (cf. [Mi-1], [Mi-2, p. 23]). But, this usage is by no means universal. In fact, the term, the *tensor coalgebra* is often used (for example, by Nichols [N-2, p. 66], Loday [Lod, p. 398, Item A.b], or Kassel [Kass, p. 6, Exercise 3; p. 68, Exercise 5(d)] to refer to the coalgebra on the underlying vector space of the tensor algebra $T(V)$ whose comultiplication and counit (rather than being given as in 3.31, above) are now given as follows: The *comultiplication* $\Delta : T(V) \to T(V) \otimes T(V)$, called *deconcatenation* or the *cut coproduct*, is given by taking $\Delta(1) := 1 \otimes 1$ and by, for $n > 1$, taking any generator $v_1 \otimes \cdots \otimes v_n \in V^{\otimes n}$ apart in all possible ways, namely, by defining

$$\begin{aligned}\Delta(v_1 \otimes \cdots \otimes v_n) := {} & 1 \otimes (v_1 \otimes \cdots \otimes v_n) + v_1 \otimes (v_2 \otimes \cdots \otimes v_n)\\ & + \sum_{p=2}^{n-1} (v_1 \otimes \cdots \otimes v_p) \otimes (v_{p+1} \otimes \cdots \otimes v_n)\\ & + (v_1 \otimes \cdots \otimes v_{n-1}) \otimes v_n + (v_1 \otimes \cdots \otimes v_n) \otimes 1.\end{aligned}$$

The *counit* $\varepsilon : T(V) \to K$ is given by setting

$$\varepsilon(1) = 1 \quad \text{and} \quad \varepsilon(v_1 \otimes \cdots \otimes v_n) = 0 \quad \text{for } n \geqslant 1.$$

It is straightforward to check (cf. [S-S, pp. 41–42]) that Δ and ε, as just defined, endow $T(V)$ with the structure of an *associative unitary coalgebra* (called by some the *cut coalgebra*). With this coalgebra structure the *only group-like element* of $T(V)$ is 1, while the *space of primitive elements* of $T(V)$ is $\bigoplus_{x \in B} K \cdot x$ where B is a *K-basis* of V. It is to be noted (cf. [N-2, p. 66] or [Blo, p. 282]) that the *tensor coalgebra on V* just defined is *not* the *cofree, associative, unitary coalgebra on V*, but rather it is (cf. [Blo, p. 282]) what is known as $[T^c(V)]^1$, the *irreducible component* of $1 = 1 + 0 + 0 + \cdots$ in $T^c(V)$, where $T^c(V)$ is the *cofree, associative, unitary coalgebra on V* (cf. [Q, p. 283, Examples 3.3]).

NOTE. Block [Blo, pp. 277–280] constructs $T^c(V)$ as a certain subspace of $\overline{TV}$, the *completion* of the *tensor algebra* $T(V)$. $\overline{TV}$ is the algebra of all infinite formal sums $\sum_{i=0}^{\infty} a_i$, where $a_i \in T_i(V) := V^{\otimes i}$. Further $\overline{TV}$ is equipped with the usual topology having as a neighborhood basis of the origin, 0, the sets $F_n(\overline{TV}) := \{a \in \overline{TV} \mid a_0 = a_1 = \cdots = a_{n-1} = 0\}$. Upon identifying $T^c(V)$ with a certain subspace of $\overline{TV}$, Block [Blo, p. 282] shows that $[T^c(V)]^1$ equals $T(V)$ with the *cut coproduct* defined above. He further shows that $T(V)$ is the *cofree, pointed, irreducible coalgebra on V*.

NOTE. By definition a *coalgebra* is *pointed* and *irreducible* in case it has a *unique minimal subcoalgebra which is one-dimensional*. Further, to say that $T(V)$ is the *cofree, pointed, irreducible coalgebra on* V is to say that $T(V)$ is *pointed* and *irreducible* and that for any given *pointed, irreducible coalgebra* C and any *linear map* $g : C \to V$ with $f(1) = 0$ (where $1 =$ the *unique group-like element* of C) there exists a *unique map* $G : C \to T(V)$ of **Coalg** such that $\pi \circ G = g$ where $\pi : T(V) \to T_1(V) = V$ is the *natural projection*. Nichols [N-2, p. 66] and Block [Blo, p. 282] further show that $T(V)$ with the above coalgebra structure can be equipped with the structure of an algebra making $T(V)$ a bialgebra, even a Hopf algebra with a suitably defined antipode. The resulting Hopf algebra is what Sweedler [Sw-1, pp. 243–272] calls the *shuffle algebra*. The product of two elements of $T(V)$ is given (cf. Block [Blo, p. 282] or Loday [Lod, p. 238]) by

$$\begin{aligned}
&\varphi\big[(v_1 \otimes \cdots \otimes v_p) \otimes (v_{p+1} \otimes \cdots \otimes v_{p+q})\big] \\
&\quad := \sum_{\substack{p+q=n \\ \sigma=(p,q)\text{-shuffle}}} v_{\sigma^{-1}(1)} \otimes \cdots \otimes v_{\sigma^{-1}(n)} \in v^{\otimes n}.
\end{aligned}$$

NOTE. $T(V)$ with *comultiplication* given by the *cut coproduct* (or *deconcatenation*) and *the usual multiplication on the tensor algebra* (*juxtaposition* or *concatenation*) is *not* a *bialgebra* since, for example, for any $x_1, x_2 \in V$,

$$\begin{aligned}
\Delta(x_1 \cdot x_2) &=: \Delta(x_1 \otimes x_2) \\
&:= 1 \otimes (x_1 \otimes x_2) + x_1 \otimes x_2 + (x_1 \otimes x_2) \otimes 1 \\
&\neq 1 \otimes (x_1 \otimes x_2) + x_2 \otimes x_1 + x_1 \otimes x_2 + (x_1 \otimes x_2) \otimes 1 \\
&:= 1 \otimes (x_1 \cdot x_2) + x_2 \otimes x_1 + x_1 \otimes x_2 + (x_1 \cdot x_2) \otimes 1 \\
&=: (1 \otimes x_1 + x_1 \otimes 1) \cdot (1 \otimes x_2 + x_2 \otimes 1) \\
&=: \Delta(x_1) \cdot \Delta(x_2).
\end{aligned}$$

(See also [Bour-1, p. A III.150, Remarque] or [Bour-2, p. 587, Remark].) – in essence because concatenation and deconcatenation encode the same information.

REMARK 3.54. Historically, the *tensor coalgebra on a vector space* V, considered as $T(V)$ together with the *comultiplication* given by the *cut coproduct*, appeared as the *coalgebra structure* of the *bar construction* (cf. [D-3, pp. 475–478]). For a modern exposition, see [Sta-1, p. 24], [Hain-2, p. 47, Definition 5.13], [Kane, pp. 58–60] or [Fe-Ha-Th-2, p. 849, Example 4.6].

REMARK 3.55. The commutative Hopf algera of K-valued functions $f : G \to K$ on a *finite* group G can be identified with the topological cohomology of the underlying discrete space of G. Similarly, the group algebra $K[G]$, a cocommutative Hopf algebra, can be identified with topological homology of the underlying discrete space of G whether G is finite or not. To see this, recall that for any discrete space X, the homology groups $H_n(X)$ are given by the fact (cf. [Dold, p. 36, Corollary 4.13]) that $H_0(X) = \bigoplus_{x \in X} \mathbb{Z}$ while $H_n(X) = 0$ for all integers $n > 0$. More simply, $H_*(X) = \bigoplus_{x \in X} \mathbb{Z}$. Thus, for any field K, we have $H_*(X;K) = \bigoplus_{x\in X} K$. Taking X to be a *discrete* group G, one finds that $H_*(G;K)$ is isomorphic, as a K-vector space, to the group algebra $K[G]$; i.e., $H_*(G;K) \cong K[G]$. The multiplication of G induces an algebra structure on its homology so that as algebras, in fact as Hopf algebras, we have $H_*(G;K) \cong K[G]$. On the other hand, by the Universal Coefficient Theorem for Cohomology (cf. [Mac-1, p. 78, Corollary 4.2]) $H^*(G;K) \cong \mathrm{Hom}_K(H_*(X);K)$ so, for *finite* G, one has, as Hopf algebras, that

$$H^*(G;K) \cong \mathrm{Hom}_K\big(H_*(X);K\big) \cong \mathrm{Hom}_K\big(K[G];K\big) = \big(K[G]\big)^* = K^G.$$

NOTE. The topological (co)homology of a discrete group is quite different from the group-theoretic (co)homology of that group (cf. [Mac-1, p. 122, Theorem 7.1]).

4. An overview of the theory

Now that we have the definition of a coalgebra and – more importantly – lots of interesting examples, it is natural to inquire how one might begin to build a theory.

As a first approach, we exploit the categorical duality between finite-dimensional coalgebras and finite-dimensional algebras furnished by the linear dual functor

$$*^{op} : \big(\mathbf{Coalg}_{\mathbf{f.d.}}\big)^{op} \to \mathbf{Alg}_{\mathbf{f.d.}}$$

that goes from the opposite category of the category of finite-dimensional coalgebras to the category of finite-dimensional algebras. For it follows from the anti-equivalence of $\mathbf{Coalg}_{\mathbf{f.d.}}$ and $\mathbf{Alg}_{\mathbf{f.d.}}$ that the *subobjects* of $\mathbf{Coalg}_{\mathbf{f.d.}}$ are in one-to-one correspondence with the *quotient objects* of $\mathbf{Alg}_{\mathbf{f.d.}}$ and that the *quotient objects* of $\mathbf{Coalg}_{\mathbf{f.d.}}$ are in one-to-one correspondence with the *subobjects* of $\mathbf{Alg}_{\mathbf{f.d.}}$. But the *quotient algebras* of a given algebra are in one-to-one correspondence with its (two-sided) *ideals*, whereas the *quotient coalgebras* of a given coalgebra are in one-to-one correspondence with its (two-sided) *coideals*. Indeed, just as *ideals* are what one factors algebras by to get *quotient algebras*, so *coideals* are what one factors coalgebras by to get *quotient coalgebras*. These facts suggest that a correspondence should exist between special subspaces of a coalgebra and special subspaces of its dual algebra, and more particularly that to a subcoalgebra of a given coalgebra there should correspond an ideal of the dual algebra, and, conversely, that to a coideal of a coalgebra there should correspond a subalgebra of the dual algebra and conversely. Such a correspondence does, in fact, exist. Schematically, it may be pictured as follows:

$$\text{subcoalgebra} \leftrightarrow \text{ideal} \quad \text{and} \quad \text{coideal} \leftrightarrow \text{subalgebra}.$$

The actual correspondence (which is not necessarily one-to-one in the infinite-dimensional case) is given by the *annihilator transformations* which are defined as follows:

Following Grünenfelder [Gr-1, p. 61] we adopt the following notation.

NOTATION 4.1. For any vector space V, let $s(V)$ denote the lattice of all subspaces of V. Then the *annihilator transformations* are the functions $\perp : s(V) \to s(V^*)$ and $\perp\!\!\!\perp : s(V^*) \to$

$s(V)$ (inclusion reversing) from the lattice, $s(V)$, of subspaces of a vector space V to the lattice, $s(V^*)$, of subspaces of the dual (vector) space, V^*, of V, and from $s(V^*)$ to $s(V)$ defined as follows: if U is a subspace of V and if $i_U : U \hookrightarrow V$ denotes the inclusion, then

$$U^{\perp} := \operatorname{Ker}\big[(i_U)^*\big] = \big\{f \in V^* \mid f(U) = \{0\}\big\} := \big\{f \in V^* \mid f(u) = 0,\ \forall u \in U\big\},$$

while if W is a subspace of V^* and if

$$i_W : W \hookrightarrow V^*$$

denotes the inclusion, then

$$\begin{aligned} W^{\perp\!\!\perp} &:= \operatorname{Ker}\big[(i_W)^* \circ \chi_V\big] = \chi_V^{-1}\big(W^{\perp}\big) = \big\{v \in V \mid W(v) = \{0\}\big\} \\ &:= \big\{v \in V \mid f(v) = 0,\ \forall f \in W\big\} \end{aligned}$$

where $\chi = \chi_V : V \to V^{**}$ is the natural injection of V into its double dual, V^*, defined for all $v \in V$ and for all $f \in V^*$ by $\chi(v)(f) := f(v)$. $U^{\perp}$ is called *the annihilator of* U *in* V^* while $W^{\perp\!\!\perp}$ is called *the annihilator of* W *in* V. We read $U^{\perp}$ as "U-perp," and $W^{\perp\!\!\perp}$ as "W-double-line-perp." Note that $\perp\!\!\perp$ is distinct from $\perp\perp$ which in our notation makes no sense. For us, the usual "double-perp" is either $\perp\ \perp\!\!\perp$ or $\perp\!\!\perp\ \perp$. (Actually, for U a subspace of V, $U^{\perp\perp}$ would make sense as a subspace of V^{**} but that is not what usually is intended since if U is a subspace of V we want $U^{\perp\perp\!\!\perp} \subseteq V$, and, likewise, if W is a subspace of V^*, we want $W^{\perp\!\!\perp\perp} \subseteq V^*$. We could, however, extend our notation as follows: If U is a subspace of V, then right now $U^{\perp} \subseteq V^*$, so we could instead write $U^{\perp(V^*)}$ and call $U^{\perp(V^*)}$ the *annihilator of* U *in* V^*. Likewise, if W is a subspace of V^*, then right now $W^{\perp\!\!\perp} \subseteq V$, so we could instead write $W^{\perp\!\!\perp(V)}$ and call $W^{\perp\!\!\perp(V)}$ the *annihilator of* W *in* V. Given a subspace U of V, we could then consider $U^{\perp(V^*)\perp(V^{**})} \subseteq V^{**}$ or $U^{\perp(V^*)\perp\!\!\perp(V)} \subseteq V$. With this understanding, the abbreviation $U^{\perp\perp}$ for $U^{\perp\perp} \subseteq V^{**}$ makes sense, but, as previously stated, this is not what usually is intended.)

If W is a subspace of V^*, then $W^{\perp\!\!\perp\perp}$ is called the *orthogonal closure* of W and W is said to be *dense* in V^* (or to be a *dense subspace* of V^*) in case $W^{\perp\!\!\perp\perp} = V^*$ or, equivalently (by 4.2(c)2, 4.2(a)5, and 4.2(a)6), in case $W^{\perp\!\!\perp} = \{0_V\}$.

REMARK 4.2. The basic properties of the annihilator transformations, $\perp$ and $\perp\!\!\perp$, are contained in the following list (cf. Abe [Ab], Brieskorn [Bri, pp. 310–312], Grünenfelder [Gr-1, pp. 61–63], Köthe [Köt, pp. 70–72], Lambe and Radford [L-R], and Sweedler [Sw-1]).

(a) Both $\perp$ and $\perp\!\!\perp$ are inclusion reversing, i.e.,

1. If $\{U_1, U_2\} \subseteq s(V)$ and $U_1 \subseteq U_2$, then $U_2^{\perp} \subseteq U_1^{\perp}$.
2. If $\{W_1, W_2\} \subseteq s(V^*)$ and $W_1 \subseteq W_2$, then $W_2^{\perp\!\!\perp} \subseteq W_1^{\perp\!\!\perp}$.
3. If $\{U_1, U_2\} \subseteq s(V)$ and if $U_1 \subseteq U_2$, then $U_1^{\perp\perp\!\!\perp} \subseteq U_2^{\perp\perp\!\!\perp}$.
4. If $\{W_1, W_2\} \subseteq s(V^*)$ and if $W_1 \subseteq W_2$, then $W_1^{\perp\!\!\perp\perp} \subseteq W_2^{\perp\!\!\perp\perp}$.
5. $0^{\perp} := \{0_V\}^{\perp} = V^*$; while $V^{\perp} = \{0_{V^*}\} =: 0$.
6. $0^{\perp\!\!\perp} := \{0_{V^*}\}^{\perp\!\!\perp} = V$; while $(V^*)^{\perp\!\!\perp} = \operatorname{Ker} \chi_V = \{0_V\} =: 0$.

7. For each $U \in s(V)$, $U^{\perp} \cong (V/U)^*$, while $V^*/U^{\perp} \cong U^*$. Hence

$$U \text{ cofinite} \quad \Rightarrow \quad U^{\perp} \text{ is finite-dimensional,}$$

while

$$U^{\perp} \text{ cofinite} \quad \Rightarrow \quad U \quad \text{is finite-dimensional.}$$

8. If $L \in s(V^*)$, then $\chi_V(L^{\perp\!\!\perp}) \subseteq L^{\perp}$, so $\chi_V : V \to V^{**}$ induces a linear map $\chi'_V : L^{\perp\!\!\perp} \to L^{\perp}$ making the following diagram commute. (Here $i_L : L \hookrightarrow V^*$ denotes the inclusion.)

$$\begin{tikzcd}
0 \arrow[r] & L^{\perp} \arrow[r, hook] & V^{**} \arrow[r, "(i_L)^*"] & L^* \\
0 \arrow[r] & L^{\perp\!\!\perp} \arrow[r, hook] \arrow[u, "\chi'_V"'] & V \arrow[r, "(i_L)\circ\chi_V"] \arrow[u, "\chi_V"'] & L^* \arrow[u, equal]
\end{tikzcd}$$

(b)
1. $U = U^{\perp\perp\!\!\perp}$ for any $U \in s(V)$.
2. $W \subseteq W^{\perp\!\!\perp\perp}$ for any $W \in s(V^*)$ with $W = W^{\perp\!\!\perp\perp}$ whenever W is a finite-dimensional subspace of V^*.

(c)
1. $U^{\perp\perp\!\!\perp\perp} = U^{\perp}$ for any $U \in s(V)$.
2. $W^{\perp\!\!\perp\perp\perp\!\!\perp} = W^{\perp\!\!\perp}$ for any $W \in s(V^*)$.

(d)
1. $\sum_{\lambda\in\Lambda}[U_\lambda^{\perp}] \subseteq [\bigcap_{\lambda\in\Lambda} U_\lambda]^{\perp}$ for any $\{U_\lambda\}_{\lambda\in\Lambda} \subseteq s(V)$, with

$$\sum_{\lambda\in\Lambda}[U_\lambda^{\perp}] = \left[\bigcap_{\lambda\in\Lambda} U_\lambda\right]^{\perp}$$

whenever Λ is a finite set.

2. $\sum_{\lambda\in\Lambda}[W_\lambda^{\perp\!\!\perp}] \subseteq [\bigcap_{\lambda\in\Lambda} W_\lambda]^{\perp\!\!\perp}$ for any $\{W_\lambda\}_{\lambda\in\Lambda} \subseteq s(V^*)$.

(e)
1. $\bigcap_{\lambda\in\Lambda}[U_\lambda^{\perp}] = [\sum_{\lambda\in\Lambda} U_\lambda]^{\perp}$ for any $\{U_\lambda\}_{\lambda\in\Lambda} \subseteq s(V)$.
2. $\bigcap_{\lambda\in\Lambda}[W_\lambda^{\perp\!\!\perp}] = [\sum_{\lambda\in\Lambda} W_\lambda]^{\perp\!\!\perp}$ for any $\{W_\lambda\}_{\lambda\in\Lambda} \subseteq s(V^*)$.
3. If $\{W_\lambda\}_{\lambda\in\Lambda} \subseteq s(V^*)$ and if $W_\lambda^{\perp\!\!\perp\perp} = W_\lambda$, $\forall\lambda \in \Lambda$, then

$$\left[\bigcap_{\lambda\in\Lambda} W_\lambda\right]^{\perp\!\!\perp\perp} = \bigcap_{\lambda\in\Lambda} W_\lambda.$$

4. If $V = U_1 \oplus U_2$ for $U_1, U_2 \in s(V)$, then

$$V^* = U_1^{\perp} \oplus U_2^{\perp}.$$

(f)

1. If $f: V \to W$ is a K-vector space map, then, for any $I \in s(V)$, $[f(I)]^{\perp} = (f^*)^{-1}(I^{\perp})$ and so $f(I) = [(f^*)^{-1}(I^{\perp})]^{\perp\!\!\perp}$. In particular, $\mathrm{Im}(f) = [\mathrm{Ker}(f^*)]^{\perp\!\!\perp}$.
2. If $f: V \to W$ is a K-vector space map, then, for any $J \in s(W)$,

$$f^{-1}(J) = \left[f^*\left(J^{\perp}\right)\right]^{\perp\!\!\perp}.$$

In particular, $\mathrm{Ker}(f) = [\mathrm{Im}(f^*)]^{\perp\!\!\perp}$.

3. If $f: V \to W$ is a K-vector space map, then, for any $J \in s(W)$,

$$f^*\left(J^{\perp}\right) = \left[f^{-1}(J)\right]^{\perp}.$$

4. If $f: V \to W$ is a K-vector space map, then, for any $A \in s(W^*)$,

$$\left[f^*(A)\right]^{\perp\!\!\perp} = f^{-1}\left(A^{\perp\!\!\perp}\right).$$

5. If $\{I, J\} \subseteq s(V)$ and $I \subseteq J$, then

$$I \text{ cofinite} \quad \Rightarrow \quad J \text{ cofinite}.$$

6. If $U \in s(V)$ with $i: U \hookrightarrow V$ the inclusion; if $F := i^*: V^* \to U^*$; and if $U^{\perp} = I \subseteq J$ for some $J \in s(V^*)$, then $J = F^{-1}[F(J)]$.

(g) Let U, V and W be K-vector spaces; let σ, σ_1, $\sigma_2: U \to V$ and $\phi: W \to V^*$ be K-linear maps; and let $\phi(W)$ be a *dense* subspace of V^* (i.e., let $[\phi(W)]^{\perp\!\!\perp} = \{0_V\}$). Then

1. If $\sigma^* \circ \phi = 0: W \xrightarrow{\phi} V^* \xrightarrow{\sigma^*} U^*$, then $\sigma \equiv 0: U \to V$ (and so $\sigma^* \equiv 0: V^* \to U^*$).
2. If

$$\sigma_1^* \circ \phi = \sigma_2^* \circ \phi: W \xrightarrow{\phi} V^* \overset{\sigma_1^*}{\underset{\sigma_2^*}{\rightrightarrows}} U^*,$$

then $\sigma_1 = \sigma_2: U \to V$ (and so $\sigma_1^*: V^* = \sigma_2^* \to U^*$).

(h)

1. Let V be a K-vector space, and let $\chi_V: V \to V^{**}$ be the canonical injection of V into its double dual (given, for all $v \in V$ and $f \in V^*$, by $\chi_V(v)(f) := v^{**}(f) := f(v)$). Then $\chi_V(V)$ is a dense subspace of V^{**}, i.e., $[\chi_V(V)]^{\perp\!\!\perp} = \{0_{V^*}\}$.
2. Let $f: V \to W$ be a K-vector space *injection*, then $f^*: W^* \to V^*$ carries dense subspaces of W^* to dense subspaces of V^*, i.e., if A is a subspace of W^* such that $A^{\perp\!\!\perp} = \{0_W\}$, then $[f^*(A)]^{\perp\!\!\perp} = \{0_V\}$.
3. Let U, V and W be K-vector spaces and let $\rho_{U,V}: U^* \otimes V^* \to (U \otimes V)^*$ and

$$\rho_{U,V,W}: U^* \otimes V^* \otimes W^* \to (U \otimes V \otimes W)^*$$

be the natural K-linear injections defined, for all $f \in U^*$, $g \in V^*$, $h \in W^*$ and for all $u \in U$, $v \in V$, $w \in W$, respectively, by

$$\rho_{U,V}(f \otimes g)(u \otimes v) := f(u) \cdot g(v) \in K$$

and by

$$\rho_{U,V,W}(f \otimes g \otimes h)(u \otimes v \otimes w) := f(u) \cdot g(v) \cdot h(w) \in K.$$

Then $\rho_{U,V}(U^* \otimes V^*)$ is *dense* in $(U \otimes V)^*$ and $\rho_{U,V,W}(U^* \otimes V^* \otimes W^*)$ is *dense* in $(U \otimes V \otimes W)^*$, i.e.,

$$\left[\rho_{U,V}(U^* \otimes V^*)\right]^{\perp\!\!\perp} = \{0_{U \otimes V}\}$$

and

$$\left[\rho_{U,V,W}(U^* \otimes V^* \otimes W^*)\right]^{\perp\!\!\perp} = \{0_{U \otimes V \otimes V \otimes W}\}.$$

4. Let U, V and W be K-vector spaces, and suppose that $P \subseteq U^*$ is a subspace of U, that $Q \subseteq V^*$ is a subspace of V^*, and that $R \subseteq W^*$ is a subspace of W^*. Then, with reference to the K-linear injections

$$P \otimes Q \hookrightarrow U^* \otimes V^* \xrightarrow{\rho_{U,V}} (U \otimes V)^*$$

and

$$P \otimes Q \otimes R \hookrightarrow U^* \otimes V^* \otimes W^* \xrightarrow{\rho_{U,V,W}} (U \otimes V \otimes W)^*,$$

we have that

$$\left[\rho_{U,V}(P \otimes Q)\right]^{\perp\!\!\perp(U \otimes V)} = P^{\perp\!\!\perp(U)} \otimes V + U \otimes Q^{\perp\!\!\perp(V)}$$

and that

$$\begin{aligned}&\left[\rho_{U,V,W}(P \otimes Q \otimes R)\right]^{\perp\!\!\perp(U \otimes V \otimes W)} \\ &\quad = P^{\perp\!\!\perp(U)} \otimes V \otimes W + U \otimes Q^{\perp\!\!\perp(V)} \otimes W + U \otimes V \otimes R^{\perp\!\!\perp(W)},\end{aligned}$$

or, more briefly, that

$$\left[\rho_{U,V}(P \otimes Q)\right]^{\perp\!\!\perp} = P^{\perp\!\!\perp} \otimes V + U \otimes Q^{\perp\!\!\perp}$$

and that

$$\left[\rho_{U,V,W}(P \otimes Q \otimes R)\right]^{\perp\!\!\perp} = P^{\perp\!\!\perp} \otimes V \otimes W + U \otimes Q^{\perp\!\!\perp} \otimes W + U \otimes V \otimes R^{\perp\!\!\perp}.$$

The proofs of all these statements, though not always trivial, are straightforward. They are available as a separate file from the author or editor.

REMARK 4.3. There is a hidden yet present topological structure on the vector space dual of any K-vector space that accounts for what is going on in 4.2(g) from the perspective of topology, and it is to an exposition of this that we now direct our attention. The fact of the matter is that the vector space dual V^* of any K-vector space V always carries a natural topology, the so-called *weak-star* topology, relative to which the *transpose* $f^*: W^* \to V^*$ of any K-linear map $f: V \to W$ is *continuous*, and relative to which a subspace W of V^* is *dense* (in V^*) if and only if $W^{\perp\!\!\perp} = 0 =: \{0_V\}$ (or equivalently, by 4.2(a)5, 4.2(a)6, and 4.2(c)2, if and only if $W^{\perp\!\!\perp\perp} = V^*$). Then (if we take $\phi: W \hookrightarrow V^*$ to be the inclusion, or, alternatively, replace W by $\phi(W)$ and (then) ϕ by the inclusion $\phi(W) \hookrightarrow V^*$), 4.2(g)1 simply states that if a continuous function $\sigma^*: V^* \to U^*$ is zero on a *dense* subspace of V^*, it is zero everywhere; while 4.2(g)2 just says that if two continuous functions agree on a dense subspace of a topological space then they agree all over. By definition, the *weak-star topology* on the vector space dual of a K-vector space V is the weakest (i.e., smallest) topology (the one contained in all others) on V^* relative to which all elements of V considered as elements of V^{**} are continuous when K has the discrete topology, i.e., it is the smallest topology on V^* relative to which all elements of V^{**} of the form x^{**} for $x \in V$ are continuous (where $x^{**} := \chi_V(x) \in V^{**}$ is defined for all $x \in V$ and for all $f \in V^*$ via $x^{**}(f) = f(x) \in K$). Since $x^{**}: V^* \to K$ is K-linear, $x^{**}(0_{V^*}) = 0_K$, and since K has the discrete topology, $\{0_K\}$ is open (and closed) in K, so $(x^{**})^{-1}(\{0_K\})$ is open (and closed) in V^* for each $x \in V$. Now

$$\begin{aligned}(x^{**})^{-1}\big(\{0_K\}\big) &= \big\{f \in V^* \mid 0_K =: 0 = x^{**}(f) = f(x)\big\}\\ &= \big\{f \in V^* \mid f(x) = 0_K\big\}.\end{aligned}$$

Hence, if U is a finite-dimensional subspace of V with *basis* $\{e_1, \ldots, e_n\}$, then

$$\begin{aligned}\bigcap_{i=1}^{n}(e_i^{**})^{-1}\big(\{0_K\}\big) &= \big\{f \in V^* \mid f(e_1) = \cdots = f(e_n) = 0_K\big\}\\ &= \big\{f \in V^* \mid f(u) = 0_K,\ \forall u \in U\big\}\\ &= U^{\perp}.\end{aligned}$$

Since each $(e_i^{**})^{-1}(\{0_K\})$ is both open and closed in V^*, and since in any topological space the intersection of finitely many open sets is again open while the intersection of arbitrarily many closed sets is again closed, it follows that any subspace of V^* of the form $U^{\perp}$, where U is a finite-dimensional subspace of V, is both open and closed in V^*. Further, if U is an infinite-dimensional subspace of V with basis $\{e_i\}_{i \in I}$, for some index set I, then

$$\begin{aligned}\bigcap_{i\in I}(e_i^{**})^{-1}\big(\{0_K\}\big) &= \big\{f \in V^* \mid f(e_i) = 0_K,\ \forall i \in I\big\}\\ &\subseteq \big\{f \in V^* \mid f(u) = 0_K,\ \forall u \in U\big\}\\ &\subseteq \bigcap_{i\in I}(e_i^{**})^{-1}\big(\{0_K\}\big)\end{aligned}$$

whence

$$\begin{aligned}\bigcap_{i\in I}(e_i^{**})^{-1}\big(\{0_K\}\big) &= \big\{f \in U^* \mid f(u) = 0_K,\ \forall u \in U\big\}\\ &= U^{\perp}\end{aligned}$$

thereby proving that subspaces of V^* of the form $U^\perp$, where U is an arbitrary subspace of V, are *closed*. We have thus shown that in the *weak-star* topology on the vector space dual V^* of a K-vector space V, the *annihilator* $U^\perp$ of any *finite-dimensional* subspace U of V is *open* in V^*, whereas the *annihilator* of an(y) *arbitrary subspace* of V is *closed* in V^*. (In particular, the annihilator of any finite-dimensional subspace of V is closed as well as open in V^* which is as it should be since (as we'll review below) in any topological group, open sets are automatically closed.) We shall now show that conversely, in the weak-star topology on V^*, each closed subspace of V^* is always the annihilator of an associated subspace of V whereas each open subspace of V^* is always the annihilator of an associated finite-dimensional subspace of V. In fact, we'll show that the claim about the form of the open subspaces of V^* follows from that about the closed subspaces of V^*.

We now turn to the matter of showing that each closed subspace of V^* is necessarily of the form $U^\perp$ for some subspace U of V. For this it will be enough to show that a subspace W of V^* is closed in the weak-star topology on V^* if and only if it is (what is called) *orthogonally closed*, i.e., if and only if $W = W^{\perp\!\!\!\perp\perp}$, since, from the foregoing discussion, each orthogonally closed subspace W of V^* is necessarily closed in the weak-star topology (being the "$\perp$" of a subspace, namely, $W^{\perp\!\!\!\perp}$, of V). But, since a subspace W of V^* is closed in the weak-star topology if and only if $W = \overline{W}$ where $\overline{W} =$ the closure of W in the weak-star topology, it is, therefore, enough (in order to show that any closed subspace of V^* is the annihilator of some subspace of V) to show that

$$\overline{W} = W^{\perp\!\!\!\perp\perp},$$

i.e., that $\overline{W}$, the closure of W in the weak-star topology, coincides with $W^{\perp\!\!\!\perp\perp}$, the (so-called) *orthogonal closure* of W. Before embarking on this, a few words are in order.

By definition (cf. Köthe [Köt, p. 71]) a subspace W of V^* is said to be *orthogonally closed* iff $W = W^{\perp\!\!\!\perp\perp}$. By 4.2(e)3, the intersection of all orthogonally closed subspaces W_λ of V^* containing a given subspace W of V^* is again an orthogonally closed subspace of V^* containing W and is, therefore, the smallest such. On the other hand, if M is an orthogonally closed subspace of V^* containing W, then $M^{\perp\!\!\!\perp\perp} = M$ and also $W \subseteq M$, so (by 4.2(b)2 and 4.2(a)3, respectively)

$$W \subseteq W^{\perp\!\!\!\perp\perp} \subseteq M^{\perp\!\!\!\perp\perp} = M.$$

Further (by 4.2(c)2),

$$\left(W^{\perp\!\!\!\perp\perp}\right)^{\perp\!\!\!\perp\perp} = \left(W^{\perp\!\!\!\perp\perp\perp\!\!\!\perp}\right)^{\perp} = W^{\perp\!\!\!\perp\perp}$$

so $W^{\perp\!\!\!\perp\perp}$ is an orthogonally closed subspace of V^* containing every other orthogonally closed subspace of V^* containing W, and hence is the smallest orthogonally closed subspace of V^* containing W. For this reason, for any subspace W of V^*, one calls $W^{\perp\!\!\!\perp\perp}$ the *orthogonal closure* of W.

Now, with regard to proving that $\overline{W} = W^{\perp\!\!\perp\perp}$ for any subspace W of V^*, observe, first, that

$$\overline{W} \subseteq W^{\perp\!\!\perp\perp}$$

since (1) $W^{\perp\!\!\perp\perp}$ is closed in the weak-star topology on V^* (being the annihilator of a subspace, viz., $W^{\perp\!\!\perp}$, of V), (2) $W^{\perp\!\!\perp\perp} \supseteq W$, and (3) by definition, $\overline{W} =$ the smallest weak-star closed subset of V^* containing W.

REMARK. By Köthe [Köt, p. 84, Item (6.a)], the closure of a linear subspace W is again a linear subspace: For if $g_0, h_0 \in \overline{W}$, then for each linear neighborhood U of 0 there are elements $g, h \in W$ for which $g \in g_0 + U$ and $h \in h_0 + U$. But then, for any $\alpha, \beta \in K$,

$$\alpha \cdot g + \beta \cdot h \in \alpha \cdot g_0 + \beta \cdot h_0 + U$$

whence $\alpha \cdot g_0 + \beta \cdot h_0 \in \overline{W}$. We may, therefore, without loss of generality, think of W as the smallest weak-star closed *linear subspace* (as opposed to *subset*) of V^* containing W.

To establish the reverse inclusion, i.e., to show that $W^{\perp\!\!\perp\perp} \subseteq \overline{W}$, we shall need to make use of the fact (requiring a proof which we'll furnish below) that any element of V^{**} that is continuous as a function from V^* to K, when V^* has the weak-star topology and K the discrete topology, is necessarily in the image $\chi_V(V)$ of V under the natural embedding $\chi_V : V \to V^{**}$ given, for all $x \in V$ and all $f \in V^{**}$, by $\chi_V(x)(f) := x^{**}(f) := f(x)$. Of course, by the definition of the weak-star topology on V^*, any element of V^{**} lying in $\chi_V(V)$ is automatically continuous. Here, the claim is that in fact more is true, specifically, that an element of V^{**} is continuous with respect to the weak-star topology on V^* and the discrete topology on K if and only if it comes from an element of V considered as an element of V^{**} under the evaluation map $\chi_V : V \to V^{**}$, i.e., if and only if it lies in the image of $\chi_V : V \to V^{**}$. This fact accounts for the fact (and is, in fact, equivalent to it) that the category of profinite vector spaces (= the category of linearly compact vector spaces, of which more will be said below) is the dual category of the category of K-vector spaces. Now, once one has the result that $\overline{W} = W^{\perp\!\!\perp\perp}$ for any subspace W of V^*, it is then immediate (as noted earlier) that a subspace W of V^* is closed if and only if it is the "$\perp$" of a subspace U of V, viz., $U = W^{\perp\!\!\perp}$, since, more generally, a subspace of a topological space is closed if and only if it coincides with its closure in the given topology. Then, to show that a subspace W of V^* is open if and only if it is the "$\perp$" of a finite-dimensional subspace U of V, one uses the above result (about closed subspaces of V^*) together with the fact that any open subgroup of a topological group is automatically, also, closed.

NOTE. If V is any K-vector space, then V^* with the *weak-star* topology is a profinite (or *linearly compact* vector space). The idea is this: Write V as a direct limit of (all) its finite-dimensional subspaces:

$$V = \varinjlim U_\alpha$$

where U_α runs through the direct set of the finite-dimensional subspaces of V. With U_α a finite-dimensional subspace of V, we have the exact sequences

$$0 \to U_\alpha \to V \to V/U_\alpha \to 0,$$
$$0 \to (V/U_\alpha)^* \to V^* \to U_\alpha^* \to 0,$$

and

$$0 \to U_\alpha^\perp \to V^* \to U_\alpha^* \to 0.$$

From the *direct* (= *inductive*) *limit*

$$V = \varinjlim U_\alpha$$

one obtains the *inverse* (= *projective*) *limit*

$$V^* = \varprojlim U_\alpha^* = \varprojlim V^*/U_\alpha^\perp.$$

When V has, additionally, the structure of a coalgebra (associative, with counit) over a field K and we write C as the direct limit of its finite-dimensional subcoalgebras (cf. Theorem 4.12, below) then C^* becomes the *projective limit* of *finite-dimensional algebras*, hence the name "profinite algebra," cf. Serre [Se-3, pp. 43–45] and [Se-4, pp. 282–284] as well as the other references given in Remark 2.95, above. As noted in that remark, the categorical duality given by

$$(-)^* : \mathbf{Vect} \rightleftarrows \mathbf{Profinite\ Vect} : (-)'$$

induces a categorical duality

$$(-)^* : \mathbf{Coalg} \rightleftarrows \mathbf{Profinite\ Alg} : (-)'$$

known (cf. [D-2, pp. 18–22]) as *Cartier duality*.

REMARK. From the fact that $\overline{W} = W^{\perp\!\!\perp\perp}$ for any K-vector subspace W of V^* it follows that the *closed* linear subspaces of V^* in the *weak-star* topology (those $W \subseteq V^*$ for which $W = \overline{W}$:= the weak-star closure of W) are just the *orthogonally closed* linear subspaces of V^* (those $W \subseteq V^*$ such that $W = W^{\perp\!\!\perp\perp}$) for if W is a subspace of V^* and $\overline{W} = W^{\perp\!\!\perp\perp}$, then, obviously

$$W = \overline{W} \quad \Leftrightarrow \quad W = W^{\perp\!\!\perp\perp}.$$

Consequently, a K-vector subspace $W \subseteq V^*$ is closed in the weak-star topology on $V^* \Leftrightarrow W = U^\perp$ for some K-vector subspace $U \subseteq V$. Can one say something in the case of the open subspaces of V^*? Well, we saw earlier that each K-vector subspace of V^* of the form $U^\perp$ where U is a finite-dimensional K-vector subspace of V is open in the weak-star topology on V^*. Conversely, if a K-vector subspace W of V^* is *open* in the weak-star topology on V^*, then necessarily $W = U^\perp$ for some *finite-dimensional* K-vector subspace $U \subseteq V$.

NOTE. In developing 4.3, I benefited from helpful conversations with K.H. Hofmann and W.H. Schikhof.

REMARK. In Remark 2.59, we said that the assignment given at the object level by $C \mapsto C^*$ induces a (covariant) functor

$$(-)^* : (\mathbf{Coalg})^{op} \to \mathbf{Alg}.$$

This fact follows at once from the functoriality of

$$(-)^* : (\mathbf{Vect})^{op} \to \mathbf{Vect}$$

and the following theorem.

THEOREM 4.4. (a) *If V is a K-vector space and if*

$$\Delta : V \to V \otimes V \quad \textit{and} \quad \varepsilon : V \to K$$

are K-linear maps, then

$$(V, \Delta, \varepsilon) \text{ is a coalgebra} \quad \Leftrightarrow \quad \big(V^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda\big) \text{ is an algebra.}$$

(b) *If V is a finite-dimensional K-vector space and if*

$$\varphi : V \otimes V \to V \quad \textit{and} \quad \eta : K \to V$$

are K-linear maps, then

$$(V, \varphi, \eta) \text{ is an algebra} \quad \Leftrightarrow \quad (V^*, \rho^{-1} \circ \varphi^*, \lambda^{-1} \circ \eta^*) \text{ is a coalgebra.}$$

(c) *If $f : V \to W$ is a morphism of* **Vect**, *then*

$$f : (V, \Delta_V, \varepsilon_V) \to (W, \Delta_W, \varepsilon_W) \text{ is a morphism of } \mathbf{Coalg} \quad \Leftrightarrow$$
$$f^* : \big(W^*, \Delta_W^* \circ \rho_W, \varepsilon_W^* \circ \lambda\big) \to \big(V^*, \Delta_V^* \circ \rho_V, \varepsilon_V^* \circ \lambda\big) \text{ is a morphism of } \mathbf{Alg}.$$

(d) *If V and W are finite-dimensional K-vector spaces, and if $f : V \to W$ is a morphism of* **Vect**, *then*

$$f : (V, \varphi_V, \eta_V) \to (W, \varphi_W, \eta_W) \text{ is a morphism of } \mathbf{Alg}_{\mathbf{f.d.}} \quad \Leftrightarrow$$
$$f^* : \big(W^*, \rho_W^{-1} \circ \varphi_W^*, \lambda^{-1} \circ \eta_W^*\big) \to \big(V^*, \rho_V^{-1} \circ \varphi_V^*, \lambda^{-1} \circ \eta_V^*\big)$$
$$\text{is a morphism of } \mathbf{Coalg}_{\mathbf{f.d.}}.$$

The following two propositions are basic. Their proofs are straightforward so we shall exhibit only one of them as a sample. For a proof of the other proposition, the reader is referred to [Sw-1, pp. 16–17, Proposition 1.4.3] or to [Ab, p. 78, Theorem 2.3.1].

PROPOSITION 4.5. *Let* (C, Δ, ε) *be a coalgebra and let* $(C^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda)$ *be its dual algebra. Then the following results hold*:

(a) *If* D *is a subcoalgebra of* C, *then* $D^\perp$ *is an ideal of* C^*.
(b) *If* I *is an ideal of* C^*, *then* $I^{\perp\!\!\perp}$ *is a subcoalgebra of* C.
(c) D *is a subcoalgebra of* $C \Leftrightarrow D^\perp$ *is an ideal of* C^*. *In this case there exists an algebra isomorphism* $C^*/D^\perp \cong D^*$.

PROPOSITION 4.6. *Let* (C, Δ, ε) *be a coalgebra and let* $(C^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda)$ *be its dual algebra. Then the following results hold*:

(a) *If* J *is a coideal of* C, *then* $J^\perp$ *is a subalgebra of* C^*.
(b) *If* B *is a subalgebra of* C^*, *then* $B^{\perp\!\!\perp}$ *is a coideal of* C.
(c) J *is a coideal of* $C \Leftrightarrow J^\perp$ *is a subalgebra of* C^*.

PROOF OF 4.6(a). Assume that J is a coideal of C; let $i : J \hookrightarrow C$ be the inclusion; and let $\pi : C \to C/J$ be the natural projection. Then

$$J^\perp := \mathrm{Ker}(i^*) = \mathrm{Im}(\pi^*).$$

Since J is a coideal, π is a map of coalgebras, so π^* is a map of algebras. Thus $J^\perp$ is a subalgebra of C^*, being the image of an algebra under an algebra map.

(b) Assume B is a subalgebra of C^*. To show that $B^{\perp\!\!\perp}$ is a coideal of C, we must show two things, namely, that

(1) $\Delta(B^{\perp\!\!\perp}) \subseteq B^{\perp\!\!\perp} \otimes C + C \otimes B^{\perp\!\!\perp}$

and that

(2) $\varepsilon(B^{\perp\!\!\perp}) = 0$.

The second condition is easily dispensed with as follows: B is a subalgebra of C^*, and, therefore, $1_{C^*} \in B$. But (cf. 2.58(a)),

$$1_{C^*} = \varepsilon : C \to K.$$

Consequently, $\varepsilon \in B$, and, therefore, $\varepsilon(B^{\perp\!\!\perp}) = 0$ since

$$B^{\perp\!\!\perp} = \{c \in C \mid B(c) = 0\}.$$

To establish the first condition, let $x \in B^{\perp\!\!\perp}$. We must show that

$$\Delta(x) \in B^{\perp\!\!\perp} \otimes C + C \otimes B^{\perp\!\!\perp}.$$

If $B^{\perp\!\!\perp} = C$, we are done; otherwise $B^{\perp\!\!\perp} \subsetneqq C$ and therefore

$$C = B^{\perp\!\!\perp} \oplus A$$

for some complementary subspace A. Then

$$\Delta(x) = \sum_{i=1}^{n} x_i \otimes (y_i + a_i) \quad \text{with } y_i \in B^{\perp\!\!\perp} \text{ and } a_i \in A,$$

so

$$\Delta(x) - \sum_{i=1}^{n} x_i \otimes a_i = \sum_{i=1}^{n} x_i \otimes y_i \in C \otimes B^{\perp\!\!\perp}.$$

To complete the proof, it suffices to show that

$$\sum_{i=1}^{n} x_i \otimes a_i \in B^{\perp\!\!\perp} \otimes C.$$

To do so, we let $\{e_k\}_{k=1}^{m}$ be a set of *linearly independent elements* of A such that, for each $i \in \{1, \ldots, n\}$,

$$a_i = \sum_{k=1}^{m} \lambda_{ik} \cdot e_k \quad \text{with } \lambda_{ik} \in K.$$

Then

$$\sum_{i=1}^{n} x_i \otimes a_i = \sum_{k=1}^{m} u_k \otimes e_k, \quad \text{where } u_k = \sum_{i=1}^{n} \lambda_{ik} \cdot x_i.$$

We claim that each $u_k \in B^{\perp\!\!\perp}$ or, equivalently, that, for each $k \in \{1, \ldots, m\}$

$$f(u_k) = 0, \quad \text{for all } f \in B.$$

To prove this, let $f \in B$ be given, and set

$$v_f := \sum_{k=1}^{m} f(u_k) \cdot e_k.$$

Since B is a subalgebra of C^* and $f \in B \subset C^*$, we have that $f * g \in B$ for all $g \in B$; hence $f * g(x) = 0$ for all $g \in B$ (because $x \in B^{\perp\!\!\perp} \subset C$). But

$$f * g := \varphi_K \circ (f \otimes g) \circ \Delta_C = \varphi_K \circ (f \otimes g) \circ \Delta$$

where $\varphi_K : K \otimes K \xrightarrow{\cong} K$ is the multiplication of the ground field K. Thus, for all $g \in B$,

$$0 = f * g(x) = \varphi_K \circ (f \otimes g)[\Delta(x)]$$

$$
\begin{aligned}
&= \varphi_K \circ (f \otimes g)\left[\sum_{i=1}^{n} x_i \otimes y_i + \sum_{i=1}^{n} x_i \otimes a_i\right] \\
&= \varphi_K \circ (f \otimes g)\left[\sum_{i=1}^{n} x_i \otimes y_i + \sum_{k=1}^{m} u_k \otimes e_k\right] \\
&= \sum_{i=1}^{n} f(x_i) \cdot g(y_i) + \sum_{k=1}^{m} f(u_k) \cdot g(e_k) \\
&= \sum_{i=1}^{n} f(x_i) \cdot 0 + \sum_{k=1}^{m} f(u_k) \cdot g(e_k) \\
&= g\left[\sum_{k=1}^{m} f(u_k) \cdot e_k\right] = g(v_f).
\end{aligned}
$$

Since $g(v_f) = 0$ for all $g \in B$, it follows that $v_f \in B^{\perp\!\!\perp}$. But

$$
v_f = \sum_{k=1}^{m} f(u_k) \cdot e_k \in A,
$$

so $v_f = 0$ (since $C = B^{\perp\!\!\perp} \oplus A$ and hence $0 = B^{\perp\!\!\perp} \cap A$). Since $e_1, \ldots, e_m$ are linearly independent, this means that $f(u_k) = 0$ for all k; thus $u_k \in B^{\perp\!\!\perp}$ for each k. Hence

$$
\sum_{i=1}^{n} x_i \otimes a_i = \sum_{k=1}^{m} u_k \otimes e_k \in B^{\perp\!\!\perp} \otimes C
$$

as was to be shown.

(c) This follows from (a) and (b) since $J^{\perp\perp\!\!\perp} = J$. □

By using the properties of the annihilator transformations in conjunction with the two propositions above, one can show, similarly, that the facts listed in the following two propositions correspond to familiar facts about algebras if one interchanges "$\bigcap$" and "$\sum$", "subcoalgebra" and "ideal", "coideal" and "subalgebra", and "image" and "preimage".

PROPOSITION 4.7. *Let C be a coalgebra, let $\{D_\lambda\}_{\lambda \in \Lambda}$ be a set of subcoalgebras of C, and let $\{I_\lambda\}_{\lambda \in \Lambda}$ be a set of coideals of C. Then the following results hold*:

(a) $\sum_{\lambda \in \Lambda} D_\lambda$ *is a subcoalgebra of C.*
(b) $\sum_{\lambda \in \Lambda} I_\lambda$ *is a coideal of C.*
(c) $\bigcap_{\lambda \in \Lambda} D_\lambda$ *is a subcoalgebra of C.*
(d) *In general, $\bigcap_{\lambda \in \Lambda} I_\lambda$ is* not *a coideal of C.*

PROPOSITION 4.8. *Let $f : C_1 \to C_2$ be a morphism of* **Coalg**. *Then the following results hold*:

(a) *If D_1 is a subcoalgebra of C_1, then $f(D_1)$ is a subcoalgebra of C_2.*

(b) *If I is a coideal of C_1, then $f(I)$ is a coideal of C_2.*
(c) *If J is a coideal of C_2, then $f^{-1}(J)$ is a coideal of C_1.*
(d) *If D_2 is a subcoalgebra of C_2, and if $f: C_1 \to C_2$ is injective, then $f^{-1}(D_2)$ is a subcoalgebra of C_1. In general, however, $f^{-1}(D_2)$ is not a subcoalgebra of C_1.*

PROOF OF 4.7(a). *Method* 1.

$$\begin{aligned}\Delta\Big(\sum_\lambda D_\lambda\Big) &\subseteq \sum_\lambda \Delta(D_\lambda) \subseteq \sum_\lambda (D_\lambda \otimes D_\lambda)\\ &\subseteq \sum_\lambda \Big[\Big(\sum_\mu D_\mu\Big) \otimes \Big(\sum_\mu D_\mu\Big)\Big]\\ &= \Big(\sum_\lambda D_\lambda\Big) \otimes \Big(\sum_\lambda D_\lambda\Big).\end{aligned}$$

Method 2.

$$\begin{aligned}\forall\lambda,\ D_\lambda = \text{a subcoalgebra of } C \quad &\Rightarrow \quad D_\lambda^\perp = \text{an ideal of } C^*\\ &\Rightarrow \quad \Big[\sum_\lambda D_\lambda\Big]^\perp = \bigcap_\lambda D_\lambda^\perp = \text{an ideal of } C^*\\ &\Rightarrow \quad \sum_\lambda D_\lambda = \Big[\sum_\lambda D_\lambda\Big]^{\perp\perp\!\!\perp}\\ &\qquad\qquad = \text{a subcoalgebra of } C. \qquad \square\end{aligned}$$

PROOF OF 4.7(b). *Method* 1.

$$\varepsilon\Big(\sum_\lambda I_\lambda\Big) \subseteq \sum_\lambda \varepsilon(I_\lambda) = 0$$

while

$$\begin{aligned}\Delta\Big(\sum_\lambda I_\lambda\Big) &\subseteq \sum_\lambda \Delta(I_\lambda) \subseteq \sum_\lambda (I_\lambda \otimes C + C \otimes I_\lambda)\\ &= \sum_\lambda (I_\lambda \otimes C) + \sum_\lambda (C \otimes I_\lambda)\\ &= \Big(\sum_\lambda I_\lambda\Big) \otimes C + C \otimes \Big(\sum_\lambda I_\lambda\Big).\end{aligned}$$

Method 2.

$$\forall\lambda,\ I_\lambda = \text{a coideal of } C \quad \Rightarrow \quad I_\lambda^\perp = \text{a subalgebra of } C^*$$

$$\Rightarrow \quad \left[\sum_{\lambda\in\Lambda} I_\lambda\right]^{\perp} = \bigcap_{\lambda\in\Lambda} I^{\perp} = \text{a subalgebra of } C^*$$
$$\Rightarrow \quad \sum_{\lambda\in\Lambda} I_\lambda = \left[\sum_{\lambda\in\Lambda} I\right]^{\perp\,\perp\!\!\perp} = \text{a coideal of } C. \qquad \square$$

PROOF OF 4.7(c).

$$\begin{aligned}
\forall\lambda,\ D_\lambda = \text{a subcoalgebra of } C \quad &\Rightarrow \quad D_\lambda^{\perp} = \text{an ideal of } C^*\\
&\Rightarrow \quad \sum_\lambda D_\lambda^{\perp} = \text{an ideal of } C^*\\
&\Rightarrow \quad \bigcap_\lambda D_\lambda = \bigcap_\lambda D_\lambda^{\perp\,\perp\!\!\perp} = \left[\sum_\lambda D_\lambda^{\perp}\right]^{\perp\!\!\perp}\\
&\qquad\qquad = \text{a subcoalgebra of } C. \qquad \square
\end{aligned}$$

PROOF OF 4.7(d). The intersection of even two coideals of a coalgebra need not be a coideal because "dually" the sum of two subalgebras of a finite-dimensional algebra need not be a subalgebra. For example, in the algebra $A := \mathcal{M}_4(\mathbb{R})$ of all 4-by-4 matrices with real entries, let A_1 and A_2 denote, respectively, the subspaces of A defined by setting

$$A_1 := \left\{ \begin{pmatrix} a & 0 & e & g\\ 0 & b & 0 & f\\ 0 & 0 & c & 0\\ 0 & 0 & 0 & d \end{pmatrix} \,\middle|\, a, b, c, d, e, f, g \in \mathbb{R} \right\}$$

and

$$A_2 := \left\{ \begin{pmatrix} t & 0 & 0 & 0\\ 0 & u & 0 & 0\\ x & 0 & v & 0\\ z & y & 0 & w \end{pmatrix} \,\middle|\, t, u, v, w, x, y, z \in \mathbb{R} \right\}.$$

One readily checks that A_1 and A_2 are subalgebras of A. On the other hand, the sum

$$A_1 + A_2$$

of A_1 and A_2 is *not* a subalgebra of A since, for example,

$$\begin{pmatrix} a & 0 & e & g\\ 0 & b & 0 & f\\ h & 0 & c & 0\\ j & i & 0 & d \end{pmatrix} \cdot \begin{pmatrix} p & 0 & t & v\\ 0 & q & 0 & u\\ w & 0 & r & 0\\ y & x & 0 & s \end{pmatrix} = \begin{pmatrix} ap + ew + gy & \boxed{gx} & \bullet & \bullet\\ \boxed{fy} & bq + fx & 0 & \bullet\\ \bullet & 0 & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet \end{pmatrix}.$$

Set $C := A^* =$ *the dual coalgebra* of A, defined in 2.58(b). By 2.86 there is an algebra isomorphism

$$A \cong (A^*)^* =: C^*.$$

Since, for each $i \in \{1, 2\}$, A_i is a *subalgebra* of A, we have, by 4.6(b), that

$$A_i^{\perp\!\!\perp} \text{ is a } \textit{coideal} \text{ of } C := A^*, \quad \forall i \in \{1, 2\}.$$

If $A_1^{\perp\!\!\perp} \cap A_2^{\perp\!\!\perp}$ were a *coideal* of C, then by 4.6(a), $[A_1^{\perp\!\!\perp} \cap A_2^{\perp\!\!\perp}]^{\perp}$ would be a *subalgebra* of C^*. But, by 4.2(d)1 and 4.2(b)2,

$$\big[A_1^{\perp\!\!\perp} \cap A_2^{\perp\!\!\perp}\big]^{\perp} = A_1^{\perp\!\!\perp\perp} + A_2^{\perp\!\!\perp\perp} = A_1 + A_2$$

which is not a subalgebra of C^*. This contradiction shows that the intersection of even two coideals of a coalgebra need not be a coideal. □

PROOF OF 4.8(a). *Method* 1.

Because D_1 is a subcoalgebra of C_1, the inclusion $i : D_1 \hookrightarrow C_1$ is a coalgebra map and therefore the composite $f \circ i : D_1 \to C_2$ is a coalgebra map. From the relevant though not displayed commutative diagram, we find that for any given $d \in D$,

$$\Delta_{C_2}\big[f(d)\big] = \Delta_{C_2}\big[f \circ i(d)\big] = \big[(f \circ i) \otimes (f \circ i)\big] \circ \Delta_{D_1}(d).$$

But $\Delta_{D_1}(d) = \sum_d d_1 \otimes d_2$, for some $d_1, d_2 \in D_1$, so

$$\Delta_{C_2}\big[f(D_1)\big] \subseteq f(D_1) \otimes f(D_1)$$

thereby proving that $f(D_1)$ is a subcoalgebra of C_2.

Method 2.

coalgebra map

$$f(D_1) = \Big[\overbrace{(f^*)^{-1}\big(D_1^{\perp}\big)}^{\text{ideal}}\Big]^{\perp\!\!\perp} = \text{subcoalgebra}$$

subcoagebra algebra map ideal □

NOTE. If $g : A \to B$ is an algebra map and if I is an ideal of B, then

$$g^{-1}(I) := \big\{x \in A \mid g(x) \in I\big\}$$

is an ideal of A. Indeed, one readily checks that $g^{-1}(I)$ is a subspace of A. Moreover, $x \in f^{-1}(I) \Leftrightarrow f(x) \in I$. Then, for all $a \in A$,

$$f(a \cdot x) = f(a) \cdot f(x) \in B \cdot I \subseteq I \quad \Rightarrow \quad ax \in f^{-1}(I).$$

Similarly $x \cdot a \in f^{-1}(I)$.

NOTE. In like manner, with reference to the proof of 4.8(b), presented in Method 2, below, one can show that the inverse image of a subalgebra under an algebra map is again a subalgebra.

PROOF OF 4.8(b). *Method* 1.

$$\begin{aligned}\Delta_{C_2}\big[f(I)\big] &= \Delta_{C_2} \circ f(I) \\ &= (f \otimes f) \circ \Delta_{C_1}(I) = (f \otimes f)\big[\Delta_{C_1}(I)\big] \\ &\subseteq (f \otimes f)[I \otimes C_1 + C_1 \otimes I] \\ &\subseteq f(I) \otimes f(C_1) + f(C_1) \otimes f(I) \\ &\subseteq f(I) \otimes C_2 + C_2 \otimes f(I)\end{aligned}$$

and

$$\begin{aligned}I \subseteq \operatorname{Ker} \varepsilon_{C_1} \quad &\Rightarrow \quad \varepsilon_{C_2}\big[f(I)\big] = \varepsilon_{C_2} \circ f(I) = \varepsilon_{C_1}(I) = 0 \\ &\Rightarrow \quad f(I) \subseteq \operatorname{Ker} \varepsilon_{C_2}.\end{aligned}$$

Method 2.

I is a coideal of $C_1 \Rightarrow I^{\perp}$ is a subalgebra of C_1^*. Now $f : C_1 \to C_2$ is a coalgebra map, so $f^* : C_2^* \to C_1^*$ is an algebra map, so $(f^*)^{-1}(I^{\perp})$ is a subalgebra of C_2^*, so

$$f(I) = \big[(f^*)^{-1}\big(I^{\perp}\big)\big]^{\perp\!\!\perp} \text{ is a coideal of } C_2. \qquad \square$$

PROOF OF 4.8(c). $J =$ a coideal of C_2

$$\begin{aligned}&\Rightarrow \quad J^{\perp} = \text{a subalgebra of } C_2^* \\ &\Rightarrow \quad f^*(J^{\perp}) = \text{a subalgebra of } C_1^* \text{ as the image of a} \\ &\qquad\qquad\qquad \text{subalgebra under an algebra map} \\ &\Rightarrow \quad f^{-1}(J) = \big[f^*\big(J^{\perp}\big)\big]^{\perp\!\!\perp} = \text{a coideal of } C_1.\end{aligned} \qquad \square$$

PROOF OF 4.8(d). If $f : C_1 \to C_2$ is an injective coalgebra map, then

$$\begin{aligned}&D_2 = \text{a subcoalgebra of } C_2 \\ &\quad\Rightarrow \quad D_2^{\perp} = \text{an ideal of } C_2^* \\ &\quad\Rightarrow \quad f^*\big(D_2^{\perp}\big) = \text{an ideal of } C_1^* \\ &\quad\Rightarrow \quad f^{-1}(D_2) = \big[f^*\big(D_2^{\perp}\big)\big]^{\perp\!\!\perp} \text{ a subcoalgebra of } C_1.\end{aligned} \qquad \square$$

NOTE. The above proof uses two facts.

(1) $f^* : C_2^* \to C_1^*$ is a surjective algebra map because $f : C_1 \to C_2$ is an injective coalgebra map.

(2) The image $g(I)$ of an ideal I of an algebra A under a surjective algebra map $g : A \to B$ is an ideal of B since

$$g(I) \cdot B = g(I) \cdot g(A) \subseteq g(I \cdot A) = g(I)$$

and similarly,

$$B \cdot g(I) \subseteq g(I).$$

As a further illustration of how results developed in this section may be applied, we consider the following.

PROPOSITION 4.9. *Let C be a coalgebra and let $x \in C$. If*

$$\Delta(x) = \sum_{i=1}^{k} y_i \otimes z_i, \quad \textit{with } \{y_1, \ldots, y_k\} \textit{ linearly independent},$$

then $\{z_1, \ldots, z_k\} \subseteq D$ for any subcoalgebra D of C containing x. Similarly, $\{y_1, \ldots, y_k\} \subseteq D$ if $\{z_1, \ldots, z_k\}$ is linearly independent. In consequence, $\{y_1, \ldots, y_k, z_1, \ldots, z_k\} \subseteq D$ whenever $\{y_1, \ldots, y_k\}$ is linearly independent and $\{z_1, \ldots, z_k\}$ is linearly independent.

PROOF. Let $g \in D^{\perp}$ and $f \in C^*$ be arbitrary. Then $f * g \in D^{\perp}$ because $D^{\perp}$ is an ideal of C^*. Therefore,

$$0 = f * g(x) = \sum_{i=1}^{k} f(y_i) \cdot g(z_i) = f\left[\sum_{i=1}^{k} g(z_i) \cdot y_i\right].$$

Since f is arbitrary, $\sum_i g(z_i) \cdot y_i = 0$. Therefore $g(z_i) = 0$. Since g is arbitrary, $z_i \in D^{\perp \perp\!\!\perp} = D$. $\square$

DEFINITION 4.10. If (C, Δ, ε) is a coalgebra and $c \in C$, then the intersection of all subcoalgebras of C containing c is called *the smallest subcoalgebra of C containing c*. It is also called *the minimal subcoalgebra of C containing c*.

REMARK 4.11. We now establish the following basic result (due to Pierre Cartier, cf. [Mum, p. 25]).

THEOREM 4.12 (The Fundamental Theorem of Coalgebras). *All associative unitary coalgebras over a field are locally finite.*

PROOF. Let (C, Δ, ε) be an associative unitary coalgebra over a field, and let $(C^*, \Delta^* \circ \rho, \varepsilon^* \circ \lambda)$ be its *dual algebra*. By Definition 2.18, we must prove that each element of C lies in a finite-dimensional subcoalgebra. For this, let $c \in C$ and let

$$B = \text{the } \textit{minimal subcoalgebra of } C \textit{ containing } c.$$

Let

$$\{e_i \mid i \in I\} = \textit{a basis for } B \textit{ over } K, \textit{for some index set } I,$$

and let $\{e^j \mid j \in I\} =$ *the dual basis of* B^* so that, for all $j \in I$, $e^j \in B^*$ is defined by $e^j(e_i) = \delta_{ij} \cdot 1_K$ where

$$1_K = \text{the identity element of the ground field } K.$$

Next, for any given $j \in I$, let

$$L^j = B^* * e^j * B^* := \left\{ \sum_{i=1}^{n} f_i * e^j * g_i \;\middle|\; \begin{array}{l} 1 \leqslant n \in \mathbb{Z}, \\ f_i, g_i \in B^* \end{array} \right\}$$

where the "$*$" at mid-level denotes the *convolution product* of C^* defined, for all $f, g \in C^*$ and for all $c \in C$, by

$$\begin{aligned} f * g(c) = \Delta^* \circ \rho(f \otimes g)(c) = \Delta^*\big[\rho(f \otimes g)\big](c) = \rho(f \otimes g) \circ \Delta(c) \\ = \varphi_K \circ (f \otimes g) \circ \Delta(c) = \sum_c f(c_1) \cdot g(c_2) \end{aligned}$$

where

$$\Delta(c) = \sum_c c_1 \otimes c_2.$$

Clearly

$$L^j = \text{the (2-sided!) } \textit{ideal} \text{ of the associative } \textit{unitary} \text{ algebra } B^* \text{ generated by } e^j,$$

so

$$\big(L^j\big)^{\perp\!\!\!\perp} := \big\{b \in B \mid L^j(b) = 0\big\} \subseteq B$$

is a *subcoalgebra* of B and hence of C. Now $e^j \in L^j := B^* * e^j * B^*$ (recall that B^* has a 2-sided identity!) and $e^j(e_j) = 1 \neq 0$, so

$$e_j \in B \quad \text{while } e_j \notin \big(L^j\big)^{\perp\!\!\!\perp}$$

thereby proving that

$$\big(L^j\big)^{\perp\!\!\!\perp} \subsetneqq B.$$

It now follows that

$$L^j(c) \neq 0;$$

for, if $L^j(c) = 0$, then

$$c \in \left(L^j\right)^{\perp\!\!\perp}$$

and therefore $(L^j)^{\perp\!\!\perp}$ would be a subcoalgebra of C containing c and *strictly smaller than* B contrary to the hypothesis that B is the *minimal* subcoalgebra of C containing c. Now since, for each $j \in I$,

$$B^* * e^j * B^*(c) =: L^j(c) \neq 0,$$

without loss of generality, it follows that, for all $j \in I$, there exists a *generator* h_j of L^j such that $h_j(c) \neq 0$. Thus, without loss of generality, for any given $j \in I$, there exist elements $f_j, g_j \in B^*$ and elements $k_{pqr} \in K$ (for some $p, q, r \in I$) such that

$$\begin{aligned} 0 \neq h_j(c) := f_j * e^j * g_j(c) &= \sum_{p,q,r} k_{pqr} \cdot f_j(e_p) \cdot e^j(e_q) \cdot g_j(e_r) \\ &= \sum_{p,r} k_{pjr} \cdot f_j(e_p) \cdot g_j(e_r) \end{aligned}$$

where

$$\sum_{p,q,r} k_{pqr} \cdot e_p \otimes e_q \otimes e_r =: (1 \otimes \Delta) \circ \Delta(c) = (\Delta \otimes 1) \circ \Delta(c).$$

In other words, for any given $j \in I$, there exist elements $p, r \in I$ and an element $k_{pjr} \in K$ such that

$$k_{pjr} \neq 0.$$

But, since only finitely many of the coefficients k_{pqr} can be non-zero and each $j \in I$ leads to such a non-zero k_{pjr} it follows that the set I indexing the basis element of B must be finite thereby proving that B is finite-dimensional. $\square$

REMARK 4.13. Our proof follows Grünenfelder ([Gr-1, p. 65, Lemma III.1.8(b)] and [Gr-2, p. 172, Lemma 2.1]) who in turn presents a non-topological version of the proof given by Larson [Lar-1, p. 351, Proposition 2.5]. There is another type of proof given by Heyneman–Sweedler [Hey-Sw-1, p. 203, Corollary 1.4.3] and by Sweedler [Sw-1, p. 46, Theorem 2.2.1] which in turn has been modified and presented in a more direct way by Yanagihara [Y, p. 39, Corollary 3.9] and by Winter [Win, pp. 177–178]. Abe [Ab, p. 77, Corollary 2.2.14(i)] has an altogether different approach. The sketch of another – this time constructively explicit proof – is to be found in Kaplansky [Kap, p. 7, Theorem 2]. Mumford, Waterhouse, Montgomery (in that order, chronologically) and subsequently [Dăs-Năs-Rai, p. 24, Theorem 1.2.4] all give variants of another type of proof (cf. Appendix 5.3, below). Interestingly enough, Mumford [Mum, p. 25] writes (of his proof) in a footnote that "This lemma was pointed out to me by Cartier." Cartier really seems to have been in on the ground floor: Recall (cf. 2.15) that Milnor and Moore [M-M-1, p. 4, Footnote 3] attribute the term "coalgebra" to Cartier. But whereas Mumford's proof is explicitly scheme-theoretic, the idea appearing in Mumford [Mum, p. 25, Lemma] appears in Waterhouse [Wa-2, p. 24, Theorem] but stripped of the scheme-theoretic setting of Mumford. Waterhouse's proof reappears with credit and minor notational modifications in Montgomery [Mon-1, p. 56, Theorem 5.11(2)]. A final reference is that of Serre [Se-3, p. 43, Théorème 1] in his article "Gèbres" which appeared in the 1989 edition of *L'Enseignment Mathematique*. As Serre himself notes, his article was originally written for (and intended to appear in) the Bourbaki Series of 1965. It appears in [Se-3, pp. 33–85] in its original form – with permission.

REMARK 4.14. The reader will observe that the proof of (4.12) definitely made use of the hypothesis that the coalgebra had a counit. In fact, however, one need not assume the existence of a counit in order to establish the local finiteness of the coalgebra. To understand why, simply observe that any associative coalgebra can always be obtained as a quotient of an associative unitary coalgebra – and this in a universal way. Indeed, given an associative coalgebra (C, Δ_C), define C_ε by

$$C_\varepsilon = C \oplus K,$$

define addition and scalar multiplication coordinatewise, and define

$$\Delta : C_\varepsilon \to C_\varepsilon \otimes C_\varepsilon \quad \text{and} \quad \varepsilon : C_\varepsilon \to K$$

by

$$\begin{aligned}\Delta(c,k) := {} & \sum_c (c_1, 0) \otimes (c_2, 0) + (c, 0) \otimes (0, 1) + (0, 1) \otimes (c, 0) \\ & + k(0, 1) \otimes (0, 1)\end{aligned}$$

and by

$$\varepsilon(c,k) = k,$$

respectively, where

$$\Delta_C(c) = \sum_c c_1 \otimes c_2.$$

(As before, we consistently use the notation $\sum_c c_1 \otimes c_2$ for the finite sum $\sum_{i=1}^n c_{1i} \otimes c_{2i}$ which depends on c.) One readily checks that $(C_\varepsilon, \Delta, \varepsilon)$ is an associative unitary coalgebra and that the natural projection $p : C_\varepsilon \to C$, given by $p[(c,k)] = c$, respects the diagonals, i.e., $\Delta_C \circ p = (p \otimes p) \circ \Delta$. Consequently (by 4.8(a)), the local finiteness of C follows from that of C_ε. One should further observe that (C_ε, p) satisfies the following *universal mapping property*: If $(D, \Delta_D, \varepsilon_D)$ is any associative unitary coalgebra and if $f : D \to C$ is any linear map compatible with the diagonals (i.e., a morphism of associative coalgebras), then there is a unique morphism $F : D \to C_\varepsilon$ of associative unitary coalgebras making the diagram

$$\begin{tikzcd} C & & C_\varepsilon \arrow[ll, "p"'] \\ & D \arrow[ul, "f"] \arrow[ur, dashed, "F"'] & \end{tikzcd}$$

commute. Here, $F : D \to C_\varepsilon$ is given by setting

$$F(d) := \big(f(d), \varepsilon_D(d)\big), \quad \text{for all } d \in D.$$

In consequence, one obtains a functor from the category of associative coalgebras to the category of associative unitary coalgebras which is right adjoint to the forgetful functor. It is in this sense that the process (just) sketched above is “dual” to that of embedding an associative algebra in an associative unitary algebra in a universal way (cf. [J-1, pp. 84–85] or [B-J-N, pp. 193–194, Example 2.7.(c)])!

NOTE. Recall that in the case of algebras, any associative algebra A may always be embedded in an associative unitary algebra A_η in a universal way, universal meaning that the functor $A \mapsto A_\eta$ so obtained is *left adjoint* to the forgetful functor from the category **Alg** of associative unitary algebras to the category of associative algebras. Specifically, if (A, φ_A) is an associative algebra, set

$$A_\eta = A \oplus K,$$

define addition and scalar multiplication coordinatewise, and define a multiplication

$$\varphi : A_\eta \otimes A_\eta \to A_\eta$$

and a unit

$$\eta : K \to A_\eta$$

by

$$\begin{aligned} \varphi\big[(a,m) \otimes (b,n)\big] &:= (\varphi_A(a \otimes b) + n \cdot a + m \cdot b, m \cdot n) \\ &:= (a \cdot b + n \cdot a + m \cdot b, m \cdot n) \end{aligned}$$

and by

$$\eta(1_K) := (0_A, 1_K)$$

respectively. One readily checks that (A_η, φ, η) is an object of **Alg**, and that the K-linear injection $i : A \to A_\eta$ given by

$$i(a) := (a, 0)$$

is an algebra map. One also sees that if (B, φ_B, η_B) is an object of **Alg**, and if $f : A \to B$ is any linear map compatible with the multiplications on A and on B (in the sense that $\varphi_B \circ (f \otimes f) = f \circ \varphi_A$), then there exists a unique morphism $F : A_\eta \to B$ of **Alg** such that $F \circ i = f$. Here, $F(a, k) := f(a) + k \cdot 1_B$ where $1_B = \eta_B(1_K)$, 1_K being the identity element of the ground field K.

REMARKS 4.15. (a) *Nonassociative* coalgebras need *not* be *locally finite*. For example, in [Mi-7, p. 346, Theorem 15] an example is given of an *infinite-dimensional Lie coalgebra* in which no element except 0 is contained in a finite-dimensional sub Lie coalgebra.

(b) If K is a *commutative ring* rather than a *field*, and if C is a K-module, then the counterpart of Theorem 4.12 does *not* hold in general (cf. [Ber-Hau, pp. 172–173] or Appendix 5.3, below) though it may hold in special cases (cf. [Haz-2] or Appendix 5.3, below).

REMARK. We now establish the remark of M. Takeuchi [Tak-3, p. 232] that we previously mentioned in 2.84(b). The proof that follows is based on suggestions in an e-mail from L. Grünenfelder.

PROPOSITION 4.16. *Every finitely generated bialgebra is a matric bialgebra, i.e., it has a set of algebra generators x_{ij} such that $\Delta(x_{ij}) = \sum_k x_{ik} \otimes x_{kj}$ and $\varepsilon(x_{ij}) = \delta_{ij} \cdot 1_K$.*

PROOF. Let H be a bialgebra that is generated as an algebra by elements $h_1, \ldots, h_q \in H$. Then for any given $h \in H$ there exists elements $k_i \in K$ and integers i_j satisfying $0 \leqslant i_j < \infty$ for which

$$h = \sum_{i=1}^{p} k_i \cdot h_1^{i_1} \cdot h_2^{i_2} \ldots h_q^{i_q}.$$

Because H is *locally finite* as a coalgebra, there exist finite-dimensional subcoalgebras C_j of H, one for each integer $j \in \{1, \ldots, q\}$, such that $h_j \in C_j$. Set

$$C = \sum_{j=1}^{q} C_j.$$

Because the sum of subcoalgebras is again a subcoalgebra (cf. 4.7(a)), and because each C_j is finite-dimensional, C is an n-dimensional subcoalgebra of H for some positive integer $n < \infty$. Let

$$A = C^* = \text{the } \textit{dual algebra} \text{ of } C.$$

Then A is an n-dimensional algebra and we may embed A in the $(n \times n)$-dimensional *endomorphism algebra* $\mathrm{End}(A) = \mathrm{Hom}_K(A, A)$ by means of the *left regular representation*

$$\lambda : A \hookrightarrow \mathrm{End}(A)$$

defined, for all $a, a' \in A$, by $\lambda(a)(a') = a \cdot a'$. As usual, *we make the identification* $\mathrm{End}(A) =$ the matrix algebra spanned by the linearly independent "matrix units" E_{ij}, where, by definition, E_{ij} is the n-by-n matrix having a 1 in the (ij)th spot and 0's elsewhere. One readily verifies that

$$E_{ij} \cdot E_{kl} = \delta_{jk} \cdot E_{il}.$$

Let

$$D := \big[\mathrm{End}(A)\big]^* = \text{the vector space dual of } \mathrm{End}(A)$$

and let $X_{ij} \in D$ be defined by

$$X_{kl}(E_{ij}) = \delta_{ki} \cdot \delta_{\ell j} \cdot 1_K$$

so that $\{X_{ij}\}$ is the *basis* for D *dual to the basis* $\{E_{ij}\}$ for $\mathrm{End}(A)$. A direct calculation then shows that in the coalgebra D, the *dual coalgebra* of $\mathrm{End}(A)$,

$$\Delta(X_{ij}) = \sum_{k=1}^{n} X_{ik} \otimes X_{kj} \quad \text{and} \quad \varepsilon(X_{ij}) = \delta_{ij} \cdot 1_K.$$

Indeed, if B^* is the dual coalgebra of the finite-dimensional algebra $B := \mathrm{End}(A)$, then $\Delta = \Delta_{B^*}$ is given by the commutative diagram

$$\begin{tikzcd} B^* \arrow[r, "(\varphi_B)^*"] \arrow[dr, dashed, "\Delta"'] & (B \otimes B)^* \arrow[d, "(\rho_B)^{-1}"] \\ & B^* \otimes B^* \end{tikzcd}$$

so, for all $f \in B^*$, we have (by 2.58(b)) that

$$\Delta(f) = \sum_f f_1 \otimes f_2 \quad \Leftrightarrow \quad f(b_1 \cdot b_2) = \sum_f f_1(b_1) \cdot f_2(b_2) \text{ for all } b_i \in B.$$

But, for all $S \in \mathrm{End}(A) \equiv \mathcal{M}_n(K)$,

$$\begin{aligned} X_{ij}(S) &= X_{ij}\left(\sum_{p=1}^{n}\sum_{q=1}^{n} S_{pq} \cdot E_{pq}\right) = \sum_{p=1}^{n}\sum_{q=1}^{n} S_{pq} \cdot X_{ij}(E_{pq}) \\ &= \sum_{p=1}^{n}\sum_{q=1}^{n} S_{pq} \cdot \delta_{ip} \cdot \delta_{jq} \cdot 1_K = S_{ij} \cdot 1_K = S_{ij} \end{aligned}$$

so that, for all $S, T \in \mathrm{End}(A) \equiv \mathcal{M}_n(K)$,

$$X_{ij}(S \cdot T) = (S \cdot T)_{ij} = \sum_{k=1}^{n} S_{ik} \cdot T_{kj} = \sum_{k=1}^{n} X_{ik}(S) \cdot X_{kj}(T)$$

so

$$\Delta(X_{ij}) = \sum_{k=1}^{n} X_{ik} \otimes X_{kj}.$$

As to $\varepsilon := \varepsilon_{B^*}$, ε_{B^*} is induced from $\eta : K \to B$ via the commutative diagram

$$\begin{tikzcd} B^* \arrow[r, "\eta^*"] \arrow[dr, dashed, "\varepsilon_{B^*}"'] & K^* \arrow[d, "\lambda^{-1}"] \\ & K \end{tikzcd}$$

so that, for all $f \in B^*$,

$$\varepsilon(f) = \lambda^{-1} \circ \eta^*(f) = \lambda^{-1}[f \circ \eta] = f \circ \eta(1_K) = f\big[\eta(1_K)\big] = f(1_B).$$

Thus, in the present case

$$\begin{aligned} \varepsilon(X_{ij}) = X_{ij}(I_n) = X_{ij}\left[\sum_{k=1}^{n} E_{kk}\right] &= \sum_{k=1}^{n} X_{ij}(E_{kk}) = \sum_{k=1}^{n} \delta_{ik} \cdot \delta_{jk} \cdot 1_K \\ &= \delta_{ij} \cdot 1_K, \end{aligned}$$

so indeed

$$\Delta(X_{ij}) = \sum_{k=1}^{n} X_{ik} \otimes X_{kj}$$

and

$$\varepsilon(X_{ij}) = \delta_{ij} \cdot 1_K.$$

Now, the embedding

$$\lambda : A \hookrightarrow \mathrm{End}(A) =: B$$

is an *injective algebra map*, so

$$f := \chi^{-1} \circ \lambda^* : D := B^* := \big[\mathrm{End}(A)\big]^* \xrightarrow{\lambda^*} A^* = C^{**} \xrightarrow{\chi^{-1}} C$$

is a *surjective coalgebra map*.

Since $\{X_{ij}\}$ is a basis for D, and since the image of a spanning set under a linear surjection is a spanning set,

$$\{f(X_{ij})\} \text{ spans } C.$$

Set $x_{ij} := f(X_{ij})$. Then $\{x_{ij}\}$ spans C. We claim that the fact that $f : D \to C$ is a surjective map of **Coalg** implies both that

$$\Delta_C(x_{ij}) = \sum_{k=1}^{n} x_{ik} \otimes x_{kj}$$

and that

$$\varepsilon_C(x_{ij}) = \delta_{ij} \cdot 1_K.$$

To establish these claims, consider, first, the commutative diagram

$$\begin{tikzcd}
D = [\mathrm{End}(A)]^* = D \arrow[r, "\Delta_D"] \arrow[d, "\lambda^*"'] \arrow[ddd, "f"'] & D \otimes D \arrow[r, equal] \arrow[d, "\lambda^* \otimes \lambda^*"] & D \otimes D \arrow[dddl, "f \otimes f"] \\
A^* \arrow[r, "\Delta_{A^*}"'] \arrow[d, equal] & A^* \otimes A^* \arrow[d, equal] & \\
C^{**} \arrow[r, "\Delta_{C^{**}}"'] \arrow[d, "\chi^{-1}"'] & C^{**} \otimes C^{**} \arrow[d, "\chi^{-1} \otimes \chi^{-1}"] & \\
C \arrow[r, "\Delta_C"'] & C \otimes C &
\end{tikzcd}$$

and notice that, for any given $x_{ij} \in C = f(D)$, we may pull $x_{ij} \in C$ back to $X_{ij} \in D$ under f to find that

$$\begin{aligned}
\Delta_C\big[f(X_{ij})\big] &= \Delta_C \circ f(X_{ij}) = (f \otimes f) \circ \Delta_D(X_{ij}) \\
&= \sum_{k=1}^{n} f(X_{ik}) \otimes f(X_{kj})
\end{aligned}$$

and therefore that

$$\Delta_C(x_{ij}) = \sum_{k=1}^{n} x_{ik} \otimes x_{kj}.$$

To establish the second identity, observe that from the commutativity of the diagram

$$\begin{tikzcd} D \arrow[r, "\varepsilon_D=\varepsilon_{B^*}"] \arrow[d] & K \arrow[d, equal] \\ C \arrow[r, "\varepsilon_C"'] & K \end{tikzcd}$$

we find that

$$\varepsilon_C\big[f(X_{ij})\big] = \varepsilon_C \circ f(X_{ij}) = \varepsilon_D(X_{ij}) = \varepsilon_{B^*}(X_{ij}) = \delta_{ij} \cdot 1_K$$

and hence that

$$\varepsilon_C(x_{ij}) = \delta_{ij} \cdot 1_K.$$

Thus, C and hence H contains a set of elements x_{ij} such that

$$\Delta_C(x_{ij}) = \sum_k x_{ik} \otimes x_{kj} \quad \text{and} \quad \varepsilon(x_{ij}) = \delta_{ij} \cdot 1_K;$$

so to complete the proof we must just show that $\{x_{ij}\}$ generates H as an algebra. But this is clear because $\{x_{ij}\}$ spans the subcoalgebra $C = \sum C_j$ of H and thus each algebra generating $h_j \in C_j$ of H may be written as a finite linear combination of the x_{ij}'s. □

5. Appendices

5.1. *A Theorem of Harish-Chandra*

The Harish-Chandra theorem of the title is the one [H-C, p. 905, Theorem 1] asserting the existence of sufficiently many representations of a finite-dimensional Lie algebra over a field of characteristic zero. More precisely, it says that if L is a finite-dimensional Lie algebra over a field of characteristic zero, and if UL is its universal enveloping algebra, then there are sufficiently many (finite-dimensional) representations of L to separate the elements of L. A representation of L is the same as a representation of the associative algebra UL and thus the statement is equivalent, [Mi-5, p. 20, Lemma 14], to the statement that the intersection $\mathcal{R}(UL)$ of all cofinite two-sided ideals of UL is zero. Harish-Chandra's original proof is long and involved, but one can give a short, non-computational, Hopf-theoretic proof of this result (cf. [Mi-5]) by observing that $\mathcal{R}(UL)$ is a coideal of UL having zero intersection with the primitives of UL.

Indeed, if I is a coideal of a primitively generated Hopf algebra H (*one generated by* $P(H)$ *as an algebra*) or, more generally, of a *pointed irreducible coalgebra* (*one having a unique group-like element*) – of which UL is an example – then a basic result from the theory of associative unitary coalgebras (cf. [Hey-Sw-1, p. 221, Definition 3.1; p. 232, Corollary 3.2.7(a)], [Sw-1, p. 218, Corollary 11.0.2] or [Ab, p. 101, Corollary 2.4.14]) guarantees that

$$I \cap P(H) = 0 \quad \Rightarrow \quad I = 0$$

where $P(H)$ denotes the space of primitives of H. Now, in characteristic zero, $P(UL) = iL$ ($\equiv$ the image of L in UL) by Friedrichs' theorem (cf. [J-2, p. 170, Theorem 9]) so

$$\mathcal{R}(UL) \cap P(UL) = \mathcal{R}(UL) \cap iL = 0$$

by the theorem of Ado and Iwasawa (cf. [J-2, pp. 202 and 204]), which asserts that for any given finite-dimensional Lie algebra L there exists a cofinite two-sided ideal I of UL such that $I \cap iL = 0$ (i.e., that there exists a faithful representation of L). Thus aside from Friedrichs' theorem, the Ado–Iwasawa theorem, and the result from the theory of coalgebras just cited, the above proof of Harish-Chandra's theorem hinges on the observation that the intersection of all cofinite two-sided ideals of a bialgebra, obviously an *ideal*, is also a *coideal* – a result easily established (cf. [Mi-5, pp. 18–20, Theorem 5]) by making use of the correspondence between special subspaces of a coalgebra and its dual algebra furnished by the annihilator transformations (described in Section 4).

5.2. *The Jacobian Conjecture*

The Jacobian conjecture, also known as the Keller problem, deals with polynomial mappings $F = (F_1, \ldots, F_n) : \mathbb{C}^n \to \mathbb{C}^n$, i.e., each F_i is a polynomial in n variables, or equivalently with endomorphisms of the polynomial ring $\mathbb{C}[X_1, \ldots, X_n]$, or more generally of $K[X_1, \ldots, X_n]$ for K a field of characteristic zero[2]. It says that such a polynomial mapping (endomorphism) has a polynomial inverse if and only if its Jacobian (= determinant of the Jacobian matrix) is a nonzero constant.

What follows is based on an e-mail [Sw-4] from Moss E. Sweedler sent December 10, 1996. The (re)solution of this problem is left as a challenge to new-comers and oldsters alike.

The easy and well known direction of the Jacobian conjecture is that an algebra automorphism of the polynomial ring has a non-zero constant Jacobian. That follows directly from the chain rule. Let's consider the converse.

An algebra map from the polynomial ring to itself can be normalized by a translation so that it carries the ideal J generated by $X_1, \ldots, X_n$ to itself. This doesn't change the problem. Call the normalized algebra map L.

With this normalization, L extends to an algebra map M of the power series ring $K[[X_1, \ldots, X_n]]$ to itself. The normalization insures that the coalgebra map on the Hopf dual carries the connected component, C_0, of the group-like element associated to J, to itself. Call this coalgebra map M_0^0.

[2]For an up-to-date summary of the current state of affairs as regards the Jacobian conjecture see A. van den Essen, *Jacobi conjecture*, in: M. Hazewinkel, ed., Encyclopaedia of Mathematics, Vol. 12 (= Supplement II), Kluwer Academic Publishers (2000), pp. 283–284.

The "nonzero constant Jacobian condition" insures that M is an algebra automorphism. Equivalently, M_0^0 is a coalgebra automorphism.

The linear dual to C_0 is (naturally isomorphic to) $K[[X_1, \ldots, X_n]]$. Also C_0 has a natural topological coalgebra structure whereby the continuous dual is the polynomial subalgebra $K[X_1, \ldots, X_n]$ of $K[[X_1, \ldots, X_n]]$. Also, M_0^0 is continuous and the induced map on the continuous dual is the map L from $K[X_1, \ldots, X_n]$ to itself. The induced algebra map on the full dual is M.

M_0^0 is both a coalgebra automorphism and a continuous map of the topological coalgebra C_0.

"The big question: Does the non-zero constant Jacobian condition insure that the coalgebra inverse $(M_0^0)^{-1}$ of M_0^0 is continuous?"

If $(M_0^0)^{-1}$ is continuous, then the induced algebra map on the continuous dual is the algebra map inverse to L from $K[X_1, \ldots, X_n]$ to itself. This would prove the open direction of the Jacobian conjecture. It would be amusing if this approach helped to establish the Jacobian conjecture, or led to a counterexample.[3]

5.3. *The Fundamental Theorem of Coalgebras*

Here we present a direct proof of 4.12 based on [Dăs-Năs-Rai, p. 24, Theorem 1.2.4]. We show that if (C, Δ, ε) is an *associative unitary coalgebra over a field* K, and if $c \in C$, then there is a *finite-dimensional subcoalgebra* D of C that contains c. Apart from elementary facts about tensor products of K-vector spaces, the proof uses just the fact that $\varepsilon : C \to K$ is a *two-sided counit* and that $\Delta : C \to C \otimes C$ is *associative*. Given $c \in C$, let

$$\Delta_2(c) = (1 \otimes \Delta) \circ \Delta(c) = (\Delta \otimes 1) \circ \Delta(c).$$

It is easy to see that there is a *subset* $\{e_i\}_{i=1}^n$ of a *basis* of C such that

$$\Delta_2(c) = \sum_{i=1}^{n} \sum_{j=1}^{n} e_i \otimes d_{ij} \otimes e_j$$

(since if we write

$$\Delta_2(c) = \sum_{i=1}^{n} \sum_{k=1}^{n} \sum_{j=1}^{n} \lambda_{ikj} e_i \otimes e_k \otimes e_j$$

[3]There is a real chance that this will lead to some insight since the automatic continuity of inverses is by no means an unknown phenomenon in functional analysis. This matter has been seriously investigated, notably in the setting of Banach algebras. **Note:** By the *Open Mapping Theorem* (cf. [Simmons, p. 236, Theorems A and B], [Hof-Mor, p. 650], or [Palm, p. 549]), a *surjective continuous linear transformation* from one Banach space *onto* another is automatically *open*. Consequently, if a *bijective* linear transformation from one Banach space *onto* another is *continuous*, then *its inverse is automatically continuous*. Open Mapping Theorems are a significant aspect of topological group theory and hold in the context of homomorphisms between locally compact groups satisfying mild additional conditions.

with $\lambda_{ikj} \in K$, then $\Delta_2(c)$ has the desired form with $d_{ij} := \sum_{k=1}^{n} \lambda_{ikj} e_k$). Observe that if we write

$$\Delta(c) = \sum_{i=1}^{n} e_i \otimes d_i \quad \text{and} \quad \Delta(d_i) = \sum_{j=1}^{n} d_{ij} \otimes e_j$$

(whence $\Delta_2(c) = \sum_i \sum_j e_i \otimes d_{ij} \otimes e_j$) then

$$c = \sum_{i=1}^{n} \varepsilon(e_i) \cdot d_i \quad \text{and} \quad d_i = \sum_{j=1}^{n} d_{ij} \cdot \varepsilon(e_j)$$

since $\varepsilon : C \to K$ is, respectively, a *left counit* and a *right counit*. Accordingly,

$$c = \sum_{i=1}^{n} \sum_{j=1}^{n} \varepsilon(e_i) \cdot d_{ij} \cdot \varepsilon(e_j)$$

so, upon setting

$$D := \mathrm{Span}_K \{d_{ij} \mid 1 \leqslant i, j \leqslant n\}$$

we clearly have that $c \in D$. We now show that D (clearly *finite-dimensional*) is a *subcoalgebra* of C. By repeated application of the *associativity* of Δ, it is straightforward to see that

$$(\Delta \otimes 1 \otimes 1) \circ \Delta_2(c) = (1 \otimes \Delta \otimes 1) \circ \Delta_2(c)$$

so, reading from right to left, we find that

$$\sum_j \left[\sum_i e_i \otimes \Delta(d_{ij}) \right] \otimes e_j = \sum_j \left[\sum_i \Delta(e_i) \otimes d_{ij} \right] \otimes e_j.$$

By the *linear independence* of the e_j,

$$\sum_i e_i \otimes \Delta(d_{ij}) = \sum_i \Delta(e_i) \otimes d_{ij} \in C \otimes C \otimes D$$

for all j. Then, by the *linear independence* of the e_i,

$$\Delta(d_{ij}) \in C \otimes D \quad \text{for all } i, j.$$

By repeated application of the *associativity* of Δ, it is straightforward to see that

$$(1 \otimes \Delta \otimes 1) \circ \Delta_2(c) = (1 \otimes 1 \otimes \Delta) \circ \Delta_2(c)$$

from which we find that

$$\sum_i e_i \otimes \left[\sum_j \Delta(d_{ij}) \otimes e_j\right] = \sum_i e_i \otimes \left[\sum_j d_{ij} \otimes \Delta(e_j)\right].$$

By the *linear independence* of the e_i, it follows that

$$\sum_j \Delta(d_{ij}) \otimes e_j = \sum_j d_{ij} \otimes \Delta(e_j) \in D \otimes C \otimes C$$

for all i. Because the e_j are *linearly independent*, it follows that

$$\Delta(d_{ij}) \in D \otimes C \quad \text{for all } i, j.$$

Hence $\Delta(d_{ij}) \in (D \otimes C) \cap (C \otimes D) = D \otimes D$ thereby proving that $D := \mathrm{Span}_K\{d_{ij}\}$ is a *subcoalgebra* of C.

NOTE. In the above, we used the following easy-to-prove facts:

(1) If A and B are subspaces of V, if $x_i, y_i \in V$ for all $i \in \{1, \ldots, n\}$, and if $w := \sum_{i=1}^n x_i \otimes y_i \in A \otimes B$, then $y_i \in B$, for all i, if $\{x_i\}_{i=1}^n$ is *linearly independent*, while $x_i \in A$, for all i, if $\{y_i\}_{i=1}^n$ is *linearly independent*.

(2) If either $\sum_{i=1}^n x_i \otimes y_i = \sum_{i=1}^n x_i \otimes z_i$ or $\sum_{i=1}^n y_i \otimes x_i = \sum_{i=1}^n z_i \otimes x_i$ for some $x_i, y_i, z_i \in V$, then $y_i = z_i$, for all i, whenever $\{x_i\}_{i=1}^n$ is *linearly independent*.

To prove (1) note that if $\{x^k\}_{k=1}^n$ is the *dual basis* of $\{x_i\}_{i=1}^n$, then, under

$$f_k := \lambda \circ \left(x^k \otimes 1_B\right) : A \otimes B \xrightarrow{x^k \otimes 1_B} K \otimes B \xrightarrow{\cong} B,$$

$y_k = f_k(w) \in B$. Similarly if $\{y^k\}_{k=1}^n$ is the *dual basis* of $\{y_i\}_{i=1}^n$, then, under

$$g_k := \rho \circ \left(1_A \otimes y^k\right) : A \otimes B \xrightarrow{1_A \otimes y^k} A \otimes K \xrightarrow{\cong} A,$$

$x_k = g_k(w) \in A$. Further (2) clearly follows from (1) by noting, respectively, that

$$\sum_{i=1}^n x_i \otimes (y_i - z_i) = 0 \in V \otimes \{0\}$$

while

$$\sum_{i=1}^n (y_i - z_i) \otimes x_i = 0 \in \{0\} \otimes V.$$

In 4.15(b), above, we remarked that, for associative unitary coalgebras defined over a commutative unitary ring [Hil-Wu, p. 125] rather than a field, the counterpart of the

Fundamental Theorem of Coalgebras need *not* hold though it may hold in certain special cases. We now expand on this. For a brief discussion of difficulties that may arise when one considers coalgebras defined over a commutative ring with 1 rather than a field, see, for example, [N-S, pp. 56–57].

DEFINITION. A coalgebra C defined over a commutative unitary ring R is said to be *locally finite* in case, for any given element $c \in C$, there exists a subcoalgebra D of C that contains c and is *finitely generated* as an R-module.

We now present examples of *non locally finite associative* (*unitary*) *coalgebras* defined over a commutative unitary ring.

EXAMPLES. Following [Ber-Hau, p. 173], let $p :=$ a prime number, let $R := \mathbb{Z}_{p^2} :=$ the *ring of integers modulo* p^2, and let C be the R-module defined by setting

$$C := R \cdot 1 \oplus \left(\bigoplus_{n=1}^{\infty} R \cdot x_n \right)$$

where $1 := 1_R$ and where the x_n are indeterminates. Define R-module maps

$$\Delta : C \to C \otimes_R C \quad \text{and} \quad \varepsilon : C \to R$$

by setting

$$\begin{aligned} \Delta(1) &:= 1 \otimes 1, \\ \Delta(x_n) &:= 1 \otimes x_n + x_n \otimes 1 + p \cdot (x_{n+1} \otimes x_{n+1}), \\ \varepsilon(1) &:= 1_R \quad \text{and} \quad \varepsilon(x_n) := 0. \end{aligned}$$

It is trivial to check that $\varepsilon : C \to R$ is a two-sided counit for $\Delta : C \to C \otimes_R C$. Because (cf. [Mac-Bir, p. 88]) the ring $\mathbb{Z}_{p^2}$ has *characteristic* p^2, one has, for all $x, y, z \in C$, that

$$p^2 \cdot (x \otimes y \otimes z) = \left(p^2 \cdot x\right) \otimes y \otimes z = 0.$$

The associativity of $\Delta : C \to C \otimes_R C$ therefore follows from the fact that

$$(1 \otimes \Delta) \circ \Delta(1) = 1 \otimes 1 \otimes 1 = (\Delta \otimes 1) \circ \Delta(1)$$

and that

$$\begin{aligned} &(1 \otimes \Delta) \circ \Delta(x_n) \\ &\quad = (1 \otimes \Delta)\left[1 \otimes x_n + x_n \otimes 1 + p \cdot (x_{n+1} \otimes x_{n+1})\right] \\ &\quad = 1 \otimes \left[1 \otimes x_n + x_n \otimes 1 + p \cdot (x_{n+1} \otimes x_{n+1})\right] + x_n \otimes 1 \otimes 1 \\ &\qquad + p \cdot x_{n+1} \otimes \left[1 \otimes x_{n+1} + x_{n+1} \otimes 1 + p \cdot (x_{n+2} \otimes x_{n+2})\right] \end{aligned}$$

$$
\begin{aligned}
&= 1 \otimes 1 \otimes x_n + 1 \otimes x_n \otimes 1 + p \cdot [1 \otimes x_{n+1} \otimes x_{n+1}] \\
&\quad + x_n \otimes 1 \otimes 1 + p \cdot [x_{n+1} \otimes 1 \otimes x_{n+1} + x_{n+1} \otimes x_{n+1} \otimes 1] \\
&\quad + p^2 \cdot [x_{n+1} \otimes x_{n+2} \otimes x_{n+2}] \\
&= 1 \otimes 1 \otimes x_n + 1 \otimes x_n \otimes 1 + x_n \otimes 1 \otimes 1 \\
&\quad + p \cdot [x_{n+1} \otimes x_{n+1} \otimes 1 + 1 \otimes x_{n+1} \otimes x_{n+1} + x_{n+1} \otimes 1 \otimes x_{n+1}] \\
&\quad + p^2 \cdot [x_{n+2} \otimes x_{n+2} \otimes x_{n+1}] \\
&= 1 \otimes 1 \otimes x_n + \big[1 \otimes x_n + x_n \otimes 1 + p \cdot (x_{n+1} \otimes x_{n+1})\big] \otimes 1 \\
&\quad + p \cdot \big[1 \otimes x_{n+1} + x_{n+1} \otimes 1 + p \cdot (x_{n+2} \otimes x_{n+2})\big] \otimes x_{n+1} \\
&= (\Delta \otimes 1)\big[1 \otimes x_n + x_n \otimes 1 + p \cdot (x_{n+1} \otimes x_{n+1})\big] \\
&= (\Delta \otimes 1) \circ \Delta(x_n).
\end{aligned}
$$

One can prove (cf. [Ber-Hau, p. 173]) that the associative unitary coalgebra (C, Δ, ε) is *not* locally finite.

Here is a nice way to understand (hence recall) how to construct this and related examples – via a two step process.

First construct a clearly *non locally finite* coalgebra (V, Δ_V) which, in the words of Bergman and Hausknecht (*loc. cit.*), is associative "by default" by virtue of the fact that, for each generator x of the R-module V,

$$
(\Delta_V \otimes 1) \circ \Delta_V(x) = 0 = (1 \otimes \Delta_V) \circ \Delta_V(x) \in V \otimes_R V \otimes_R V.
$$

Then *adjoin a counit* to (V, Δ_V), as in 4.14, to get the desired *non locally finite* associative unitary coalgebra $C = V_\varepsilon := V \oplus R$. In the present case, take the ring R to be $\mathbb{Z}_{p^2}$, the *ring of integers modulo* p^2; take V to be the R-module defined by setting

$$
V := \bigoplus_{n=1}^{\infty} R \cdot x_n
$$

for indeterminates x_n; and take $\Delta_V : V \to V \otimes_R V$ to be the R-module map defined by setting

$$
\Delta_V(x_n) := p \cdot (x_{n+1} \otimes x_{n+1}).
$$

Because the ring $\mathbb{Z}_{p^2}$ has *characteristic* p^2, one has, for each integer $n \geqslant 1$, that

$$
\begin{aligned}
&\big(p^2 \cdot x_{n+1}\big) \otimes x_{n+2} \otimes x_{n+2} = 0 \\
&\quad = \big(p^2 \cdot x_{n+2}\big) \otimes x_{n+2} \otimes x_{n+1} \in V \otimes_R V \otimes_R V
\end{aligned}
$$

whence

$$
\begin{aligned}
(1 \otimes \Delta_V) \circ \Delta_V(x_n) &= \big(p^2 \cdot x_{n+1}\big) \otimes x_{n+2} \otimes x_{n+2} = 0 \\
&= \big(p^2 \cdot x_{n+2}\big) \otimes x_{n+2} \otimes x_{n+1} \\
&= (\Delta_V \otimes 1) \circ \Delta_V(x_n),
\end{aligned}
$$

so (V, Δ_V) is an associative R-coalgebra. From the way $\Delta_V : V \to V \otimes_R V$ is defined, it is very plausible that (V, Δ_V) is *not* locally finite. In fact, one can prove this (cf. [Ber-Hau, p. 173]) as we shall see below. Next, following the procedure of 4.14, let C be the R-module defined by setting

$$C := V_\varepsilon := V \oplus R$$

and let

$$\Delta : C \to C \otimes_R C \quad \text{and} \quad \varepsilon : C \to R$$

be the R-module maps defined by setting

$$\begin{aligned} \Delta(x_n, 0) &:= p \cdot (x_{n+1}, 0) \otimes (x_{n+1}, 0) \\ \Delta(0, 1) &:= (0, 1) \otimes (0, 1), \\ \varepsilon(x_n, 0) &:= 0 \quad \text{and} \quad \varepsilon(0, 1) := 1 := 1_R. \end{aligned}$$

M. Hazewinkel[4] has constructed a slightly more elaborate example of a similar type; namely, he takes C to be the $\mathbb{Z}$-module defined by setting

$$C := \mathbb{Z} \cdot 1 \oplus \left(\bigoplus_{n=2}^{\infty} \mathbb{Z}_n \cdot x_n \right) := \mathbb{Z} \cdot 1 \oplus \mathbb{Z}_2 \cdot x_2 \oplus \mathbb{Z}_3 \cdot x_3 \oplus \cdots$$

where $1 := 1_{\mathbb{Z}}$ and where the x_n are indeterminates. He then takes

$$\Delta : C \to C \otimes_{\mathbb{Z}} C \quad \text{and} \quad \varepsilon : C \to \mathbb{Z}$$

to be the $\mathbb{Z}$-module maps defined by setting

$$\begin{aligned} \Delta(1) &:= 1 \otimes 1, \\ \Delta(x_n) &:= 1 \otimes x_n + x_n \otimes 1 + n \cdot (x_{n^2} \otimes x_{n^2}), \\ \varepsilon(1) &:= 1 := 1_{\mathbb{Z}} \quad \text{and} \quad \varepsilon(x_n) := 0. \end{aligned}$$

This example may be constructed, as above, by taking

$$V := \bigoplus_{n=2}^{\infty} \mathbb{Z}_n \cdot x_n$$

with

$$\Delta_V : V \to V \otimes_{\mathbb{Z}} V$$

[4] M. Hazewinkel (1993), Letter to Giovanna Carnovale.

the $\mathbb{Z}$-module map defined by setting

$$\varDelta_V(x_n) := n \cdot [x_{n^2} \otimes x_{n^2}].$$

Then $\varDelta_V : V \to V \otimes_{\mathbb{Z}} V$ is associative because

$$\begin{aligned}(1 \otimes \varDelta_V) \circ \varDelta_V(x_n) &= n \cdot x_{n^2} \otimes \left[n^2 \cdot (x_{n^4} \otimes x_{n^4})\right] \\ &= n \cdot \left(n^2 \cdot x_{n^2}\right) \otimes x_{n^4} \otimes x_{n^4} = (n \cdot 0) \otimes x_{n^4} \otimes x_{n^4} \\ &= 0 \in (\mathbb{Z}_{n^2} \cdot x_{n^2}) \otimes_{\mathbb{Z}} (\mathbb{Z}_{n^4} \cdot x_{n^4}) \otimes_{\mathbb{Z}} (\mathbb{Z}_{n^4} \cdot x_{n^4}) \\ &\subseteq V \otimes_{\mathbb{Z}} V \otimes_{\mathbb{Z}} V\end{aligned}$$

and similarly,

$$\begin{aligned}(\varDelta_V \otimes 1) \circ \varDelta_V(x_n) &= n \cdot \left[n^2 \cdot (x_{n^4} \otimes x_{n^4})\right] \otimes x_{n^2} \\ &= x_{n^4} \otimes x_{n^4} \otimes (n \cdot 0) \\ &= 0 \in (\mathbb{Z}_{n^4} \cdot x_{n^4}) \otimes_{\mathbb{Z}} (\mathbb{Z}_{n^4} \cdot x_{n^4}) \otimes_{\mathbb{Z}} (\mathbb{Z}_{n^2} \cdot x_{n^2}) \\ &\subseteq V \otimes_{\mathbb{Z}} V \otimes_{\mathbb{Z}} V.\end{aligned}$$

Bergman and Hausknecht's "recipe" (cf. [Ber-Hau, p. 173]) for constructing non locally finite associative unitary coalgebras over a commutative unitary ring is as follows: Take a non-associative coalgebra over the field $\mathbb{Z}_p$, p a *prime*, that provides a counterexample to the Fundamental Theorem of Coalgebras; "lift this to a coalgebra over" the ring $\mathbb{Z}_{p^2}$; then multiply the comultiplication of the counterexample by p to make a new comultiplication that is associative "by default." Then "get a counital version... by adjoining a counit."

In [Mi-7, p. 343, Definition 2, Proposition 3], I constructed *Lie coalgebras* $(E_n, \varDelta_n)$, one for each integer $n \geqslant 1$, such that $\dim_K[\mathrm{Loc}(E_n)] = n$, where, for any coalgebra C, $\mathrm{Loc}(C)$ is defined (as in 3.52) to be the sum of all of the finite-dimensional subcoalgebras of C. Applying the above *recipe*, for example, to the Lie coalgebra $(E, \varDelta)$ of [Mi-1, p. 4, Example (I.1.3.d)] or [Mi-2, p. 9], which is just the Lie coalgebra $(E_1, \varDelta_1)$ of [Mi-7, p. 343, Definition 2], we obtain the non locally finite associative unitary coalgebra $(C, \varDelta, \varepsilon)$ defined over the ring R of integers modulo p^2, where p is a prime, by taking C to be the R-module defined by setting

$$C := R \cdot 1 \oplus \left(\bigoplus_{n=0}^{\infty} R \cdot x_n\right)$$

where $1 := 1_R$, where the x_n are indeterminates, and by taking

$$\varDelta : C \to C \otimes_R C \quad \text{and} \quad \varepsilon : C \to R$$

to be the R-module maps defined by setting

$$\varDelta(1) := 1 \otimes 1,$$

$$\begin{aligned}
&\Delta(x_0) := 1 \otimes x_0 + x_0 \otimes 1,\\
&\Delta(x_n) := 1 \otimes x_n + x_n \otimes 1 + p \cdot (x_0 \otimes x_{n+1} - x_{n+1} \otimes x_0), \quad \forall n \geqslant 1,\\
&\quad \varepsilon(1) := 1 := 1_R, \quad \text{and} \quad \varepsilon(x_n) := 0, \quad \forall n \geqslant 0.
\end{aligned}$$

REMARK. Expanding on the idea in [Ber-Hau, p. 173], we now show that, for $R := \mathbb{Z}_{p^2}$, $V := \bigoplus_{i=1}^{\infty} R \cdot x_i$, and $\Delta_V(x_i) := p \cdot (x_{i+1} \otimes x_{i+1})$, the coalgebra (V, Δ_V) is *not locally finite*. To do so, *filter V by R-submodules V_k*, one for each integer $k \geqslant 1$, defined by setting

$$V_k := \bigoplus_{i=1}^{k} R \cdot x_i.$$

Clearly, for each integer $n > 2$,

$$V_1 \subset V_2 \subset \cdots \subset V_n \subset V_{n+1} \subset \cdots$$

and

$$V = \bigcup_{k=1}^{\infty} V_k.$$

Next, for each integer $k \geqslant 1$, set

$$M_k := V_k + p \cdot V.$$

Then, for each integer $n > 2$,

$$M_1 \subset M_2 \subset \cdots \subset M_n \subset M_{n+1} \subset \cdots$$

and, because $p \cdot V \subseteq V$,

$$\begin{aligned}
V \subseteq V + p \cdot V &:= \bigcup_{k=1}^{\infty} V_k + p \cdot V = \bigcup_{k=1}^{\infty} (V_k + p \cdot V)\\
&=: \bigcup_{k=1}^{\infty} M_k = V + p \cdot V \subseteq V + V \subseteq V
\end{aligned}$$

whence

$$V = \bigcup_{k=1}^{\infty} V_k = \bigcup_{k=1}^{\infty} (V_k + p \cdot V) = \bigcup_{k=1}^{\infty} M_k.$$

To prove that (V, Δ_V) is *not* locally finite, we shall first prove that if (C, Δ_C) is any associative coalgebra with C *finitely generated* as *an* R*-module*, then, for each *morphism* $f:(C, \Delta_C) \to (V, \Delta_V)$ *of associative coalgebras*, we have that

$$f(C) \subseteq p \cdot V.$$

Suppose, therefore, that (C, Δ_C) is an associative coalgebra with C *finitely generated* as *an* R*-module*. We then have, for each of the *finitely many* R*-generators* c_j of C, that

$$f(c_j) \in V = \bigoplus_{i=1}^{\infty} R \cdot x_i.$$

Consequently, each $f(c_j)$ is an R*-linear combination of finitely many of the* x_i's so

$$f(C) \subseteq V_n \subseteq V_n + p \cdot V$$

for some *positive integer* n. If $f(C) \nsubseteq p \cdot V$, let m be the *least positive integer* n such that $f(C) \subseteq V_n + p \cdot V$. Either $m = 1$ or $m > 1$. Suppose, first, that $m > 1$. Then

$$f(C) \subseteq V_m + p \cdot V \quad \text{but} \quad f(C) \nsubseteq V_{m-1} + p \cdot V.$$

Hence, there exists $c \in C$ for which

$$f(c) \in V_m + p \cdot V \quad \text{but} \quad f(c) \notin V_{m-1} + p \cdot V.$$

For that c, we can write

$$f(c) = r_1 \cdot x_1 + \cdots + r_m \cdot x_m + p \cdot x$$

for some $x \in V$ and some $r_i \in R$ for which $r_m \neq p \cdot s$ for any $s \in R$ (for otherwise

$$f(c) = r_1 \cdot x_1 + \cdots + r_{m-1} \cdot x_{m-1} + p \cdot (s \cdot x_m + x) \in V_{m-1} + p \cdot V$$

contrary to the assumption that $f(c) \notin V_{m-1} + p \cdot V$). Since $f:(C, \Delta_C) \to (V, \Delta_V)$ is a *map of coalgebras*, we have that

$$\begin{aligned}
\Delta_V\big[f(c)\big] &= (f \otimes f)\big[\Delta_C(c)\big] = (f \otimes f)\Big[\sum_c c_1 \otimes c_2\Big] \\
&= \sum_c f(c_1) \otimes f(c_2) \in f(C) \otimes f(C) \\
&\subseteq (V_m + p \cdot V) \otimes (V_m + p \cdot V) \\
&= V_m \otimes V_m + p \cdot [V_m \otimes V + V \otimes V_m] + (p^2 \cdot V) \otimes V \\
&= V_m \otimes V_m + p \cdot [V_m \otimes V + V \otimes V_m] + 0.
\end{aligned}$$

Because

$$f(c) = r_m \cdot x_m + r_{m-1} \cdot x_{m-1} + \cdots + r_1 \cdot x_1 + p \cdot x$$

with $r_m \in R$ *not divisible* by p, it follows that

$$\Delta_V\big[f(c)\big] = r_m \cdot p \cdot (x_{m+1} \otimes x_{m+1}) + \cdots + r_1 \cdot p \cdot (x_2 \otimes x_2) + 0$$

with $r_m \cdot p$, the coefficient of $x_{m+1} \otimes x_{m+1}$ on the *left-hand-side* of the equation

$$\Delta_V\big[f(c)\big] = \sum_c f(c_1) \otimes f(c_2), \tag{$*$}$$

non-zero, divisible by p, but *not divisible by* p^2, whereas each summand $f(c_1) \otimes f(c_2)$ on the *right-hand-side* of $(*)$ is of the form

$$(s_m \cdot x_m + \cdots + s_1 \cdot x_1 + p \cdot u) \otimes (t_m \cdot x_m + \cdots + t_1 \cdot x_1 + p \cdot v) \tag{$**$}$$

with, say,

$$u := \sum_i d_i \cdot x_i \in \bigoplus_{i=1}^{\infty} R \cdot x_i =: V,$$

and

$$v := \sum_i e_i \cdot x_i \in \bigoplus_{i=1}^{\infty} R \cdot x_i =: V,$$

so that the *coefficient of* $x_{m+1} \otimes x_{m+1}$ in $(**)$ must come from

$$(p \cdot d_{m+1} \cdot x_{m+1}) \otimes (p \cdot e_{m+1} \cdot x_{m+1}) = p^2 \cdot d_{m+1} \cdot e_{m+1} \cdot (x_{m+1} \otimes x_{m+1}),$$

from which it follows that the *coefficient of* $x_{m+1} \otimes x_{m+1}$ on the *right-hand-side* of the equation $(*)$ is either 0 or is *divisible* by p^2 which is impossible since, as previously shown, the coefficient of $x_{m+1} \otimes x_{m+1}$ on the left-hand side of the equation

$$\Delta_V\big[f(c)\big] = \sum_c f(c_1) \otimes f(c_2), \tag{$*$}$$

is *not zero* and *not divisible* by p^2. In case $m = 1$,

$$f(c) = r_1 x_1 + p \cdot x$$

for some $x \in V$ and some $r_1 \in R$ with $r_1 \neq p \cdot s$ for any $s \in R$. Then

$$\Delta_V\big[f(c)\big] = r_1 \cdot p \cdot (x_2 \otimes x_2) + 0$$

with $p \cdot r_1$ *non-zero, divisible* by p, but *not divisible by* p^2 whereas the *coefficient of* $x_2 \otimes x_2$ in $f(c_1) \otimes f(c_2)$ is *either zero* or *is divisible by* p^2 since

$$f(c_1) \otimes f(c_2) = (s_1 \cdot x_1 + p \cdot u) \otimes (t_1 \cdot x_1 + p \cdot v)$$

for some $s_1, t_1 \in R$ (*not* 0 and *not divisible by* p) and some

$$u, v \in V := \bigoplus_{i=1}^{\infty} R \cdot x_i$$

with, say,

$$u := \sum_i d_i \cdot x_i \quad \text{and} \quad v := \sum_i e_i \cdot x_i,$$

and hence the coefficient of $x_2 \otimes x_2$ in the expansion of

$$(s_1 \cdot x_1 + p \cdot u) \otimes (t_1 \cdot x_1 + p \cdot v)$$

is $p^2 \cdot d_2 \cdot e_2$ (and therefore is either 0 or divisible by p^2) since it must come from

$$(p \cdot d_1 \cdot x_2) \otimes (p \cdot e_2 \cdot x_2).$$

The contradiction proves that the assumption that $f(C) \nsubseteq p \cdot V$ is untenable, thereby proving that $f(C) \subseteq p \cdot V$ for any *coalgebra map* $f : (C, \Delta_C) \to (V, \Delta_V)$ with C *finitely generated* as an R-module. From this it follows that (V, Δ_V) is *not* locally finite. For suppose, to the contrary, that (V, Δ_V) were locally finite. Then, any given x_i would lie in some *subcoalgebra* C_i *finitely generated as an* R*-submodule*. Because the *inclusion map* $C_i \hookrightarrow V$ is a *coalgebra map*, it would then follow that

$$x_i \in C_i \subseteq p \cdot V \subseteq V.$$

But this is impossible since then

$$x_i = p \cdot x = p \cdot \left(\sum_j r_j \cdot x_j \right) = \sum_j p \cdot r_j \cdot x_j$$

from which it would follow that

$$\Delta(x_i) = \sum_j p^2 \cdot r_j \cdot (x_{j+1} \otimes x_{j+1}) = 0$$

contrary to the fact that

$$\Delta(x_i) := p \cdot (x_{i+1} \otimes x_{i+1}) \neq 0.$$

REMARK. Associative unitary coalgebras over particular commutative unitary rings can turn out to be locally finite as illustrated by the following discussion based on correspondence from M. Hazewinkel. [**Note:** According to [Ber-Hau, p. 173], any associative unitary coalgebra over a semisimple Artinian ring is locally finite.]

Let R be a *principal* ideal domain, let M be a *free* R-module, and let V be a finite-dimensional K-vector subspace of W where K is the *quotient field* of R. Then, it may be shown that

$$(M \otimes M) \cap (V \otimes V) = (M \cap V) \otimes (M \cap V) \tag{5.3.1}$$

and consequently that coalgebras over a *principal ideal domain* whose underlying module is *free* are *locally finite*.

In general, formula (5.3.1) fails if R is *not* a principle ideal domain. A module M over a ring R is called *algebraically reflexive* if the canonical double dual homomorphism $\chi_M : M \to M^{**}$ is an isomorphism. Here the *algebraic dual* is taken both times (no topology!). Thus, an infinite-dimensional vector space over a field is never reflexive in this sense. But a free abelian group of countably infinite rank is algebraically reflexive.[5]

More generally, a free abelian group of cardinality $\mathbf{m}$ is algebraically reflexive if $\mathbf{m}$ is a non-ω-measurable cardinal.[6] A set is ω-measurable if it has a nonprinciple ultrafilter $\mathcal{D}$ such that for all countable sets D_n, with $n \in \mathbb{N}$ and $D_n \in \mathcal{D}$, the intersection of all the D_n is in $\mathcal{D}$. The countable set $\mathbb{N}$ is obviously non-ω-measurable. For results on reflexivity of modules over rings other than the ring of integers (and there are many of them), see *Almost Free Modules*[7] by Eklof and Mekler and/or P.C. Eklof's *Whitehead modules*, in this volume.

Using duality methods much like those in Section 4, above, one can now show that any associative unitary coalgebra, defined over a commutative unitary ring R, whose underlying R-module is *algebraically reflexive* is *locally finite*.[8]

Finally, let $CoF(M)$ be the *cofree associative unitary coalgebra on* M where M is an R-module with R an *integral domain*. Then $CoF(M)$ is *locally finite* in case M is a *free R-module* or the *dual of a free R-module* (or in a number of other cases).[9]

The general question of whether each associative unitary coalgebra whose underlying R-module is *free* (or *torsion-free*) is *locally finite* is *open*.

Acknowledgements

First and foremost, I would like to thank my typist, Meredith R. Mickel, for the superb job of TEXing this chapter including the beautiful diagrams and more revisions than she would

[5]E. Specker, *Additive Gruppen von Folgen ganzer Zahlen*, Portugaliae Math. **9** (1950), 131–140.

[6]S. Balcerzyk, *On groups of functions defined on Boolean algebras*, Fund. Math. **50** (1961/1962), 347–367.

[7]P.C. Eklof and A.H. Mekler, *Almost Free Modules. Set-Theoretic Methods*, North-Holland Mathematical Library, Vol. 46 (1990); *Almost Free Modules. Set-Theoretic Methods*, Revised edition, North-Holland Mathematical Library, Vol. 65 (2002).

[8]G. Carnovale, *Le coalgebre sull' anello degli interi*, Master's Thesis, Univ. di Roma 'La Sapienza' (1993).

[9]M. Hazewinkel, *Cofree coalgebras and multivariable recursiveness*, Preprint (2001). See [Haz-2].

care to count. Without her help, the present work would never have seen the light of day. Next, I would like to thank Jim Stasheff for critically reading the entire manuscript and making many constructive suggestions even if I did not always follow them. I would like to thank Karl Heinrich Hofmann for helpful discussions regarding the *weak-star* topology; Michiel Hazewinkel, Moss Sweedler, Richard Block, Luzius Grünenfelder, and Jörg Feldvoss for various helpful suggestions; my parents, teachers, friends, and colleagues, recent and past, for their positive influence; the Mathematics Department of the University of Washington for pleasant working conditions over the course of numerous summers; and the Tulane University Mathematics Department for the use of its library. Finally, I would like to thank Zigmantas Kryžius, Rimas Maliukevičius, Michiel Hazewinkel, Joyce Happee and all others at VTEX and Elsevier Science for their cooperation, patience, and help in the preparation of this work.

References

[Ab] Abe, E., *Hopf Algebras*, Cambridge University Press, Cambridge, UK (1980).

[Abella] Abella, A., *Grupos cuánticos compactos matriciales y grupos cuánticos generales lineales de tipo ortogonal a partir de coálgebras de Yang–Baxter*, Ph.D. Thesis, Universidad de la República, Montevideo, Uruguay (Setiembre de 1999).

[Abella-Andrus-1] Abella, A. and Andruskiewitsch, N., *Compact quantum groups arising from the* FRT-*construction*, Bol. Acad. Nacional de Ciencias, Córdoba, Argentina **63** (1999), 14–44.

[Abella-Andrus-2] Abella, A. and Andruskiewitsch, N., *Compact matrix quantum groups arising from Hermitian Yang–Baxter coalgebras*, Comm. Algebra **30** (7) (2002), 3107–3142.

[Abr-1] Abrams, L., *Frobenius algebra structures in topological quantum field theory and quantum cohomology*, Ph.D. Thesis, The Johns Hopkins University, Baltimore, MD (1997).

[Abr-2] Abrams, L., *Two-dimensional topological quantum field theories and Frobenius algebras*, J. Knot Theory Ramifications **5** (1996), 569–587.

[Abr-3] Abrams, L., *Modules, comodules, and cotensor products over Frobenius algebras*, J. Algebra **219** (1999), 201–213.

[Ada-Her-Str] Adámek, J., Herrlich, H. and Strecker, G.E., *Abstract and Concrete Categories: The Joy of Cats*, Wiley-Interscience, New York (1990).

[Ad-1] Adams, J.F., *On the nonexistence of elements of Hopf invariant one*, Bull. Amer. Math. Soc. **64** (1958), 279–282.

[Ad-2] Adams, J.F., *On the non-existence of elements of Hopf invariant one*, Ann. of Math. (2) **72** (1960), 20–104.

[Aguiar] Aguiar, M., *Internal categories and quantum groups*, Ph.D. Thesis, Cornell University, Ithaca, NY (1997).

[Al] Albert, A.A., *Modern Higher Algebra*, The University of Chicago Press, Chicago (1937).

[Al-Fer] Allen, H.P. and Ferrar, J.C., *Weighted incidence coalgebras*, J. Algebra **51** (1978), 1–11.

[Al-Tr] Allen, H.P. and Trushin, D., *Coproper coalgebras*, J. Algebra **54** (1978), 203–215.

[Al-Sw] Allen, H.P. and Sweedler, M.E., *A theory of linear descent based on Hopf algebraic techniques*, J. Algebra **12** (1969), 242–294.

[A-M-V-W] Allouch, D., Mézard, A., Vaillant, J.-C. and Weil, J., *Solutions Développées des Exercices du Traité Algèbre par S. Mac Lane et G. Birkhoff, 2^e Partie, Algèbre Linéaire*, Bordas (Gauthier-Villars) Paris (1973).

[Alp-Be] Alperin, J.L. and Bell, R.B., *Groups and Representations*, Graduate Texts in Mathematics, Vol. 162, Springer-Verlag, New York (1995).

[An] André, M., *Hopf algebras with divided powers*, J. Algebra **18** (1971), 19–50.

[Andrus] Andruskiewitsch, N., *Notes on extensions of Hopf algebras*, Canad. J. Math. **48** (1) (1996), 3–42.

[Andrus-Graña] Andruskiewitsch, N. and Graña, M., *Braided Hopf algebras over non abelian finite groups*, Bol. Acad. Nac. Cs., Córdoba, Argentina **63** (1999), 45–78.

[Andrus-Natale-1] Andruskiewitsch, N. and Natale, S., *Plancherel identity for semisimple Hopf algebras*, Comm. Algebra **25** (10) (1997), 3239–3254.

[Andrus-Natale-2] Andruskiewitsch, N. and Natale, S., *Examples of self-dual Hopf algebras*, J. Math. Sci. Univ. Tokyo **6** (1) (1999), 181–215.

[Andrus-Natale-3] Andruskiewitsch, N. and Natale, S., *Harmonic analysis on semisimple Hopf algebras*, Algebra i Analiz **12** (5) (2000) 3–27; translation: St. Petersburg Math. J. **12** (5) (2001), 713–732.

[Andrus-Natale-4] Andruskiewitsch, N. and Natale, S., *Counting arguments for Hopf algebras of low dimension*, Tsukuba J. Math. **25** (1) (2001), 187–201.

[Andru-Schn-1] Andruskiewitsch, N. and Schneider, H.-J., *Lifting of quantum linear spaces and pointed Hopf algebras of order* p^3, J. Algebra **209** (1998), 658–691.

[Andru-Schn-2] Andruskiewitsch, N. and Schneider, H.-J., *Lifting of Nichols algebras of type* A_2 *and pointed Hopf algebras of order* p^4, Hopf Algebras and Quantum Groups, S. Caenepeel et al., eds, Marcel Dekker (2000), 1–14.

[Andrus-Schnei-1] Andruskiewitsch, N. and Schneider, H.-J., *Finite quantum groups over abelian groups of prime exponent*, Ann. Sci. École Norm. Sup. (4) **35** (1) (2002), 1–26.

[Andrus-Schnei-2] Andruskiewitsch, N. and Schneider, H.-J., *Pointed Hopf algebras*, New Directions in Hopf Algebras, Montgomery, S. and Schneider, H.-J., eds, Mathematical Sciences Research Institute Publication, Vol. 43, Cambridge University Press, Cambridge, UK (2002), 1–68.

[Angwin] Angwin, G.T., *Constructing algebras from Lie and Hopf algebras*, Ph.D. Thesis, The University of Chicago, Chicago, IL (1973).

[Ap] Apostol, T., *Introduction to Analytic Number Theory*, Undergraduate Texts in Mathematics, Springer-Verlag, New York (1976).

[Arb-Man] Arbib, M.A. and Manes, E.G., *Arrows, Structures, and Functors: The Categorical Imperative*, Academic Press, New York (1975).

[A-GZ-V] Arnol'd, V.I., Gusein-Zade, S.M. and Varchenko, A.N., *Singularities of Differentiable Maps, Vol. I*, Birkhäuser Boston, Boston (1985).

[Art] Artin, M., *Algebra*, Prentice-Hall, Englewood Cliffs, NJ (1991).

[As-1] Assmus, E.F., Jr., *On the homology of local rings*, Ph.D. Thesis, Harvard University, Cambridge, MA (1958).

[As-2] Assmus, E.F., Jr., *On the homology of local rings*, Illinois J. Math. **3** (1959), 187–199.

[Av] Avramov, L.L., *The Hopf algebra of a local ring*, Math. USSR-Izv. **8** (1974), 259–284.

[Barr] Barr, M., *Coalgebras over a commutative ring*, J. Algebra **32** (1974), 600–610.

[Ba-Co-Wr] Bass, H., Connell, E.H. and Wright, D., *The Jacobian Conjecture: Reduction of degree and formal expansion of the inverse*, Bull. Amer. Math. Soc. (New Series) **7** (2) (1982), 287–330.

[Bäu-deK] Bäuerle, G.G.A. and de Kerf, E.A., *Lie Algebras, Part I, Finite and Infinite Dimensional Lie Algebras and Applications in Physics*, Studies in Mathematical Physics Series, Vol. 1, van Groessen, E. and de Jager, E.M., eds, Elsevier, Amsterdam (1990).

[B-D-G] Beattie, M., Dăscălescu, S. and Grünenfelder, L., *On the number of types of finite dimensional Hopf algebras*, Invent. Math. **136** (1999), 1–7.

[Bei-Mar-Mik] Beidar, K.I., Martindale, W.S., III and Mikhalev, A.V., *Rings with Generalized Identities*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. 196, Marcel Dekker, New York (1996).

[Bén] Bénabou, J., *Catégories avec multiplication*, C. R. Acad. Sci. Paris **256** (1963), 1887–1890.

[Ber] Bergman, G.M., *Everybody knows what a Hopf algebra is*, Group Actions on Rings, Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference on Group Actions on Rings held at Bowdoin College in Brunswick, Maine, July 8–14, 1984, Montgomery, S., ed., Contemporary Mathematics, Vol. 43, American Mathematical Society, Providence, RI (1985), 25–48.

[Ber-Hau] Bergman, G.M. and Hausknecht, A.O., *Cogroups and Co-rings in Categories of Associative Rings*, Mathematical Surveys and Monographs, Vol. 45, American Mathematical Society, Providence, RI (1996).

[Bernstein] Bernstein, I., *On co-groups in the category of graded algebras*, Trans. Amer. Math. Soc. **115** (1965), 257–269.

[B-J-N] Bhattacharya, P.B., Jain, S.K. and Nagpaul, S.R., *Basic Abstract Algebra*, 2nd edn, Cambridge University Press, Cambridge (1994).

[Bir-Mac] Birkhoff, G. and Mac Lane, S., *A Survey of Modern Algebra*, 4th edn, Macmillan, New York (1977).

[Bis-Joy] Bisson, T.P. and Joyal, A., *Q-rings and the homology of the symmetric groups*, Operads: Proceedings of Renaissance Conferences, Special Session and International Conference on Moduli Spaces, Operads, and Representation Theory/Operads and Homotopy Algebra, March 1995/May–June 1995, Hartford, Connecticut/Luminy, France, Loday, J.-L., Stasheff, J.D. and Voronov, A.A., eds, Contemporary Mathematics, Vol. 202, American Mathematical Society, Providence, RI (1997), 235–286.

[Blo] Block, R.E., *Commutative Hopf algebras, Lie coalgebras, and divided powers*, J. Algebra **96** (1) (1985), 275–306.

[Block] Block, R.E., *Determination of the irreducible divided power Hopf algebras*, J. Algebra **96** (1) (1985), 307–317.

[Blo-Gri] Block, R.E. and Griffing, G., *Recognizable formal series on trees and cofree coalgebraic systems*, J. Algebra **215** (1999), 543–573.

[Blo-Ler] Block, R.E. and Leroux, P., *Generalized dual coalgebras of algebras, with applications to cofree coalgebras*, J. Pure Appl. Algebra **36** (1985), 15–21.

[Bly] Blythe, T.S., *Module Theory, An Approach to Linear Algebra*, 2nd edn, Oxford University Press (Clarendon Press), Oxford (1990).

[Boa] Boardman, J.M., *Stable operations in generalized cohomology*, Handbook of Algebraic Topology, James, I.M., ed., Elsevier, Amsterdam (1995), 585–686 (Chapter 14).

[Boa-Joh-Wil] Boardman, J.M., Johnson, D.C. and Wilson, W.S., *Unstable operations in generalized cohomology*, Handbook of Algebraic Topology, James, I.M., ed., Elsevier, Amsterdam (1995), 687–828 (Chapter 15).

[Boca] Boca, I., *Filtrations and projective representations of Hopf algebras: Applications to quantum groups*, Ph.D. Thesis, University of Southern California, Los Angeles, CA (1995).

[B-F-G-P] Bonneau, P., Flato, M., Gerstenhaber, M. and Pinczon, G., *The hidden group structure of quantum groups: strong duality, rigidity and preferred deformations*, Comm. Math. Phys. **161** (1994), 125–156.

[Bo] Borel, A., *Linear Algebraic Groups*, W.A. Benjamin, New York (1969); *Linear Algebraic Groups*, 2nd enlarged edn, Graduate Texts in Mathematics, Vol. 126, Springer-Verlag, New York (1991).

[Bor-Gab-Ren] Borho, W., Gabriel, P. and Rentschler, R., *Primideale in Einhüllenden auflösbarer Lie-Algebren*, Lecture Notes in Mathematics, Vol. 357, Springer-Verlag, Berlin (1973).

[Bour-1] Bourbaki, N., *Éléments de Mathématique, Algèbre I, Chapitres 1 à 3*, Hermann, Paris (1970).

[Bour-2] Bourbaki, N., *Elements of Mathematics, Algebra, Part I, Chapters 1–3*, Addison-Wesley, Reading, MA (1974).

[Bour-3] Bourbaki, N., *Elements of Mathematics, Lie Groups and Lie Algebras, Part I: Chapters 1–3*, Addison-Wesley, Reading, MA (1975).

[Bri] Brieskorn, E., *Lineare Algebra und Analytische Geometrie, I,* F. Vieweg & Sohn, Braunschweig/Wiesbaden (1983).

[Broa-Krei-1] Broadhurst, D.J. and Kreimer, D., *Renormalization automated by Hopf algebra*, J. Symbolic Comput. **27** (6) (1999), 581–600.

[Broa-Krei-2] Broadhurst, D.J. and Kreimer, D., *Combinatoric explosion of renormalization tamed by Hopf algebra: 30-loop Padé–Borel resummation*, Phys. Lett. B **475** (1–2) (2000), 63–70.

[B-t.D] Bröcker, T. and tom Dieck, T., *Representations of Compact Lie Groups*, Graduate Texts in Mathematics, Vol. 98, Springer-Verlag, New York (1985).

[Brow] Browder, W., *On differential Hopf algebras*, Trans. Amer. Math. Soc. **107** (1963), 153–176.

[Bro] Brown, E.H., Jr., *Twisted tensor products, I*, Ann. of Math. (2) **69** (1959), 223–246.

[Brzez] Brzeziński, T., *On modules associated to coalgebra Galois extensions*, J. Algebra **215** (1999), 290–317.

[Buc-Del] Bucur, I. and Deleanu, A., *Introduction to the Theory of Categories and Functors*, Pure and Applied Mathematics: A Series of Texts and Monographs, Vol. XIX, A Wiley-Interscience Publication, John Wiley & Sons Ltd., London (1968).

[Burton] Burton, C.L., *Hopf algebras and Dieudonné modules*, Ph.D. Thesis, University of Washington, Seattle, WA (1998).

[Cae] Caenepeel, S., *Brauer Groups, Hopf Algebras and Galois Theory*, K-Monographs in Mathematics, Vol. 4, Kluwer Academic Publishers, Dordrecht (1998).

[Caen-Dăsc-1] Caenepeel, S. and Dăscălescu, S., *Pointed Hopf algebras of dimension* p^3, J. Algebra **209** (1998), 622–634.

[Caen-Dăsc-2] Caenepeel, S. and Dăscălescu, S., *On pointed Hopf algebras of dimension* 2^n, Bull. London Math. Soc. **31** (1999), 17–24.

[Caen-Mili-Zhu] Caenepeel, S., Militaru, G. and Zhu, S., *Frobenius and Separable Functors for Generalized Module Categories and Nonlinear Equations*, Lecture Notes in Mathematics, Vol. 1787, Springer-Verlag, Berlin (2002).

[Cartan] Cartan, H., *Puissances divisées*, Séminaire Henri Cartan, 7e année, 1954/1955, Algèbres d'Eilenberg–MacLane et Homotopie, Ecole Normale Supérieure, Paris (1956), Expose n°7 (3-1-1955), 7-01–7-11.

[C-E] Cartan, H. and Eilenberg, S., *Homological Algebra*, Princeton Mathematical Series, Vol. 19, Princeton University Press, Princeton, New Jersey (1956).

[Cartier] Cartier, P., *Dualité de Tannaka des groupes et des algèbres de Lie*, C.R. Acad. Sci. Paris **242** (1956), 322–325.

[Ca-1] Cartier, P., *Hyperalgèbres et groupes de Lie formels*, Séminaire "Sophus Lie", 2e année: 1955/56, especially, Exposé n°2 (*Hyperalgèbres et groupes formules*, pp. 2-01–2-06), Exposé n°3 (*Exemples d'hyperalgèbres*, pp. 3-01–3-15) and Exposé n°4 (*Modules sur une coalgèbre*, pp. 4-01–4-07).

[Ca-2] Cartier, P., *Groupes algébriques et groupes formels*, Colloque sur la Théorie des Groupes Algébriques Tenu à Bruxelles du 5 au 7 Juin 1962, Centre Belge de Recherches Mathématiques (CBRM), Librairie Universitaire, Louvain-Belgique, et Librairie Gauthier-Villars, Paris, France (1962), 87–111.

[Ca-3] Cartier, P., *An Introduction to Quantum Groups,* Proceedings of Symposia in Pure Mathematics, Vol. 56, Part 2, Algebraic Groups and Their Generalizations: Quantum and Infinite-Dimensional Methods, Summer Research Institute on Algebraic Groups and Their Generalizations, July 6–26, 1991, Pennsylvania State University, University Park, Pennsylvania, Haboush, W.J. and Parshall, B.J., eds, American Mathematical Society, Providence, RI (1994).

[Chai-Demi] Chaichian, M. and Demichev, A., *Introduction to Quantum Groups*, World Scientific, Singapore (1996).

[Char-Pres] Chari, V. and Pressley, A., *A Guide to Quantum Groups*, Cambridge University Press, Cambridge, UK (1994).

[Chas-Swee] Chase, S.U. and Sweedler, M.E., *Hopf Algebras and Galois Theory*, Lecture Notes in Mathematics, Vol. 97, Springer-Verlag, Berlin (1969).

[Chevalley-1] Chevalley, C., *Theory of Lie Groups: I*, Princeton Mathematical Series, Vol. 8, Princeton University Press, Princeton, NJ (1946).

[Chevalley-2] Chevalley, C., *Fundamental Concepts of Algebra*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. VII, Academic Press Inc., New York (1956).

[Cheng] Cheng, Y., *Hopf algebras with antipodes of finite orders*, Ph.D. Thesis, Rutgers University, New Brunswick, NJ (1984).

[Chil] Childs, L.N., *Taming Wild Extensions: Hopf Algebras and Local Galois Module Theory*, Mathematical Surveys and Monographs, Vol. 80, American Mathematical Society, Providence, RI (2000).

[C-G-M-S-Z] Childs, L.N., Greither, C., Moss, D.J., Sauerberg, J. and Zimmermann, K., *Hopf Algebras, Polynomial Formal Groups, and Raynaud Orders*, Mem. Amer. Math. Soc., Vol. 136, no. 651, American Mathematical Society, Providence, RI (1998).

[Chin-1] Chin, W., *Spectra of smash products*, Israel J. Math. **72** (1–2) (1990), 84–98.

[Chin-2] Chin, W., *Crossed products of semisimple cocomutative Hopf algebras*, Proc. Amer. Math. Soc. **116** (2) (1992), 321–327.

[Chin-3] Chin, W., *Hereditary and path coalgebras*, Comm. Algebra **30** (4) (2002), 1829–1831.

[Chin-Gold] Chin, W. and Goldman, J., *Bialgebras of linearly recursive sequences*, Comm. Algebra **21** (11) (1993), 3935–3952.

[Chin-Krop] Chin, W. and Krop, L., *Injective comodules for* 2×2 *quantum matrices*, Comm. Algebra **28** (4) (2000), 2043–2057.

[Chin-Mont] Chin, W. and Montgomery, S., *Basic coalgebras*, Modular Interfaces: Modular Lie Algebras, Quantum Groups, and Lie Superalgebras, Chari, V. and Penkov, I.B., eds, AMS/IP Studies in Advanced Mathematics, Vol. 4N, American Mathematical Society, Providence, RI, International Press, Cambridge, MA (1997), 41–47.

[Chin-Muss] Chin, W. and Musson, I.M., *Hopf algebra duality, injective modules and quantum groups*, Comm. Algebra **22** (12) (1994), 4661–4692.

[Chin-Quin] Chin, W. and Quinn, D., *Actions of divided power Hopf algebras*, J. Algebra **144** (2) (1991), 371–389.

[Cibils] Cibils, C., *A quiver quantum group*, Comm. Math. Phys. **157** (3) (1993), 459–477.

[Cibils-Rosso-1] Cibils, C. and Rosso, M., *Algèbres des chemins quantiques*, Adv. in Math. **125** (2) (1997), 171–199.

[Cibils-Rosso-2] Cibils, C. and Rosso, M., *Hopf bimodules are modules*, J. Pure Appl. Algebra **128** (3) (1998), 225–231.

[Cibils-Rosso-3] Cibils, C. and Rosso, M., *Hopf quivers*, J. Algebra **254** (2) (2002), 241–251.

[Coh-1] Cohen, M., *Smash products, inner actions and quotient rings*, Pacific J. Math. **125** (1) (1986), 45–66.

[Coh-2] Cohen, M., *Hopf algebra actions – revisited*, Deformation Theory and Quantum Groups with Applications to Mathematical Physics (Proceedings of an AMS-IMS-SIAM 1990 Joint Summer Research Conference, June 14–20, 1990 University of Massachusetts, Amherst with support from the National Science Foundation), Gerstenhaber, M. and Stasheff, J., eds, Contemporary Mathematics, Vol. 134, American Mathematical Society, Providence, RI (1992), 1–18.

[Coh-Fis-1] Cohen, M. and Fishman, D., *Hopf algebra actions*, J. Algebra **100** (2) (1986), 363–379.

[Coh-Fis-2] Cohen, M. and Fishman, D., *Semisimple extensions and elements of trace 1*, J. Algebra **149** (2) (1992), 419–437.

[Coh-Fis-Mon] Cohen, M., Fishman, D. and Montgomery, S., *Hopf Galois extensions, smash products, and Morita equivalence*, J. Algebra **133** (2) (1990), 351–372.

[Coh-Mon] Cohen, M. and Montgomery, S., *Group-graded rings, smash products, and group actions*, Trans. Amer. Math. Soc. **282** (1) (1984), 237–258.

[Cohn] Cohn, P.M., *Sur le critère de Friedrichs pour les commutateurs dans une algèbre associative libre*, C. R. Acad. Sci. Paris **239** (1954), 743–745.

[Conn-Krei-1] Connes, A. and Kreimer, D., *Lessons from quantum field theory: Hopf algebras and spacetime geometries*, Moshé Flato (1937–1998), Lett. Math. Phys. **48** (1) (1999), 85–96.

[Conn-Krei-2] Connes, A. and Kreimer, D., *From local perturbation theory to Hopf and Lie algebras of Feynman graphs*, Mathematical Physics in Mathematics and Physics (Siena, 2000), Fields Inst. Commun., Vol. 30, Amer. Math. Soc., Providence, RI (2001), 105–114.

[Conn-Krei-3] Connes, A. and Kreimer, D., *From local perturbation theory to Hopf and Lie algebras of Feynman graphs*, EuroConférence Moshé Flato 2000, Part I (Dijon), Lett. Math. Phys. **56** (1) (2001), 3–15.

[Cor-Gel] Corwin, L. and Gelfand, I.M., *Hopf algebra structures for the Heisenberg algebra: I*, The Gelfand Mathematical Seminars, 1990–1992, Corwin, L., Gelfand, I. and Lepowsky, J., eds, Birkhaüser, Boston (1993), 10–17.

[Cox-Paul] Cox-Paul, M.P., *The image of the Picard invariant map for Hopf Galois extensions*, Ph.D. Thesis, State University of New York at Albany, Albany, NY (1994).

[C-R-1] Curtis, C.W. and Reiner, I., *Representation Theory of Finite Groups and Associative Algebras*, Pure and Applied Mathematics: A Series of Texts and Monographs, Vol. XI, Wiley-Interscience, New York (1962).

[C-R-2] Curtis, C.W. and Reiner, I., *Methods of Representation Theory with Applications to Finite Groups and Orders*, Vol. I, Wiley-Interscience, New York (1981).

[Dăs-Năs-Rai] Dăscălescu, S., Năstăsescu, C. and Raianu, Ş., *Hopf Algebras: An Introduction*, Pure and Applied Mathematics Series of Monographs and Textbooks, Vol. 235, Marcel Dekker, New York (2001).

[Deligne-Milne] Deligne, P. and Milne, J.S., *Tannakien Categories*, Lecture Notes in Mathematics, Vol. 900, Springer-Verlag, Berlin (1982).

[Diarra-1] Diarra, B., *La dualité de Tannaka dans les corps valués ultramétriques complets*, C. R. Acad. Sci. Paris Sér. A **279** (1974), 907–909.

[Diarra-2] Diarra, B., *Algèbres de Hopf ultramétriques complètes*, Proceedings of the Conference on p-adic Analysis held in Nijmegen, January 16–20, 1978, Schikhof, W., ed., Report 7806, Katholieke Universiteit, Mathematisch Instituut, Nijmegen, The Netherlands (1978), 57–71.

[Diarra-3] Diarra, B., *Algèbres de Hopf et fonctions presque périodiques ultramétriques*, Séminaire d'Analyse, 6, 7. 1990–1991, 1991–1992, Papers from the seminars held at the Université Blaise Pascal, Aubière, 1990–1991 and 1991–1992, Université Blaise Pascal, Laboratoires de Mathématiques Pures et Appliquées, Clermont-Ferrand, Exp. No. 6.7 (1995), 30 pp.

[Diarra-4] Diarra, B., *Representative subalgebra of a complete ultrametric Hopf algebra*, Proceedings of the 3rd International Conference on p-adic Functional Analysis held in Aubière, April 24–29, 1994, Université Blaise Pascal, Laboratoires de Mathématiques Pures et Appliquées, Aubière, Ann. Math. Blaise Pascal **2** (1) (1995), 99–115.

[Diarra-5] Diarra, B., *On the definition of the dual Lie coalgebra of a Lie algebra*, Publ. Mat. **39** (2), (1995), 349–354.

[Diarra-6] Diarra, B., *Algèbres de Hopf et fonctions presque périodiques ultramétriques*, Riv. Mat. Pura Appl. **17** (1995), 113–132.

[Diarra-7] Diarra, B., *Complete ultrametric Hopf algebras which are free Banach spaces*, p-adic Functional Analysis (Nijmegen, 1996), Proceedings of the 4th International Conference held at the University of Nijmegen, Nijmegen, June 3–7, 1996, Schikhof, W. H., Perez-Garcia, C. and Kąkol, J., eds, Lecture Notes in Pure and Applied Mathematics, Vol. 192, Marcel Dekker, New York (1997), 61–80.

[Diarra-8] Diarra, B., *On the integrity of the dual algebra of some complete ultrametric Hopf algebras*, p-adic Functional Analysis (Poznań, 1998), Proceedings of the 5th International Conference held at Adam Mickiewicz University, Poznań, 1998, Kąkol, J., De Grande-De Kimpe, N. and Perez-Garcia, C., eds, Lecture Notes in Pure and Applied Mathematics, Vol. 207, Marcel Dekker, New York (1999), 45–64.

[Diarra-9] Diarra, B., *Continuous coderivations of complete ultrametric Hopf algebras*, p-adic Functional Analysis (Ioannina, 2000), Proceedings of the 6th International Conference held at the University of Ioannina, Ioannina, July 3–7, 2000, Katsaras, A.K., Schikhof, W.H. and Van Hamme, L., eds, Lecture Notes in Pure and Applied Mathematics, Vol. 222, Marcel Dekker, New York (2001), 75–90.

[Dem-Gab-1] Demazure, M. and Gabriel, P., *Groupes Algébriques*, Tome I, Masson & Cie, Paris (1970).

[Dem-Gab-2] Demazure, M. and Gabriel, P., *Introduction to Algebraic Geometry and Algebraic Groups*, Translation of Chapters 1 & 2 of [Dem-Gab-1], North-Holland Mathematics Studies, Vol. 39, North-Holland, Amsterdam (1970).

[Didt-1] Didt, D., *Differential calculus on inhomogeneous quantum groups*, Proceedings of the 1999 Euroconference: On Non-commutative Geometry and Hopf Algebras in Field Theory and Particle Physics held in Torino, September 20–30, 1999, Castellani, L., Landi, G. and Lizzi, F., eds, Internat. J. Modern Phys. B **14** (22–23), World Scientific, Singapore (2000), 2299–2303.

[Didt-2] Didt, D., *Linkable Dynkin diagrams*, J. Algebra **255** (2) (2002), 373–391.

[Didt-3] Didt, D., *Linkable Dynkin diagrams and quasi-isomorphisms for finite dimensional pointed Hopf algebras*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (2003).

[D-1] Dieudonné, J., *Linearly compact spaces and double vector spaces over sfields*, Amer. J. Math. **73** (1951), 13–19.

[D-2] Dieudonné, J., *Introduction to the Theory of Formal Groups*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. 20, Marcel Dekker, New York (1973).

[D-3] Dieudonné, J., *A History of Algebraic and Differential Topology 1900–1960*, Birkhäuser, Boston, Basel (1989).

[Dit] Ditters, E.J., *Curves and exponential series in the theory of noncommutative formal groups*, Ph.D. Thesis, University of Nijmegen, Nijmegen, The Netherlands (1969).

[Dixmier-1] Dixmier, J., *C^*-algebras et Leur Représentations*, Cahiers Scientifiques, Fasc. XXIX: 2[éme] édition, Bordas (Gauthier-Villars), Paris (1969).

[Dixmier-2] Dixmier, J., *Algèbres Enveloppantes*, Cahiers Scientifiques, Fasc. XXXIX: Fasc. XXXVII, Bordas (Gauthier-Villars), Paris (1974).

[Dix] Dixmier, J., *Enveloping Algebras*, North-Holland Mathematical Library, Vol. 14, North-Holland, Amsterdam (1977).

[Doi] Doi, Y., *Homological coalgebra*, J. Math. Soc. Japan **33** (1) (1981), 31–50.

[Dold] Dold, A., *Lectures on Algebraic Topology*, 2nd edn, Grundlehren der mathematischen Wissenschaften, Vol. 200, Springer-Verlag, Berlin (1980).

[Dou] Doubilet, P., *A Hopf algebra arising from the lattice of partitions of a set*, J. Algebra **28** (1974), 127–132.

[D-R-S] Doubilet, P., Rota, G.-C. and Stanley, R., *The idea of a generating function*, Finite Operator Calculus, Rota, G.-C., ed., Academic Press, New York (1975), 83–134. Originally: *On the foundations of combinatorial theory, VI: The idea of a generating function*, Proceedings 6th Berkeley Symposium on Mathematical Statistics and Probability, Vol. II, Probability Theory (1972), 267–318.

[Dr-1] Drinfel'd, V.G., *Hopf algebras and the quantum Yang–Baxter equation*, Soviet Math. Dokl. **32** (1985), 254–258.

[Dr-2] Drinfel'd, V.G., *Quantum groups*, Proceedings International Congress of Mathematicians, August 3–11, 1986, Berkeley, CA, Vol. 1, Gleason, A.M., ed., Amer. Math. Soc., Providence, RI (1987), 798–820.

[Dr-3] Drinfel'd, V.G., *A new realization of Yangians and quantized affine algebras*, Soviet Math. Dokl. **36** (1988), 212–216.

[Dr-4] Drinfel'd, V.G., *On almost cocommutative Hopf algebras*, Leningrad Math. J. **1** (1990), 321–342.

[Dr-5] Drinfel'd, V.G., *Quasi-Hopf algebras*, Leningrad Math. J. **1** (1990), 1419–1457.

[Duf] Duflot, J., *Lots of Hopf algebras*, J. Algebra **204** (1998), 69–94.

[Dum-Foo] Dummit, D.S. and Foote, R.M., *Abstract Algebra*, 2nd edn, Prentice-Hall, Upper Saddle River, NJ (1999).

[Dy] Dyer, E., *Cohomology Theories*, W.A. Benjamin, New York (1969).

[Eberwein] Eberwein, M., *Cosemisimple Hopf algebras*, Ph.D. Thesis, The Florida State University, Tallahassee, FL (1992).

[E] Eckmann, B., *Selecta*, Knus, M.-A., Mislin, G. and Stammbach, U., eds, Springer-Verlag, Berlin (1987), 392–406 and 407–435.

[E-H-1] Eckmann, B. and Hilton, P.J., *Structure maps in group theory*, Fund. Math. **50** (1961), 207–221.

[E-H-2] Eckmann, B. and Hilton, P.J., *Group-like structures in general categories I. Multiplications and comultiplications*, Math. Ann. **145** (1962), 227–255.

[E-H-3] Eckmann, B. and Hilton, P.J., *Group-like structures in general categories III. Primitive categories*, Math. Ann. **150** (1963), 165–187.

[E-H-4] Eckmann, B. and Hilton, P.J., *Group-like structures in general categories II. Equalizers, limits, lengths*, Math. Ann. **151** (1963), 150–186.

[Eil-Kel] Eilenberg, S. and Kelly, G.M., *Closed categories*, Proceedings of the Conference on Categorical Algebra at La Jolla, held June 7–12, 1965, Eilenberg, S., Harrison, D.K., Mac Lane, S. and Röhrl, H., eds, Springer-Verlag, New York (1965), 421–562.

[Eil-Moo-1] Eilenberg, S. and Moore, J.C., *Adjoint functors and triples*, Illinois J. Math. **9** (1965), 381–398.

[Eil-Moo-2] Eilenberg, S. and Moore, J.C., *Homology and fibrations I. Coalgebras, cotensor product and its derived functors*, Comment. Math. Helv. **40** (1965–66), 199–236.

[Enoc-Schw] Enock, M. and Schwartz, J.-M., *Kac Algebras and Duality of Locally Compact Groups*, Springer-Verlag, Berlin (1992).

[Er] Ernest, J., *Hopf–von Neumann algebras*, Functional Analysis (Proceedings of a Conference held at the University of California, Irvine), Gelbaum, B.R., ed., Thompson Book Company, Washington, DC, Academic Press, London (1967), 195–215.

[E-G-1] Etingof, P. and Gelaki, S., *Some properties of finite-dimensional semisimple Hopf algebras*, Math. Res. Lett. **5** (1998), 191–197.

[E-G-2] Etingof, P. and Gelaki, S., *Semisimple Hopf algebras of dimension pq are trivial*, J. Algebra **210** (1998), 664–669.

[Etin-Schi] Etingof, P. and Schiffmann, O., *Lectures on Quantum Groups*, Lectures in Mathematical Physics, International Press Incorporated, Boston, Cambridge, MA (1998).

[Ey] Eyre, T.M.W., *Quantum Stochastic Calculus and Representations of Lie Superalgebras*, Lecture Notes in Mathematics, Vol. 1692, Springer-Verlag, Berlin (1998).

[F-R-T] Faddeev, L., Reshetikhin, N. and Takhtajan, L., *Quantum groups*, Braid Group, Knot Theory and Statistical Mechanics, Yang, C.N. and Ge, M.-L., eds, Advanced Series in Mathematical Physics, Vol. 9, World Scientific, Singapore (1989), 97–110.

[Farn] Farnsteiner, R., *On Frobenius extensions defined by Hopf algebras*, J. Algebra **166** (1) (1994), 130–141.

[Feld] Feldvoss, J., *Representations of Lie colour algebras*, Adv. in Math. **157** (2) (2001), 95–137.

[Fe-Ha-Th-1] Felix, Y., Halperin, S. and Thomas, J.-C., *Adams' cobar equivalence*, Trans. Amer. Math. Soc. **329** (1992), 531–549.

[Fe-Ha-Th-2] Felix, Y., Halperin, S. and Thomas, J.-C., *Differential graded algebras in topology*, Handbook of Algebraic Topology, James, I.M., ed., Elsevier, Amsterdam (1995), 829–865 (Chapter 16).

[FS-1] Ferrer Santos, W.R., *Cohomology of comodules*, Ph.D. Thesis, University of California, Berkeley, CA (1980).

[FS-2] Ferrer Santos, W.R., *Cohomology of comodules*, Pacific J. Math. **109** (1) (1983), 179–213.

[Fischman] Fischman, D.D., *Actions of Hopf algebras*, Ph.D. Thesis, Ben Gurion University of the Negev, Beer Sheva, Israel (1989).

[F-M-S] Fischman, D., Montgomery, S. and Schneider, H.-J., *Frobenius extensions of subalgebras of Hopf algebras*, Trans. Amer. Math. Soc. **349** (1997), 4857–4895.

[Fl] Fliess, M., *Matrices de Hankel*, J. Math. Pures Appl. **53** (1974), 197–224.

[Fog] Fogarty, J., *Invariant Theory*, W. A. Benjamin, New York (1969).

[Fox-Mar] Fox, T.F. and Markl, M., *Distributive laws, bialgebras, and cohomology*, Operads: Proceedings of Renaissance Conferences (Special Session and International Conference on Moduli Spaces, Operads, and Representation Theory/Operads and Homotopy Algebra, March 1995/May–June 1995, Hartford, Connecticut/Luminy, France), Loday, J.-L., Stasheff, J.D. and Voronov, A.A., eds, Contemporary Mathematics, Vol. 202, American Mathematical Society, Providence, RI (1997), 167–205.

[Fr] Freyd, P., *Abelian Categories: An Introduction to the Theory of Functors*, Harper's Series in Modern Mathematics, Harper & Row, New York (1964).

[Fr-Sc] Freyd, P.J. and Scedrow, A., *Categories, Allegories*, North-Holland Mathematical Library, Vol. 39, Elsevier Science, Amsterdam (1990).

[Frie] Friedrichs, K.O., *Mathematical aspects of the quantum theory of fields, Part V*, Comm. Pure and Appl. Math. **6** (1953), 1–72.

[Ga] Gabriel, P., *Etude infinitésimale des schémas en groupes; Exposé VIII$_B$: Groupes formels, Schémas en groupes, Tome I: Propriétés générales des schémas en groupes*, Séminaire de Géométrie Algébrique du Bois-Marie 1962/64 (SGA 3) (Un séminaire dirigé par M. Demazure et A. Grothendieck avec la collaboration de M. Artin, J. Bertin, P. Gabriel, M. Raynaud et J.P. Serre), Lecture Notes in Mathematics, Vol. 151, Springer-Verlag, Berlin (1970), 476–562.

[Gelaki-1] Gelaki, S., *Hopf algebras, quantum groups and invariants of knots and* 3*-manifolds*, Ph.D. Thesis, Ben Gurion University of the Negev, Beer Sheva, Israel (1997).

[Gel] Gelaki, S., *Pointed Hopf algebras and Kaplansky's 10th conjecture*, J. Algebra **209** (1998), 635–657.

[Gelaki-2] Gelaki, S., *On the classification of finite-dimensional Hopf algebras*, New Directions in Hopf Algebras, Montgomery, S. and Schneider, H.-J., eds, Mathematical Sciences Research Institute Publication, Vol. 43, Cambridge University Press, Cambridge, UK (2002), 69–116.

[Giaq] Giaquinto, A., *Deformation methods in quantum groups*, Ph.D. Thesis, University of Pennsylvania, Philadelphia, PA (1991).

[Gol] Goldblatt, R., *Topoi: The Categorical Analysis of Logic*, rev. edn, Studies in Logic and the Foundations of Mathematics, Vol. 98, North-Holland, Elsevier Science, Amsterdam (1984).

[Goo-Wal] Goodman, R. and Wallach, N.R., *Representations and Invariants of the Classical Groups*, Encyclopedia of Mathematics and Its Applications, Vol. 68, Rota, G.-C., ed., Cambridge University Press, Cambridge, UK (1998).

[G-K-P] Graham, R.L., Knuth, D.E. and Patashnik, O., *Concrete Mathematics: A Foundation for Computer Science*, 2nd edn, Addison-Wesley, Reading, MA (1994).

[Graña-1] Graña, M., *Álgebras de Nichols y álgebras de Hopf punteadas de dimensión finita*, Ph.D. Thesis, Universidad de Buenos Aires, Buenos Aires, Argentina (May 29, 2000).

[Graña-2] Graña, M., *On Nichols algebras of low dimension*, New Trends in Hopf Algebra Theory. Proceedings of the Colloquium on Quantum Groups and Hopf Algebras held in La Falda, August 9–13, 1999, Andruskiewitsch, N., Ferrer Santos, W.R. and Schneider, H.-J., eds, Contemporary Mathematics, Vol. 267, Amer. Math. Soc., Providence, RI (2000), 111–134.

[Graña-3] Graña, M., *A freeness theorem for Nichols algebras*, J. Algebra **231** (1) (2000), 235–257.

[Graña-4] Graña, M., *On pointed Hopf algebras of dimension p^5*, Glasgow Math. J. **42** (3) (2000), 405–419.

[Graña-5] Graña, M., *Pointed Hopf algebras of dimension 32*, Comm. Algebra **28** (6) (2000), 2935–2976.

[Gray] Gray, M., *A Radical Approach to Algebra*, Addison-Wesley, Reading, MA (1970).

[Greipel] Greipel, K.-P., *Induzierte Darstellungen punktierter Hopfalgebren*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1985).

[Griffing] Griffing, G., *Cofree Lie and other coalgebras*, Ph.D. Thesis, University of California, Riverside, CA (1986).

[Gri-Hil] Griffiths, H.B. and Hilton, P.J., *A Comprehensive Textbook of Classical Mathematics; A Contemporary Interpretation*, Springer-Verlag, New York (1978); originally published by the Van Nostrand Reinhold Company (1970).

[Gril] Grillet, P.A., *Algebra,* Wiley-Interscience, New York (1999).

[Gr-1] Grünenfelder, L., *Über die Struktur von Hopf-Algebren*, Dissertation No. 4194, ETH Zürich (1969).

[Gr-2] Grünenfelder, L., *Hopf-Algebren und Coradikal*, Math. Z. **116** (1970), 166–182.

[Gr-Ko-1] Grünenfelder, L. and Košir, T., *An algebraic approach to multiparameter spectral theory*, Trans. Amer. Math. Soc. **348** (1996), 2983–2998.

[Gr-Ko-2] Grünenfelder, L. and Košir, T., *Coalgebras and spectral theory in one and several variables*, Oper. Th. Adv. Appl. **87** (1996), 177–192.

[Gr-Ko-3] Grünenfelder, L. and Košir, T., *Koszul cohomology for finite families of comodule maps and applications*, Comm. Algebra **25** (1997), 459–479.

[Gr-Om] Grünenfelder, L. and Omladič, M., *Linearly recursive sequences and operator polynomials*, Linear Alg. Appl. **182** (1993), 127–145.

[Gr-P] Grünenfelder, L. and Paré, R., *Families parametrized by coalgebras*, J. Algebra **107** (1987), 316–375.

[Gug] Gugenheim, V.K.A.M., *On extensions of algebras, co-algebras and Hopf algebras, I*, Amer. J. Math. **84** (2) (1962), 349–382.

[Gugen] Gugenheim, V.K.A.M., *Cohomology theory in the category of Hopf algebras*, Colloque de Topologie Tenu à Bruxelles du 7 au 10 Septembre 1964, Centre Belge de Recherches Mathématiques (CBRM), Librairie Universitaire, Louvain–Belgique, et Librairie Gauthier-Villars, Paris (1966), 137–148.

[G-L] Gugenheim, V.K.A.M. and Lambe, L.A., *Perturbation theory in differential homological algebra. I*, Illinois J. Math. **33** (4) (1989), 566–582.

[G-S] Gugenheim, V.K.A.M. and Stasheff, J.D., *On perturbations and A_∞-structures*, Bull. Soc. Math. Belg. Sér. A **38** (1986), 237–246.

[Gui] Guichardet, A., *Groupes Quantiques: Introduction au Point de Vue Formel*, Savoirs Actuels, InterÉditions/CNRS Éditions, Paris (1995).

[Gul-Lev] Gulliksen, T.H. and Levin, G., *Homology of Local Rings*, Queen's Papers in Pure and Applied Mathematics, No. 20, Queen's University, Kingston, ON (1969).

[Günther] Günther, R., *Verschränkte Produkte für punktierte Hopfalgebren*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1999).

[Hai-Sch] Haiman, M. and Schmitt, W., *Incidence algebra antipodes and Lagrange inversion in one and several variables*, J. Combin. Theory A **50** (1989), 172–185.

[Hain-1] Hain, R.M., *Iterated integrals, minimal models and rational homotopy theory*, Ph.D. Thesis, University of Illinois at Urbana–Champaign, Urbana, IL (1980).

[Hain-2] Hain, R.M., *Iterated integrals and homotopy periods*, Mem. Amer. Math. Soc. **47** (291) (third of six numbers), Providence, RI (1984).

[Hal] Halpern, E., *Twisted Polynomial Hyperalgebras*, Mem. Amer. Math. Soc., Vol. 29, Providence, RI (1958).

[H-C] Harish-Chandra, *On representations of Lie algebras*, Ann. of Math. (2) **50** (1949), 900–915.

[Hausknecht] Hausknecht, A.O., *Coalgebras in categories of associative algebras*, Ph.D. Thesis, University of California, Berkeley, CA (1975).

[Haz-1] Hazewinkel, M., *Formal Groups and Applications*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. 78, Academic Press, New York (1978).

[Haz-2] Hazewinkel, M., *Cofree coalgebras and multivariable recursiveness*, J. Pure Appl. Algebra, to appear.

[He] Henderson, D.W., *A short proof of Wedderburn's theorem*, Amer. Math. Monthly **72** (1965), 385–386.

[Herr-Stre] Herrlich, H. and Strecker, G.E., *Category Theory: An Introduction*, Allyn and Bacon Inc., Boston (1973); 2nd edn, Heldermann Verlag, Berlin (1979).

[Hey-Rad] Heyneman, R.G. and Radford, D.E., *Reflexivity and coalgebras of finite type*, J. Algebra **28** (2) (1974), 215–246.

[Hey-Sw-1] Heyneman, R.G. and Sweedler, M.E., *Affine Hopf algebras, I*, J. Algebra **13** (2) (1969), 192–241.

[Hey-Sw-2] Heyneman, R.G. and Sweedler, M.E., *Affine Hopf algebras, II*, J. Algebra **16** (2) (1970), 271–297.

[Hil] Hilton, P., *Homotopy Theory and Duality*, Notes on Mathematics and its Applications, Gordon and Breach Science Publishers, New York (1965).

[Hil-Sta] Hilton, P.J. and Stammbach, U., *A Course in Homological Algebra*, Graduate Texts in Mathematics, Vol. 4, Springer-Verlag, New York (1971).

[Hil-Wu] Hilton, P. and Wu, Y.-C., *A Course in Modern Algebra*, Wiley-Interscience, New York (1974).

[Hirs-Raph] Hirschhorn, P.S. and Raphael, L.A., *Coalgebraic foundations of the method of divided differences*, Adv. in Math. **91** (1) (1992), 75–135.

[Hoch] Hochschild, G., *Representation Theory of Lie Algebras* (Notes by G. Leger), The University of Chicago, Department of Mathematics, Lecture Notes in Mathematics, Chicago, IL (1959).

[H-1] Hochschild, G., *Algebraic Lie algebras and representative functions*, Illinois J. Math. **3** (1959), 499–523.

[H-1*] Hochschild, G., *Algebraic Lie algebras and representative functions, Supplement*, Illinois J. Math. **4** (1960), 609–618.

[H-2] Hochschild, G., *The Structure of Lie Groups*, Holden-Day, San Francisco, CA (1965).

[H-3] Hochschild, G., *Algebraic groups and Hopf algebras*, Illinois J. Math. **14** (1970), 52–65.

[H-4] Hochschild, G., *Introduction to Affine Algebraic Groups*, Holden-Day, San Francisco, CA (1971).

[H-5] Hochschild, G., *Basic Theory of Algebraic Groups and Lie Algebras*, Graduate Texts in Mathematics, Vol. 75, Springer-Verlag, New York (1981).

[Hoc-Mos-1] Hochschild, G. and Mostow, G.D., *Representations and representative functions of Lie groups*, Ann. of Math. (2) **66** (1957), 495–542.

[Hoc-Mos-2] Hochschild, G. and Mostow, G.D., *Representations and representative functions of Lie groups, II*, Ann. of Math. (2) **68** (1958), 295–313.

[Hoc-Mos-3] Hochschild, G. and Mostow, G.D., *Representations and representative functions of Lie groups, III*, Ann. of Math. (2) **70** (1959), 85–100.

[Hoc-Mos-4] Hochschild, G. and Mostow, G.D., *On the algebra of representative functions of an analytic group*, Amer. J. Math. **83** (1961), 111–136.

[Hof] Hofmann, K.H., *The Duality of Compact Semigroups and C^*-Bialgebras*, Lecture Notes in Mathematics, Vol. 129, Springer-Verlag, Berlin (1970).

[Hof-Mos-1] Hofmann, K.H. and Mostert, P.S., *Elements of Compact Semigroups*, Charles E. Merrill Books Inc., Columbus, OH (1966).

[Hof-Mos-2] Hofmann, K.H. and Mostert, P.S., *Cohomology Theories for Compact Abelian Groups* (*With an Appendix by Nummela, E.C.*), Springer-Verlag, Berlin (1973).

[Hof-Mor] Hofmann, K.H. and Morris, S.A. *The Structure of Compact Groups: A Primer for the Student – A Handbook for the Expert*, de Gruyter Studies in Mathematics, Vol. 25, Walter de Gruyter, Berlin (1998).

[Hoffmann] Hoffmann, K., *Coidealunteralgebren in endlich dimensionalen Hopfalgebren*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1991).

[Hofstetter] Hofstetter, I., *Erweiterungen von Hopf-Algebren und ihre kohomologische Beschreibung*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1991).

[Hopf-1] Hopf, H., *Über die Topologie der Gruppen-Mannigfaltigkeiten und ihre Verallgemeinerungen*, Ann. of Math. (2) **42** (1941), 22–52 (also in [Hopf-2], pp. 119–151).

[Hopf-2] Hopf, H., *Selecta*, Eckmann, B., ed., Springer-Verlag, Berlin (1964), 119–151.

[Hu-1] Hu, S.-T., *Elements of Modern Algebra*, Holden-Day, San Francisco, CA (1965).

[Hu-2] Hu, S.-T., *Introduction to Homological Algebra*, Holden-Day, San Francisco, CA (1968).

[Hueb] Huebschmann, J., *Extensions of Lie–Rinehart algebras and the Chern–Weil construction*, Higher Homotopy Structures in Topology and Mathematical Physics, Proceedings of the International Conference Held in Honor of the 60th Birthday of Jim Stasheff at Vassar College, Poughkeepsie, NY, June 13–15, 1996, McCleary, J., ed., Contemporary Mathematics, Vol. 227, American Mathematical Society, Providence, RI (1999), 145–176.

[Hughes] Hughes, J., *A study of categories of algebras and coalgebras*, Ph.D. Thesis, Carnegie–Mellon University, Pittsburg, PA (2001).

[Hun] Hungerford, T.W., *Algebra*, Holt, Rinehart and Winston, Inc., New York (1974); Graduate Texts in Mathematics, Vol. 73, Springer-Verlag New York, Inc., New York, New York (Third Printing, 1984).

[Hurley] Hurley, S., *Tame and Galois Hopf objects with normal bases*, Ph.D. Thesis, State University of New York at Albany, Albany, NY (1984).

[Husemoller] Husemoller, D., *The structure of the Hopf algebra $H_*(BU)$ over a $Z_{(p)}$-algebra*, Amer. J. Math. **93** (1971), 329–349.

[Hus] Husemoller, D., *Homology of certain H-spaces as group ring objects*, Algebra, Topology, and Category Theory: A Collection of Papers in Honor of Samuel Eilenberg, Heller, A. and Tierney, M., eds, Academic Press, New York (1976), 77–94.

[J-1] Jacobson, N., *Lectures in Abstract Algebra, Vol. I – Basic Concepts*, D. Van Nostrand Company, Princeton, NJ (1951); Graduate Texts in Mathematics, Vol. 30, Springer-Verlag, New York (1984).

[J-2] Jacobson, N., *Lie Algebras*, Wiley-Interscience, New York (1962); Dover, New York (1979).

[J-3] Jacobson, N., *Basic Algebra I*, 2nd edn, W.H. Freeman, New York (1985).

[Jac] Jacobson, N., *Basic Algebra II*, 2nd edn, W.H. Freeman, New York (1989).

[J-4] Jacobson, N., *Lectures in Abstract Algebra, Volume II – Linear Algebra*, D. Van Nostrand, Princeton, NJ (1953); 2nd printing, Graduate Texts in Mathematics, Vol. 31, Springer-Verlag, New York (1984).

[Jam-Lie] James, G. and Liebeck, M., *Representations and Characters of Groups*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, UK (1993).

[Jensen] Jensen, A.-L., *Grothendieck rings and integral representation rings of Hopf-algebra orders*, Ph.D. Thesis, University of Illinois at Chicago, Chicago, IL (1981).

[Jia] Jia, H.-Q., *Classification of QC Hopf algebras*, Ph.D. Thesis, The Florida State University, Tallahassee, FL (2000).

[Jimbo] Jimbo, M., *Introduction to the Yang–Baxter equation*, Braid Group, Knot Theory and Statistical Mechanics, Yang, C.N. and Ge, M.L., eds, Advanced Series in Mathematical Physics, Vol. 9, World Sci. Publishing, Teaneck, NJ (1989), 111–134.

[Jon] Jonah, D.W., *Cohomology of Coalgebras*, Mem. Amer. Math. Soc., Vol. 82, Providence, RI (1968).

[J-R] Joni, S.A. and Rota, G.-C., *Coalgebras and bialgebras in combinatorics*, Umbral Calculus and Hopf Algebras (Proceedings of the Conference on Umbral Calculus and Hopf Algebras held at the University of Oklahoma, Norman, Oklahoma, May 15–20, 1978), Morris, R., ed., Contemporary Mathematics, Vol. 6, Amer. Math. Soc., Providence, RI (1982), 1–47.

[Josek] Josek, M., *Zerlegung induzierter Darstellungen für Hopf-Galois-Erweiterungen*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1992).

[Jos] Joseph, A., *Quantum Groups and Their Primitive Ideals*, Ergebnisse der Mathematik und ihrer Grenzgebiete, 3. Folge-Band 29, A Series of Modern Surveys in Mathematics, Springer-Verlag, Berlin (1995).

[J-S-1] Joyal, A. and Street, R., *The geometry of tensor calculus, I.*, Adv. in Math. **88** (1991), 55–112.

[J-S-2] Joyal, A. and Street, R., *An introduction to Tannaka duality and quantum groups*, Category Theory (Como, 1990), Lecture Notes in Mathematics, Vol. 1488, Springer-Verlag, Berlin (1991), 413–492.

[J-S-3] Joyal, A. and Street, R., *Braided tensor categories*, Adv. in Math. **102** (1993), 20–78.

[Kac-1] Kac, G.I., *Ring groups and the duality principle*, Trans. Moscow Math. Soc. **12** (1963), 291–339.

[Kac-2] Kac, G.I., *Ring groups and the duality principle. II*, Trans. Moscow Math. Soc. **13** (1965), 94–126.

[Kac-3] Kac, G.I., *Certain arithmetic properties of ring groups*, Funct. Anal. Appl. **6** (1972), 158–160.

[Kac-Pal-1] Kac, G.I. and Paljutkin, V.G., *An example of a ring group generated by Lie groups*, Ukrain. Math. J. **16** (1964), 99–105.

[Kac-Pal-2] Kac, G.I. and Paljutkin, V.G., *Finite ring groups*, Trans. Moscow Math. Soc. **15** (1966), 251–294.

[Kan] Kan, D.M., *Adjoint functors*, Trans. Amer. Math. Soc. **87** (1958), 294–329.

[Kane] Kane, R.M., *The Homology of Hopf Spaces*, North-Holland Mathematical Library, Vol. 40, North-Holland, Elsevier Science, Amsterdam (1988).

[Kap] Kaplansky, I., *Bialgebras*, University of Chicago, Department of Mathematics, Lecture Notes in Mathematics, Chicago, IL (1975).

[Kash-1] Kashina, Y., *Studies of finite-dimensional semisimple Hopf algebras*, Ph.D. Thesis, University of Southern California, Los Angeles, CA (August 1999).

[Kash-2] Kashina, Y., *On the order of the antipode of Hopf algebras in* ${}^H_H YD$, Comm. Algebra **27** (3) (1999), 1261–1273.

[Kash-3] Kashina, Y., *A generalized power map for Hopf algebras*, Hopf Algebras and Quantum Groups: Proceedings of the Brussels Conference (Brussels, 1998), Caenepeel, S. and Van Oystaeyen, F., eds, Lecture Notes in Pure and Appl. Math., Vol. 209, Marcel Dekker (2000), 159–175.

[Kash-4] Kashina, Y., *Classification of semisimple Hopf algebras of dimension 16*, J. Algebra **232** (2) (2000), 617–663.

[Kass] Kassel, C., *Quantum Groups*, Graduate Texts in Mathematics, Vol. 155, Springer-Verlag, New York (1995).

[K-R-T] Kassel, C., Rosso, M. and Turaev, V., *Quantum Groups and Knot Invariants*, Panoramas et Synthèses, Noméro 5, Société Mathématique de France, Maison de la S.M.F., Marseille Cedex 9, France (1997).

[Kauffman-1] Kauffman, L.H., *Knots and Physics*, Series on Knots and Everything, Vol. 1, 1st edn, World Scientific, Singapore (1991); 2nd edn (1993); 3rd edn (2001).

[Kauffman-2] Kauffman, L.H., *Gauss codes, quantum groups and ribbon Hopf algebras*, Rev. Math. Phys. **5** (4) (1993), 735–773.

[Kauf-Radf-1] Kauffman, L.H. and Radford, D.E., *A necessary and sufficient condition for a finite-dimensional Drinfel'd double to be a ribbon Hopf algebra*, J. Algebra **159** (1) (1993), 98–114.

[Kauf-Radf-2] Kauffman, L.H. and Radford, D.E., *Invariants of* 3*-manifolds derived from finite-dimensional Hopf algebras*, J. Knot Theory Ramifications **4** (1) (1995), 131–162.

[Kauf-Radf-3] Kauffman, L.H. and Radford, D.E., *Quantum algebra structures on $n \times n$ matrices*, J. Algebra **213** (2) (1999), 405–436.

[Kauf-Radf-4] Kauffman, L.H. and Radford, D.E., *A separation result for quantum coalgebras with an application to pointed quantum coalgebras of small dimension*, J. Algebra **225** (1) (2000), 162–200.

[Kauf-Radf-5] Kauffman, L.H. and Radford, D.E., *Quantum algebras, quantum coalgebras, invariants of* 1–1 *tangles and knots*, Comm. Algebra **28** (2000), 5101–5156.

[Kelley] Kelley, J.L., *General Topology*, D. Van Nostrand, Princeton, NJ (1955).

[Khal] Khalkhali, M., *An approach to operations on cyclic homology*, J. Pure Appl. Algebra **107** (1) (1996), 47–59.

[Kir] Kirillov, A.A., *Elements of the Theory of Representations*, Grundlehren der mathematischen Wissenschaften, Band 220, Springer-Verlag, Berlin (1976).

[Klein-1] Klein, F., *Elementarmathematik vom höheren Standpunkte aus I; Erster Band, Arithmetik, Algebra, Analysis*, Die Grundlehren der Mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete, Dritte Auflage, Band XIV, Verlag von Julius Springer, Berlin (1924).

[Klein-2] Klein, F., *Elementary Mathematics from an Advanced Standpoint, Vol. 1: Arithmetic, Algebra, Analysis*, Dover, New York (1945). Translated from the third German edition by E.R. Hedrick and C.A. Noble.

[Kleis] Kleisli, H., *Resolutions in Additive and Non-Additive Categories*, Queen's Papers in Pure and Applied Mathematics, No. 32, Queen's University, Kingston, ON (1973).

[Kli-Sch] Klimyk, A. and Schmüdgen, K., *Quantum Groups and Their Representations*, Texts and Monographs in Physics, Springer-Verlag, Berlin (1997).

[K-M-R-T] Knus, M.-A., Merkurjev, A., Rost, M. and Tignol, J.-P., *The Book of Involutions*, American Mathematical Society Colloquium Publications, Vol. 44, Amer. Math. Soc., Providence, RI (1998).

[Koc] Kock, A., *Synthetic Differential Geometry*, London Mathematical Society Lecture Note Series, Vol. 51, Cambridge University Press, Cambridge (1981).

[Kohl] Kohl, T., *Classification of abelian Hopf algebra forms acting on radical extensions*, Ph.D. Thesis, State University of New York at Albany, Albany, NY (1996).

[Kohno] Kohno, T., *Integrable systems related to braid groups and Yang–Baxter equation*, Braid Group, Knot Theory and Statistical Mechanics, Yang, C.N. and Ge, M.L., eds, Advanced Series in Mathematical Physics, Vol. 9, World Sci. Publishing, Teaneck, NJ (1989), 135–149.

[Kor-Soi] Korogodski, L.I. and Soibelman, Y.S., *Algebras of Functions on Quantum Groups: Part I*, Mathematical Surveys and Monographs, Vol. 56, Amer. Math. Soc., Providence, RI (1988).

[Kostant] Kostant, B., *Groups over $\mathbb{Z}$*, Algebraic Groups and Discontinuous Subgroups, Borel, A. and Mostow, G.D., eds, Proceedings of Symposia in Pure Mathematics, Vol. IX, Amer. Math. Soc., Providence, RI (1966), 90–98.

[Kos] Kostrikin, A.I., *Introduction to Algebra*, Universitext, Springer-Verlag, New York (1982). Translated from the Russian by Neal Koblitz.

[Köt] Köthe, G., *Topological Vector Spaces I*, Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Band 159, Springer-Verlag, New York (1969).

[Krei-Del] Kreimer, D. and Delbourgo, R., *Using the Hopf algebra structure of AFT in calculations*, Phys. Rev. D **60** (3) (1999), 627–670.

[L] Lam, T.Y., *A First Course in Noncommutative Rings*, Graduate Texts in Mathematics, Vol. 131, Springer-Verlag, New York (1991).

[L-R] Lambe, L.A. and Radford, D.E., *Introduction to the Quantum Yang–Baxter Equation and Quantum Groups: An Algebraic Approach*, Kluwer Academic, Dordrecht (1997).

[La] Lambek, J., *Lectures on Rings and Modules*, Blaisdell-Ginn, Waltham, MA (1966); Chelsea, New York (1976).

[Larson-1] Larson, R.G., *Hopf algebras and group algebras*, Ph.D. Thesis, The University of Chicago, Chicago, IL (1965).

[Larson-2] Larson, R.G., *Group rings over Dedekind domains*, J. Algebra **5** (3) (1967), 358–361.

[Larson-3] Larson, R.G., *Group rings over Dedekind domains. II*, J. Algebra **7** (2) (1967), 278–279.

[Lar-1] Larson, R.G., *Cocommutative Hopf algebras*, Canad. J. Math. **19** (2) (1967), 350–360.

[Lar-2] Larson, R.G., *The order of the antipode of a Hopf algebra*, Proc. Amer. Math. Soc. **21** (1) (1969), 167–170.

[Lar-3] Larson, R.G., *Characters of Hopf algebras*, J. Algebra **17** (2) (1971), 352–368.

[Lar-4] Larson, R.G., *Orders in Hopf algebras*, J. Algebra **22** (2) (1972), 201–210.

[Lar-Rad-1] Larson, R.G. and Radford, D.E., *Semisimple cosemisimple Hopf algebras*, Amer. J. Math. **110** (1988), 187–195.

[Lar-Rad-2] Larson, R.G. and Radford, D.E., *Finite dimensional cosemisimple Hopf algebras in characteristic* 0 *are semisimple*, J. Algebra **117** (1988), 267–289.

[Lar-Rad-3] Larson, R.G. and Radford, D.E., *Semisimple Hopf algebras*, J. Algebra **171** (1995), 5–35.

[Lar-Sw] Larson, R.G. and Sweedler, M.E., *An associative orthogonal bilinear form for Hopf algebras*, Amer. J. Math. **XCI** (1) (1969), 75–94.

[Lar-Ta] Larson, R.G. and Taft, E.J., *The algebraic structure of linearly recursive sequences under Hadamard product*, Israel J. Math. **72** (1990), 118–132.

[Lar-Tow] Larson, R.G. and Towber, J., *Two dual classes of bialgebras related to the concepts of "quantum group" and "quantum Lie algebra"*, Comm. Algebra **19** (1991), 3295–3345.

[Law] Lawvere, F.W., *Functorial semantics of algebraic theories*, Proc. Nat. Acad. Sci. USA **50** (1963), 869–872.

[Lawvere] Lawvere, F.W., *Algebraic theories, algebraic categories, and algebraic functors*, The Theory of Models: Proceedings of the 1963 International Symposium at Berkeley, Addison, J.W., Henkin, L. and Tarski, A., eds, North-Holland, Amsterdam (1965), 413–418.

[Law-Sch] Lawvere, F.W. and Schanuel, S.H., *Conceptual Mathematics: A First Introduction to Categories*, Cambridge University Press, Cambridge, UK (1997).

[Laz] Lazard, M., *Groupes Analytiques p-adiques*, Publications Mathématiques de ℓ'I.H.E.S. (Institut des Hautes Études Scientifiques), N°26, Le Bois-Marie-Bures-sur-Yvette (1965).

[Lee] Lee, D.H., *The Structure of Complex Lie Groups*, Chapman & Hall/CRC Research Notes in Mathematics, Vol. 429, Chapman & Hall/CRC, Boca Raton, FL (2002).

[Lef] Lefschetz, S., *Algebraic Topology*, American Mathematical Society, Colloquium Publications, Vol. XXVII, Amer. Math. Soc., Providence, RI (1942).

[Lengyel] Lengyel, F., *Recursion categories of coalgebras*, Ph.D. Thesis, City University of New York, New York, NY (2002).

[Lev] Levin, G., *Homology of local rings*, Ph.D. Thesis, The University of Chicago, Chicago, IL (1965).

[Lin] Lin, J.P., *H-spaces with finiteness conditions*, Handbook of Algebraic Topology, James, I.M., ed., Elsevier, Amsterdam (1995), 1095–1141 (chapter 22).

[Lin-B] Lin, B.I.-P., *On homological properties of coalgebras*, Ph.D. Thesis, University of Illinois at Chicago, Chicago, IL (1981).

[Linchenko] Linchenko, V., *Some properties of Hopf algebras and H-module algebras*, Ph.D. Thesis, University of Southern California, Los Angeles, CA (2001).

[Lod] Loday, J.-L., *Cyclic Homology*, Grundlehren der mathematischen Wissenschaften, Vol. 301, Springer-Verlag, Berlin (1992).

[Lot] Lothaire, M., *Combinatorics on Words*, Encyclopedia of Mathematics and its Applications, Vol. 17, Rota, G.-C., ed., Cohn, P.M. and Lyndon, R., Algebra Section eds, Addison-Wesley/ World Science Division, Reading, MA (1983).

[Lu-Yan-Zhu] Lu, J.-H., Yan, M. and Zhu, Y., *On Hopf algebras with positive bases*, J. Algebra **237** (2) (2001), 421–445.

[Mac D] MacDuffee, C.C., *The Theory of Matrices*, Ergebnisse der Mathematik und Ihrer Grenzgebiete, Verlag von Julius Springer, Berlin (1933); corrected reprint, Chelsea, New York (1956).

[Mac-1] Mac Lane, S., *Homology*, Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Band 114, Springer-Verlag, New York (1963); 3rd corr. printing (1975), and Classics in Mathematics (reprint of the 1975 edition, 4th printing), Springer-Verlag, Berlin (1995).

[Mac-2] Mac Lane, S., *Some additional advances in algebra*, Studies in Modern Algebra, Albert, A.A., ed., Studies in Mathematics, Vol. 2, Curtis, C.W., ed., Mathematical Association of America (1963), 35–58.

[Mac-3] Mac Lane, S., *Categorical algebra*, Bull. Amer. Math. Soc. **71** (1965), 40–106.

[Mac-4] Mac Lane, S., *Categories for the Working Mathematician*, Graduate Texts in Mathematics, Vol. 5, Springer-Verlag, New York (1971).

[Mac-5] Mac Lane, S., *Mathematics: Form and Function*, Springer-Verlag, New York (1986).

[Mac-6] Mac Lane, S., *Concepts and categories in perspective*, A Century of Mathematics in America (Part I), P. Duren, ed., American Mathematical Society History of Mathematics, Vol. 1, Amer. Math. Soc., Providence, RI (1988), 323–366.

[Mac-Bir] Mac Lane, S. and Birkhoff, G., *Algebra*, 3rd edn, Chelsea, New York (1988).

[Mac-Moe] Mac Lane, S. and Moerdijk, I., *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*, Universitext, Springer-Verlag New York (1992).

[Mag] Magnus, W., *On the exponential solution of differential equations for a linear operator*, Comm. Pure and Appl. Math. **7** (1954), 649–673.

[Mag-Kar-Sol] Magnus, W., Karrass, A. and Solitar, D., *Combinatorial Group Theory*, Wiley-Interscience, New York (1966).

[Majid-1] Majid, S., *Non-commutative-geometric groups by a bicrossproduct construction: Hopf algebras at the Planck scale*, Ph.D. Thesis, Harvard University, Cambridge, MA (1988).

[Majid-2] Majid, S., *Hopf algebras for physics at the Planck scale*, Classical Quantum Gravity **5** (12) (1988), 1587–1606.

[Majid-3] Majid, S., *Physics for algebraists: noncommutative and noncocommutative Hopf algebras by a bicrossproduct construction*, J. Algebra **130** (1) (1990), 17–64.

[Maj-1] Majid, S., *Foundations of Quantum Group Theory*, Cambridge University Press, Cambridge, UK (1995).

[Maj-2] Majid, S., *A Quantum Groups Primer*, London Mathematical Society Lecture Notes Series, Vol. 292, Cambridge University Press, Cambridge, UK (2002).

[Manes] Manes, E.G., *Algebraic Theories*, Graduate Texts in Mathematics, Vol. 26, Springer-Verlag, New York (1976).

[Man-1] Manin, Y.I., *Quantum Groups and Non-Commutative Geometry*, Centre de Recherches Mathématiques (CRM), Université de Montréal, Montréal, Canada (1988).

[Man-2] Manin, Y.I., *Topics in Noncommutative Geometry*, M.B. Porter Lectures, Rice University, Princeton University Press, Princeton, NJ (1991).

[Marc] Marcus, M., *Introduction to Modern Algebra*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. 47, Marcel Dekker, New York (1978).

[Marg] Margolis, H.R., *Spectra and the Steenrod Algebra*, North-Holland Mathematical Library, Vol. 29, North-Holland, Elsevier Science, Amsterdam (1983).

[Marlo] Marlowe, T.J., *Pointed coalgebras and incidence coalgebras*, Ph.D. Thesis, Rutgers University, New Brunswick, NJ (1975).

[Mastnak] Mastnak, M., *On Hopf algebra extensions and cohomologies*, Ph.D. Thesis, Dalhousie University, Halifax, Nova Scotia, Canada (2002).

[Masuoka-1] Masuoka, A., *Self-dual Hopf algebras of dimension* p^3, J. Algebra **178** (1995), 791–806.

[Masuoka-2] Masuoka, A., *Extensions of Hopf algebras* (Lecture notes taken by Matías Graña), Trabajos de Matemática [Mathematical Works], Notas Mat. No. 41/99, Universidad Nacional de Córdoba, Facultad de Matemática, Astronomía y Física, Córdoba (1999).

[Masuoka-3] Masuoka, A., *Extensions of Hopf algebras and Lie bialgebras*, Trans. Amer. Math. Soc. **352** (2000), 3837–3879.

[Masuoka-4] Masuoka, A., *Hopf algebra extensions and cohomology*, New Directions in Hopf Algebras, Montgomery, S. and Schneider, H.-J., eds, Mathematical Sciences Research Institute Publication, Vol. 43, Cambridge University Press, Cambridge, UK (2002), 167–209.

[McR] McRae, D.G., *Coherent rings and homology*, Ph.D. Thesis, University of Washington, Seattle, WA (1967).

[Me] Meyer, P.-A., *Quantum Probability for Probabilists*, Lecture Notes in Mathematics, Vol. 1538, Springer-Verlag, Berlin (1993).

[Mi-1] Michaelis, W.J., *Lie coalgebras*, Ph.D. Thesis, University of Washington, Seattle, WA (1974).

[Mi-2] Michaelis, W., *Lie coalgebras*, Adv. in Math. **38** (1) (1980), 1–54.

[Mi-3] Michaelis, W., *The dual Poincaré–Birkhoff–Witt theorem*, Adv. in Math. **57** (2) (1985), 93–162.

[Mi-4] Michaelis, W., *Properness of Lie algebras and enveloping algebras, II*, in: Lie Algebras and Related Topics (Proceedings of a Summer Seminar held June 26–July 6, 1984), Britten, D.J., Lemire, F.W., and Moody, R.V., eds, Canad. Math. Soc. Conf. Proc., Vol. 5, Amer. Math. Soc., Providence, R.I. (1986), 265–280.

[Mi-5] Michaelis, W., *Properness of Lie algebras and enveloping algebras, I*, Proc. Amer. Math. Soc. **101** (1987), 17–23.

[Mi-6] Michaelis, W., *On the dual Lie coalgebra of the Witt algebra*, XVII International Colloquium on Group Theoretical Methods in Physics, June 27–July 2, 1988, Sainte-Adèle, PQ, Canada, Saint-Aubin, Y. and Vinet, L., eds, Word Scientific, Teaneck, NJ (1989), 435–439.

[Mi-7] Michaelis, W., *An example of a non-zero Lie coalgebra M for which* $\mathrm{Loc}(M) = 0$, J. Pure Appl. Algebra **68** (1990), 341–348.

[Mi-8] Michaelis, W., *The primitives of the continuous linear dual of a Hopf algebra as the dual Lie algebra of a Lie coalgebra*, Lie Algebras and Related Topics, Proceedings of a Research Conference held May 22–June 2, 1988, University of Wisconsin, Madison, WI, Benkart, G. and Osborn, J.M., eds, Contemporary Mathematics, Vol. 110, American Mathematical Society, Providence, RI (1990), 125–176.

[Mi-9] Michaelis, W., *A class of infinite-dimensional Lie bialgebras containing the Virasoro algebra*, Adv. in Math. **107** (2) (1994), 365–392.

[Mi-10] Michaelis, W., *The dual Lie bialgebra of a Lie bialgebra*, Modular Interfaces: Modular Lie Algebras, Quantum Groups, and Lie Superalgebras, Chari, V. and Penkov, I.B., eds, AMS/IP Studies in Advanced Mathematics, Vol. 4N, American Mathematical Society, Providence, RI/International Press, Cambridge, MA (1997), 81–93.

[Mi-11] Michaelis, W., *Lie coalgebras*, Handbook of Algebra, Hazewinkel, M., ed., Vol. 4 (or Vol. 5), North-Holland, Elsevier Science, Amsterdam, to appear.

[Miller] Miller, H., *The Sullivan conjecture on maps from classifying spaces*, Ann. of Math. (2) **120** (1984), 39–87.

[Milinski] Milinski, A., *Operationen punktierter Hopfalgebren auf primen Algebren*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1995).

[Mil] Milnor, J.W., *The Steenrod algebra and its dual*, Ann. of Math. (2) **67** (1958), 150–171.

[M-M-1] Milnor, J.W. and Moore, J.C., *On the structure of Hopf algebras*, Report of the Air Force Office of Scientific Research Preprint for [M-M-2].

[M-M-2] Milnor, J.W. and Moore, J.C., *On the structure of Hopf algebras*, Ann. of Math. (2) **81** (1965), 211–264.

[Molnar-1] Molnar, R.K., *Smash coproducts and decompositions of Hopf algebras*, Ph.D. Thesis, The University of North Carolina at Chapel Hill, Chapel Hill, NC (1973).

[Molnar-2] Molnar, R.K., *Semi-direct products of Hopf algebras*, J. Algebra **47** (1977), 29–51.

[Molnar-3] Molnar, R.K., *Coradical splittings and a generalization of Kostant's theorem*, J. Pure Appl. Algebra **68** (1990), 349–357.

[Mon-1] Montgomery, S., *Hopf Algebras and Their Actions on Rings*, Conference Board of the Mathematical Sciences (CBMS) Regional Conference Series in Mathematics, No. 82, American Mathematical Society, Providence, RI (1993).

[Mon-2] Montgomery, S., *Classifying finite-dimensional semisimple Hopf algebras*, Trends in the Representation Theory of Finite Dimensional Algebras, Proceedings of an AMS–IMS–SIAM Joint Summer Research Conference held July 20–24, 1997, University of Washington, Seattle, WA, Green, E.L. and Huisgen-Zimmermann, B., eds, Contemporary Mathematics, Vol. 229, American Mathematical Society, Providence, RI (1998), 265–279.

[Montgomery] Montgomery, S., *Indecomposable coalgebras, simple comodules, and pointed Hopf algebras*, Proc. Amer. Math. Soc. **123** (1995), 2343–2351.

[Mon-Sch] Montgomery, S. and Schneider, H.-J., eds, *New Directions in Hopf Algebras*, Mathematical Sciences Research Institute Publications, Vol. 43, Cambridge University Press, Cambridge, UK (2002).

[M-P] Moody, R.V. and Pianzola, A., *Lie Algebras with Triangular Decompositions*, Canadian Mathematical Society Series of Monographs and Advanced Texts, Wiley-Interscience, New York (1995).

[Moo-1] Moore, J.C., *Complements sur les algèbres de Hopf*, Séminaire H. Cartan–J.C. Moore, 12e année, 1959/60, Exposé n°4 (30 novembre 1959), 4-01–4-12.

[Moo-2] Moore, J.C., *Algèbres de Hopf universelles*, Séminaire H. Cartan–J.C. Moore, 12e année, 1959/60, Exposé n°10 (18 janvier 1960), 10-01–10-11.

[Moo-3] Moore, J.C., *What it means for a coalgebra to be simply connected*, Algebra, Topology, and Category Theory: A Collections of Papers in Honor of Samuel Eilenberg, Heller, A. and Tierney, M. eds, Academic Press, New York (1974), 145–148.

[Moo-Smi-1] Moore, J.C. and Smith, L., *Hopf algebras and multiplicative fibrations, I*, Amer. J. Math. **90** (1968), 752–780.

[Moo-Smi-2] Moore, J.C. and Smith, L., *Hopf algebras and multiplicative fibrations, II*, Amer. J. Math. **90** (1968), 1113–1150.

[M-T] Mosher, R.E. and Tangora, M.C., *Cohomology Operations and Applications in Homotopy Theory*, Harper's Series in Modern Mathematics, Harper & Row, New York (1968).

[Müller-1] Müller, E., *Quantengruppen im Einheitswurzelfall*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1998).

[Müller-2] Müller, E., *Cosemisimple Hopf algebras and the FRT construction*, Comm. Algebra **29** (10) (2001), 4377–4394.

[Mum] Mumford, D., *Geometric Invariant Theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 34, Springer-Verlag, Berlin (1965).

[Mumford] Mumford, D., *Abelian Varieties*, Tata Institute of Fundamental Research Studies in Mathematis, No. 5, 2nd edn, Oxford University Press, London (1975).

[M-F-K] Mumford, D., Fogarty, J. and Kirwan, F., *Geometric Invariant Theory*, 3rd enlarged edn, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 34, Springer-Verlag, Berlin (1994).

[Mun] Munkres, J.R., *Elements of Algebraic Topology*, Addison-Wesley, Menlo Park, CA (1984).

[Mus] Musson, I.M., *Finite quantum groups and pointed Hopf algebras*, Quantum Groups and Lie Theory, Pressley, A., ed., London Mathematical Society Lecture Notes Series, Vol. 290, Cambridge University Press, Cambridge, UK (2001), 191–205.

[Nagata] Nagata, M., *Complete reducibility of rational representations of a matric group*, J. Math. Kyoto Univ. **1** (1961), 87–99.

[Na-1] Natale, S., *Álgebras de Hopf semisimples*, Ph.D. Thesis, Universidad Nacional de Córdoba, Córdoba, Argentina (1999).

[Na-2] Natale, S., *On semisimple Hopf algebras of dimension pq^2*, J. Algebra **221** (1999), 242–278.

[Neisen] Neisendorfer, J., *Lie algebras, coalgebras and rational homotopy theory for nilpotent spaces*, Pacific J. Math. **74** (1978), 429–460.

[Netzsch] Netzsch, R., *Bialgebren in Endomorphismenringen*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1979).

[Neuchl] Neuchl, M., *Representation theory of Hopf-categories*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1997).

[New-1] Newman, K.W., *Topics in the theory of irreducible Hopf algebras*, Ph.D. Thesis, Cornell University, Ithaca, NY (1970).

[New-2] Newman, K., *Sequences of divided powers in irreducible, cocommutative Hopf algebras*, Trans. Amer. Math. Soc. **163** (1972), 25–34.

[New-3] Newman, K., *The structure of free irreducible, cocommutative Hopf algebras*, J. Algebra **29** (1974), 1–26.

[New-4] Newman, K., *A correspondence between bi-ideals and sub-Hopf algebras in cocommutative Hopf algebras*, J. Algebra **36** (1975), 1–15.

[New-5] Newman, K., *The coalgebra structure of Hopf algebras*, J. Algebra **50** (1978), 245–264.

[New-Rad] Newman, K. and Radford, D.E., *The cofree irreducible Hopf algebra on an algebra*, Amer. J. Math. **101** (1979), 1024–1045.

[N-1] Nichols, W.D., *Bialgebras*, Ph.D. Thesis, The University of Chicago, Chicago, IL (1975).

[N-2] Nichols, W.D., *Pointed irreducible bialgebras*, J. Algebra **57** (1) (1979), 64–76.

[N-3] Nichols, W.D., *The structure of the dual Lie coalgebra of the Witt algebra*, J. Pure Appl. Algebra **68** (1990), 359–364.

[N-4] Nichols, W.D., *Cosemisimple Hopf algebras*, Advances in Hopf Algebras, Bergen, J. and Montgomery, S., eds, Lecture Notes in Pure and Applied Mathematics, Vol. 158, Marcel Dekker, New York (1994), 135–151.

[N-R-1] Nichols, W.D. and Richmond (Zoeller), M.B., *Freeness of infinite dimensional Hopf algebras*, Comm. Algebra **20** (1992), 1489–1492.

[N-R-2] Nichols, W.D. and Richmond, M.B., *The Grothendieck group of a Hopf algebra*, J. Pure Appl. Algebra **106** (1996), 297–306.

[N-R-3] Nichols, W.D. and Richmond, M.B., *The Grothendieck algebra of a Hopf algebra. I*, Comm. Algebra **26** (1998), 1081–1095.

[N-S] Nichols, W. and Sweedler, M., *Hopf algebras and combinatorics*, Umbral Calculus and Hopf Algebras (Proceedings of the Conference on Umbral Calculus and Hopf Algebras held at the University of Oklahoma, Norman, Oklahoma, May 15–20, 1978), Morris, R., ed., Contemporary Mathematics, Vol. 6, American Mathematical Society, Providence, RI (1982), 49–84.

[N-Z-1] Nichols, W.D. and Zoeller, M.B., *A Hopf algebra freeness theorem*, Amer. J. Math. **111** (1989), 381–385.

[N-Z-2] Nichols, W.D. and Zoeller, M.B., *Freeness of infinite dimensional Hopf algebras over grouplike subalgebras*, Comm. Algebra **17** (1989), 413–424.

[N-Z-3] Nichols, W.D. and Zoeller, M.B., *Finite-dimensional Hopf algebras are free over grouplike subalgebras*, J. Pure Appl. Algebra **56** (1989), 51–57.

[Ni] Nicholson, W.K., *Introduction to Abstract Algebra*, The Prindle, Weber and Schmidt Series in Mathematics, PWS-Kent Publishing Company (a division of Wadsworth, Inc.), Boston (1993).

[No] Noether, E., *Hyperkomplexe Grössen und Darstellungstheorie*, Math. Z. **30** (1929), 641–692.

[Nor] Northcott, D.G., *Multilinear Algebra*, Cambridge University Press, Cambridge, UK (1984).

[Oud] Oudom, J.-M., *Coproducts and cogroups in the category of graded dual Leibniz algebras*, Operads: Proceedings of Renaissance Conferences (Special Session and International Conference on Moduli Spaces, Operads, and Representation Theory/Operads and Homotopy Algebra, March 1995/May–June 1995, Hartford, Connecticut/Luminy, France), Loday, J.-L., Stasheff, J.D. and Voronov, A.A., eds, Contemporary Mathematics, Vol. 202, American Mathematical Society, Providence, RI (1997), 115–135.

[Ouyang] Ouyang, M., *Actions of Hopf algebras*, Ph.D. Thesis, University of Southern California, Los Angeles, CA (1997).

[Palm] Palmer, T.W., *Banach Algebras and the General Theory of $*$-Algebras*, Encyclopedia of Mathematics and its Applications, Rota, G.-C., ed., Vol. 49, Cambridge University Press, Cambridge, UK (1994).

[Pareig] Pareigis, B., *When Hopf algebras are Frobenius algebras*, J. Algebra **18** (1971), 588–596.

[Par-1] Pareigis, B., *A non-commutative non-cocommutative Hopf algebra in "nature"*, J. Algebra **70** (1981), 356–374.

[Par-2] Pareigis, B., *Fourier transforms over finite quantum groups*, Seminarberichte BF Mathematik Fern Universitätt Hagen, Bd. 63 (1998), 561–570(3).

[Parker] Parker, D.B., *Hopf Galois extensions and forms of coalgebras and Hopf algebras*, Ph.D. Thesis, University of California, Berkeley, CA (1975).

[Pas] Passman, D.S., *The Algebraic Structure of Group Rings*, Wiley-Interscience, New York (1977).

[Patras-1] Patras, F., *A Leray theorem for the generalization to operads of Hopf algebras with divided powers*, J. Algebra **218** (1999), 528–542.

[Patras-2] Patras, F., *Generic algebras and iterated Hochschild homology*, J. Pure Appl. Algebra **162** (2001), 337–357.

[P-T] Peterson, B. and Taft, E.J., *The Hopf algebra of linearly recursive sequences*, Aequationes Math. **20** (1980), 1–17.

[Phung] Phung, H.H., *Über Quantengruppen und koquasitrianguläre Hopf-Algebren in verzopften Kategorien*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1996).

[Pi] Pierce, R.S., *Associative Algebras*, Graduate Texts in Mathematics, Vol. 88, Springer-Verlag, New York (1982).

[Pop-1] Pop, H.C., *Quantum group constructions in a symmetric monoidal category*, Ph.D. Thesis, University of Southern California, Los Angeles, CA (1995).

[Pop-2] Pop, H.C., *Sur les matrices quantiques*, C.R. Acad. Sci. Paris Ser. I Math. **324** (7) (1997), 753–757.

[Q] Quillen, D., *Rational homotopy theory*, Ann. of Math. (2) **90** (1969), 205–295.

[Quinn] Quinn, F., *Lectures on axiomatic topological quantum field theory*, Geometry and Quantum Field Theory, IAS/Park City Mathematics Series, Vol. 1, Amer. Math. Soc., Providence, RI (1995), 323–453.

[R-1] Radford, D.E., *Rationality and the theory of coalgebras*, Ph.D. Thesis, The University of North Carolina, Chapel Hill, NC (1970).

[R-2] Radford, D.E., *A free rank* 4 *Hopf algebra with antipode of order* 4, Proc. Amer. Math. Soc. **30** (1) (1971), 55–58.

[R-3] Radford, D.E., *Coreflexive coalgebras*, J. Algebra **26** (1973), 512–535.

[R-4] Radford, D.E., *The order of the antipode of a finite dimensional Hopf algebra is finite*, Amer. J. Math. **98** (1976), 333–355.

[R-5] Radford, D.E., *Operators on Hopf algebras*, Amer. J. Math. **99** (1977), 139–158.

[R-6] Radford, D.E. *On the structure of pointed coalgebras*, J. Algebra **77** (1) (1982), 1–14.

[Rad-1] Radford, D.E., *A natural ring basis for the shuffle algebra and an application to group schemes*, J. Algebra **58** (1979), 432–454.

[Rad-2] Radford, D.E., *On quantum algebras and coalgebras, oriented quantum algebras and coalgebras, invariants of* 1–1 *tangles, knots and links*, New Directions in Hopf Algebras, Montgomery, S. and Schneider, H.-J., eds, Mathematical Sciences Research Institute Publication, Vol. 43, Cambridge University Press, Cambridge, UK (2002), 263–319.

[R-T-W] Radford, D.E., Taft, E.J. and Wilson, R.L., *Forms of certain Hopf algebras*, Manuscripta Math. **17** (1975), 333–338.

[Ragozzine] Ragozzine, Jr., C.B., *On Hopf algebras generated by coalgebras*, Ph.D. Thesis, The Ohio State University, Columbus, OH (1999).

[Ra] Ravenel, D.C., *Complex Cobordism and Stable Homotopy Groups of Spheres*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. 121, Academic Press, Orlando, FL (1986).

[Ree] Ree, R., *Lie elements and an algebra associated with shuffles*, Ann. of Math. **68** (1958), 210–220.

[Reshetikhin] Reshetikhin, N., *Quantization of Lie bialgebras*, Internat. Math. Res. Notices **7** (1992), 143–151.

[Ri] Ribenboim, P., *Rings and Modules*, Interscience Tracts in Pure and Applied Mathematics, Vol. 24, Wiley-Interscience, New York (1969).

[Roby] Roby, N., *Les algèbres à puissances divisées*, Bull. Sci. Math. (2) **89** (1965), 75–91.

[R-R] Roman, S.M. and Rota, G.-C., *The umbral calculus*, Adv. in Math. **27** (2) (1978), 95–188.

[Ross-1] Ross, L.E., *On representations and cohomology of graded Lie algebras*, Ph.D. Thesis, University of California, Berkeley, CA (1964).

[Ross-2] Ross, L.E., *Representations of graded Lie algebras*, Trans. Amer. Math. Soc. **120** (1965), 17–23.

[Rosso-1] Rosso, M., *Groupes quantiques, représentations linéaires et applications*, Ph.D. Thesis, Université Denis Diderot – Paris VII, Paris (1990).

[Rosso-2] Rosso, M., *Quantum groups and quantum shuffles*, Invent. Math. **133** (2) (1998), 399–416.

[Rota] Rota, G.-C., with the collaboration of Doubilet, P., Greene, C., Kahaner, D., Odlyzko, A. and Stanley, R., *Finite Operator Calculus*, Academic Press, New York (1975).

[R-K-O] Rota, G.-C., Kahaner, D. and Odlyzko, A., *Finite Operator Calculus*, Finite Operator Calculus, Rota, G.-C., ed., Academic Press, New York (1975), 7–82.

[Rotm] Rotman, J.J., *An Introduction to Homological Algebra*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. 85, Academic Press, Inc., New York, NY (1979).

[Row-1] Rowen, L.H., *Ring Theory, Vol. I*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. 127, Academic Press, San Diego, CA (1988).

[Row-2] Rowen, L.H., *Ring Theory, Vol. II*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. 128, Academic Press, San Diego, CA (1988).

[S-R] Saavedra-Rivano, N., *Catégories Tannakiennes*, Lecture Notes in Mathematics, Vol. 265, Springer-Verlag, Berlin (1972).

[Sa] Samelson, H., *Topology of Lie groups*, Bull. Amer. Math. Soc. **58** (1952), 2–37.

[Scharfschwerdt] Scharfschwerdt, B., *Endlichdimensionale Yetter–Drinfeld Hopfalgebren*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (2000).

[Schauenburg-1] Schauenburg, P., *Zur nichtkommutativen Differentialgeometrie von Hauptfaserbündeln-Hopf-Galois-Erweiterungen von De Rham-Komplexen*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1993).

[Schauenburg-2] Schauenburg, P., *Hopf algebra extensions and monoidal categories*, New Directions in Hopf Algebras, Montgomery, S. and Schneider, H.-J., eds, Mathematical Sciences Research Institute Publication, Vol. 43, Cambridge University Press, Cambridge, UK (2002), 321–381.

[Schik] Schikhof, W.H., *Non-archimedean representations of compact groups*, Compositio Math. **23** (1971), 215–232.

[Schmitt] Schmitt, W.R., *Antipodes and incidence coalgebras*, Ph.D. Thesis, MIT, Cambridge, MA (1986).

[Schmi] Schmitt, W., *Antipodes and incidence coalgebras*, J. Combin. Theory A **46** (1987), 264–290.

[Schmü] Schmüdgen, K., *Eine Einführung in die Theorie der Quantenmatrixgruppen*, J. Lie Theory **1** (1991), 3–21 (formerly Seminar Sophus Lie Darmstadt Erlangen Greifswald Leipzig).

[Sch-1] Schneider, H.-J., *Bemerkung zu einer Arbeit von Tate–Oort*, Manuscripta Math. **8** (1973), 319–322.

[Sch-2] Schneider, H.-J., *Lectures on Hopf algebras* (Notes by Sonia Natale), Trabajos de Matemática [Mathematical Works], Vol. 31/95, Universidad Nacional de Córdoba, Facultad de Matemática, Astronomía y Física, Córdoba (1995).

[Sch-3] Schneider, H.-J., *Representation theory of Hopf Galois extensions*, Israel J. Math **72** (1990), 196–231.

[S-S] Schnider, S. and Sternberg, S., *Quantum Groups: From Coalgebras to Drinfeld Algebras, a Guided Tour*, International Press, Cambridge, MA (1993).

[Schub] Schubert, H., *Categories*, Springer-Verlag, Berlin (1972); Revised and enlarged translation (by E. Gray) of "Kategorien I" and "Kategorien II", Heidelberger Taschenbücher, Vols. 65, 66, Springer-Verlag, Berlin (1970).

[Schür] Schürmann, M., *White Noise on Bialgebras*, Lectures Notes in Mathematics, Vol. 1544, Springer-Verlag, Berlin (1993).

[Selick-Wu] Selick, P. and Wu, J., *On Natural Coalgebra Decompositions of Tensor Algebras and Loop Suspensions*, Mem. Amer. Math. Soc., Vol. 148, no. 701, American Mathematical Society, Providence, RI (2000).

[Sel] Seligman, G.B., *Modular Lie Algebras*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Vol. 40, Springer-Verlag, New York (1967).

[Se-1] Serre, J.-P., *Lie Algebras and Lie Groups*, W.A. Benjamin, New York (1965).

[Se-2] Serre, J.-P., *Lie Algebras and Lie Groups*, Lecture Notes in Mathematics, Vol. 1500, Springer-Verlag, Berlin (1992).

[Se-3] Serre, J.-P., *Gèbres*, Enseig. Math. **39** (1–2) (1993), 33–85.

[Se-4] Serre, J.-P., *Gèbres*, Item 160 in J.-P. Serre, Œuvres, Collected Papers, Vol. IV: 1985–1998, Springer-Verlag, Berlin (2000), 272–324.

[Simmons] Simmons, G.F., *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York (1964).

[Sing] Singer, W.M., *Extension theory for connected Hopf algebras*, J. Algebra **21** (1972), 1–16.

[Smir] Smirnov, V.A., *Simplicial and Operad Methods in Algebraic Topology*, Translations of Mathematical Monographs, Vol. 198, American Mathematical Society, Providence, RI (2001).

[Smit] Smith, S.P., *Quantum groups: an introduction and survey for ring theorists*, Noncommutative Rings, Montgomery, S. and Small, L., eds, Mathematical Sciences Research Institute Publications, Vol. 24, Springer-Verlag, New York (1992), 131–178.

[Smith] Smith III, H.H., *Constructing Hopf orders in elementary Abelian group rings*, Ph.D. Thesis, State University of New York at Albany, Albany, NY (1997).

[Som] Sommerhäuser, Y., *On Kaplansky's fifth conjecture*, J. Algebra **204** (1) (1998), 202–224.

[Sommer-1] Sommerhäuser, Y., *Yetter–Drinfeld Hopf algebras over groups of prime order*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1999).

[Sommer-2] Sommerhäuser, Y., *On Kaplansky's conjectures*, Interactions Between Ring Theory and Representations of Algebra, Van Oystaeyen, F. and Saorin, M., eds, Lecture Notes in Pure and Applied Mathematics, Vol. 210, Marcel Dekker, New York (2000), 393–412.

[Sommer-3] Sommerhäuser, Y., *Yetter–Drinfel'd Hopf Algebras over Groups of Prime Order*, Lecture Notes in Mathematics, Vol. 1789, Springer-Verlag, Berlin (2002).

[Smo] Smoke, W., *Invariant differential operators*, Trans. Amer. Math. Soc. **127** (1967), 460–494.

[S-O] Spiegel, E. and O'Donnell, C.J., *Incidence Algebras*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. 206, Marcel Dekker, New York (1997).

[Spin] Spindler, K.H., *Abstract Algebra with Applications, Vol. I, Vector Spaces and Groups*, Marcel Dekker, New York (1994).

[Sta-1] Stasheff, J., *H-Spaces from a Homotopy Point of View*, Lecture Notes in Mathematics, Vol. 161, Springer-Verlag, Berlin (1970).

[Sta-2] Stasheff, J., *Milnor's work from the perspective of algebraic topology*, Topological Methods in Modern Mathematics: Proceedings of a Symposium in Honor of John Milnor's Sixtieth Birthday Held at the State University of New York at Stony Brook, June 14–June 21, 1991, Goldberg, L.R. and Phillips, A.V., eds, Publish or Perish, Inc., Houston, TX (1993), 5–22.

[Stef] Ştefan, D., *Hopf algebras of low dimension*, J. Algebra **211** (1999), 343–361.

[Stein] Stein, J.A., *Hopf algebra structures on tensor products of Z-modules*, Ph.D. Thesis, Harvard University, Cambridge, MA (1980).

[Ste] Sternberg, S., *Notes from the Seminar on Drinfeld Algebras*, Tel Aviv (1991).

[Stolberg] Stolberg, H.J., *Group cohomologies for Hopf algebras and other categories*, Ph.D. Thesis, Cornell University, Ithaca, NY (1969).

[Strad-Farns] Strade, J. and Farnsteiner, R., *Modular Lie Algebras and their Representations*, Pure and Applied Mathematics, Vol. 116, Marcel Dekker, New York (1988).

[Străt] Strătilă, Ş., *Modular Theory in Operator Algebras*, Academiei, Bucareşti, România/Abacus Press, Tunbridge Wells, Kent, UK (1979/1981).

[Sull-1] Sullivan, J.B., *The classification of Hopf algebras with integrals*, Ph.D. Thesis, Cornell University, Ithaca, NY (1971).

[Sull-2] Sullivan, J.B., *The uniqueness of integrals for Hopf algebras and some existence theorems of integrals for commutative Hopf algebras*, J. Algebra **19** (1971), 426–440.

[Sweed-1] Sweedler, M.E., *Cocommutative Hopf algebras with antipode*, Ph.D. Thesis, MIT, Cambridge, MA (1965).

[Sweed-2] Sweedler, M.E., *The Hopf algebra of an algebra applied to field theory*, J. Algebra **8** (3) (1968), 262–276.

[Sweed-3] Sweedler, M.E., *Integrals for Hopf algebras*, Ann. of Math. (2) **89** (1969), 323–335.

[Sweed-4] Sweedler, M.E., *Connected fully reducible affine group schemes in positive characteristic are abelian*, J. Math. Kyoto Univ. **11** (1) (1971), 51–70.

[Sw-1] Sweedler, M.E., *Hopf Algebras*, W.A. Benjamin, New York (1969).

[Sw-2] Sweedler, M.E., *Book review of "Hopf algebras" by Eiichi Abe*, Bull. Amer. Math. Soc. (N.S.) **5** (3) (1981), 349–354.

[Sw-3] Sweedler, M.E., Private communication in Sweedler's office in White Hall at Cornell University (October 11, 1996).

[Sw-4] Sweedler, M.E., E-mail from December 10, 1996.

[Ta-1] Taft, E.J., *The order of the antipode of a finite-dimensional Hopf algebra*, Proc. Nat. Acad. Sci. USA **68** (11) (1971), 2631–2633.

[Ta-2] Taft, E.J., *Linearly recursive tableaux*, Congressus Numerantium **107** (1995), 33–36.

[Ta-3] Taft, E.J., *Hadamard invertibility of linearly recursive sequences in several variables*, Discrete Math. **139** (1995), 393–397.

[Ta-W-1] Taft, E.J. and Wilson, R.L., *On antipodes in pointed Hopf algebras*, J. Algebra **29** (1974), 27–32.

[Ta-W-2] Taft, E.J. and Wilson, R.L., *There exist finite-dimensional Hopf algebras with antipodes of arbitrary even order*, J. Algebra **62** (1980), 283–291.

[Taka] Takahashi, S., *A characterization of group rings as a special class of Hopf algebras*, Canad. Math. Bull. **8** (1965), 465–475.

[Tak-1] Takeuchi, M., *Free Hopf algebras generated by coalgebras*, J. Math. Soc. Japan **23** (1971), 561–582.

[Tak-2] Takeuchi, M., *Topological coalgebras*, J. Algebra **97** (1985), 505–539.

[Tak-3] Takeuchi, M., *Matric bialgebras and quantum groups*, Israel J. Math. **72** (1–2) (1990), 232–251.

[Tak-4] Takeuchi, M., *Finite dimensional representations of the quantum Lorentz group*, Comm. Math. Phys. **144** (1992), 557–580.

[Take-1] Takeuchi, M., *A Hopf algebraic approach to the Picard–Vessiot theory*, J. Algebra **122** (2) (1989), 481–509.

[Take-2] Takeuchi, M., *A short course on quantum matrices*, New Directions in Hopf Algebras, Montgomery, S. and Schneider, H.-J., eds, Mathematical Sciences Research Institute Publication, Vol. 43, Cambridge University Press, Cambridge, UK (2002), 383–435.

[Tannaka] Tannaka, T., *Über den Dualitätssatz der nicht kommutativen topologischen Gruppen*, Tôhoku J. Math. **45** (1938), 1–12.

[Tat] Tate, J., *Homology of Noetherian rings and local rings*, Illinois J. Math. **1** (1957), 14–27.

[T-O] Tate, J. and Oort, F., *Group schemes of prime order*, Ann. Scient. Éc. Norm. Sup. (4) **3** (1970), 1–21.

[Ter] Terras, A., *Harmonic Analysis on Symmetric Spaces and Applications I*, Springer-Verlag, New York (1985).

[Tie] Tietze, H., *Famous Problems of Mathematics. Solved and Unsolved Mathematical Problems from Antiquity to Modern Times*, Authorized translation of the second (1959) revised German edition, Graylock Press (1965).

[Tonge] Tonge, A.M., *Algebras and coalgebras in analysis*, Ph.D. Thesis, University of Cambridge, Cambridge, UK (1976).

[Tru-1] Trushin, D.S., *Coinduced comodules and applications to the representation theory of coalgebras*, Ph.D. Thesis, The Ohio State University, Columbus, OH (1975).

[Tru-2] Trushin, D., *A theorem on induced corepresentations and applications to finite group theory*, J. Algebra **42** (1) (1976), 173–183.

[Tse] Tse, M.Y., *Hopf algebra actions on elementary abelian extensions of degree $P(2)$*, Ph.D. Thesis, State University of New York at Albany, Albany, NY (1997).

[Ulm] Ulm, V., *Operationen von Hopfalgebren in Kategorien von Yetter–Drinfeld Modulen*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (2000).

[Underwood] Underwood, R.G., *Hopf algebra orders over a complete discrete valuation ring, their duals, and extensions of R-groups*, Ph.D. Thesis, State University of New York at Albany, Albany, NY (1992).

[Val] Vallin, J.-M., *C^*-algèbres de Hopf et C^*-algèbres de Kac*, Proc. London Math. Soc. (3) **50** (1985), 131–174.

[Van] van den Essen, A., *Polynomial Automorphisms and the Jacobian Conjecture*, Progress in Mathematics, Vol. 190, Birhäuser-Verlag, Basel (2000).

[Van Ost] Van Ostaeyen, F., *Algebraic Geometry for Associative Algebras*, Pure and Applied Mathematics: A Series of Monographs and Textbooks, Vol. 232, Marcel Dekker, New York (2000).

[Ve-S-1] Verde-Star, L., *Biorthogonal polynomial bases and Vandermonde-like matrices*, Stud. Appl. Math. **95** (1995), 269–295.

[Ve-S-2] Verde-Star, L., *A Hopf algebra structure on rational functions*, Adv. in Math. **116** (1995), 377–388.

[Ve-S-3] Verde-Star, L., *Divided differences and linearly recurrent sequences*, Stud. Appl. Math. **95** (1995), 433–456.

[Ve-S-4] Verde-Star, L., *An algebraic approach to convolutions and transform methods*, Adv. in Appl. Math. **19** (1997), 117–143.

[Verdier] Verdier, J.-L., *Groupes quantiques* [*d'apres V.G. Drinfel'd*], Seminaire Nicholas Bourbaki 39 ème année, Vol. 1986–87, no. 685.

[Vin] Vinberg, E.B., *Linear Representations of Groups*, Basler Lehrbücher, A Series of Advanced Textbooks in Mathematics, Vol. 2, Birkhäuser-Verlag, Basel (1989).

[Wai] Waismann, F., *Introduction to Mathematical Thinking*, Frederick Ungar Publishing Co., New York (1951).

[Wa-1] Waterhouse, W.C., *Antipodes and group-likes in finite Hopf algebras*, J. Algebra **37** (1975), 290–295.

[Wa-2] Waterhouse, W.C., *Introduction to Affine Group Schemes*, Graduate Texts in Mathematics, Vol. 66, Springer-Verlag, New York (1979).

[Wea] Weaver, N., *Mathematical Quantization*, Studies in Advanced Mathematics, Chapman & Hall/CRC, Boca Raton, FL (2001).

[Wei] Weibel, C.A., *An Introduction to Homological Algebra*, Cambridge Studies in Advanced Mathematics, Vol. 38, Cambridge Univ. Press, Cambridge, UK (1994).

[Weinraub] Weinraub, D., *Cofinite induction and Noether's theorem for Hopf orders in group algebras*, Ph.D. Thesis, State University of New York at Albany, Albany, NY (1990).

[Wen] Wen, F., *Topics in quantum groups*, Ph.D. Thesis, The Florida State University, Tallahassee, FL (1996).

[Wenninger] Wenninger, C.H., *Galois-Algebren zu Hopf-Algebren und verallgemeinerte Quaternionen*, Ph.D. Thesis, Ludwig-Maximilians-Universität München, Munich, Germany (1989).

[Weyl] Weyl, H., *The Classical Groups*, Princeton Mathematical Series, Vol. 1, Princeton University Press, Princeton, NJ (1946).

[Wichmann] Wichmann, J., *The theory of Banach coalgebras*, Ph.D. Thesis, Texas Christian University, Ft. Worth, TX (1975).

[Wh-1] Whitehead, G.W., *Homotopy Theory*, MIT Press, Cambridge, MA (1966).

[Wh-2] Whitehead, G.W., *Elements of Homotopy Theory*, Graduate Texts in Mathematics, Vol. 61, Springer-Verlag, New York (1978).

[Williams] Williams, R.E., *Finite dimensional Hopf algebras*, Ph.D. Thesis, The Florida State University, Tallahassee, FL (1988).

[Wils] Wilson, W.S., *Hopf rings in algebraic topology*, Expo. Math. **18** (2000), 369–388.

[Win] Winter, D., *The Structure of Fields*, Graduate Texts in Mathematics, Vol. 16, Springer-Verlag, New York (1974).

[Y] Yanagihara, H., *Theory of Hopf Algebras Attached to Group Schemes*, Lecture Notes in Mathematics, Vol. 614, Springer-Verlag, Berlin (1977).

[Yetter] Yetter, D.N., *Quantum groups and representations of monoidal categories*, Math. Proc. Cambridge Philos. Soc. **108** (2) (1990), 261–290.

[Zas] Zassenhaus, H., *A theorem of Friedrichs*, Proc. and Trans. of the Royal Soc. of Canada, 3rd ser. **51** (1957), Section 3, 55–64.

[Zhang-1] Zhang, J.J., *H-algebras*, Adv. in Math. **889** (2) (1991), 144–191.

[Zhang-2] Zhang, J.J., *The quantum Cayley–Hamilton theorem* J. Pure Appl. Algebra **129** (1) (1998), 101–109.

[Zhu] Zhu, Y., *Hopf algebras of prime dimension*, IMRN (Internat. Math. Res. Notices) **1** (1994), 53–59.

[Zis] Zisman, M., *Espaces de Hopf, algèbres de Hopf*, Séminaire H. Cartan–J.C. Moore, 12e année, Exposé n°2 (1959/60).

[Zoeller-1] Zoeller, M.B., *Freeness of Hopf algebras over grouplike subalgebras*, Ph.D. Thesis, The Florida State University, Tallahassee, FL (1985).

[Zoeller-2] Zoeller, M.B., *Freeness of Hopf algebras over semisimple grouplike subalgebras*, J. Algebra **118** (1988), 102–108.

Section 4A
Lattices and Partially Ordered Sets

This Page Intentionally Left Blank

Frames

Aleš Pultr
Department of Applied Mathematics and Institute of Theoretical Computer Science (ITI), Charles University, Prague, Czech Republic
E-mail: pultr@kam-enterprise.ms.mff.cuni.cz

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

Preface

Have you ever seen a point? That is, an entity with position but no extent? Is it not so that when contemplating a space you rather see "places" or "spots" of non-trivial extent, albeit sometimes very small, and their interrelations (for instance, do they meet or not, how do they group)? Even when thinking of a number on the real line as a solution of a concrete task, we often think of a small interval rather than of a crisp approximating point (what else is viewing $\sqrt{2}$ as "1.414-and-I-don't-know-what-comes-next"?). Building general geometry (topology) on this intuition is one motivation for the theory of frames to be discussed in this chapter. Not the only one; hence it is not surprising that the theory has developed beyond the purely geometric scope. Still, we can think of a frame as of a kind of space, more general than the classical one, allowing us to see topological phenomena in a new perspective. Although the other aspects (connections with domain theory [139, 135] and continuous lattices [10,59], logic [150], etc.) should by no means be ignored, we will, hopefully, be excused if we will treat the subject primarily from this point of view.

An author who is not a specialist in history walks on thin ice when trying to outline the development of a field. But we cannot do without such an outline and can only hope that omitting, by ignorance, important milestones will be pardoned. For a more precise treatment we can advise the reader to see the excellent survey in Johnstone's [73] or [72].

Modern topology originates, in principle, from Hausdorff's "Mengenlehre" [55] in 1914 (one year earlier there was a paper by Caratheodory [38] containing the idea of a point as an entity localized by a special system of diminishing sets; this is also of relevance for the modern point-free thinking). In the twenties and thirties the importance of (the lattice of) open sets became gradually more and more apparent (see, e.g., Alexandroff [2] or Sierpinski [141]). In [144] and [145] (1934, 1936), M.H. Stone presented his famous duality theorem from which it followed that compact zero-dimensional spaces and continuous maps are well represented by the Boolean algebras of closed open sets and lattice homomorphisms. Although zero-dimensional spaces are rather special, and not very geometric, this was certainly an encouragement for those who endeavoured to treat topology other than as a structure on a given system of points (Wallman 1938 [151], Menger 1940 [95], McKinsey and Tarski 1944 [94]). In the Ehresmann seminar in the late fifties [46,33,101], we encounter frame theory already in the form we know today (it should be noted that almost at the same time, independently, there appeared two important papers – Bruns [37], Thron [147] – on homeomorphism of spaces with isomorphic lattices of open sets, under weak separation axioms). After that, many authors got interested (C.H. Dowker, D. Papert (Strauss), J. Isbell, B. Banaschewski, etc.) and the field started to develop rapidly. The pioneering paper by J. Isbell [64], which opened several topics, merits particular mentioning. In 1983, P.T. Johnstone published his monograph [72] which is still a primary source of reference. Since the mid eighties, intensive research has been done in enriched point-free structures such as uniform and nearness frames, or metric frames. It should be noted that this also has its origins in [64].

It has already been said that the notion of a frame can be viewed as an extension of the notion of a (topological) space. Extending or generalizing a notion calls for justification.

Several questions naturally arise: When abandoning points, do we not lose too much information? Is the broader range of "spaces" we now have desirable at all? That is, is the theory in this context, in whatever sense, more satisfactory? And is it not so that the new techniques obscure the geometric contents? Here are some answers:

- Starting with very low separation axioms (sobriety, T_D) the point-free representation contains all the information.
- The broader context does yield, in some areas, better theory. For instance, paracompactness, which is important in applications of topology, is in the classical context badly behaved under constructions. Even products of paracompact and metric spaces are not necessarily paracompact. This is not the case in the point-free context: the category of paracompact frames is coreflective in that of all frames. For more examples see below.
- The class of locally compact spaces is represented equivalently ("no undesirables") by distributive continuous lattices, also a very satisfactory fact.
- The techniques are sometimes less intuitive than the classical one; but it can be argued that they are very often simpler. And, perhaps surprisingly, they often yield constructive results where the classical counterparts cannot. For instance, the Tychonoff product theorem is fully constructive (meaning: no choice principle and no excluded middle, see [70,11]).

The extent of the chapter is naturally limited, however generous the editors have been with the number of pages allowed. Besides the basic facts, both of plain frames and of the enriched ones, I have included in detail, as examples of the merits of the theory, a constructive compactification (implying the constructive Tychonoff product theorem at least for the regular case), a constructive completion of uniform frames, and a discussion of paracompactness. Let me mention a few other interesting items that have not been included – both for volume and for technical reasons:

- (One form of) realcompactness is equivalent to the Lindelöf property (Madden and Vermeer [92], see also [15,137]).
- Connected locally connected frames are in a well-defined sense pathwise connected (Moerdijk and Wraith [96]), and the rift between locally connected connectedness and plain connectedness is also more pronounced as seen from other points of view [86,87].
- Subgroups of localic groups (analogues of topological groups in the point-free context) are always closed [69,76] and localic groups are always complete in the natural uniformity [32].

Other topics that could not have been discussed here are for instance the descriptive theory [68], point-free metric (e.g., [122,119,30]), results relevant for functional analysis [99,100, 107] or the rapid development of non-symmetric uniformity theory (e.g., [49,108,110]).

Just in passing let me mention quantales [34,36,32] which are also partly motivated by point-free topology. These will be given, I am told, a separate chapter in this Handbook.

In almost the whole of the text we keep the frame (algebraic) notation and reasoning. The reader should keep in mind that the localic (geometric) interpretation reads backwards. Thus, the discussion of coproducts of frames concerns products of locales analogous to products of spaces, coreflectivity results are reflectivity ones in the localic interpretation, etc.

1. Preliminaries

The aim of this section is to agree on terminology and notation, and on a few facts assumed.

1.1. *Posets.* The *opposite* of a poset $P = (X, \leqslant)$, that is, P with the order inverted, will be denoted by P^{op}. The least (resp. largest) element, if it exists, will be referred to as the *bottom* or *zero* (resp. *top*) and often denoted by 0 (resp. 1). An element x of P is an *atom* if $x \neq 0$ and $y < x \;\Rightarrow\; y = 0$; a co-atom of P is an atom of P^{op}.

If $M \subseteq P$ is a subset we set

$$\begin{aligned}&{\downarrow}M = \{x \mid \exists y \in M,\ x \leqslant y\},\\&{\uparrow}M = \{x \mid \exists y \in M,\ x \geqslant y\},\\&{\downarrow}x = {\downarrow}\{x\} \quad \text{and} \quad {\uparrow}x = {\uparrow}\{x\}.\end{aligned}$$

A non-void $D \subseteq P$ is said to be *directed* if for any two $x, y \in D$ there is a $z \in D$ such that $x, y \leqslant z$.

A supremum (resp. infimum) of $M \subseteq P$, if it exists, is denoted by $\bigvee M$ (resp. $\bigwedge M$) and often referred to as the *join* (resp. *meet*) of M. We will also use the symbols $a \vee b$ for $\bigvee\{a, b\}$, $\bigvee_{i \in J} x_i$ for $\bigvee\{x_i \mid i \in J\}$, etc.

1.2. *Semilattices, lattices, complete lattices.* A poset P such that for any two $a, b \in P$ there is the infimum $a \wedge b$ will be called (*meet-*)*semilattice*. If we also always have the supremum $a \vee b$ we speak of a lattice. An element $a \neq 1$ of a semilattice is *meet-irreducible* if

$$a = b \wedge c \quad \Rightarrow \quad a = b \text{ or } a = c. \tag{1.2.1}$$

A *complete lattice* is a poset in which each subset has an infimum and a supremum (recall that this is equivalent to requiring, say, only suprema, while a semilattice is not necessarily a lattice). Thus in particular a complete lattice has the bottom $0 = \bigvee \emptyset$ and top $1 = \bigwedge \emptyset$.

A *filter* F in a lattice L is a non-void subset such that

$$\begin{aligned}&a, b \in F \quad \Rightarrow \quad a \wedge b \in F, \quad \text{and}\\&a \in F \text{ and } a \leqslant b \quad \Rightarrow \quad b \in F.\end{aligned}$$

As a rule we will work with non-trivial filters, that is, $F \neq L$. A filter is *prime* if $a \vee b \in F$ implies that either $a \in F$ or $b \in F$, and *completely prime* (or, simply, *complete*) if $\bigvee M \in F$ implies that $M \cap F \neq \emptyset$.

1.3. *Galois adjunction.* Monotone maps $f : P \to Q$, $g : Q \to P$ are (*Galois*) *adjoint* (f is the *left adjoint* of g, g is the *right adjoint* of f) if

$$f(x) \leqslant y \quad \text{iff} \quad x \leqslant g(y). \tag{GA}$$

This condition is equivalent to

$$fg(y) \leqslant y \quad \text{and} \quad gf(x) \geqslant x \quad \text{for all } x \in P, y \in Q. \tag{GA$'$}$$

The following are well known (and easy) facts:

- *g is uniquely determined by f, and vice versa,*
- *a left adjoint preserves all existing suprema and a right one preserves all existing infima,*
- *if P and Q are complete lattices then each suprema preserving $f : P \to Q$ is a left adjoint and each infima preserving $g : Q \to P$ is a right adjoint.*

1.4. *Distributive lattices.* A lattice L is distributive if

$$\forall a, b, c \in L,\ a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c), \tag{1.4.1}$$

which is equivalent to

$$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c). \tag{1.4.2}$$

Another characteristic of distributive lattices is that

the system of equations $a \vee x = b,\ a \wedge x = c$ has, for any a, b, c, at most one solution.

In a distributive lattice, *a meet irreducible element $a \neq 1$ is characterized by the implication*

$$b \wedge c \leqslant a \quad \Rightarrow \quad b \leqslant a \text{ or } c \leqslant a \tag{1.4.3}$$

which is often handier than (1.2.1) above.

1.5. *Topology.* The structure of a topological space will be as a rule described by the system $\Omega(X)$ of all the open sets of X. Recall that $\Omega(X)$ is a complete lattice with $\bigvee U_i = \bigcup U_i$ and $U \wedge V = U \cap V$ (while the infima of infinite systems typically differ from the intersections).

Continuous maps $f : X \to Y$ are characterized by the property that

$$\text{for each } U \in \Omega(Y), \quad f^{-1}(U) \in \Omega(X).$$

If also $f[U] \in \Omega(Y)$ for each $U \in \Omega(X)$ we speak of an *open map.*

The reader is certainly acquainted with the standard separation axioms T_0, T_1, T_2 (Hausdorff axiom), T_3 (regularity), $T_{3\frac{1}{2}}$ (complete regularity) and T_4 (normality). Furthermore we will use the axiom

$$\text{for each } x \in X \text{ there is an open } U \ni x \text{ such that } U \backslash \{x\} \text{ is open} \tag{T_D}$$

and the axiom of *sobriety*

$$\text{each meet-irreducible } U \neq X \text{ in } \Omega(X) \text{ is of the form } X \backslash \overline{\{x\}} \text{ with a unique } x \in X. \tag{Sob}$$

(Note that $X\backslash\overline{\{x\}}$ is meet irreducible in any space; thus, a space is sober if there are no others and if it is T_0.) T_D is stronger than T_0 and weaker than T_1, (Sob) is weaker than T_2 and incomparable with either of T_D, T_1.

1.5.1. *Scott topology.* Let $P = (X, \leqslant)$ be a poset. A subset $U \subseteq P$ is *Scott-open* if $\uparrow U = U$ and if $U \cap D \neq \emptyset$ whenever D is directed and $\bigvee D \in U$. (Roughly speaking, the Scott topology is the topology in which suprema of directed sets appear as limits.)

1.5.2. *Uniform spaces.* Uniformities can be introduced in various ways. We will think of a uniformity as a filter $\mathcal{U}$ of open covers (with respect to the refinement preorder) such that for each $U \in \mathcal{U}$ there is a star-refinement $V \in \mathcal{U}$, and that the topology induced by $\mathcal{U}$ coincides with the original one.

For more about topology, the reader may consult [84,35] or [47].

1.6. *Categories.* The reader is certainly acquainted with basic categorical notions (category, subcategory, full subcategory and full embedding, functor covariant and contravariant, the dual category $\mathcal{C}^{\mathrm{op}}$, mono-, epi- and isomorphism, transformation, natural equivalence, limit and colimit of a diagram). Recall that a category is *complete* (resp. *cocomplete*) if every diagram has a limit (resp. colimit) and that for completeness (resp. cocompleteness) it suffices to have products and equalizers (resp. coproducts and coequalizers).

The functors $L : \mathcal{A} \to \mathcal{B}$ and $R : \mathcal{B} \to \mathcal{A}$ are adjoint (L is a left adjoint of R, R is a right adjoint of L; the situation is often indicated by writing $L \dashv R$) if the functors in two variables $\mathcal{B}(L(-), -)$ and $\mathcal{A}(-, R(-))$ are naturally equivalent. This is equivalent to the existence of *adjunction units*, transformations

$$\eta : \mathrm{Id} \to RL, \qquad \varepsilon : LR \to \mathrm{Id}$$

such that the composite transformations

$$L \xrightarrow{L\eta} LRL \xrightarrow{\varepsilon L} L \quad \text{and} \quad R \xrightarrow{\eta R} RLR \xrightarrow{R\varepsilon} R$$

are the identities of L resp. R.

A *(co)reflective subcategory* $\mathcal{A}$ of $\mathcal{B}$ is a full subcategory such that the embedding $\mathcal{A} \subseteq \mathcal{B}$ is a right (left) adjoint. A (co)reflective subcategory $\mathcal{A}$ of $\mathcal{B}$ is closed under all (co)limits taken in $\mathcal{B}$.

For more about categories the reader may consult, e.g., [89] or [1].

2. Heyting and pseudocomplemented algebras

2.1. A *Heyting algebra* is a non-empty lattice L with an additional binary operation $a \to b$ satisfying

$$a \wedge b \leqslant c \quad \text{iff} \quad a \leqslant b \to c. \tag{H}$$

If L, M are Heyting algebras, a lattice homomorphism $h : L \to M$ is said to be *Heyting* if $h(a \to b) = h(a) \to h(b)$.

Note that obviously

$$a \to b \text{ increases in the second variable and decreases in the first one.} \tag{2.1.1}$$

A lattice L may or may not admit a Heyting operation. A necessary condition for the existence of a Heyting operation is that

$$\Bigl(\bigvee a_i\Bigr) \wedge b = \bigvee (a_i \wedge b) \tag{2.1.2}$$

whenever the supremum $\bigvee a_i$ exists (indeed, the formula (H) makes $(x \mapsto x \wedge b) : L \to L$ a left Galois adjoint – recall 1.3). In particular,

a Heyting algebra is always distributive.

If the lattice L is complete, the stronger distributive law (2.1.2) is in fact (recall 1.3) a necessary and sufficient condition for the existence of a Heyting operation.

The fact that $\to$ is determined by the lattice structure, however, does not mean that a lattice homomorphism should be Heyting. For a general lattice homomorphism h we have only

$$h(a \to b) \leqslant h(a) \to h(b)$$

which will immediately follow from (H8) below.

2.2. *Some easy Heyting formulas.* First note that the formula (H), besides the Galois adjunction between $(x \mapsto x \wedge b)$ and $(x \mapsto (b \to x))$ also implies that

$$a \leqslant b \to c \quad \text{iff} \quad b \leqslant a \to c \tag{H$'$}$$

and hence a Galois adjunction between $(x \mapsto (x \to c)) : L \to L^{\mathrm{op}}$ on the left and $(x \mapsto (x \to c)) : L^{\mathrm{op}} \to L$ on the right. Thus we have

(H1) $a \to \bigwedge b_i = \bigwedge (a \to b_i)$ and

(H2) $(\bigvee a_i) \to b = \bigwedge (a_i \to b)$.

A FEW MORE FORMULAS.

(H3) *A Heyting algebra has a top* 1 *and one has* $a \to a = 1$ *and* $1 \to a = a$.

(H4) $a \leqslant b$ *iff* $a \to b = 1$.

(H5) $a \leqslant b \to a$.

(H6) $a \to b = a \to (b \wedge a)$.

(H7) $(a \wedge b) \to c = a \to (b \to c) = b \to (a \to c)$.

(H8) $a \wedge (a \to b) \leqslant b$.

(H9) $a \wedge (a \to b) = a \wedge b$.

(H10) $a \leqslant (a \to b) \to b$.

(H11) $a \to b = ((a \to b) \to b) \to b$.
(H12) $a \wedge b = a \wedge c$ *iff* $a \to b = a \to c$.
(H13) $x = (x \vee a) \wedge (a \to x)$.

PROOF. (H3), (H4), (H5) and (H6) immediately follow from (H) (as for the first statement in (H3), $x \wedge a \leqslant a$ and hence $x \leqslant a \to a$ for all x; (H6) since $x \wedge a \leqslant b$ iff $x \wedge a \leqslant b \wedge a$).

(H7): We have $a \wedge b \leqslant a$; $x \leqslant a \to b$ iff $x \wedge a \leqslant b$ iff $x \wedge a \leqslant a \wedge b$; $x \leqslant (a \wedge b) \to c$ iff $x \wedge a \wedge b \leqslant c$ iff $x \wedge b \leqslant a \to c$ iff $x \leqslant a \to (b \to c)$.

(H8) follows from (H) and from $a \to b \leqslant a \to b$.

(H9) follows from (H8) and (H5), and (H10) follows immediately from (H8).

(H11): Substituting $a \to b$ for a in (H10) we obtain $\leqslant$, applying (H10) in the first variable of $x \to b$ we obtain $\geqslant$ (recall (2.1.1)).

(H12) follows from (H6) and (H9).

(H13): We have $x \leqslant (x \vee a) \wedge (a \to x)$ by (H5); by distributivity and (H8), $(x \vee a) \wedge (a \to x) = (x \wedge (a \to x)) \vee (a \wedge (a \to x)) \leqslant x$. □

NOTE (*Modus ponens*). However trivial, (H8) is a very important fact. Heyting algebras model intuitionistic logic in the sense in which Boolean algebras model the classical one; when we think of the inequality as inference and of the Heyting operation as implication, we see that (H8) is the *modus ponens* rule $a \ \& \ (a \Rightarrow b) \ \vdash \ b$.

2.3. *Pseudocomplements and pseudocomplemented lattices.* Let L be a lattice with zero (bottom) 0, let $a \in L$. An element a^* is a *pseudocomplement* of a if

$$x \wedge a = 0 \quad \text{iff} \quad x \leqslant a^*. \tag{PC}$$

(Using the operational symbol a^* is correct since there is obviously at most one pseudocomplement of a: if (PC) holds and if we have also $x \wedge a = 0$ iff $x \leqslant b$ then $x \leqslant b$ iff $x \leqslant a^*$ and hence $b = a^*$.)

A lattice L with 0 is said to be *pseudocomplemented* if each element of L has a pseudocomplement.

A pseudocomplemented lattice necessarily has a top 1, namely 0^*, since $x \leqslant 0^*$, that is, $x \wedge 0 = 0$, holds for any x. On the other hand, $x \leqslant 1^*$ iff $x = x \wedge 1 = 0$ and hence we have
(PC1) $0^* = 1$ and $1^* = 0$.
Here are a few easy but important rules
(PC2) $a \leqslant b \Rightarrow b^* \leqslant a^*$,
(PC3) $a \leqslant a^{**}$,
(PC4) $a^* = a^{***}$, and
(PC5) $a \wedge b = 0$ iff $a^{**} \wedge b = 0$
(if $a \leqslant b$ then $a \wedge b^* = 0$; (PC3) follows from $a \wedge a^* = 0$; (PC4) is obtained from (PC3) using, first, (PC2) and, second, substituting a^* for a – compare with (H10) and (H11); (PC5) is a reformulation of (PC4)).

Next, since $x \leqslant y^*$ iff $x \wedge y = 0$ iff $y \leqslant x^*$ we have the Galois adjunction between $(x \mapsto x^*) : L^{\mathrm{op}} \to L$ and $(x \mapsto x^*) : L \to L^{\mathrm{op}}$ and hence we have the DeMorgan law
(PC6) $(\bigvee a_i)^* = \bigwedge a_i^*$
whenever $\bigvee a_i$ makes sense. Note that the other DeMorgan law does not hold in general.

2.4. OBSERVATION. *Each Heyting algebra with* 0 *is pseudocomplemented and we have*

$$x^* = x \to 0.$$

2.5. *Complements.* Let L be a distributive lattice with 0 and 1. The *complement* of an element $a \in L$ is an element $\bar{a} \in L$ satisfying

$$a \wedge \bar{a} = 0 \quad \text{and} \quad a \vee \bar{a} = 1.$$

If $\bar{a}$ exists we say that a is *complemented.*

Since in a distributive lattice the couple of equations $a \wedge x = b$, $a \vee x = c$ has at most one solution, we see that

(1) *An element of a distributive lattice has at most one complement.*

(2) *If a is complemented, so is $\bar{a}$ and we have $\bar{\bar{a}} = a$.*

(3) *For any* $0, 1$*-preserving lattice homomorphism $h : L \to M$ the image of a complemented element is complemented, and we have*

$$h(\bar{a}) = \overline{h(a)}.$$

Note that for pseudocomplements and general 0, 1-preserving lattice homomorphisms we only have the obvious inequality

$$h(a^*) \leqslant h(a)^*. \tag{2.5.1}$$

OBSERVATION. *In a distributive lattice, each complement is a pseudocomplement.*

(If $x \wedge a = 0$ then $x = x \wedge (a \vee \bar{a}) = x \wedge \bar{a}$ and hence $x \leqslant \bar{a}$.)

2.6. *Boolean algebras.* A *Boolean algebra* is a distributive lattice with 0 and 1 in which every element has a complement.

2.6.1. PROPOSITION. *A Boolean algebra is a Heyting algebra, with the Heyting operation given by $a \to b = \bar{a} \vee b$.*

PROOF. If $a \wedge b \leqslant c$ then $a = a \wedge (b \vee \bar{b}) = (a \wedge b) \vee (a \wedge \bar{b}) \leqslant c \vee \bar{b}$. If $a \leqslant \bar{b} \vee c$ then $a \wedge b \leqslant a \wedge c \leqslant c$. □

2.6.2. COROLLARY. *In a Boolean algebra we have*

$$\left(\bigvee a_i\right) \wedge b = \bigvee (a_i \wedge b)$$

whenever the supremum $\bigvee a_i$ exists.

2.6.3. By 2.5(2), if L is Boolean, the mapping $(a \mapsto a^*) : L \cong L^{\mathrm{op}}$ is an isomorphism. Hence we also have

$$\left(\bigwedge a_i\right) \vee b = \bigwedge (a_i \vee b)$$

whenever the infimum $\bigwedge a_i$ exists, and besides the DeMorgan law (PC6) we also have the dual one,

$$\overline{\bigwedge a_i} = \bigvee \overline{a_i}.$$

2.7. *Booleanization.* Let L be a distributive pseudocomplemented lattice (for instance, a Heyting algebra with 0). Recall that by (PC2)–(PC4), $(a \mapsto a^{**})$ is monotone and $a \leqslant a^{**} = a^{****}$. The following formula is easy although slightly less obvious:

$$(a \wedge b)^{**} = a^{**} \wedge b^{**} \tag{2.7.1}$$

($\leqslant$ follows from the monotonicity; further, $a \wedge b \wedge (a \wedge b)^* = 0$ and hence, using (PC5) twice we obtain $a^{**} \wedge b^{**} \wedge (a \wedge b)^* = 0$ so that $a^{**} \wedge b^{**} \leqslant (a \wedge b)^{**}$).

PROPOSITION. *Let L be a distributive pseudocomplemented lattice* (*for instance, a Heyting algebra with* 0). *A system $\{a_i \mid i \in J\} \subseteq \mathcal{B}L = \{x \mid x = x^{**}\}$ has a supremum in $\mathcal{B}L$ whenever $\bigvee a_i$ exists in L, namely $\bigvee' a_i = (\bigvee a_i)^{**}$. The mapping*

$$\beta_L = (x \mapsto x^{**}) : L \to \mathcal{B}L$$

preserves (*finite*) *meets and all existing suprema, and $\mathcal{B}L$ is a Boolean algebra.*

PROOF. If $x \geqslant a_i$ for all i and $x \in \mathcal{B}L$ then $x \geqslant \bigvee a_i$ and $x = x^{**} \geqslant (\bigvee a_i)^{**}$. We have $\beta(a \wedge b) = \beta(a) \wedge \beta(b)$ by (2.7.1), and

$$\beta\left(\bigvee a_i\right) = \left(\bigvee a_i\right)^{**} \leqslant \left(\bigvee a_i^{**}\right)^{**} = \bigvee{}' \beta(a_i) \leqslant \beta\left(\bigvee a_i\right).$$

Since β is onto, $\mathcal{B}L$ satisfies the equalities in $\vee$ and $\wedge$ satisfied generally in L, in particular the distributivity. Finally, in $\mathcal{B}L$ we have $a \wedge a^* = 0$ and $(a \vee a^*)^* = a^* \wedge a^{**} = 0$ so that $a \vee' a^* = (a \vee a^*)^{**} = 0^* = 1$. Thus a^* is the complement of a in $\mathcal{B}L$. □

The homomorphism $\beta : L \to \mathcal{B}L$, or just the Boolean algebra $\mathcal{B}L$, is often called the *Booleanization* of L.

3. Frames

3.1. A *frame* is a complete lattice L satisfying the distributivity law

$$\left(\bigvee A\right) \wedge b = \bigvee \{a \wedge b \mid a \in A\} \tag{$*$}$$

for each subset $A \subseteq L$ and $b \in L$.

A *frame homomorphism* $h : L \to M$ is a mapping preserving all suprema (including the bottom 0) and all finite infima (including the top 1). The category of frames and frame homomorphisms will be denoted by

$$\mathbf{Frm}.$$

NOTE. The condition $(*)$, stating that the mappings $m_b = (x \mapsto x \wedge b) : L \to L$ preserve all suprema, is by 1.3 equivalent with the existence of the Heyting operation $\to$ on L. Thus, a frame and a complete Heyting algebra is basically the same. What makes the structure special is the choice of morphisms. Frame homomorphisms are not necessarily Heyting (indeed, they seldom are – see Section 9 below), neither are they complete lattice homomorphisms in the sense of preserving also the infinite meets.

3.2. *Spaces and frames.* The lattice $\Omega(X)$ of open sets of a topological space X is a frame; if $f : X \to Y$ is a continuous map we have a frame homomorphism $\Omega(f) = (U \mapsto f^{-1}(U)) : \Omega(Y) \to \Omega(X)$. Thus we have a contravariant functor

$$\Omega : \mathbf{Top} \to \mathbf{Frm}.$$

The dual of the category **Frm**, usually denoted as **Loc**, is called the category of *locales*. Thus we have a (covariant) functor

$$\Omega : \mathbf{Top} \to \mathbf{Loc}$$

which, as we will see in 4.1.1, restricts to a full embedding on a very sizable part of **Top**. This justifies thinking of **Loc** as a category of generalized spaces.

A frame L isomorphic to an $\Omega(X)$ is said to be *spatial*.

3.3. *Another important case: Boolean algebras.* A complete Boolean algebra is a frame (recall 2.6.2). Note that a frame homomorphism $h : B \to B'$ between complete Boolean algebras is always a complete Boolean homomorphism (the complements are preserved by 2.5(3), and preserving general infima is readily deduced using the DeMorgan formula). Thus, the standard category of complete Boolean algebras (often referred to as *Boolean frames*) is a full subcategory of **Frm**. We will see (4.7 below) that Boolean frames are typically non-spatial; in fact, the classes of spatial frames and Boolean frames intersect exactly in the class of (complete) atomic Boolean algebras (which represent the discrete spaces).

3.4. *Meet-semilattices, down-set functor and free frames.* Denote by

$$\mathbf{SLat}$$

the category of bounded meet-semilattices (that is, meet semilattices with (bottom) 0 and (top) 1) with $(0, 1, \wedge)$-homomorphisms. For $S \in \mathbf{SLat}$ set

$$\mathfrak{D}S = \big(\{A \subseteq S \mid \emptyset \neq A = {\downarrow}A\}, \subseteq\big)$$

and for a homomorphism $h : S \to T$ define

$$\mathfrak{D}h = \big(A \mapsto {\downarrow}h[A]\big) : \mathfrak{D}S \to \mathfrak{D}T.$$

Obviously $\mathfrak{D}S$ is a frame and $\mathfrak{D}h$ is a frame homomorphism, and we obtain a functor (often referred to as the *down-set functor*)

$$\mathfrak{D} : \mathbf{SLat} \to \mathbf{Frm}.$$

(The meets in $\mathfrak{D}S$ are the intersections. Note that the suprema in $\mathfrak{D}S$ are the unions, *with one important exception*: namely $\bigvee \emptyset = \{0\}$, not the empty set which does not appear in $\mathfrak{D}S$.)

3.4.1. PROPOSITION. *The functor $\mathfrak{D}$ is a left adjoint to the natural forgetful functor $\mathcal{U} : \mathbf{Frm} \to \mathbf{SLat}$ (forgetting the join part of the structure).*

PROOF. Set

$$\eta_S = (a \mapsto {\downarrow}a) : S \to \mathcal{U}\mathfrak{D}S,$$
$$\varepsilon_L = \Big(A \mapsto \bigvee A\Big) : \mathfrak{D}\mathcal{U}L \to L.$$

Obviously each η_S is a homomorphism and each ε_L is a frame homomorphism (by distributivity and since A, B are down-sets, $\varepsilon(A) \wedge \varepsilon(B) = \bigvee A \wedge \bigvee B = \bigvee\{a \wedge b \mid a \in A, b \in B\} \leqslant \bigvee\{c \mid c \in A \cap B\} = \varepsilon(A \cap B) \leqslant \varepsilon(A) \cap \varepsilon(B)$; clearly, $\varepsilon(\bigcup A_i) = \bigvee \varepsilon(A_i)$ and $\varepsilon(\{0\}) = 0$). Checking that the collections $\eta = (\eta_S)_S$ and $\varepsilon = (\varepsilon_L)_L$ constitute natural transformations is trivial.

We have

$$\mathcal{U}\varepsilon_L\big(\eta_{\mathcal{U}L}(a)\big) = \bigvee({\downarrow}a) = a$$

and

$$\varepsilon_{\mathfrak{D}S}\big(\mathfrak{D}\eta_S(A)\big) = \bigvee\{{\downarrow}a \mid a \in A\} = A$$

and hence η and ε are adjunction units. □

3.4.2. In other words we have

PROPOSITION. *For every (bounded) meet-semilattice S, for every frame L and for every $(0, 1, \wedge)$-homomorphism $h : S \to L$ there is exactly one frame homomorphism $g : \mathfrak{D}S \to L$ such that $g({\downarrow}a) = h(a)$, namely the mapping given by $g(A) = \bigvee\{h(a) \mid a \in A\}$.*

The frame $\mathfrak{D}S$ is (therefore) usually referred to as the *free frame over S*.

4. Reconstruction of a space. Points, spectra

4.1. *Reconstructing a sober space.* The question naturally arises as to how much information is lost when considering the lattice $\Omega(X)$ instead of the space X. (For the indiscrete space, $\Omega(X)$ is always just the two-point Boolean algebra

$$\mathbf{2} = \{0, 1\}$$

regardless the size of X.) We will see that sober spaces, and continuous maps between them, can be fully reconstructed.

4.1.1. PROPOSITION. *Let Y be a sober space and X a general one. Then for each frame homomorphism $h: \Omega(Y) \to \Omega(X)$ there is exactly one continuous map $f: X \to Y$ such that $h = \Omega(f)$. Thus, the restriction*

$$\Omega: \mathbf{Sob} \to \mathbf{Loc}$$

of Ω is a full embedding.

PROOF. The uniqueness immediately follows from Y being T_0. Now let $h: \Omega(Y) \to \Omega(X)$ be a frame homomorphism. For $x \in X$ set

$$\mathcal{F}_x = \bigl\{U \in \Omega(Y) \mid x \notin h(U)\bigr\} \quad \text{and} \quad F_x = \bigcup \mathcal{F}_x.$$

Since joins are preserved we have $x \notin h(F_x)$ and hence, for $U \in \Omega(Y)$,

$$x \notin h(U) \quad \text{iff} \quad U \subseteq F_x. \tag{$*$}$$

F_x is meet irreducible: Indeed, since $x \notin h(F_x)$, $F_x \neq X$; if $F_x = U \cap V$ we have $x \notin h(U) \cap h(V)$ and hence, say $x \notin h(U)$ and $U \subseteq F_x$. Thus, by sobriety, $F_x = X \backslash \overline{\{y\}}$ for a unique $y \in Y$ and if we chose such y for $f(x)$ we can rewrite $(*)$ to

$$x \notin h(U) \quad \text{iff} \quad U \subseteq X \backslash \overline{\{y\}} \quad (\text{iff } f(x) \notin U, \text{ since } U \text{ is open})$$

and hence

$$x \in h(U) \quad \text{iff} \quad f(x) \in U, \text{ that is, } x \in f^{-1}(U).$$

Thus, $f^{-1}(U) = h(U) \in \Omega(X)$ and hence f is continuous and $h = \Omega(f)$. □

4.1.2. Using 4.1.1 we can reconstruct a sober space X as follows:

Denote by P the one point space $\{\cdot\}$. The $x \in X$ are in the natural one–one correspondence with the (continuous) maps $f_x = (\cdot \mapsto x): P \to X$, and hence with the frame homomorphisms

$$h: \Omega(X) \to \Omega(P) \cong 2.$$

An element x belongs to an open set U iff $h(U) = f_x^{-1}(U) \neq \emptyset$. Thus, X is homeomorphic with

$$\big(\{h \mid h : \Omega(X) \to 2\}, \{\widetilde{U} = \{h \mid h(U) = 1\} \mid U \in \Omega(X)\}\big).$$

4.2. *Spectrum.* The construction above leads to the following definitions:

A *point* of a frame L is a frame homomorphism $h : L \to 2$. Denote by ΣL the set of all points of L. For $a \in L$ set $\Sigma_a = \{h : L \to 2 \mid h(a) = 1\}$.

OBSERVATION. *$\Sigma_{a \wedge b} = \Sigma_a \cap \Sigma_b$, $\Sigma_0 = \emptyset$, $\Sigma_1 = \Sigma L$ and $\Sigma_{\bigvee a_i} = \bigcup \Sigma_{a_i}$.*

Consequently $\{\Sigma_a \mid a \in L\}$ constitutes a topology on ΣL. From now on, ΣL will be always considered as thus obtained space, and called the *spectrum* of L.

For a frame homomorphism $h : L \to M$ consider the mapping

$$\Sigma h : \Sigma M \to \Sigma L$$

defined by $(\Sigma h)(\alpha) = \alpha h$.

LEMMA. *For each $a \in L$,*

$$(\Sigma h)^{-1}(\Sigma_a) = \Sigma_{h(a)}. \tag{2.1}$$

PROOF. Indeed, $(\Sigma h)(\alpha) \in \Sigma_a$ iff $(\Sigma h)(\alpha)(a) = \alpha(h(a)) = 1$. $\square$

COROLLARY. *$\Sigma h : \Sigma M \to \Sigma L$ is a continuous mapping and we have obtained a (contravariant) functor*

$$\Sigma : \mathbf{Frm} \to \mathbf{Top}.$$

4.3. *Alternative descriptions of the spectrum.*

4.3.1. Obviously there is a one–one correspondence

$$F \mapsto \alpha_F,\ \alpha_F(x) = 1 \quad \text{iff} \quad x \in F \text{ and } \alpha \mapsto F_\alpha = \{x \mid \alpha(x) = 1\}$$

between frame homomorphisms $\alpha : L \to 2$ and complete filters (recall 1.2) on L. Thus we can represent the spectrum as the space

$$\Sigma L = \big(\{F \mid F \text{ complete filter in } L\}, \{\Sigma_a \mid a \in L\}\big)$$

with $\Sigma_a = \{F \mid a \in F\}$, $(\Sigma f)(F) = f^{-1}(F)$.

4.3.2. Another representation of the spectrum is based on the following very easy

LEMMA. *Set* $\mathsf{mirr}(L) = \{a \mid a \in L,\ a \text{ meet irreducible}\}$. *Then the formulas*

$$x \mapsto \alpha_x,\ \alpha_x(a) = 0 \quad \textit{iff} \quad a \leqslant x, \qquad \alpha \mapsto x_\alpha = \bigvee\{a \mid \alpha(a) = 0\}$$

constitute a one–one correspondence between ΣL *and* $\mathsf{mirr}(L)$.

4.4. PROPOSITION. *Each* ΣL *is a sober space.*

PROOF. Use the standard representation of ΣL. We have

$$\alpha \in \overline{\{\beta\}} \quad \text{iff} \quad \alpha \leqslant \beta$$

(indeed, the first formula says that if $\alpha \in \Sigma_a$, that is, $\alpha(a) = 1$ then $\beta \in \Sigma_a$, that is, $\beta(a) = 1$).

Let Σ_a be a meet irreducible open set in ΣL. Set $b = \bigvee\{c \mid \Sigma_c \subseteq \Sigma_a\}$; hence in particular $\Sigma_b = \Sigma_a$. If $x \wedge y \leqslant b$ then $\Sigma_x \cap \Sigma_y \subseteq \Sigma_b = \Sigma_a$ and hence, say, $\Sigma_x \subseteq \Sigma_b$ so that $x \leqslant b$. Thus, b is meet irreducible. Consider the $\alpha = \alpha_b$ from 4.3.2. We have

$$\begin{aligned}&\beta \notin \overline{\{\alpha\}} \quad \text{iff} \quad (\exists c, \beta(c) = 1 \text{ and } \alpha(c) = 0) \quad \text{iff}\\ &(\exists c, \beta(c) = 1 \text{ and } c \leqslant b) \quad \text{iff} \quad \beta(b) = 1 \quad \text{iff} \quad \beta \in \Sigma_b = \Sigma_a.\end{aligned}$$

Thus, $\Sigma_a = \Sigma L \backslash \overline{\{\alpha\}}$. □

4.5. *Spectrum is adjoint to* Ω.

THEOREM. $\Sigma : \mathbf{Loc} \to \mathbf{Top}$ *is a right adjoint to* $\Omega : \mathbf{Top} \to \mathbf{Loc}$.

PROOF. The covariant ("localic") formulation enables us to say which of the functors is to be viewed as the right adjoint and which one as the left adjoint. The proof will be done, however, using the algebraic (frame) reasoning.

For a topological space X define

$$\eta_X : X \to \Sigma\Omega X$$

by setting $\eta_X(x)(U) = 1$ iff $x \in U$ (checking that each $\eta_X(x)$ is a frame homomorphism is straightforward, and we have

$$\eta_X^{-1}(\Sigma_U) = \bigl\{x \mid \eta_X(x) \in \Sigma_U\bigr\} = U) \tag{4.5.1}$$

so that each η_X is continuous.

For a frame L define

$$\varepsilon_L : L \to \Omega\Sigma L \quad (\Omega\Sigma L \to L \text{ in } \mathbf{Loc})$$

by setting $\varepsilon_L(a) = \Sigma_a$ (by the Observation in 4.2 it is a frame homomorphism).

If $f : X \to Y$ is a continuous map ($h : L \to M$ a frame homomorphism), we have $\Sigma\Omega f(\eta_X(x))(U) = \eta_X(x)(\Omega f(U)) = \eta_X(x)(f^{-1}(U)) = 1$ iff $x \in f^{-1}(U)$ iff $f(x) \in U$ iff $\eta_Y(f(x))(U) = 1$ ($\Omega\Sigma h(\varepsilon_L(a)) = (\Sigma h)^{-1}(\Sigma_a) = \Sigma_{h(a)} = \varepsilon_M(h(a))$, by (2.1)) so that we have natural transformations $\eta : \mathrm{Id} \to \Sigma\Omega$ and $\varepsilon : \mathrm{Id} \to \Omega\Sigma$. These transformations are adjunction units: We have

$$\big(\Sigma\varepsilon_L\big(\eta_{\Sigma L}(\alpha)\big)\big)(U) = \eta_{\Sigma L}(\alpha)\big(\varepsilon_L(U)\big) = 1 \quad \text{iff} \quad \alpha \in \Sigma_U \quad \text{iff} \quad \alpha(U) = 1,$$

hence $\Sigma\varepsilon_L \cdot \eta_{\Sigma L} = \mathrm{id}$, and, by (4.5.1)

$$\Omega(\eta_X)\big(\varepsilon_{\Omega X}(U)\big) = \eta_X^{-1}(\Sigma U) = U. \tag{4.5.2}$$

□

4.6. *A spatiality criterion.* The unit ε yields an easy spatiality criterion.

PROPOSITION. *The following statements are equivalent*:

(1) *L is spatial,*
(2) *ε_L is one–one,*
(3) *ε_L is an isomorphism.*

PROOF. Since each ε_L is onto, (2) ⇔ (3). Trivially (3) ⇒ (1).

(1) ⇒ (2): Let $h : L \to \Omega(X)$ be an isomorphism. Then $\varepsilon_L = (\Omega\Sigma h)^{-1} \cdot \varepsilon_{\Omega X} \cdot h$, and $\varepsilon_{\Omega X}$ is one–one since $\Omega\eta_X \cdot \varepsilon_{\Omega X} = \mathrm{id}$. □

Thus, *if ε_L is not an isomorphism, L is isomorphic to no $\Omega(X)$ whatsoever.*

4.7. *Spatial Boolean algebras.* The intersection of the classes of frames introduced in 3.2 and 3.3 consists of (frame representation of) discrete spaces only.

LEMMA. *Each meet irreducible element in a Boolean algebra is a co-atom.*

PROOF. Let a be meet-irreducible and let $a < x$. We have $a = a \vee (x \wedge x^*) = x \wedge (a \vee x^*)$, hence by meet irreducibility $a \vee x^* \leqslant a$, that is, $x^* \leqslant a$ and consequently $a^* \leqslant x$. Thus, $x \geqslant a \vee a^* = 1$. □

PROPOSITION. *Let B be a Boolean algebra. Then the following statements are equivalent*:

(1) *B is spatial,*
(2) *each $b \in B$ is a join of atoms* (*equivalently, each $b \in B$ is a meet of co-atoms*),
(3) *B is isomorphic to the set of all subsets of a set, that is, to $\Omega(X)$ with X discrete.*

PROOF. (2) ⇔ (3) is a well-known (and easy) fact and (3) ⇒ (1) is trivial.

(1) ⇒ (2): By 4.6 and 4.3.2, if $x \nleqslant y$ there is a meet irreducible (and hence, by Lemma, co-atom) a such that $x \nleqslant a$ and $y \leqslant a$. Thus, $y = \bigwedge\{a \mid a \text{ co-atom}, y \leqslant a\}$. □

4.8. *Spectrum and sober reflection.* The unit η constitutes a reflection of **Top** to **Sob** (and a sobriety criterion).

PROPOSITION. *The following statements are equivalent*:

(1) *X is sober,*

(2) *η_X is an invertible mapping,*

(3) *η_X is a homeomorphism.*

PROOF. (1) $\Rightarrow$ (2) by 4.1.1, and (3) $\Rightarrow$ (1) follows from 4.4.

(2) $\Rightarrow$ (3): By (4.5.1), $\eta_X[U] = \eta_X[\eta_X^{-1}(\Sigma_U)] = \Sigma_U$ so that an invertible η_X is also open. $\square$

5. Sublocales, congruences, nuclei

5.1. *Various ways of representing generalized subspaces.*

5.1.1. One of the ways to represent subobjects in a category (as opposed to just one–one structure-preserving maps that are usually adequately represented as monomorphisms) is using the *extremal monomorphisms*, that is, monomorphisms μ that cannot be decomposed as $\mu = \alpha\beta$ with β an epimorphism unless β is an isomorphism. The dual notion is, of course, an *extremal epimorphism*, an epimorphism ε such that in each decomposition $\varepsilon = \alpha\beta$ with α monomorphic, α is an isomorphism. It is easy to see that

> the extremal epimorphisms in **Frm** (the extremal monomorphisms in **Loc**) are exactly the frame homomorphisms onto.

(It should be noted that while this is absolutely straightforward, the structure of plain epimorphisms in **Frm** is very complex – see also 5.8 below.)

Furthermore, taking into account the fact that with the embedding $j : Y \subseteq X$ of spaces we have the associated onto homomorphisms $\Omega(j) = (U \mapsto U \cap Y) : \Omega(X) \to \Omega(Y)$ it is natural to define *sublocales* (more exactly, *sublocale maps*) of L as

> onto frame homomorphisms $h : L \to M$.

On the class of sublocales we have the natural preorder

$$h \sqsubseteq k \quad \text{iff} \quad \text{there is an } l \text{ such that } h = lk.$$

The sublocales h, k are equivalent if $h \sqsubseteq k$ and $k \sqsubseteq h$. The ensuing partially ordered set will be denoted by

$$\mathcal{S}(L)$$

and the suprema resp. infima in $\mathcal{S}(L)$ will be denoted by $h \sqcup k$, $\bigsqcup h_i$ resp. $h \sqcap k$, $\bigsqcap h_i$.

5.1.2. There is the obvious invertible correspondence between $\mathcal{S}(X)$ and the set

$$\mathbf{C}L$$

of all congruences on L, associating with a sublocale $h : L \to M$ the congruence $E_h = \{(x, y) \mid h(x) = h(y)\}$, and with a congruence $E \subseteq L \times L$ the natural projection $p_E = (x \mapsto xE) : L \to L/E$. Since obviously

$$h \sqsubseteq k \quad \text{iff} \quad E_k \subseteq E_h,$$

and since any intersection of congruences is a congruence, we see that

$\mathcal{S}(L)$ is a complete lattice isomorphic to $\mathbf{C}L^{\mathrm{op}}$.

5.1.3. *Nuclei.* A *nucleus* on a frame L is a mapping $\nu : L \to L$ such that
(N1) $a \leqslant \nu(a)$,
(N2) $\nu\nu(a) = \nu(a)$, and
(N3) $\nu(a \wedge b) = \nu(a) \wedge \nu(b)$.
(Note that the Booleanization $\beta_L = (a \mapsto a^{**}) : L \to \mathfrak{B}$ from 2.9 is an example of nucleus.)

LEMMA. (1) *If* $a_i = \nu(a_i)$ *then* $\nu(\bigwedge a_i) = \bigwedge a_i$.
(2) *If* $b = \nu(b)$ *then for any* a, $\nu(a \to b) = a \to b$.

PROOF. (1) $\nu(\bigwedge a_i) \leqslant \bigwedge \nu(a_i) = \bigwedge a_i$.
(2) By (N3) and 2(H8), $\nu(a \to b) \wedge a \leqslant \nu(a \wedge (a \to b)) \leqslant \nu(b) = b$ and hence $\nu(a \to b) \leqslant a \to b$. □

The nucleus ν as a mapping $L \to L$ is typically not a frame homomorphism, but the restriction $L \to \nu[L]$ is. We have

PROPOSITION. *The subset* $\nu[L] \subseteq L$ *is a frame with the infima coinciding with those of* L *and the suprema given by* $\bigvee' a_i = \nu(\bigvee a_i)$; *the restriction* $\nu : L \to \nu[L]$ *is a frame homomorphism.*

PROOF. By the lemma above, $\nu[L]$ is closed under infima. We have $a_j \leqslant \nu(\bigvee a_i)$ for all j, and if $a_j \leqslant b \in \nu[L]$ for all j, $\bigvee a_i \leqslant b$ and $\bigvee' a_i = \nu(\bigvee a_i) \leqslant \nu(b) = b$. The mapping $\nu : L \to \nu[L]$ preserves finite meets by (N1) and (N3), and $\nu(\bigvee a_i) \leqslant \nu(\bigvee \nu(a_i)) = \bigvee' \nu(a_i) \leqslant \nu(\bigvee a_i)$. Finally, as $\nu : L \to \nu[L]$ is onto and as it preserves all joins and all finite meets, $\nu[L]$ satisfies the frame distributive law. □

Denote by

$$\mathcal{N}(L)$$

the system of all nuclei on L endowed with the natural order.

PROPOSITION. *The correspondences*

$$E \mapsto \nu_E, \qquad \nu_E(x) = \bigvee xE,$$
$$\nu \mapsto E_\nu, \qquad xE_\nu y \quad \textit{iff} \quad \nu(x) = \nu(y)$$

establish an isomorphism of the partially ordered sets $\mathbf{C}L$ *and* $\mathcal{N}(L)$. *Thus,* $\mathcal{N}(L)$ *is a complete lattice.*

PROOF. E_ν is a congruence by the previous proposition. Trivially, $\nu_E(x) \geqslant x$ and $(\bigvee xE)Ex$ and hence also $\nu_E\nu_E(x) = \nu_E(x)$. Obviously ν_E is monotone and we have

$$\begin{aligned}\nu_E(x) \wedge \nu_E(y) &= \bigvee\{u \wedge v \mid uEx, uEy\} \leqslant \bigvee\big\{z \mid zE(x \wedge y)\big\}\\ &= \nu_E(x \wedge y) \leqslant \nu_E(x) \wedge \nu_E(y).\end{aligned}$$

The correspondence $E \mapsto \nu_E$ is obviously monotone. Finally, $xE_{\nu_E}y$ iff $\bigvee xE = \bigvee yE$ iff xEy, and $\nu_{E_\nu}(x) = \bigvee xE_\nu = \bigvee\{u \mid \nu(u) = \nu(x)\} = \nu(x)$. □

5.1.4. *Sublocale sets.* A subset S of a frame L is said to be a *sublocale set* if

(S1) for each $A \subseteq S$, $\bigwedge A \in S$ (in particular, $1 = \bigwedge \emptyset \in S$), and

(S2) for each $a \in L$ and $b \in S$, $a \to b \in S$.

The system of all sublocale sets ordered by inclusion will be denoted by

$$\mathcal{S}'(L).$$

PROPOSITION. *The correspondences* $\nu \mapsto \nu[L]$, $S \mapsto \nu_S$, $\nu_S(a) = \bigwedge\{s \mid s \in S, a \leqslant s\}$ *constitute an isomorphism* $\mathcal{S}'(L) \cong \mathcal{N}(L)^{\mathrm{op}}$.

PROOF. By the lemma in 5.1.3, $\nu[L]$ is a sublocale set. On the other hand, obviously $a \leqslant \nu_S(a) = \nu_S\nu_S(a)$ and ν_S is monotone (and hence $\nu_S(a \wedge b) \leqslant \nu_S(a) \wedge \nu_S(b)$). Since $a \wedge b \leqslant \nu_S(a \wedge b)$ we have $a \leqslant b \to \nu_S(a \wedge b) \in S$ and hence $\nu_S(a) \leqslant b \to \nu_S(a \wedge b)$ and $b \wedge \nu_S(a) \leqslant \nu_S(a \wedge b)$. Repeating the procedure we obtain $\nu_S(a) \wedge \nu_S(b) \leqslant \nu_S(a \wedge b)$.

Obviously, $\nu_S[L] = \{x \mid x = \bigwedge\{s \in S \mid x \leqslant s\}\} = S$, and since $(s = \nu(s)\ \&\ a \leqslant s)$ iff $\nu(a) \leqslant \nu(s) = s$, we have $\nu_{\nu[L]}(a) = \nu(a)$. □

As a trivial consequence of (S2) we obtain that

$$\text{if } a \in S \text{ then for any } x,\ \nu_S(x) \to a = x \to a. \tag{5.1.4}$$

5.2. *Congruence frame.* The lattice structure is particularly transparent in $\mathcal{S}'(L)$: since obviously any intersection of sublocale sets is a sublocale set, the meets in $\mathcal{S}'(L)$ are simply the intersections; the supremum is given by the formula

$$\bigvee_{i \in J} S_i = \Big\{\bigwedge A \mid A \subseteq \bigcup_{i \in J} S_i\Big\}$$

(indeed, a sublocale set containing all S_i necessarily contains $\{\bigwedge A \mid A \subseteq \bigcup_{i\in J} S_i\}$; on the other hand, this set is obviously closed under meets, and for any $x \in L$, $x \to \bigwedge A = \bigwedge\{x \to a \mid a \in A\}$ by 2.(H1), and each $x \to a$ with $a \in A$ is in $\bigcup S_i$).

THEOREM. *$\mathcal{S}'(L)$ is a co-frame. Consequently, also $\mathcal{S}(L)$ is a co-frame, and $\mathcal{N}(L)$ and* $\mathbf{C}L$ *are frames.*

PROOF. Since the inclusion $(\bigcap_J A_i) \vee B \subseteq \bigcap_J (A_i \vee B)$ is trivial, we just have to prove the reverse one. We can assume that $J \neq \emptyset$. If $x \in \bigcap(A_i \vee B)$ we have, for each i, $a_i \in A_i$ and $b_i \in B$ such that $x = a_i \wedge b_i$. Consequently, if we set $b = \bigwedge b_i$, we have $x = (\bigwedge a_i) \wedge b \leqslant a_i \wedge b \leqslant a_i \wedge b_i = x$ so that $x = a_i \wedge b$ for all i. By 2.(H9), $x = (b \to a_i) \wedge b$ and by (H12) $b \to a_i = a$ does not depend on i. Thus, $x = a \wedge b$ with $b \in B$ and $a \in \bigcap A_i$ as, for each i, $a = b \to a_i \in A_i$. □

5.3. *Open and closed sublocales.* Let a be an element of L. We have the sublocales

$$\hat{a} = (x \mapsto a \wedge x) : L \to {\downarrow}a, \quad \text{with the associated congruence } \Delta_a,$$
$$\check{a} = (x \mapsto a \vee x) : L \to {\uparrow}a \quad \text{with the associated congruence } \nabla_a.$$

The former will be referred to as *open* sublocales and the latter as *closed* ones.

In the sublocale set representation we obtain the open resp. closed sublocale sets

$$\mathbf{o}(a) = \{a \to x \mid x \in L\} = \{x \mid x = a \to x\} \quad \text{resp. } \mathbf{c}(a) = {\uparrow}a$$

(indeed: $\bigvee\{y \mid y \wedge a = x \wedge a\} = \bigvee\{y \mid y \wedge a \leqslant x\} = \bigvee\{y \mid y = a \to x\} = a \to x$; note that $\mathbf{o}(a)$ is closed under meets by 2.(H1) and if u is general, $u \to (a \to x) = a \to (u \to x) \in \mathbf{o}(a)$ by (H7)).

PROPOSITION. (1) *∇_0 is the minimal congruence and ∇_1 the maximal one*; $\nabla_a \cap \nabla_b = \nabla_{a\wedge b}$ *and* $\bigvee_{i\in J} \nabla_{a_i} = \nabla_{\bigvee a_i}$. *Thus, $\nabla = (a \mapsto \nabla_a) : L \to \mathbf{C}L$ is a frame homomorphism.*

(2) ∇_a *and* Δ_a ($\check{a}$ *and* $\hat{a}$, $\mathbf{c}(a)$ *and* $\mathbf{o}(a)$) *are complements in* $\mathbf{C}L$ ($\mathcal{S}(L)$, $\mathcal{S}'(L)$).

PROOF. (1) Set $a = \bigvee a_i$. Obviously $\nabla_{a_i} \subseteq \nabla_a$ for all i; if $\nabla_{a_i} \subseteq E$ for all i, we have $a_i E 0$ for all i and hence $a = (\bigvee a_i) E 0$, and hence if $x \vee a = y \vee a$ we have $x E (x \vee a) = (y \vee a) E y$ so that $\nabla_a \subseteq E$. If $x \vee a = y \vee a$ and $x \vee b = y \vee b$ then $x \vee (a \wedge b) = (x \vee a) \wedge (x \vee b) = (y \vee a) \wedge (y \vee b) = y \vee (a \wedge b)$, and $x \vee (a \wedge b) = y \vee (a \wedge b)$ obviously implies $x \vee a = y \vee a$ and $x \vee b = y \vee b$.

(2) A direct proof is easy, but the fact is particularly evident in $\mathcal{S}'(L)$: If $y = a \to x \in {\uparrow}a$ we have $a \leqslant a \to x$, hence $a \leqslant x$ and $y = a \to x = 1$ by 2.(H4). Thus, $\mathbf{o}(a) \cap \mathbf{c}(a) = \{1\}$. On the other hand, for any $x \in L$, $x = (x \vee a) \wedge (a \to x)$ by (H3) so that $x \in \mathbf{c}(a) \vee \mathbf{o}(a)$. □

5.4. *Closure.* Obviously, $\overline{S} = {\uparrow}\bigwedge S$ is the least closed sublocale set containing S. It will be called the *closure* of S. In the language of sublocale maps, the closure of h (the smallest closed sublocale $\check{a}$ such that $h \sqsubseteq \check{a}$) is obtained as the $\check{c}$ with $c = \bigvee\{x \mid h(x) = 0\}$; here it is often denoted as $\mathcal{C}\ell(h)$; checking that this corresponds to the closure $\overline{S}$ above is straightforward and can be left to the reader.

PROPOSITION. *We have* $\overline{\{1\}} = \{1\}$, $\overline{\overline{S}} = \overline{S}$ *and* $\overline{S \vee T} = \overline{S} \vee \overline{T}$.

PROOF. The first two facts are trivial, and the third one very easy: Set $a = \bigwedge S$, $b = \bigwedge T$. Then $\overline{S} \vee \overline{T} = \uparrow a \vee \uparrow b = \{x \wedge y \mid x \geqslant a,\ y \geqslant b\} = \uparrow(a \wedge b) = \uparrow \bigwedge(S \vee T) = \overline{S \vee T}$. $\square$

5.5. *Dense sublocales.* A sublocale (more generally, a frame homomorphism) is said to be *dense* if $a \neq 0 \Rightarrow h(a) \neq 0$ (this agrees with the homonymous notion concerning spaces: $Y \subseteq X$ is dense iff $Y \cap U \neq \emptyset$ for each non-void open U in X). In the language of sublocale sets this translates to the condition that

$$S \text{ is dense} \quad \text{iff} \quad 0 \in S$$

(indeed, the congruence class of 0 should be $\{0\}$, and hence $\nu(0) = 0$ for the corresponding nucleus). Thus

$$S \text{ is dense} \quad \text{iff} \quad \overline{S} = L.$$

The Booleanization $\beta_L = (a \mapsto a^{**})$ (recall 2.7) is obviously dense. The following is in a strong contrast with the situation in classical spaces.

PROPOSITION. *Each frame has the smallest dense sublocale, namely the Booleanization.*

PROOF using sublocale sets. The sublocale set corresponding to the β_L is obviously $\{x^{**} \mid x \in L\}$. Let S contain 0. Then for each $y \in L$, $y \to 0 \in S$ and consequently $x^{**} = x^* \to 0 \in S$. $\square$

5.6. *Making a relation into a congruence.* Often we want to construct a new frame by identifying prescribed couples of elements. In other words, a relation $R \subseteq L \times L$ is given and we look for the least congruence containing R and for the corresponding quotient frame (sublocale) of L. This can be easily done as follows:

An element $s \in L$ is *saturated* (more precisely, R-*saturated*) if

$$\forall a, b, c,\ aRb \Rightarrow (a \wedge c \leqslant s \text{ iff } b \wedge s \leqslant s).$$

Obviously

any meet of saturated sets is saturated.

Consequently, we have the saturated

$$\nu(a) = \nu_R(a) = \bigwedge\{s \text{ saturated} \mid a \leqslant s\}.$$

5.6.1. PROPOSITION. ν_R *is a nucleus.*

PROOF. We have that

for any saturated s and any $x \in L$, $x \to s$ is saturated

(indeed, $a \wedge c \leqslant x \to s$ iff $a \wedge c \wedge x \leqslant s$ iff $b \wedge c \wedge x \leqslant s$ iff $b \wedge c \leqslant x \to s$).

Thus, the system of all saturated elements is a sublocale set and we can apply Proposition 5.1. □

Set

$$L/R = \nu_R[L] = \{x \mid \nu_R(x) = x\}.$$

5.6.2. THEOREM. *L/R is a frame and the restriction of ν_R to $L \to L/R$ is a sublocale. If aRb then $\nu_R(a) = \nu_R(b)$ and for every frame homomorphism $h : L \to M$ such that $aRb \Rightarrow h(a) = h(b)$ there is a frame homomorphism $\bar{h} : L/R \to M$ such that $\bar{h} \cdot \nu_R = h$. Moreover, $\bar{h}(a) = h(a)$ for all $a \in L/R$.*

PROOF. The first statement follows from 5.1.3 and 5.6.1. Further, if aRb then $b \leqslant \nu(a)$ since $a \leqslant \nu(a)$ and $\nu(a)$ is saturated. Hence $\nu(b) \leqslant \nu(a)$ and by symmetry $\nu(b) = \nu(a)$.

Let $h : L \to M$ be such that $aRb \Rightarrow h(a) = h(b)$. Set

$$\sigma(x) = \bigvee \{y \mid h(y) \leqslant h(x)\}.$$

Obviously

$$x \leqslant \sigma(x) \quad \text{and} \quad h\sigma(x) = h(x). \tag{$*$}$$

Let aRb and $a \wedge c \leqslant \sigma(x)$. Then $h(b \wedge c) = h(a \wedge c) \leqslant h\sigma(x) = h(x)$ and hence $b \wedge c \leqslant \sigma(x)$. Thus, σ is saturated. Combining this fact with ($*$) we obtain that $x \leqslant \nu(x) \leqslant \sigma(x)$ and hence

$$h(x) \leqslant h\nu(x) \leqslant h\sigma(x) = h(x)$$

so that $h\nu(x) = h(x)$ and the statement follows. □

5.6.3. NOTE. If L is join-generated by a subset M then the saturation condition is obviously equivalent to

$$\forall (a, b) \in R \text{ and } \forall c \in M \ (a \wedge c \leqslant s \text{ iff } b \wedge c \leqslant s).$$

5.7. *The image of a sublocale under a frame homomorphism.* Let $h : L \to M$ be a frame homomorphism. Let $\gamma : M \to N$ be a sublocale. The *image* of γ under h, denoted

$$h[\gamma],$$

is the projection homomorphism $(x \mapsto xE) : L \to L/E$ of the congruence

$$xEy \quad \equiv_{\mathrm{df}} \quad \gamma h(x) = \gamma h(y).$$

Note that if $f : X \to Y$ is a continuous map and $j : A \subseteq X$ a subspace embedding, $\Omega(f)[\Omega(j)]$ is equivalent to the $\Omega(k)$ where k is the embedding $f[A] \subseteq Y$. Indeed, we have $\Omega(j)\Omega(f)(U) = A \cap f^{-1}(U)$, and $A \cap f^{-1}(U) = A \cap f^{-1}(V)$ holds iff $f[A] \cap U = f[A] \cap V$.

5.8. *Aside: More about congruence frames; strange epimorphisms.* If $a \nleq b$ then $\nabla_a \ni (a, 0) \notin \nabla_b$. Hence, the homomorphism

$$\nabla = \nabla^L : L \to \mathbf{C}L$$

from 5.3 is one–one, that is, a frame embedding.

PROPOSITION. *For every $C \in \mathbf{C}L$ we have*

$$C = \bigvee \{\nabla_a \cap \Delta_b \mid aCb\}.$$

PROOF. If aCb and $(x, y) \in \nabla_a \cap \Delta_b$ we have

$$\begin{aligned} x &= x \wedge (x \vee a) = x \wedge (y \vee a) = (x \wedge y) \vee (x \wedge a) C (x \wedge y) \vee (x \wedge b) \\ &= (x \wedge y) \vee (y \wedge b) = (x \vee b) \wedge y C (x \vee a) \wedge y = (y \vee a) \wedge y = y \end{aligned}$$

so that $\nabla_a \cap \Delta_b \subseteq C$. On the other hand, let $E \supseteq \nabla_a \cap \Delta_b$ for all $(a, b) \in C$. Let $(a, b) \in C$. We have $(b, a \vee b) \in \nabla_a \cap \Delta_b \subseteq E$ and since also $(b, a) \in C$, we have $(a, a \vee b) \in E$ as well. Thus, $aE(a \vee b)Eb$, and $C \subseteq E$. $\square$

COROLLARY. *The embedding $\nabla^L : L \to \mathbf{C}L$ is an epimorphism.*

(Indeed, if $f, g : \mathbf{C}L \to M$ coincide in the ∇_a's, they coincide, by Proposition in 5.3 and 2.5(3), also in the Δ_a's, and hence by the proposition above, in every $C \in \mathbf{C}L$.)

The embeddings can be combined to

$$L \xrightarrow{\nabla^L} \mathbf{C}L \xrightarrow{\nabla^{\mathbf{C}L}} \mathbf{C}^2 L \to \cdots \to \mathbf{C}^n L,$$

and also the transfinite step is easy. Thus, one has epimorphisms $\nabla^\alpha : L \to \mathbf{C}^\alpha L$ for all ordinals α. For some frames L the $\mathbf{C}^\alpha L$ never stop growing (in fact, in all the known cases it is either this, or the growth stops before the fourth step; whether it can be otherwise is an open problem). Thus, for such a fixed frame L one has epimorphisms $\varepsilon : L \to M$ with arbitrarily large M.

6. Behaviour of the category of frames; coproducts of frames (products of locales)

6.1. *Coproducts of meet-semilattices.* First observe that in the category **SLat** (recall 3.4) the coproducts are obtained as follows: Set

$$\prod_{i\in J}{}' A_i = \left\{ (a_i)_{i\in J} \in \prod_{i\in J} A_i \mid a_i = 1 \text{ for all but finitely many } i \right\} \cup \left\{ (0)_{i\in J} \right\}$$

and define

$$\gamma_j : A_j \to \prod_{i\in J}{}' A_i \quad \text{by setting } \bigl(\gamma_j(a)\bigr)_i = \begin{cases} a & \text{for } i = j, \\ 1 & \text{otherwise.} \end{cases}$$

Obviously, if $h_j : A_j \to B$ are morphisms in **SLat** we have a uniquely defined $h : \prod' A_i \to B$ such that $h\gamma_j = h_j$, namely that given by $h((a_i)_{i\in J}) = \bigwedge_{i\in J} h_i(a_i)$ (the meet is finite, all but finitely many $h_i(a_i)$ being 1).

6.2. *Coproducts of frames.* Recall the $\eta_A : A \to \mathfrak{D}A$ from 3.4. Let L_i, $i \in J$, be frames. View them, for a moment, as objects of **SLat**, and consider $A = \prod'_{i\in J} L_i$. On the frame $\mathfrak{D}(\prod'_{i\in J} L_i)$ take the relation

$$R = \left\{ \left(\eta\gamma_j\Bigl(\bigvee_{m\in M} a_m \Bigr), \bigvee_{m\in M} \eta\gamma_j(a_m) \right) \mid j \in J,\ M \text{ any set, } a_m \in L_j \right\}$$

and set

$$\bigoplus_{i\in J} L_i = \mathfrak{D}\Bigl(\prod_{i\in J}{}' L_i \Bigr) \Big/ R.$$

Let $\nu : \mathfrak{D}(\prod'_{i\in J} L_i) \to \bigoplus_{i\in J} L_i$ be the nucleus homomorphism from 5.6.

OBSERVATION. *The $\iota_j = \nu\eta\gamma_j$ are frame homomorphisms.*

(Indeed, 0, 1 and $\wedge$ are preserved trivially and $\eta\gamma_j(\bigvee_m a_m) R \bigvee_m \eta\gamma_j(a_m)$ and hence $\iota_j(\bigvee a_m) = \bigvee \iota_j(a_m)$.)

6.2.1. PROPOSITION. *The system $(\gamma_j : L_j \to \bigoplus_{i\in J} L_i)_{j\in J}$ is a coproduct in* **Frm**.

PROOF. Consider the diagram

$$\begin{array}{ccccccc}
L_j & \xrightarrow{\gamma_j} & \prod'_{i\in J} L_i & \xrightarrow{\eta} & \mathfrak{D}(\prod'_{i\in J} L_i) & \xrightarrow{\nu} & \bigoplus_{i\in J} L_i \\
\downarrow{\scriptstyle h_j} & & \downarrow{\scriptstyle h'} & & \downarrow{\scriptstyle h''} & & \downarrow{\scriptstyle h} \\
M & = & M & = & M & = & M
\end{array}$$

where h_j are some frame homomorphisms, h' is the **SLat**-homomorphism obtained by 6.1, and h'' is the frame homomorphism from 3.4. We have

$$\begin{aligned} h''\Big(\bigvee_m \eta\gamma_j(a_m)\Big) &= \bigvee\big\{h'\big((b_i)_{i\in J}\big) \mid (b_i)_{i\in J} \leqslant \gamma_j(a_m) \text{ for some } m\in M\big\} \\ &= \bigvee_m h'\gamma_j(a_m) = \bigvee_m h_j(a_m) = h_j\Big(\bigvee_m a_m\Big) \\ &= h'\gamma_j\Big(\bigvee_m a_m\Big) = h''\eta\gamma_j\Big(\bigvee_m a_m\Big) \end{aligned}$$

and hence by 5.6 there is a frame homomorphism h such that $h\cdot\nu = h''$. Thus,

$$h\cdot\iota_j = h\nu\eta\gamma_j = h''\eta\gamma_j = h'\gamma_j = h_j.$$

The uniqueness follows from the obvious fact that all the elements of $\mathfrak{D}(\prod'_{i\in J} L_i)$ are joins of finite meets of the $\eta\gamma_j(a)$ and hence all the elements of $\bigoplus_{i\in J} L_i$ are joins of finite meets of the $\iota_j(a)$ $(j\in J,\ a\in L_j)$. □

NOTATION. For finite systems we write $L\oplus M$, $L_1\oplus L_2\oplus L_3$, etc.

6.3. *More about the structure of coproduct.* Recalling that the join $\bigvee_m U_m$ in $\mathfrak{D}A$ is equal to the union $\bigcup_m U_m$ if $M\neq\emptyset$, and to the set $\{0\}$ if the index set M is void, we see that in particular

$$\big(\downarrow\gamma_j(0), \big\{(0)_{i\in J}\big\}\big)\in R \quad \text{for all } j.$$

Set

$$\mathsf{O} = \Big\{(a_i)_{i\in j}\in \prod{}' L_i \mid \exists i, a_i = 0\Big\}.$$

Since for $M\neq\emptyset$ obviously $\bigcup_m\downarrow\gamma_j(a_m) = \downarrow\gamma_j(\bigcup_m a_m)$ and since $\mathfrak{D}(\prod' L_i)$ is generated by the $\downarrow(b_i)_i$, we easily infer (recall 5.6) that

$U\in\mathfrak{D}(\prod' L_j)$ is saturated iff
(1) $\mathsf{O}\subseteq U$, *and*
(2) *for $M\neq\emptyset$, whenever $x_{im} = x_i$ for $i\neq j$, $x_j = \bigvee_{m\in M} x_{jm}$ and $(x_{im})_{i\in J}\in U$ for all m then $(x_i)_{i\in J}\in U$.*

6.3.1. LEMMA. *For any $(a_i)_i\in\prod' L_i$, the set*

$$\oplus_{i\in J} a_i = \downarrow(a_i)_{i\in J}\cup\mathsf{O}$$

is saturated.

PROOF. Let $(x_{im})_i$ and x_j be as in the condition above. If $x_i = 0$ for some i then $(x_i)_i \in \oplus_i a_i$. Else all $x_{im} \neq 0$ for $i \neq j$ and $\bigvee x_{jm} \neq 0$. Hence $x_{jn} \neq 0$ for some n, $(x_{in})_i$ is not in O and hence $x_i \leqslant a_i$ for all $i \neq j$; but then also all $x_{jm} \leqslant a_j$ and $x_j = \bigvee x_{jm} \leqslant a_j$. $\square$

COROLLARY. *If $\oplus_i a_i \leqslant \oplus_i b_i$ and $a_i \neq 0$ for all i then $a_i \leqslant b_i$ for all i.*

NOTATION. For finite index sets we write

$$a \oplus b,\ a_1 \oplus a_2 \oplus a_3,\ \text{etc.}$$

6.3.2. Note that for the $(a_i)_i \in \prod' L_i$ we have

$$\bigwedge_{i \in J} \iota_i(a_i) = \oplus_i a_i.$$

Thus the set of the elements of the form $\oplus_i a_i$ is a (join-)basis of $\bigoplus L_i$ and we have, for any $u \in \bigoplus L_i$,

$$u = \bigvee \{\oplus_i a_i \mid (a_i)_i \in u\} = \bigvee \{\oplus_i a_i \mid \oplus_i a_i \leqslant u\}.$$

REMARK. The first construction of coproducts in frames was presented in [43].

6.4. The category **Frm** is obviously complete: The products are the Cartesian products $\prod_{i \in J} L_i$ (with the standard projections $\prod L_i \to L_j$), the order (and hence the meets and joins) defined coordinatewise; if $h_1, h_2 : L \to M$ are frame homomorphisms then $K = \{x \mid h_1(x) = h_2(x)\}$ is obviously a subframe of L and the embedding $j : K \subseteq L$ is the equalizer of h_1, h_2.

Coequalizers in **Frm** are also easy: Consider the relation $R = \{(h_1(x), h_2(x)) \mid x \in L\}$ on M; then the $\kappa : M \to M/R$ from 5.6 is the coequalizer of h_1, h_2 (for special frames we will have, later, a more explicit formula). In 6.2 we have presented a construction of coproducts (the only tricky one among the basic limits and colimits). Thus, **Frm** is cocomplete.

The *monomorphisms* in **Frm** are easily seen to be exactly the one–one homomorphisms. The structure of *epimorphisms* in **Frm**, on the contrary, is not very transparent (recall 5.8). On the other hand, the *extremal epimorphisms* are easy, being exactly the onto homomorphisms (sublocale maps). Thus we have (extremal epi, mono)-decompositions

$$(h : L \to M) = \left(L \xrightarrow{h'} h[L] \xrightarrow{\subseteq} M\right)$$

with $h'(x) = h(x)$.

6.5. We will conclude this section with a technical statement which will be used later on.

PROPOSITION. *Let L_i, $i = 1, 2$, be frames and $a_i \in L_i$. Then we have, for the open sublocales,*

$$\downarrow a_1 \oplus \downarrow a_2 = \downarrow (a_1 \oplus a_2).$$

More precisely, if $\iota_i : L_i \to L_1 \oplus L_2$ are the coproduct injections then the

$$\iota_i' = \big(x \mapsto \iota_i(x) \wedge (a_1 \oplus a_2)\big) : \downarrow a_i \to \downarrow (a_1 \oplus a_2)$$

constitute a coproduct in **Frm**.

PROOF. Let $h_i : \downarrow a_i \to M$ be frame homomorphisms. Consider the $g : L_1 \oplus L_2 \to M$ such that $g\iota_i = h_i \hat{a}_i$. We have $g(x_1 \oplus x_2) = g(\iota_1(x_1) \wedge \iota_2(x_2)) = h_1(x_1 \wedge a_1) \wedge h_2(x_2 \wedge a_2)$ and hence if $(x_1 \oplus x_2) \wedge (a_1 \oplus a_2) = (y_1 \oplus y_2) \wedge (a_1 \oplus a_2)$ then $g(x_1 \oplus x_2) = g(y_1 \oplus y_2)$. Thus there is a frame homomorphism $h : \downarrow (a_1 \oplus a_2) \to M$ such that $h \cdot \widehat{a_1 \oplus a_2} = g$. For $x \in \downarrow a_i$ we have $h(\iota_i'(x)) = h(\iota_i(x) \wedge (a_1 \oplus a_2)) = g(\iota_i(x)) = h_i(x)$. The unicity of such an h is obvious. $\square$

7. Separation axioms. I: Not quite like the classical case

Recall the standard separation axioms of classical topology ($T_0, T_1, \dots, T_4$) and let us dismiss, right away T_0, which cannot have any point-free counterpart: two points violating T_0 cannot be told apart in the language of open sets.

In this section we will discuss properties loosely related to T_1 and T_2 (and one more). Here the classical situation is so heavily dependent on points that we cannot expect exact counterparts; the next section will be concerned with regularity, complete regularity and normality which, on the contrary, are indeed extensions of the classical homonymous properties.

7.1. *Subfitness.* A frame L is said to be *subfit* ([64], *conjunctive* in [142]) if

$$a \nleq b \quad \Rightarrow \quad \exists c,\ a \vee c = 1 \neq b \vee c. \tag{Sfit}$$

OBSERVATION. *Let X be T_1. Then $\Omega(X)$ is subfit.*

(Indeed, if $a \nleq b$ choose an $x \in a \backslash b$ and set $c = X \backslash \{x\}$.)

On the other hand, if $\Omega(X)$ is subfit, X is not necessarily T_1. But the relation of the two requirements is close: a T_D space is T_1 iff $\Omega(X)$ is subfit.

7.1.1. PROPOSITION. *The following statements about a frame L are equivalent*:

(1) *L is subfit.*

(2) *If $E1 = \{1\}$ for a congruence E on L then E is trivial. In other words, if a frame homomorphism $h : L \to M$ is onto and $h^{-1}(\{1\}) = \{1\}$ then h is an isomorphism.*

(3) *If $h : L \to M$ is a sublocale not equivalent to* id *then there is a closed $\check{x} \neq \hat{0}$ such that $h \sqcap \check{x} = \hat{0}$.*
(4) *Each open sublocale is an intersection of closed ones.*

PROOF: will be done in the language of sublocale sets; in this the statements (2), (3) and (4) are translated to

(2′) For a sublocale set $S \subseteq L$, the subset $S \backslash \{1\}$ is cofinal in $L \backslash \{1\}$ (in other words, $\nu_S^{-1}(\{1\}) = \{1\}$) iff $S = L$,
(3′) If $S \neq L$ for a sublocale set $S \subseteq L$ then there is a closed $\mathbf{c}(x) \neq \{1\}$ such that $S \cap \mathbf{c}(x) = \{1\}$,

(ν_S is the nucleus associated with S) and

(4′) Each open sublocale set is an intersection of closed sublocale sets. More exactly, $\mathbf{o}(a) = \bigcap\{\mathbf{c}(x) \mid x \vee a = 1\}$.

(1) $\Rightarrow$ (2′): Let $b \in L$ and $a = \nu_S(b)$. If $a \vee c = 1$ we have $\nu(b \vee c) \geqslant a \vee c = 1$ and hence $b \vee c = 1$. Thus, $a \leqslant b$, that is, $b \in S$.

(2′) $\Leftrightarrow$ (3′): (3′) is just an immediate reformulation of (2′).

(3′) $\Rightarrow$ (4′): For $a \in L$ set $S = \mathbf{c}(a) \vee \bigvee\{\mathbf{c}(x) \mid x \vee a = 1\}$. If $\mathbf{c}(y) \cap S = \{1\}$ we have $\mathbf{c}(y) \cap \mathbf{c}(a) = \{1\}$ and $\{1\} = \mathbf{c}(y) \cap \bigvee\{\mathbf{c}(x) \mid x \vee a = 1\}$. By the former and proposition in 5.3, $\mathbf{c}(y) \subseteq \mathbf{o}(a)$, and $y \vee a = 1$, and consequently, by the latter, $\{1\} = \mathbf{c}(y)$. Thus, $S = L$ and (see 5.3 again) $\mathbf{o}(a) = \mathbf{o}(a) \cap S = \mathbf{o}(a) \cap \bigvee\{\mathbf{c}(x) \mid x \vee a = 1\}$ so that $\mathbf{o}(a) \subseteq \bigvee\{\mathbf{c}(x) \mid x \vee a = 1\}$ ($\subseteq \mathbf{o}(a)$ by 5.3).

(4′) $\Rightarrow$ (1): If $a \nleqslant b$ we have $\mathbf{c}(b) \nsubseteq \mathbf{c}(a)$ (as $b \in \mathbf{c}(b) \backslash \mathbf{c}(a)$) and hence, by 5.3, $\mathbf{o}(a) \nsubseteq \mathbf{o}(b)$. Thus, by (4′) there is a c such that $c \vee a = 1$ and $\mathbf{c}(c) \nsubseteq \mathbf{o}(b)$, that is, $c \vee b \neq 1$. □

7.2. *Fitness.* A frame L is said to be *fit* [64] if

$$a \nleqslant b \quad \Rightarrow \quad \exists c,\ a \vee c = 1 \text{ and } c \to b \neq b. \tag{Fit}$$

7.2.1. PROPOSITION. *Fitness is hereditary. That is, if L is fit and $h : L \to M$ a sublocale then M is fit.*

PROOF. Let $h(a) \nleqslant h(b)$; set $b_0 = \bigvee\{x \mid h(x) \leqslant h(b)\}$. Then we have $h(a) \nleqslant h(b_0)$ (and hence $a \nleqslant b_0$) and there is a $c \in L$ such that $a \vee c = 1$ and $c \to b_0 \neq b_0$. Then also $h(c) \to h(b) \neq h(b)$ since else $h(c \to b_0) \leqslant h(c) \to h(b_0) = h(b_0) = h(b)$ and $b_0 \leqslant c \to b_0 \leqslant b_0$. □

7.2.2. PROPOSITION. *The following statements about a frame L are equivalent*:
(1) *L is fit.*
(2) *If E_1, E_2 are congruences on L such that $E_1 1 = E_2 1$ then $E_1 = E_2$.*
(3) *Each sublocale is an intersection of open sublocales.*

PROOF. Again, we will work in the language of sublocale sets. First, for a sublocale set S in a general L define

$$S' = {\downarrow}\big(S \backslash \{1\}\big) \quad (= \{x \in L \mid \nu_S(x) \neq 1\}).$$

Now we can translate the statements (2) and (3) to

(2′) For any two sublocale sets $S, T \subseteq L$, $S' = T' \Rightarrow S = T$. In other words, $\nu_S^{-1}(1) = \nu_T^{-1}(1) \Rightarrow S = T$,

and

(3′) Each sublocale set is an intersection of open sublocale sets.

(1) $\Rightarrow$ (2′): Let $T' = S'$ and let $b \in T$, $b \neq 1$. Set $a = \nu_S(b)$. We will prove that $a \vee c = 1$ implies $c \to b = b$ so that, by (Fit), $a \leqslant b$ and hence $b = \nu_S(b) \in S$.

If $a \vee c = 1$ and $b \vee c \leqslant a_1 \in S$ then $a_1 \geqslant a \vee c = 1$ so that $b \vee c \notin S' = T'$. By 2.(H13), however, $(b \vee c) \wedge (c \to b) \leqslant b$ and hence $b \vee c \leqslant (c \to b) \to b \in T$ so that $(c \to b) \to b = 1$ and $c \to b = b$ by (H4) and (H5).

(2′) $\Rightarrow$ (3′): If $a \in S$ and $\nu(x) = 1$ then $x \to a = a$ by 5.1.4 and 2.(H3). Thus, $a \in \mathbf{o}(x)$ and we have the inclusion $\subseteq$. If $a \in T = \bigcap\{\mathbf{o}(x) \mid \nu(x) = 1\}$ we have $x \to a = a$ whenever $\nu_S(x) = 1$. Thus, if $\nu_S(a) = 1$ we have $a = a \to a = 1$. Hence $T \setminus \{1\} \subseteq S'$ and we conclude that $T = S$.

(3″) $\Rightarrow$ (1): In particular, $\mathbf{c}(a) = {\uparrow}a = \bigcap\{\mathbf{o}(c) \mid a \vee c = 1\}$ (since obviously $\Delta_c \subseteq \nabla_a$ iff $a \vee c = 1$) and hence, if $c \to b = b$ for all c such that $a \vee c = 1$ then $a \leqslant b$. □

7.2.3. PROPOSITION. *A frame L is fit iff each of its sublocales is subfit.*

PROOF. Comparing 7.1.1(2) and 7.2.2(2) we see that fit implies subfit. Hence, by 7.2.1, the condition is necessary.

Now let each sublocale of L be subfit and let E_1, E_2 be congruences on L such that $E_1 1 = E_2 1$. Set $M = L/(E_1 \cap E_2)$ and write $[x] \in M$ for the congruence class of $x \in L$. Define congruences E_i' on M by setting $[x]E_i'[y]$ if $x E_i y$. Then $[x]E_i'[1]$ iff $[x] = [1]$ and by 7.1.1 both E_i' are trivial. Hence $x E_i y$ iff $[x] = [y]$ and $E_1 = E_2$. □

7.3. *A Hausdorff type property.* Obviously, a topological space X is Hausdorff if and only if the diagonal $\{(x, x) \mid x \in X\}$ is closed in $X \times X$. One of the variants of the Hausdorff-type axioms for frames imitates this statement.

We say that a frame is *Isbell–Hausdorff* (*strongly Hausdorff* in [64]) if the codiagonal

$$\nabla : L \oplus L \to L$$

(that is, the homomorphism ∇ such that $\nabla\iota_i = \mathrm{id}$ for both coproduct injections) is a closed sublocale.

NOTE. $\Omega(X \times X)$ is not generally isomorphic to $\Omega(X) \oplus \Omega(X)$. Hence, the Isbell–Hausdorff property is only an imitation of the classical Hausdorff axiom, not an extension. In fact, if X is T_0 and $\Omega(X)$ is Isbell–Hausdorff then X is Hausdorff, but the reverse need not be true – see [72].

7.3.1. For a frame L set

$$d_L = \bigvee\{x \oplus y \mid x \wedge y = 0\} \in L \oplus L.$$

We have

PROPOSITION. *L is Isbell–Hausdorff iff for any $a, b \in L$,*

$$a \oplus b \leqslant \big((a \wedge b) \oplus (a \wedge b)\big) \vee d_L.$$

PROOF. Since obviously $\nabla(x \oplus y) = x \wedge y$ and since the $x \oplus y$ generate $L \oplus L$, $\check{d}_L : L \oplus L \to {\uparrow} d_L$ is the closure of ∇. Thus, the Isbell–Hausdorff condition amounts to the existence of a frame homomorphism

$$\alpha : L \to {\uparrow} d_L \quad \text{such that} \quad \alpha \cdot \nabla = \check{d}_L. \tag{7.3.1}$$

Let such an α exist. Then $(a \oplus b) \vee d_L = \alpha(\nabla(a \oplus b)) = \alpha(\nabla((a \wedge b) \oplus (a \wedge b))) = ((a \wedge b) \oplus (a \wedge b)) \vee d_L$.

Now let the condition be satisfied. Set

$$\alpha(x) = (x \oplus x) \vee d_L.$$

As $x_i \oplus x_j \leqslant ((x_i \wedge x_j) \oplus (x_i \wedge x_j)) \vee d_L$ we have $(x_i \oplus x_j) \vee d_L \leqslant (x_i \oplus x_i) \vee d_L$ and hence

$$\begin{aligned} \alpha\Big(\bigvee x_i\Big) &= \Big(\bigvee x_i \oplus \bigvee x_i\Big) \vee d_L = \bigvee_{i,j} (x_i \oplus x_j) \vee d_L \\ &= \bigvee (x_i \oplus x_i) \vee d_L = \bigvee \alpha(x_i). \end{aligned}$$

Trivially, α preserves finite meets and hence $\alpha : L \to {\uparrow} d_L$ is a frame homomorphism. We have $\check{d}_L(a \oplus b) = (a \oplus b) \vee d_L = ((a \wedge b) \oplus (a \wedge b)) \vee d_L = \alpha(\nabla(a \oplus b))$. As the $a \oplus b$ generate $L \oplus L$, $\check{d}_L = \alpha \cdot \nabla$. □

7.3.2. One of the important properties of Hausdorff spaces is that if $f_1, f_2 : X \to Y$ are continuous maps and Y is Hausdorff then the (equalizer) set $\{x \mid f_1(x) = f_2(x)\}$ is closed. The following is a point-free counterpart of this statement.

PROPOSITION. *Let L be Isbell–Hausdorff and $h_1, h_2 : L \to M$ frame homomorphisms. Set $c = \bigvee\{h_1(x) \wedge h_2(y) \mid x \wedge y = 0\}$. Then*

$$\check{c} : M \to {\uparrow} c$$

is the coequalizer of h_1 and h_2.

PROOF. Obviously, $c = \nabla((h_1 \oplus h_2)(d_L))$. Consider the homomorphism α from (7.3.1). We have

$$(a \oplus 1) \vee d_L = \alpha \nabla(a \oplus 1) = \alpha(a) = \alpha \nabla(1 \oplus a) = (1 \oplus a) \vee d_L$$

and hence

$$\begin{aligned} \check{c} h_1(a) = h_1(a) \vee c &= \nabla(h_1 \oplus h_2)\big((a \oplus 1) \vee d_L\big) \\ &= \nabla(h_1 \oplus h_2)\big((1 \oplus a) \vee d_L\big) = h_2(a) \vee c = \check{c} h_2(a). \end{aligned}$$

On the other hand, if $\varphi h_1 = \varphi h_2 = h$ for a $\varphi : M \to K$, we have

$$\varphi(c) = \bigvee \{h(x \wedge y) \mid x \wedge y = 0\} = 0$$

and hence we can define $\overline{\varphi} : {\uparrow} c \to K$ by $\overline{\varphi}(x) = \varphi(x)$ to obtain $\bar{\varphi} \cdot \check{c} = \varphi$. □

7.3.3. COROLLARY. *If $g : M \to K$ is dense, if L is Isbell–Hausdorff and if $gh_1 = gh_2$ for $g_1, g_2 : L \to M$ then $h_1 = h_2$.*

(Indeed, for the c from 7.3.2 we have $g(c) = 0$ and hence $c = 0$. Thus, $h_1(x) = h_1(x) \vee c = h_2(x) \vee c = h_2(x)$.)

8. Separation axioms. II: Quite like the classical case

8.1. *The relation "rather below".* Let x, y be elements of a frame L. We say that x is *rather below* y (in L) and write

$$x \prec y$$

if $x^* \vee y = 1$.

8.1.1. LEMMA. (1) *$a \prec b \Rightarrow a \leqslant b$, and for any a, $0 \prec a \prec 1$.*
(2) *$x \leqslant a \prec b \leqslant y \Rightarrow x \prec y$.*
(3) *If $a \prec b$ then $b^* \prec a^*$.*
(4) *If $a \prec b$ then $a^{**} \prec b$.*
(5) *If $a_i \prec b_i$ for $i = 1, 2$ then $a_1 \vee a_2 \prec b_1 \vee b_2$ and $a_1 \wedge a_2 \prec b_1 \wedge b_2$.*

PROOF. Recall 2.3. The statement (1) is trivial, (2) follows from the fact that $x \leqslant y \Rightarrow y^* \leqslant x^*$, (3) from $b \leqslant b^{**}$ and (4) from $a^{***} = a^*$.

(5) $(a_1 \vee a_2)^* \vee (b_1 \vee b_2) = (a_1^* \wedge a_2^*) \vee (b_1 \vee b_2) \geqslant (a_1^* \vee b_1) \wedge (a_2^* \vee b_2)$ and $(a_1 \wedge a_2)^* \vee (b_1 \wedge b_2) \geqslant (a_1^* \vee a_2^*) \vee (b_1 \wedge b_2) \geqslant (a_1^* \vee b_1) \wedge (a_2^* \vee b_2)$. □

8.2. *Regularity.* A frame L is said to be *regular* if

$$\text{for each } a \in L,\ a = \bigvee \{b \mid b \prec a\}. \tag{8.2.1}$$

Here, the notion is an extension of the classical one. We have

8.2.1. PROPOSITION. *A space X is regular iff the frame $\Omega(X)$ is regular.*

PROOF. The pseudocomplement in $\Omega(X)$ is obviously given by the formula $U^* = X \setminus \overline{U}$ so that $V \prec U$ iff $\overline{V} \subseteq U$. The classical regularity is equivalent to the condition that for each open U, $U = \bigcup \{V \mid V \text{ open}, \overline{V} \subseteq U\}$ (a well-known fact and a very easy exercise). □

The formula (8.2.1) is easily translated to

$$a \nleqslant b \quad \Rightarrow \quad \exists c,\ a \vee c = 1 \text{ and } c \to 0 = c^* \nleqslant b. \tag{8.2.2}$$

(Indeed, (8.2.1) says that if $a \nleqslant b$ there is an x such that $x^* \vee a = 1$ and $x \nleqslant b$. Set $c = x^*$ and use the fact that $c \leqslant c^{**}$ and $x \leqslant x^{**}$.)

Since $c \to 0 \leqslant c \to b$ we immediately obtain (recall 7.2)

8.2.2. COROLLARY. *Each regular frame is fit.*

From 7.2.2 we further infer

8.2.3. COROLLARY. *Let L be regular and let E_1, E_2 be congruences on L such that $E_1 1 = E_2 1$. Then $E_1 = E_2$.*

8.2.4. PROPOSITION. *Each regular frame is Isbell–Hausdorff.*

PROOF. Use 7.3.1. Let $x \prec a$ and $y \prec b$. Then

$$\begin{aligned} x \oplus y &= \big(x \wedge (y^* \vee b)\big) \oplus \big(y \wedge (x^* \vee a)\big) \\ &= \big((x \wedge b) \vee (y^* \wedge x)\big) \oplus \big((a \wedge y) \vee (x^* \wedge y)\big) \\ &\leqslant \big((a \wedge b) \oplus (a \wedge b)\big) \vee (x \oplus x^*) \vee (y^* \oplus y) \\ &\leqslant \big((a \wedge b) \oplus (a \wedge b)\big) \vee d_L. \end{aligned}$$

Now since L is regular we have $a \oplus b = \bigvee\{x \mid x \prec a\} \oplus \bigvee\{y \mid y \prec b\} = \bigvee\{x \oplus y \mid x \prec a, y \prec b\} \leqslant ((a \wedge b) \oplus (a \wedge b)) \vee d_L$. $\square$

8.3. *The category* **RegFrm**. First observe that for any frame homomorphism $h : L \to M$,

$$x \prec y \quad \Rightarrow \quad h(x) \prec h(y). \tag{8.3.1}$$

(Use the fact that $h(x^*) \leqslant h(x)^*$.)

For a frame L set

$$L' = \Big\{a \mid a = \bigvee\{x \mid x \prec a\}\Big\}.$$

OBSERVATION. *L' is a subframe of L.*

PROOF. Obviously $1 \in L'$ and L' is closed under joins. Further, if $a, b \in L'$ we have by 8.1.1.5

$$a \wedge b = \bigvee\{x \wedge y \mid x \prec a, y \prec b\} \leqslant \bigvee\{z \mid z \prec a \wedge b\} \leqslant a \wedge b. \qquad \square$$

Now define $L^{(\alpha)}$ for ordinals α by setting

$$L^{(0)} = L, \quad L^{(\alpha+1)} = \left(L^{(\alpha)}\right)' \quad \text{and} \quad L^{(\alpha)} = \bigcap_{\beta<\alpha} L^{(\beta)} \quad \text{for limit } \alpha$$

and denote by $L^{(\infty)}$ the $L^{(\alpha)}$ for which $L^{(\alpha)} = L^{(\alpha+1)}$.

Denote by

RegFrm

the full subcategory of **Frm** generated by regular frames. We have

8.3.1. PROPOSITION. **RegFrm** *is* (*mono*)*coreflective in* **Frm**, *with the coreflection given by the embeddings* $\gamma_L : L^{(\infty)} \subseteq L$.

PROOF. Let M be regular and L general and let $h : M \to L$ be a frame homomorphism. By (8.3.1), for any $a \in M$, $h(a) = \bigvee\{h(x) \mid x \prec a\} \leqslant \bigvee\{y \mid y \prec h(a)\} \leqslant h(a)$ and hence $h(a) \in L'$. Thus, we have an $h' : M \to L'$ with $h = (L' \subseteq L) \cdot h'$, and consequently, by induction we have that $h[M] \subseteq L^{(\infty)}$, and we can define $\bar{h} : M \to L^{(\infty)}$ by $\bar{h}(a) = h(a)$ to obtain $\gamma_L \bar{h} = h$. Obviously, $L^{(\infty)}$ is regular. □

(Note that the "rather below" in L' is generally not the same as the "rather below" in L.)

COROLLARY. **RegFrm** *is complete and cocomplete, and the colimits in* **RegFrm** *coincide with those in* **Frm**.

NOTE. In particular, sublocales and coproducts of regular locales, taken in **Frm** are regular. A direct proof of the former statement is a trivial consequence of 8.3.1; also the latter can be easily proved directly, which can be left to the reader as an exercise. (Hint: realize, first, that $(a \oplus 1)^* = a^* \oplus 1$.)

8.3.2. From 8.2.3, 8.2.4, 7.2.2 and 7.3 we easily obtain

COROLLARY. (1) *In* **RegFrm**, *dense frame homomorphisms are monomorphisms.*

(2) *Each codense* $h : L \to M$ *in* **RegFrm** (*that is, homomorphism* h *such that* $h(x) = 1 \Rightarrow x = 1$) *is one–one.*

(3) *The coequalizer of* $h_1, h_2 : L \to M$ *in* **RegFrm** *is the closed* $\check{c}$ *with*

$$c = \bigvee\left\{h_1(x) \wedge h_2(y) \mid x \wedge y = 0\right\}.$$

NOTE. In fact, monomorphisms in **RegFrm** are *exactly* the dense homomorphisms.

8.4. *Complete regularity.* Let x, y be elements of a frame L. We say that x is *completely below* y (in L) and write

$$x \prec\!\!\prec y$$

if there are $x_r \in L$ for r dyadic rational in the interval $\langle 0, 1\rangle$ such that

$$x_0 = x, \quad x_1 = y \quad \text{and} \quad x_r \prec x_s \quad \text{for } r < s.$$

A relation R is said to be *interpolative* if

$$aRb \quad \Rightarrow \quad \exists c,\ aRcRb.$$

The following is a trivial

8.4.1. OBSERVATION. *$\prec\!\!\prec$ is the largest interpolative $R \subseteq \prec$.*

From 8.1.1 we immediately obtain

8.4.2. LEMMA. (1) *$a \prec\!\!\prec b \ \Rightarrow\ a \leqslant b$, and for any a, $0 \prec\!\!\prec a \prec\!\!\prec 1$.*
(2) *$x \leqslant a \prec\!\!\prec b \leqslant y \ \Rightarrow\ x \prec\!\!\prec y$.*
(3) *If $a \prec\!\!\prec b$ then $b^* \prec\!\!\prec a^*$.*
(4) *If $a \prec\!\!\prec b$ then $a^{**} \prec\!\!\prec b$.*
(5) *If $a_i \prec\!\!\prec b_i$ for $i = 1, 2$ then $a_1 \vee a_2 \prec\!\!\prec b_1 \vee b_2$ and $a_1 \wedge a_2 \prec\!\!\prec b_1 \wedge b_2$.*

8.4.3. A frame L is said to be *completely regular* if

$$\text{for each } a \in L,\ a = \bigvee\{b \mid b \prec\!\!\prec a\}. \tag{8.4.1}$$

8.4.4. FACT. *A space X is completely regular iff the frame $\Omega(X)$ is completely regular.*

(Here the proof is not quite so straightforward as in 8.2.1. The implication $\Rightarrow$ is immediate. For the implication $\Leftarrow$ one has to construct suitable real functions; this can be done by a procedure imitating the standard proof of the well-known Urysohn lemma of classical topology.)

Denote by

CRegFrm

the full subcategory of **Frm** generated by completely regular frames. We have

8.4.5. PROPOSITION. **CRegFrm** *is* (*mono*)*coreflective in* **Frm**.

PROOF. Follows by the same reasoning as in 8.3.1. The situation is, however, simpler. In fact, here the construction stops already at the first step (see also the proposition in 8.5). □

8.5. *Aside: Strong inclusion.* An interpolative relation $\lhd$ on a frame L satisfying the properties observed for $\prec\!\!\prec$ in 8.4.2 is called a *strong inclusion*. It is said to be *admissible* if $a = \bigvee\{x \mid x \lhd a\}$ for each $a \in L$. A cover A (that is, $A \subseteq L$ such that $\bigvee A = 1$) is said to be a $\lhd$-*strong refinement* of a cover B if for each $a \in B$ there is a $b \in B$ such that $a \lhd b$. The following proposition, due to Banaschewski and Brümmer, will be useful in the sequel.

PROPOSITION. *Let $\lhd$ be a strong inclusion on a frame L. Then it is an admissible strong inclusion on the subframe*

$$M = \left\{a \mid a = \bigvee\{x \mid x \lhd a\}\right\}$$

and each cover of L which has a $\lhd$-strong refinement in L has a $\lhd$-strong refinement by a cover of M.

PROOF. For couples x, a such that $x \lhd a$ choose $b_{x,a}$ with $x \lhd b_{x,a} \lhd a$ and set $\varphi(x, a) = \bigvee\{y \mid x \lhd y \lhd b_{x,a}\}$. Then

(1) $\varphi(x, a) \in M$, and

(2) $x \leqslant \varphi(x, a) \leqslant b_{x,a} \lhd a$.

If $a \in M$ we obviously have $a = \bigvee\{\varphi(x, a) \mid x \lhd a\}$ and hence $\lhd$ is admissible in M. If C is a $\lhd$-strong refinement of A then $B = \{\varphi(x, a) \mid c \in C, a \in A\}$ is a strong refinement of A by a cover of M. $\square$

8.6. *Normality.* We say that a frame L is *normal* if

> whenever $a \vee b = 1$ for $a, b \in L$, there exist $u, v \in L$ such that
> $u \wedge v = 0$, $u \vee b = 1$ and $a \vee v = 1$.

Of all the separation conditions discussed so far, this is the most immediate one, and we have a trivial

8.6.1. FACT. *A space X is normal iff the frame $\Omega(X)$ is normal.*

8.6.2. PROPOSITION. *A frame is normal iff for each finite system $a_1, \ldots, a_n \in L$ such that $\bigvee_{i=1}^{n} a_i = 1$ there are $u_i \prec a_i$ such that $\bigvee_{i=1}^{n} u_i = 1$.*

PROOF. I. The u in the definition of normality is rather below a: as $u \wedge v = 0$ we have $v \leqslant u^*$ and $u^* \vee a \geqslant a \vee v = 1$. Thus, if $a_1 \vee a_2 \vee \cdots \vee a_n = 1$ we can choose, inductively, $u_1 \prec a_1$ such that $u_1 \vee a_2 \vee \cdots \vee a_n = 1$, $u_2 \prec a_2$ such that $u_1 \vee u_2 \vee a_3 \vee \cdots \vee a_n = 1$, etc.

II. If the condition holds and $a \vee b = 1$, choose $u \prec a$ with $u \vee b = 1$. It suffices to put $v = u^*$. $\square$

8.7. *The relation of normality to (complete) regularity.*

8.7.1. LEMMA. *The relation $\prec$ in a normal frame is interpolative.*

PROOF. Let $a \prec b$ in a normal frame L. Then there are u, v such that $u \leqslant v^*$, $u \vee b = 1$ and $a^* \vee v = 1$. Thus, $a \prec v$ and $v^* \vee b \geqslant u \vee b = 1$ so that also $v \prec b$. $\square$

8.7.2. PROPOSITION. *Let L be normal and subfit. Then it is completely regular.*

NOTE. This is a (slightly stronger) counterpart of the standard classical fact that normal T_1-spaces are completely regular.

PROOF. By 8.7.1 and 8.4.1, $\prec = \prec\!\!\prec$. Thus, it suffices to show that L is regular. For $a \in L$ set $b = \bigvee\{x \mid x \prec a\} = \bigvee\{x \mid x^* \vee a = 1\}$. Let $a \vee c = 1$. By normality we have a u such that $u \vee c = 1$ and $a \vee u^* = 1$. Thus, $u \leqslant b$ and $b \vee c = 1$. Hence $a \vee c = 1 \Rightarrow b \vee c = 1$ and by subfitness $a \leqslant b$ $(\leqslant a)$. $\square$

REMARK. In [42], Dowker and Strauss presented an extensive study of separation axioms in point-free context. The Hausdorff type axioms discussed there are weaker than the Isbell–Hausdorff property. See also [88].

9. Intermezzo: Open homomorphisms, open maps, and T_D

9.1. A frame homomorphism $h : L \to M$ is said to be *open* if the image (recall 5.7) of each open sublocale is open. More explicitly, h is open iff

for each $b \in M$ there is an $a \in L$ such that

$$b \wedge h(x) = b \wedge h(y) \quad \text{iff} \quad a \wedge x = a \wedge y. \tag{Op}$$

PROPOSITION [82]. *Open frame homomorphisms are exactly the complete Heyting homomorphisms.*

PROOF. The assumption of the existence, for each $b \in M$, of the $a \in L$ satisfying the (Op) above can be rewritten as that

there is a mapping $\varphi : M \to L$ such that

$$b \wedge h(x) = b \wedge h(y) \quad \text{iff} \quad \varphi(b) \wedge x = \varphi(b) \wedge y \tag{$*$}$$

which in turn is equivalent to

$$b \wedge h(x) \leqslant h(y) \quad \text{iff} \quad \varphi(b) \wedge x \leqslant y. \tag{$**$}$$

Setting $x = 1$ we see that

φ is a left Galois adjoint to h and hence, in particular, h preserves all meets.

Further, from $(**)$ we obtain that $b \leqslant h(x) \to h(y)$ iff $\varphi(b) \leqslant x \to y$ iff $b \leqslant h(x \to y)$. Thus, $h(x) \to h(y) = h(x \to y)$ and h is Heyting. On the other hand, if h is a complete

Heyting homomorphism and if φ is its left Galois adjoint, we have $b \wedge h(x) \leqslant h(y)$ iff $b \leqslant h(x) \to h(y) = h(x \to y)$ iff $\varphi(b) \leqslant x \to y$ iff $\varphi(b) \wedge x \leqslant y$ and $(**)$ is satisfied. $\square$

9.2. PROPOSITION. *Let L be fit (in particular, regular). Then the following statements on an $h : L \to M$ are equivalent:*
(1) *h is an open frame homomorphism,*
(2) *h is a complete Heyting homomorphism,*
(3) *h is a complete lattice homomorphism.*

PROOF. (1) $\Rightarrow$ (2) is in 9.1 and (2) $\Rightarrow$ (3) is trivial.

(3) $\Rightarrow$ (1): Recall 7.2.2 (and 8.2.2). The congruences on the left-hand side and on the right-hand side of $(*)$ are equivalent iff the classes of 1 coincide. Thus, $(*)$ is equivalent to

$$b \wedge h(x) = b \text{ (that is, } b \leqslant h(x)) \quad \text{iff} \quad \varphi(b) \wedge x = \varphi(b) \text{ (that is, } \varphi(b) \leqslant x). \qquad \square$$

9.3. In case of spaces and continuous maps the condition $(**)$ implies that there is a $\varphi : \Omega(X) \to \Omega(Y)$ such that

$$\varphi(U) \subseteq V \quad \text{iff} \quad U \subseteq f^{-1}(V).$$

Since we have for the images of subsets $f[U] \subseteq V$ iff $U \subseteq f^{-1}(V)$ we can be tempted to infer that $f[U] = \varphi(U)$ and hence $f[U]$ is always open. But this is not a correct conclusion: $(U \mapsto f[U])$ is a left adjoint to $(V \mapsto f^{-1}(V)) : \exp Y \to \exp X$, not to $(V \mapsto f^{-1}(V)) : \Omega(Y) \to \Omega(X)$ and hence the use of the unicity of the adjoint has been premature.

Of course, if f is an open continuous map then $(U \mapsto f[U])$ can be used as the required φ, and $\Omega(f)$ is an open homomorphism. On the other hand, the statement that $\Omega(f)$ open should imply f open is not generally true. But it is true for a fairly large class of spaces; namely, we have

PROPOSITION. *Let Y be T_D (in particular, T_1). Then a continuous $f : X \to Y$ is open iff $\Omega(f)$ is an open homomorphism.*

This has to do with the following fact: For a general subset A of a topological space X consider the congruence E_A on $\Omega(X)$ defined by

$$U E_A V \quad \text{iff} \quad U \cap A = V \cap A. \tag{$*$}$$

For T_D-spaces we have $E_A = E_B \;\Rightarrow\; A = B$, but generally this implication does not hold (thus, *in spaces that are not T_D the sublocales do not represent subspaces properly* – similarly as in spaces that were not sober the frame homomorphisms did not properly represent continuous maps). In particular in our problem above one had only $E_{f[U]} = E_{\varphi(U)}$ following immediately from $(*)$. That is, $f[U]$ and $\varphi(U)$ cannot be distinguished by testing them by meets with open sets, but do not necessarily coincide.

10. Compactness and compactification

10.1. *Covers, compactness, Lindelöf property.* A *cover* of a frame L is a subset $A \subseteq L$ such that $\bigvee A = 1$, and a *subcover* B of A is a subset $B \subseteq A$ which is still a cover. These notions obviously correspond to those of open covers and subcovers in classical topology. Hence we also have the following immediate extensions of classical notions:

A frame is said to be *compact* (resp. *Lindelöf*) if each cover A of L has a finite (resp. at most countable) subcover.

The following statement is trivial:

10.1.1. PROPOSITION. (1) *Each subframe of a compact* (*resp. Lindelöf*) *frame is compact* (*resp. Lindelöf*).

(2) *Each closed sublocale* $\uparrow c$ *of a compact* (*resp. Lindelöf*) *frame is compact* (*resp. Lindelöf*).

(Note that the first statement is related to the classical fact that quotients of compact (resp. Lindelöf) spaces are compact (resp. Lindelöf).)

10.2. *Regular Lindelöf implies normal, like in classical spaces.*

PROPOSITION. *Each regular Lindelöf* (*in particular, each regular compact*) *frame is normal. Consequently* (*recall* 8.7) *it is completely regular.*

PROOF. Let $a \vee b = 1$ in L. By regularity, $a = \bigvee\{x \mid x \prec a\}$ and $b = \bigvee\{y \mid y \prec b\}$. Thus, $\bigvee\{x \mid x \prec a\} \vee b = 1 = a \vee \bigvee\{y \mid y \prec b\}$ and by the Lindelöf property there are $x_1, x_2, \ldots \prec a,\ y_1, y_2, \ldots \prec b$ such that $\bigvee_{i=1}^{\infty} x_i \vee b = 1 = a \vee \bigvee_{i=1}^{\infty} y_i$. By 8.1.1.5 we can assume that

$$x_1 \leqslant x_2 \leqslant \cdots \quad \text{and} \quad y_1 \leqslant y_2 \leqslant \cdots.$$

Set $u_i = x_i \wedge y_i^*$ and $v_i = y_i \wedge x_i^*$. Then we have $a \vee v_i = a \vee (x_i^* \wedge y_i) = (a \vee x_i^*) \wedge (a \vee y_i) = a \vee y_i$ and similarly $u_i \vee b = x_i \vee b$. Consequently, if we set $u = \bigvee u_i$ and $v = \bigvee v_i$ we have

$$a \vee v = \bigvee (a \vee y_i) = a \vee \bigvee y_i = 1 \quad \text{and} \quad u \vee b = 1.$$

Finally, $u_i \wedge v_j = x_i \wedge y_i^* \wedge x_j^* \wedge y_j = 0$ for any two i, j since if $i \leqslant j$ then $x_i \wedge x_j^* = 0$ and if $i \geqslant j$ then $y_j \wedge y_i^* = 0$. Thus, $u \wedge v = \bigvee_{i,j} u_i \wedge v_j = 0$. □

NOTE. However easy the proof was, for a compact L it is even much easier: If $a \vee b = 1$ we have $\bigvee\{x \mid x \prec a\} \vee b = 1$ and hence by compactness and 8.1.1.5 there is an $x \prec a$ (and hence $x^* \vee a = 1$) such that $x \vee b = 1$.

10.3. *Two counterparts of statements on regular compact spaces.*

10.3.1. PROPOSITION. *Let L be regular and let M be compact. Then each dense $h: L \to M$ is one–one.*

PROOF. In view of the second corollary in 8.3.1 it suffices to show that h is co-dense. Suppose $h(a) = 1$. Since $a = \bigvee\{x \mid x \prec a\}$, the set $\{h(x) \mid x \prec a\}$ is a cover of M and hence there are $x_1, \ldots, x_n \prec a$ such that $\bigvee h(x_i) = 1$. By 8.1.1.5 $x = x_1 \vee \cdots \vee x_n \prec a$ and we have

$$h(x) = 1 \quad \text{and} \quad x^* \vee a = 1.$$

Since $h(x^*) \leqslant h(x)^* = 0$, $x^* = 0$ and finally $a = 1$. □

10.3.2. PROPOSITION. *Let L be regular and $h: L \to M$ a sublocale with compact M. Then h is closed.*

PROOF. Recall the closure $\check{c}: L \to \uparrow c$ of h from 5.4 ($c = \bigvee\{x \mid h(x) = 0\}$). The homomorphism $g: \uparrow c \to M$ such that $g \cdot \check{c} = h$ is dense onto. Hence, by 10.3.1, it is an isomorphism. □

10.4. *An easy "compactification".* In this paragraph we will introduce an easy construction of compact spaces naturally related to general ones. Although it is not a compactification in the strict sense (in particular, the construction yields a new different frame even if the original has been already compact), it will be a basis for the satisfactory construction in the next paragraph.

An *ideal* in a frame L is a non-void subset $J \subseteq L$ such that

(I1) $b \leqslant a \in J \Rightarrow b \in J$, and

(I2) $a, b \in J \Rightarrow a \vee b \in J$.

Denote by

$$\mathfrak{J}L$$

the set of all ideals in L ordered by inclusion.

10.4.1. PROPOSITION. *$\mathfrak{J}L$ is a compact frame.*

PROOF. Obviously, intersection of ideals is an ideal. For the supremum we have the formula

$$\bigvee J_i = \left\{\bigvee X \mid X \text{ finite}, X \subseteq \bigcup J_i\right\}$$

(obviously the set is an ideal, and each ideal J containing all J_i has to contain all the $\bigvee X$). If J_i, K are ideals and $x = x_1 \vee \cdots \vee x_n \in (\bigvee J_i) \cap K$, $x_j \in J_{i_j}$, then by (I1) $x_j \in J_{i_j} \cap K$ and $x \in \bigvee (J_i \cap K)$. The inclusion $\bigvee (J_i \cap K) \subseteq (\bigvee J_i) \cap K$ is trivial and hence $\mathfrak{J}L$ is a frame. Now let $\{J_i \mid i \in I\}$ be a cover. Thus, $1 \in L = \bigvee J_i$ and there are $x_j \in J_{i_j}$ such that $1 = x_1 \vee \cdots \vee x_n$. Then we have $1 \in \bigvee_{j=1}^{n} J_{i_j}$ and by (I1) $L = \bigvee_{j=1}^{n} J_{i_j}$. □

10.4.2. *The homomorphisms v_L.* Define a mapping $v_L : \mathfrak{J}L \to L$ by setting $v_L(J) = \bigvee J$.

LEMMA. *v_L is a dense sublocale homomorphism.*

PROOF. Define a mapping $\alpha : L \to \mathfrak{J}L$ by setting $\alpha(a) = {\downarrow}a$. Obviously, $v\alpha(a) = a$ and $\alpha v(J) = {\downarrow}\bigvee J \supseteq J$. Thus, first, v is onto and, second, v is a left Galois adjoint and hence it preserves all suprema. We have $v(L) = 1$ and, by (I1),

$$\begin{aligned} v(J_1) \wedge v(J_2) &= \bigvee \{x \wedge y \mid x \in J_1, y \in J_2\} \\ &\leqslant \bigvee \{z \mid z \in J_1 \cap J_2\} = v(J_1 \cap J_2) \quad (\leqslant v(J_1) \wedge v(J_2)) \end{aligned}$$

so that v preserves finite meets.

Finally, if $v(J) = \bigvee J = 0$ then necessarily $J = \{0\}$, the bottom of $\mathfrak{J}L$ (ideals are non-void). □

10.4.3. *$\mathfrak{J}$ as a functor.* For a frame homomorphism $h : L \to M$ define $\mathfrak{J}h : \mathfrak{J}L \to \mathfrak{J}M$ by setting

$$\mathfrak{J}h(L) = {\downarrow}h[J].$$

The following is a matter of trivial immediate checking:

PROPOSITION. *$\mathfrak{J}$ is a functor* **Frm** $\to$ **Frm** *and $v = (v_L)_L$ is a natural transformation $\mathfrak{J} \to \mathrm{Id}$.*

10.5. *The real thing: Stone–Čech compactification.* In this paragraph we will obtain, following the construction from [18], a compactification of completely regular frames by an easy modification of the previous construction.

An ideal $J \subseteq L$ is said to be *regular* if

$$\text{for each } a \in J \text{ there is a } b \in J \text{ such that } a \prec\!\!\prec b. \tag{IR}$$

The set of all regular ideals in L will be denoted by

$$\mathfrak{R}L.$$

LEMMA. *$\mathfrak{R}L$ is a subframe of $\mathfrak{J}L$. In particular, it is compact.*

PROOF. Intersection of regular ideals is obviously regular. Now let J_i be regular and let $a \in \bigvee J_i$; then $a = x_1 \vee \cdots \vee x_n$ with some $x_j \in J_{i_j}$. There are $y_j \in J_{i_j}$ such that $x_j \prec\!\!\prec y_j$ and hence $b = y_1 \vee \cdots \vee y_n \in \bigvee J_i$, and $a \prec\!\!\prec b$ by 8.4.2. □

For an element a of a frame L set

$$\sigma(a) = \{x \mid x \prec\!\!\prec a\}.$$

Using 8.4.2 and the interpolativity of $\prec\!\!\prec$ we immediately obtain

FACT. *$\sigma(a)$ is a regular ideal.*

PROPOSITION. *$\mathfrak{R}L$ is a completely regular compact frame.*

PROOF. By lemma and by 10.2 it suffices to show that $\mathfrak{R}L$ is regular. Since, further, for a regular ideal J obviously $J = \bigcup\{\sigma(a) \mid a \in J\} = \bigvee\{\sigma(a) \mid a \in J\}$ it suffices to show that

$$b \prec\!\!\prec a \text{ in } L \quad \Rightarrow \quad \sigma(b) \prec \sigma(a) \text{ in } \mathfrak{R}L.$$

Interpolate $b \prec\!\!\prec x \prec\!\!\prec y \prec\!\!\prec a$. Since $\sigma(b^*) \cap \sigma(b) = \{0\}$ we have $\sigma(b^*) \subseteq \sigma(b)^*$, and by 8.4.2.3, $x^* \in \sigma(b^*) \subseteq \sigma(b)^*$. Thus, $1 = x^* \vee y \in \sigma(b)^* \vee \sigma(a)$ and hence $\sigma(b)^* \vee \sigma(a) = L$, the top of $\mathfrak{R}L$. □

THEOREM (Stone–Čech compactification). *Define $\mathfrak{R}h = \mathfrak{J}h$ for homomorphisms $h : L \to M$ and $v_L : \mathfrak{R}L \to L$ by $v_L(J) = \bigvee J$. These formulas yield a functor $\mathfrak{R}$:* **CRegFrm** $\to$ **CRegFrm** *and a natural transformation $v : \mathfrak{R} \to \mathrm{Id}$ such that*

(1) *each $\mathfrak{R}L$ is (regular and) compact,*
(2) *each v_L is a dense sublocale homomorphism, and*
(3) *v_L is an isomorphism iff L is compact.*

PROOF. If L is completely regular we have $v_L\sigma(a) = a$ for each a, and obviously $\sigma(v_L(J)) \subseteq J$. Thus (again) v_L is a left Galois adjoint and hence it preserves suprema. Preserving finite meets is seen by the same procedure as in 10.4.2. Obviously v_L is dense.

If J is regular in L, $\mathfrak{J}h(L) = {\downarrow}h[L]$ is regular by (8.3.1).

Thus, the only statement left to be proved is that if L is compact then v_L is an isomorphism. (At this moment, the reader may wonder why we do not simply use 10.3.1. This would indeed yield the result, but we wish to have everything very explicitly constructive – see next paragraph. Therefore we prefer to show directly that σ is the inverse of v_L.) We already know that $J \subseteq \sigma v_L(J)$. Now let L be compact and let $x \in \sigma v_L(J)$. Then $x \prec \bigvee J$, hence $x^* \vee \bigvee J = 1$ and by compactness there are $a_1, \ldots, a_n \in J$ such that $x^* \vee a_1 \vee \cdots \vee a_n = 1$. Then $a = a_1 \vee \cdots \vee a_n \in J$, and $x \prec a$ and hence $x \in J$; thus also $\sigma v_L(J) \subseteq J$ and σ is the inverse of v_L. □

10.6. *A note on the Tychonoff theorem.* Denote by

KRegFrm

the category of compact regular frames. By 10.5, **KRegFrm** is a coreflective subcategory in **CRegFrm** (which in turn is coreflective in **Frm**) and hence in particular

coproducts of compact frames spaces are compact.

This is a counterpart of a (substantial) part of the well-known Tychonoff theorem.

It is a remarkable fact that while the classical Tychonoff theorem is heavily non-constructive (the general one is equivalent to the axiom of choice, the special one – concerning regular spaces – is equivalent to the Boolean ultrafilter theorem), in the point-free reasoning one obtains the compactness of coproducts constructively (the reader may check that we have used no choice principle and not even the excluded middle; it should be noted that one can also present a fully constructive proof of the general point-free Tychonoff theorem, of course using much subtler methods than here – see [70]). Putting the Tychonoff theorem into the right perspective is one of the achievements of point-free topology: now we see, roughly speaking, that it is not the compactness but the existence of points that needs the choice principle.

11. Locally compact (continuous) frames

11.1. *The relation "well below".* The relation to be discussed has its use in much more general context. We will restrict ourselves, however, to complete lattices and in particular to frames.

We say that a is *well below* b in a complete lattice L, and write

$$a \ll b,$$

if for each directed $D \subseteq L$,

$$b \leqslant \bigvee D \quad \Rightarrow \quad \exists d \in D,\ a \leqslant d. \tag{WB}$$

Equivalently, $a \ll b$ iff for each non-void $X \subseteq L$,

$$b \leqslant \bigvee X \ \Rightarrow \ \text{there is a non-void finite } Y \subseteq X \text{ such that } a \leqslant \bigvee Y. \tag{WB$'$}$$

The following statements are immediate.

11.1.1. LEMMA. (1) $0 \ll a$ *and if* $x \leqslant a \ll b \leqslant y$ *then* $x \ll y$.

(2) *If* $a_1, a_2 \ll b$ *then* $a_1 \vee a_2 \ll b$.

(3) *Consequently, the set* $\{x \mid x \ll a\}$ *is directed.*

11.2. *Continuous frames.* A complete lattice (or frame) L is said to be *continuous* if

$$\text{for each } a \in L,\ a = \bigvee \{x \mid x \ll a\}.$$

11.2.1. OBSERVATION. *Let* X *be a locally compact topological space. Then* $\Omega(X)$ *is a continuous frame.*

(Indeed, let $U \subseteq X$ be open. For $x \in U$ consider a compact Y and an open $V \ni x$ such that $V \subseteq Y \subseteq U$. Then $V \ll U$.)

11.2.2. LEMMA. (1) *In any frame,* $a \prec b \ll 1 \;\Rightarrow\; a \ll b$.
(2) *If* L *is* (*completely*) *regular then* $a \ll b \;\Rightarrow\; a \prec b$ $(a \prec\!\!\prec b)$.

PROOF. (1) If $b \leqslant \bigvee D$ for a directed D we have $1 \leqslant a^* \vee \bigvee D$ and as $b \ll 1$ there is a $d \in D$ with $b \leqslant a^* \vee d$. Thus, $a = a \wedge b \leqslant a \wedge d$, that is, $a \leqslant d$.
(2) We have $b = \bigvee\{x \mid x \prec (\prec\!\!\prec) b\}$ and since $\{x \mid x \prec (\prec\!\!\prec) b\}$ is directed, there is an $x \prec (\prec\!\!\prec)\, b$ such that $a \leqslant x$. □

11.2.3. A general compact frame is not necessarily continuous. But we have

PROPOSITION. *Any open sublocale* $\downarrow u$ (*more precisely, the open sublocale is* $\hat{u} : L \to \downarrow u$) *of a regular compact frame is continuous. Moreover,* $\prec = \prec\!\!\prec = \ll$.

PROOF. In a compact frame $1 \ll 1$ and hence, by 11.2.2, $a \prec b \;\Rightarrow\; a \ll b$; furthermore, if $a \ll b$ in L and $b \leqslant u$ then also $a \ll b$ in $\downarrow u$. □

11.2.4. PROPOSITION. *In a continuous lattice the relation* $\ll$ *interpolates.*

PROOF. We have $b = \bigvee\{x \mid x \ll b\} = \bigvee\{\bigvee\{y \mid y \ll x\} \mid x \ll b\} = \bigvee\{y \mid \exists x,\ y \ll x \ll b\}$. Let $a \ll b$. By 11.1.1.2, $\{y \mid \exists x,\ y \ll x \ll b\}$ is directed and hence we have a y and an x such that $a \leqslant y \ll x \ll b$. □

11.2.5. PROPOSITION. *A completely regular frame is continuous iff it is open in its Stone–Čech compactification* (*that is, iff the* $\upsilon_L : \mathfrak{R}L \to L$ *from* 10.5 *is an open sublocale map*).

PROOF. We have to prove that if L is continuous then there is a regular ideal A in L such that for any two regular ideals J_1, J_2,

$$\upsilon_L(J_1)\left(=\bigvee J_1\right) = \upsilon_L(J_2)\left(=\bigvee J_2\right) \quad \text{iff} \quad J_1 \cap A = J_2 \cap A.$$

From 11.2.2 we immediately infer that $A = \{x \mid x \ll 1\}$ is a regular ideal. Since υ_L preserves meets, and $\upsilon_L(A) = 1$ by continuity, $J_1 \cap A = J_2 \cap A \;\Rightarrow\; \upsilon_L(J_1) = \upsilon_L(J_2)$. Now let $\bigvee J_1 = \bigvee J_2$ and let $a \in J_1 \cap A$. Choose $b \in J_1 \cap A$ such that $a \prec\!\!\prec b$ (and hence $a \ll b$ by 11.2.2). Since ideals are directed and $b \leqslant \bigvee J_2 = \bigvee J_1$, there is an $x \in J_2$ with $a \leqslant x$ and hence, finally, $a \in J_2$. □

11.2.6. From 11.2.3, 11.2.5 and 6.5 we immediately infer

COROLLARY. *Coproduct* (*in* **Frm**) *of two completely regular continuous frames is continuous.*

11.3. *Separating elements by open filters and by complete ones.* Recall the Scott topology from 1.5.1. We have

LEMMA. *A filter* F *in a frame is complete iff it is prime and Scott open.*

PROOF. The implication $\Rightarrow$ is trivial, and so is the other as well: Let F be prime and Scott open and let $\bigvee_{i \in J} a_i \in F$. Since F is open, there are $a_{i_1}, \dots, a_{i_n}$ with $a_{i_1} \vee \cdots \vee a_{i_n} \in F$. Since F is prime, some of the a_{i_j} is in F. $\square$

11.3.1. PROPOSITION. *Let F be a Scott open filter in a frame L such that $b \notin F \ni a$. Then there is a complete filter $P \supseteq F$ such that $b \notin P \ni a$.*

PROOF. This is just the famous Birkhoff theorem with the openness added. Using Zorn's lemma in the standard way (taking into account that unions of open sets are open) we obtain an open filter $P \supseteq F$ maximal with respect to the condition that $b \notin P \ni a$. We will prove that it is prime (and hence, by the lemma, complete). Suppose it is not; then there are $u, v \notin P$ such that $u \vee v \in P$. Set

$$G = \{x \mid x \vee v \in P\}.$$

G is obviously a Scott open filter and, because of the u, $P \subsetneq G$. Thus, $b \in G$, $b \vee v \in P$ and we can repeat the procedure with $v, b \notin P$, $v \vee b \in P$ and $H = \{x \mid x \vee b \in P\}$ to obtain the contradiction $b = b \vee b \in P$. $\square$

11.3.2. COROLLARY. *Each Scott open filter in a frame is an intersection of complete filters.*

11.4. *Spatiality of continuous frames. Hofmann–Lawson duality.*

PROPOSITION. *Each continuous frame is spatial.*

PROOF. By 4.6 statement (2), using the spectrum in the form from 4.3.1, it suffices to prove that if $a \leqslant b$ then there is a complete filter P such that $b \notin P \ni a$. By 11.3.1 it suffices to find a Scott open filter F such that $b \notin F \ni a$. Since $a \nleqslant b$ and L is continuous, there is a c such that

$$c \ll a \quad \text{and} \quad c \nleqslant b.$$

Interpolate (recall 11.2.4) inductively

$$a \gg x_1 \gg x_2 \gg \cdots \gg x_n \gg \cdots \gg c \tag{11.4.1}$$

and set

$$F = \{x \mid x \geqslant x_k \text{ for some } k\}. \tag{11.4.2}$$

Then F is obviously a Scott open filter, and $b \notin F \ni a$. $\square$

From 11.2.3 we now obtain

COROLLARY. *Each open sublocale of a regular compact frame is spatial.*

11.4.2. LEMMA. *Let L be a frame. A subset $K \subseteq \Sigma L$ is compact iff $\bigcap\{P \mid P \in K\}$ is Scott open.*

PROOF. Let $\bigcap\{P \mid P \in K\}$ be Scott open and let $K \subseteq \bigcup\{\Sigma_a \mid a \in A\}$. Then $\bigvee A \in \bigcap\{P \mid P \in K\}$ since for each $P \in K$ there is an $a \in A$ such that $a \in P$, and hence $\bigvee A \in P$. Thus, there are $a_1, \ldots, a_n \in A$ with $a_1 \vee \cdots \vee a_n \in \bigcap\{P \mid P \in K\}$ and hence $K \subseteq \Sigma_{a_1 \vee \cdots \vee a_n} = \bigcup_{i=1}^n \Sigma_{a_i}$. If K is compact and $\bigvee A \in \bigcap\{P \mid P \in K\}$ then $K \subseteq \Sigma_{\bigvee A} = \bigcup\{\Sigma_a \mid a \in A\}$ and there are $a_1, \ldots, a_n \in A$ such that $K \subseteq \Sigma_{a_1 \vee \cdots \vee a_n} = \bigcup_{i=1}^n \Sigma_{a_i}$ and finally $a_1 \vee \cdots \vee a_n \in \bigcap\{P \mid P \in K\}$. □

11.4.3. PROPOSITION. *Let L be a continuous frame. Then ΣL is locally compact.*

PROOF. Let $P \in \Sigma L$ and $P \in \Sigma_a$, that is, $a \in P$. Since $a = \bigvee\{x \mid x \ll a\}$ there is a $c \ll a$ such that $c \in P$. Consider the F constructed as in (11.4.1) and (11.4.2) and set

$$K = \{Q \in \Sigma L \mid F \subseteq Q\}.$$

By 11.3.2, $F = \bigcap K$ and hence, by 11.5, K is compact. Now if $c \in Q$ we have $F \subseteq Q$, and if $F \subseteq Q$ then $a \in Q$. Thus, $P \in \Sigma_c \subseteq K \subseteq \Sigma_a$. □

11.4.4. From 11.2.1, 11.4 and 11.6.1 we conclude

THEOREM (Hofmann–Lawson duality). *The functors Ω and Σ (recall Section 4) restrict to a dual equivalence of the category of sober locally compact spaces and the category of continuous frames.*

(This justifies speaking of continuous frames as of *locally compact frames*.)

NOTES. (1) By 11.2.3 this further restricts to a duality between the category of compact Hausdorff spaces and **KRegFrm**.

(2) The duality appeared, first, in [59]. In the present proof we have followed the reasoning from [9].

11.5. *A case of preserving products by the functor Ω.* It is a well-know fact that a product of two locally compact spaces in **Top** is locally compact. By 11.2.6, a product of two completely regular continuous frames is continuous. Thus, as a consequence of 11.4.4 we also obtain

PROPOSITION. *The functor $\Omega :$ **Top** $\to$ **Loc** preserves finite products of sober completely regular locally compact spaces.*

12. Uniform and nearness frames

12.1. *Systems of covers and the relation* $\lhd$. For covers (and, more generally, for subsets) A, B of L we write $A \leqslant B$ if A refines B (that is, if for every $a \in A$ there is a $b \in B$ such that $a \leqslant b$). Further, set $A \wedge B = \{a \wedge b \mid a \in A,\ b \in B\}$. For a cover A of L and an element $x \in L$ set

$$Ax = \bigvee\{a \mid a \in A,\ a \wedge x \neq 0\}$$

and for two covers A, B set

$$AB = \{Ab \mid b \in B\}.$$

It is easy to check that

(1) $A \leqslant B$ and $x \leqslant y \Rightarrow Ax \leqslant By$,
(2) $A(Bx) \leqslant (AB)x = A(B(Ax))$,
(3) $(A_1 \wedge \cdots \wedge A_n)(B_1 \wedge \cdots \wedge B_n) \leqslant (A_1B_1) \wedge \cdots \wedge (A_nB_n)$, and
(4) $A(\bigvee x_i) = \bigvee Ax_i$; the correspondence $(x \mapsto Ax) : L \to L$ has a right adjoint $(x \mapsto x/A) : L \to L$ (namely, $x/A = \bigvee\{z \mid Az \leqslant x\}$).

12.1.1. For a system of covers $\mathcal{A}$ on L define the relation $\lhd_{\mathcal{A}}$ (the subscript will be often omitted) by

$$x \lhd_{\mathcal{A}} y \quad \equiv_{\mathrm{df}} \quad \exists A \in \mathcal{A},\ Ax \leqslant y.$$

LEMMA. (1) $x \leqslant x' \lhd_{\mathcal{A}} y' \leqslant y \Rightarrow x \lhd_{\mathcal{A}} y$.

(2) *If for any two* $A, B \in \mathcal{A}$ *there is a common refinement* $C \in \mathcal{A}$ *then* $x \lhd_{\mathcal{A}} y_1, y_2 \Rightarrow x \lhd_{\mathcal{A}} (y_1 \wedge y_2)$ *and* $x_1, x_2 \lhd_{\mathcal{A}} y \Rightarrow (x_1 \vee x_2) \lhd_{\mathcal{A}} y$.

(3) $x \lhd_{\mathcal{A}} y \Rightarrow x \prec y$.

(4) $x \lhd_{\mathcal{A}} y \Rightarrow x^{**} \lhd_{\mathcal{A}} y$.

PROOF. (1) and (2) are immediate consequences of the formulas above.

(3) If $Ax \leqslant y$ then $1 = \bigvee A = \bigvee\{a \mid a \leqslant x^*\} \vee Ax \leqslant x^* \vee y$.

(4) follows from 2.(PC5) $\square$

REMARK. Since $\{x^*, y\}x \leqslant y$, $\lhd$ coincides with $\prec$ whenever $\mathcal{A}$ contains all finite covers.

12.1.2. A system of covers $\mathcal{A}$ of L is said to be *admissible* if

$$\forall x \in L,\ x = \bigvee\{y \mid y \lhd_{\mathcal{A}} x\}. \tag{12.1.2}$$

Using the right adjoint above, this can be expressed as

$$\forall x \in L,\ x = \bigvee\{x/A \mid A \in \mathcal{A}\}.$$

PROPOSITION. (1) *L is regular iff there exists an admissible $\mathcal{A}$ on L.*

(2) *If $\mathcal{A}$ is admissible and $\mathcal{B} \supseteq \mathcal{A}$ then $\mathcal{B}$ is admissible.*

(3) *If $\mathcal{A}$ is admissible on L then $\bigcup \mathcal{A}$ generates L by joins.*

PROOF. (1) If $\mathcal{A}$ is admissible then L is regular by 12.1.1.3. If L is regular we can take for $\mathcal{A}$ the system of all covers.

(2) is obvious.

(3) Set $\mathcal{C}(\mathcal{A}, x) = \{a \mid \exists A \in \mathcal{A},\ \exists y \in L,\ a \in A \text{ and } Ay \leqslant x\}$. By (12.1.2), $x = \bigvee \mathcal{C}(\mathcal{A}, x)$. □

12.2. *Uniformity and nearness.* A *uniformity* on a frame L is an admissible system of covers $\mathcal{A}$ such that

(U1) $A \in \mathcal{A}$ and $A \leqslant B \ \Rightarrow\ B \in \mathcal{A}$,

(U2) $A, B \in \mathcal{A} \ \Rightarrow\ A \wedge B \in \mathcal{A}$,

(U3) for every $A \in \mathcal{A}$ there is a $B \in \mathcal{A}$ such that $BB \leqslant A$.

(A cover B such that $BB \leqslant A$ is usually called a *star-refinement* of A.)

If (U3) is not required we speak of a *nearness* on L and if, instead of (U3), we require

(SN) for every $A \in \mathcal{A}$, $\{b \mid b \lhd_{\mathcal{A}} a \in A\} \in \mathcal{A}$

we speak of a *strong nearness* (if $BB \leqslant A$ then $B \leqslant \{b \mid b \lhd_{\mathcal{A}} a \in A\} \in \mathcal{A}$; hence (U3) $\Rightarrow$ (SN)).

NOTE. If X is a topological space then a uniformity on $\Omega(X)$ is the same as the classical uniformity defined by (open) covers (the admissibility makes sure that the topology induced by the uniformity coincides with the original one). A strong nearness on $\Omega(X)$ is what is known as *regular nearness* on X (see [57]).

A *uniform frame* is a couple $(L, \mathcal{A})$ where $\mathcal{A}$ is a uniformity on a frame L. If $\mathcal{A}$ is just a nearness, or a strong nearness, we speak of a *nearness frame* resp. a *strong nearness frame*.

12.2.1. A nearness or uniformity is often described by a *basis*, that is, an admissible system of covers $\mathcal{B}$ such that

(U2′) $A, B \in \mathcal{B} \ \Rightarrow\ \exists C \in \mathcal{B},\ C \leqslant A, B$

and in the case of uniformity satisfying, moreover, (U3).

Then $\mathcal{A} = \{B \mid \exists B \in \mathcal{B},\ B \leqslant A\}$ is obviously a nearness resp. uniformity, and $\lhd_{\mathcal{A}} = \lhd_{\mathcal{B}}$.

Sometimes, one constructs a uniformity as $\{A \mid \exists B_1, \dots, B_n \in \mathcal{S},\ B_1 \wedge \cdots \wedge B_n \leqslant A\}$ from a *subbasis* $\mathcal{S}$ for which only (U3) is required; here, however, one has to be careful since $\lhd_{\mathcal{A}}$ in general does not coincide with $\lhd_{\mathcal{S}}$.

The following is an immediate consequence of (U3), 8.4.1 and 12.1.1:

OBSERVATION. *If $\mathcal{A}$ is a (basis of) uniformity on L then $\lhd_{\mathcal{A}}$ interpolates. Consequently,*

(1) *$\lhd_{\mathcal{A}}$ is a strong inclusion in the sense of* 8.5, *and*

(2) $x \lhd_{\mathcal{A}} y \ \Rightarrow\ x \prec\!\!\prec y$.

12.2.2. PROPOSITION. (1) *A frame L admits a nearness iff it admits a strong nearness iff it is regular.*

(2) *A frame L admits a uniformity iff it is completely regular.*

PROOF. (1) is already contained in 12.1.2. Observe that if A is a cover of a regular L then also $\{x \mid x \lhd a \in A\}$ is a cover.

(2) If L admits a uniformity then it is completely regular by the observation above. Now let L be completely regular. For a sequence $a_1 \lll a_2 \lll \cdots \lll a_n$ define a cover

$$A(a_1, \ldots, a_n) = \{a_2, a_1^* \wedge a_3, a_2^* \wedge a_4, \ldots, a_{n-2}^* \wedge a_n, a_{n-1}^*\}.$$

If we interpolate $a_1 \lll u_1 \lll v_1 \lll a_2 \lll u_2 \cdots \lll v_{n-1} \lll a_n$ and set $B = A(a_1, u_1, v_1, a_2, u_2, \ldots, v_{n-1}, a_n)$ we easily check that $BB \leqslant A(a_1, a_2, \ldots, a_n)$. Thus, the system $\mathcal{S}$ of all the $A(a_1, a_2, \ldots, a_n)$ is a subbasis of a uniformity $\mathcal{A}$: $\mathcal{S}$ is admissible since if $x \lll a$ then $A(x, a)x \leqslant a$, and since $\mathcal{S} \subseteq \mathcal{A}$, $\mathcal{A}$ is admissible as well. $\square$

12.3. *Fine uniformity, full normality, normal covers.* A union of uniformities is obviously a uniformity. Thus,

each completely regular frame L has a largest uniformity.

This uniformity is called the *fine uniformity of* L. Unlike the largest (strong) nearness on a regular L, it is not necessarily the system of all covers since a cover does not necessarily have a star refinement. A frame L such that each cover has a star refinement is called *fully normal*.

A cover A of a general L is said to be *normal* if there are covers A_n, $n = 1, 2, \ldots,$ such that

$$A = A_1 \quad \text{and} \quad A_n \geqslant A_{n+1}A_{n+1} \quad \text{for all } n.$$

Thus, the fine uniformity of L consists of all the normal covers of L.

12.3.1. The term "fully normal" alludes to the following characteristics of normality:

PROPOSITION. *The following statements on a frame L are equivalent*:

(1) *L is normal,*
(2) *each finite cover of L has a finite star-refinement,*
(3) *each finite cover of L has a star-refinement.*

PROOF. (1) $\Rightarrow$ (2): Recall 8.6.2. Let $A = \{a_1, \ldots, a_n\}$ be a cover of L and let $B = \{b_1, \ldots, b_n\}$ be a cover such that $b_i \prec a_i$. Set $C_i = \{b_i^*, a_i\}$ and $C = B \wedge C_1 \wedge \cdots \wedge C_n$. Then $CC \leqslant A$: Indeed, fix $c = b_i \wedge \bigwedge_{j=1}^n c_j \in C$ (c_j either b_j^* or a_j) and consider another $x = b_k \wedge \bigwedge_{j=1}^n x_j \in C$ (x_j either b_j^* or a_j). If $x \wedge c \neq 0$ then $x_i \neq b_i^*$ and hence $x_i = a_i$ and $x \leqslant a_i$. Thus, $Cc \leqslant a_i$.

(2) $\Rightarrow$ (3) is trivial.

(3) $\Rightarrow$ (1): Let B be a star-refinement of $\{a_1, \ldots, a_n\}$. Set $b_i = \bigvee\{b \in B \mid Bb \leqslant a_i\}$. Then $\{b_1, \ldots, b_n\}$ is a cover and $Bb_i \leqslant a_i$ and hence $b_i \prec a_i$. Use 8.6.2. $\square$

12.4. *The unique uniformity on a compact regular frame*

PROPOSITION. *A compact regular frame L admits exactly one uniformity, namely the system of all covers of L.*

Since L is completely regular (recall 10.2) it admits a uniformity $\mathcal{A}$. Let U be an arbitrary cover of L. Set $V = \{v \mid v \lhd_{\mathcal{A}} u \in U\}$. Then, by admissibility, V is again a cover and we can choose $v_1, \ldots, v_n \in V$ such that $\bigvee v_i = 1$. Choose $A_i \in \mathcal{A}$ and $u_i \in U$ such that $A_i v_i \leqslant u_i$. Set $A = A_1 \wedge \cdots \wedge A_n$. Then $A \in \mathcal{A}$ and $A v_i \leqslant u_i$ for all i. For $a \in A$, $a \neq 0$ we have $a = a \wedge \bigvee v_i$ and hence there is a v_i such that $a \wedge v_i \neq 0$. Then $a \leqslant u_i$ and we see that $A \leqslant U$ and $U \in \mathcal{A}$.

NOTE. We have used only the properties (U1) and (U2) of $\mathcal{A}$. Thus, we have also proved that

the system of all covers is the only nearness on a regular compact frame.

12.5. *Uniform homomorphisms.* Let $(L, \mathcal{A}), (M, \mathcal{B})$ be uniform (or just nearness) frames. A frame homomorphism is said to be *uniform* if for each $A \in \mathcal{A}$, $f[A] \in \mathcal{B}$. (If the uniformities resp. nearnesses are represented by bases $\mathcal{A}, \mathcal{B}$ we require, of course, that for each $A \in \mathcal{A}$ there is a $B \in \mathcal{B}$ such that $B \leqslant f[A]$.)

NOTE. From 12.4 we immediately obtain

COROLLARY. *Let $(L, \mathcal{A})$ be an arbitrary nearness frame and let M be regular compact. Then each frame homomorphism $h : L \to M$ is uniform.*

The resulting categories will be denoted by

UniFrm (or **NearFrm**, or **SNearFrm**).

These categories are complete and cocomplete (the proof is straightforward). From 12.2.2 and corollary in 8.3.2 follows an important

12.5.1. OBSERVATION. *In any of the above categories, the dense morphisms are monomorphisms.*

12.5.2. LEMMA. *If $h : (L, \mathcal{A}) \to (M, \mathcal{B})$ is uniform then*

$$x \lhd_{\mathcal{A}} y \quad \Rightarrow \quad h(x) \lhd_{\mathcal{B}} h(y).$$

(If $h(a) \wedge h(x) = h(a \wedge x) \neq 0$ then $a \wedge x \neq 0$. Hence $h[A]h(x) \leqslant h(Ax)$.)

12.6. *Uniform embeddings.* A *uniform embedding* is a uniform sublocale mapping $h:(L,\mathcal{A}) \to (M,\mathcal{B})$ such that $\mathcal{B} = \{h[A] \mid A \in \mathcal{A}\}$.

12.6.1. LEMMA. *A uniform sublocale h is a uniform embedding iff $\{h[A] \mid A \in \mathcal{A}\}$ is a basis of $\mathcal{B}$.*

PROOF. For the right adjoint h_+ of h we have $x \leqslant h_+h(x)$ and, as h is onto, $hh_+ = \mathrm{id}$. Thus, if $B \geqslant h[A]$ we have $A' = h_+[B] \geqslant h_+h[A] \geqslant A$ and hence $A' \in \mathcal{A}$ and $h[A'] = hh_+[B] = B$. □

12.6.2. In the sequel, dense uniform embeddings will play a particular role. Here is a useful criterion:

PROPOSITION. *Let $h:(L,\mathcal{A}) \to (M,\mathcal{B})$ be a uniform dense sublocale map and let $\mathcal{A}$ be a strong nearness. Then h is a uniform embedding iff $\{h_+[B] \mid B \in \mathcal{B}\}$ is a basis of $\mathcal{A}$.*

PROOF. Let h be a uniform embedding. For the right adjoint h_+ of h we have $h(h_+h(x) \wedge x^*) = h(x) \wedge h(x^*) = 0$ and since h is dense, $h_+h(x) \wedge x^* = 0$, that is, $h_+h(x) \leqslant x^{**}$. For $A \in \mathcal{A}$ and $C = \{c \mid c \lhd a \in A\}$ $(\in \mathcal{A})$ we have $C \leqslant h_+h[C] \leqslant \{c^{**} \mid c \lhd a \in A\} \leqslant A$ by 12.1.1.4. Thus, $\{h_+[B] \mid B \in \mathcal{B}\}$ is a basis of $\mathcal{A}$. On the other hand, if $\{h_+[B] \mid B \in \mathcal{B}\}$ is a basis of $\mathcal{A}$, each $B \in \mathcal{B}$ is $h[h_+[B]]$ with $h_+[B] \in \mathcal{A}$. □

12.7. *Metric and metrizable frames.* Imitating the well-known metrizability theorem for uniform spaces, *metrizable uniform frames* were defined in [64] as the $(L,\mathcal{A})$ in which $\mathcal{A}$ has a countable basis. In fact, in the point-free context one can define a metric structure that justifies this definition. A *diameter* on a frame L is a mapping $d: L \to \mathbb{R}_+$ (non-negative reals) such that

$$\begin{aligned} &d(0) = 0,\ a \leqslant b \quad \Rightarrow \quad d(a) \leqslant d(b), \quad \text{and} \\ &a \wedge b \neq 0 \quad \Rightarrow \quad d(a \vee b) \leqslant d(a) + d(b). \end{aligned}$$

If we set $\mathcal{A}(d) = \{\{a \mid d(a) < \varepsilon\} \mid \varepsilon > 0\}$ we obtain a system satisfying (U2′) and (U3). We say that a *diameter d is admissible* if $\mathcal{A}(d)$ is admissible. We have

PROPOSITION [119]. *$(L,\mathcal{A})$ has a countable basis of the uniformity iff L admits a diameter such that $\mathcal{A}(d)$ is a basis of $\mathcal{A}$.*

One can assume that the diameter d has, furthermore, some useful additional properties making the structure of (L,d) even closer to that of a metric space; also, one has metrizability theorems for (plain) frames quite analogous to classical metrizability theorems for topological spaces (Bing, Moore, and other criteria, see [30]).

The diameter structure allows to study, in the point-free context, metric-based topics like contractions, Lipschitz property, etc. [120,122].

13. Completeness and completion

13.1. A uniform frame $(L, \mathcal{A})$ is said to be *complete* if every dense uniform embedding $(M, \mathcal{B}) \to (L, \mathcal{A})$ is an isomorphism in **UniFrm**.

NOTE. In other words, $(L, \mathcal{A})$ is complete if every uniform embedding $(M, \mathcal{B}) \to (L, \mathcal{A})$ is closed. Thus, the notion is a counterpart of the standard classical completeness.

13.1.1. From 12.4 we immediately see that

a regular compact frame (with the unique uniformity) is complete.

13.2. *Regular Cauchy mappings.* A mapping $\varphi : (L, \mathcal{A}) \to (M, \mathcal{B})$ preserving 0 and finite meets is said to be *Cauchy* if $\varphi[C]$ is a cover for each $C \in \mathcal{A}$. It is said to be *regular* if for each $a \in L$

$$\varphi(a) = \bigvee \{\varphi(x) \mid x \lhd a\}.$$

LEMMA. *Let $h : (M, \mathcal{B}) \to (L, \mathcal{A})$ be a dense uniform embedding. Then the right Galois adjoint $h_+ : (M, \mathcal{B}) \to (L, \mathcal{A})$ is a regular Cauchy mapping.*

PROOF. By density, $h_+(0) = 0$, and as h_+ is a right adjoint it preserves all meets. By 12.6, each $h_+[B]$, $B \in \mathcal{B}$, is a cover. Thus, it remains to be proved that $h_+(a) = \bigvee\{h_+(x) \mid x \lhd a\}$. We have

$$z \lhd h_+(a) \quad \Rightarrow \quad h(z) \lhd a$$

(indeed, $h(z) \lhd hh_+(a) \leqslant a$ by 12.5.2). Thus,

$$\begin{aligned} h_+(a) &= \bigvee \{z \mid z \lhd h_+(a)\} \leqslant \bigvee \{z \mid h(z) \lhd a\} \\ &\leqslant \bigvee \{h_+h(z) \mid h(z) \lhd a\} \leqslant \bigvee \{h_+(x) \mid x \lhd a\} \leqslant h_+(a). \end{aligned}$$ □

13.3. *Construction.* Let $(L, \mathcal{A})$ be a uniform frame. For $a \in L$ set $\mathfrak{k}(a) = \{x \mid x \lhd a\}$. Take the down-set construction from 3.4 and regard, for a moment, the L as the underlying meet-semilattice. On the frame $\mathfrak{D}L$ consider the relation R consisting of all the couples

$$(\downarrow a, \mathfrak{k}(a)),\ a \in L \quad \text{and} \quad (L, \downarrow C),\ C \in \mathcal{A}$$

and set

$$\mathbf{C}(L, \mathcal{A}) = \mathfrak{D}L/R.$$

It is easy to see that $U \in \mathfrak{D}L$ is saturated iff
(R1) $\mathfrak{k}(a) \subseteq U \ \Rightarrow\ a \in U$, and
(R2) $\{a\} \wedge C \subseteq U$ for a $C \in \mathcal{A} \ \Rightarrow\ a \in U$.

13.3.1. LEMMA. *The homomorphism ε_L from* 3.4.1 *equalizes the relation R. Consequently we have a frame homomorphism $v_{(L,\mathcal{A})}:\mathbf{C}(L,\mathcal{A})\to L$ defined by $v_{(L,\mathcal{A})}(U)=\bigvee U$.*

PROOF. $\varepsilon(\mathfrak{k}(a))=\bigvee\{x\mid x\lhd a\}=a=\varepsilon(\downarrow a)$ and $\varepsilon(\downarrow C)=\bigvee C=1=\bigvee L$. $\square$

13.3.2. LEMMA. (1) *Each $\downarrow b$ is in $\mathbf{C}(L,\mathcal{A})$.*
(2) *Each $C^{\downarrow}=\{\downarrow c\mid c\in C\}$ with $C\in\mathcal{A}$ is a cover of $\mathbf{C}(L,\mathcal{A})$.*

PROOF. (1) is trivial. (2): $\bigvee\{\downarrow c\mid c\in C\}\supseteq\bigcup\{\downarrow c\mid c\in C\}=\downarrow C$. Apply (R2). $\square$

13.3.3. LEMMA. *The system $\mathcal{A}^{\downarrow}=\{C^{\downarrow}\mid C\in\mathcal{A}\}$ is an admissible system of covers.*

PROOF. We have $\downarrow c\cap\downarrow a\neq\{0\}$ iff $c\wedge a\neq 0$ and hence if $Ca\leqslant b$ then $C^{\downarrow}\downarrow a\subseteq\downarrow b$ and we have the implication

$$a\lhd_{\mathcal{A}}b\quad\Rightarrow\quad\downarrow a\lhd_{\mathcal{A}^{\downarrow}}\downarrow b$$

which together with the obvious $U=\bigcup\{\downarrow a\mid a\in U\}=\bigvee\{\downarrow a\mid a\in U\}$ yields the statement. $\square$

CONVENTION. Since obviously $C^{\downarrow}\wedge D^{\downarrow}=(C\wedge D)^{\downarrow}$ and $C^{\downarrow}C^{\downarrow}\leqslant(CC)^{\downarrow}$, the system $\mathcal{A}^{\downarrow}$ is a basis of uniformity on $\mathbf{C}(L,\mathcal{A})$. From now on, $\mathbf{C}(L,\mathcal{A})$ will be considered as the resulting uniform frame.

13.3.4. PROPOSITION. *The homomorphism $v_{(L,\mathcal{A})}:\mathbf{C}(L,\mathcal{A})\to(L,\mathcal{A})$ is a dense uniform embedding and $\lambda_L=(x\mapsto\downarrow x):(L,\mathcal{A})\to\mathbf{C}(L,\mathcal{A})$ is a regular Cauchy mapping.*

PROOF. We already know that $v_{(L,\mathcal{A})}$ is a frame homomorphism; it is obviously onto and dense. Also obviously, λ_L is the right Galois adjoint of $v_{(L,\mathcal{A})}$. By 12.6, since $C^{\downarrow}=\lambda[C]$, $v_{(L,\mathcal{A})}$ is a dense uniform embedding, and λ_L is regular Cauchy by 13.2. $\square$

13.4. $\mathbf{C}(L,\mathcal{A})$ *is complete. The completion functor.*

13.4.1. LEMMA. *Let $\varphi:(L,\mathcal{A})\to(M,\mathcal{B})$ be a regular Cauchy mapping. Then there is a frame homomorphism $f:\mathbf{C}(L,\mathcal{A})\to M$ such that $f(U)=\bigvee\varphi[U]$ for all $U\in\mathbf{C}(L,\mathcal{A})$* (*and consequently $f\cdot\lambda_L=\varphi$*).

PROOF. Consider the frame homomorphism $h:\mathfrak{D}L\to M$ from 3.4.2, $h(\downarrow a)=\varphi(a)$. Then $h(U)=\bigvee\varphi[U]$. We will prove that this h equalizes R and hence, by 5.6.2, restricts to the desired f. We have $h(\mathfrak{k}(a))=\bigvee\varphi[\mathfrak{k}(a)]=\bigvee\{\varphi(x)\mid x\lhd a\}=\varphi(a)=h(\downarrow a)$ and $h(\downarrow C)=h(\bigcup\{\downarrow c\mid c\in C\}=\bigvee\{h(\downarrow c)\mid c\in C\}=\bigvee\varphi[C]=1$. $\square$

13.4.2. LEMMA. *Let $h:(M,\mathcal{B})\to(L,\mathcal{A})$ be a dense uniform embedding. Then there is a dense uniform embedding $f:\mathbf{C}(L,\mathcal{A})\to(M,\mathcal{B})$ such that $hf=v_{(L,\mathcal{A})}$.*

PROOF. By 13.4.1 and 13.2 we have a frame homomorphism $f:\mathbf{C}(L,\mathcal{A})\to M$ such that $f(U)=\bigvee h_+[U]$. Then $hf(U)=h(\bigvee h_+[U])=\bigvee hh_+[U]=\bigvee U=v(U)$. Now since $\{h_+[C]\mid C\in\mathcal{A}\}$ is admissible, $h_+[L]\supseteq\bigcup\{h_+[C]\mid C\in\mathcal{A}\}$ generates M (recall 12.6) and since $f[\mathbf{C}(L,\mathcal{A})]\supseteq f[\lambda[L]]=h_+[L]$ we have $f[\mathbf{C}(L,\mathcal{A})]=M$.

If $f(U)=\bigvee h_+[U]=0$ we have $h_+(u)=0$ for all $u\in U$ and since $hh_+(u)=u$, $U=\{0\}$.

Finally, from $h\cdot f=v$ we obtain for the right adjoints that $f_+h_+=\lambda$. Thus, $f_+[h_+[C]]=C^{\downarrow}$ and since the covers $h_+[C]$ generate $\mathcal{B}$ we see that f is a uniform embedding by 12.6. □

13.4.3. PROPOSITION. $\mathbf{C}(L,\mathcal{A})$ *is complete.*

PROOF. Let $g:(M,\mathcal{B})\to\mathbf{C}(L,\mathcal{A})$ be a dense uniform embedding. Set $h=vg$ and apply 4.2. We have a dense surjection f such that $hf=vgf=v$ and since v is dense and hence a monomorphism (recall 12.5.1), $gf=\mathrm{id}$. Using the density of g we obtain that also $fg=\mathrm{id}$. □

13.4.4. For a Cauchy mapping $\varphi:(L,\mathcal{A})\to(M,\mathcal{B})$ set

$$\varphi^\circ(a)=\bigvee\bigl\{\varphi(x)\mid x\lhd a\bigr\}.$$

LEMMA. (1) *If* $\varphi,\psi:(L,\mathcal{A})\to(M,\mathcal{B})$ *are Cauchy mappings and if* $\psi\leqslant\varphi$ *then* $\varphi^\circ\leqslant\psi$.

(2) φ° *is a regular Cauchy mapping.*

PROOF. (1): If $x\lhd a$ in L then $Cx\leqslant a$ for a $C\in\mathcal{A}$ and we have $C\leqslant\{a,x^*\}$. Thus, $\{a,x^*\}\in\mathcal{A}$ and hence $\psi(a)\vee\psi(x^*)=1$. Since $\varphi(x)\wedge\psi(x^*)\leqslant\varphi(x)\wedge\varphi(x^*)=0$ we have $\varphi(x)=\varphi(x)\wedge(\psi(x^*)\vee\psi(a))\leqslant\psi(a)$ and conclude that $\varphi^\circ(a)\leqslant\psi(a)$.

(2): Take a $D\in\mathcal{A}$ such that $D\leqslant\{x\mid x\lhd c\in C\}$ (for instance, a D such that $DD\leqslant C$). Then

$$\bigvee\varphi^\circ[C]\geqslant\bigvee\varphi[D]=1.$$

Obviously $\varphi^\circ(0)=0$ and by distributivity

$$\begin{aligned}\varphi^\circ(a)\wedge\varphi^\circ(b)&=\bigvee\bigl\{\varphi(x)\mid x\lhd a\bigr\}\wedge\bigvee\bigl\{\varphi(y)\mid y\lhd b\bigr\}\\&=\bigvee\bigl\{\varphi(x\wedge y)\mid x\lhd a,\,y\lhd b\bigr\}\\&\leqslant\bigvee\bigl\{\varphi(z)\mid z\lhd a\wedge b\bigr\}=\varphi^\circ(a\wedge b)\leqslant\varphi^\circ(a)\wedge\varphi^\circ(b).\end{aligned}$$ □

13.4.5. For a uniform frame homomorphism $h:(L,\mathcal{A})\to(M,\mathcal{B})$ consider, first $\varphi=\lambda_{(M,\mathcal{B})}h$ and then define $\mathbf{C}h:\mathbf{C}(L,\mathcal{A})\to\mathbf{C}(M,\mathcal{B})$ as the frame homomorphism such that $\mathbf{C}h\cdot\lambda=\varphi^\circ$ from 13.4.1 and 13.4.4.

THEOREM. $\mathbf{C}: \mathbf{UniFrm} \to \mathbf{UniFrm}$ *is a functor and* $v = (v_{(L,\mathcal{A})})_{(L,\mathcal{A})}: \mathbf{C} \to \mathrm{Id}$ *is a natural transformation constituting a reflection of* **UniFrm** *onto the full subcategory of complete uniform frames.*

PROOF. We have $\mathbf{C}h[C^{\downarrow}] = \varphi^{\circ}[C] = \{\bigvee\{\downarrow h(x) \mid x \lhd c\} \mid c \in C\}$; if we take a $D \in \mathcal{A}$ such that $D \leqslant \{x \mid x \lhd c \in C\}$ (for instance, if $D \cdot D \leqslant C$), the last is refined by $h[D]^{\downarrow}$. Thus, $\mathbf{C}h$ is a uniform homomorphism.

We have $v_{(M,\mathcal{B})}\mathbf{C}h(\downarrow a) = v\varphi^{\circ}(a) = \bigvee\{v\lambda h(x) \mid x \lhd a\} = \bigvee\{h(x) \mid x \lhd a\} = h(a) = hv_{(L,\mathcal{A})}(\downarrow a)$. Since the $\downarrow a$'s generate $\mathbf{C}(L, \mathcal{A})$, we see that

$$v_{(M,\mathcal{B})} \cdot \mathbf{C}h = h \cdot v_{(L,\mathcal{A})}.$$

Since $v_{(M,\mathcal{B})}$ is dense and hence a monomorphism, and hence there is only one g satisfying $v_{(M,\mathcal{B})} \cdot g = h \cdot v_{(L,\mathcal{A})}$, we now easily infer that $\mathbf{C}$ is functorial and $v: \mathbf{C} \to \mathrm{Id}$ is a transformation. □

13.5. *An easy completeness criterion.* Since a frame admitting a uniformity is regular, $v_{(L,\mathcal{A})}$ is an isomorphism iff it is codense. Thus,

$$(L, \mathcal{A}) \text{ is complete} \quad \text{iff} \quad (U \text{ saturated and } \bigvee U = 1) \ \Rightarrow\ U = L.$$

We can do better, though:

PROPOSITION. $(L, \mathcal{A})$ *is complete iff* L *is the only* $U \in \mathfrak{D}L$ *satisfying* (R2) *and* $\bigvee U = 1$.

PROOF. Let U satisfy (R2). Then $\widetilde{U} = \{a \mid \mathfrak{k}(a) \subseteq U\}$ satisfies (R2) as well: Let $\{a\} \wedge C \subseteq \widetilde{U}$. Choose $D \in \mathcal{A}$ such that $DD \leqslant C$. Fix an $x \in \mathfrak{k}(a)$. For $d \in D$ we have a $c \in C$ such that $Dd \leqslant c$, and hence $x \wedge d \in \mathfrak{k}(a \wedge c) \subseteq U$, as $a \wedge c \in \widetilde{U}$. Thus, $\{x\} \wedge D \subseteq U$ and hence $x \in U$. Hence $\mathfrak{k}(a) \subseteq U$ and $a \in \widetilde{U}$.

Furthermore, $\widetilde{U}$ also satisfies (R1) (by the interpolativity, $\mathfrak{k}(a) = \bigcup\{\mathfrak{k}(b) \mid b \lhd a\}$) and $\bigvee \widetilde{U} \geqslant \bigvee U = 1$. Thus, $1 \in \widetilde{U}$, that is, $1 \in \mathfrak{k}(1) \subseteq U$, and hence $U = L$. □

13.6. *More generally: Completion of nearness frames.* The construction of $\mathbf{C}(L, \mathcal{A})$ works for a general nearness frame ($\mathcal{A}^{\downarrow}$ is a nearness resp. strong nearness resp. uniformity in accordance with what the $\mathcal{A}$ was), and results in a nearness frame complete in the sense of 13.1. In the case of strong nearnesses, also the construction of $\mathbf{C}h$ works well (the reader may check that we were careful to use just the strong nearness property in the proofs), and Theorem 13.4.5 can be extended accordingly. On the other hand, for general nearness frames one has uniform homomorphisms h without a suitable $\mathbf{C}h$, that is, such that there is no g with $vg = hv$.

In particular, the criterion 13.5 also holds generally and we obtain

COROLLARY. *Each regular frame admits a* complete *strong nearness* (*namely the system of all covers*).

(Indeed, if U is saturated and $\bigvee U = 1$ then U is a cover and $\{1\} \wedge U \subseteq U$ and hence $1 \in U$.)

On the other hand, a completely regular frame does not necessarily admit a complete uniformity – see Section 15.

13.7. *Cauchy completion.* Imitating the classical completion of spaces one can construct for a uniform frame the uniform space $\Psi(X, \mathcal{A})$ of *Cauchy points* of $(L, \mathcal{A})$ ($\equiv$ regular Cauchy filters, that is, filters F in L such that

(1) for each $a \in F$ there is a $b \lhd_{\mathcal{A}} a$, $b \in F$, and

(2) for each $C \in \mathcal{A}$, $F \cap C \neq \emptyset$.

In the case of a uniform space, $\Psi(\Omega(X), \mathcal{A})$ coincides with the standard classical completion. The following facts can be of interest:

- in general, even for a spatial $(L, \mathcal{A})$, $\Omega\Psi(L, \mathcal{A})$ is not necessarily isomorphic to $\mathbf{C}(L, \mathcal{A})$; thus, a uniform space can be complete in the classical sense without being complete as a uniform locale;
- but if $(L, \mathcal{A})$ is metrizable (that is, if $\mathcal{A}$ is countably generated), $\Omega\Psi(L, \mathcal{A})$ is isomorphic to $\mathbf{C}(L, \mathcal{A})$;
- thus in particular, the completion of a metrizable frame $(L, \mathcal{A})$ is always spatial, even if L is not.

(See, e.g., [20,22,64].)

14. Paracompactness. I. Similar like in spaces

14.1. A subset X of a frame L is said to be *locally finite* (resp. *discrete*) *witnessed by a cover W* if for each $w \in W$ there are only finitely many (resp. there is at most one) $x \in X$ such that $x \wedge w \neq 0$.

$X \subseteq L$ is *σ-locally finite* (resp. *σ-discrete*) if it is a union of countably many locally finite (resp. discrete) sets.

14.2. A frame L is said to be *paracompact* if it is regular and if each cover of L has a locally finite refinement.

14.2.1. THEOREM. *Let L be a regular frame. Then the following statements are equivalent*:

(1) *L is paracompact,*

(2) *each cover of L has a σ-locally finite refinement,*

(3) *each cover of L has a σ-discrete refinement,*

(4) *L is fully normal.*

PROOF. We will add one more statement:

$$\text{For each cover } A \text{ of } L \text{ there is a locally finite } X \subseteq L \text{ such that } X \leqslant A \text{ and } \left(\bigvee X\right)^* = 0. \qquad (*)$$

I. (2) $\Rightarrow$ $(*)$: Let A be a cover and let $B = \bigcup_{n=1}^{\infty} B_n$ refine A, with locally finite B_n witnessed by W_n. We can assume that $B_n \subseteq B_{n+1}$. For $b \in B$ set $n(b) = \min\{n \mid b \in B_n\}$ and choose an antireflexive R well-ordering B such that $n(c) < n(b) \Rightarrow cRb$. Further set $V_n = W_n \wedge B_n$ and set $V = \bigcup V_n$. Since $\bigvee V_n = \bigvee W_n \wedge \bigvee B_n = \bigvee B_n$, we have $\bigvee V = \bigvee B = 1$. Finally, define for $b \in B$

$$\tilde{b} = b \wedge \left(\bigvee\{c \mid cRb\}\right)^*$$

and set $X = \{\tilde{b} \mid b \in B\}$. Obviously $X \leqslant A$.

If $0 \neq x \in L$ then $x \wedge b \neq 0$ for some $b \in B$. Consider the first b, in the order R, with this property. Then $x \leqslant (\bigvee\{c \mid cRb\})^*$ and hence $x \wedge \tilde{b} \neq 0$. Thus, $(\bigvee X)^* = 0$.

X is locally finite witnessed by V: Take a $v \in V$, say, $v = w \wedge c$ with $w \in W_n$ and $c \in B_n$. Whenever $n(b) > n$ $(\geqslant n(c))$, we have cRb and hence $v \leqslant \bigvee\{u \mid uRb\}$ and $v \wedge \tilde{b} = 0$. Thus, if $v \wedge \tilde{b} \neq 0$ then $b \in B_n$ and as $v \leqslant w \in W_n$ there are only finitely many such b.

II. $(*) \Rightarrow$ (4): For a cover B define $B^{\times} = \{\bigvee S \mid S \subseteq B,\ \bigwedge S \neq 0\}$. Since obviously $BB \leqslant (B^{\times})^{\times}$ it suffices to prove that for each cover A there is a cover B such that $B^{\times} \leqslant A$.

Let $X \leqslant \{c \mid c \prec a \in A\}$ such that $(\bigvee X)^* = 0$ be locally finite witnessed by W. For $x \in X$ choose $a(x) \in A$ such that $x \prec a(x)$, and define

$$B = \big\{b \in L \mid \forall x \in X\ (b \leqslant a(x) \text{ or } b \leqslant x^*)\big\}.$$

Take a $w \in W$ and consider the system $x_1, \ldots, x_n$ of all the elements of X such that $x \wedge w \neq 0$. We have

$$1 = \bigwedge_{i=1}^{n} \big(a(x_i) \vee v_i^*\big) = \bigvee \left\{ \bigwedge_{i=1}^{n} \xi_i \mid \xi \in \big\{a(x_1), x_1^*\big\} \times \cdots \times \big\{a(x_n), x_n^*\big\} \right\} = s.$$

Now since $w \leqslant x^*$ for $x \in X\backslash\{x_1, \ldots, x_n\}$ we have $w \wedge \bigvee B = w \wedge \bigvee\{b \in L \mid \forall i\ (b \leqslant a(x_i) \text{ or } b \leqslant x_i^*)\} = w \wedge s = w$, and since W is a cover we infer that $\bigvee B = 1$.

Finally let $S \subseteq B$, $\bigwedge S \neq 0$. Then there is an $x \in X$ such that $x \wedge \bigwedge S \neq 0$ and hence $s \neq x^*$ for all $s \in S$. Consequently $s \leqslant a(x)$ for all $s \in S$, $\bigvee S \leqslant a(x)$, and we conclude that $B^{\times} \leqslant A$.

III. (4) $\Rightarrow$ (3): For a cover A choose A_n such that $A_0 = A$ and $A_n A_n \leqslant A_{n-1}$. Let A_1 be well ordered by an antireflexive R; write $b\overline{R}a$ for (bRa or $b = a$). For an $a \in A_1$ set

$$a_1 = a,\ a_{n+1} = A_{n+1} a_n, \quad \text{and} \quad \varphi_n(a) = \bigvee\{b_n \mid bRa\}.$$

Note that $A_{n+1}\varphi_n(a) = \varphi_{n+1}(a)$ so that $\varphi_n(a) \prec \varphi_{n+1}(a)$. Set $\widetilde{A}_n = \{\tilde{a}_n = a_n \wedge \varphi_{n+1}(a)^* \mid a \in A_1\}$ and $\widetilde{A} = \bigcup \widetilde{A}_n$. We have

$$\bigvee_{n=1}^{\infty} \left(\bigvee\{\tilde{b}_n \mid b\overline{R}a\}\right) = \bigvee_{n=1}^{\infty} \left(\bigvee\{b_n \mid b\overline{R}a\}\right). \tag{14.2.1}$$

Indeed: It obviously holds for the a first in R. Now if it holds for all cRa we have

$$\begin{aligned}\bigvee_{n=1}^{\infty}\bigvee\{\tilde{b}_n \mid b\overline{R}a\} &= \bigvee_{n=1}^{\infty}\Big(\bigvee_{cRa}\bigvee\{\tilde{b}_n \mid b\overline{R}c\}\vee\tilde{a}_n\Big)\\ &= \bigvee_{n=1}^{\infty}\Big(\bigvee\{\tilde{b}_n \mid bRa\}\vee\tilde{a}_n\Big)\\ &= \bigvee_n\big(\varphi_n(a)\vee\big(a_n\wedge\varphi_{n+1}(a)^*\big)\big)\\ &= \bigvee_n\big(\varphi_{n+2}(a)\vee\big(a_n\wedge\varphi_{n+1}(a)^*\big)\big)\\ &= \bigvee_n\big(\varphi_{n+2}(a)\vee a_n\big) = \bigvee_n\big(\varphi_n(a)\vee a_n\big)\\ &= \bigvee_n\Big(\bigvee\{b_n \mid b\overline{R}a\}\Big).\end{aligned}$$

From (14.2.1) we immediately infer that $\bigvee\widetilde{A}\geqslant\bigvee A_1=1$. Thus, $\widetilde{A}$ is a cover. For an $a\in A_1$ we have a $b\in A$ such that $A_1a=A_1a_1\leqslant b$. Since $A_{n+1}a_{n+1}=A_{n+1}(A_{n+1}a_n)\leqslant(A_{n+1}A_{n+1})a_n\leqslant A_na_n$ we see that $A_na_n\leqslant b$ for all n and hence $\tilde{a}_n\leqslant a_n\leqslant b$. Thus, $\widetilde{A}$ refines A.

Finally, $\widetilde{A}_n$ is discrete witnessed by A_{n+1}: For $x\in A_{n+1}$ take the $a\in A_1$ first in R such that $x\wedge\tilde{a}_n\neq 0$. If aRb we have $x\leqslant A_{n+1}a_n=a_{n+1}\leqslant\varphi_{n+1}(b)$ so that $x\wedge\varphi_{n+1}(b)^*=0$ and $x\wedge\tilde{b}_n=0$.

IV. Since (3) $\Rightarrow$ (2) is trivial we have already closed the cycle (4) $\Rightarrow$ (3) $\Rightarrow$ (2) $\Rightarrow$ $(*)\Rightarrow$ (4). Since also (1) $\Rightarrow$ (2) is trivial we can conclude the proof by showing that (4)&$(*)\Rightarrow$ (1). For a cover A of L choose a cover B such that $BB\leqslant A$, and for B choose an $X\subseteq L$ as in $(*)$, locally finite witnessed by W. Further, let V be a cover such that $VV\leqslant W$. The set $D=\{(V\wedge B)x \mid x\in X\}$ is obviously a cover, and since $V\wedge B\leqslant B$ and $X\leqslant B$ it refines A. Finally, D is locally finite witnessed by V: If $v\wedge(V\wedge B)x\neq 0$ then obviously $Vv\wedge x\neq 0$. Choose a $w\in W$ such that $Vv\leqslant w$ Then $w\wedge x\neq 0$, and there are only finitely many such $x\in X$. □

14.2.1. From the equivalence (1) $\equiv$ (4), 8.7.2 and 12.3.1 we immediately obtain

COROLLARY. *A paracompact frame is completely regular.*

14.2.2. Taking into account (2) (or (3)) we infer

COROLLARY. *Each Lindelöf frame is paracompact.*

14.3. *Paracompactness of metrizable frames*

14.3.1. LEMMA. *If $\mathcal{A}$ is an admissible system of covers then also $\{AA \mid A\in\mathcal{A}\}$ is admissible.*

PROOF. Since $(AA)x = A(A(Ax))$ (recall 12.1) and since if $A_1 y \leqslant y_1$, $A_2 y_1 \leqslant y_2$ and $A_3 y_2 \leqslant x$ we have $A(A(Ay)) \leqslant x$ for $A = A_1 \wedge A_2 \wedge A_3$, we have $x = \bigvee\{y \mid \exists y_1, y_2,\ y \lhd_{\mathcal{A}} y_1 \lhd_{\mathcal{A}} y_2 \lhd_{\mathcal{A}} x\} = \bigvee\{y \mid \exists A \in \mathcal{A},\ (AA)y \leqslant x\}$. $\square$

14.3.2. For a cover A of L and $x \in L$ set

$$b(A, x) = \big\{a \in A \mid a \wedge \big(x/(AA)(AA)\big) \neq 0\big\}.$$

We have

LEMMA. $\bigvee b(A, x) = A(x/(AA)(AA))$ *and* $A(\bigvee b(A, x)) \leqslant x/(AA)$.

PROOF. The equality is trivial. Further, $A(\bigvee b(A, x)) = A(A(x/(AA)(AA))) \leqslant (AA)(x/(AA)(AA)) \leqslant (AA)((x/(AA))(AA)) \leqslant x/(AA)$. $\square$

14.3.3. THEOREM. *Each frame admitting a countable system of covers is paracompact.*

PROOF. Let L admit $\{A'_n \mid n = 1, 2, \ldots\}$. Then it admits $\{A_n = A'_1 \wedge \cdots \wedge A'_n \mid n = 1, 2, \ldots\}$ and we have $A_1 \geqslant A_2 \geqslant \cdots$. We will prove that L is fully normal.

Let C be a general cover. Set $B = \bigcup\{b(A_n, c) \mid c \in C,\ n = 1, 2, \ldots\}$.

B is a cover: By 14.3.1,

$$\begin{aligned} \bigvee B &= \bigvee_{c \in C} \bigvee_n \bigvee b(A_n, c) = \bigvee_{c \in C} \bigvee_n A_n\big(c/(A_n A_n)(A_n A_n)\big) \\ &\geqslant \bigvee_{c \in C} \bigvee_n \big(c/(A_n A_n)(A_n A_n)\big) = \bigvee_{c \in C} c = 1. \end{aligned}$$

$BB \leqslant C$: Take an $a \in B$, say, $a \in b(A_k, c)$. Set

$$n = \min\big\{i \mid \exists x \in C\ \exists y \in b(A_i, x) \text{ such that } y \wedge a \neq 0\big\}.$$

This is correct (since $a \in b(A_k, c)$ the set is non-void), and $n \leqslant k$. Choose $c_0 \in C$ and $y_0 \in b(A_n, c_0)$ such that $y_0 \wedge a \neq 0$. Since $A_k \leqslant A_n$ we have a $b \in A_n$ such that $a \leqslant b$, and hence $y_0 \wedge b \neq 0$ and $b \leqslant A_n y_0$. Now if $u \wedge a \neq 0$ for some $u \in B$, say $u \in b(A_i, d)$, we have $A_i \leqslant A_n$ and hence there is a $v \in A_n$ such that $u \leqslant v$. Then $v \wedge b \geqslant u \wedge a \neq 0$, and since $b \leqslant A_n y_0$ we obtain, using 14.3.2,

$$\begin{aligned} u \leqslant v \leqslant A_n(A_n y_0) &\leqslant A_n\Big(A_n\Big(\bigvee b_n(A_n, c_0)\Big)\Big) \\ &\leqslant A_n(c_0/A_n A_n) \leqslant (A_n A_n)(c_0/A_n A_n) \leqslant c_0 \end{aligned}$$

so that $Ba \leqslant c_0$. $\square$

NOTE. Thus, every metrizable frame (recall 12.7) is paracompact. In fact, the assumption in the theorem above is equivalent to metrizability. We can assume $A_1 \geqslant A_2 \geqslant \cdots$ (see the proof) and then we can set $B_1 = A_1$ and if B_n is already chosen, we can choose, using full normality, B'_{n+1} such that $B'_{n+1}B'_{n+1} \leqslant B_n$ and set $B_{n+1} = B'_{n+1} \wedge A_{n+1}$. Then $\{B_1, B_2, \dots\}$ is an admissible basis of uniformity.

15. Paracompactness. II. Different (and better) than in spaces

15.1. The trouble with paracompact spaces (which play an important rôle in applications) is that they do not behave very well under standard constructions. For instance, even a product of a paracompact space with a metric one is not necessarily paracompact.

The point-free situation is entirely different. The category of paracompact frames is coreflective in **CRegFrm** (and hence in **Frm**) so that, in particular, paracompactness is preserved under frame coproducts (localic products).

Also note the elegant characterization of paracompactness in 15.4 below (due to Isbell) from which the coreflection result follows.

15.2. LEMMA. *In any frame L, each normal cover has a normal locally finite refinement witnessed by a normal cover.*

PROOF. Let A be a normal cover and let $\mathcal{A} = \{A = A_1, A_2, \dots\}$ be such that $A_n A_n \leqslant A_{n-1}$. Then $\lhd = \lhd_{\mathcal{A}}$ is a strong inclusion. Recall 8.5 and consider the subframe $M = \{a \mid a = \bigvee\{x \mid x \lhd a\}\}$. Since each A_n has a strong refinement, namely A_{n+1}, there exist B_n covers of M such that $B_n \leqslant A_n$. Set $\mathcal{B} = \{B_1, B_2, \dots\}$. Since obviously $x \lhd y \Rightarrow x \lhd_{\mathcal{B}} y$, $\mathcal{B}$ is admissible, and by 14.3 M is paracompact. Thus, each cover of M is normal and has a locally finite refinement. Since every cover of M is also a cover of L the statement follows. $\square$

15.3. LEMMA. *Let A be a locally finite normal cover and let C_a, $a \in A$, be normal covers. Then $C = \bigcup\{\{a\} \wedge C_a \mid a \in A\}$ is a normal cover.*

PROOF. If A' refines A choose for $b \in A'$ an $a \in A$ with $a \geqslant b$, and set $C'_b = C_a$. Then $\bigcup\{\{b\} \wedge C'_b \mid b \in A'\}$ refines $\bigcup\{\{a\} \wedge C_a \mid a \in A\}$ and consequently we can assume, using 15.2, that A is locally finite witnessed by a normal W. By 15.2 again there is a locally finite B such that

$$BB \leqslant A \quad \text{and} \quad B \leqslant W.$$

For $a \in A$ choose normal covers U_a such that $U_a U_a \leqslant C_a$. Since $B \leqslant W$, the system $\{a \in A \mid a \wedge b \neq 0\}$ is finite for any $b \in B$ and hence we can define

$$D_b = \bigwedge\{U_a \mid a \in A,\ a \wedge b \neq 0\} \quad \text{and} \quad D = \bigcup\big\{\{b\} \wedge D_b \mid b \in B\big\}.$$

We will show that $DD \leqslant C$ which will finish the proof since the structure of D allows to proceed by induction. Fix a $b \in B$ and a $d \in D_k$ and consider a general $x \in B$ and $y \in D_x$ such that $(b \wedge d) \wedge (x \wedge y) \neq 0$. If $Bb \leqslant a \in A$ then $x \leqslant a$. Further, $y = y_1 \wedge \cdots \wedge y_n$ for some $y_i \in U_{a_i}$ where $a_1, \dots, a_n$ are the elements of A that meet b. Hence a is one of the a_k and $y \leqslant y_k \in U_a$. By the same reasoning we also have a $d \leqslant d_k \in U_a$. Now take a $c \in C_a$ such that $U_a d_k \leqslant c$. Since $d_k \wedge y_k \geqslant d \wedge y \neq 0$ we have $y_k \leqslant c$ and $x \wedge y \leqslant a \wedge y_k \leqslant a \wedge c$. Thus, $D(b \wedge d) \leqslant a \wedge c$. □

15.4. *Another characterization of paracompact frames*

THEOREM. *A frame L is paracompact iff it admits a complete uniformity.*

PROOF. ($\Rightarrow$) follows from full normality and corollary in 13.6.

($\Leftarrow$): By 13.5.1 we can assume that the fine uniformity $\mathcal{F}$ (that is, the system of all normal covers) of L is complete. Let A be a general cover. Set

$$U = \big\{u \in L \mid \exists B \in \mathcal{F},\ \{u\} \wedge B \leqslant A\big\}.$$

Obviously $U \in \mathfrak{D}L$ and $A \subseteq U$ (since $\{a\} \wedge \{1\} = \{a\}$).

Now let $\{x\} \wedge C \subseteq U$ for some $C \in \mathcal{F}$. By 15.2 we can assume that C is locally finite. For each $c \in C$ we have a normal cover B_c such that $\{x \wedge c\} \wedge B_c \leqslant A$. Now

$$\{x\} \wedge \bigcup \big\{\{c\} \wedge B_b \mid c \in C\big\} = \bigcup \big\{\{x \wedge c\} \wedge B_c \mid c \in C\big\} \leqslant A$$

and since $\bigcup\{\{c\} \wedge B_b \mid c \in C\}$ is normal by 15.3, $x \in U$. Thus, U satisfies (R2) in $(L, \mathcal{F})$. Moreover, $\bigvee U \geqslant \bigvee A$ (as $A \subseteq U$) and hence, by 13.5, $1 \in U$. But this means that there is a normal B such that $B = \{1\} \wedge B \leqslant A$, making A itself normal. □

15.5. *Paracompact coreflection.* Denote by $\mathcal{F}(L)$ the fine uniformity on L. The image of a normal cover under any frame homomorphism is normal; hence we have

OBSERVATION. *Each frame homomorphism $L \to M$ is a uniform homomorphism $(L, \mathcal{F}(L)) \to (M, \mathcal{F}(M))$.*

15.5.1. For a completely regular frame L define $\mathbf{P}(L)$ as the underlying frame of $\mathbf{C}(L, \mathcal{F}(L))$ and

$$\pi_L : \mathbf{P}(L) \to L$$

as the underlying frame homomorphism of $v_{L,\mathcal{F}(L)}$. Finally, denote by

$$\mathbf{P}(h) : \mathbf{P}(L) \to \mathbf{P}(M)$$

the underlying frame homomorphism of $\mathbf{C}(h) : \mathbf{C}(L, \mathcal{F}(L)) \to \mathbf{C}(M, \mathcal{F}(M))$.

THEOREM. *The subcategory* **ParFrm** *of paracompact frames is coreflective in the category* **CRegFrm** *of completely regular frames (with the coreflection functor* **P** *and coreflection transformation $\pi = (\pi_L)_L$), and hence also in* **Frm**.

PROOF. Since $\mathbf{C}(L, \mathcal{F}(L))$is complete (and uniform), $\mathbf{P}(L)$ is paracompact by 15.4. If L is paracompact, $\mathcal{F}(L)$ is the system of all covers and hence $(L, \mathcal{F}(L))$ is complete (13.5) and π_L is an isomorphism. $\square$

Acknowledgement

I am indebted to Marcel Erné for very valuable comments.

References

In the text we have cited only a few papers. Therefore it may be useful to specify the topics of some more of the citations below.

Categorical and algebraic questions: [4–7,10,12–14,16,19,24,28,41,43,48,63,65–67,71,74,80,91,111,114–116, 118,122,125,127,149,154]
Spatiality: [9,59,88,102,116,117,124]
Separation: [3,37,42,60,78,83,98,104,106,128,133,143]
Compactness: [8,18,75,97,126,137,148]
Paracompactness: [23,44,105,131,146]
Symmetric uniformity, nearness, completion: [17,20–22,27,29,44,45,53,54,56,58,85,119,129,154]
Nonsymmetric uniformity: [40,50,51,61,62,112]
Metric frames: [20,22,105,121]
Groups: [32,77,85,156,157]
Booleanization: [25,26]
σ-frames: [14,39,90,93,153,155]

[1] J. Adámek, H. Herrlich and G. Strecker, *Abstract and Concrete Categories: The Joy of Cats*, Pure and Applied Mathematics, John Wiley & Sons (1990).

[2] P.S. Alexandroff, *Zur Begründung der N-dimensionalen mengentheoretischen Topologie*, Math. Ann. **94** (1925), 296–308.

[3] C.E. Aull and W.J. Thron, *Separation axioms between T_0 and T_1*, Indag. Math. **24** (1963), 26–37.

[4] R.N. Ball, *Distributive Cauchy lattices*, Algebra Universalis **18** (1984).

[5] R.N. Ball and A.W. Hager, *On the localic Yoshida representation of an Archimedean lattice ordered group with weak unit*, J. Pure Appl. Algebra **70** (1991), 17–43.

[6] R.N. Ball, T. Hager and A.T. Molitor, *Spaces with filters*, Festschrift on the Occasion of the 60th Birthday of Guillaume Brümmer, Proc. Symp. on Cat. Top., Cape Town, 1994, University of Cape Town (1999).

[7] R.N. Ball and J.L. Walters-Wayland, *C- and C^*-embedded sublocales*, submitted to: Mem. Amer. Math. Soc.

[8] B. Banaschewski, *Frames and compactifications*, Extension Theory of Topological Structures and its Applications, Deutscher Verlag der Wissenschaften (1969), 29–33.

[9] B. Banaschewski, *The duality of distributive continuous lattices*, Canad. J. Math. **32** (1980), 385–394.

[10] B. Banaschewski, *Coherent frames*, Continuous Lattices, Lecture Notes in Math., Vol. 871, Springer, Berlin (1981), 1–11.

[11] B. Banaschewski, *Another look at the localic Tychonoff theorem*, Comment. Math. Univ. Carolinae **29** (1988), 647–656.

[12] B. Banaschewski, *On pushing out frames*, Comment. Math. Univ. Carolinae **31** (1990), 13–21.

[13] B. Banaschewski, *Singly generated frame extensions*, J. Pure Appl. Algebra **83** (1990), 1–21.

[14] B. Banaschewski, *The frame envelope of a σ-frame*, Quaest. Math. **16** (1993), 51–60.

[15] B. Banaschewski, *A uniform view of localic realcompactness*, J. Pure Appl. Algebra **143** (1999), 49–68.

[16] B. Banaschewski and H.S. Hong, *Filters and strict extension of frames*, Kyungpook Math. J. **39** (1999), 215–230.
[17] B. Banaschewski, H.S. Hong and A. Pultr, *On the completion of nearness frames*, Quaest. Math. **21** (1998), 19–37.
[18] B. Banaschewski and C.J. Mulvey, *Stone–Čech compactification of locales. I*, Houston J. Math. **6** (1980), 301–312.
[19] B. Banaschewski and S.B. Niefield, *Projective and supercoherent frames*, J. Pure Appl. Algebra **70** (1991), 45–51.
[20] B. Banaschewski and A. Pultr, *Cauchy points of metric locales*, Canad. J. Math. **41** (5) (1989), 830–854.
[21] B. Banaschewski and A. Pultr, *Samuel compactification and completion of uniform frames*, Math. Proc. Cambridge Philos. Soc. **108** (1990), 63–78.
[22] B. Banaschewski and A. Pultr, *A Stone duality for metric spaces*, Canad. Math. Soc. Conf. Proc. (Int. Summer Categ. Theory Meeting, June 1991) **13** (1992), 33–42.
[23] B. Banaschewski and A. Pultr, *Paracompactness revisited*, Appl. Cat. Structures **1** (1993), 181–190.
[24] B. Banaschewski and A. Pultr, *Variants of openness*, Appl. Cat. Structures **2** (1994), 331–350.
[25] B. Banaschewski and A. Pultr, *Booleanization*, Cahiers Topologie Géom. Différentielle Catégoriques **37** (1996), 41–60.
[26] B. Banaschewski and A. Pultr, *Booleanization of uniform frames*, Comment. Math. Univ. Carolinae **37** (1996), 135–146.
[27] B. Banaschewski and A. Pultr, *Cauchy points of uniform and nearness frames*, Quaest. Math. **19** (1–2) (1996), 101–127.
[28] B. Banaschewski and A. Pultr, *Universal categories of uniform and metric locales*, Algebra Universalis **35** (1996), 556–569.
[29] B. Banaschewski and A. Pultr, *On Cauchy homomorphisms of nearness frames*, Math. Nachr. **183** (1997), 5–18.
[30] B. Banaschewski and A. Pultr, *A new look at pointfree metrization theorems*, Comment. Math. Univ. Carolinae **39** (1) (1998), 167–175.
[31] B. Banaschewski and A. Pultr, *Scott information systems, frames and domains*, Preprint (2000).
[32] B. Banaschewski and J.C.C. Vermeulen, *On the completeness of localic groups*, Comment. Math. Univ. Carolinae **40** (1999), 293–307.
[33] J. Bénabou, *Treillis locaux et paratopologies*, Séminaire Ehresmann, 1re année, exposé 2 (1958).
[34] U. Berni-Canani, F. Borceux, R. Succi-Crucini and G. Van den Bossche, *A theory of quantal sets*, J. Pure Appl. Algebra **62** (1989), 123–136.
[35] N. Bourbaki, *Topologie générale, Les structures fondamentales de l'analyse, Livre III*, Hermann & Cie, Paris.
[36] F. Borceux and G. Van den Bossche, *Quantales and their sheaves*, Order **3** (1986), 61–87.
[37] G. Bruns, *Darstellungen und Erweiterungen geordneter Mengen, II*, J. Math. **210** (1962), 1–23.
[38] C. Caratheodory, *Über die Begrenzung einfach zusamenhängender Gebiete*, Math. Ann. **73** (1913).
[39] M.G. Charalambous, *Dimension theory for σ-frames*, J. London Math. Soc. **8** (1974), 149–160.
[40] D. Doichinov, *Some reflections on quasi-uniform frames*, Bolyai Soc. Math. Studies **4** (1993), 151–158.
[41] C.H. Dowker and D. Papert, *Quotient frames and subspaces*, Proc. London Math. Soc. **16** (1966), 275–296.
[42] C.H. Dowker and D. Strauss, *Separation axioms for frames*, Colloq. Mat. Soc. J. Bolyai **8** (1972), 223–240.
[43] C.H. Dowker and D. Strauss, *Sums in the category of frames*, Houston J. Math. **3** (1977), 7–15.
[44] T. Dube, *Paracompact and locally fine nearness frames*, Topology Appl. **62** (1995), 247–253.
[45] T. Dube, *The Tamano–Dowker type theorems for nearness frames*, J. Pure Appl. Algebra **99** (1995), 1–7.
[46] C. Ehresmann, *Gattungen von lokalen Strukturen*, Jber. Deutsch. Math. Verein **60** (1957), 59–77.
[47] R. Engelking, *General Topology*, Series in Pure Math., Vol. 6, Heldermann Verlag, Berlin (1989).
[48] M. Erné, *Lattice representations for categories of closure spaces*, Categorical Topology, Proc. Conference Toledo, Ohio, 1983, Heldermann Verlag, Berlin (1984), 197–222.
[49] P. Fletcher and W. Hunsaker, *Entourage uniformities for frames*, Monatsh. Math. **112** (1991), 271–279.
[50] P. Fletcher, W. Hunsaker and W. Lindgren, *Characterizations of frame uniformities*, Quaest. Math. **16** (1993), 371–383.
[51] P. Fletcher, W. Hunsaker and W. Lindgren, *Frame quasi-uniformities*, Monatsh. Math. **117** (1994), 223–236.

[52] H. Freundenthal, *Neuaufbau der Endentheorie*, Ann. Math. **43** (1942), 261–279.
[53] J.L. Frith, *Structured frames*, PhD Thesis, Univ. of Cape Town (1987).
[54] A.M. Hager, *Minimal covers of topological spaces*, Papers on General Topology and Related Category Theory, Ann. New York Acad. Sci. **552** (1989), 44–59.
[55] F. Hausdorff, *Grundzüge der Mengenlehre*, Veit & Co., Leipzig (1914).
[56] H. Herrlich, *A concept of nearness*, Gen. Topology Appl. **4** (1974), 191–212.
[57] H. Herrlich, *Topologie II: Uniforme Räume*, Heldermann Verlag, Berlin (1988).
[58] H. Herrlich and A. Pultr, *Nearness, subfitness and sequential regularity*, Appl. Cat. Structures **8** (2000), 67–80.
[59] K.H. Hofmann and J.D. Lawson, *The spectral theory of distributive continuous lattices*, Trans. Amer. Math. Soc. **246** (1978), 285–310.
[60] R.-E. Hoffmann, *On the sobrification remainder* $sX\setminus X$, Pacific J. Math. **83** (1979), 145–156.
[61] W. Hunsaker and J. Picado, *A note on total boundedness*, Acta Math. Hung. **88** (2000), 25–34.
[62] W. Hunsaker and J. Picado, *Frames with transitive structure*, to appear in Appl. Math. Structures.
[63] J.M.E. Hyland, *Function spaces in the category of locales*, Continuous Lattices, Lecture Notes in Math., Vol. 871, Springer, Berlin (1981), 264–281.
[64] J.R. Isbell, *Atomless parts of spaces*, Math. Scand. **31** (1972), 5–32.
[65] J.R. Isbell, *Function spaces and adjoints*, Math. Scand. **36** (1975), 317–339.
[66] J.R. Isbell, *Product spaces in locales*, Proc. Amer. Math. Soc. **81** (1981), 116–118.
[67] J.R. Isbell, *Direct limits of meet-continuous lattices*, J. Pure Appl. Algebra **23** (1982), 33–35.
[68] J.R. Isbell, *First steps in descriptive theory of locales*, Trans. Amer. Math. Soc. **327** (1991), 353–371; ibid. **341** (1994), 462–468.
[69] J. Isbell, I. Kříž, A. Pultr and J. Rosický, *Remarks on localic groups*, Categorical Algebra and its Applications (Louvain-la-Neuve, 1987), Lecture Notes in Math., Vol. 1348, Springer, Berlin (1988), 154–172.
[70] P.T. Johnstone, *Tychonoff Theorem without the axiom of choice*, Fund. Math. **113** (1981), 21–35.
[71] P.T. Johnstone, *Factorization theorems for geometric morphisms II*, Lecture Notes in Math., Vol. 915, Springer, Berlin (1982), 216–233.
[72] P.T. Johnstone, *Stone Spaces*, Cambridge Studies in Advanced Math., No. 3, Cambridge Univ. Press (1983).
[73] P.T. Johnstone, *The point of pointless topology*, Bull. Amer. Math. Soc. (N.S.) **8** (1983), 41–53.
[74] P.T. Johnstone, *Open locales and exponentiation*, Mathematical Applications of Category Theory, Contemporary Math., Vol. 30, Amer. Math. Soc., Providence, RI (1984), 84–116.
[75] P.T. Johnstone, *Wallman compactification of locales*, Houston J. Math. **10** (1984), 201–206.
[76] P.T. Johnstone, *A simple proof that localic groups are closed*, Cahiers de Topologie Géom. Différentielle Catégoriques **29** (1988), 157–161.
[77] P.T. Johnstone, *A constructive closed subgroup theorem*, Cahiers de Topologie Géom. Différentielle Catégoriques **30** (1989), 3–23.
[78] P.T. Johnstone, *Fibrewise separation axioms for locales*, Math. Proc. Cambridge Philos. Soc. **108** (1990), 247–256.
[79] P.T. Johnstone, *The art of pointless thinking: A student's guide to the category of locales*, Category Theory at Work, H. Herrlich and H.-E. Porst, eds, Research and Exposition in Math., Vol. 18, Heldermann Verlag (1991), 85–107.
[80] P.T. Johnstone and S.-H. Sun, *Weak products and Hausdorff locales*, Categorical Algebra and its Applications (Louvain-la-Neuve, 1987), Lecture Notes in Math., Vol. 1348, Springer, Berlin (1988), 154–172.
[81] P.T. Johnstone and S.J. Vickers, *Preframe presentations present*, Proceedings of the 1990 Como Categ. Conference, Lecture Notes in Math., Vol. 1488, Springer, Berlin (1990), 193–212.
[82] A. Joyal and M. Tierney, *An extension of the Galois theory of Grothendieck*, Mem. Amer. Math. Soc., Vol. 51, No. 309 (1984).
[83] T. Kaiser, *A sufficient condition of full normality*, Comment. Math. Univ. Carolinae **37** (1996), 381–389.
[84] J.L. Kelley, *General Topology*, The University Series in Higher Mathematics, Van Nostrand (1955).
[85] I. Kříž, *A direct description of uniform completion in locales and a characterization of LT-groups*, Cahiers de Topologie Géom. Différentielle Catégoriques **27** (1986), 19–34.
[86] I. Kříž and A. Pultr, *Products of locally connected locales*, Proc. 13th Winter School, Suppl. ai Rend. del Circ. Matem. di Palermo, Serie II **11** (1985), 61–70.

[87] I. Křížand A. Pultr, *Peculiar behaviour of connected locales*, Cahiers de Topologie Géom. Différentielle Catégoriques **30** (1989), 25–43.

[88] I. Křížand A. Pultr, *A spatiality criterion and an example of a quasitopology which is not a topology*, Houston J. Math. **15** (2) (1989), 215–234.

[89] S. MacLane, *Categories for the Working Mathematician*, Springer-Verlag, New York (1971).

[90] J.J. Madden, *κ-frames*, J. Pure Appl. Algebra **70** (1991), 107–127.

[91] J.J. Madden and A.T. Molitor, *Epimorphisms of frames*, J. Pure Appl. Algebra **70** (1991), 128–132.

[92] J.J. Madden and J. Vermeer, *Lindelöf locales and realcompactness*, Math. Proc. Cambridge Philos. Soc. **99** (1986), 473–480.

[93] P. Matutu, *Stably continuous σ-frames*, Preprint, Cape Town (1999).

[94] J.C.C. McKinsey and A. Tarski, *The algebra of topology*, Ann. Math. **45** (1944), 141–191.

[95] K. Menger, *Topology without points*, Rice Institute Pamphlet **27** (1940), 80–107.

[96] I. Moerdijk and G.C. Wraith, *Connected locally connected toposes are path-connected*, Trans. Amer. Math. Soc. **295** (1986), 849–859.

[97] A.T. Molitor, *A localic construction of some covers of compact Hausdorff spaces*, Proc. Northeast Conf. on Topology Appl., 1988, Marcel Dekker (1988).

[98] A.T. Molitor, *Normal filters in regular frames*, Preprint (1989).

[99] C.J. Mulvey and J.W. Pelletier, *The dual locale of seminormed space*, Cahiers de Topologie Géom. Différentielle Catégoriques **23** (1982), 73–92.

[100] C.J. Mulvey and J.W. Pelletier, *A globalization of the Hahn–Banach theorem*, Adv. in Math. **89** (1991), 1–59.

[101] C.J. Mulvey and J.W. Pelletier, *A quantisation of the calculus of relations*, Canad. Math. Soc. Conf. Proc. **13** (1992), 345–360.

[102] S.B. Niefield and K.I. Rosenthal, *Spatial sublocales and essential primes*, Topology Appl. **26** (1987), 263–269.

[103] D. Papert and S. Papert, *Sur les treillis des ouverts et paratopologies*, Séminaire Ehresmann, 1re année, exposé 1 (1958).

[104] J. Paseka, *T_2-separation axioms on frames*, Acta Univ. Carolinae Mat. Phys. **28** (1987), 95–98.

[105] J. Paseka, *Paracompact locales and metric spaces*, Math. Proc. Cambridge Philos. Soc. **110** (1991), 251–256.

[106] J. Paseka and B. Šmarda, *T_2-frames and almost compact frames*, Czech. Math. J. **42** (117) (1992), 385–402.

[107] J.W. Pelletier, *Locales in functional analysis*, J. Pure Appl. Algebra **70** (1991), 133–145.

[108] J. Picado, *Weil uniformities for frames*, Comment. Math. Univ. Carolinae **36** (1995), 357–370.

[109] J. Picado, *Join-continuous frames, Priestley's duality and biframes*, Appl. Cat. Structures **2** (1996), 297–313.

[110] J. Picado, *Weil nearness spaces*, Port. Math. **55** (1998), 233–254.

[111] J. Picado, *Two theorems of Efremovič in point-free context*, J. Pure Appl. Algebra **141** (1999), 289–298.

[112] J. Picado, *Frame quasi-uniformities by entourages*, Proc. Symp. on Cat. Top., Cape Town, 1994 (1999), 161–175.

[113] J. Picado, *Structured frames by Weil entourages*, Appl. Cat. Structures **8** (2000), 25–34.

[114] T. Plewe, *Quotient maps of locales*, Preprint (1996).

[115] T. Plewe, *Regular epimorphisms of spatial Hausdorff locales*, Preprint.

[116] T. Plewe, *Localic products of spaces*, Proc. London Math. Soc. **73** (1996), 642–678.

[117] T. Plewe, *Countable products of absolute C_δ-spaces*, Topology Appl. **74** (1996), 39–50.

[118] T. Plewe, A. Pultr and A. Tozzi, *Regular monomorphisms of Hausdorff frames*, to appear in Appl. Math. Structures.

[119] A. Pultr, *Pointless uniformities I, II*, Comment. Math. Univ. Carolinae **25** (1984), 91–104, 105–120.

[120] A. Pultr, *Remarks on metrizable locales*, Proc. 12th Winter School, Suppl. ai Rend. del Circ. Matem. di Palermo, Serie II **6** (1984), 247–258.

[121] A. Pultr, *Diameters in locales: How bad they can be*, Comment. Math. Univ. Carolinae **29** (1988), 731–742.

[122] A. Pultr, *Categories of diametric frames*, Math. Proc. Cambridge Philos. Soc. **105** (1989), 285–297.

[123] A. Pultr and J. Sichler, *Frames in Priestley duality*, Cahiers de Topologie Géom. Différentielle Catégoriques **29** (1988), 193–202.

[124] A. Pultr and J. Sichler, *A Priestley view on spatialization of frames*, Cahiers de Topologie Géom. Différentielle Catégoriques **41** (2000), 225–238.

[125] A. Pultr and W. Tholen, *Localic enrichments of categories*, Cahiers de Topologie Géom. Différentielle Catégoriques **41** (2000), 121–142.

[126] A. Pultr and A. Tozzi, *Notes on Kuratowski–Mrówka theorems in point-free context*, Cahiers de Topologie Géom. Différentielle Catégoriques **33** (1992), 3–14.

[127] A. Pultr and A. Tozzi, *Equationally closed subframes and representation of quotient spaces*, Cahiers de Topologie Géom. Différentielle Catégoriques **34** (1993), 167–183.

[128] A. Pultr and A. Tozzi, *The role of separation axioms in algebraic representation of some topological facts*, Appl. Cat. Structures **2** (1994), 107–118.

[129] A. Pultr and A. Tozzi, *Completion and coproducts of nearness frames*, Festschrift on the Occasion of the 60th Birthday of Guillaume Brümmer, Proc. Symp. on Cat. Top., Cape Town, 1994, University of Cape Town (1999), 177–186.

[130] A. Pultr and A. Tozzi, *A note on reconstruction of spaces and maps from lattice data*, to appear in Quaest. Math.

[131] A. Pultr and J. Úlehla, *Notes on characterization of paracompact spaces*, Comment. Math. Univ. Carolinae **30** (1989), 377–384.

[132] K.I. Rosenthal, *Quantales and their Applications*, Pitman Research Notes in Math., Vol. 234, Longman Scientific & Technical (1990).

[133] J. Rosický and B. Šmarda, *T_1-locales*, Math. Proc. Cambridge Math. Soc. **98** (1985), 81–86.

[134] G. Sambin, *Intuitionistic formal spaces – a first communication*, Mathematical Logic and its Applications, D. Skordev, ed., Plenum, New York (1987), 187–204.

[135] G. Sambin, *Formal topology and domains*, Preprint (1999).

[136] G. Sambin, S. Valentini and P. Virgili, *Constructive domain theory as a branch of intuitionistic pointfree topology*, Theoret. Comput. Sci. **159** (1996), 319–341.

[137] G. Schlitt, *The Lindelöf Tychonoff theorem and choice principles*, Math. Proc. Cambridge Math. Soc. **110** (1991), 57–65.

[138] G. Schlitt, **N**-*compact frames*, Comment. Mat. Univ. Carolinae **32** (1991), 173–187.

[139] D. Scott, *Continuous lattices*, Toposes, Geometry and Logic, Lecture Notes in Math., Vol. 247, Springer, Berlin (1972), 97–136.

[140] D. Scott, *Domains for denotational semantics*, Int. Colloq. on Automata, Languages and Programs, Lecture Notes in Comput. Sci., Vol. 140 (1982), 579–613.

[141] W. Sierpinski, *La notion de derivée comme base d'une théorie des ensembles abstraits*, Mat. Ann. **27** (1927), 321–337.

[142] H. Simmons, *The lattice theoretic part of topological separation properties*, Proc. Edinburgh Math. Soc. (2) **21** (1978), 41–48.

[143] H. Simmons, *A framework for topology*, Logic Colloq. '77, North-Holland, Amsterdam (1978), 239–251.

[144] M.H. Stone, *Boolean algebras and their application in topology*, Proc. Nat. Acad. Sci. USA **20** (1934), 197–202.

[145] M.H. Stone, *The theory of representations for Boolean algebras*, Trans. Amer. Mat. Soc. **40** (1936), 37–111.

[146] S.H. Sun, *On paracompact locales and metric locales*, Comment. Math. Univ. Carolinae **30** (1989), 101–107.

[147] W.J. Thron, *Lattice-equivalence of topological spaces*, Duke Math. J. **29** (1962), 671–679.

[148] J.J.C. Vermeulen, *Weak compactness in constructive spaces*, Preprint (1989).

[149] J.J.C. Vermeulen, *Proper maps of locales*, J. Pure Appl. Algebra **92** (1994), 79–107.

[150] S. Vickers, *Topology via Logic*, Cambridge Tracts Theoret. Comput. Sci., No. 5, Cambridge University Press, Cambridge (1985).

[151] H. Wallman, *Lattices of topological spaces*, Ann. of Math. **39** (1938), 112–126.

[152] J. Walters, *Compactifications and uniformities of σ-frames*, Comment. Math. Univ. Carolinae **32** (1991), 189–198.

[153] J.L. Walters-Wayland, *Metric fine uniform frames*, Comment. Math. Univ. Carolinae **39** (1998), 617–632.

[154] J.L. Walters-Wayland, *A Shirota theorem for frames*, Appl. Cat. Structures **7** (1999), 271–277.
[155] J.L. Walters-Wayland, *Cozfine and measurable uniform frames*, J. Pure Appl. Algebra **145** (2000), 199–213.
[156] G.C. Wraith, *Localic groups*, Cahiers Topologie Géom. Différentielle Catégoriques **12** (1981), 61–66.
[157] G.C. Wraith, *Unsurprising results on localic groups*, J. Pure Appl. Algebra **67** (1990), 95–100.

This Page Intentionally Left Blank

Section 4D
Varieties of Algebras, Groups, . . .

This Page Intentionally Left Blank

Quasivarieties

V.A. Artamonov

Department of Algebra, Faculty of Mechanics and Mathematics, Moscow State University, 119899 Moscow, Russia
E-mail: artamon@mech.math.msu.su

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

Introduction

In this chapter we aim to explain some of the principal results on quasivarieties of algebras. The notion of a quasivariety is mainly due to A.I. Malcev [18]. He was studying classes of semigroups with the property of cancellation that were embedded into groups. It was shown that the class of these semigroups could be defined in terms of quasi-identities. These investigations inspired an interest in the study of quasivarieties in different classes of algebras.

1. Basic notions of universal algebra

In this section we shall briefly recall some necessary basic notions and results in universal algebra.

1.1. *Algebras, subalgebras, direct products*

For any set A and a non-negative integer n denote by A^n the n-th direct power of A. In particular, if $n = 0$, then A^0 is the one-element set. Any map $t : A^n \to A$ is called an *algebraic* operation on A. In particular, a map $t : A^0 \to A$ fixes an element $t(A^0)$ in A.

DEFINITION 1.1. A *signature* or a *system of operations* is a family of sets $T = \{T_n \mid n \geqslant 0\}$. A *general algebra of a signature* T or a T*-algebra* (A, T) is a set A, called the *support*, such that T acts as algebraic operations on A. This means that each $t \in T_n$ determines an n-ary operation on A which is also denoted by t. If $a = (a_1, \ldots, a_n) \in A^n$ and $t \in T_n$, then the image of a under the action of t will be denoted by $t(a) = t(a_1, \ldots, a_n) \in A$.

Usually it is assumed that a signature T is fixed. A slightly different definition of a general algebra is used in [10, §7].

DEFINITION 1.2. A *type* is a sequence of non-negative integers

$$\tau = (n_0, n_1, \ldots, n_\gamma, \ldots), \quad \gamma < o(\tau),$$

where $o(\tau)$ is a given cardinal. An *algebra of type* τ is a set A, the support, such that $\gamma < o(\tau)$ defines an n_γ algebraic operation on A.

It is clear that Definition 1.1 and Definition 1.2 are equivalent. In what follows we shall mainly use the definition of an algebra in the form 1.1.

DEFINITION 1.3. A subset B in an algebra (A, T) is a *subalgebra* if $t(b_1, \ldots, b_n) \in B$ for any $t \in T_n$, $n \geqslant 0$, and any set of elements $b_1, \ldots, b_n \in B$. Denote by $\operatorname{Sub} A$ the partially ordered set with respect to inclusion of all subalgebras in an algebra A.

According to Definition 1.3 an empty subset of an algebra (A, T) is always a subalgebra. It follows that for any subset Y in $\operatorname{Sub} A$ the intersection of all members of Y belongs to $\operatorname{Sub} A$. Observe that if T_0 is nonempty then any subalgebra in (A, T) always contains the elements $t(A^0)$, so any subalgebra in (A, T) is nonempty and there exists a least subalgebra in A which is nonempty. Since $\operatorname{Sub} A$ is closed under intersections of any subsets and there exists the largest element of $\operatorname{Sub} A$ – the algebra A itself, the partially ordered set $\operatorname{Sub} A$ is a complete lattice.

DEFINITION 1.4. For any subset Y in an algebra A denote by $\langle Y\rangle$ the least subalgebra in A containing Y. The subalgebra $\langle Y\rangle$ is called the subalgebra *generated* by the set Y. An algebra A is *finitely generated* if $A = \langle Y\rangle$ for some finite subset Y in A.

We shall now recall one basic notion from lattice theory.

DEFINITION 1.5. An element a of a complete lattice L is *compact* if for any subset Y in L an inequality $a \leqslant \sup Y$ implies $a \leqslant \sup Y'$ for some finite subset Y' in Y. A complete lattice L is an *algebraic* lattice if for every element a of L there exists a set of compact elements $Z \subseteq L$ such that $a = \sup Z$.

THEOREM 1.6 (G. Birkhoff and O. Frink [10, §9]). *A subalgebra B in an algebra A is compact in* $\operatorname{Sub} A$ *if and only if B is a finitely generated algebra. The lattice* $\operatorname{Sub} A$ *is algebraic. Moreover any algebraic lattice is isomorphic to a lattice of the form* $\operatorname{Sub} A$ *for some algebra A.*

For any signature T and a set (an alphabet) X there always exists a nontrivial algebra of T-words $W(X)$ which is defined as follows.

DEFINITION 1.7. The notion of a T-*word* and the functions d_z, $z \in X$, are defined by induction:

- the elements $x \in X$ are T-words and

$$d_z(x) = \begin{cases} 1 & \text{if } z = x, \\ 0 & \text{otherwise;} \end{cases}$$

- the elements $t \in T_0$ are T-words and $d_z(t) = 0$;
- if $t_1, \dots, t_n$ are T-words and $t \in T_n,\ n > 0$, then $t(t_1, \dots, t_n)$ is a T-word and

$$d_z\bigl(t(t_1, \dots, t_n)\bigr) = d_z(t_1) + \cdots + d_z(t_n).$$

It is assumed that the set of T-words $W(X)$ contains only the elements which are obtained by the previous construction. The value $d_z(t)$ is called the number of *occurrences* of the variable z in t.

It is an easy exercise to observe that for any $t \in W(X)$ there exists finitely many elements $x \in X$ such that $d_x(t) > 0$.

NOTATION 1.8. Let $t \in W(X)$ and let $x_1, \dots, x_n \in X$ be the only elements such that $d_{x_i}(t) > 0,\ i = 1, \dots, n$. In that case we write $t = t(x_1, \dots, x_n)$.

Now we are going to introduce the notion of a direct product. It seems likely that the reader is quite familiar with it especially in the case of direct products of finitely many factors. Nevertheless to avoid some ambiguities with the foundations of mathematics we shall define direct products in the following way.

DEFINITION 1.9. Let there be given a set $A_i,\ i \in I$, of T-algebras. Denote by Q the disjoint union $\bigcup_{i \in I} A_i$. The *direct product* A of $A_i,\ i \in I$, is the set of all maps $f : I \to Q$ such that $f(i) \in A_i$, for each $i \in I$. If $t \in T_n,\ n \geqslant 0$, and $f_1, \dots, f_n \in A$, then we define $t(f_1, \dots, f_n)$ pointwise, i.e.

$$\big[t(f_1, \dots, f_n)\big](i) = t\big(f_1(i), \dots, f_n(i)\big) \in A_i. \tag{1}$$

It is clear that $t(f_1, \dots, f_n) \in A$. Thus A is a T-algebra called the *direct product* of the algebras $A_i,\ i \in I$, notation $A = \prod_{i \in I} A_i$. In particular, if $A_i = C$ for each $i \in I$, then A can be identified with the set all maps $I \to C$ and with the action of T given by (1). In this particular case the algebra A is called a *direct power* of C, notation $A = C^I$.

If $I = \{1, 2, \dots, n\}$ is a finite set then the direct product A of algebras $A_1, \dots, A_n$ is also denoted by $A = A_1 \times \cdots \times A_n$. The elements f of $A_1 \times \cdots \times A_n$ can be identified with n-tuples $(f(1), \dots, f(n))$.

1.2. *Homomorphisms and congruences*

DEFINITION 1.10. Let A, B be two algebras. A map $f : A \to B$ is a *homomorphism* if

$$f\big(t(a_1, \dots, a_n)\big) = t\big(f(a_1), \dots, f(a_n)\big),$$

for any elements $a_1, \dots, a_n \in A$. A homomorphism of an algebra to itself is an endomorphism of the algebra. A bijective homomorphism of an algebra A to an algebra B is an *isomorphism*, notation $A \simeq B$. An isomorphism of an algebra onto itself is an *automorphism* of an algebra.

It is clear that a product of homomorphisms is again a homomorphism. The identity map is an automorphism of an algebra. So all T-algebras and their homomorphisms form a category of T-algebras. Moreover an image of any homomorphism is always a subalgebra.

There are some special homomorphisms associated with direct products. Let $A = \prod_{i \in I} A_i$.

NOTATION 1.11. Denote by $p_i : A \to A_i$ the map $p_i(f) = f(i),\ i \in I$.

PROPOSITION 1.12. *The map p_i from Notation* 1.11 *is a homomorphism. Let B be a T-algebra and $h_i : A \to A_i,\ i \in I$, a set of homomorphisms. Then there exists a unique homomorphism $h : B \to A$ such that $p_i h = h_i$, for any $i \in I$.*

The algebra of T-words $W(X)$ enjoys some special universal property.

PROPOSITION 1.13. *Let B be a T-algebra and $h : X \to B$ an arbitrary map of sets. Then there exists a unique homomorphism $f : W(X) \to B$ such that $f(x) = h(x)$, for any $x \in X$. The image* $\mathrm{im}(f)$ *is the subalgebra $\langle h(X)\rangle$ of B generated by the set $h(X)$. In particular any T-algebra is a homomorphic image of an algebra $W(X)$, for some set X.*

The next notion is an analog of a normal subgroup in group theory and an ideal in ring theory.

DEFINITION 1.14. A *binary relation* $\wp$ on a set A is a subset of the direct square $A^2 = A \times A$. An *equivalence* is a binary relation $\wp$ on A such that

(1) $(a, a) \in \wp$;
(2) $(a, b) \in \wp$ implies $(b, a) \in \wp$;
(3) $(a, b), (b, c) \in \wp$ implies $(a, c) \in \wp$,

for any $a, b, c \in A$. A *congruence* $\wp$ in a T-algebra A is a subalgebra of A^2 which is at the same time an equivalence.

THEOREM 1.15 [10, §18]. *The set* $\mathrm{Con}\, A$ *of all congruences is a partially ordered set with respect to inclusion. The diagonal $\Delta = \{(a, a) \mid a \in A\}$ is the least element of* $\mathrm{Con}\, A$. *The greatest element of* $\mathrm{Con}\, A$ *is the congruence A^2. An intersection of any set of congruences is again a congruence. Thus* $\mathrm{Con}\, A$ *is a complete lattice. Moreover,* $\mathrm{Con}\, A$ *is an algebraic lattice. Any algebraic lattice is isomorphic to a lattice of the form* $\mathrm{Con}\, A$ *for some algebra A.*

Each congruence $\wp$ in an algebra A induces a partition of the set A into disjoint classes of equivalent elements.

NOTATION 1.16. For an element $a \in A$ and a congruence denote by $\wp(a)$ the class containing a. Denote by $A/\wp$ the set of all classes of the congruence $\wp$.

Let $t \in T_n \geqslant 0$ and $a_1, \ldots, a_n \in A$. It is not difficult to check that the algebraic operation

$$t\bigl(\wp(a_1), \ldots, \wp(a_n)\bigr) = \wp\bigl(t(a_1, \ldots, a_n)\bigr) \tag{2}$$

is well defined. So $A/\wp$ is again a T-algebra. The map

$$\pi : A \to A/\wp, \quad a \rightsquigarrow \wp(a), \tag{3}$$

is a homomorphism.

DEFINITION 1.17. The T-algebra $A/\wp$ is called the quotient algebra or *factor algebra* of A by $\wp$. The homomorphism π from (3) is called a *natural* homomorphism.

There is a natural duality between homomorphisms and congruences.

DEFINITION 1.18. If $f : A \to B$ is a homomorphism of T-algebra, then the kernel $\ker f$ is the set of all pairs $(a, b) \in A \times A$ such that $f(a) = f(b)$.

PROPOSITION 1.19. *A kernel of any homomorphism is a congruence. If* $\wp \in \operatorname{Con} A$ *and* $\pi : A \to A/\wp$ *is the natural homomorphism, then* $\ker \pi = \wp$.

THEOREM 1.20 (first homomorphism theorem). *Let* $f : A \to B$ *is a homomorphism, then* $\operatorname{im} f \simeq A/\ker f$.

THEOREM 1.21 (second homomorphism theorem). *Let* $\wp \subseteq \mathfrak{q} \in \operatorname{Con} A$. *Denote by* $\mathfrak{q}/\wp$ *the set of all pairs* $(\wp(a), \wp(b)) \in (A/\wp) \times (A/\wp)$, *such that* $(a, b) \in \mathfrak{q}$. *Then* $\mathfrak{q}/\wp \in \operatorname{Con}(A/\wp)$ *and*

$$(A/\wp)/(\mathfrak{q}/\wp) \simeq A/\mathfrak{q}.$$

A join in the lattice $\operatorname{Con} A$ coincides with the set-theoretic intersection of congruences as subsets in $A \times A$. The union in the lattice $\operatorname{Con} A$ is a rather complicated.

DEFINITION 1.22. Let θ, θ' be binary relations in a set A. The *product* $\theta \circ \theta'$ is the set of all elements $(a, c) \in A^2$ such that there exist an element $b \in A$ such that $(a, b) \in \theta$ and $(b, c) \in \theta'$. The *inverse* relation $\wp^{-1}$ consists of all pairs (a, b) such that $(b, a) \in \wp$. An algebra A is *congruence-permutable* if $c \circ c' = c' \circ c$ for any two congruences on A.

PROPOSITION 1.23. *The set of all binary relations in an algebra A form a semigroup with respect to* $\circ$ *with zero and unit element. A binary relation* $\wp$ *in* A *is an equivalence if and only if*

$$\wp \supseteq \Delta, \qquad \wp^{-1} = \wp, \qquad \wp \circ \wp = \wp.$$

If $\theta, \theta' \in \operatorname{Con} A$, *then and* $\theta \vee \theta'$ *is the union in the lattice* $\operatorname{Con} A$, *then*

$$\theta \vee \theta' = \bigcup_{n \geqslant 1} (\theta \circ \theta')^n.$$

In particular, if A *is congruence permutable then* $\theta \vee \theta' = \theta \circ \theta'$. *In that case the lattice* $\operatorname{Con} A$ *is modular.*

An intersection in $\operatorname{Con} A$ can be easily interpreted in terms of direct products.

PROPOSITION 1.24. *Let* B *and* A_i, $i \in I$, *be* T*-algebras. Suppose that there are given homomorphisms of* T*-algebras* $h_i : B \to A_i$, $i \in I$, *and let* $h : B \to A = \prod_{i \in I} A_i$ *be the homomorphism from Proposition* 1.12 *such that* $\pi_i h = h_i$, $i \in I$. *Then* $\ker h = \bigcap_{i \in I} \ker h_i$.

DEFINITION 1.25. An algebra B is a *subdirect product* of algebras A_i, $i \in I$, if there exists an embedding $h : B \to A = \prod_{i \in I} A_i$ such that $\pi_i(B) = A_i$, for any $i \in I$. An algebra B is *subdirectly irreducible* if there exists a least congruence in B, different from the diagonal Δ. An algebra A is *residually finite* if A is a subdirect product of finite algebras.

It follows immediately from Proposition 1.24 that B is a subdirect product of algebras $A_i,\ i \in I$, if there exists a family of surjective homomorphisms $h_i : B \to A_i,\ i \in I$, such that $\Delta = \bigcap_{i \in I} \ker h_i$.

THEOREM 1.26 (G. Birkhoff). *Let B be algebra. Then B is a subdirect product of subdirectly irreducible algebras.*

1.3. *Filtered products, ultraproducts*

DEFINITION 1.27. A nonempty family $\mathcal{F}$ of subsets of a set X is a *filter* if

(1) the empty subset does not belong to $\mathcal{F}$;

(2) if $I \in \mathcal{F}$ and $X \supseteq J \supseteq I$, then $J \in \mathcal{F}$;

(3) if $K, L \in \mathcal{F}$, then $K \cap L \in \mathcal{F}$.

A filter $\mathcal{F}$ is trivial, if it consists only of the set X. A filter $\mathcal{F}$ is principal, if there exists a nonempty subset D in X such that $\mathcal{F}$ consists of all subsets containing D, notation (D). The *Frechet* filter $\mathcal{F}$ on an infinite set X consists of all subsets I in X such that $X \setminus I$ is finite.

Similarly one can define a filter in any lattice.

It is clear that if X is a finite set then any filter over X is principal. In fact let $\mathcal{F} = \{X_1, \dots, X_k\}$. Then $\mathcal{F} = (X_1 \cap \dots \cap X_k)$.

PROPOSITION 1.28. *The Frechet filter $\mathcal{F}$ on an infinite set X is a filter and it is not principal.*

PROOF. It is pretty clear that the Frechet filter is a filter. Suppose that there exists a subset $D \subseteq X$ such that $\mathcal{F} = (D)$. Then D coincides with the intersection of all members of $\mathcal{F}$. It is not hard to see that this intersection is in fact empty. □

DEFINITION 1.29. A set of all filters on a set X form a partially ordered set by inclusion. Maximal elements of this set are called *ultrafilters*.

PROPOSITION 1.30. *A filter $\mathcal{F}$ on a set X is an ultrafilter if and only if for any subset I in X either I or $X \setminus I$ belongs to $\mathcal{F}$. Any filter is contained in a ultrafilter. If X is an infinite set then there exists a nonprincipal ultrafilter on X. An ultrafilter $\mathcal{F}$ on X is principal if and only if there exists an element $x \in X$ such that $\mathcal{F} = (x)$.*

Suppose that $\mathcal{F}$ is a filter over a set X and there is given a set of T-algebras $A_i,\ i \in X$. Put $A = \prod_{i \in X} A_i$.

NOTATION 1.31. Denote by $\widetilde{\mathcal{F}}$ the set of all pairs $(f, g) \in A \times A$, such that the set of all indices $i \in X$ for which $f(i) = g(i)$ belongs to $\mathcal{F}$.

PROPOSITION 1.32. *$\widetilde{\mathcal{F}}$ is a congruence in A.*

DEFINITION 1.33. The factor algebra $A/\widetilde{\mathcal{F}}$ is called an *filtered product* of algebras A_i, $i \in X$. If $\mathcal{F}$ is an ultrafilter, then $A/\widetilde{\mathcal{F}}$ is called an *ultra product* or a *prime product*.

1.4. *Prevarieties, quasivarieties and varieties*

We shall recall some necessary notations and results on prevarieties from [4, §1.1].

DEFINITION 1.34. A class K of T-algebras is abstract if it is closed under isomorphisms of algebras. This means that if $A \in K$ and $B \simeq A$, then $B \in K$.

NOTATION 1.35. Let K be an abstract class of algebras. Then
- SK is the class of algebras isomorphic to subalgebras of algebras from K;
- HK is the class of all homomorphic images of algebras from K;
- ΠK is the class of algebras isomorphic to direct products of algebras from K;
- $\Pi_u K$ is the class of algebras isomorphic to ultraproducts of algebras from K;
- $\Pi_f K$ is the class of algebras isomorphic to filter products of algebras from K.

An abstract class K of T-algebras is a *prevariety* if it contains the one-element algebra and K is closed under operators S, Π. An abstract nonempty class K of T-algebras is a *variety* if K is closed under operators S, H, Π. An abstract class K of T-algebras is a *quasivariety* if it is closed under S, Π_f and contains the one-element algebra.

THEOREM 1.36. *Let L be a class of T-algebras. Then $S\Pi(L) \cup E$ is the least prevariety of T-algebras containing L. Here E is the class of the one-element algebras, which are of course isomorphic. In particular if L consists of one finite algebra A, then $S\Pi(A) \cup E$ is locally finite that is finitely generated algebras in $S\Pi(A) \cup E$ are finite. If K is a prevariety then $H(K)$ is a variety. In particular if L is a nonempty class of T-algebras, then $HS\Pi(L)$ is the least variety containing L. The class $S\Pi_f(L) \cup E$ is the least quasivariety containing LE. An abstract class K is a quasivariety if and only if K is a prevariety and K is closed under Π_u.*

Let A be a T-algebra and K a prevariety. It follows immediately from Proposition 1.24 that there exists the least congruence $K(A)$ in A such that $A/K(A) \in K$. If $h : A \to B$ is a homomorphism and $B \in K$, then $\ker h \supseteq K(A)$. By Theorem 1.20, $\operatorname{im} h \simeq A/\ker h$. Since K is closed under subalgebras we conclude $\operatorname{im} h \in K$. By Theorem 1.21 there exists a homomorphism $h' : A/K(A) \to \operatorname{im} K$ such that $\ker h' = \ker h/K(A)$. Then $h = h'r$, where r is the natural homomorphism $A \to A/K(A)$. These considerations lead to the proof of

THEOREM 1.37 (A.I. Malcev). *Let K be an abstract class of T-algebras containing the one element algebra. Then the following are equivalent*:

(1) *K is a prevariety*;

(2) *any T-algebra A has a least congruence $K(A)$ such that $A/K(A) \in K$*;

(3) *for any T-algebra A there exists an algebra $R(A) \in K$ and a surjective homomorphism $r : A \to R(A)$ such that for any K-algebra B and a homomorphism $h : A \to B$ there exists a homomorphism $h' : R(A) \to B$ for which $h'r = h$.*

It is not too difficult to see that if α is an endomorphism of an T-algebra A and $(a, b) \in K(A)$, then $(\alpha(a), \alpha(b)) \in K(A)$.

NOTATION 1.38. Let K be a prevariety of T-algebras and A an arbitrary T-algebra. Denote by $\mathrm{Con}_K(A)$ the set of all congruences c on A such that $A/c \in K$.

PROPOSITION 1.39. $\mathrm{Con}_K(A)$ *is closed under intersection that is* $\mathrm{Con}_K(A)$ *is a complete lower semilattice. Moreover, the largest congruence* $A \times A \in \mathrm{Con}_K(A)$. *Hence* $\mathrm{Con}_K(A)$ *is a complete lattice.*

PROOF. Suppose that c_i, $i \in I$, are elements of $\mathrm{Con}_K(A)$. Then by Proposition 1.24 $\bigcap_{i \in I} c_i \in \mathrm{Con}_K(A)$. The rest of the proof can be deduced from that fact that a complete lower semilattice with the greatest element is a complete lattice. □

DEFINITION 1.40. Let X be a set and K a prevariety of T-algebras. An algebra $F_K(X) \in K$ is a *free* K-algebra with basis X if $F_K(X)$ is generated by the set X and any map from X to any algebra $B \in K$ can be extended to a homomorphism $F_K(X) \to B$.

Since X generates $F_K(X)$ the extension mentioned is always unique. It is not hard to see that the algebra $F_K(X)$ is unique up to an isomorphism if it exists. If K contains only the one-element algebra then certainly K has no free algebras with basis X if X contains at least two distinct elements. If K is the class of all T-algebras then $F_K(X)$ coincides with the algebra of T-words $W(X)$ from Definition 1.7.

THEOREM 1.41. *Let a prevariety K contains an algebra with at least two elements and let $W(X)$ be the free algebra on the alphabet X. Then $W(X)/K(W(X)) = F_K(X)$ and the natural homomorphism $r : W(X) \to W(X)/K(W(X))$ is an embedding. If $V = HL$ is the least variety containing K, see Theorem* 1.36, *then a K-free algebra is also a V-free algebra.*

Let $F_K(X)$ be the free K-algebra with basis $X = \{x_1, x_2, \ldots\}$.

NOTATION 1.42. Suppose that $t \in F_K(X)$ and t is an image of a word w from $W(X)$ such that $w = w(x_1, \ldots, x_n)$ in the sense of Notation 1.8. Then we write $t = t(x_1, \ldots, x_n)$ and $v_z(t) = d_z(w)$. Let $B \in K$ and $h : F_k(X) \to B$ a homomorphism such that $h(x_1) = b_1, \ldots, h(x_n) = b_n$. Put $t(b_1, \ldots, b_n) = h(t(x_1, \ldots, x_n)) \in B$.

This notation is correct because $F_K(X)$ is a K-free algebra. This means that any element of $F_K(X)$ can be viewed as an n-ary operation induced by t. Observe that w, d_z are not uniquely determined by t.

DEFINITION 1.43. The elements of the algebra $F_K(X)$ with a base $X = \{x_1, x_2, \ldots\}$ are called K-*terms* and the operations induced by terms are called *term operations*. A variable $z \in X$ *occurs* in a term t if $v_z(t) > 0$.

2. First order languages, semantic characterizations

In this section we shall a semantic characterization of quasivarieties. First we need some basic notions related to logical aspects of algebraic theories.

2.1. *The first order language*

DEFINITION 2.1. Let K be a fixed nontrivial prevariety of T-algebras with K-free algebra $F_K(X)$ with basis $X = \{x_1, x_2, \ldots\}$ as in Definition 1.43. The *first order language* L_K contains the following objects:

(1) the terms $f \in F_K(X)$, in particular the (individual) variables $x_1, x_2, \ldots$,
(2) the symbol for equality,
(3) the logical connectives, $\forall$, $\exists$, $\vee$, $\wedge$, $\to$, $\neg$,
(4) parentheses ( ) and commas.

A *formula* of L_K is defined by the following rules:

(1) if f, g are terms, then $f = g$ is a formula called an *atomic formula* of L_K; we put $v_i(f = g) = v_{x_i}(f) + v_{x_i}(g)$ as in Notation 1.42;
(2) if Φ_1, Φ_2 are formulas, then so are $\Phi_1 \vee \Phi_2$, $\Phi_1 \wedge \Phi_2$, $\Phi_1 \to \Phi_2$; we put

$$v_i(\Phi_1 \vee \Phi_2) = v_i(\Phi_1 \wedge \Phi_2) = v_i(\Phi_1 \to \Phi_2) = v_{x_i}(f) + v_{x_i}(g);$$

(3) if Φ is a formula, then so are $\neg\Phi$, $(\forall x_k)\Phi$, $(\exists x_k)\Phi$

$$v_i(\neg\Phi) = v_i(\Phi), \qquad v_i\big((\forall x_k)\Phi\big) = v_i\big((\exists x_k)\Phi\big) = \begin{cases} 0 & \text{if } i = k, \\ v_i(\Phi) & \text{otherwise.} \end{cases}$$

A variable x_i is *free* in a formula Φ if $v_i(\Phi) > 0$. A formula is *closed* or it is a *sentence* if it has no free variables.

As in Notation 1.42 we write $\Phi = \Phi(x_1, \ldots, x_n)$ for a formula Φ if $x_1, \ldots, x_n$ are the only free variables in Φ.

Let now A be an algebra of the prevariety K and Φ is a formula of the language L_K. We shall define inductively the notion of a *satisfiability* of $\Phi = \Phi(x_1, \ldots, x_n)$ in A.

DEFINITION 2.2. Let $f(x_1, \ldots, x_n) = g(x_1, \ldots, x_n)$ be an atomic formula. It is satisfied at an n-tuple $a = (a_1, \ldots, a_n) \in A^n$ if $f(a_1, \ldots, a_n) = g(a_1, \ldots, a_n)$. A formula $\Phi_1 \vee \Phi_2$ is satisfied at a if either Φ_1 or Φ_2 is satisfied. A formula $\Phi_1 \wedge \Phi_2$ is satisfied by a if both Φ_1 and Φ_2 are satisfied. A formula $\neg\Phi$ is satisfied if Φ is not satisfied. A formula $\Phi_1 \to \Phi_2$ is satisfied at a if either $\neg\Phi_1$ or Φ_2 are satisfied. A formula $(\exists x_k)\Phi$ is satisfied if there exists an element $b \in A$ such that Φ is satisfied at an element $a = (a_1, \ldots, a_n) \in A^n$ where $b = a_k$. A formula $(\forall x_k)\Phi$ is satisfied if for any element $b \in A$ the formula Φ is satisfied at an element $a = (a_1, \ldots, a_n) \in A^n$ where $a_k = b$.

A formula Φ is satisfied in a class $M \subseteq K$ if for any algebra $A \in M$ and n-tuple $a \in A^n$ the formula Φ is satisfied at a. The *elementary theory* $\mathrm{Th}(M)$ of the class M is the set of all sentences which are satisfied in M. Two K-algebras are *elementary equivalent* if $\mathrm{Th}(A) = \mathrm{Th}(B)$.

Two formulas Φ_1, Φ_2 are *equivalent* in K if $(\Phi_1 \to \Phi_2) \wedge (\Phi_2 \to \Phi_1)$ is satisfied in any algebra K-algebra.

It is not hard to observe that the following pairs of formulas are equivalent:
- $\Phi_1 \wedge \Phi_2$ and $\neg((\neg\Phi_1) \vee (\neg\Phi_2))$;
- $\Phi_1 \to \Phi_2$ and $(\neg\Phi_1) \vee \Phi_2$;
- $(\forall x_k)\Phi$ and $\neg((\exists x_k)\Phi)$.

These equivalences lead to the proof of

THEOREM 2.3. *Any formula in the language L_K is equivalent to a formula in prenex form*

$$(Q_1 x_{i_1}) \cdots (Q_n x_{i_n})\Phi,$$

where Q_i are quantifiers $\forall, \exists$ and the formula Φ does not contain quantifiers.

DEFINITION 2.4. A sentence Φ of a language L_K is *universal* if it has a prenex form

$$(\forall x_1) \dots (\forall x_n)\Phi.$$

A class of algebras is *universal* if it consists of all algebras in which a given set of universal sentences is satisfied.

Some natural families of universal classes are considered in the next subsection.

2.2. *Semantic properties of quasivarieties*

In this subsection we shall apply these notions to the study of ultraproducts.

THEOREM 2.5 (J. Łoś). *Let K be a quasivariety and A an ultraproduct of algebras $A_i \in K, i \in I$, corresponding to an ultrafilter $\mathcal{F}$ on I. Suppose that Φ is a formula of the language L_K. Then Φ is satisfied in A if and only if the set of all indices $i \in I$ such that Φ is satisfied in A_i belongs to $\mathcal{F}$.*

COROLLARY 2.6. *Let $K, A, A_i, I, \mathcal{F}$ be from Theorem* 2.5. *Suppose that Φ is a sentence of the formula L_K which is satisfied in each algebra A_i, $i \in I$. Then Φ is satisfied in A.*

THEOREM 2.7. *Let K be a quasivariety and $A \in K$. Then A and any of its ultrapowers are elementary equivalent.*

Isomorphic algebras are of course elementary equivalent. It can be deduced from Theorem 2.7 that if A is an infinite algebra then there exists an ultrapower of A with an arbitrary large cardinality which is elementary equivalent but nonisomorphic to A.

Now we are able to present a semantic characterization of quasivarieties and varieties.

DEFINITION 2.8. A *quasi-identity* is a sentence

$$(\forall x_1)\ldots(\forall x_n)\big((f_1 = g_1) \wedge \cdots \wedge (f_m = g_m) \rightarrow (f = g)\big)$$

in the language L_K where $f_1, g_1, \ldots, f_n, g_m, f, g$ are terms. An *identity* is a sentence

$$(\forall x_1)\ldots(\forall x_n)(f = g)$$

in the language L_K where f, g are terms.

It is clear that any identity is a quasi-identity.

THEOREM 2.9 (G. Birkhoff). *A abstract class K of T-algebras is a variety if and only there exists a set of identities Λ such that K consists of all algebras in which every identity from Λ is satisfied.*

THEOREM 2.10 (A.I. Malcev). *A abstract class K of T-algebras is a quasivariety if and only there exists a set of quasi-identities Λ such that K consists of all algebras in which every quasi-identity from Λ is satisfied.*

For example, a free group satisfies the set of quasi-identities

$$(\forall x)\big((x^n = 1) \rightarrow x = 1\big), \quad \text{for any } n \geqslant 1.$$

This example shows that a free algebra in a variety may generate a proper subquasivariety.

The following remark was the motivation for A.I. Malcev for introducing the notion of a quasivariety.

Suppose that K is an arbitrary quasivariety of T-algebras and $T' \subseteq T$. Let K' be the class of all T'-algebras which are embedable into K-algebras. Then K' is a quasivariety of T'-algebras. In particular all semigroups embeddable into a group form a quasivariety.

Any quasi-identity is a universal sentence in the L_K.

3. General theory

In this section we shall present some recent results of quasivarieties of algebras. The survey is intended to acquaint the reader with some modern trends in the theory of quasivarieties.

3.1. *Lattices of quasivarieties*

NOTATION 3.1. If K is a quasivariety then all subquasivarieties in K form a lattice $L_q(K)$ with respect to inclusion. If K is a variety then $L(V)$ denotes the lattice of subvarieties in K.

A. Shafaat [22] showed that for a quasivariety K the following are equivalent:

(1) if M is a subquasivariety in K then it has finitely many finite algebras such that each M-algebra is a subdirect product of these algebras;

(2) K is locally finite and each lattice $L_q(M)$ has ACC for any subquasivariety M in K.

THEOREM 3.2 [6]. *Let V be a quasivariety and N an abstract subclass in V which is closed under subalgebras and finite direct products. The following are equivalent*:

(1) *V as a quasivariety is generated by the class N*;

(2) *each finitely presented V-algebra is a subdirect product of N-algebras*;

(3) *each V-algebra is a subalgebra of an ultraproduct of N-algebras.*

THEOREM 3.3 [6]. *Let V be a quasivariety. The lattice $L_q(V)$ is finite if and only if $L_q(V)$ has descending and ascending chain conditions.*

THEOREM 3.4 [6]. *Let V be a quasivariety. The lattice $L_q(V)$ is modular if and only if it is distributive.*

V.A. Gorbunov studied systematically the structure of lattices of the form $L_q(V)$. His results are expounded in [9].

THEOREM 3.5 (V.P. Belkin and V.A. Gorbunov). *Let Q_T be a quasivariety of all T-algebras where either $|T_1| \leqslant 2$ or T_n is nonempty for some $n > 1$. Then any filter in the lattice $L_q(Q_T)$ contains at least continuum many elements. A similar result holds for the quasivarieties of all loops, quasigroups and groups.*

Any subdirectly irreducible finitely presented group satisfies some nontrivial group quasi-identity.

THEOREM 3.6 (V.A. Gorbunov). *The lattice quasi-identity*

$$x \vee y = x \vee z \to x \vee y = x \vee (y \wedge z) \tag{4}$$

holds in any lattice of the form $L_q(K)$.

DEFINITION 3.7. A lattice L is *relatively complemented* if for any elements $a \geqslant c$ in L there exists an element $d \in L$ such that $b \vee d = a$ and $b \wedge d = c$. A lattice L with zero element is pseudocomplemented is for any element $a \in L$ there exists the maximal element $a^* \in L$ such that $a \wedge a^* = 0$.

V.A. Gorbunov has shown that if a lattice of the form $L_q(K)$ is relatively complemented then it is Boolean. A Boolean algebra L is isomorphic to a lattice of the form $L_q(K)$ if and only if it is countable and it is isomorphic to a lattice of subsets of some set.

A finite lattice L of the form $L_q(K)$ is distributive if and only if a certain fixed 7-element lattice is not embedable into L.

THEOREM 3.8 (V.A. Gorbunov). *For a quasivariety K the following are equivalent*:

(1) $L_q(K)$ *is distributive*;

(2) *finitely generated subdirectly irreducible K-algebras are projective in K*;

(3) *any subquasivariety in K is defined by identities.*

Let K be a locally finite quasivariety of T-algebras where T is finite.

THEOREM 3.9 (V.A. Gorbunov). *The lattice $L_q(K)$ is pseudocomplemented if and only if $L_q(K)$ enjoys the quasi-identities which are dual to* (4). *Moreover the following are equivalent*

- *for any nonmaximal element $M \in L_q(K)$ there exists a cover N that is $N > M$ and there exists no element N' such that $N > N' > M$*;
- *K is generated by finitely many finite algebras.*

A 2-element T-algebra with finite T generates a minimal quasivariety.

Suppose that K is a locally finite (quasi)variety of algebraic systems (in particular of algebras). It is shown in [9] that if V is a variety, then $L(V)$ is isomorphic to an inverse limit of some family of finite lattices that are upper semidistributive at zero. If K is a quasivariety, then $L_q(V)$ is isomorphic to an inverse limit of some family of finite lower bounded lattices. There exists a family of quasi-identities S_n^0 which are satisfied in any lattice of subquasivarieties $L_q(K)$ if K is of a finite type.

It has been shown that if K is a directed family with respect to surjections of l-projective systems of a prevariety V then there exists a complete homomorphism φ from the lattice $L_q(V)$ to the inverse limit of lattices of congruences $\mathrm{Con}_V A$ of systems A from K. Here as above $\mathrm{Con}_V(A)$ denotes the lattice of all congruences $c \in \mathrm{Con}\, A$ such that $A/c \in V$. There are sufficient conditions under which φ is an isomorphisms. Similar results have been obtained for lattices of V-subvarieties $L_v(V)$. A prevariety is locally Noetherian if each lattice $\mathrm{Con}_V(A)$ has ACC for any finitely generated V-system A. It has been shown that $L_q(V) = L_q(Q(V))$ for any locally Noetherian prevariety V, where $Q(V)$ is the quasivariety generated by V. Moreover, if V is a union of directed family of V-varieties V_i, then under some restrictions $L_q(V)$ is an inverse limit of $L_q(V_i)$. The lattice $L_q(V)$ of a locally Noetherian quasivariety V satisfies the quasi-identities

$$(S_n):\qquad \&_{0\leqslant i<n}[x_i \leqslant x_{i+1} \vee y_i \,\&\, x_i \wedge y_i \leqslant x_{i+1}]\,\& \\ \bigwedge_{0\leqslant i<n} x_i \leqslant \bigwedge_{0\leqslant i<n} y_i \to x_0 \approx x_1,$$

$$(T_{n,j}):\qquad \&_{0\leqslant i<n}[x_i \leqslant x_{i+1} \vee y_i \,\&\, x_{i+1} \wedge y_{i+1} \leqslant y_i]\,\& \\ \bigwedge_{0\leqslant i<n} x_i \leqslant y_j \to x_1 \leqslant y_1$$

for all $n \geqslant 2,\ 0 \leqslant j < n$. Observe that the free lattice of countable rank satisfies $S_n, T_{n,j}$. If V is a locally finite prevariety of systems with finitely many basic relations then $L_v(V)$

is an inverse limit of finite lattices. There is an example of a lattice of a variety L such that L is not isomorphic to a lattice $L_v(W)$ for any locally finite variety W. If V is a locally finite variety then $L_v(V)$ is an inverse limit of semi-distributive finite lattices. A class of finite algebraic systems V is a pseudovariety if it is closed under subsystems and finite direct products. Let V be a locally finite quasivariety of systems with finitely many basic relations. Then $L_q(V)$ is isomorphic to the lattice of pseudovarieties $L_{pq}(V_{fin})$, where V_{fin} is the class of all finite V-systems. $L_{pq}(V_{fin})$ always satisfies the quasi-identities S_n, T_{nj}. If V is a variety then any universal (positive universal) proposition which is true in $L_q(V)$ is always true in $L_{pq}(V_{fin})$.

THEOREM 3.10 (V.I. Tumanov). *A finite distributive lattice is isomorphic to a lattice $L_q(K)$ for some locally finite quasivarieties of T-algebras for some finite T.*

The next result is proved in [24].

THEOREM 3.11. *Let K be a quasivariety in which either $L_q(K)$ satisfies the quasi-identity*

$$x \wedge y = 0 \to (z \vee x) \wedge y = z \not\supset y,$$

or K is congruence-permutable. Then the following are equivalent:
- *if M is a subquasivariety in K, then $L_q(M)$ is self-dual*;
- *$L_q(K)$ is a direct product of finitely many finite chains.*

Suppose that in each K-algebra every nontrivial congruence has a class which is in fact a subalgebra. Then the following are equivalent [25]:

(1) the lattice $L_q(K)$ is complemented;
(2) the lattice $L_q(K)$ is Boolean;
(3) K is a union of finitely many quasivarieties.

The next result is due to S. Tulipani.

THEOREM 3.12 [23]. *For a variety of algebras K the following are equivalent*:

(1) *K as a quasivariety is generated by its finite members*;
(2) *any finitely presented K-algebra is residually-finite*;
(3) *K as a universal class is generated by its finite members.*

A quasivariety K is *Q-universal* if for any quasivariety M a lattice $L_q(M)$ is a factor-lattice of a sublattice in $L_q(K)$. Denote by $P_{fin}(\omega)$ the lattice of all finite subsets in ω, and by $P(X)$ the lattice of all subsets of a set X.

THEOREM 3.13 [2]. *Suppose that there exists a family of finite algebras A_X, $X \in P_{fin}(\omega)$ in a quasivariety K such that*

(1) *$A_\emptyset$ is a one-element algebra*;
(2) *if $X = Y \cup Z$, then A_X belongs to a subquasivariety generated by A_Y, A_Z*;

(3) *if A_X belongs to a quasivariety generated by A_Y and the set X is nonempty, then $Y = X$;*

(4) *if A_X be a subalgebra of $B \times C$, where B, C are in the quasivariety generated by all $A_Y, Y \in P_{fin}(\omega)$, then there exist subsets $Z, T \in P_{fin}(\omega)$ such that $X = Z \cup T$ and A_Z, A_T belong to the quasivariety, generated by B and C, respectively.*

Then K is Q-universal.

THEOREM 3.14 [2]. *A quasivariety K is Q-universal if one the following conditions holds*:

(1) *there is an infinite set of hereditary simple finite simple algebras in K which are nonembedable in each other and every congruence in a direct product of K-algebras is decomposable*;

(2) *K is a quasivariety of algebras with two unary operations f, g and one nullary operation e which is defined by quasi-identities*

$$fg(x) = x, \qquad gf(x) = x,$$
$$\big(f(x) = x\big) \wedge \big(g(x) = x\big) \Leftrightarrow (x = e).$$

Some interesting new approaches to the study of quasivarieties are exposed in the paper [12].

DEFINITION 3.15. A quasivariety of algebras V is *categorical* if for any cardinal α any two V-algebras of cardinality α are isomorphic.

A complete description of categorical quasivarieties of algebras (and algebraic systems) was obtained by E.A. Palyutin [21]. Let G be a group and E a family of subgroups in E such that

- if $H \in E$ and $g \in G$, then $gHg^{-1} \in E$;
- if $H_1, H_2 \in E$ and $H_1 \cap H_2 \neq 1$, then $H_1 = H_2$;
- if $H \in E$ and $g \in G$, then $gHg^{-1} \neq H$.

Suppose that n is a positive integer and κ is an cardinal. Given the data (n, κ, G, E) we can define a system of operations Ω consisting of unary operations $g \in G$ and $g_1, \dots, g_n$, nullary operations c_H, $H \in E$ and c_α, $\alpha < \kappa$, and an n-ary operation f. The quasivarieties of Ω-algebras $A_1, \dots, A_{15}$ are defined respectively by the following quasi-identities:

$A_1:$ $h_1(h_2(g_1(x))) = (h_1h_2)(g_1(x))$, where $h_1, h_2 \in G$;
$A_2:$ $1(g_1(x)) = g_1(x)$;
$A_3:$ $h(c_H) = c_{hHh^{-1}}$, where $h \in G, H \in E$;
$A_4:$ $h(x) = x \to x = c_H$, where $h \in H \setminus 1, H \in E$;
$A_5:$ $h(x) = x \to x = y$, where h does not belong to any $H \in E$;
$A_6:$ $h(c_\alpha) = c_\beta \to x = c_H$, where $h \in G \setminus 1$ and $\alpha, \beta < \kappa$;
$A_7:$ $h(c_\alpha) = c_H \to x = c_H$, where $h \in G, \alpha < \kappa$ and $H \in E$;
$A_8:$ $c_{H_1} = c_{H_2} \to x = y$, where $H_1 \neq H_2 \in G$;
$A_9:$ $g_1(c_\tau) = c_\tau$, where $\tau \in E \cup \kappa$;
$A_{10}:$ $g_1(h(x)) = h(x)$, where $h \in G$;
$A_{11}:$ $h(x) = h(g_1(x))$, where $h \in G$;

$A_{12}:\ f(g_1(x),\dots,g_1(x))=x$;
$A_{13}:\ g_i(f(g_1(x_1),\dots,g_1(x_n)))=g_1(x_i)$, where $i=1,\dots,n$;
$A_{14}:\ g-i(g_j(x))=g_j(x)$, where $1\leqslant i,j\leqslant n$;
$A_{15}:\ f(x_1,\dots,x_n)=f(g_1(x_1),\dots,g_n(x_n))$.

THEOREM 3.16 [21]. *Each quasivariety $A_1,\dots,A_{15}$ of Ω-algebras is categorical.*

We shall now present second series B of categorical quasivarieties. Let n,κ,f be as above. The set of basic operations Ω consists now of a binary addition $+$, an n-ary operation f, unary operations $g_1,\dots,g_n$ and from a fixed division ring D, nullary operations $c_\alpha,\alpha<\kappa$. The quasivarieties of Ω-algebras $B_1,\dots,B_{15}$ are defined respectively by the following quasi-identities:
$B_1:\ x+y=y+x$;
$B_2:\ x+(y+z)=(x+y)+z$;
$B_3:\ x+c_0=g_1(x)$;
$B_4:\ 0(x)=c_0$;
$B_5:\ \lambda_1(\lambda_2(x))=(\lambda_1\lambda_2(x))$, where $\lambda_1,\lambda_2\in D$;
$B_6:\ \lambda(x+y)=\lambda x+\lambda y$, where $\lambda\in D$;
$B_7:\ \lambda_1(x)+\lambda_2(x)=(\lambda_1+\lambda_2)(x)$, where $\lambda_1,\lambda_2\in D$;
$B_8:\ 1(x)=g_1(x)$;
$B_9:\ \lambda_1c_{\alpha_1}+\dots+\lambda_kc_{\alpha_k}=c_0\to x=y$, where $0<\alpha_1<\dots<\alpha_k<\kappa$ and $\lambda_1,\dots,\lambda_k\in D\setminus 0$;
$B_{10}:\ g_1(x+y)=x+y$;
$B_{11}:\ g_1(\lambda(x))=\lambda(x)$, where $\lambda\in D$;
$B_{12}:\ g_1(c_\alpha)=c_\alpha,\ \alpha<\kappa$;
$B_{13}:\ x+y=g_1(x)+g_1(y)$;
$B_{14}:\ \lambda(x)=\lambda(g_1(x)),\ \lambda\in D$;
$B_{15}:$ axioms A_{12}–A_{15}.

THEOREM 3.17 [21]. *Each quasivariety B_1–B_{15} is categorical.*

Now we consider a final series C_1–C_{29} of categorical quasivarieties. Let n,κ,D,f be as above and $\varepsilon=\pm1$. The set Ω of basic operations consists of a ternary operation S, an n-ary operation f, binary operations $g_1,\dots,g_n$ and $P_\lambda,\ \lambda\in D$, unary operations $R_\alpha,\ \alpha<\kappa$. If $\varepsilon=1$, then there is also a nullary operation 0. The quasivarieties of Ω-algebras $C_1,\dots,C_{29}$ are defined in [21, pp. 164, 165]. We shall omit the precise definition because it is too complicated. It is proved that these quasivarieties are also categorical.

THEOREM 3.18 [21]. *Let Q be a categorical quasivariety of algebras. Then Q is rationally equivalent* (*in the sense of A.I. Malcev*) *to one of quasivarieties $A_1,\dots,A_{15}$, $B_1,\dots,$ $B_{15},C_1,\dots,C_{29}$.*

3.2. *Quasivarieties of special algebras*

Let A be the variety of all Abelian groups. Then the lattice of subquasivarieties $L_q(A)$ is distributive [A.A. Vinogradov, 1965]. Let V be an *almost Abelian* quasivariety of groups, that is V contains a non-Abelian group and any proper subquasivariety in V contains only Abelian groups.

THEOREM 3.19 [6]. *The lattice $L_q(V)$ is distributive. A locally finite quasivariety of groups V is almost Abelian if and only if V coincides with one of the following quasivarieties*:

(1) *the quasivariety generated by the group*

$$F = \langle a, b \mid a^{p^n} = b^{p^n} = [a, b]^p = [a, b, a] = [a, b, b] = 1\rangle,$$

where p is a prime and p is odd if $n = 1$;

(2) *the quasivariety generated by an group F which is generated by the elements*

$$\{a_0, \dots, a_{p^n-1}\}$$

of a Galois field $\mathbb{F}_{p^n}$ and an element b subject to the defining relations

$$a_i^p = b^{q^m} = 1, \quad a_i a_j = a_{i+j}, \quad b a_i b^{-1} = \omega(a_i),$$

where p, q are distinct primes, n is the minimal integer such that $q|(p^n - 1)$, and the order of $\omega \in \mathbb{F}^*_{p^n}$ is equal to q.

There exists a locally finite quasivariety of nilpotency class 2 groups whose lattice of subquasivarieties is not modular (and distributive) [6].

DEFINITION 3.20. A *unary* algebra is a general algebra whose system of operations T consists only of unary operations that is $T = T_1$. A *unar* is a unary algebra with one unary operation f.

Several authors considered properties of quasivarieties of unary algebras.

DEFINITION 3.21. A system of quasi-identities Φ_j, $j \in J$, is *independent* in a quasivariety K if for any $j \in J$ there exists an algebra $A_j \in K$ such that all quasi-identities except Φ_j are satisfied in A_j.

Varieties of unars can be easily classified since any identity of unars has one of the forms $f^m(x) = f^n(x)$ or $f^n(x) = f^n(y)$, [19, §13]. In contrast there exist continuum many quasivarieties of unars, having no independent basis of quasi-identities [15]. Any finitely generated unar has an independent basis of quasi-identities [14]. There is also given a classification of quasivarieties of unars K such that the lattice $L_q(K)$ satisfies (4), is Boolean, or is pseudocomplemented.

Kurinnoi [16] considered properties of congruences of unars in quasivarieties. He classifies all unars A which satisfy the following equivalent properties:

(1) if B is an unar and $\operatorname{Con} A \simeq \operatorname{Con} B$ then the quasivarieties generated by B and A coincide;

(2) if B is an unar and $\operatorname{Con} A \simeq \operatorname{Con} B$ then the varieties generated by B and A coincide.

Some special results are obtained in [13]. There are given 3-element algebras A_1, A_2, A_3 such that A_{i+1} is obtained from A_i by addition of one nullary operation to the system of operations of A_i. It is shown that the quasivariety generated by A_2 has infinitely many subquasivarieties while each quasivariety generated by A_1, or A_3 has finitely many subquasivarieties. The varieties generated by these algebras are congruence-distributive.

Let $C_4 = \{0 < a < a' < 1\}$ be the 4-element de Morgan algebra. It is shown in [8] that the quasivariety generated by C_4 is minimal. The variety of de Morgan algebras generated by C_4 consists of Kleene algebras. It has some special properties for congruences of its algebras.

Let $A = \{0, a, 1\}$ be a chain with unary operation $\bar{x}$, where $\bar{0} = 1,\ \bar{1} = 0,\ \bar{a} = a$. Then A is a Kleene algebra. Denote by K the quasivariety of Kleene algebras generated by A. It is shown in [1] that the lattice $L_q(K)$ has continuum many members and a free lattice of a countable rank is embedable in $L_q(K)$.

3.3. *Generalization*

An important tool in the study of general algebras and classes of algebras is the notion of a clone.

NOTATION 3.22. Let X be a set and $\mathcal{O}_n(X)$ a set of all n-ary algebraic operations on the set X. Put

$$\mathcal{O}(X) = \mathcal{O} = \bigl\{\mathcal{O}_n(X) = \mathcal{O}_n \mid n \geqslant 0\bigr\}.$$

We can define in $\mathcal{O}$ an operation of superposition

$$\mathcal{O}_m \times \underbrace{\mathcal{O}_n \times \cdots \times \mathcal{O}_n}_{m} \to \mathcal{O}_n, \qquad (f, g_1, \dots, g_m) \rightsquigarrow f(g_1, \dots, g_m),$$

by setting

$$f(g_1, \dots, g_m)(x_1, \dots, x_n) = f\bigl(g_1(x_1, \dots, x_n), \dots, g_m(x_1, \dots, x_n)\bigr).$$

There are also special elements $p_{in} \in \mathcal{O}_n,\ i = 1, \dots, n$, called *projections* which are defined as follows

$$p_{in}(x_1, \dots, x_n) = x_i.$$

It is an easy exercise to check that

$$f(g_1,\dots,g_m)(h_1,\dots,h_n)=f\big(g_1(h_1,\dots,h_n),\dots,g_m(h_1,\dots,h_n)\big),$$
$$f=f(p_{1m},\dots,p_{mm})=f,\qquad p_{in}(g_1,\dots,g_n)=g_i \tag{5}$$

for any $f\in\mathcal{O}_m$ $g_1,\dots,g_m\in\mathcal{O}_n$, $h_1,\dots,h_n\in\mathcal{O}_s$.

DEFINITION 3.23. A clone A is a family of sets $A=\{A_n\mid n\geqslant 0\}$ with operations of superposition $f(g_1,\dots,g_m)$ as above and fixed elements $p_{in}\in A_n$, $i=1,\dots,n$, which satisfy the identities (5).

One of the main results concerning clones is due to I. Rosenberg (see [3]). Let X be a finite set. Then any subclone in $\mathcal{O}(X)$ is contained in a maximal one. I. Rosenberg classified maximal subclones in $\mathcal{O}(X)$. Suppose that R is an n-ary relation in the set X that is a subset in X^n.

DEFINITION 3.24. An operation $f\in\mathcal{O}$ *preserves* a relation R if R is a subalgebra of $(X,f)^n$.

THEOREM 3.25 (I. Rosenberg). *A subclone C in $\mathcal{O}$ is maximal if and only if it preserves one of the following relations*:

(1) *a relation R of partial order with largest and smallest elements*;
(2) *a relation R of all pairs $(x,p(x))$, where p is a permutation of prime order on X without fixed elements*;
(3) *a relation R consisting of all quadruples (x,y,z,t) related to some structure of an elementary Abelian p-group in X, such that $a+z=y+t$*;
(4) *an equivalence relation R on X*;
(5) *an m-ary relation R, $1\leqslant m\leqslant |X|$, such that*
   (a) *if $(x_1,\dots,x_m)\in R$ and s is a permutation of degree m then*
   $$(x_{s1},\dots,x_{sm})\in R;$$
   (b) *R contains all m-tuples $(x_1,x_2,\dots,x_m)$ with $x_1=x_2$*;
   (c) *there exists a nonempty subset Y in X such that $Y\times X^{m-1}\subseteq R$*;
(6) *a relation $R\subseteq X^m$, $3\leqslant m\leqslant |X|$, such that there exist equivalences*
   $$R_1,\dots,R_m$$
   on X with the following properties:
   (a) *if $H_1,\dots,H_m$ are classes of $R_1,\dots,R_m$ equivalent elements, respectively, then $H_1\cap\dots\cap H_m$ is nonempty*;
   (b) *R consists of all tuples $(x_1,\dots,x_m)\in X^m$ such that for any index $i=1,\dots,m$ there exist indices $1\leqslant p\neq q\leqslant m$ such that $(x_p,x_q)\in R_i$.*

DEFINITION 3.26. A finite algebra A is *preprimal* if the clone of term operations $T(A)$ on A is maximal in the clone $O(A)$ of all operations on A.

K. Denecke and O. Lüders [7] considered a variety V generated by a preprimal algebra. Suppose that a prevariety K as a category is equivalent to V. Then K is a variety generated by a preprimal algebra. Moreover preprimal algebras generating V and K have the same type in Rosenberg's classification of maximal clones of operations on finite sets.

Let K be a prevariety generated by a preprimal algebra A. Then K as a category is equivalent to a prevariety generated by a preprimal algebra B of the same type as A and in addition

(1) if A has type (1) from Rosenberg's classification, then $|B| = 2$;
(2) if A has type either (2) or (3) then $|B| = 3$;
(3) if A has type (4) then $|B|$ is prime.

There is given some detailed information on B of a similar nature for the other types.

A related result was proved in [17]. In 1941 E.L. Post classified all two-element algebras up to term equivalence. It terms of clones this means that subclones in $\mathcal{O}(X)$ for a two-element set X were classified. It was shown that there exist 31 classes of these algebras. Two algebras A, B are categorically equivalent if the varieties $V(A), V(B)$ generated by A and B, respectively, are equivalent as categories. The paper presents a list of algebras B which are categorical equivalent to some two element algebra A.

An interesting result is published in [26]. For any positive integer n there exists a calculus G_n, extending the language L_K of the class of all T-algebras with the following properties: if A is a finite T-algebra of order at most n, then A is finitely axiomatizable within the language G_n.

There some interesting results concerning pseudovarieties of algebras that are related to quasivarieties.

DEFINITION 3.27. An abstract class of T-algebras K is a *pseudovariety* if K is nonempty and it is closed under subalgebras, homomorphic images and finite direct products.

Let M, N be two monoids such that N acts M. Denote by $M \rtimes N$ the corresponding semidirect product. If V, W are pseudovarieties of monoids, then VW denotes the pseudovariety generated by all semidirect products $M \rtimes N$ where $M \in V$ and $N \in W$. It is shown in [5] that if W is the pseudovariety of all finite p-groups then there exists a set of identities $\Sigma_{p,k},\ k \geqslant 1$, such that a monoid X belongs to VW if and only if X satisfies the identities $\Sigma_{p,k}$ for almost all k.

Let H be a pseudovariety of groups which is closed under semidirect products and P the pseudovariety of semigroups whose groups belong to H. Then the lattice of all pseudovarieties the member P is indecomposable into a direct product and into a Malcev product [20]. In particular there is no proper pseudovariety of semigroups.

A pseudovariety V is *regular* if there exists a base of identities in V such that in every identity $p = q$ the sets of variables occurring in p and in q are the same. The set of regular pseudovarieties of T-algebras forms a sublattice in the lattice of all pseudovarieties of T-algebras. For any pseudovariety V denote by $R(V)$ the least regular pseudovariety V containing V. The properties of $R(V)$ are investigated in [11].

References

[1] M.E. Adams, *Lattices of quasivarieties of 3-element algebras*, J. Algebra **166** (1) (1994), 181–210.
[2] M.E. Adams and W. Dziobiak, *Q-universal quasivarieties of algebras*, Proc. Amer. Math. Soc. **120** (4) (1994), 1053–1059.
[3] V.A. Artamonov, *Universal algebra*, Handbook in Algebra, Vol. 2, Nauka, Moscow (1991), Ch. VI.
[4] V.A. Artamonov, *Varieties of algebras*, Handbook of Algebra, Vol. 2, M. Hazewinkel, ed., Elsevier, Amsterdam (2000), 3–29.
[5] F. Blanchet-Sardi, *On the semidirect product of the pseudovariety of semilattices by a locally finite pseudovariety of groups*, Inf. Theor. Appl. **31** (3) (1997), 237–250.
[6] A.I. Budkin and V.A. Gorbunov, *On the theory of quasivarieties of algebraic systems*, Algebra i Logika **14** (2) (1975), 123–142.
[7] K. Denecke and O. Lüders, *Category equivalences and dualities of varieties and prevarieties generated by single preprimal algebras*, Acta Sci. Math. **58** (1–4) (1993), 75–92.
[8] H. Gaitan, *Quasivarieties of De Morgan algebras*, Notas Log. Mat. **39** (1994), 121–129.
[9] V.A. Gorbunov, *Algebraic Theory of Quasivarieties*, NIIMIOO, Novosibirsk (1998).
[10] G. Gratzer, *Universal Algebra*, Springer, Berlin (1979).
[11] E. Graczyńska, *On some operators on pseudovarieties. II*, Bull Sect. Logic Univ. Lodź. **24** (2) (1995), 80–88.
[12] H.-J. Hoenke, *Remarks and problems on varieties and quasi-varieties, clones and quasi-clones*, Discuss. Math. Algebra Stochast. Methods **16** (1) (1996), 11–56.
[13] K. Idziak, *Quasivarieties of equational algebras*, Rep. Math. Logic **29** (1995), 81–85.
[14] V.K. Kartashov, *Quasivarieties of unars with finitely many cycles*, Algebra and Logic **19** (2) (1980), 173–193.
[15] V.K. Kartashov, *On lattices of quasivarieties of unars*, Siberian Math. J. **26** (3) (1985), 49–62.
[16] G.Ch. Kurinnoi, *Congruences and quasi-identities in unars*, Ukrainian Math. J. **49** (6) (1997), 852–856.
[17] O. Lüders, *Categorical equivalence*, General Algebra and Applications in Discrete Mathematics, K. Denecke et al., eds, Proceedings of the Conference on General Algebra and Discrete Mathematics, Potsdam, Germany, June 1996, Shaker Verlag, Aachen, Berichte aus der Mathematik (1997), 137–152.
[18] A.I. Malcev, *Selected Proceedings, Vol. 2, Mathematical Logic and General Theory of Algebraic Systems*, Nauka, Moscow (1976).
[19] A.I. Malcev, *Algebraic Systems*, Nauka, Moscow (1970).
[20] S. Margolis, M. Sapir and P. Weil, *Irreducibility of certain pseudovarieties*, Comm. Algebra **26** (3) (1998), 779–792.
[21] E.A. Palyutin, *A description of categorical quasivarieties*, Algebra i Logika **14** (2) (1975), 145–185.
[22] A. Shafaat, *On implicationally defined classes of algebras*, J. London Math. Soc. **44** (1) (1969), 137–140.
[23] S. Tulipani, *On the universal theory of classes of finite modules*, Trans. Amer. Math. Soc. **284** (1) (1984), 163–170.
[24] B.M. Vernikov, *Hereditary self-dual varieties and quasivarieties*, Izv. Vyssh. Uchebn. Zaved. Mat. **9** (1984), 25–29.
[25] B.M. Vernikov and M.V. Volkov, *Complementations in lattices of varieties and quasivarieties*, Izv. Vyssh. Uchebn. Zaved. Mat. **11** (1982), 17–20.
[26] A. Wroński, *Gentzen-type calculi for quasi-identities of finite algebras*, Rep. Math. Logic **28** (1994), 73–79.

This Page Intentionally Left Blank

Section 4E
Lie Algebras

This Page Intentionally Left Blank

Free Lie Algebras

Christophe Reutenauer

Département de mathématiques, Université du Québec à Montréal, Case postale 8888, succursale Centre-Ville, Montréal (Québec), Canada H3C 3P8

E-mail: christo@math.uqam.ca

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1. Introduction

Lie polynomials appeared at the end of the 19th century and the beginning of the 20th century in the work of Campbell, Baker and Hausdorff on exponential mapping in a Lie group, which has lead to the so-called Campbell–Baker–Hausdorff formula. Around 1930, Witt showed that the Lie algebra of Lie polynomials is actually the free Lie algebra, and that its enveloping algebra is the associative algebra of noncommutative polynomials. He proved what is now called the Poincaré–Birkhoff–Witt theorem, and showed how the free Lie algebra is related to the lower central series of a free group. About at the same time P. Hall and Magnus, with their commutator calculus, opened the way to bases of the free Lie algebra, by M. Hall.

In this chapter, our aim is to prove that the Lie algebra of Lie polynomials is indeed the free Lie algebra. We shall do this without using the Poincaré–Birkhoff–Witt theorem, but by constructing Hall bases of the free Lie algebras. The bases we consider are more general than the original ones; they include the Lyndon basis. We follow the method of Schützenberger (with improvements from Melançon). We shall also follow him to compute the dual basis.

In the Notes, we give statement of other results, and references for them, together with indications on recent work related to free Lie algebras.

In all that follows, **A** *is a commutative ring with unit*. In Section 14, we assume that it contains **Q**.

2. Free Lie algebra

Recall that a *Lie algebra* over **A** is an **A**-module L equipped with a bilinear mapping $L \times L \to L$, $(x, y) \mapsto [x, y]$, satisfying the two following properties, for any x, y, z in L:

$$[x, x] = 0,$$
$$\bigl[[x, y], z\bigr] + \bigl[[y, z], x\bigr] + \bigl[[z, x], y\bigr] = 0.$$

The latter identity is called the *Jacobi identity*. Note that the first one implies *antisymmetry*, that is,

$$[x, y] = -[y, x],$$

since, using bilinearity, we have $0 = [x + y, x + y] = [x, x] + [x, y] + [y, x] + [y, y] = [x, y] + [y, x]$. In view of antisymmetry, we may rewrite the Jacobi identity as

$$\bigl[[x, y], z\bigr] = \bigl[x, [y, z]\bigr] + \bigl[[x, z], y\bigr].$$

This identity, which will be useful later, means that $x \mapsto [x, z]$ is a derivation of the Lie algebra L.[1]

[1] A *derivation* of L is a linear mapping $L \to L, x \mapsto x'$, that satisfies the Leibnitz identity, that is $[x, y]' = [x, y'] + [x', y]$.

Homomorphisms, isomorphisms and Lie subalgebras of Lie algebras are defined as usual.

Given a set X, a *free Lie algebra* on X over $\mathbf{A}$ is a Lie algebra L over $\mathbf{A}$, together with a mapping $i: X \to L$, with the following *universal property*: for each Lie algebra K and each mapping $f: X \to K$, there exists a unique Lie algebra homomorphism $F: L \to K$ such that $f = F \circ i$.

A standard argument shows that a free Lie algebra on X is necessarily unique, up to Lie algebra isomorphism.

Its existence is shown as follows: a *tree* on X is a well-formed expression on X, recursively defined by: each x in X is a tree; if t_1, t_2 are trees, then so is $t = (t_1, t_2)$ (the terminology comes from the fact that each tree may be identified with a binary, complete, rooted and planary tree with leaves labelled by X). Let $M(X)$ denote the set of all trees on X. The mapping which sends the pair of trees t_1, t_2 to t as above is a binary law on $M(X)$. Let $D(X)$ denote the free $\mathbf{A}$-module with basis $M(X)$. Then the previous law defines a bilinear mapping $D(X) \times D(X) \to D(X)$, still denoted $(\ ,\ )$. An ideal I of $D(X)$ is a submodule such that for any $u \in I,\ v \in M(X)$, one has $(u, v), (v, u) \in I$. Let I be the ideal generated by all the elements of the form $[t, t]$, or $[[t_1, t_2], t_3] + [[t_2, t_3], t_1] + [[t_3, t_1], t_2]$, $t, t_1, t_2, t_3 \in M(X)$. Then the quotient $D(X)/I$ is clearly a Lie algebra. Furthermore, it is the free Lie algebra on X over $\mathbf{A}$. Indeed, define $i: X \to D(X)/I$ by $i(x) = x \bmod I$; now, if K is a Lie algebra with a mapping $f: X \to K$, then f clearly extends uniquely to a mapping $M(X) \to K$, by replacing each tree by the corresponding Lie bracketing, with $x \in X$ replaced by $f(x)$. By linear extension, this mapping extends to $D(X)$; the kernel of the latter clearly contains I, since K is a Lie algebra, hence we obtain a Lie homomorphism $F: D(X)/I \to K$, which satisfies $f = F \circ i$.

In other words, a Lie algebra over $\mathbf{A}$ is free on X if and only if it is generated by all possible Lie bracketings of elements of X (these bracketings are in bijection with trees), and if the only possible relations existing among these bracketings are consequence of the bilinearity of the bracketing, of the Jacobi identity and the identity $[u, u] = 0$.

3. Noncommutative polynomials

Given a set X of *noncommuting variables* (also called *letters*), denote by X^* the set of words on X, where by *word* we mean a finite sequence of elements of X, or equivalently, a *noncommutative monomial*; we include the *empty word*, denoted 1. With the product defined by concatenation of words, X^* becomes a monoid, which is the *free monoid* on X.

Indeed, for any mapping $f: X \to M$, where M is any monoid, there is a unique homomorphism of monoids $F: X^* \to M$ such that $f = F \circ i$, where i is the natural injection $X \to X^*$.

A *noncommutative polynomial* (we shall say polynomial) is a linear combination of words with coefficients in $\mathbf{A}$. The set they form is denoted $\mathbf{A}\langle X\rangle$. With the product inherited from the free monoid, it becomes an $\mathbf{A}$-algebra; note that it is the $\mathbf{A}$-algebra of the monoid X^*. Equivalently, it is the *tensor algebra* of the free $\mathbf{A}$-module with basis X.

This $\mathbf{A}$-algebra $\mathbf{A}\langle X\rangle$ is the *free associative* $\mathbf{A}$*-algebra* on X; indeed, for any mapping $f: X \to A$, where A is any $\mathbf{A}$-algebra, there is a unique homomorphism of $\mathbf{A}$-algebras $F: \mathbf{A}\langle X\rangle \to A$ such that $f = F \circ i$, where i is the natural injection $X \to \mathbf{A}\langle X\rangle$.

4. Lie polynomials

For any polynomials P, Q in $\mathbf{A}\langle X\rangle$, define their *Lie bracket* by $[P, Q] = PQ - QP$. This defines a structure of Lie algebra on $\mathbf{A}\langle X\rangle$. Indeed, antisymmetry is immediate and a simple calculation shows that the Lie bracket satisfies the Jacobi identity.

Let $L_{\mathbf{A}}(X)$ (or simply $L(X)$) denote the smallest Lie subalgebra of $\mathbf{A}\langle X\rangle$ that contains each letter in X. It will be shown that this Lie algebra is free. A *Lie polynomial* is an element of $L(X)$.

5. Hall sets

Each tree t in the free magma $M(X)$ is either a letter $t \in X$, or is of the form $t = (t_1, t_2)$, for some trees t_1, t_2. We then write $t' = t_1$, $t'' = t_2$, and call t', t'' the *left, right immediate subtree* of t.

Consider a total order $>$ on $M(X)$ such that for any tree t, one has $t < t''$. Note that such orders surely exist. We define the *Hall set* H relative to this order recursively by: each letter is in H; if a tree t is not a letter, then it is in H if and only if t', t'' are in H, $t' < t''$ and: either t' is in X or $(t')'' \geqslant t''$. Elements of H are called *Hall trees*.

In the sequel, we fix a Hall set H.

6. Standard sequences

Given a Hall set H, a *standard sequence* is a sequence of Hall trees $(t_1, \dots, t_n)$ with $n \geqslant 0$ and: for any i, either t_i is a letter, or $t_i'' \geqslant t_{i+1}, \dots, t_n$.

Clearly, each sequence of letters is standard. Moreover, a *decreasing sequence* (that is, a sequence such that $t_1 \geqslant \cdots \geqslant t_n$) of Hall trees is always standard: indeed, if t_i is not a letter, then $t_i'' > t_i \geqslant t_{i+1}, \dots, t_n$.

We call a *rise* of a sequence an index i such that $t_i < t_{i+1}$; in that case, we also say that (t_i, t_{i+1}) is a rise. An *inversion* of the sequence is a couple (i, j) such that $i < j$ and $t_i < t_j$ (note the opposite inequality to the classical one for inversions of permutations); here also, we say that (t_i, t_j) is an inversion. Observe that a sequence is decreasing if and only if it has no inversion, or equivalently, no rise. A *legal rise* is a rise i such that $t_{i+1} \geqslant t_{i+2}, \dots, t_n$.

We now define a rewriting system on the set of standard sequences. Let $s = (t_1, \dots, t_n)$ be a standard sequence and i some legal rise of s. We write $s \to s'$ if $s' = (t_1, \dots, t_{i-1}, (t_i, t_{i+1}), t_{i+2}, \dots, t_n)$. In other words, s' is obtained from s by multiplying in the free magma $M(X)$ the two trees that form the chosen legal rise. Note that s' is a sequence of Hall trees: indeed, either t_i is in X, or $t_i = (t_i', t_i'')$ and then $t_i'' \geqslant t_{i+1}$ since s is standard; this shows that (t_i, t_{i+1}) is in H since $t_i < t_{i+1}$, i being a rise. Furthermore, s' is standard: indeed, if $j = 1, \dots, i-1$, then either t_j is a letter, or $t_j'' \geqslant t_{i+1}$ (since s is standard) $> (t_i, t_{i+1})$, by assumption on the order $>$. Moreover, $t_{i+1} \geqslant t_{i+2}, \dots, t_n$, since i is a legal rise; thus s' is a standard sequence, since s is.

We denote by $\to^*$ the reflexive and transitive closure of the binary relation $\to$.

THEOREM 1.

(1) *For any standard sequences s, s_1, s_2 such that $s \to^* s_1$ and $s \to^* s_2$, there exists a standard sequence r such that $s_1 \to^* r$ and $s_2 \to^* r$.*
(2) *For each standard sequence s, there exists a sequence of letters r such that $r \to^* s$.*
(3) *For each standard sequence s, there exists a decreasing standard sequence t such that $s \to^* t$.*

PROOF. (1) Assume that $s \to s_1$ and $s \to s_2$. Write $s = (t_1, \ldots, t_n)$, $s_1 = (t_1, \ldots, t_{i-1}, (t_i, t_{i+1}), t_{i+2}, \ldots, t_n)$ and $s_2 = (t_1, \ldots, t_{j-1}, (t_j, t_{j+1}), t_{j+2}, \ldots, t_n)$, where i, j are legal rises of s. We may assume that $i < j$. Then $i + 1 < j$: indeed, otherwise, $i + 1 = j$; hence $t_{i+1} = t_j$, and $t_{i+1} < t_{j+1}$ since j is a rise; moreover, i is a legal rise, so that $t_{i+1} \geqslant t_{i+2} = t_{j+1}$, a contradiction.

Thus (t_i, t_{i+1}) is a rise of s_2 and (t_j, t_{j+1}) is a rise of s_1. We show that (t_j, t_{j+1}) is a legal rise of s_1, and that (t_i, t_{i+1}) is a legal rise of s_2. The first assertion is clear, since if t is at the right of t_{j+1} in s_1, then it is also in s; hence $t_{j+1} \geqslant t$, because j is a legal rise of s. For the second assertion, we have $t_{i+1} \geqslant t_{j+1}$ (since i is a legal rise of s) $> (t_j, t_{j+1})$ by the property of the order; hence t_{i+1} is greater or equal to each tree at its right in the sequence s_2, since this is true in s.

Define now $r = (t_1, \ldots, t_{i-1}, (t_i, t_{i+1}), t_{i+2}, \ldots, t_{j-1}, (t_j, t_{j+1}), t_{j+2}, \ldots, t_n)$. Then we have $s_1 \to r$ and $s_2 \to r$ by definition of $\to$.

This being done, the first assertion of the theorem follows by a straightforward induction on the lengths of the chains from s to s_1 and s to s_2.

(2) Let $s = (t_1, \ldots, t_n)$. If s is not a sequence of letters, then consider i such that t_i is not a letter, and for any $j = 1, \ldots, i-1$, t_j is a letter: such an i surely exists. Then let $r' = (t_1, \ldots, t_{i-1}, t_i', t_i'', t_{i+1}, \ldots, t_n)$. The sequence r' is standard: indeed, either t_i' is a letter, or $(t_i')'' \geqslant t_i''$ (since (t_i', t_i'') is a Hall tree) $\geqslant t_{i+1}, \ldots, t_n$, because s is standard; moreover, either t_i'' is a letter, or $(t_i'')'' > t_i''$ (by the property of the order) $\geqslant t_{i+1}, \ldots, t_n$; finally, for $j = 1, \ldots, i-1$, t_j is a letter. Thus r' is standard since s is. Moreover, (t_i', t_i'') is a rise of r', because $t_i' < t_i''$ (since (t_i', t_i'') is a Hall tree), which is legal, since $t_i'' \geqslant t_{i+1}, \ldots, t_n$. Thus $r' \to s$. We conclude by induction on the maximum degree of the trees in s.

(3) If s is not decreasing, it has at least one rise. Choose i to be the rightmost one. Then $t_i < t_{i+1} \geqslant t_{i+2} \geqslant \cdots \geqslant t_n$. Hence this rise is legal and we obtain a shorter sequence s' such that $s \to s'$. We conclude by induction on the length of the sequence. □

7. Hall words

There is a canonical mapping from $M(X)$ onto X^*, denoted f; it is defined inductively by $f(x) = x$ if $x \in X$, and $f(t) = f(t')f(t'')$ if $t = (t', t'')$. For example, $f((y, ((x, y), z))) = yxyz$.

THEOREM 2. *Each word in X^* has a unique factorization $f(t_1) \ldots f(t_n)$, with $n \geqslant 0$, $t_i \in H$ and $t_1 \geqslant \cdots \geqslant t_n$.*

PROOF. Extend the mapping f to sequences by $f(t_1, \ldots, t_n) = f(t_1) \ldots f(t_n)$. Clearly, $s \to^* r$ implies $f(s) = f(r)$. For a word w, the sequence s of its letters is standard. By

Theorem 1, we have $s \to^* t$ for some decreasing sequence of Hall trees; hence w admits a factorization as in the statement.

Suppose it has another such factorization: $w = f(h_1)\dots f(h_p)$. Consider the standard sequences $s = (t_1, \dots, t_n)$ and $r = (h_1, \dots, h_p)$. By the theorem, we find sequences of letters s', r' such that $s' \to^* s$ and $r' \to^* r$. We have $f(s') = f(s) = w = f(r) = f(r')$. Thus $s' = r'$. Hence by Theorem 1 again, we have $s \to^* u$ and $r \to^* u$ for some standard sequence u. Since s, r are nondecreasing, hence have no rise, this is possible only if $s = r$. This proves unicity. □

Given a Hall set H, we call *Hall word* each word of the form $f(t)$, $t \in H$.

COROLLARY 1. *For each Hall word w, there exists exactly one Hall tree t such that $w = f(t)$.*

COROLLARY 2. *Each word has exactly one decreasing factorization into Hall words.*

8. Poincaré–Birkhoff–Witt basis

There is a canonical mapping g from $M(X)$ into $L(X)$. It is defined by $g(x) = x$ if $x \in X$, and $g(t) = [g(t'), g(t'')]$ if $t = (t', t'')$. In other words, $g(t)$ is obtained by replacing in t parentheses by Lie bracketing. For example, $g((y, ((x, y), z))) = [y, [[x, y], z]]$.

THEOREM 3. *The products $g(t_1)\dots g(t_n)$, $n \geqslant 0$, $t_1 \geqslant \dots \geqslant t_n$, form a basis of the $\mathbf{A}$-module $\mathbf{A}\langle X\rangle$.*

PROOF. It is enough to show this when X is finite.

For each standard sequence $s = (t_1, \dots, t_n)$, with a legal rise i, define

$$\lambda_i(s) = \bigl(t_1, \dots, t_{i-1}, (t_i, t_{i+1}), t_{i+2}, \dots, t_n\bigr)$$

and

$$\rho_i(s) = (t_1, \dots, t_{i-1}, t_{i+1}, t_i, t_{i+2}, \dots, t_n).$$

The first sequence is the same as for the definition of the rewriting system in Section 6, and is obtained by multiplying in $M(X)$ the two trees that form the legal rise; the second is obtained by interchanging them.

Extend naturally g to sequences by: $g(s) = g(t_1)\dots g(t_n)$ if $s = (t_1, \dots, t_n)$. Since clearly $ab = [a, b] + ba$, we obtain that $g(s) = g(\lambda_i(s)) + g(\rho_i(s))$.

Note that $\lambda_i(s)$ is shorter than s, and that $\rho_i(s)$ has one inversion less.

Thus we obtain by induction that $g(s)$ is equal to a sum of products as in the statement. This being true for each sequence of letters, we deduce that these elements span the module $\mathbf{A}\langle X\rangle$.

Note that these elements are homogeneous. Now Corollary 2 shows that for fixed degree d, the elements that have degree d are as numerous as the words of length d; but the latter form a basis of the submodule of homogeneous polynomials of degree d. Hence the polynomials of the statement which are of degree d form a basis of that submodule: indeed, in a free $\mathbf{A}$-module of rank N, if a generating set has cardinality N, then it is a basis (since a right invertible square matrix over $\mathbf{A}$ is invertible).

This show that the whole collection of polynomials as in the statement form a basis of $\mathbf{A}\langle X\rangle$. □

COROLLARY 3. *Each word, when written in the basis of the theorem, has coefficients in* $\mathbf{N}$.

9. Hall bases

We can now prove that $L(X)$ is a free $\mathbf{A}$-module with basis $g(H)$.

THEOREM 4. *The elements $g(t)$, $t \in H$, form a basis of $L(X)$.*

PROOF. By Theorem 3, the polynomials $g(t)$ are linearly independent. It is enough to show that they span $L(X)$. Again, we may suppose that X is finite. Since $L(X)$ is generated as a Lie algebra by X, and X is contained in $g(H)$, it is enough to show that: for any t_1, t_2 in H, the polynomial $g((t_1, t_2)) = [g(t_1), g(t_2)]$ is a linear combination over $\mathbf{Z}$ of polynomials $g(t)$, $t \in H$, with $t'' \leqslant \max(t_1, t_2)$. Let $|t|$ denote the *degree* of a tree t, which is defined by: $|t| = 1$ if t is in X; and $|t| = |t'| + |t''|$ if $t = (t', t'')$. We prove the previous statement by induction on the couple $(|t_1| + |t_2|, \max(t_1, t_2))$ where these couples are ordered by: $(d, u) < (e, v)$ if and only if either $d < e$ or $d = e$ and $u < v$; note that this is correct, since there are only finitely many Hall trees of a given degree.

By antisymmetry of the Lie bracket, we may assume that $t_1 < t_2$. Now, if t_1 is in X or if $t_1 = (t_1', t_1'')$ with $t_1'' \geqslant t_2$, then $t = (t_1, t_2)$ is a Hall tree, and $g(t)$ is in $g(H)$; moreover, $t'' = t_2 \leqslant \max(t_1, t_2)$.

So we may assume that $t_1 = (t_1', t_1'')$ and that $t_1'' < t_2$. By the property of the order, we have $t_1 < t_1''$, hence $t_1 < t_1'' < t_2$. Moreover, $t_1' < t_1''$, hence $t_1' < t_1'' < t_2$.

By the Jacobi identity, we have

$$\begin{aligned} g\big((t_1, t_2)\big) &= \big[g(t_1), g(t_2)\big] = \big[\big[g(t_1'), g(t_1'')\big], g(t_2)\big] \\ &= \big[\big[g(t_1'), g(t_2)\big], g(t_1'')\big] + \big[g(t_1'), \big[g(t_1''), g(t_2)\big]\big]. \end{aligned}$$

Since $|t_1'| + |t_2|$ and $|t_1''| + |t_2|$ are both strictly smaller than $|t_1| + |t_2|$, the induction hypothesis shows that $[g(t_1'), g(t_2)] = \sum * g(u_i)$ and $[g(t_1''), g(t_2)] = \sum * g(v_j)$, where the $*$ indicate integers whose value is of no importance here, with the property that: $u_i'' \leqslant \max(t_1', t_2) = t_2$ and $v_j'' \leqslant \max(t_1'', t_2) = t_2$; moreover, by homogeneity, $|u_i| = |t_1'| + |t_2|$ and $|v_j| = |t_1''| + |t_2|$.

Thus we obtain

$$g\big((t_1, t_2)\big) = \sum * \big[g(u_i), g(t_1'')\big] + \sum * \big[g(t_1'), g(v_j)\big].$$

We have $|u_i| + |t_1''| = |t_1'| + |t_2| + |t_1''| = |t_1| + |t_2|$, and since $u_i < u_i'' \leqslant t_2$, $\max(u_i, t_1'') < t_2 = \max(t_1, t_2)$; thus by the induction hypothesis, we deduce that $[g(u_i), g(t_1'')]$ is a linear combination over $\mathbf{Z}$ of $g(t)$ with t in H and $t'' \leqslant \max(u_i, t_1'') < \max(t_1, t_2)$. Similarly, we have $|t_1'| + |v_j| = |t_1'| + |t_1''| + |t_2| = |t_1| + |t_2|$ and $\max(t_1', v_j) < t_2 = \max(t_1, t_2)$ since $v_j < v_j'' \leqslant t_2$. Thus by induction, $[g(t_1'), g(v_j)]$ is a linear combination of $g(t)$ with $t \in H$ and $t'' \leqslant \max(t_1', v_j) < \max(t_1, t_2)$. □

10. The algorithm

If we have a closer look to the previous proof, we see that it gives an algorithm which given any Lie bracketing of the letters, writes it as a linear combination over $\mathbf{Z}$ of Hall bracketings, that is, elements of the Hall basis $g(H)$.

More formally, the algorithm takes as input a linear combination of trees $S = \sum \alpha_t t$ and gives as output a linear combination of Hall trees $\sum \beta_h h$ such that $\sum \alpha_t g(t) = \sum \beta_h g(h)$. It works as follows: as a first step, look if each tree appearing in S is a Hall tree; then there is nothing to do and the algorithm stops. Otherwise, take some tree t appearing in S, which is not a Hall tree, and consider a subtree $s = (s', s'')$ of t which is not a Hall tree but such that s', s'' are Hall trees (s surely exists since letters are Hall trees). If $s' > s''$, replace s by (s'', s') in t, and replace α_t by $-\alpha_t$ in S. If $s' = s''$ then remove t from the linear combination. If $s' < s''$, note that since s is not in H, s' is not in X, and $s' = (a, b)$ with $b < s''$; now replace t in S by the sum of the two trees t_1, t_2 obtained as follows: t_1 is obtained by replacing in t the subtree s by $((a, s''), b)$ and t_2 is obtained by replacing s by $(a, (b, s''))$. Now go back to the first step of the algorithm.

Then this algorithm stops and does the desired job, as follows from the proof in the previous section.

Observe that if $S = \sum \alpha_t t$ and if $\sum \alpha_t g(t) = 0$, then, since the polynomials $g(h)$ are linearly independent, the algorithm outputs the 0 linear combination. As a consequence, we obtain the following result.

THEOREM 5. *$L(X)$ is the free Lie algebra on X.*

PROOF. There is a canonical surjective linear mapping $G : D(X) \to L(X)$, that sends each tree t onto the corresponding Lie bracketing $g(t)$ in $\mathbf{A}\langle X\rangle$. It suffices to show that $\operatorname{Ker} G = I$, with the notations of Section 2. Since $L(X)$ is a Lie algebra, $I \subset \operatorname{Ker} G$.

For the reverse inclusion, consider an element in $\operatorname{Ker} G$. It may be written as a linear combination $S = \sum \alpha_t t$ of trees. Then $0 = G(S) = \sum \alpha_t g(t)$. The previous algorithm shows the existence of a sequence of elements $S_0, \ldots, S_n$ of $D(X)$ such that $S_0 = S$, $S_n = 0$ and that $S_i \equiv S_{i+1} \bmod I$ for each i. Thus $S \in I$. □

11. Dimensions

We assume that X has cardinality q. Let α_n denote the dimension of the homogeneous part of degree n of $L(X)$. Equivalently, α_n is the number of Hall trees of degree n.

THEOREM 6. *One has* $\alpha_n = \frac{1}{n}\sum_{d|n}\mu(d)q^{n/d}$.

PROOF. This identity is equivalent to $n\alpha_n = \sum_{d|n}\mu(d)q^{n/d}$, hence by Möbius inversion to $q^n = \sum_{d|n} d\alpha_d$. By taking generating functions, this in turn is the same as

$$\sum_{n\geqslant 1} q^n s^n = \sum_{n\geqslant 1}\Bigl(\sum_{d|n} d\alpha_d\Bigr)s^n.$$

The latter is equal to $\sum_{k\geqslant 1} k\alpha_k \sum_{i\geqslant 1} s^{ki}$. Thus all we have to verify is $\frac{qs}{1-qs} = \sum_{k\geqslant 1}\frac{k\alpha_k s^k}{1-s^k}$. But this is a consequence of

$$\frac{1}{1-qs} = \prod_{k\geqslant 1}\frac{1}{(1-s^k)^{\alpha_k}}, \tag{$*$}$$

by taking logarithmic derivatives with respect to s and multiplying by s. Note that by Corollary 2, one has the following equality of noncommutative formal power series: $(1-\sum_{x\in X} x)^{-1} = \prod_h (1-f(h))^{-1}$, where the product is over all Hall words in strictly decreasing order. The identity $(*)$ follows from the latter by applying the homomorphism sending every letter to the same variable s. $\square$

12. The dual basis 1

Given any word w in X^*, it has a unique decreasing factorization $w = f(t_1)\dots f(t_n)$, where the t_i are Hall trees. We define $P_w = g(t_1)\dots g(t_n)$. Then the polynomials P_w form a basis of the $\mathbf{A}$-module $\mathbf{A}\langle X\rangle$ and the polynomials P_h, h a Hall word, form a basis of $L(X)$. In order to determine the dual basis, consider the scalar product $(\,,\,)$ on $\mathbf{A}\langle X\rangle$ for which X^* is an orthonormal basis. Since the previous basis is homogeneous, and even finely homogeneous with respect to each partial degree, there exist polynomials S_w which represent the dual basis with respect to this scalar product, that is, such that $P = \sum_{w\in X^*}(S_w, P)P_w$, for any polynomial P.

Note that $S_1 = 1$.

THEOREM 7. *If* $h = xv$ *is a Hall word with* $x\in X$, $v\in X^*$, *then* $S_h = xS_v$.

We know that Hall words are in one-to-one correspondence with Hall trees; we use this bijection to transfer definitions on standard sequences of Hall trees to sequences of Hall words. In particular, we transfer the total order on Hall trees to Hall words; moreover, if $h = f(t)$ is a Hall word, the image under f of the Hall tree t, we put $g(h) = g(t)$, and for a standard sequence of Hall words $s = (h_1,\dots,h_n)$, we extend g by $g(s) = g(h_1)\dots g(h_n)$, which is also equal to $P_{h_1}\dots P_{h_n}$ with our previous notation.

Furthermore for each standard sequence of Hall words s, which has at least one rise, choose a fixed rise $i(s)$. Now, define a relation $\Rightarrow$ on standard sequences by: $s\Rightarrow s'$ if $i = i(s)$ exists and if $s' = \lambda_i(s) = (h_1,\dots,h_{i-1},h_ih_{i+1},h_{i+2},\dots,h_n)$ or $s' = \rho_i(s) =$

$(h_1, \dots, h_{i-1}, h_{i+1}, h_i, h_{i+2}, \dots, h_n)$. Denote by $\Rightarrow^*$ the reflexive and transitive closure of $\Rightarrow$.

LEMMA 1. *Let s be a standard sequence $(h_1, \dots, h_n)$ of Hall words, with $n \geqslant 2$ and $h_2 \geqslant \cdots \geqslant h_n$.*

- *If $h_1 \dots h_n$ is a Hall word, there is exactly one chain $s \Rightarrow \cdots \Rightarrow (h_1 \dots h_n)$.*
- *Given a standard sequence s' such that $s \Rightarrow^* s'$ and such that $s' \neq (h_1 \dots h_n)$ (this inequality holds certainly if $h_1 \dots h_n$ is not a Hall word), s' is of length at least 2.*

PROOF. We claim that if $u = (u_1, \dots, u_m)$ is a standard sequence of Hall words with $m \geqslant 2$ and $u_1 \geqslant u_2, \dots, u_m$, then for each standard sequence v with $u \Rightarrow^* v$, v is of length at least 2. Indeed, if i is a rise of u, then $i \geqslant 2$. Thus $\rho_i(s)$ satisfies the same hypothesis as u. Moreover, so does $\lambda_i(s)$, since $u_{i+1} > u_i u_{i+1}$, because $u_i u_{i+1}$ is a Hall word, image under f of the Hall tree $t = (t', t'')$ with $f(t') = u_i$ and $f(t'') = u_{i+1}$. Thus we conclude by induction on the length of the chain from u to v.

We now prove the lemma. If $h_1 \geqslant h_2$, then there is no nontrivial chain starting from s. Thus we may assume that $h_1 < h_2$. This will be the only rise, so that $i(s) = 1$; moreover, $\rho_1(s) = (h_2, h_1, h_3, \dots, h_n)$ and $\lambda_1(s) = (h_1 h_2, h_3, \dots, h_n)$. Since by assumption, the first element of $\rho_1(s)$ is its maximum, the claim implies that the sequences of each chain starting at $\rho_1(s)$ are all of length at least 2. Moreover, $\lambda_1(s)$ is shorter than s and either is of length 1, or satisfies the same hypothesis as s. Thus we conclude by induction. $\square$

PROOF OF THE THEOREM. For any standard sequence $s = (h_1, \dots, h_n)$ of Hall words, $P_{h_1} \dots P_{h_n}$ is equal to the sum of all $g(s')$, for all possible decreasing sequences of Hall words s' such that $s \Rightarrow^* s'$. Indeed, this follows from Section 8.

The equality in the theorem is equivalent to $(S_h, yu) = \delta_{x,y}(S_v, u)$ ($x, y \in X$, $u, v \in X^*$), since S_h has no constant term. We have $u = \sum_{w \in X^*} (S_w, u) P_w$, so that $yu = \sum_{w \in X^*} (S_w, u) y P_w$. Let w be a word with $w = h_1 \dots h_n$, h_i Hall word, $h_1 \geqslant \cdots \geqslant h_n$. Then the sequence $s = (y, h_1, \dots, h_n)$ is standard and

$$y P_w = \sum_{s'} \alpha_{s'} g(s'),$$

where the summation is over all decreasing sequences s' such that $s \Rightarrow^* s'$ and where $\alpha_{s'}$ is the number of chains $s \Rightarrow \cdots \Rightarrow s'$.

By Lemma 1, each such s' is of length at least 2, except when $yw = yh_1 \dots h_n$ is a Hall word and $s' = (yw)$, in which case there is exactly one chain from s to s'. This implies that yu is equal to $\sum_{yw \text{ Hall word}} (S_w, u) P_{yw}$ + a sum of decreasing products $g(h'_1) g(h'_2) \dots$ with at least two factors, $h'_i \in H$. Hence the coefficient of $P_h = P_{xv}$ in this sum is equal to 0 if $x \neq y$ and to (S_v, u) if $x = y$. In other words, $(S_h, yu) = \delta_{x,y}(S_v, u)$. $\square$

13. Shuffle product

For each word $w = x_1 \dots x_n$ of length n on X and each subset I of $\{1, \dots, n\}$, denote by $w|I$ the word $x_{i_1} \dots x_{i_k}$, with $I = \{i_1 < \cdots < i_k\}$. Given p words $u_1, \dots, u_p$ whose lengths $n_1, \dots, n_p$ add up to n, their *shuffle product* is the polynomial $u_1 \times \cdots \times u_p = \sum w(I_1, \dots, I_p)$, where the sum is over all p-tuples $(I_1, \dots, I_p)$ of disjoint subsets of $\{1, \dots, n\}$, whose union is $\{1, \dots, n\}$, with $|I_j| = n_j$, and where the word $w(I_1, \dots, I_p)$ is defined by $w(I_1, \dots, I_p)|I_j = u_j$. This product extends linearly to polynomials, since words form a basis of $\mathbf{A}\langle X\rangle$. It is easy to see that the 2-ary shuffle product $u \times v$ is an associative product, with as neutral element the empty word.

The shuffle product may be alternatively defined as follows. Let δ_p denote the homomorphism from the free associative algebra $\mathbf{A}\langle X\rangle$ into the p-fold tensor product $\mathbf{A}\langle X\rangle^{\otimes p}$, defined by $\delta_p(x) = x \otimes 1 \otimes \cdots \otimes 1 + 1 \otimes x \otimes \cdots \otimes 1 + \cdots + 1 \otimes 1 \otimes \cdots \otimes x$ for any letter x.

PROPOSITION 1. *For each polynomial P, one has*

$$\delta_p(P) = \sum_{u_1, \dots, u_p \in X^*} (P, u_1 \times \cdots \times u_p) u_1 \otimes \cdots \otimes u_p.$$

Equivalently, $(\delta_p(P), P_1 \otimes \cdots \otimes P_p) = (P, P_1 \times \cdots \times P_p)$ for any polynomials P, P_1, $\dots$, P_p, where the scalar product is naturally extended to the tensor product.

PROOF. It is enough to prove this when $P = w = x_1 \dots x_n$ is a word in X^*. Then by definition $\delta_p(w)$ is equal to the product $\delta_p(x_1) \dots \delta_p(x_n)$. Since $\delta_p(x_i) = x_i \otimes 1 \otimes \cdots \otimes 1 + 1 \otimes x_i \otimes \cdots \otimes 1 + \cdots + 1 \otimes 1 \otimes \cdots \otimes x_i$, the proposition follows by inspection. □

We also need the following result.

PROPOSITION 2. *If P is a Lie polynomial then*

$$\delta_p(P) = P \otimes 1 \otimes \cdots \otimes 1 + 1 \otimes P \otimes \cdots \otimes 1 + \cdots + 1 \otimes 1 \otimes \cdots \otimes P.$$

PROOF. This holds by definition when P is a letter. A simple computation shows that if it holds for two polynomials, then it holds also for their Lie bracket. Hence the proposition follows, since the set of polynomials for which the formula is true is a submodule closed under the Lie bracket. □

14. The dual basis 2

In Section 12, we have determined all S_h when h is a Hall word, knowing S_w for shorter words w. The next theorem completely solves the problem of recursively determining the dual basis (S_w) of the Poincaré–Birkhoff–Witt basis. We assume that the ring $\mathbf{A}$ contains $\mathbf{Q}$. We denote by ${}^{\times i}$ the shuffle exponentiation.

THEOREM 8. *For any word $w = h_1^{i_1} \dots h_k^{i_k}$, where the h_j are Hall words with $h_1 > \cdots > h_k$ and i_j in $\mathbf{N}$, one has*

$$S_w = \frac{1}{i_1! \dots i_k!} S_{h_1}^{\times i_1} \times \cdots \times S_{h_k}^{\times i_k}.$$

PROOF. Note that by definition of the dual basis, one has $(S_u, P_v) = \delta_{u,v}$. In particular, if u is a Hall word and v is not, $(S_u, P_v) = 0$.

Following Corollary 2, we may write w as a decreasing product of Hall words: $w = w_1 \dots w_i$, $w_1 \geqslant \cdots \geqslant w_i$; hence $i = i_1 + \cdots + i_k$. By the remark after Proposition 1, we have $(S_{w_1} \times \cdots \times S_{w_i}, P_u) = (S_{w_1} \otimes \cdots \otimes S_{w_i}, \delta_i(P_u))$.

Write $u = u_1 \dots u_n$ as decreasing product of Hall words. Then $P_u = P_{u_1} \dots P_{u_n}$ and the P_{u_j} are Lie polynomials. By Proposition 2, we have

$$\delta_i(P_{u_j}) = P_{u_j} \otimes 1 \otimes \cdots \otimes 1 + 1 \otimes P_{u_j} \otimes \cdots \otimes 1 + \cdots + 1 \otimes 1 \otimes \cdots \otimes P_{u_j}.$$

Moreover, $\delta_i(P_u) = \delta_i(P_{u_1}) \dots \delta_i(P_{u_n})$. Thus, by inspection, we find that $\delta_i(P_u)$ is a sum of terms $Q_1 \otimes \cdots \otimes Q_i$ and, correspondingly, $(S_{w_1} \times \cdots \times S_{w_i}, P_u)$ is a sum of products $(S_{w_1}, Q_1) \dots (S_{w_i}, Q_i)$: If $i > n$, then in each term at least one Q_j is equal to 1, hence since S_{w_i} has no constant term, we have $(S_{w_1} \times \cdots \times S_{w_i}, P_u) = 0$. If $i < n$, then in each term, at least one Q_j is a decreasing product $P_{u'} = P_{u_{l_1}} \dots P_{u_{l_r}}$ with $r \geqslant 2$, so that $(S_{w_j}, P_{u'}) = 0$, since w_j is a Hall word and $u' = u_{l_1} \dots u_{l_r}$ is not; thus we also have $(S_{w_1} \times \cdots \times S_{w_i}, P_u) = 0$. If finally $i = n$, then we obtain, again because S_{w_i} has no constant term,

$$\begin{aligned}(S_{w_1} \times \cdots \times S_{w_i}, P_u) &= \sum_{\sigma \in S_n} (S_{w_1}, P_{u_{\sigma(1)}}) \dots (S_{w_n}, P_{u_{\sigma(n)}}) \\ &= \sum_{\sigma \in S_n} \delta_{w_1, u_{\sigma(1)}} \dots \delta_{w_n, u_{\sigma(n)}}.\end{aligned}$$

If $w \neq u$, then $(w_1, \dots, w_n) \neq (u_1, \dots, u_n)$ and since both sequences are decreasing, the right-hand side of the summation vanishes. If $w = u$, then both sequences are equal by Corollary 2; since $(w_1, \dots, w_n) = (h_1, \dots, h_1, \dots, h_k, \dots, h_k)$, each h_j repeated i_j times, the right-hand side is equal to the number of permutations fixing the previous sequence, that is $i_1! \dots i_k!$. Hence we obtain that

$$\left(\frac{1}{i_1! \dots i_k!} S_{h_1}^{\times i_1} \times \cdots \times S_{h_k}^{\times i_k}, P_u \right) = \delta_{w,u},$$

which proves the theorem, by definition of the dual basis. □

COROLLARY 4. *The polynomials of the dual basis (S_w) have all coefficients in* $\mathbf{N}$.

COROLLARY 5. *The shuffle algebra is a free commutative* $\mathbf{A}$*-algebra over the S_h ($h =$ Hall word).*

15. Notes

Hall bases are due to Marshall Hall Jr. [16]. Similar constructions of "basic commutators" in a free group had been done previously by Philip Hall [18] and Wilhelm Magnus [25]. M. Hall's construction was generalized by Meier-Wunderli [27], Witt [51], Schützenberger

[37], Širšov [40], Michel [31], Viennot [47]. Unlike the original Hall basis, these generalizations include the Lyndon basis constructed by Viennot (loc. cit.; see also Lothaire [24]), following the lines of the commutator calculus of Chen, Fox, Lyndon [8], and the Širšov basis [39].

Theorem 1 is due to Melançon [29], who extended a method of Schützenberger (loc. cit.), itself related to the "collecting process" of P. Hall (loc. cit.; see also M. Hall [17]). This theorem immediately implies Theorem 2, Corollaries 1 and 2, which constitute the combinatorial facts underlying Hall bases. The latter corollary also easily implies Theorem 6, which is the Witt formula (loc. cit.). Note that this formula gives also the number of *primitive circular words* on the alphabet X; an explanation of this fact is the following result: each primitive conjugation class of words contains exactly one Hall word (see, e.g., the book by Lothaire, loc. cit.).

The main result on Hall bases is Theorem 4; for the proof we have followed Schützenberger's proof (loc. cit.), which is algorithmic, and gives as a byproduct Theorem 5: the Lie algebra of Lie polynomials is the free Lie algebra; this theorem is due to Witt (loc. cit.). Note that this method does not use the theorem of Poincaré–Birkhoff–Witt (PBW). The counterpart for Hall bases of this theorem (Theorem 3) is established here directly. For the latter, Theorems 7 and 8, we have also followed Schützenberger (loc. cit.), with improvements from Melançon and Reutenauer [30] and Melançon [28].

Note that other bases, which are not Hall bases, are constructed by Kukin [23], Blessenohl and Laue [3], Garsia [13].

Proposition 2 characterizes Lie elements when $i = 2$: this is Friedrichs' criterion [12]. An equivalent version says that P is a Lie polynomial if and only if P is orthogonal to each shuffle product of two nonempty words for the scalar product of Section 12. In relation with Lie polynomials, the shuffle product was introduced by Ree [33], who simplified previous work of Chen on iterated integrals [7]. An alternative formulation of Friedrichs' criterion is due to Garsia (loc. cit.): say that a noncommutative polynomial $P(x, y, \ldots)$ on the alphabet $X = \{x, y, \ldots\}$ is *linear* if, taking a second alphabet $X' = \{x', y', \ldots\}$, in bijection with the previous one, such that each letter in X commutes with each letter in X', one has: $P(x + x', y + y', \ldots) = P(x, y, \ldots) + P(x', y', \ldots)$. Then Lie polynomials are exactly the linear ones.

Another well-known and earlier characterization of Lie polynomials is the following: denote by l the linear endomorphism of $\mathbf{A}\langle X\rangle$ "Lie bracketing from left to right" defined by $l(w) = [\ldots[[x_1, x_2], x_3], \ldots x_n]$ for any word $w = x_1 \ldots x_n$ of length n. Then a homogeneous polynomial P of degree n is a Lie polynomial if an only if $l(P) = nP$ (we assume that $\mathbf{A}$ is of characteristic 0). This characterization was found at the same time by three authors: Dynkin [11], Specht [44], Wever [48].

Actually, Lie polynomials appear implicitly in the work of Campbell [5,6], Baker [1] and Hausdorff [19], leading to their famous formula. It asserts that the noncommutative series $\mathrm{e}^x \mathrm{e}^y$ is the exponential of a Lie series (a Lie series is a series whose homogeneous components are Lie polynomials). Another famous Lie series is the Drinfeld *associator*, see [10].

Lie subalgebras of the free Lie algebra are again free, see Širšov [38], Witt [50,51]. Automorphism of a free Lie algebra are always tame (Cohn [9]) and are characterized by a Jacobian condition, see Shpilrain [41], Reutenauer [35], Umirbaev [46].

Representation-theoretic studies on the free Lie algebra began with the work of Thrall [45] and Brandt [4]. The full linear group acts on Lie polynomials, and the symmetric group acts on those which are multilinear. A result of Klyachko [21] characterizes the irreducible representations that occur, and another one of Kraskiewicz and Weyman [22] gives the exact multiplicities, in terms of the *major index* of Young tableaux. The Lie representation of the symmetric group is induced from any faithful representation of a subgroup generated by a full cycle (Klyachko, loc. cit.).

In relation with representation theory, many idempotents of the symmetric group algebra, called *Lie idempotents* may be found in the literature: the Dynkin–Specht–Wever idempotent, implicit in the work of these three authors (loc. cit.); the canonical idempotent, related to the PBW theorem, see Solomon [42], Mielnik, Plebanski [32], Reutenauer [34], Helmstetter [20]; the idempotent of Klyachko (loc. cit.), that has a fabulous definition using the major index of permutations, and its generalization by Bergeron, Bergeron, Garsia [2].

There are other idempotents related to the PBW decomposition of the tensor algebra; these idempotents constitute all primitive idempotents of the *descent algebra* of the symmetric group, see Garsia, Reutenauer [14]; the descent algebra was introduced by Solomon [43] for each finite Coxeter group. All previous Lie idempotents lie in the descent algebra. The latter is itself a Hopf algebra, isomorphic to a free associative algebra, and its primitive elements are exactly the quasi-Lie idempotents that lie in it, see Gelfand, Krob, Lascoux, Leclerc, Retakh, Thibon [15]. The descent algebra is dual to the ring of quasi-symmetric functions, see [26], which is therefore a free commutative algebra.

See the author's book on Free Lie Algebras [36] for more on the subject.

References

[1] H.F. Baker, *Alternants and continuous groups*, Proc. London Math. Soc. **3** (1905), 24–47.

[2] F. Bergeron, N. Bergeron and A. Garsia, *Idempotents for the free Lie algebra and q-enumeration*, Invariant Theory and Tableaux (Minneapolis, MN, 1988), IMA Vol. Math. Appl., Vol. 19, Springer, New York (1990), 166–190.

[3] D. Blessenohl and H. Laue, *A basis construction for free Lie algebras*, Exposition. Math. **11** (1993), 145–152.

[4] A. Brandt, *The free Lie ring and representations of the full linear group*, Trans. Amer. Math. Soc. **56** (1944), 528–536.

[5] J.E. Campbell, *On a law of combination of operators bearing on the theory of continuous transformation groups*, Proc. London Math. Soc. **28** (1897), 381–390.

[6] J.E. Campbell, *On a law of combination of operators bearing on the theory of continuous transformation groups*, Proc. London Math. Soc. **29** (1898), 14–32.

[7] K.T. Chen, *Integration of paths, geometric invariants and a generalized Baker–Hausdorff formula*, Ann. of Math. **65** (1957), 163–178.

[8] K.T. Chen, R.H. Fox and R.C. Lyndon, *Free differential calculus IV: The quotient groups of the lower central series*, Ann. of Math. **68** (1958), 81–95.

[9] P.M. Cohn, *Subalgebras of free associative algebras*, Proc. London Math. Soc. **14** (1964), 618–632.

[10] V.G. Drinfeld, *On quasitriangular quasi-Hopf algebras and on a group that is closely connected with* $\mathrm{Gal}(\overline{Q}/Q)$, Algebra i Analiz **2** (1990), 149–181; transl.: Leningrad Math. J. **2** (1991), 829–860.

[11] E.B. Dynkin, *Calculation of the coefficients in the Campbell–Hausdorff formula*, Dokl. Akad. Nauk SSSR **57** (1947), 323–326.

[12] K.O. Friedrichs, *Mathematical aspects of the quantum theory of fields. V: Fields modified by linear homogeneous forces*, Comm. Pure Appl. Math. **6** (1953), 1–72.

[13] A. Garsia, *Combinatorics of the free Lie algebra and the symmetric group*, Analysis, et Cetera, Academic Press, Boston, MA (1990), 309–382.

[14] A. Garsia and C. Reutenauer, *A decomposition of Solomon's descent algebra*, Adv. Math. **77** (1989), 189–262.

[15] I.M. Gelfand, D. Krob, A. Lascoux, B. Leclerc, V. Retakh and J.-Y. Thibon, *Noncommutative symmetric functions*, Adv. Math. **112** (1995), 218–348.

[16] M. Hall, *A basis for free Lie rings and higher commutators of the free group*, Proc. Amer. Math. Soc. **1** (1950), 575–580.

[17] M. Hall, *The Theory of Groups*, Macmillan, New York (1959).

[18] P. Hall, *A contribution to the theory of groups of prime power order*, Proc. London Math. Soc. **36** (1933), 29–95.

[19] F. Hausdorff, *Die symbolische Exponentialformel in der Gruppentheorie*, Leipziger Ber. **58** (1906), 19–48.

[20] J. Helmstetter, *Série de Hausdorff d'une algèbre de Lie et projections canoniques dans l'algèbre enveloppante*, J. Algebra **120** (1989), 170–199.

[21] A.A. Klyachko, *Lie elements in the tensor algebra*, Sibirsk. Mat. Zh. (translation) **15** (1974), 1296–1304.

[22] W. Kraskiewicz and J. Weyman, *Algebra of coinvariants and the action of Coxeter elements*, Manuscript (1987).

[23] G.P. Kukin, *Bases of a free Lie algebra*, Mat. Zametki **24** (1978), 375–382.

[24] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, Reading, MA (1983).

[25] W. Magnus, *Über Beziehungen zwischen höheren Kommutatoren*, J. Reine Angew. Math. **177** (1937), 105–115.

[26] C. Malvenuto and C. Reutenauer, *Duality between quasi-symmetric functions and the Solomon descent algebra*, J. Algebra **177** (1995), 967–982.

[27] H. Meier-Wunderli, *Note on a basis of P. Hall for the higher commutators in free groups*, Comment. Math. Helv. **26** (1951), 1–5.

[28] G. Melançon, *Récritures dans l'algèbre de Lie libre, le groupe libre et l'algèbre associative libre*, Thèse Math., Univ. du Québec à Montréal (1991).

[29] G. Melançon, *Combinatorics of Hall trees and Hall words*, J. Combin. Theory A **59** (1992), 285–308.

[30] G. Melançon and C. Reutenauer, *Lyndon words, free algebras and shuffles*, Canad. J. Math. **41** (1989), 577–591.

[31] J. Michel, *Bases des algèbres de Lie libres, études des coefficients de la formule de Campbell–Hausdorff*, Thèse de 3$^{\text{ème}}$ cycle, Université Paris XI (1974).

[32] B. Mielnik and J. Plebanski, *Combinatorial approach to Baker–Campbell–Hausdorff exponents*, Ann. Inst. H. Poincaré A **12** (1970), 215–254.

[33] R. Ree, *Lie elements and an algebra associated with shuffles*, Ann. of Math. **68** (1958), 210–220.

[34] C. Reutenauer, *Theorem of Poincaré–Birkhoff–Witt, logarithm and representations of the symmetric group whose orders are the Stirling numbers*, Combinatoire Énumérative, Proc., Montréal, P. Leroux and G. Labelle, eds, Lecture Notes in Math., Vol. 1234, Springer, Berlin (1986).

[35] C. Reutenauer, *Applications of a noncommutative Jacobian matrix*, J. Pure Appl. Algebra **77** (1992), 169–181.

[36] C. Reutenauer, *Free Lie Algebras*, Oxford University Press (1993).

[37] M.-P. Schützenberger, *Sur une propriété combinatoire des algèbres de Lie libres pouvant être utilisée dans un problème de mathématiques appliquées*, Séminaire P. Dubreil, Faculté des Sciences, Paris (1958).

[38] A.I. Širšov, *Subalgebras of free Lie algebras*, Mat. Sb. **33** (1953), 441–452.

[39] A.I. Širšov, *Free Lie rings*, Mat. Sb. **45** (1958), 113–122.

[40] A.I. Širšov, *On the bases of a free Lie algebra*, Algebra i Logika **1** (1962), 14–19.

[41] V. Shpilrain, *On generators of L/R^2 Lie algebras*, Proc. Amer. Math. Soc. **119** (1993), 1039–1043.

[42] L. Solomon, *On the Poincaré–Birkhoff–Witt theorem*, J. Combin. Theory A **4** (1968), 363–375.

[43] L. Solomon, *A Mackey formula in the group ring of a finite Coxeter group*, J. Algebra **41** (1976), 255–268.

[44] W. Specht, *Die linearen Beziehungen zwischen höheren Kommutatoren*, Mat. Z. **51** (1948), 367–376.

[45] R.M. Thrall, *On symmetrized Kronecker powers and the structure of the free Lie ring*, Amer. J. Math. **64** (1942), 371–388.

[46] U.U. Umirbaev, *Partial derivations and endomorphisms of some relatively free Lie algebras*, Sibirsk. Mat. Zh. **34** (1993), 179–188; transl.: Siberian Math. J. **34** (1993), 1161–1170.

[47] G.X. Viennot, *Algèbres de Lie libres et monoïdes libres*, Lecture Notes in Math., Vol. 691, Springer, Berlin (1978).

[48] F. Wever, *Über Invarianten in Lieschen Ringen*, Math. Ann. **120** (1949), 563–580.

[49] E. Witt, *Treue Darstellungen Liescher Ringe*, J. Reine Angew. Math. **177** (1937), 152–160.

[50] E. Witt, *Über freie Liesche Ringe und ihre Unterringe*, Math. Z. **58** (1953), 113–114.

[51] E. Witt, *Die Unterringe der freien Lieschen Ringe*, Math. Z. **64** (1956), 195–216.

This Page Intentionally Left Blank

Section 4H
Rings and Algebras with Additional Structure

This Page Intentionally Left Blank

Yangians and their Applications

A.I. Molev

School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia
E-mail: alexm@maths.usyd.edu.au

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

1. Introduction

The term *Yangian* was introduced by Drinfeld (in honor of C.N. Yang) in his fundamental paper [35] (1985). In that paper, besides the Yangians, Drinfeld defined the *quantized Kac–Moody algebras* which together with the work of Jimbo [64], who introduced these algebras independently, marked the beginning of the era of *quantum groups*. The Yangians form a remarkable family of quantum groups related to rational solutions of the classical Yang–Baxter equation. For each simple finite-dimensional Lie algebra $\mathfrak{a}$ over the field of complex numbers, the corresponding Yangian $\mathrm{Y}(\mathfrak{a})$ is defined as a canonical deformation of the universal enveloping algebra $\mathrm{U}(\mathfrak{a}[x])$ for the polynomial current Lie algebra $\mathfrak{a}[x]$. Importantly, the deformation is considered in the class of Hopf algebras which guarantees its uniqueness under some natural 'homogeneity' conditions. An alternative description of the algebra $\mathrm{Y}(\mathfrak{a})$ was given later in Drinfeld [38].

Prior to the introduction of the Hopf algebra $\mathrm{Y}(\mathfrak{a})$ in [35], the algebra, which is now called the *Yangian for the general linear Lie algebra* $\mathfrak{gl}_n$ and denoted $\mathrm{Y}(\mathfrak{gl}_n)$, was considered in the work of Faddeev and the St.-Petersburg school in relation with the *inverse scattering method*; see for instance Takhtajan and Faddeev [161], Kulish and Sklyanin [89], Tarasov [157,158]. The latter algebra is a deformation of the universal enveloping algebra $\mathrm{U}(\mathfrak{gl}_n[x])$.

For any simple Lie algebra $\mathfrak{a}$ the Yangian $\mathrm{Y}(\mathfrak{a})$ contains the universal enveloping algebra $\mathrm{U}(\mathfrak{a})$ as a subalgebra. However, only in the case $\mathfrak{a} = \mathfrak{sl}_n$ does there exist an evaluation homomorphism $\mathrm{Y}(\mathfrak{a}) \to \mathrm{U}(\mathfrak{a})$ which is the identity on the subalgebra $\mathrm{U}(\mathfrak{a})$; see Drinfeld [35, Theorem 9]. In this chapter we concentrate on this distinguished Yangian which is closely related to $\mathrm{Y}(\mathfrak{gl}_n)$. For each of the classical Lie algebras $\mathfrak{a} = \mathfrak{o}_{2n+1},\ \mathfrak{sp}_{2n},\ \mathfrak{o}_{2n}$ Olshanski [139] introduced another algebra called the *twisted Yangian*. Namely, the Lie algebra $\mathfrak{a}$ can be regarded as a fixed point subalgebra of an involution σ of the appropriate general linear Lie algebra $\mathfrak{gl}_N$. Then the twisted Yangian $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$ can be defined as a subalgebra of $\mathrm{Y}(\mathfrak{gl}_N)$ which is a deformation of the universal enveloping algebra for the twisted polynomial current Lie algebra

$$\mathfrak{gl}_N[x]^\sigma = \big\{A(x) \in \mathfrak{gl}_N[x] \mid \sigma\big(A(x)\big) = A(-x)\big\}. \tag{1.1}$$

For each $\mathfrak{a}$ the twisted Yangian contains $\mathrm{U}(\mathfrak{a})$ as a subalgebra, and an analog of the evaluation homomorphism $\mathrm{Y}(\mathfrak{gl}_N, \sigma) \to \mathrm{U}(\mathfrak{a})$ does exist. Moreover, the twisted Yangian turns out to be a (left) coideal of the Hopf algebra $\mathrm{Y}(\mathfrak{gl}_N)$.

The defining relations of the Yangian $\mathrm{Y}(\mathfrak{gl}_n)$ can be written in a form of a single *ternary* (or *RTT*) relation on the matrix of generators. This relation has a rich and extensive background. It originates from quantum inverse scattering theory; see, e.g., Takhtajan and Faddeev [161], Kulish and Sklyanin [89], Drinfeld [37]. The Yangians were primarily regarded as a vehicle for producing rational solutions of the Yang–Baxter equation which plays a key role in the theory of integrable models; cf. Drinfeld [35]. Conversely, the ternary relation was used in Reshetikhin, Takhtajan and Faddeev [149] as a tool for studying quantum groups. Moreover, the Hopf algebra structure of the Yangian can also be conveniently described in a matrix form.

Similarly, the twisted Yangians can be equivalently presented by generators and defining relations which can be written as a *quaternary* (or *reflection*) equation for the matrix of generators, together with a *symmetry* relation. Relations of this type appeared for the first time in Cherednik [28] and Sklyanin [152], where integrable systems with boundary conditions were studied.

This remarkable form of the defining relations for the Yangian and twisted Yangians allows special algebraic techniques (the so-called *R-matrix formalism*) to be used to describe the algebraic structure and study representations of these algebras. On the other hand, the evaluation homomorphisms to the corresponding classical enveloping algebras allow one to use these results to better understand the classical Lie algebras themselves. In particular, new constructions of the Casimir elements can be obtained in this way. These include the noncommutative characteristic polynomials for the generator matrices and the Capelli-type determinants. Some other applications include the constructions of Gelfand–Tsetlin bases and commutative subalgebras. Moreover, as was shown by Olshanski [138,139], the Yangian and the twisted Yangians can be realized as some projective limit subalgebras of a sequence of centralizers in the classical enveloping algebras. This is known as the *centralizer construction*; see also [120].

The representation theory of the Yangians $\mathrm{Y}(\mathfrak{a})$ is a very much nontrivial and fascinating area. Although the finite-dimensional irreducible representations of $\mathrm{Y}(\mathfrak{a})$ are completely described by Drinfeld [38], their general structure still remains unknown. This part of the theory of the Yangians will have to be left outside this chapter. We give, however, some references in the bibliography which we hope cover at least some of the most important results in the area.

The Yangians, as well as their super and q-analogs, have found a wide variety of applications in physics. This includes the theory of integrable models in statistical mechanics, conformal field theory, quantum gravity. We do not attempt to review all the relevant physics literature but some references are given as a guide to such applications.

At the end of each section we give brief bibliographical comments pointing towards the original articles and to the references where the proofs or further details can be found.

2. The Yangian for the general linear Lie algebra

As we pointed out in the Introduction, the discovery of the Yangians was motivated by quantum inverse scattering theory. It is possible, however, to "observe" the Yangian defining relations from a purely algebraic viewpoint. We start by showing that they are satisfied by certain natural elements of the universal enveloping algebra $\mathrm{U}(\mathfrak{gl}_n)$. Then we show that these relations can be written in a matrix form which provides a starting point for special algebraic techniques to study Yangian structures.

2.1. *Algebraic motivations and definitions*

Consider the general linear Lie algebra $\mathfrak{gl}_n$ with its standard basis $E_{ij},\ i, j = 1, \ldots, n$. The commutation relations are given by

$$[E_{ij}, E_{kl}] = \delta_{kj} E_{il} - \delta_{il} E_{kj},$$

where δ_{ij} is the Kronecker delta; see, e.g., [58]. Introduce the $n \times n$-matrix E whose ij-th entry is E_{ij}. The traces of the powers of the matrix E

$$g_s = \operatorname{tr} E^s, \quad s = 1, 2, \dots$$

are central elements of the universal enveloping algebra $\mathrm{U}(\mathfrak{gl}_n)$ known as the *Gelfand invariants*; see [44]. Moreover, the first n of them are algebraically independent and generate the center. A proof of the centrality of the g_s is easily deduced from the following relations in the enveloping algebra

$$\big[E_{ij}, (E^s)_{kl}\big] = \delta_{kj} (E^s)_{il} - \delta_{il} (E^s)_{kj}. \tag{2.1}$$

One could wonder whether a more general closed formula exists for the commutator of the matrix elements of the powers of E. The answer to this question turns out to be affirmative and the following generalization of (2.1) can be verified by induction:

$$\big[(E^{r+1})_{ij}, (E^s)_{kl}\big] - \big[(E^r)_{ij}, (E^{s+1})_{kl}\big] = (E^r)_{kj} (E^s)_{il} - (E^s)_{kj} (E^r)_{il},$$

where $r, s \geqslant 0$ and $E^0 = 1$ is the identity matrix. We can axiomatize these relations by introducing the following definition.

DEFINITION 2.1. The *Yangian for* $\mathfrak{gl}_n$ is a unital associative algebra with countably many generators $t_{ij}^{(1)},\ t_{ij}^{(2)}, \dots$ where $1 \leqslant i, j \leqslant n$, and defining relations

$$\big[t_{ij}^{(r+1)}, t_{kl}^{(s)}\big] - \big[t_{ij}^{(r)}, t_{kl}^{(s+1)}\big] = t_{kj}^{(r)} t_{il}^{(s)} - t_{kj}^{(s)} t_{il}^{(r)}, \tag{2.2}$$

where $r, s \geqslant 0$ and $t_{ij}^{(0)} = \delta_{ij}$. This algebra is denoted by $\mathrm{Y}(\mathfrak{gl}_n)$, or $\mathrm{Y}(n)$ for brevity.

Introducing the generating series,

$$t_{ij}(u) = \delta_{ij} + t_{ij}^{(1)} u^{-1} + t_{ij}^{(2)} u^{-2} + \cdots \in \mathrm{Y}(n)[\![u^{-1}]\!],$$

we can write (2.2) in the form

$$(u - v)\big[t_{ij}(u), t_{kl}(v)\big] = t_{kj}(u) t_{il}(v) - t_{kj}(v) t_{il}(u). \tag{2.3}$$

Divide both sides by $u - v$ and use the formal expansion

$$(u - v)^{-1} = \sum_{r=0}^{\infty} u^{-r-1} v^r$$

to get an equivalent form of the defining relations

$$\big[t_{ij}^{(r)}, t_{kl}^{(s)}\big] = \sum_{a=1}^{\min(r,s)} \big(t_{kj}^{(a-1)} t_{il}^{(r+s-a)} - t_{kj}^{(r+s-a)} t_{il}^{(a-1)}\big). \tag{2.4}$$

The previous discussion implies that the algebra $\mathrm{Y}(n)$ is nontrivial, as the mapping

$$t_{ij}^{(r)} \mapsto (E^r)_{ij} \tag{2.5}$$

defines an algebra homomorphism $\mathrm{Y}(n) \to \mathrm{U}(\mathfrak{gl}_n)$.

Alternatively, the generators of the Yangian can be realized as "Capelli minors". Keeping the notation E for the matrix of the basis elements of $\mathfrak{gl}_n$ introduce the *Capelli determinant*

$$\begin{aligned}\det(1+Eu^{-1})\\ = \sum_{\sigma\in\mathfrak{S}_n} \operatorname{sgn}\sigma\cdot(1+Eu^{-1})_{\sigma(1),1}\cdots\big(1+E(u-n+1)^{-1}\big)_{\sigma(n),n}.\end{aligned} \tag{2.6}$$

When multiplied by $u(u-1)\cdots(u-n+1)$ this determinant becomes a polynomial in u whose coefficients constitute another family of algebraically independent generators of the center of $\mathrm{U}(\mathfrak{gl}_n)$. The value of this polynomial at $u=n-1$ is a distinguished central element whose image in a natural representation of $\mathfrak{gl}_n$ by differential operators is given by the celebrated *Capelli identity* [19]; see also [57]. For a positive integer $m \leqslant n$ introduce the subsets of indices $\mathcal{B}_i = \{i, m+1, m+2, \dots, n\}$ and for any $1 \leqslant i, j \leqslant m$ consider the Capelli minor $\det(1+Eu^{-1})_{\mathcal{B}_i\mathcal{B}_j}$ defined as in (2.6), whose rows and columns are respectively enumerated by $\mathcal{B}_i$ and $\mathcal{B}_j$. These minors turn out to satisfy the Yangian defining relations, i.e. there is an algebra homomorphism

$$\mathrm{Y}(m) \to \mathrm{U}(\mathfrak{gl}_n), \qquad t_{ij}(u) \mapsto \det(1+Eu^{-1})_{\mathcal{B}_i\mathcal{B}_j}.$$

These two interpretations of the Yangian defining relations (which will reappear in Sections 2.4, 2.12 and 2.13) indicate a close relationship between the representation theory of the algebra $\mathrm{Y}(n)$ and the conventional representation theory of the general linear Lie algebra. Many applications of the Yangian are based on the following simple observation.

PROPOSITION 2.2. *The mapping*

$$\pi: t_{ij}(u) \mapsto \delta_{ij} + E_{ij}u^{-1} \tag{2.7}$$

defines an algebra epimorphism $\mathrm{Y}(n) \to \mathrm{U}(\mathfrak{gl}_n)$. *Moreover,*

$$E_{ij} \mapsto t_{ij}^{(1)}$$

is an embedding $\mathrm{U}(\mathfrak{gl}_n) \hookrightarrow \mathrm{Y}(n)$.

In particular, any $\mathfrak{gl}_n$-module can be extended to the algebra $\mathrm{Y}(n)$ via (2.7). This plays an important role in Yangian representation theory.

2.2. *A matrix form of the defining relations*

Introduce the $n \times n$ matrix $T(u)$ whose ij-th entry is the series $t_{ij}(u)$. It is convenient to regard it as an element of the algebra $\mathrm{Y}(n)[\![u^{-1}]\!] \otimes \mathrm{End}\,\mathbb{C}^n$ given by

$$T(u) = \sum_{i,j=1}^{n} t_{ij}(u) \otimes e_{ij}, \tag{2.8}$$

where the e_{ij} denote the standard matrix units. For any positive integer m we shall be using algebras of the form

$$\mathrm{Y}(n)[\![u^{-1}]\!] \otimes \mathrm{End}\,\mathbb{C}^n \otimes \cdots \otimes \mathrm{End}\,\mathbb{C}^n, \tag{2.9}$$

with m copies of $\mathrm{End}\,\mathbb{C}^n$. For any $a \in \{1, \ldots, m\}$ we denote by $T_a(u)$ the matrix $T(u)$ which acts on the a-th copy of $\mathrm{End}\,\mathbb{C}^n$. That is, $T_a(u)$ is an element of the algebra (2.9) of the form

$$T_a(u) = \sum_{i,j=1}^{n} t_{ij}(u) \otimes 1 \otimes \cdots \otimes 1 \otimes e_{ij} \otimes 1 \otimes \cdots \otimes 1, \tag{2.10}$$

where the e_{ij} belong to the a-th copy of $\mathrm{End}\,\mathbb{C}^n$ and 1 is the identity matrix. Similarly, if

$$C = \sum_{i,j,k,l=1}^{n} c_{ijkl}\, e_{ij} \otimes e_{kl} \in \mathrm{End}\,\mathbb{C}^n \otimes \mathrm{End}\,\mathbb{C}^n,$$

then for distinct indices $a, b \in \{1, \ldots, m\}$ we introduce the element C_{ab} of the algebra (2.9) by

$$C_{ab} = \sum_{i,j,k,l=1}^{n} c_{ijkl} 1 \otimes 1 \otimes \cdots \otimes 1 \otimes e_{ij} \otimes 1 \otimes \cdots \\ \otimes 1 \otimes e_{kl} \otimes 1 \otimes \cdots \otimes 1, \tag{2.11}$$

where the e_{ij} and e_{kl} belong to the a-th and b-th copies of $\mathrm{End}\,\mathbb{C}^n$, respectively. Consider now the permutation operator

$$P = \sum_{i,j=1}^{n} e_{ij} \otimes e_{ji} \in \mathrm{End}\,\mathbb{C}^n \otimes \mathrm{End}\,\mathbb{C}^n.$$

The rational function

$$R(u) = 1 - P u^{-1} \tag{2.12}$$

with values in $\operatorname{End}\mathbb{C}^n \otimes \operatorname{End}\mathbb{C}^n$ is called the *Yang R-matrix*. (Here and below we write 1 instead of $1 \otimes 1$ for brevity.) An easy calculation in the group algebra $\mathbb{C}[\mathfrak{S}_3]$ shows that the following identity holds in the algebra $(\operatorname{End}\mathbb{C}^n)^{\otimes 3}$

$$R_{12}(u)R_{13}(u+v)R_{23}(v) = R_{23}(v)R_{13}(u+v)R_{12}(u). \tag{2.13}$$

This is known as the *Yang–Baxter equation*. The Yang R-matrix is its simplest nontrivial solution. In the following we regard $T_1(u)$ and $T_2(u)$ as elements of the algebra (2.9) with $m = 2$.

PROPOSITION 2.3. *The defining relations* (2.2) *can be written in the equivalent form*

$$R(u-v)T_1(u)T_2(v) = T_2(v)T_1(u)R(u-v). \tag{2.14}$$

The relation (2.14) is known as the *ternary* or *RTT relation*.

2.3. *Automorphisms and anti-automorphisms*

Consider an arbitrary formal series which begins with 1,

$$f(u) = 1 + f_1 u^{-1} + f_2 u^{-2} + \cdots \in \mathbb{C}[\![u^{-1}]\!].$$

Also, let $a \in \mathbb{C}$ be an arbitrary element and B an arbitrary nondegenerate complex matrix.

PROPOSITION 2.4. *Each of the mappings*

$$T(u) \mapsto f(u)T(u), \tag{2.15}$$

$$T(u) \mapsto T(u+a), \tag{2.16}$$

$$T(u) \mapsto BT(u)B^{-1}$$

defines an automorphism of $\mathrm{Y}(n)$.

The matrix $T(u)$ can be regarded as a formal series in u^{-1} whose coefficients are matrices over $\mathrm{Y}(n)$. Since this series begins with the identity matrix, $T(u)$ is invertible and we denote by $T^{-1}(u)$ the inverse element. Also, denote by $T^t(u)$ the transposed matrix for $T(u)$.

PROPOSITION 2.5. *Each of the mappings*

$$T(u) \mapsto T(-u),$$
$$T(u) \mapsto T^t(u),$$
$$T(u) \mapsto T^{-1}(u)$$

defines an anti-automorphism of $\mathrm{Y}(n)$.

2.4. *Poincaré–Birkhoff–Witt theorem*

THEOREM 2.6. *Given an arbitrary linear order on the set of the generators $t_{ij}^{(r)}$, any element of the algebra $\mathrm{Y}(n)$ is uniquely written as a linear combination of ordered monomials in the generators.*

OUTLINE OF THE PROOF. There are two natural ascending filtrations on the algebra $\mathrm{Y}(n)$. Here we use the one defined by

$$\deg_1 t_{ij}^{(r)} = r.$$

(The other filtration will be used in Section 2.7.) It is immediate from the defining relations (2.4) that the corresponding graded algebra $\mathrm{gr}_1\,\mathrm{Y}(n)$ is commutative. Denote by $\bar{t}_{ij}^{(r)}$ the image of $t_{ij}^{(r)}$ in the r-th component of $\mathrm{gr}_1\,\mathrm{Y}(n)$. It will be sufficient to show that the elements $\bar{t}_{ij}^{(r)}$ are algebraically independent.

The composition of the automorphism $T(u) \mapsto T^{-1}(-u)$ of $\mathrm{Y}(n)$ and the homomorphism (2.7) yields another homomorphism $\mathrm{Y}(n) \to \mathrm{U}(\mathfrak{gl}_n)$ such that

$$T(u) \mapsto (1 - Eu^{-1})^{-1}. \tag{2.17}$$

The image of the generator $t_{ij}^{(r)}$ is given by (2.5). For any nonnegative integer m consider the Lie algebra $\mathfrak{gl}_{n+m}$ and now let E denote the corresponding matrix formed by its basis elements E_{ij}. Then formula (2.5) still defines a homomorphism

$$\mathrm{Y}(n) \to \mathrm{U}(\mathfrak{gl}_{n+m}). \tag{2.18}$$

Consider the canonical filtration on the universal enveloping algebra $\mathrm{U}(\mathfrak{gl}_{n+m})$ and observe that the homomorphism (2.18) is filtration-preserving. So, it defines a homomorphism of the corresponding graded algebras

$$\mathrm{gr}_1\,\mathrm{Y}(n) \to \mathrm{S}(\mathfrak{gl}_{n+m}), \tag{2.19}$$

where $\mathrm{S}(\mathfrak{gl}_{n+m})$ is the symmetric algebra of $\mathfrak{gl}_{n+m}$. One can show that for any finite family of elements $\bar{t}_{ij}^{(r)}$ there exists a sufficiently large parameter m such that their images under (2.19) are algebraically independent. $\square$

COROLLARY 2.7. $\mathrm{gr}_1\,\mathrm{Y}(n)$ *is the algebra of polynomials in the variables* $\bar{t}_{ij}^{(r)}$.

2.5. *Hopf algebra structure*

A *Hopf algebra* A (over $\mathbb{C}$) is an associative algebra equipped with a *coproduct* (or *comultiplication*) $\Delta : A \to A \otimes A$, an *antipode* $\mathrm{S} : A \to A$ and a *counit* $\varepsilon : A \to \mathbb{C}$ such that Δ and ε are algebra homomorphisms, S is an anti-automorphism and some other axioms are satisfied. These can be found in any textbook on the subject. In [25, Chapter 4] the Hopf algebra axioms are discussed in the context of quantum groups.

THEOREM 2.8. *The Yangian* $\mathrm{Y}(n)$ *is a Hopf algebra with the coproduct*

$$\Delta : t_{ij}(u) \mapsto \sum_{a=1}^{n} t_{ia}(u) \otimes t_{aj}(u),$$

the antipode

$$\mathrm{S} : T(u) \mapsto T^{-1}(u),$$

and the counit

$$\varepsilon : T(u) \mapsto 1.$$

PROOF. We only verify the most nontrivial axiom that $\Delta : \mathrm{Y}(n) \to \mathrm{Y}(n) \otimes \mathrm{Y}(n)$ is an algebra homomorphism. The remaining axioms follow directly from the definitions. We slightly generalize the notation used in Section 2.2. Let p and m be positive integers. Introduce the algebra

$$\big(\mathrm{Y}(n)[\![u^{-1}]\!]\big)^{\otimes p} \otimes (\operatorname{End} \mathbb{C}^n)^{\otimes m}$$

and for all $a \in \{1, \dots, m\}$ and $b \in \{1, \dots, p\}$ consider its elements

$$T_{[b]a}(u) = \sum_{i,j=1}^{n} \big(1^{\otimes b-1} \otimes t_{ij}(u) \otimes 1^{\otimes p-b}\big) \otimes \big(1^{\otimes a-1} \otimes e_{ij} \otimes 1^{\otimes m-a}\big).$$

The definition of Δ can now be written in a matrix form,

$$\Delta : T(u) \mapsto T_{[1]}(u) T_{[2]}(u),$$

where $T_{[b]}(u)$ is an abbreviation for $T_{[b]1}(u)$. We need to show that $\Delta(T(u))$ satisfies the ternary relation (2.14), i.e.

$$\begin{aligned} &R(u-v) T_{[1]1}(u) T_{[2]1}(u) T_{[1]2}(v) T_{[2]2}(v) \\ &\quad = T_{[1]2}(v) T_{[2]2}(v) T_{[1]1}(u) T_{[2]1}(u) R(u-v). \end{aligned}$$

However, this is implied by the ternary relation (2.14) and the observation that $T_{[2]1}(u)$ and $T_{[1]2}(v)$, as well as $T_{[1]1}(u)$ and $T_{[2]2}(v)$, commute. $\square$

2.6. *Quantum determinant and quantum minors*

Consider the rational function $R(u_1,\dots,u_m)$ with values in the algebra $(\operatorname{End}\mathbb{C}^n)^{\otimes m}$ defined by

$$R(u_1,\dots,u_m)=(R_{m-1,m})(R_{m-2,m}R_{m-2,m-1})\cdots(R_{1m}\cdots R_{12}), \tag{2.20}$$

where we abbreviate $R_{ij}=R_{ij}(u_i-u_j)$. Applying the Yang–Baxter equation (2.13) and the fact that R_{ij} and R_{kl} commute if the indices are distinct, we can write (2.20) in a different form. In particular,

$$R(u_1,\dots,u_m)=(R_{12}\cdots R_{1m})\cdots(R_{m-2,m-1}R_{m-2,m})(R_{m-1,m}).$$

As before, we use the notation $T_a(u_a)$ for the matrix $T(u_a)$ of the Yangian generators corresponding to the a-th copy of $\operatorname{End}\mathbb{C}^n$.

PROPOSITION 2.9. *We have the relation*

$$R(u_1,\dots,u_m)\,T_1(u_1)\cdots T_m(u_m)=T_m(u_m)\cdots T_1(u_1)R(u_1,\dots,u_m).$$

We let the $e_i,\ i=1,\dots,n$, denote the canonical basis of $\mathbb{C}^n$, and A_m the antisymmetrizer in $(\mathbb{C}^n)^{\otimes m}$ given by

$$A_m(e_{i_1}\otimes\cdots\otimes e_{i_m})=\sum_{\sigma\in\mathfrak{S}_m}\operatorname{sgn}\sigma\cdot e_{i_{\sigma(1)}}\otimes\cdots\otimes e_{i_{\sigma(m)}}. \tag{2.21}$$

Note that this operator satisfies $A_m^2=m!A_m$.

PROPOSITION 2.10. *If $u_i-u_{i+1}=1$ for all $i=1,\dots,m-1$ then*

$$R(u_1,\dots,u_m)=A_m.$$

By Propositions 2.9 and 2.10 we have the identity

$$A_m\,T_1(u)\cdots T_m(u-m+1)=T_m(u-m+1)\cdots T_1(u)\,A_m. \tag{2.22}$$

Suppose now that $m=n$. Then the antisymmetrizer is a one-dimensional operator in $(\mathbb{C}^n)^{\otimes n}$. Therefore, the element (2.22) equals A_n times a scalar series with coefficients in $\mathrm{Y}(n)$ which prompts the following definition.

DEFINITION 2.11. The *quantum determinant* is the formal series

$$\operatorname{qdet}T(u)=1+d_1u^{-1}+d_2u^{-2}+\cdots\in\mathrm{Y}(n)[[u^{-1}]] \tag{2.23}$$

such that the element (2.22) (with $m=n$) equals $A_n\operatorname{qdet}T(u)$.

PROPOSITION 2.12. *For any permutation $\rho \in \mathfrak{S}_n$ we have*

$$\operatorname{qdet} T(u) = \operatorname{sgn} \rho \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sgn} \sigma \cdot t_{\sigma(1),\rho(1)}(u) \cdots t_{\sigma(n),\rho(n)}(u-n+1) \tag{2.24}$$

$$= \operatorname{sgn} \rho \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sgn} \sigma \cdot t_{\rho(1),\sigma(1)}(u-n+1) \cdots t_{\rho(n),\sigma(n)}(u). \tag{2.25}$$

EXAMPLE 2.13. In the case $n = 2$ we have

$$\begin{aligned}\operatorname{qdet} T(u) &= t_{11}(u)t_{22}(u-1) - t_{21}(u)t_{12}(u-1)\\ &= t_{22}(u)t_{11}(u-1) - t_{12}(u)t_{21}(u-1)\\ &= t_{11}(u-1)t_{22}(u) - t_{12}(u-1)t_{21}(u)\\ &= t_{22}(u-1)t_{11}(u) - t_{21}(u-1)t_{12}(u).\end{aligned}$$

More generally, assuming that $m \leqslant n$ is arbitrary, we can define $m \times m$ quantum minors as the matrix elements of the operator (2.22). Namely, the operator (2.22) can be written as

$$\sum t^{c_1 \cdots c_m}_{d_1 \cdots d_m}(u) \otimes e_{c_1 d_1} \otimes \cdots \otimes e_{c_m d_m},$$

summed over the indices $c_i, d_i \in \{1, \dots, n\}$, where $t^{c_1 \cdots c_m}_{d_1 \cdots d_m}(u) \in \mathrm{Y}(n)[\![u^{-1}]\!]$. We call these elements the *quantum minors* of the matrix $T(u)$. The following formulas are obvious generalizations of (2.24) and (2.25),

$$\begin{aligned}t^{c_1 \cdots c_m}_{d_1 \cdots d_m}(u) &= \sum_{\sigma \in \mathfrak{S}_m} \operatorname{sgn} \sigma \cdot t_{c_{\sigma(1)} d_1}(u) \cdots t_{c_{\sigma(m)} d_m}(u-m+1)\\ &= \sum_{\sigma \in \mathfrak{S}_m} \operatorname{sgn} \sigma \cdot t_{c_1 d_{\sigma(1)}}(u-m+1) \cdots t_{c_m d_{\sigma(m)}}(u).\end{aligned}$$

It is clear from the definition that the quantum minors are skew-symmetric with respect to permutations of the upper indices and of the lower indices.

PROPOSITION 2.14. *The images of the quantum minors under the coproduct are given by*

$$\Delta\big(t^{c_1 \cdots c_m}_{d_1 \cdots d_m}(u)\big) = \sum_{a_1 < \cdots < a_m} t^{c_1 \cdots c_m}_{a_1 \cdots a_m}(u) \otimes t^{a_1 \cdots a_m}_{d_1 \cdots d_m}(u), \tag{2.26}$$

summed over all subsets of indices $\{a_1, \dots, a_m\}$ from $\{1, \dots, n\}$.

PROOF. Using the notation of Section 2.5 we can write the image of the left-hand side of (2.22) under the coproduct Δ as

$$A_m T_{[1]1}(u) T_{[2]1}(u) \cdots T_{[1]m}(u-m+1) T_{[2]m}(u-m+1).$$

Since $m!A_m = A_m^2$, by (2.22) this coincides with

$$\frac{1}{m!} A_m T_{[1]1}(u) \cdots T_{[1]m}(u-m+1) A_m T_{[2]1}(u) \cdots T_{[2]m}(u-m+1).$$

Taking here the matrix elements and using the skew-symmetry of the quantum minors we find (2.26).

COROLLARY 2.15. *We have*

$$\Delta : \operatorname{qdet} T(u) \mapsto \operatorname{qdet} T(u) \otimes \operatorname{qdet} T(u).$$

2.7. *The center of* $\mathrm{Y}(n)$

PROPOSITION 2.16. *We have the relations*

$$\begin{aligned}&\big[t_{ab}(u), t^{c_1 \cdots c_m}_{d_1 \cdots d_m}(v)\big] \\ &\quad = \frac{1}{u-v} \left(\sum_{i=1}^{m} t_{c_i b}(u) t^{c_1 \cdots a \cdots c_m}_{d_1 \cdots d_m}(v) - \sum_{i=1}^{m} t^{c_1 \cdots c_m}_{d_1 \cdots b \cdots d_m}(v) t_{a d_i}(u) \right),\end{aligned}$$

where the indices a and b in the quantum minors replace c_i and d_i, respectively.

COROLLARY 2.17. *For any indices i, j we have*

$$\big[t_{c_i d_j}(u), t^{c_1 \cdots c_m}_{d_1 \cdots d_m}(v)\big] = 0.$$

Recall the elements $d_i \in \mathrm{Y}(n)$ are defined by (2.23).

THEOREM 2.18. *The coefficients $d_1, d_2, \ldots$ of the series* $\operatorname{qdet} T(u)$ *belong to the center of the algebra* $\mathrm{Y}(n)$. *Moreover, these elements are algebraically independent and generate the center of* $\mathrm{Y}(n)$.

OUTLINE OF THE PROOF. The first claim follows from Corollary 2.17. To prove the second claim introduce a filtration on $\mathrm{Y}(n)$ by setting

$$\deg_2 t_{ij}^{(r)} = r - 1.$$

The corresponding graded algebra $\operatorname{gr}_2 \mathrm{Y}(n)$ is isomorphic to the universal enveloping algebra $\mathrm{U}(\mathfrak{gl}_n[x])$ where $\mathfrak{gl}_n[x]$ is the Lie algebra of polynomials in an indeterminate x with coefficients in $\mathfrak{gl}_n$. Indeed, denote by $\bar{t}_{ij}^{(r)}$ the image of $t_{ij}^{(r)}$ in the $(r-1)$-th component of $\mathrm{gr}_2\,\mathrm{Y}(n)$. Then $E_{ij}x^{r-1} \mapsto \bar{t}_{ij}^{(r)}$ is an algebra homomorphism $\mathrm{U}(\mathfrak{gl}_n[x]) \to \mathrm{gr}_2\,\mathrm{Y}(n)$. Its kernel is trivial by Theorem 2.6.

We observe now from (2.24) (with $\rho = 1$) that the coefficient d_r of $\mathrm{qdet}\, T(u)$ has the form

$$d_r = t_{11}^{(r)} + \cdots + t_{nn}^{(r)} + \text{terms of degree} < r - 1.$$

This implies that the elements d_r, $r \geqslant 1$ are algebraically independent. Furthermore, the image of d_r in the $(r-1)$-th component of $\mathrm{gr}_2\,\mathrm{Y}(n)$ coincides with Zx^{r-1} where $Z = E_{11} + \cdots + E_{nn}$. It remains to note that the elements Zx^{r-1} with $r \geqslant 1$ generate the center of $\mathrm{U}(\mathfrak{gl}_n[x])$. The latter follows from the fact that the center of $\mathrm{U}(\mathfrak{sl}_n[x])$ is trivial [22]; see also [119, Proposition 2.12]. □

2.8. *The Yangian for the special linear Lie algebra*

For any simple Lie algebra $\mathfrak{a}$ over $\mathbb{C}$ the corresponding Yangian $\mathrm{Y}(\mathfrak{a})$ is a deformation of the universal enveloping algebra $\mathrm{U}(\mathfrak{a}[x])$ in the class of Hopf algebras. Two different presentations of $\mathrm{Y}(\mathfrak{a})$ are given in [35] and [38]. The type A Yangian $\mathrm{Y}(\mathfrak{sl}_n)$ can also be realized as a Hopf subalgebra of $\mathrm{Y}(n)$, as well as a quotient of $\mathrm{Y}(n)$.

For any series $f(u) \in 1 + u^{-1}\mathbb{C}[\![u^{-1}]\!]$ denote by μ_f the automorphism (2.15) of $\mathrm{Y}(n)$.

DEFINITION 2.19. The *Yangian for* $\mathfrak{sl}_n$ is the subalgebra $\mathrm{Y}(\mathfrak{sl}_n)$ of $\mathrm{Y}(n)$ which consists of the elements stable under all automorphisms μ_f.

We let $\mathrm{Z}(n)$ denote the center of $\mathrm{Y}(n)$.

THEOREM 2.20. *The subalgebra* $\mathrm{Y}(\mathfrak{sl}_n)$ *is a Hopf algebra whose coproduct, antipode and counit are obtained by restricting those from* $\mathrm{Y}(n)$. *Moreover,* $\mathrm{Y}(n)$ *is isomorphic to the tensor product of its subalgebras*

$$\mathrm{Y}(n) = \mathrm{Z}(n) \otimes \mathrm{Y}(\mathfrak{sl}_n). \tag{2.27}$$

OUTLINE OF THE PROOF. It is easy to verify that there exists a unique formal power series $\tilde{d}(u)$ in u^{-1} with coefficients in $\mathrm{Z}(n)$ which begins with 1 and satisfies

$$\tilde{d}(u)\tilde{d}(u-1)\cdots\tilde{d}(u-n+1) = \mathrm{qdet}\, T(u).$$

Then by Proposition 2.12, the image of $\tilde{d}(u)$ under μ_f is given by

$$\mu_f : \tilde{d}(u) \mapsto f(u)\tilde{d}(u). \tag{2.28}$$

This implies that all coefficients of the series

$$\tau_{ij}(u) = \tilde{d}(u)^{-1} t_{ij}(u)$$

belong to the subalgebra $\mathrm{Y}(\mathfrak{sl}_n)$. In fact, they generate this subalgebra. Furthermore, the coefficients of the series $\tilde{d}(u)$ are algebraically independent over $\mathrm{Y}(\mathfrak{sl}_n)$ which gives (2.27). Corollary 2.15 implies that

$$\Delta : \tilde{d}(u) \mapsto \tilde{d}(u) \otimes \tilde{d}(u)$$

and so the image of $\mathrm{Y}(\mathfrak{sl}_n)$ under the coproduct is contained in $\mathrm{Y}(\mathfrak{sl}_n) \otimes \mathrm{Y}(\mathfrak{sl}_n)$. Using Definition 2.11, we find that the image of $\operatorname{qdet} T(u)$ under the antipode S is $(\operatorname{qdet} T(u))^{-1}$ and so,

$$\mathrm{S} : \tilde{d}(u)^{-1}\, T(u) \mapsto \tilde{d}(u)\, T^{-1}(u).$$

Due to (2.28), $\mathrm{Y}(\mathfrak{sl}_n)$ is stable under S. $\square$

COROLLARY 2.21. *The algebra* $\mathrm{Y}(\mathfrak{sl}_n)$ *is isomorphic to the quotient of* $\mathrm{Y}(n)$ *by the ideal generated by the center, i.e.*

$$\mathrm{Y}(\mathfrak{sl}_n) \cong \mathrm{Y}(n)/\big(\operatorname{qdet} T(u) = 1\big).$$

2.9. *Two more realizations of the Hopf algebra* $\mathrm{Y}(\mathfrak{sl}_2)$

DEFINITION 2.22. Let $\mathcal{A}$ denote the associative unital algebra with six generators $e, f, h, J(e), J(f), J(h)$ and the defining relations

$$\begin{aligned} &[e, f] = h, \qquad [h, e] = 2e, \qquad [h, f] = -2f, \\ &\big[x, J(y)\big] = J\big([x, y]\big), \qquad J(ax) = aJ(x), \end{aligned}$$

where $x, y \in \{e, f, h\},\ a \in \mathbb{C}$, and

$$\big[[J(e), J(f)], J(h)\big] = \big(J(e)f - eJ(f)\big)h.$$

The Hopf algebra structure on $\mathcal{A}$ is defined by

$$\begin{aligned} &\Delta : x \mapsto x \otimes 1 + 1 \otimes x, \\ &\Delta : J(x) \mapsto J(x) \otimes 1 + 1 \otimes J(x) + \tfrac{1}{2}[x \otimes 1, t], \\ &\mathrm{S} : x \mapsto -x, \qquad J(x) \mapsto -J(x) + x, \\ &\varepsilon : x \mapsto 0, \qquad J(x) \mapsto 0, \end{aligned}$$

where $t = e \otimes f + f \otimes e + \frac{1}{2} h \otimes h$.

This definition generalizes to any simple Lie algebra $\mathfrak{a}$. Given a basis $e_1, \dots, e_n$ of $\mathfrak{a}$, the corresponding Yangian $\mathrm{Y}(\mathfrak{a})$ is a Hopf algebra generated by $2n$ elements $e_i,\ J(e_i)$, $i = 1, \dots, n$, as originally defined by Drinfeld [35].

THEOREM 2.23. *The mapping*

$$
\begin{aligned}
&e \mapsto t_{12}^{(1)}, \qquad f \mapsto t_{21}^{(1)}, \qquad h \mapsto t_{11}^{(1)} - t_{22}^{(1)},\\
&J(e) \mapsto t_{12}^{(2)} - \tfrac{1}{2}\big(t_{11}^{(1)} + t_{22}^{(1)} - 1\big)t_{12}^{(1)},\\
&J(f) \mapsto t_{21}^{(2)} - \tfrac{1}{2}\big(t_{11}^{(1)} + t_{22}^{(1)} - 1\big)t_{21}^{(1)},\\
&J(h) \mapsto t_{11}^{(2)} - t_{22}^{(2)} - \tfrac{1}{2}\big(t_{11}^{(1)} + t_{22}^{(1)} - 1\big)\big(t_{11}^{(1)} - t_{22}^{(1)}\big)
\end{aligned}
$$

defines a Hopf algebra isomorphism $\mathcal{A} \to \mathrm{Y}(\mathfrak{sl}_2)$.

DEFINITION 2.24. Let $\mathcal{B}$ be the associative algebra with generators e_k, f_k, h_k where $k = 0, 1, 2, \ldots$ and the defining relations given in terms of the generating series

$$
\begin{aligned}
&e(u) = \sum_{k=0}^{\infty} e_k u^{-k-1}, \qquad f(u) = \sum_{k=0}^{\infty} f_k u^{-k-1},\\
&h(u) = 1 + \sum_{k=0}^{\infty} h_k u^{-k-1}
\end{aligned}
$$

as follows

$$
\begin{aligned}
&\big[h(u), h(v)\big] = 0, \qquad \big[e(u), f(v)\big] = -\frac{h(u) - h(v)}{u - v},\\
&\big[e(u), e(v)\big] = -\frac{(e(u) - e(v))^2}{u - v},\\
&\big[f(u), f(v)\big] = \frac{(f(u) - f(v))^2}{u - v},\\
&\big[h(u), e(v)\big] = -\frac{\{h(u), e(u) - e(v)\}}{u - v},\\
&\big[h(u), f(v)\big] = \frac{\{h(u), f(u) - f(v)\}}{u - v},
\end{aligned}
$$

where we have used the notation $\{a, b\} = ab + ba$. The Hopf algebra structure on $\mathcal{B}$ is defined by the coproduct

$$
\begin{aligned}
&\Delta : e(u) \mapsto e(u) \otimes 1 + \sum_{k=0}^{\infty} (-1)^k f(u+1)^k h(u) \otimes e(u)^{k+1},\\
&\Delta : f(u) \mapsto 1 \otimes f(u) + \sum_{k=0}^{\infty} (-1)^k f(u)^{k+1} \otimes h(u) e(u+1)^k,\\
&\Delta : h(u) \mapsto \sum_{k=0}^{\infty} (-1)^k (k+1) f(u+1)^k h(u) \otimes h(u) e(u+1)^k,
\end{aligned}
$$

the antipode

$$\begin{aligned}
&\mathrm{S}: e(u) \mapsto -\big(h(u) + f(u+1)e(u)\big)^{-1} e(u),\\
&\mathrm{S}: f(u) \mapsto -f(u)\big(h(u) + f(u)e(u+1)\big)^{-1},\\
&\mathrm{S}: h(u) \mapsto \big(h(u) + f(u+1)e(u)\big)^{-1}\\
&\qquad\qquad \cdot \big(1 - f(u+1)\big(h(u) + f(u+1)e(u)\big)^{-1} e(u)\big),
\end{aligned}$$

and the counit

$$\varepsilon: e(u) \mapsto 0, \qquad f(u) \mapsto 0, \qquad h(u) \mapsto 1.$$

Explicitly, the defining relations of $\mathcal{B}$ can be written in the form

$$\begin{aligned}
&[h_k, h_l] = 0, \qquad [e_k, f_l] = h_{k+l},\\
&[h_0, e_k] = 2e_k, \qquad [h_0, f_k] = -2f_k,\\
&[e_{k+1}, e_l] - [e_k, e_{l+1}] = e_k e_l + e_l e_k,\\
&[f_{k+1}, f_l] - [f_k, f_{l+1}] = -f_k f_l - f_l f_k,\\
&[h_{k+1}, e_l] - [h_k, e_{l+1}] = h_k e_l + e_l h_k,\\
&[h_{k+1}, f_l] - [h_k, f_{l+1}] = -h_k f_l - f_l h_k.
\end{aligned}$$

Such a realization exists for an arbitrary Yangian $\mathrm{Y}(\mathfrak{a})$. Some authors call it the *new realization* following the title of Drinfeld's paper [38] where it was introduced. This presentation of the Yangian is most convenient to describe its finite-dimensional irreducible representations [38]. However, in the case of an arbitrary simple Lie algebra $\mathfrak{a}$ no explicit formulas for the coproduct and antipode are known.

THEOREM 2.25. *The mapping*

$$\begin{aligned}
&e(u) \mapsto t_{22}(u)^{-1} t_{12}(u),\\
&f(u) \mapsto t_{21}(u) t_{22}(u)^{-1},\\
&h(u) \mapsto t_{11}(u) t_{22}(u)^{-1} - t_{21}(u) t_{22}(u)^{-1} t_{12}(u) t_{22}(u)^{-1},
\end{aligned}$$

defines a Hopf algebra isomorphism $\mathcal{B} \to \mathrm{Y}(\mathfrak{sl}_2)$.

Combining the two above theorems we obtain a Hopf algebra isomorphism $\mathcal{A} \to \mathcal{B}$ given by

$$e \mapsto e_0, \qquad f \mapsto f_0, \qquad h \mapsto h_0,$$

and

$$\begin{aligned}
&J(e) \mapsto e_1 - \tfrac{1}{4}(e_0 h_0 + h_0 e_0),\\
&J(f) \mapsto f_1 - \tfrac{1}{4}(f_0 h_0 + h_0 f_0),\\
&J(h) \mapsto h_1 + \tfrac{1}{2}(e_0 f_0 + f_0 e_0 - h_0^2).
\end{aligned}$$

As we have seen in the proof of Theorem 2.18, the graded algebra $\mathrm{gr}_2\,\mathrm{Y}(\mathfrak{sl}_2)$ is isomorphic to the universal enveloping algebra $\mathrm{U}(\mathfrak{sl}_2[x])$. The images of the generators of the algebra $\mathcal{A}$ in the graded algebra clearly correspond to e, f, h, ex, fx, hx while the images of the generators e_k, f_k, h_k of $\mathcal{B}$ correspond to ex^k, fx^k, hx^k.

2.10. *Quantum Liouville formula*

Here we give another family of generators of the center of $\mathrm{Y}(n)$. Introduce the series $z(u)$ with coefficients from $\mathrm{Y}(n)$ by the formula

$$z(u)^{-1} = \tfrac{1}{n}\,\mathrm{tr}\big(T(u)T^{-1}(u-n)\big), \tag{2.29}$$

so that

$$z(u) = 1 + z_2u^{-2} + z_3u^{-3} + \cdots, \quad z_i \in \mathrm{Y}(n).$$

DEFINITION 2.26. The *quantum comatrix* $\widehat{T}(u)$ is defined by

$$\widehat{T}(u)T(u-n+1) = \mathrm{qdet}\,T(u). \tag{2.30}$$

PROPOSITION 2.27. *The matrix elements* $\hat{t}_{ij}(u)$ *of the matrix* $\widehat{T}(u)$ *are given by*

$$\hat{t}_{ij}(u) = (-1)^{i+j}\,t^{1\cdots\hat{j}\cdots n}_{1\cdots\hat{i}\cdots n}(u), \tag{2.31}$$

where the hats on the right-hand side indicate the indices to be omitted. Moreover, we have the relation

$$\widehat{T}^t(u-1)T^t(u) = \mathrm{qdet}\,T(u). \tag{2.32}$$

PROOF. Using Definition 2.11 we derive from (2.30)

$$A_nT_1(u)\cdots T_{n-1}(u-n+2) = A_n\widehat{T}_n(u).$$

Taking the matrix elements we come to (2.31). Further, consider the automorphism $\varphi: T(u) \mapsto T^t(-u)$ of $\mathrm{Y}(n)$; see Proposition 2.5. Using Proposition 2.12 we find

$$\varphi: \mathrm{qdet}\,T(u) \mapsto \mathrm{qdet}\,T(-u+n-1),$$
$$\varphi: \widehat{T}(u) \mapsto \widehat{T}^t(-u+n-2).$$

Now applying φ to (2.30) and replacing $-u+n-1$ by u we get (2.32). □

The following is a 'quantum' analog of the classical Liouville formula; see [119, Remark 5.8] for more comment.

THEOREM 2.28. *We have the relation*

$$z(u) = \frac{\operatorname{qdet} T(u-1)}{\operatorname{qdet} T(u)}. \tag{2.33}$$

PROOF. From (2.29) and (2.30) we find

$$z(u)^{-1} = \tfrac{1}{n} \operatorname{tr}\big(T(u)\widehat{T}(u-1)\big(\operatorname{qdet} T(u-1)\big)^{-1}\big).$$

Using the centrality of $\operatorname{qdet} T(u)$ and (2.32) we get (2.33). $\square$

COROLLARY 2.29. *The coefficients $z_2, z_3, \dots$ of $z(u)$ are algebraically independent generators of the center of* $\mathrm{Y}(n)$.

PROPOSITION 2.30. *The square of the antipode* S *is the automorphism of* $\mathrm{Y}(n)$ *given by*

$$\mathrm{S}^2 : T(u) \mapsto z(u+n)T(u+n).$$

In particular, $\operatorname{qdet} T(u)$ *is stable under* S^2.

OUTLINE OF THE PROOF. The series $z(u)$ can be defined equivalently in a way similar to the quantum determinant; cf. Definition 2.11. Multiply the ternary relation (2.14) by $T_2^{-1}(v)$ from both sides and take the transposition with respect to the second copy of $\operatorname{End}\mathbb{C}^n$. We obtain the relation

$$R^t(u-v)\widetilde{T}_2(v)T_1(u) = T_1(u)\widetilde{T}_2(v)R^t(u-v), \tag{2.34}$$

where $\widetilde{T}(v) = (T^{-1}(v))^t$ and

$$R^t(u) = 1 - Qu^{-1}, \qquad Q = \sum_{i,j=1}^{n} e_{ij} \otimes e_{ij}.$$

Observe that Q is a one-dimensional operator satisfying $Q^2 = nQ$. Therefore, $R^t(u)^{-1}$ has a simple pole at $u = n$ with residue Q. Relation (2.34) now implies

$$QT_1(u)\widetilde{T}_2(u-n) = \widetilde{T}_2(u-n)T_1(u)Q, \tag{2.35}$$

and this element equals $Qz(u)^{-1}$. It now suffices to apply the antipode S to both sides of the identity $T(u)T^{-1}(u) = 1$. The second claim follows from Theorem 2.28. $\square$

2.11. *Factorization of the quantum determinant*

DEFINITION 2.31. Let X be a square matrix over a ring with 1. Suppose that there exists the inverse matrix X^{-1} and its ji-th entry $(X^{-1})_{ji}$ is an invertible element of the ring. Then the *ij-th quasi-determinant of X* is defined by the formula

$$|X|_{ij} = \big((X^{-1})_{ji}\big)^{-1}.$$

For any $1 \leqslant m \leqslant n$ denote by $T^{(m)}(u)$ the submatrix of $T(u)$ corresponding to the first m rows and columns.

THEOREM 2.32. *The quantum determinant* $\operatorname{qdet} T(u)$ *admits the following factorization in the algebra* $\mathrm{Y}(n)[[u^{-1}]]$

$$\operatorname{qdet} T(u) = t_{11}(u)\big|T^{(2)}(u-1)\big|_{22} \cdots \big|T^{(n)}(u-n+1)\big|_{nn}. \tag{2.36}$$

Moreover, the factors on the right-hand side are permutable.

PROOF. By Definition 2.26 we have

$$\widehat{T}(u) = \operatorname{qdet} T(u) T^{-1}(u-n+1).$$

Taking the nn-th entry we see

$$\operatorname{qdet} T(u)\big(T^{-1}(u-n+1)\big)_{nn} = \hat{t}_{nn}(u).$$

Proposition 2.27 gives

$$\operatorname{qdet} T(u) = \operatorname{qdet} T^{(n-1)}(u)\big|T^{(n)}(u-n+1)\big|_{nn}.$$

Note that the factors here commute by the centrality of the quantum determinant. An obvious induction completes the proof. $\square$

There is a generalization of this result providing a block factorization of $\operatorname{qdet} T(u)$. For subsets $\mathcal{P}$ and $\mathcal{Q}$ in $\{1,\dots,n\}$ and an $n \times n$-matrix X we shall denote by $X_{\mathcal{P}\mathcal{Q}}$ the submatrix of X whose rows and columns are enumerated by $\mathcal{P}$ and $\mathcal{Q}$, respectively. Fix an integer $0 \leqslant m \leqslant n$ and set $\mathcal{A} = \{1,\dots,m\}$ and $\mathcal{B} = \{m+1,\dots,n\}$. We let $T^*(u)$ denote the matrix $T^{-1}(-u)$.

THEOREM 2.33. *We have the identity*

$$\operatorname{qdet} T(u) \operatorname{qdet} T^*(-u+n-1)_{\mathcal{A}\mathcal{A}} = \operatorname{qdet} T(u)_{\mathcal{B}\mathcal{B}}.$$

We keep the notation $t'_{ij}(u)$ for the matrix elements of the matrix $T^{-1}(u)$.

PROPOSITION 2.34. *We have the relations*

$$\big[t_{ij}(u), t'_{kl}(v)\big] = \frac{1}{u-v}\left(\delta_{kj}\sum_{a=1}^{n} t_{ia}(u)t'_{al}(v) - \delta_{il}\sum_{a=1}^{n} t'_{ka}(v)t_{aj}(u)\right).$$

In particular, the matrix elements of the matrices $T(u)_{\mathcal{A}\mathcal{A}}$ and $T^*(v)_{\mathcal{B}\mathcal{B}}$ commute with each other.

PROOF. It suffices to multiply the ternary relation (2.14) by $T_2^{-1}(v)$ from both sides and equate the matrix elements. □

2.12. *Quantum Sylvester theorem*

The following commutation relations between the quantum minors generalize Proposition 2.16.

PROPOSITION 2.35. *We have the relations*

$$\begin{aligned}\big[t^{a_1\cdots a_k}_{b_1\cdots b_k}(u), t^{c_1\cdots c_l}_{d_1\cdots d_l}(v)\big] = \sum_{p=1}^{\min\{k,l\}} & \frac{(-1)^{p-1}p!}{(u-v-k+1)\cdots(u-v-k+p)} \\ & \times \sum_{\substack{i_1<\cdots<i_p \\ j_1<\cdots<j_p}} \big(t^{a_1\cdots c_{j_1}\cdots c_{j_p}\cdots a_k}_{b_1\cdots b_k}(u)\, t^{c_1\cdots a_{i_1}\cdots a_{i_p}\cdots c_l}_{d_1\cdots d_l}(v) \\ & \qquad - t^{c_1\cdots c_l}_{d_1\cdots b_{i_1}\cdots b_{i_p}\cdots d_l}(v)\, t^{a_1\cdots a_k}_{b_1\cdots d_{j_1}\cdots d_{j_p}\cdots b_k}(u)\big).\end{aligned}$$

Here the p-tuples of upper indices $(a_{i_1},\dots,a_{i_p})$ and $(c_{j_1},\dots,c_{j_p})$ are respectively interchanged in the first summand on the right-hand side while the p-tuples of lower indices $(b_{i_1},\dots,b_{i_p})$ and $(d_{j_1},\dots,d_{j_p})$ are interchanged in the second summand.

As in the previous section, we fix an integer m satisfying $1 \leqslant m \leqslant n$. For any indices $1 \leqslant i, j \leqslant m$ introduce the following series with coefficients in $\mathrm{Y}(n)$

$$\tilde{t}_{ij}(u) = t^{i,m+1\cdots n}_{j,m+1\cdots n}(u)$$

and combine them into the matrix $\widetilde{T}(u) = (\tilde{t}_{ij}(u))$. The following is an analog of the classical Sylvester theorem. As before, we denote $\mathcal{B} = \{m+1,\dots,n\}$.

THEOREM 2.36. *The mapping*

$$t_{ij}(u) \mapsto \tilde{t}_{ij}(u), \quad 1 \leqslant i, j \leqslant m,$$

defines an algebra homomorphism $\mathrm{Y}(m) \to \mathrm{Y}(n)$. *Moreover, one has the identity*

$$\operatorname{qdet}\widetilde{T}(u) = \operatorname{qdet}T(u)\operatorname{qdet}T(u-1)_{\mathcal{B}\mathcal{B}}\cdots\operatorname{qdet}T(u-m+1)_{\mathcal{B}\mathcal{B}}. \tag{2.37}$$

OUTLINE OF THE PROOF. Using Proposition 2.35 we check that the series $\tilde{t}_{ij}(u)$ satisfy the Yangian defining relations (2.3) which proves the first claim. The identity (2.37) is derived by induction on m from the relation

$$\widehat{T}(u)_{\mathcal{A}\mathcal{A}}\widetilde{T}(u-m+1) = \operatorname{qdet} T(u) \operatorname{qdet} T(u-m+1)_{\mathcal{B}\mathcal{B}},$$

where $\widehat{T}(u)$ is the quantum comatrix; see Definition 2.26. $\square$

2.13. *The centralizer construction*

Fix a nonnegative integer m and for any $n \geqslant m$ denote by $\mathfrak{g}_m(n)$ the subalgebra in $\mathfrak{gl}_n$ spanned by the basis elements E_{ij} with $m+1 \leqslant i, j \leqslant n$. The subalgebra $\mathfrak{g}_m(n)$ is isomorphic to $\mathfrak{gl}_{n-m}$. Let $\mathrm{A}_m(n)$ denote the centralizer of $\mathfrak{g}_m(n)$ in the universal enveloping algebra $\mathrm{A}(n) = \mathrm{U}(\mathfrak{gl}_n)$. Let $\mathrm{A}(n)^0$ denote the centralizer of E_{nn} in $\mathrm{A}(n)$ and let $\mathrm{I}(n)$ be the left ideal in $\mathrm{A}(n)$ generated by the elements E_{in}, $i = 1, \dots, n$. Then $\mathrm{I}(n)^0 = \mathrm{I}(n) \cap \mathrm{A}(n)^0$ is a two-sided ideal in $\mathrm{A}(n)^0$ and one has a vector space decomposition

$$\mathrm{A}(n)^0 = \mathrm{I}(n)^0 \oplus \mathrm{A}(n-1).$$

Therefore the projection of $\mathrm{A}(n)^0$ onto $\mathrm{A}(n-1)$ with the kernel $\mathrm{I}(n)^0$ is an algebra homomorphism. Its restriction to the subalgebra $\mathrm{A}_m(n)$ defines a filtration preserving homomorphism

$$\pi_n : \mathrm{A}_m(n) \to \mathrm{A}_m(n-1) \tag{2.38}$$

so that one can define the algebra A_m as the projective limit with respect to this sequence of homomorphisms in the category of filtered algebras.

By the Harish-Chandra isomorphism [34, Section 7.4], the center $\mathrm{A}_0(n)$ of $\mathrm{U}(\mathfrak{gl}_n)$ is naturally isomorphic to the algebra $\Lambda^*(n)$ of polynomials in n variables $\lambda_1, \dots, \lambda_n$ which are symmetric in the shifted variables $\lambda_1, \lambda_2 - 1, \dots, \lambda_n - n + 1$. So, in the case $m = 0$ the homomorphisms π_n are interpreted as the specialization homomorphisms $\pi_n : \Lambda^*(n) \to \Lambda^*(n-1)$ such that

$$\pi_n : f(\lambda_1, \dots, \lambda_n) \mapsto f(\lambda_1, \dots, \lambda_{n-1}, 0). \tag{2.39}$$

The corresponding projective limit in the category of filtered algebras is called the *algebra of shifted symmetric functions* and denoted by Λ^*. The elements of Λ^* are well-defined functions on the set of all sequences $\lambda = (\lambda_1, \lambda_2, \dots)$ which contain only a finite number of nonzero terms. The following families of elements of Λ^* are analogs of power sums, elementary symmetric functions and complete symmetric functions:

$$p_m(\lambda) = \sum_{k=1}^{\infty} \big((\lambda_k - k)^m - (-k)^m\big), \quad m = 1, 2, \dots,$$

$$1+\sum_{m=1}^{\infty} e_m(\lambda)t^m=\prod_{k=1}^{\infty}\frac{1+(\lambda_k-k)t}{1-kt},$$
$$1+\sum_{m=1}^{\infty} h_m(\lambda)t^m=\prod_{k=1}^{\infty}\frac{1+kt}{1-(\lambda_k-k)t}.$$

Each of the families $\{p_m\}$, $\{e_m\}$, $\{h_m\}$ can be taken as a system of algebraically independent generators of the algebra Λ^*. To summarize, we have the following.

PROPOSITION 2.37. *The algebra* A_0 *is isomorphic to the algebra of shifted symmetric functions* Λ^*.

Now consider the homomorphism (2.17) and take its composition with an automorphism of $\mathrm{Y}(n)$ given by (2.16) to yield another homomorphism $\varphi_n : \mathrm{Y}(n) \to \mathrm{A}(n)$ such that

$$\varphi_n : T(u) \mapsto \left(1-\frac{E}{u+n}\right)^{-1}.$$

It follows from the defining relations (2.3) that the image of the restriction of φ_n to the subalgebra $\mathrm{Y}(m)$ is contained in the centralizer $\mathrm{A}_m(n)$ thus yielding a homomorphism $\varphi_n : \mathrm{Y}(m) \to \mathrm{A}_m(n)$.

THEOREM 2.38. *For any fixed* $m \geqslant 1$ *the sequence* $(\varphi_n \mid n \geqslant m)$ *defines an algebra embedding* $\varphi : \mathrm{Y}(m) \hookrightarrow \mathrm{A}_m$. *Moreover, one has an isomorphism*

$$\mathrm{A}_m = \mathrm{A}_0 \otimes \mathrm{Y}(m), \tag{2.40}$$

where $\mathrm{Y}(m)$ *is identified with its image under the embedding* φ.

OUTLINE OF THE PROOF. First we verify that the family $(\varphi_n \mid n \geqslant m)$ is compatible with the chain of homomorphisms (2.38). Further, to prove the injectivity of φ we consider the corresponding commutative picture replacing $\mathrm{A}(n)$ with its graded algebra $\operatorname{gr}\mathrm{A}(n) \cong \mathrm{S}(\mathfrak{gl}_n)$. This reduces the task to the description of the invariants in the symmetric algebra $\mathrm{S}(\mathfrak{gl}_n)$ with respect to the action of the group $GL(n-m)$ corresponding to the Lie algebra $\mathfrak{g}_m(n)$. This can be done with the use of the classical invariant theory [172] which also implies the decomposition (2.40). $\square$

A different embedding $\mathrm{Y}(m) \hookrightarrow \mathrm{A}_m$ can be constructed with the use of the quantum Sylvester theorem; see Section 2.12. Consider the homomorphism $\mathrm{Y}(m) \to \mathrm{Y}(n)$ provided by Theorem 2.36 and take its composition with (2.7). We obtain an algebra homomorphism $\psi_n : \mathrm{Y}(m) \to \mathrm{A}(n)$ given by

$$\psi_n : t_{ij}(u) \mapsto \operatorname{qdet}(1+Eu^{-1})_{\mathcal{B}_i\mathcal{B}_j}, \tag{2.41}$$

where $\mathcal{B}_i$ denotes the set $\{i, m+1, \dots, n\}$. Corollary 2.17 implies that this image commutes with the elements of the subalgebra $\mathfrak{g}_m(n)$ and so (2.41) defines a homomorphism

$\psi_n : \mathrm{Y}(m) \to \mathrm{A}_m(n)$. Furthermore, the family of homomorphisms $(\psi_n \mid n \geqslant m)$ is obviously compatible with the projections (2.38) and thus defines an algebra homomorphism $\psi : \mathrm{Y}(m) \to \mathrm{A}_m$. Denote by $\widetilde{\mathrm{A}}_0$ the projective limit of the sequence of the centers of the universal enveloping algebras $\mathrm{U}(\mathfrak{g}_m(n))$, where $n = m, m+1, \dots,$ defined by the corresponding homomorphisms (2.39). By Proposition 2.37, $\widetilde{\mathrm{A}}_0$ is isomorphic to the algebra of shifted symmetric functions in the variables $\lambda_{m+1}, \lambda_{m+2}, \dots$. The following is an analog of Theorem 2.38.

THEOREM 2.39. *The homomorphism $\psi : \mathrm{Y}(m) \to \mathrm{A}_m$ is injective. Moreover, one has an isomorphism*

$$\mathrm{A}_m = \widetilde{\mathrm{A}}_0 \otimes \mathrm{Y}(m),$$

where $\mathrm{Y}(m)$ is identified with its image under the embedding ψ.

2.14. *Commutative subalgebras*

Here we use the notation introduced in Section 2.2. Consider the algebra (2.9) with $m = n$. Fix an $n \times n$ matrix C with entries in $\mathbb{C}$ and for any $1 \leqslant k \leqslant n$ introduce the series $\tau_k(u, C)$ with coefficients in $\mathrm{Y}(n)$ by

$$\tau_k(u, C) = \operatorname{tr} A_n T_1(u) \cdots T_k(u - k + 1) C_{k+1} \cdots C_n,$$

where A_n is the antisymmetrizer defined by (2.21) and the trace is taken over all n copies of $\operatorname{End} \mathbb{C}^n$.

THEOREM 2.40. *All the coefficients of the series $\tau_1(u, C), \dots, \tau_n(u, C)$ commute with each other. Moreover, if the matrix C has simple spectrum then the coefficients at $u^{-1}, u^{-2}, \dots$ of these series are algebraically independent and generate a maximal commutative subalgebra of the Yangian $\mathrm{Y}(n)$.*

Consider the epimorphism $\pi : \mathrm{Y}(n) \to \mathrm{U}(\mathfrak{gl}_n)$ defined in (2.7). Clearly, the coefficients of the images of the series $\tau_k(u, C)$, $k = 1, \dots, n$, under π form a commutative subalgebra $\mathcal{C} \subseteq \mathrm{U}(\mathfrak{gl}_n)$.

THEOREM 2.41. *If the matrix C has simple spectrum then the subalgebra $\mathcal{C}$ of $\mathrm{U}(\mathfrak{gl}_n)$ is maximal commutative.*

Bibliographical notes

2.2. For the origins of the *RTT* relation and associated R-matrix formalism see for instance the papers by Takhtajan and Faddeev [161], Kulish and Sklyanin [89], Reshetikhin, Takhtajan and Faddeev [149]. The statistical mechanics background of quantum groups is also explained in the book by Chari and Pressley [25, Chapter 7].

2.4. The Poincaré–Birkhoff–Witt theorem for general Yangians is due to Drinfeld (unpublished). Another proof was given by Levendorskiĭ [97]. The details of the proof outlined here can be found in [119]. It follows the approach of Olshanski [138].

2.6. The definition of the quantum determinant $\operatorname{qdet} T(u)$ (in the case $n = 2$) originally appeared in Izergin and Korepin [63]. The basic ideas and formulas associated with the quantum determinant for an arbitrary n are contained in survey paper by Kulish and Sklyanin [89]. Detailed proofs are given in [119]. Proposition 2.14 is contained, e.g., in Iohara [59] and Nazarov and Tarasov [130].

2.8. By the general approach of Drinfeld [37], the Yangian for $\mathfrak{sl}_n$ should be defined as a quotient algebra of $\mathrm{Y}(n)$. The fact that it can also be realized as a (Hopf) subalgebra of $\mathrm{Y}(n)$ was observed by Olshanski [119].

2.9. For any simple Lie algebra $\mathfrak{a}$ the Yangian $\mathrm{Y}(\mathfrak{a})$ was defined by Drinfeld [35,38]. The two definitions given here are particular cases for $\mathfrak{a} = \mathfrak{sl}_2$. The formulas for the coproduct and antipode in Definition 2.24 are due to the author; see, e.g., Khoroshkin and Tolstoy [70]. These formulas were employed in [70] in the construction of the double of the Yangian. These results were generalized to the Yangian $\mathrm{Y}(\mathfrak{sl}_3)$ by Soloviev [154], and to $\mathrm{Y}(\mathfrak{sl}_n)$ by Iohara [59].

2.10. The series $z(u)$ was introduced by Nazarov [124]. The quantum Liouville formula is also due to him. The argument given in Section 2.10 is a simplified version of his R-matrix proof [119].

2.11. A general theory of quasi-determinants of matrices over noncommutative rings is developed by Gelfand and Retakh [47,48]. Various analogs of the classical theorems for such determinants are given. Quasi-determinant factorizations of quantum determinants for the quantized algebra $GL_q(n)$ are also proved in those papers; see also Krob and Leclerc [84]. In a more general context of Hopf algebras such factorizations are constructed by Etingof and Retakh [41].

2.12. An analog of Sylvester's theorem for the algebra $GL_q(n)$ was given by Krob and Leclerc [84] with the use of the quasi-determinant version of this theorem due to Gelfand and Retakh [47]. The approach of [84] works for the Yangians as well. The proof outlined here follows [117] where a proof of Proposition 2.35 is given. The latter result and some other quantum minor relations are known to specialists as 'folklore theorems'. Some more quantum analogs of the classical minor relations are collected in Iohara [59].

2.13. The centralizer construction is due to Olshanski [136,138]. The modified version (Theorem 2.39) based on the quantum Sylvester theorem is given in [117]. The algebra Λ^* of shifted symmetric functions is studied in detail by Okounkov and Olshanski [134].

2.14. The commutative subalgebras in the Yangian originate from integrable models in statistical mechanics (specifically, from the *six vertex* or *XXX* model); see Baxter [9]. The common eigenvectors of such a commutative subalgebra in certain standard Yangian modules can be constructed by a special procedure called the *algebraic Bethe Ansatz*; see Faddeev [42], Kulish–Sklyanin [89], Kulish–Reshetikhin [85], Kirillov–Reshetikhin [75,76]. Theorems 2.40 and 2.41 are proved by Nazarov and Olshanski [128].

Finite-dimensional irreducible representations of the Yangians were classified by Drinfeld [38] with the use of the particular case of $\mathrm{Y}(\mathfrak{gl}_2)$ considered earlier by Tarasov [157, 158]. A detailed exposition of these results for $\mathrm{Y}(\mathfrak{gl}_n)$ is contained in [112]. Cherednik [29, 30] used the Yangians to 'materialize' the second Weyl character formula. Nazarov [124, 126] and Okounkov [133] employed the Yangian techniques to obtain remarkable immanant analogs of the classical Capelli identity [19,20]. An explicit construction of all irreducible finite-dimensional representations of $\mathrm{Y}(\mathfrak{gl}_2)$ as tensor products of the evaluation modules is given by Tarasov [158] and Chari and Pressley [23,26]. In particular, this provides an irreducibility criterion of tensor products of the $\mathrm{Y}(\mathfrak{gl}_2)$ evaluation modules. A generalization of this criterion to $\mathrm{Y}(\mathfrak{gl}_n)$ with an arbitrary n was given in [116]; see also Leclerc–Nazarov–Thibon [96]. An important part of the criterion is the *binary property* established by Nazarov and Tarasov [132]; see also Kitanine, Maillet and Terras [78, 103]. A character formula for an arbitrary finite-dimensional irreducible representation of $\mathrm{Y}(\mathfrak{gl}_n)$ is given by Arakawa [3] with the use of the Drinfeld functor; see Drinfeld [36]. This formula is also implied by the earlier results of Ginzburg and Vasserot [49] combined with the work of Lusztig; see Nakajima [122], Varagnolo [171]. The irreducible characters are expressed in terms of those for the 'standard tensor product modules' via the Kazhdan–Lusztig polynomials. Bases of Gelfand–Tsetlin type for 'generic' representations of $\mathrm{Y}(\mathfrak{gl}_n)$ are constructed in [107]. More general class of 'tame' Yangian modules was introduced and explicitly constructed by Nazarov and Tarasov [130]. The earlier works of Nazarov and Tarasov [129] and the author [107] provide different constructions of the well-known Gelfand–Tsetlin basis for representations of $\mathfrak{gl}_n$. A surprising connection of the Yangian $\mathrm{Y}(\mathfrak{gl}_n)$ with the finite $\mathcal{W}$-algebras was discovered by Ragoucy and Sorba [145,146]; see also Briot and Ragoucy [18]. The Yangian actions on certain modules over the affine Lie algebras were constructed by Uglov [167].

3. The twisted Yangians

Here we describe the structure of the twisted Yangians corresponding to the orthogonal and symplectic Lie algebras $\mathfrak{o}_N$ and $\mathfrak{sp}_N$. We consider both cases simultaneously, unless otherwise stated.

3.1. *Defining relations*

Given a positive integer N, we number the rows and columns of $N \times N$ matrices by the indices $\{-n, \dots, -1, 0, 1, \dots n\}$ if $N = 2n + 1$, and by $\{-n, \dots, -1, 1, \dots n\}$ if $N = 2n$.

Similarly, in the latter case the range of indices $-n \leqslant i, j \leqslant n$ will exclude 0. It will be convenient to use the symbol θ_{ij} which is defined by

$$\theta_{ij} = \begin{cases} 1 & \text{in the orthogonal case,} \\ \operatorname{sgn} i \cdot \operatorname{sgn} j & \text{in the symplectic case.} \end{cases}$$

Whenever the double sign $\pm$ or $\mp$ occurs, the upper sign corresponds to the orthogonal case and the lower sign to the symplectic case. By $X \mapsto X^t$ we will denote the matrix transposition such that for the matrix units we have

$$(e_{ij})^t = \theta_{ij} e_{-j,-i}. \tag{3.1}$$

Introduce the following elements of the Lie algebra $\mathfrak{gl}_N$:

$$F_{ij} = E_{ij} - \theta_{ij} E_{-j,-i}, \quad -n \leqslant i, j \leqslant n.$$

The Lie subalgebra of $\mathfrak{gl}_N$ spanned by the elements F_{ij} is a realization of a simple Lie algebra $\mathfrak{g}_n$ of rank n (see, e.g., [58]). In the orthogonal case $\mathfrak{g}_n$ is of type D_n or B_n if $N = 2n$ or $N = 2n + 1$, respectively. In the symplectic case, $N = 2n$ and $\mathfrak{g}_n$ is of type C_n. Thus,

$$\mathfrak{g}_n = \mathfrak{o}_{2n}, \quad \mathfrak{o}_{2n+1} \quad \text{or} \quad \mathfrak{sp}_{2n}.$$

DEFINITION 3.1. Each of the *twisted Yangians* $\mathrm{Y}^+(2n)$, $\mathrm{Y}^+(2n+1)$ and $\mathrm{Y}^-(2n)$ corresponding to the Lie algebras $\mathfrak{o}_{2n}$, $\mathfrak{o}_{2n+1}$ and $\mathfrak{sp}_{2n}$, respectively, is a unital associative algebra with generators $s_{ij}^{(1)}, s_{ij}^{(2)}, \ldots$ where $-n \leqslant i, j \leqslant n$, and the defining relations written in terms of the generating series

$$s_{ij}(u) = \delta_{ij} + s_{ij}^{(1)} u^{-1} + s_{ij}^{(2)} u^{-2} + \cdots$$

as follows

$$\begin{aligned}
&(u^2 - v^2)\big[s_{ij}(u), s_{kl}(v)\big] \\
&\quad = (u+v)\big(s_{kj}(u)s_{il}(v) - s_{kj}(v)s_{il}(u)\big) \\
&\qquad - (u-v)\big(\theta_{k,-j} s_{i,-k}(u)s_{-j,l}(v) - \theta_{i,-l} s_{k,-i}(v)s_{-l,j}(u)\big) \\
&\qquad + \theta_{i,-j}\big(s_{k,-i}(u)s_{-j,l}(v) - s_{k,-i}(v)s_{-j,l}(u)\big)
\end{aligned}$$

and

$$\theta_{ij} s_{-j,-i}(-u) = s_{ij}(u) \pm \frac{s_{ij}(u) - s_{ij}(-u)}{2u}. \tag{3.2}$$

These relations can be conveniently presented in an equivalent matrix form analogous to (2.14). For this introduce the transposed $R^t(u)$ for the Yang R-matrix (2.12) (correcting the indices for the matrix elements appropriately) by

$$R^t(u) = 1 - Qu^{-1}, \quad Q = \sum_{i,j=-n}^{n} e_{ij}^t \otimes e_{ji}.$$

Furthermore, denote by $S(u)$ the $N \times N$ matrix whose ij-th entry is $s_{ij}(u)$. As in (2.8) we regard $S(u)$ as an element of the algebra $\mathrm{Y}^{\pm}(N) \otimes \mathrm{End}\,\mathbb{C}^N$ given by

$$S(u) = \sum_{i,j=-n}^{n} s_{ij}(u) \otimes e_{ij}.$$

PROPOSITION 3.2. *The defining relations for the twisted Yangian* $\mathrm{Y}^{\pm}(N)$ *are equivalent to the quaternary relation*

$$R(u-v)S_1(u)R^t(-u-v)S_2(v) = S_2(v)R^t(-u-v)S_1(u)R(u-v) \tag{3.3}$$

and the symmetry relation

$$S^t(-u) = S(u) \pm \frac{S(u) - S(-u)}{2u}. \tag{3.4}$$

The following relation between the twisted Yangians and the corresponding classical Lie algebras plays a key role in many applications; cf. Proposition 2.2.

PROPOSITION 3.3. *The mapping*

$$\pi : s_{ij}(u) \mapsto \delta_{ij} + F_{ij}\big(u \pm \tfrac{1}{2}\big)^{-1} \tag{3.5}$$

defines an algebra epimorphism $\mathrm{Y}^{\pm}(N) \to \mathrm{U}(\mathfrak{g}_n)$. *Moreover,*

$$F_{ij} \mapsto s_{ij}^{(1)}$$

is an embedding $\mathrm{U}(\mathfrak{g}_n) \hookrightarrow \mathrm{Y}^{\pm}(N)$.

3.2. *Embedding into the Yangian*

We keep the notation $t_{ij}^{(r)}$ for the generators of the Yangian $\mathrm{Y}(N)$. However, in accordance with the above, we now let the indices i, j run over the set $\{-n, \dots, n\}$. Also, matrix transposition is now understood in the sense (3.1).

THEOREM 3.4. *The mapping*

$$S(u) \mapsto T(u)T^t(-u) \tag{3.6}$$

defines an embedding $\mathrm{Y}^{\pm}(N) \hookrightarrow \mathrm{Y}(N)$.

OUTLINE OF THE PROOF. It is straightforward to verify that the matrix $T(u)T^t(-u)$ satisfies both relations (3.3) and (3.4). To show that the homomorphism (3.6) is injective we use the corresponding homomorphism of the graded algebras $\mathrm{gr}_1\, \mathrm{Y}^{\pm}(N) \to \mathrm{gr}_1\, \mathrm{Y}(N)$, where $\mathrm{gr}_1\, \mathrm{Y}^{\pm}(N)$ is defined by setting $\deg_1 s_{ij}^{(r)} = 1$. Then apply Corollary 2.7. □

This result allows us to regard the twisted Yangian as a subalgebra of $\mathrm{Y}(N)$. The following is an analog for the algebras $\mathrm{Y}^{\pm}(N)$ of the Poincaré–Birkhoff–Witt theorem.

COROLLARY 3.5. *Given an arbitrary linear order on the set of the generators*

$$s_{ij}^{(2k)}, \quad i+j \leqslant 0; \qquad s_{ij}^{(2k-1)}, \quad i+j < 0; \quad k = 1, 2, \ldots,$$

in the case of $\mathrm{Y}^{+}(N)$, *and the set of the generators*

$$s_{ij}^{(2k)}, \quad i+j < 0; \qquad s_{ij}^{(2k-1)}, \quad i+j \leqslant 0; \quad k = 1, 2, \ldots,$$

in the case of $\mathrm{Y}^{-}(N)$, *any element of the algebra* $\mathrm{Y}^{\pm}(N)$ *is uniquely written as a linear combination of ordered monomials in the generators.*

PROPOSITION 3.6. *The subalgebra* $\mathrm{Y}^{\pm}(N)$ *is a left coideal of* $\mathrm{Y}(N)$,

$$\Delta\big(\mathrm{Y}^{\pm}(N)\big) \subseteq \mathrm{Y}(N) \otimes \mathrm{Y}^{\pm}(N).$$

PROOF. This follows from the explicit formula

$$\Delta : s_{ij}(u) \mapsto \sum_{a,b=-n}^{n} \theta_{bj}\, t_{ia}(u) t_{-j,-b}(-u) \otimes s_{ab}(u).$$

The restrictions of some of the automorphisms and anti-automorphisms of $\mathrm{Y}(N)$ described in Section 2.3 preserve the subalgebra $\mathrm{Y}^{\pm}(N)$ and so we have the following.

PROPOSITION 3.7. *The mapping*

$$S(u) \mapsto S^t(u)$$

defines an anti-automorphism of $\mathrm{Y}^{\pm}(N)$. *For any formal series* $g(u) \in 1 + u^{-2}\mathbb{C}[\![u^{-2}]\!]$ *the mapping*

$$S(u) \mapsto g(u)\; S(u) \tag{3.7}$$

defines an automorphism of $\mathrm{Y}^{\pm}(N)$.

3.3. *Sklyanin determinant*

Here we use the notation of Section 2.6 with the usual convention on the matrix element indices. Define $R(u_1, \dots, u_m)$ by (2.20) and set

$$S_i = S_i(u_i), \quad 1 \leqslant i \leqslant m, \quad \text{and}$$
$$R^t_{ij} = R^t_{ji} = R^t_{ij}(-u_i - u_j), \quad 1 \leqslant i < j \leqslant m,$$

where $S_a(u)$ and $R^t_{ab}(u)$ are defined by an obvious analogy with (2.10) and (2.11). For an arbitrary permutation $(p_1, \dots, p_m)$ of the numbers $1, \dots, m$, we abbreviate

$$\langle S_{p_1}, \dots, S_{p_m} \rangle = S_{p_1} \big(R^t_{p_1 p_2} \cdots R^t_{p_1 p_m} \big) S_{p_2} \big(R^t_{p_2 p_3} \cdots R^t_{p_2 p_m} \big) \cdots S_{p_m}.$$

PROPOSITION 3.8. *We have the identity*

$$R(u_1, \dots, u_m) \langle S_1, \dots, S_m \rangle = \langle S_m, \dots, S_1 \rangle R(u_1, \dots, u_m).$$

Now take $m = N$ and specialize the variables u_i as

$$u_i = u - i + 1, \quad i = 1, \dots, N.$$

By Propositions 2.10 and 3.8 we have

$$A_N \langle S_1, \dots, S_N \rangle = \langle S_N, \dots, S_1 \rangle A_N. \tag{3.8}$$

DEFINITION 3.9. The *Sklyanin determinant* is the formal series

$$\operatorname{sdet} S(u) = 1 + c_1 u^{-1} + c_2 u^{-2} + \cdots \in \mathrm{Y}^{\pm}(N)[[u^{-1}]]$$

such that the element (3.8) equals $A_N \operatorname{sdet} S(u)$.

In the next theorem we regard $\mathrm{Y}^{\pm}(N)$ as a subalgebra of $\mathrm{Y}(N)$.

THEOREM 3.10. *We have*

$$\operatorname{sdet} S(u) = \gamma_n(u) \operatorname{qdet} T(u) \operatorname{qdet} T(-u + N - 1),$$

where

$$\gamma_n(u) = \begin{cases} 1 & \textit{for } \mathrm{Y}^{+}(N), \\ \dfrac{2u + 1}{2u - 2n + 1} & \textit{for } \mathrm{Y}^{-}(2n). \end{cases}$$

There is an explicit formula for the Sklyanin determinant in terms of the generators $s_{ij}(u)$. It uses a special map of the symmetric groups

$$\pi_N : \mathfrak{S}_N \to \mathfrak{S}_N, \qquad p \mapsto p', \tag{3.9}$$

defined by an inductive procedure. First of all, $p'(N) = N$ so that p' can be regarded as an element of $\mathfrak{S}_{N-1}$. Given a set of positive integers $\omega_1 < \cdots < \omega_N$ we regard $\mathfrak{S}_N$ as the group of their permutations. If $N = 2$ we define π_2 as the only map $\mathfrak{S}_2 \to \mathfrak{S}_1$. For $N > 2$ define a map from the set of ordered pairs (ω_k, ω_l) with $k \neq l$ into itself by the rule

$$\begin{aligned}
(\omega_k, \omega_l) &\mapsto (\omega_l, \omega_k), \quad k, l < N, \\
(\omega_k, \omega_N) &\mapsto (\omega_{N-1}, \omega_k), \quad k < N-1, \\
(\omega_N, \omega_k) &\mapsto (\omega_k, \omega_{N-1}), \quad k < N-1, \\
(\omega_{N-1}, \omega_N) &\mapsto (\omega_{N-1}, \omega_{N-2}), \\
(\omega_N, \omega_{N-1}) &\mapsto (\omega_{N-1}, \omega_{N-2}).
\end{aligned} \tag{3.10}$$

Let $p = (p_1, \ldots, p_N)$ be a permutation of the indices $\omega_1, \ldots, \omega_N$. Define its image $p' = (q_1, \ldots, q_{N-1})$ under the map π_N as follows. First take (q_1, q_{N-1}) as the image of the ordered pair (p_1, p_N) under the map (3.10). Then define $(q_2, \ldots, q_{N-2})$ as the image of $(p_2, \ldots, p_{N-1})$ under π_{N-2} where $\mathfrak{S}_{N-2}$ is regarded as the group of permutations of the family of indices obtained from $(\omega_1, \ldots, \omega_N)$ by deleting p_1 and p_N. To describe the combinatorial properties of the map π_N introduce the signless Stirling numbers of the first kind $c(N, k)$ by

$$\sum_{k=1}^{N} c(N, k) x^k = x(x+1)\cdots(x+N-1).$$

We also need the Boolean posets B_n. The elements of B_n are subsets of $\{1, \ldots, n\}$ with the usual set inclusion as the partial ordering. Equip the symmetric group $\mathfrak{S}_N$ with the standard Bruhat order.

THEOREM 3.11. *Each fiber of the map $\pi_N : \mathfrak{S}_N \to \mathfrak{S}_N$ is an interval in $\mathfrak{S}_N$ with respect to the Bruhat order, isomorphic to the Boolean poset B_k for some k. Moreover, for any k the number of intervals isomorphic to B_k is the Stirling number $c(N-1, k)$.*

Below we denote the matrix elements of the transposed matrix $S^t(u)$ by $s^t_{ij}(u)$. For any permutation $p \in \mathfrak{S}_N$ we denote by p' its image under the map π_N.

THEOREM 3.12. *Let $(a_1, \ldots, a_N)$ be an arbitrary permutation of the set of indices $(-n, -n+1, \ldots, n)$. Then*

$$\begin{aligned}
\operatorname{sdet} S(u) = (-1)^n \gamma_n(u) & \\
\cdot \sum_{p \in \mathfrak{S}_N} & \operatorname{sgn} pp' \cdot s^t_{-a_{p(1)}, a_{p'(1)}}(-u) \cdots s^t_{-a_{p(n)}, a_{p'(n)}}(-u+n-1) \\
& \cdot s_{-a_{p(n+1)},\, a_{p'(n+1)}}(u-n) \cdots s_{-a_{p(N)},\, a_{p'(N)}}(u-N+1)
\end{aligned}$$

and also

$$\begin{aligned}\operatorname{sdet} S(u) = (-1)^n \gamma_n(u) \\ \cdot \sum_{p\in\mathfrak{S}_N} \operatorname{sgn} pp' \cdot s_{-a_{p'(1)},a_{p(1)}}(u-N+1)\cdots s_{-a_{p'(n)},a_{p(n)}}(u-N+n) \\ \cdot s^t_{-a_{p'(n+1)},a_{p(n+1)}}(-u+N-n-1)\cdots s^t_{-a_{p'(N)},a_{p(N)}}(-u).\end{aligned}$$

OUTLINE OF THE PROOF. The key idea is to use the symmetry relation (3.4) in order to eliminate the intermediate factors R^t_{ij} in the expression (3.8) which defines the Sklyanin determinant. □

EXAMPLE 3.13. In the case $N=2$ we have

$$\begin{aligned}\operatorname{sdet} S(u) &= \frac{2u+1}{2u\pm 1}\big(s_{-1,-1}(u-1)s_{-1,-1}(-u) \mp s_{-1,1}(u-1)s_{1,-1}(-u)\big) \\ &= \frac{2u+1}{2u\pm 1}\big(s_{1,1}(-u)s_{1,1}(u-1) \mp s_{1,-1}(-u)s_{-1,1}(u-1)\big).\end{aligned}$$

3.4. *The center of the twisted Yangian*

Theorem 3.10 implies the following relation between the coefficients of the Sklyanin determinant

$$\gamma_n(u) \operatorname{sdet} S(-u+N-1) = \gamma_n(-u+N-1) \operatorname{sdet} S(u).$$

In particular, the odd coefficients of $\operatorname{sdet} S(u)$ can be expressed in terms of the even ones.

THEOREM 3.14. *All the coefficients of the series* $\operatorname{sdet} S(u)$ *belong to the center of the algebra* $\mathrm{Y}^{\pm}(N)$. *Moreover, the even coefficients* $c_2, c_4, \dots$ *are algebraically independent and generate the center of* $\mathrm{Y}^{\pm}(N)$.

OUTLINE OF THE PROOF. The first assertion is immediate from Theorem 2.18 and (3.10). Alternatively, one can also prove the centrality of $\operatorname{sdet} S(u)$ by using the matrix form of the defining relations of $\mathrm{Y}^{\pm}(N)$. To prove the second assertion introduce a filtration on $\mathrm{Y}^{\pm}(N)$ by setting $\deg_2 s^{(r)}_{ij} = r-1$. To describe the corresponding graded algebra $\operatorname{gr}_2 \mathrm{Y}^{\pm}(N)$ consider the involution σ on the Lie algebra $\mathfrak{gl}_N$ defined by $\sigma : A \mapsto -A^t$. Then the fixed point subalgebra is the classical Lie algebra $\mathfrak{g}_n$. Introduce the twisted polynomial current Lie algebra $\mathfrak{gl}_N[x]^\sigma$ by (1.1). Then we have an algebra isomorphism

$$\operatorname{gr}_2 \mathrm{Y}^{\pm}(N) \cong \mathrm{U}\big(\mathfrak{gl}_N[x]^\sigma\big).$$

Identifying these algebras we find that the image of the coefficient c_{2m} in the $(2m-1)$-st component of $\operatorname{gr}_2 \mathrm{Y}^{\pm}(N)$ coincides with Zx^{2m-1} where

$$Z = E_{-n,-n} + E_{-n+1,-n+1} + \cdots + E_{n,n}.$$

To complete the proof of the theorem we use the fact that the center of the algebra $\mathrm{U}(\mathfrak{gl}_N[x]^\sigma)$ is generated by the elements Zx^{2m-1} with $m \geqslant 1$. $\square$

3.5. *The special twisted Yangian*

In the next definition we regard $\mathrm{Y}^\pm(N)$ as a subalgebra of the Yangian $\mathrm{Y}(N)$. Recall the definition of the Yangian $\mathrm{Y}(\mathfrak{sl}_N)$ from Section 2.8.

DEFINITION 3.15. The *special twisted Yangian* $\mathrm{SY}^\pm(N)$ is the subalgebra of $\mathrm{Y}^\pm(N)$ defined by

$$\mathrm{SY}^\pm(N) = \mathrm{Y}(\mathfrak{sl}_N) \cap \mathrm{Y}^\pm(N).$$

Equivalently, $\mathrm{SY}^\pm(N)$ is the subalgebra of $\mathrm{Y}^\pm(N)$ which consists of the elements stable under all automorphisms of the form (3.7). The following result is implied by Theorem 2.20.

THEOREM 3.16. *The algebra* $\mathrm{Y}^\pm(N)$ *is isomorphic to the tensor product of its center* $\mathrm{Z}^\pm(N)$ *and the subalgebra* $\mathrm{SY}^\pm(N)$,

$$\mathrm{Y}^\pm(N) = \mathrm{Z}^\pm(N) \otimes \mathrm{SY}^\pm(N).$$

In particular, the center of $\mathrm{SY}^\pm(N)$ *is trivial.*

COROLLARY 3.17. *The subalgebra* $\mathrm{SY}^\pm(N)$ *of* $\mathrm{Y}(\mathfrak{sl}_N)$ *is a left coideal.*

3.6. *The quantum Liouville formula*

Define the series $\zeta(u)$ with coefficients in $\mathrm{Y}^\pm(N)$ by

$$\zeta(u)^{-1} = \frac{1}{N}\,\mathrm{tr}\left\{\left(\frac{2u-N}{2u-N\pm 1}S^t(-u) \pm \frac{1}{2u-N\pm 1}S(-u)\right)S^{-1}(u-N)\right\}.$$

THEOREM 3.18. *We have the relation*

$$\zeta(u) = \varepsilon_n(u)\frac{\mathrm{sdet}\, S(u-1)}{\mathrm{sdet}\, S(u)},$$

where $\varepsilon_n(u) = \gamma_n(u)\,\gamma_n(u-1)^{-1}$.

OUTLINE OF THE PROOF. The quaternary relation (3.3) implies

$$QS_1^{-1}(-u)R(2u-N)S_2(u-N) = S_2(u-N)R(2u-N)S_1^{-1}(-u)Q.$$

It is deduced from the definition of $\zeta(u)$ that this expression coincides with $\zeta(u)Q$ up to a scalar function. Combining this with the matrix definition (2.35) of $z(u)$ we obtain

$$\zeta(u) = z(u)z(-u+N)^{-1}.$$

Now the claim follows from Theorems 2.28 and 3.10. □

COROLLARY 3.19. *The coefficients of the series $\zeta(u)$ generate the center of* $\mathrm{Y}^{\pm}(N)$.

3.7. *Factorization of the Sklyanin determinant*

Let $1 \leqslant m \leqslant n$. Denote by $S^{(m)}(u)$ the submatrix of $S(u)$ corresponding to the rows and columns enumerated by $-m, -m+1, \dots, m$, and by $\widetilde{S}^{(m)}(u)$ the submatrix of $S^{(m)}(u)$ obtained by removing the row and column enumerated by $-m$. Set

$$c(u) = \frac{1}{\gamma_n(u+N/2-1/2)} \operatorname{sdet} S(u+N/2-1/2).$$

Then by Theorem 3.10, $c(u)$ is an even formal series in u^{-1}, with coefficients in the center of the $\mathrm{Y}^{\pm}(N)$. We shall use the quasi-determinants introduced in (2.31).

THEOREM 3.20. *If $N = 2n$ then*

$$\begin{aligned} c(u) = {} & \big|\widetilde{S}^{(1)}(-u-1/2)\big|_{11} \cdot \big|S^{(1)}(u-1/2)\big|_{11} \cdots \\ & \cdot \big|\widetilde{S}^{(n)}(-u-n+1/2)\big|_{nn} \cdot \big|S^{(n)}(u-n+1/2)\big|_{nn}. \end{aligned}$$

If $N = 2n+1$ then

$$\begin{aligned} c(u) = {} & s_{00}(u) \cdot \big|\widetilde{S}^{(1)}(-u-1)\big|_{11} \cdot \big|S^{(1)}(u-1)\big|_{11} \cdots \\ & \cdot \big|\widetilde{S}^{(n)}(-u-n)\big|_{nn} \cdot \big|S^{(n)}(u-n)\big|_{nn}. \end{aligned}$$

Moreover, the factors on the right side of each expression are permutable.

OUTLINE OF THE PROOF. Define the *Sklyanin comatrix* $\widehat{S}(u) = (\hat{s}_{ij}(u))$ by the formula

$$\widehat{S}(u)S(u-N+1) = \operatorname{sdet} S(u). \tag{3.11}$$

Taking the nn-th entry gives

$$\operatorname{sdet} S(u) = \hat{s}_{nn}(u)\big|S(u-N+1)\big|_{nn}.$$

Then proceed by induction with the use of the formula

$$\hat{s}_{nn}(u) = \frac{2u+1}{2u\pm 1}\big|\widetilde{S}^{(n)}(-u)\big|_{nn} \operatorname{sdet} S^{(n-1)}(u-1),$$

which is deduced from the definition of $\operatorname{sdet} S(u)$. □

PROPOSITION 3.21. *The mapping*

$$S(u) \mapsto \gamma_n(u)\widehat{S}\Big(-u + \frac{N}{2} - 1\Big),$$

defines an automorphism of the algebra $\mathrm{Y}^{\pm}(N)$.

Consider the algebra $\widetilde{\mathrm{Y}}^{\pm}(N)$ which is defined exactly as the twisted Yangian $\mathrm{Y}^{\pm}(N)$ (Definition 3.1) but with the symmetry relation (3.2) dropped. We denote the generators of $\widetilde{\mathrm{Y}}^{\pm}(N)$ by the same symbols $s_{ij}^{(r)}$. Note that the definition of the Sklyanin determinant does not use the symmetry relation; see Definition 3.9. Therefore, we can define $\operatorname{sdet} S(u)$ in the same way for $\widetilde{\mathrm{Y}}^{\pm}(N)$. One can show that all coefficients of $\operatorname{sdet} S(u)$ belong to the center of this algebra; cf. Theorem 3.14. The matrix $S^*(u) = S^{-1}(-u - N/2)$ satisfies the quaternary relation (3.3) and so, its Sklyanin determinant $\operatorname{sdet} S^*(u)$ is well-defined.

Fix a nonnegative integer $M \leqslant N$ such that $N - M$ is even and put $m = [M/2]$. Set $\mathcal{A} = \{-n, \dots, -m-1, m+1, \dots, n\}$ and $\mathcal{B} = \{-m, \dots, m\}$ and use the notation of Section 2.11 for submatrices of $S(u)$.

THEOREM 3.22. *In the algebra* $\widetilde{\mathrm{Y}}^{\pm}(N)$ *we have*

$$\operatorname{sdet} S(u) \operatorname{sdet} S^*(-u + N/2 - 1)_{\mathcal{A}\mathcal{A}} = \operatorname{sdet} S(u)_{\mathcal{B}\mathcal{B}}.$$

We shall use the notation $s'_{ij}(u)$ for the matrix elements of the matrix $S^{-1}(u)$.

PROPOSITION 3.23. *In the algebra* $\widetilde{\mathrm{Y}}^{\pm}(N)$ *the matrix elements of the matrices* $S(u)_{\mathcal{A}\mathcal{A}}$ *and* $S^{-1}(v)_{\mathcal{B}\mathcal{B}}$ *commute with each other.*

3.8. *The centralizer construction*

Let $\mathfrak{g}_n \subset \mathfrak{gl}_N$ be the classical Lie algebra of type B_n, C_n or D_n, as defined in Section 3.1. Fix an integer m satisfying $0 \leqslant m \leqslant n$ if $N = 2n$, and $-1 \leqslant m \leqslant n$ if $N = 2n + 1$. Denote by $\mathfrak{g}_m(n)$ the subalgebra of $\mathfrak{g}_n$ spanned by the elements F_{ij} subject to the condition $m+1 \leqslant |i|, |j| \leqslant n$. Let $\mathrm{A}_m(n)$ denote the centralizer of $\mathfrak{g}_m(n)$ in the universal enveloping algebra $\mathrm{A}(n) = \mathrm{U}(\mathfrak{g}_n)$. In particular, $\mathrm{A}_0(n)$ (respectively, $\mathrm{A}_{-1}(n)$) is the center of $\mathrm{A}(n)$. Let $\mathrm{A}(n)^0$ denote the centralizer of F_{nn} in $\mathrm{A}(n)$ and let $\mathrm{I}(n)$ be the left ideal in $\mathrm{A}(n)$ generated by the elements F_{in}, $i = -n, \dots, n$. Then $\mathrm{I}(n)^0 = \mathrm{I}(n) \cap \mathrm{A}(n)^0$ is a two-sided ideal in $\mathrm{A}(n)^0$ and one has a vector space decomposition

$$\mathrm{A}(n)^0 = \mathrm{I}(n)^0 \oplus \mathrm{A}(n-1).$$

Therefore the projection of $\mathrm{A}(n)^0$ onto $\mathrm{A}(n-1)$ with the kernel $\mathrm{I}(n)^0$ is an algebra homomorphism. Its restriction to the subalgebra $\mathrm{A}_m(n)$ defines a filtration preserving homomorphism

$$\pi_n : \mathrm{A}_m(n) \to \mathrm{A}_m(n-1)$$

so that one can define the algebra A_m as the projective limit with respect to this sequence of homomorphisms in the category of filtered algebras.

We denote by $\mathfrak{h}_n$ the diagonal Cartan subalgebra of $\mathfrak{g}_n$, and by $\mathfrak{n}^+$ and $\mathfrak{n}^-$ the subalgebras spanned by the upper triangular and lower triangular matrices, respectively. We identify $\mathrm{U}(\mathfrak{h}_n)$ with the algebra of polynomial functions on $\mathfrak{h}_n^*$ and let λ_i denote the function which corresponds to F_{ii}. For $i = 1, \dots, n$ denote

$$\rho_{-i} = -\rho_i = \begin{cases} i-1 & \text{for } \mathfrak{g}_n = \mathfrak{o}_{2n}, \\ i-\frac{1}{2} & \text{for } \mathfrak{g}_n = \mathfrak{o}_{2n+1}, \\ i & \text{for } \mathfrak{g}_n = \mathfrak{sp}_{2n}, \end{cases} \tag{3.12}$$

and set $l_i = -l_{-i} = \lambda_i + \rho_i$. We also set $\rho_0 = 1/2$ in the case of $\mathfrak{g}_n = \mathfrak{o}_{2n+1}$. Recall [34, Section 7.4] that the image $\chi(z) \in \mathrm{U}(\mathfrak{h}_n)$ of an element z of the center of $\mathrm{A}(n)$ under the Harish-Chandra isomorphism χ is uniquely determined by the condition

$$z - \chi(z) \in \big(\mathfrak{n}^-\mathrm{A}(n) + \mathrm{A}(n)\mathfrak{n}^+\big).$$

If we identify $\mathrm{U}(\mathfrak{h}_n)$ with the algebra of polynomials in the variables $l_1, \dots, l_n$ then $\chi(z)$ belongs to the subalgebra $\mathrm{M}^*(n)$ of those polynomials $f = f(l_1, \dots, l_n)$ which are invariant under the shifted action of the Weyl group. More precisely, f must be invariant under all permutations of the variables and all transformations $l_i \mapsto \pm l_i$, where in the case of $\mathfrak{g}_n = \mathfrak{o}_{2n}$ the number of '$-$' has to be even. In the case of the minimum value of m ($m = 0$ or $m = -1$, respectively) the homomorphisms π_n are interpreted as the specialization homomorphisms $\pi_n : \mathrm{M}^*(n) \to \mathrm{M}^*(n-1)$ such that

$$\pi_n : f(\lambda_1, \dots, \lambda_n) \mapsto f(\lambda_1, \dots, \lambda_{n-1}, 0).$$

The corresponding projective limit in the category of filtered algebras is an analog of the algebra of shifted symmetric functions which is denoted by M^*. The elements of M^* are well-defined functions on the set of all sequences $\lambda = (\lambda_1, \lambda_2, \dots)$ which contain only a finite number of nonzero terms. The following families of elements of M^* are analogs of power sums, elementary symmetric functions, and complete symmetric functions:

$$p_m(\lambda) = \sum_{k=1}^{\infty} \big(l_k^{2m} - \rho_k^{2m}\big), \quad m = 1, 2, \dots,$$
$$1 + \sum_{m=1}^{\infty} e_m(\lambda) t^m = \prod_{k=1}^{\infty} \frac{1 + l_k^2 t}{1 + \rho_k^2 t},$$
$$1 + \sum_{m=1}^{\infty} h_m(\lambda) t^m = \prod_{k=1}^{\infty} \frac{1 - \rho_k^2 t}{1 - l_k^2 t}.$$

Each of the families $\{p_m\}$, $\{e_m\}$, $\{h_m\}$ can be taken as a system of algebraically independent generators of the algebra M^*.

PROPOSITION 3.24. *The algebra* A_0 *or* A_{-1} *in the case* $N = 2n$ *or* $N = 2n + 1$, *respectively, is isomorphic to the algebra of shifted symmetric functions* M^*.

Consider the homomorphism (3.5) and take its composition with the automorphism of $\mathrm{Y}^{\pm}(N)$ given by Proposition 3.21 to yield another homomorphism $\varphi_n : \mathrm{Y}^{\pm}(N) \to \mathrm{U}(\mathfrak{g}_n)$. Set $M = 2m$ or $M = 2m + 1$ depending on whether $N = 2n$ or $N = 2n + 1$. The image of the restriction of φ_n to the subalgebra $\mathrm{Y}^{\pm}(M)$ is contained in the centralizer $\mathrm{A}_m(n)$.

THEOREM 3.25. *The sequence of homomorphisms* $(\varphi_n \mid n \geqslant m)$ *defines an algebra embedding* $\varphi : \mathrm{Y}^{\pm}(M) \hookrightarrow \mathrm{A}_m$. *Moreover, one has an isomorphism*

$$\mathrm{A}_m = \mathrm{M}^* \otimes \mathrm{Y}^{\pm}(M),$$

where $\mathrm{Y}^{\pm}(M)$ *is identified with its image under the embedding* φ.

3.9. *Commutative subalgebras*

Fix an $N \times N$ matrix C with entries in $\mathbb{C}$ such that $C^t = C$ or $C^t = -C$, where the transposition is defined in (3.1). Consider the algebra $\mathrm{Y}^{\pm}(N)[\![u^{-1}]\!] \otimes (\operatorname{End} \mathbb{C}^N)^{\otimes N}$ and for any $1 \leqslant k \leqslant N$ introduce its element

$$S(u, k) = \langle S_1, \dots, S_k \rangle$$

as in Section 3.3 with the variables u_i specialized to $u_i = u - i + 1$ for $i = 1, \dots, k$. Similarly, define the element $C(u, k)$ by

$$\begin{aligned} C(u, k) = C_{k+1} \widetilde{R}^t_{k+1,k+2} \cdots \widetilde{R}^t_{k+1,N} C_{k+2} \widetilde{R}^t_{k+2,k+3} \cdots \\ \cdot \widetilde{R}^t_{k+2,N} \cdots C_{N-1} \widetilde{R}^t_{N-1,N} C_N, \end{aligned}$$

where we abbreviate $\widetilde{R}^t_{ij} = R^t_{ij}(-2u - N + i + j + 2)$. Introduce the series $\sigma_k(u, C)$ with coefficients in $\mathrm{Y}^{\pm}(N)$ by

$$\sigma_k(u, C) = \operatorname{tr} A_N S(u, k) \Bigg(\prod_{i=1,\dots,k}^{\rightarrow} \; \prod_{j=k+1,\dots,N}^{\rightarrow} R^t_{ij} \Bigg) C(u, k),$$

where $R^t_{ij} = R^t_{ij}(-2u + i + j + 2)$, A_N is the antisymmetrizer defined by (2.21) and the trace is taken over all N copies of $\operatorname{End} \mathbb{C}^N$.

THEOREM 3.26. *All the coefficients of the series* $\sigma_1(u, C), \dots, \sigma_N(u, C)$ *commute with each other. Moreover, if the matrix* C *has simple spectrum and satisfies* $C^t = -C$ *then these coefficients generate a maximal commutative subalgebra of the twisted Yangian* $\mathrm{Y}^{\pm}(N)$.

Consider the epimorphism $\pi : \mathrm{Y}^{\pm}(N) \to \mathrm{U}(\mathfrak{g}_n)$ defined in (3.5) (recall that $\mathfrak{g}_n$ denotes the classical Lie algebra $\mathfrak{o}_{2n}$, $\mathfrak{o}_{2n+1}$ or $\mathfrak{sp}_{2n}$). Clearly, the coefficients of the images of the series $\sigma_k(u, C)$, $k = 1, \ldots, N$, under π form a commutative subalgebra $\mathcal{C} \subseteq \mathrm{U}(\mathfrak{g}_n)$.

THEOREM 3.27. *If the matrix C has simple spectrum and satisfies $C^t = -C$ then the subalgebra $\mathcal{C}$ of $\mathrm{U}(\mathfrak{g}_n)$ is maximal commutative.*

Bibliographical notes

The twisted Yangians were introduced by Olshanski in [139] where he also outlined their basic properties. A detailed exposition of the most of the results presented here can be found in [119].

3.3. Theorem 3.12 is proved in [108]. Theorem 3.11 was conjectured by Lascoux and proved by the author in [111].

3.7. Quasi-determinant factorization of the Sklyanin determinant is given in [109].

3.8. The centralizer construction originates from Olshanski [139]. A detailed proof of Theorem 3.25 is given in [120].

3.9. The commutative subalgebras in the twisted Yangians and the classical enveloping algebras were constructed in Nazarov–Olshanski [128].

Finite-dimensional irreducible representations of the twisted Yangians are classified in [112] with the symplectic case done earlier in [106]. Explicit constructions of all representations of $\mathrm{Y}^{\pm}(2)$ are also given in [112]. These results were used in [113–115] to construct weight bases of Gelfand–Tsetlin type for the classical Lie algebras. Ragoucy [144] discovered a relationship between the twisted Yangians and folded $\mathcal{W}$-algebras. A family of algebras defined by a quaternary type relation (or reflection equation) is defined by Sklyanin [152]. He also constructed commutative subalgebras and some representations of these algebras. They were also studied in the physics literature in connection with the integrable models with boundary conditions and the nonlinear Schrödinger equation; see, e.g., Kulish and Sklyanin [91], Kulish, Sasaki and Schwiebert [90], Kuznetsov, Jørgensen and Christiansen [94], Liguori, Mintchev and Zhao [100], Mintchev, Ragoucy, Sorba and Zaugg [105,104]. The action of such algebras on hypergeometric functions was studied by Koornwinder and Kuznetsov [82].

4. Applications to classical Lie algebras

Here we give constructions of families of Casimir elements for the classical Lie algebras implied by the results discussed in the previous sections. All of these constructions (including some well known) are related with the quantum determinant for the Yangian $\mathrm{Y}(n)$

or the Sklyanin determinant for the twisted Yangian $\mathrm{Y}^{\pm}(N)$. We keep the notation $\mathfrak{g}_n$ for the classical Lie algebra of type B_n, C_n or D_n as in Section 3.1. For any element z of the center of the universal enveloping algebra $\mathrm{U}(\mathfrak{gl}_n)$ or $\mathrm{U}(\mathfrak{g}_n)$ we shall denote by $\chi(z)$ its Harish-Chandra image; see Sections 2.13 and 3.8, respectively. In the case of $\mathfrak{g}_n$ we keep using the parameters ρ_i defined in (3.12).

4.1. *Newton's formulas*

Consider the case of $\mathfrak{gl}_n$ first. As in Section 2.1 we denote by E the $n \times n$-matrix whose ij-th entry if E_{ij}. Denote by $C(u)$ the *Capelli determinant*

$$C(u) = \sum_{p \in \mathfrak{S}_n} \operatorname{sgn} p \cdot (u+E)_{p(1),1} \cdots (u+E-n+1)_{p(n),n}.$$

This is a polynomial in u whose coefficients belong to the center of $\mathrm{U}(\mathfrak{gl}_n)$. The image of $C(u)$ under the Harish-Chandra isomorphism is clearly given by

$$\chi : C(u) \mapsto (u+l_1) \cdots (u+l_n), \tag{4.1}$$

where $l_i = \lambda_i - i + 1$. Another family of central elements is provided by the Gelfand invariants $\operatorname{tr} E^k$; see Section 2.1.

The following can be regarded as a noncommutative analog of the classical Newton formula which relates the elementary and power sums symmetric functions; see, e.g., Macdonald [101].

THEOREM 4.1. *We have the formula*

$$1 + \sum_{k=0}^{\infty} \frac{(-1)^k \ \operatorname{tr} E^k}{(u-n+1)^{k+1}} = \frac{C(u+1)}{C(u)}.$$

PROOF. Apply the homomorphism (2.7) to the quantum Liouville formula; Section 2.10. □

COROLLARY 4.2. *The images of the Gelfand invariants under the Harish-Chandra isomorphism are given by*

$$1 + \sum_{k=0}^{\infty} \frac{(-1)^k \chi(\operatorname{tr} E^k)}{(u-n+1)^{k+1}} = \prod_{i=1}^{n} \left(1 + \frac{1}{u+l_i}\right). \tag{4.2}$$

In the case of $\mathfrak{g}_n$ we denote by F the $N \times N$-matrix whose ij-th entry is F_{ij}. Introduce the *Capelli-type determinant*

$$C(u) = (-1)^n \sum_{p \in \mathfrak{S}_N} \operatorname{sgn} pp' \cdot (u + \rho_{-n} + F)_{-a_{p(1)}, a_{p'(1)}} \cdots \\ \cdot (u + \rho_n + F)_{-a_{p(N)}, a_{p'(N)}},$$

where $(a_1, \ldots, a_N)$ is any permutation of the indices $(-n, \ldots, n)$ and p' is the image of p under the map (3.9). Using Theorem 3.12 and applying the homomorphism (3.5) we deduce that all coefficients of the polynomial $C(u)$ belong to the center of $\mathrm{U}(\mathfrak{g}_n)$. If we take $(a_1, \ldots, a_N) = (-n, \ldots, n)$ then the image of $C(u)$ under the Harish-Chandra isomorphism is easily found. For $N = 2n$ we get

$$\chi : C(u) \mapsto \prod_{i=1}^{n} \big(u^2 - l_i^2\big),$$

and for $N = 2n+1$

$$\chi : C(u) \mapsto \big(u + \tfrac{1}{2}\big) \prod_{i=1}^{n} \big(u^2 - l_i^2\big).$$

THEOREM 4.3. *If $N = 2n$ we have*

$$1 + \frac{2u+1}{2u+1 \mp 1} \sum_{k=0}^{\infty} \frac{(-1)^k \operatorname{tr} F^k}{(u+\rho_n)^{k+1}} = \frac{C(u+1)}{C(u)},$$

where the upper sign is taken in the orthogonal case and the lower sign in the symplectic case. If $N = 2n+1$ then the same formula holds with $C(u)$ replaced by

$$\overline{C}(u) = \frac{2u}{2u+1} C(u).$$

PROOF. Apply the homomorphism (3.5) to the quantum Liouville formula; Section 3.6. □

COROLLARY 4.4. *The images of the Gelfand invariants under the Harish-Chandra isomorphism are given by*

$$1 + \frac{2u+1}{2u+1 \mp 1} \sum_{k=0}^{\infty} \frac{(-1)^k \chi(\operatorname{tr} F^k)}{(u+\rho_n)^{k+1}} = \prod_{i=-n}^{n} \left(1 + \frac{1}{u+l_i}\right), \tag{4.3}$$

where the zero index is skipped in the product if $N = 2n$, while for $N = 2n+1$ one should set $l_0 = 0$.

4.2. *Cayley–Hamilton theorem*

The polynomials $C(u)$ turn out to be the noncommutative characteristic polynomials for the matrices E and F. Consider the case of $\mathfrak{gl}_n$ first.

THEOREM 4.5. *We have the identities*

$$C(-E+n-1) = 0 \quad \textit{and} \quad C(-E^t) = 0. \tag{4.4}$$

PROOF. Applying the homomorphism (2.7) to the relations (2.30) and (2.32) and multiplying by the denominators we get

$$C(u) = \widehat{C}(u)(u + E - n + 1) \quad \text{and} \quad C(u) = \widehat{C}^t(u - 1)(u + E^t),$$

where $\widehat{C}(u)$ is polynomial in u with coefficients in $\mathrm{U}(\mathfrak{gl}_n) \otimes \operatorname{End}\mathbb{C}^n$. □

Taking the images of the identities (4.4) in a highest weight representation L of $\mathfrak{gl}_n$ with the highest weight $(\lambda_1, \dots, \lambda_n)$ we derive the *characteristic identities*.

COROLLARY 4.6. *The image of the matrix E in L satisfies the identities*

$$\prod_{i=1}^{n}(E - l_i - n + 1) = 0 \quad \textit{and} \quad \prod_{i=1}^{n}\big(E^t - l_i\big) = 0,$$

where $l_i = \lambda_i - i + 1$. □

Now turn to the case of the Lie algebras $\mathfrak{g}_n$. As before, we consider the three cases simultaneously.

THEOREM 4.7. *We have the identity*

$$C(-F - \rho_n) = 0. \tag{4.5}$$

OUTLINE OF THE PROOF. Apply (3.5) to the relation (3.11) and multiply by the denominators. □

Let L be a highest weight representation of $\mathfrak{g}_n$ with the highest weight $(\lambda_1, \dots, \lambda_n)$ with respect to the basis elements $F_{11}, \dots, F_{nn}$ of the Cartan subalgebra $\mathfrak{h}_n$. The following are the *characteristic identities* for $\mathfrak{g}_n$ which are obtained by taking the image of (4.5) in L.

COROLLARY 4.8. *The image of the matrix F in L satisfies the identities*

$$\prod_{i=-n}^{n}(F - l_i + \rho_n) = 0,$$

where $l_i = \lambda_i + \rho_i$. The zero index is skipped in the product if $N = 2n$, while for $N = 2n+1$ one should set $l_0 = \frac{1}{2}$.

4.3. *Graphical constructions of Casimir elements*

For $1 \leqslant m \leqslant n$ denote by $E^{(m)}$ the $m \times m$-matrix with the entries E_{ij}, where $i, j = 1, \dots, m$. Let $\mathcal{E}^{(m)}$ denote the complete oriented graph with the vertices $\{1, \dots, m\}$, the arrow from i to j is labelled by the ij-th matrix element of the matrix $E^{(m)} - m + 1$. Then every path in the graph defines a monomial in the matrix elements in a natural way. A path from i to j is called *simple* if it does not pass through the vertices i and j except for the beginning and the end of the path. Using this graph introduce the elements $\Lambda_k^{(m)}$, $S_k^{(m)}$, $\Psi_k^{(m)}$ and $\Phi_k^{(m)}$ of the universal enveloping algebra $\mathrm{U}(\mathfrak{gl}_n)$ as follows. For $k \geqslant 1$

$(-1)^{k-1}\Lambda_k^{(m)}$ is the sum of all monomials labelling simple paths in $\mathcal{E}^{(m)}$ of length k going from m to m;

$S_k^{(m)}$ is the sum of all monomials labelling paths in $\mathcal{E}^{(m)}$ of length k going from m to m;

$\Psi_k^{(m)}$ is the sum of all monomials labelling paths in $\mathcal{E}^{(m)}$ of length k going from m to m, the coefficient of each monomial being the length of the first return to m;

$\Phi_k^{(m)}$ is the sum of all monomials labelling paths in $\mathcal{E}^{(m)}$ of length k going from m to m, the coefficient of each monomial being the ratio of k to the number of returns to m.

THEOREM 4.9. *The center of the algebra* $\mathrm{U}(\mathfrak{gl}_n)$ *is generated by the scalars and each of the following families of elements*

$$\Lambda_k = \sum_{i_1+\cdots+i_n=k} \Lambda_{i_1}^{(1)} \cdots \Lambda_{i_n}^{(n)},$$
$$S_k = \sum_{i_1+\cdots+i_n=k} S_{i_1}^{(1)} \cdots S_{i_n}^{(n)},$$
$$\Psi_k = \sum_{m=1}^{n} \Psi_k^{(m)}, \qquad \Phi_k = \sum_{m=1}^{n} \Phi_k^{(m)},$$

where $k = 1, 2, \ldots, n$. *Moreover,* $\Psi_k = \Phi_k$ *for any* k, *and the images of* Λ_k, S_k *and* Ψ_k *under the Harish-Chandra isomorphism are, respectively, the elementary, complete and power sums symmetric functions of degree* k *in the variables* $l_1, \ldots, l_n$.

PROOF. Consider the polynomial $C(u)$ introduced in Section 4.1 and set

$$\widetilde{C}(t) = t^n C\big(t^{-1}\big).$$

Applying the homomorphism (2.7) to the decomposition (2.36) we obtain a decomposition in the algebra of formal series with coefficients in $\mathrm{U}(\mathfrak{gl}_n)$,

$$\widetilde{C}(t) = \big|1 + tE^{(1)}\big|_{11} \cdots \big|1 + t(E^{(n)} - n + 1)\big|_{nn},$$

and the factors are permutable. By [46, Proposition 7.20], the elements introduced above are now interpreted as follows

$$1 + \sum_{k=1}^{\infty} \Lambda_k^{(m)} t^k = \big|1 + t(E^{(m)} - m + 1)\big|_{mm},$$

$$1+\sum_{k=1}^{\infty} S_k^{(m)} t^k = \left|1-t(E^{(m)}-m+1)\right|_{mm}^{-1},$$

$$\sum_{k=1}^{\infty} \Psi_k^{(m)} t^{k-1} = \left|1-t(E^{(m)}-m+1)\right|_{mm} \frac{d}{dt} \left|1-t(E^{(m)}-m+1)\right|_{mm}^{-1},$$

$$\sum_{k=1}^{\infty} \Phi_k^{(m)} t^{k-1} = -\frac{d}{dt} \log\bigl(\left|1-t(E^{(m)}-m+1)\right|_{mm}\bigr).$$

Due to the relations between the classical symmetric functions [101], the second part of the theorem follows from (4.1). □

Similarly, in the case of $\mathfrak{g}_n$ for any $1 \leqslant m \leqslant n$ denote by $F^{(m)}$ the matrix with the entries F_{ij}, where $i, j = -m, -m+1, \dots, m$ (the index 0 is skipped if $N = 2n$). Consider the complete oriented graph $\mathcal{F}_m$ with the vertices $\{-m, -m+1, \dots, m\}$, the arrow from i to j is labelled by the ij-th matrix element of the matrix $F^{(m)} + \rho_m$.

Introduce the elements $\Lambda_k^{(m)}$, $\widetilde{\Lambda}_k^{(m)}$, $S_k^{(m)}$, $\widetilde{S}_k^{(m)}$, $\Phi_k^{(m)}$, $\widetilde{\Phi}_k^{(m)}$ of the universal enveloping algebra $\mathrm{U}(\mathfrak{g}_n)$ as follows: for $k \geqslant 1$

$(-1)^{k-1}\Lambda_k^{(m)}$ (resp. $-\widetilde{\Lambda}_k^{(m)}$) is the sum of all monomials labelling simple paths in $\mathcal{F}^{(m)}$ (resp. simple paths that do not pass through $-m$) of length k going from m to m;

$S_k^{(m)}$ (resp. $(-1)^k\widetilde{S}_k^{(m)}$) is the sum of all monomials labelling paths in $\mathcal{F}^{(m)}$ (resp. paths that do not pass through $-m$) of length k going from m to m;

$\Phi_k^{(m)}$ (resp. $(-1)^k\widetilde{\Phi}_k^{(m)}$) is the sum of all monomials labelling paths in $\mathcal{F}^{(m)}$ (resp. paths that do not pass through $-m$) of length k going from m to m, the coefficient of each monomial being the ratio of k to the number of returns to m.

THEOREM 4.10. *Each of the following families of elements is contained in the center of the algebra* $\mathrm{U}(\mathfrak{g}_n)$:

$$\Lambda_{2k} = \sum_{i_1+\dots+i_{2n}=2k} \widetilde{\Lambda}_{i_1}^{(1)} \Lambda_{i_2}^{(1)} \cdots \widetilde{\Lambda}_{i_{2n-1}}^{(n)} \Lambda_{i_{2n}}^{(n)},$$

$$S_{2k} = \sum_{i_1+\dots+i_{2n}=2k} \widetilde{S}_{i_1}^{(1)} S_{i_2}^{(1)} \cdots \widetilde{S}_{i_{2n-1}}^{(n)} S_{i_{2n}}^{(n)},$$

$$\Phi_{2k} = \sum_{m=1}^{n} \bigl(\widetilde{\Phi}_{2k}^{(m)} + \Phi_{2k}^{(m)}\bigr),$$

where $k = 1, 2, \dots$. *Moreover, the images of* $(-1)^k\Lambda_{2k}$, S_{2k} *and* $\Phi_{2k}/2$ *under the Harish-Chandra isomorphism are, respectively, the elementary, complete and power sums symmetric functions of degree k in the variables* $l_1^2, \dots, l_n^2$.

OUTLINE OF THE PROOF. The proof is the same as for Theorem 4.9 with the use of Theorem 3.20 and the homomorphism (3.5). □

4.4. *Pfaffians and Hafnians*

Suppose first that $\mathfrak{g}_n$ is the orthogonal Lie algebra $\mathfrak{o}_{2n}$ or $\mathfrak{o}_{2n+1}$. For any $1 \leqslant k \leqslant n$ consider a subset $I \subseteq \{-n, \dots, n\}$ of cardinality $2k$ so that the elements of I are $i_1 < \cdots < i_{2k}$. The matrix $[F_{i_p,-i_q}]$ is skew-symmetric. Introduce the corresponding *Pfaffian* $\operatorname{Pf} F^I$ by

$$\operatorname{Pf} F^I = \frac{1}{2^k k!} \sum_{\sigma \in \mathfrak{S}_{2k}} \operatorname{sgn}\sigma \cdot F_{i_{\sigma(1)},-i_{\sigma(2)}} \cdots F_{i_{\sigma(2k-1)},-i_{\sigma(2k)}}.$$

Given a subset I set $I^* = \{-i_{2k}, \dots, -i_1\}$ and denote

$$c_k = (-1)^k \sum_{|I|=2k} \operatorname{Pf} F^I \operatorname{Pf} F^{I^*}, \quad k \geqslant 1,$$

and $c_0 = 1$. Consider the Capelli-type determinant $C(u)$ introduced in Section 4.1.

THEOREM 4.11. *The elements c_k belong to the center of* $\mathrm{U}(\mathfrak{g}_n)$ *and one has a decomposition*

$$C(u) = \sum_{k=0}^{n} c_k \big(u^2 - \rho_1^2\big) \cdots \big(u^2 - \rho_{n-k}^2\big)$$

if $N = 2n$, and

$$C(u) = \big(u + \tfrac{1}{2}\big) \sum_{k=0}^{n} c_k \big(u^2 - \rho_1^2\big) \cdots \big(u^2 - \rho_{n-k}^2\big)$$

if $N = 2n + 1$. Moreover, the image of c_k under the Harish-Chandra isomorphism is given by

$$\chi(c_k) = (-1)^k \sum_{i_1 < \cdots < i_k} \big(l_{i_1}^2 - \rho_{i_1}^2\big) \cdots \big(l_{i_k}^2 - \rho_{i_k - k+1}^2\big).$$

Using the definition of the Sklyanin determinant one can write a similar expression for the Capelli-type determinant in any realization of the orthogonal or symplectic Lie algebra. Consider the realization of $\mathfrak{o}_N$ corresponding to the canonical symmetric form so that the elements of $\mathfrak{o}_N$ are skew-symmetric matrices with respect to the usual transposition. For this discussion only, we use more standard numbering $1, \dots, N$ of the rows and columns of such matrices. Here F will denote the $N \times N$-matrix whose ij entry is $F_{ij} = E_{ij} - E_{ji}$. The elements c_k are now given by

$$c_k = \sum_{|I|=2k} \big(\operatorname{Pf} F^I\big)^2,$$

where $I = \{i_1, \ldots, i_{2k}\}$ is a subset of $\{1, \ldots, N\}$ and $\operatorname{Pf} F^I$ is the Pfaffian of the skew-symmetric matrix $[F_{i_p,i_q}]$. The Capelli-type determinant is given by

$$C(u) = \sum_{p \in \mathfrak{S}_N} \operatorname{sgn} pp' \cdot (u + F + \sigma_1)_{p(1),p'(1)} \cdots (u + F + \sigma_N)_{p(N),p'(N)},$$

where $\sigma_i = N/2 - i$ for $i \leqslant n$ and $\sigma_i = N/2 - i + 1$ for $i > n$. Theorem 4.11 holds in the same form with ρ_i replaced by σ_{n-i+1} for every $i = 1, \ldots, n$. Introduce also the more standard determinant

$$D(u) = \sum_{p \in \mathfrak{S}_N} \operatorname{sgn} p \cdot (u + F + m)_{p(1),1} (u + F + m - 1)_{p(2),2} \cdots \\ \cdot (u + F - m + 1)_{p(N),N}, \tag{4.6}$$

where $m = N/2$.

THEOREM 4.12. *The coefficients of the polynomial $D(u)$ are central in $\mathrm{U}(\mathfrak{o}_N)$ and the following decomposition holds*

$$D(u) = \sum_{k=0}^{n} c_k (u + m - k)(u + m - k - 1) \cdots (u - m + k + 1). \tag{4.7}$$

In particular, we get the following two analogs of the relation $(\operatorname{Pf} A)^2 = \det A$ which holds for a numerical skew-symmetric $2n \times 2n$-matrix A.

COROLLARY 4.13. *If $N = 2n$ then*

$$(\operatorname{Pf} F)^2 = C(0) = D(0).$$

Now consider the symplectic Lie algebra $\mathfrak{sp}_{2n}$ and return to our standard notation. For any $k \geqslant 1$ consider a sequence I of indices from $\{-n, \ldots, n\}$ of cardinality $2k$ so that the elements of I are $i_1 \leqslant \cdots \leqslant i_{2k}$. Set $\widetilde{F}_{ij} = \operatorname{sgn} i \cdot F_{ij}$. Then we have $\widetilde{F}_{i,-j} = \widetilde{F}_{j,-i}$. Introduce the corresponding *Hafnian* $\operatorname{Hf} F^I$ by

$$\operatorname{Hf} F^I = \frac{1}{2^k k!} \sum_{\sigma \in \mathfrak{S}_{2k}} \widetilde{F}_{i_{\sigma(1)}, -i_{\sigma(2)}} \cdots \widetilde{F}_{i_{\sigma(2k-1)}, -i_{\sigma(2k)}}.$$

For each I let $f_{\pm 1}, \ldots, f_{\pm n}$ be the multiplicities of the indices $\pm 1, \ldots, \pm n$ in I, and let

$$\operatorname{sgn} I = (-1)^{f_{-1} + \cdots + f_{-n}}.$$

Set $I^* = \{-i_{2k}, \ldots, -i_1\}$ and denote

$$d_k = (-1)^k \sum_{|I|=2k} \frac{\operatorname{sgn} I \cdot \operatorname{Hf} F^I \operatorname{Hf} F^{I^*}}{f_1! f_{-1}! \cdots f_n! f_{-n}!}.$$

Consider the Capelli-type determinant $C(u)$ introduced in Section 4.1 and introduce the series $c(u)$ by

$$c(u) = \frac{C(u)}{(u^2-\rho_1^2)\cdots(u^2-\rho_n^2)}.$$

THEOREM 4.14. *The elements d_k belong to the center of* $\mathrm{U}(\mathfrak{sp}_{2n})$ *and one has a decomposition*

$$c(u)^{-1} = 1 + \sum_{k=1}^{\infty} \frac{d_k}{(u^2-(n+1)^2)\cdots(u^2-(n+k)^2)}.$$

Moreover, the image of d_k under the Harish-Chandra isomorphism is given by

$$\chi(d_k) = \sum_{i_1 \leqslant \cdots \leqslant i_k} \big(l_{i_1}^2 - i_1^2\big)\cdots\big(l_{i_k}^2 - (i_k+k-1)^2\big).$$

Bibliographical notes

4.1. The images of the Gelfand invariants under the Harish-Chandra isomorphism were first found by Perelomov and Popov [140–142]; see also Želobenko [173]. The observation that the formulas (4.2) and (4.3) are related with the theory of Yangians is due to Cherednik [30]. A derivation of the Perelomov–Popov formulas from the Liouville formula is contained in [110]. Of course, given the Harish-Chandra images of the polynomials $C(u)$, Theorems 4.1 and 4.3 follow from Corollaries 4.2 and 4.4. Different proofs of Newton's formulas without using the Yangians are given by Itoh and Umeda [60,61,170] for the general linear and orthogonal Lie algebras.

4.2. Theorem 4.5 is due to Nazarov and Tarasov [129]. Theorem 4.7 is proved in [108]. An independent proof is given by Nazarov (unpublished) without using the map (3.9). One more proof in the orthogonal case is given by Itoh [61] with the use of a determinant of type (4.6) instead of $C(u)$. Corollaries 4.6 and 4.8 are the remarkable characteristic identities which are due to Bracken and Green [17,52]. More general identities are obtained by Gould [50].

4.3. The elements Λ_k, S_k, Ψ_k, Φ_k are the noncommutative symmetric functions associated with a matrix; see Gelfand et al. [46]. Theorem 4.9 is proved in [46, Section 7.5]. Theorem 4.10 is contained in [109]. A different version for the orthogonal Lie algebra is given in [111].

4.4. Most of these results are contained in [118]; see also [111]. The centrality of the determinant (4.6) was first proved by Howe and Umeda [57]. Its relationship with Pfaffians (formula (4.7)) was established by Itoh and Umeda [62]. The first relation in Corollary 4.13

is proved in [108]. Both Pfaffians and Hafnians are key ingredients in the analogs of the celebrated Capelli identity [19,20] for the classical Lie algebras given in [118]; see also Weyl [172], Howe [56], Howe and Umeda [57] for the role of the Capelli identity in the classical invariant theory. The polynomials $\chi(c_k)$ and $\chi(d_k)$ are respectively the elementary and complete factorial symmetric functions; see, e.g., Okounkov and Olshanski [134, 135].

Acknowledgements

The author would like to thank A.N. Kirillov, M.L. Nazarov, G.I. Olshanski and V.O. Tarasov for reading the manuscript and valuable comments.

References

[1] C. Ahn and W.M. Koo, *$gl(n|m)$ color Calogero–Sutherland models and super Yangian algebra*, Phys. Lett. B **365** (1996), 105–112.
[2] C. Ahn and S. Nam, *Yangian symmetries in the $SU(N)_1$ WZW model and the Calogero–Sutherland model*, Phys. Lett. B **378** (1996), 107–112.
[3] T. Arakawa, *Drinfeld functor and finite-dimensional representations of Yangian*, Comm. Math. Phys. **205** (1999), 1–18.
[4] D. Arnaudon, J. Avan, L. Frappat and E. Ragoucy, *Yangian and quantum universal solutions of Gervais–Neveu–Felder equations*, Preprint math.QA/0104181.
[5] D. Arnaudon, J. Avan, L. Frappat, E. Ragoucy and M. Rossi, *On the quasi-Hopf structure of deformed double Yangians*, Lett. Math. Phys. **51** (2000), 193–204.
[6] J. Avan, A. Jevicki and J. Lee, *Yangian-invariant field theory of matrix–vector models*, Nucl. Phys. B **486** (1997), 650–672.
[7] B. Basu-Mallick and A. Kundu, *Multi-parameter deformed and nonstandard $\mathrm{Y}(\mathfrak{gl}_M)$ Yangian symmetry in a novel class of spin Calogero–Sutherland models*, Nucl. Phys. B **509** (1998), 705–728.
[8] B. Basu-Mallick, P. Ramadevi and R. Jagannathannd, *Multiparametric and coloured extensions of the quantum group $GL_q(N)$ and the Yangian algebra $Y(gl_N)$ through a symmetry transformation of the Yang–Baxter equation*, Int. J. Mod. Phys. A **12** (1997), 945–962.
[9] R.J. Baxter, *Exactly Solved Models in Statistical Mechanics*, Academic Press, New York (1982).
[10] D. Bernard, *An introduction to Yangian symmetries*, Int. J. Mod. Phys. B **7** (1993), 3517–3530.
[11] D. Bernard and A. LeClair, *Quantum group symmetries and non-local currents in 2D QFT*, Commun. Math. Phys. **142** (1991), 99–138.
[12] D. Bernard, Z. Maassarani and P. Mathieu, *Logarithmic Yangians in WZW models*, Mod. Phys. Lett. A **12** (1997), 535–544.
[13] P. Bouwknegt, A.W.W. Ludwig and K. Schoutens, *Spinon bases, Yangian symmetry and fermionic representations of Virasoro characters in conformal field theory*, Phys. Lett. B **338** (1994), 448–456.
[14] P. Bouwknegt and K. Schoutens, *The $SU(n)_1$ WZW models: Spinon decomposition and Yangian structure*, Nucl. Phys. B **482** (1996), 345–372.
[15] P. Bouwknegt and K. Schoutens, *Spinon decomposition and Yangian structure of $\widehat{\mathfrak{sl}}_n$ modules*, Geometric Analysis and Lie Theory in Mathematics and Physics, Austral. Math. Soc. Lect. Ser., Vol. 11, Cambridge Univ. Press, Cambridge (1998), 105–131.
[16] S.I. Boyarchenko and S.Z. Levendorskiĭ, *On affine Yangians*, Lett. Math. Phys. **32** (1994), 269–274.
[17] A.J. Bracken and H.S. Green, *Vector operators and a polynomial identity for $SO(n)$*, J. Math. Phys. **12** (1971), 2099–2106.
[18] C. Briot and E. Ragoucy, *RTT presentation of finite $\mathcal{W}$-algebras*, Preprint math.QA/0005111.

[19] A. Capelli, *Über die Zurückführung der Cayley'schen Operation Ω auf gewöhnliche Polar-Operationen*, Math. Ann. **29** (1887), 331–338.
[20] A. Capelli, *Sur les opérations dans la théorie des formes algébriques*, Math. Ann. **37** (1890), 1–37.
[21] V. Chari, *On the fermionic formula and the Kirillov–Reshetikhin conjecture*, Preprint math.QA/0006090.
[22] V. Chari and S. Ilangovan, *On the Harish-Chandra homomorphism for infinite-dimensional Lie algebras*, J. Algebra **90** (1984), 476–494.
[23] V. Chari and A. Pressley, *Yangians and R-matrices*, L'Enseign. Math. **36** (1990), 267–302.
[24] V. Chari and A. Pressley, *Fundamental representations of Yangians and rational R-matrices*, J. Reine Angew. Math. **417** (1991), 87–128.
[25] V. Chari and A. Pressley, *A Guide to Quantum Groups*, Cambridge University Press (1994).
[26] V. Chari and A. Pressley, *Yangians: their representations and characters,* Acta Appl. Math. **44** (1996), 39–58.
[27] V. Chari and A. Pressley, *Yangians, integrable quantum systems and Dorey's rule*, Comm. Math. Phys. **181** (1996), 265–302.
[28] I.V. Cherednik, *Factorized particles on the half-line and root systems*, Theor. Math. Phys. **61** (1) (1984), 35–44.
[29] I.V. Cherednik, *A new interpretation of Gelfand–Tzetlin bases*, Duke Math. J. **54** (1987), 563–577.
[30] I.V. Cherednik, *Quantum groups as hidden symmetries of classic representation theory*, Differential Geometric Methods in Physics, A.I. Solomon, ed., World Scientific, Singapore (1989), 47–54.
[31] T. Curtright and C. Zachos, *Supersymmetry and the nonlocal Yangian deformation symmetry*, Nucl. Phys. B **402** (1993), 604–612.
[32] X.M. Ding, B.-Y. Hou and L. Zhao, *$\hbar$ (Yangian) deformation of the Miura map and Virasoro algebra*, Internat. J. Modern Phys. A **13** (1998), 1129–1144.
[33] X.M. Ding and L. Zhao, *Free boson representation of $DY_\hbar(sl_2)_k$ and the deformation of the Feigin–Fuchs*, Commun. Theor. Phys. **32** (1999), 103–108.
[34] J. Dixmier, *Algèbres Enveloppantes*, Gauthier-Villars, Paris (1974).
[35] V.G. Drinfeld, *Hopf algebras and the quantum Yang–Baxter equation*, Soviet Math. Dokl. **32** (1985), 254–258.
[36] V.G. Drinfeld, *Degenerate affine Hecke algebras and Yangians*, Funct. Anal. Appl. **20** (1986), 56–58.
[37] V.G. Drinfeld, *Quantum Groups*, Proc. Int. Congress Math., Berkeley, 1986, Amer. Math. Soc., Providence, RI (1987), 798–820.
[38] V.G. Drinfeld, *A new realization of Yangians and quantized affine algebras*, Soviet Math. Dokl. **36** (1988), 212–216.
[39] B. Enriquez and G. Felder, *A construction of Hopf algebra cocycles for the Yangian double $DY(\mathfrak{sl}_2)$*, J. Phys. A **31** (1998), 2401–2413.
[40] B. Enriquez and G. Felder, *Coinvariants for Yangian doubles and quantum Knizhnik–Zamolodchikov equations*, Internat. Math. Res. Notices (2) (1999), 81–104.
[41] P. Etingof and V. Retakh, *Quantum determinants and quasideterminants*, Asian J. Math. **3** (1999), 345–352.
[42] L.D. Faddeev, *Integrable models in $(1+1)$-dimensional quantum field theory*, Recent Advances in Field Theory and Statistical Mechanics, Les Houches Lectures 1982, North-Holland, Amsterdam (1984), 561–608.
[43] L.D. Faddeev and L.A. Takhtajan, *Spectrum and scattering of excitations in the one-dimensional isotropic Heisenberg model*, J. Soviet Math. **24** (1984), 241–267.
[44] I.M. Gelfand, *The center of an infinitesimal group ring*, Mat. Sb. (N.S.) **26** (68) (1950), 103–112 (in Russian); English transl.: I.M. Gelfand, Collected Papers, Vol. II, Springer-Verlag, Berlin (1988).
[45] I.M. Gelfand and M.L. Tsetlin, *Finite-dimensional representations of the group of unimodular matrices*, Dokl. Akad. Nauk SSSR **71** (1950), 825–828 (in Russian); English transl.: I.M. Gelfand, Collected Papers, Vol. II, Springer-Verlag, Berlin (1988).
[46] I.M. Gelfand, D. Krob, A. Lascoux, B. Leclerc, V.S. Retakh and J.-Y. Thibon, *Noncommutative symmetric functions*, Adv. Math. **112** (1995), 218–348.
[47] I.M. Gelfand and V.S. Retakh, *Determinants of matrices over noncommutative rings*, Funct. Anal. Appl. **25** (1991), 91–102.

[48] I.M. Gelfand and V.S. Retakh, *A theory of noncommutative determinants and characteristic functions of graphs*, Funct. Anal. Appl. **26** (1992), 1–20; Publ. LACIM, UQAM, Montreal **14** (1993), 1–26.
[49] V. Ginzburg and E. Vasserot, *Langlands reciprocity for affine quantum groups of type A_n*, Internat. Math. Res. Notices **3** (1993), 67–85.
[50] M.D. Gould, *Characteristic identities for semi-simple Lie algebras*, J. Austral. Math. Soc. B **26** (1985), 257–283.
[51] M.D. Gould and Y.-Z. Zhang, *On super-RS algebra and Drinfeld realization of quantum affine superalgebras*, Lett. Math. Phys. **44** (1998), 291–308.
[52] H.S. Green, *Characteristic identities for generators of $GL(n)$, $O(n)$ and $Sp(n)$*, J. Math. Phys. **12** (1971), 2106–2113.
[53] F.D.M. Haldane, Z.N.C. Ha, J.C. Talstra, D. Bernard and V. Pasquier, *Yangian symmetry of integrable quantum chains with long-range interactions and a new description of states in conformal field theory*, Phys. Rev. Lett. **69** (1992), 2021–2025.
[54] G. Hatayama, A. Kuniba, M. Okado, T. Takagi and Y. Yamada, *Remarks on fermionic formula*, Contemp. Math. **248** (1999), 243–291.
[55] T. Hauer, *Systematic proof of the existence of Yangian symmetry in chiral Gross–Neveu models*, Phys. Lett. B **417** (1998), 297–302.
[56] R. Howe, *Remarks on classical invariant theory*, Trans. Amer. Math. Soc. **313** (1989), 539–570.
[57] R. Howe and T. Umeda, *The Capelli identity, the double commutant theorem, and multiplicity-free actions*, Math. Ann. **290** (1991), 569–619.
[58] J.E. Humphreys, *Introduction to Lie algebras and Representation Theory*, Graduate Texts in Math., Vol. 9, Springer, New York (1972).
[59] K. Iohara, *Bosonic representations of Yangian double $DY_\hbar(\mathfrak{g})$ with $\mathfrak{g} = \mathfrak{gl}_N, \mathfrak{sl}_N$*, J. Phys. A **29** (1996), 4593–4621.
[60] M. Itoh, *Explicit Newton's formulas for $\mathfrak{gl}_n$*, J. Algebra **208** (1998), 687–697.
[61] M. Itoh, *Capelli elements for the orthogonal Lie algebras*, J. Lie Theory **10** (2000), 463–489.
[62] M. Itoh and T. Umeda, *On central elements in the universal enveloping algebras of the orthogonal Lie algebras*, Compos. Math., to appear.
[63] A.G. Izergin and V.E. Korepin, *A lattice model related to the nonlinear Schrödinger equation*, Sov. Phys. Dokl. **26** (1981) 653–654.
[64] M. Jimbo, *A q-difference analogue of* $\mathrm{U}(\mathfrak{g})$ *and the Yang–Baxter equation*, Lett. Math. Phys. **10** (1985), 63–69.
[65] G.-x. Ju, S.-k. Wang and K. Wu, *The algebraic structure of the $gl(n|m)$ color Calogero–Sutherland models*, J. Math. Phys. **39** (1998), 2813–2820.
[66] S.V. Kerov, A.N. Kirillov and N.Yu. Reshetikhin, *Combinatorics, the Bethe ansatz and representations of the symmetric group*, J. Sov. Math. **41** (1988), 916–924.
[67] S. Khoroshkin, D. Lebedev and S. Pakuliak, *Traces of intertwining operators for the Yangian double*, Lett. Math. Phys. **41** (1997), 31–47.
[68] S.M. Khoroshkin, A.A. Stolin and V.N. Tolstoy, *Rational solutions of Yang–Baxter equation and deformation of Yangians*, From Field Theory to Quantum Groups, World Sci. Publishing, River Edge, NJ (1996), 53–75.
[69] S.M. Khoroshkin, A.A. Stolin and V.N. Tolstoy, *Deformation of Yangian* $\mathrm{Y}(sl_2)$, Comm. Algebra **26** (1998), 1041–1055.
[70] S.M. Khoroshkin and V.N. Tolstoy, *Yangian double*, Lett. Math. Phys. **36** (1996), 373–402.
[71] A.N. Kirillov, *Combinatorial identities and completeness of states for the Heisenberg magnet*, J. Sov. Math. **30** (1985), 2298–3310.
[72] A.N. Kirillov, *Completeness of states of the generalized Heisenberg magnet*, J. Sov. Math. **36** (1987), 115–128.
[73] A.N. Kirillov, A. Kuniba and T. Nakanishi, *Skew Young diagram method in spectral decomposition of integrable lattice models*, Comm. Math. Phys. **185** (1997), 441–465.
[74] A.N. Kirillov, A. Kuniba and T. Nakanishi, *Skew Young diagram method in spectral decomposition of integrable lattice models. II. Higher levels*, Nucl. Phys. B **529** (1998), 611–638.
[75] A.N. Kirillov and N.Yu. Reshetikhin, *Yangians, Bethe ansatz and combinatorics*, Lett. Math. Phys. **12** (1986), 199–208.

[76] A.N. Kirillov and N.Yu. Reshetikhin, *The Bethe ansatz and the combinatorics of Young tableaux*, J. Soviet Math. **41** (1988), 925–955.

[77] A.N. Kirillov and N.Yu. Reshetikhin, *Representations of Yangians and multiplicities of the inclusion of the irreducible components of the tensor product of representations of simple Lie algebras*, J. Soviet Math. **52** (1990), 3156–3164.

[78] N. Kitanine, J.-M. Maillet and V. Terras, *Form factors of the XXZ Heisenberg spin-$\frac{1}{2}$ finite chain*, Nucl. Phys. B **554** (1999), 647–678.

[79] M. Kleber, *Combinatorial structure of finite-dimensional representations of Yangians: the simply-laced case*, Internat. Math. Res. Notices **7** (1997), 187–201.

[80] H. Knight, *Spectra of tensor products of finite-dimensional representations of Yangians*, J. Algebra **174** (1995), 187–196.

[81] H. Konno, *Free field representation of level-k Yangian double $\mathcal{D}Y(sl_2)_k$ and deformation of Wakimoto*, Lett. Math. Phys. **40** (1997), 321–336.

[82] T.H. Koornwinder and V.B. Kuznetsov, *Gauss hypergeometric function and quadratic R-matrix algebras*, St. Petersburg Math. J. **6** (1994), 161–184.

[83] D. Korotkin and H. Samtleben, *Yangian symmetry in integrable quantum gravity*, Nucl. Phys. B **527** (1998), 657–689.

[84] D. Krob and B. Leclerc, *Minor identities for quasi-determinants and quantum determinants*, Comm. Math. Phys. **169** (1995), 1–23.

[85] P.P. Kulish and N.Yu. Reshetikhin, *Diagonalisation of $GL(N)$ invariant transfer matrices and quantum N-wave system (Lee model)*, J. Phys. A **16** (1983), L591–L596.

[86] P.P. Kulish and N.Yu. Reshetikhin, *GL_3-invariant solutions of the Yang–Baxter equation*, J. Soviet Math. **34** (1986), 1948–1971.

[87] P.P. Kulish, N.Yu. Reshetikhin and E.K. Sklyanin, *Yang–Baxter equation and representation theory*, Lett. Math. Phys. **5** (1981), 393–403.

[88] P.P. Kulish and E.K. Sklyanin, *On the solutions of the Yang–Baxter equation*, J. Soviet Math. **19** (1982), 1596–1620.

[89] P.P. Kulish and E.K. Sklyanin, *Quantum spectral transform method: recent developments*, Integrable Quantum Field Theories, Lecture Notes in Phys., Vol. 151, Springer, Berlin–Heidelberg (1982), 61–119.

[90] P.P. Kulish, R. Sasaki and G. Schwiebert, *Constant solutions of reflection equations and quantum groups*, J. Math. Phys. **34** (1993), 286–304.

[91] P.P. Kulish and E.K. Sklyanin, *Algebraic structures related to reflection equations*, J. Phys. A **25** (1992), 5963–5975.

[92] P.P. Kulish and A.A. Stolin, *Deformed Yangians and integrable models*, Czechoslovak J. Phys. **47** (1997), 1207–1212.

[93] A. Kuniba and J. Suzuki, *Analytic Bethe ansatz for fundamental representations of Yangians*, Commun. Math. Phys. **173** (1995), 225–264.

[94] V.B. Kuznetsov, M.F. Jørgensen and P.L. Christiansen, *New boundary conditions for integrable lattices*, J. Phys. A **28** (1995), 4639–4654.

[95] A. LeClair and F. Smirnov, *Infinite quantum group symmetry of fields in massive 2D quantum field theory*, Internat. J. Modern Phys. A **7** (1992), 2997–3022.

[96] B. Leclerc, M. Nazarov and J.-Y. Thibon, *Induced representations of affine Hecke algebras and the canonical bases for quantum groups*, Preprint math.QA/0011074.

[97] S.Z. Levendorskiĭ, *On PBW bases for Yangians*, Lett. Math. Phys. **27** (1993), 37–42.

[98] S.Z. Levendorskiĭ, *On generators and defining relations of Yangians*, J. Geom. Phys. **12** (1993), 1–11.

[99] S.Z. Levendorskiĭ and A. Sudbery, *Yangian construction of the Virasoro algebra*, Lett. Math. Phys. **37** (1996), 243–247.

[100] A. Liguori, M. Mintchev and L. Zhao, *Boundary exchange algebras and scattering on the half line*, Comm. Math. Phys. **194** (1998), 569–589.

[101] I.G. Macdonald, *Symmetric Functions and Hall Polynomials*, Oxford University Press, 2nd edn. (1995).

[102] J.M. Maillet and J. Sanchez de Santos, *Drinfeld twists and algebraic Bethe ansatz*, L.D. Faddeev's Seminar on Mathematical Physics, Amer. Math. Soc. Transl. Ser. 2, Vol. 201, Amer. Math. Soc., Providence, RI (2000), 137–178.

[103] J.M. Maillet and V. Terras, *On the quantum inverse scattering problem*, Nucl. Phys. B **575** (2000), 627–644.

[104] M. Mintchev, E. Ragoucy and P. Sorba, *Spontaneous symmetry breaking in the $gl(N)$-NLS hierarchy on the half line*, Preprint hep-th/0104079.

[105] M. Mintchev, E. Ragoucy, P. Sorba and Ph. Zaugg, *Yangian symmetry in the nonlinear Schrödinger hierarchy*, J. Phys. A **32** (1999), 5885–5900.

[106] A.I. Molev, *Representations of twisted Yangians*, Lett. Math. Phys. **26** (1992), 211–218.

[107] A.I. Molev, *Gelfand–Tsetlin basis for representations of Yangians*, Lett. Math. Phys. **30** (1994), 53–60.

[108] A.I. Molev, *Sklyanin determinant, Laplace operators and characteristic identities for classical Lie algebras*, J. Math. Phys. **36** (1995), 923–943.

[109] A.I. Molev, *Noncommutative symmetric functions and Laplace operators for classical Lie algebras*, Lett. Math. Phys. **35** (1995), 135–143.

[110] A.I. Molev, *Yangians and Laplace operators for classical Lie algebras*, Confronting the Infinite, Proceedings of the Conference in Celebration of the 70th Years of H.S. Green and C.A. Hurst, World Scientific, Singapore (1995), 239–245.

[111] A.I. Molev, *Stirling partitions of the symmetric group and Laplace operators for the orthogonal Lie algebra*, Discrete Math. **180** (1998), 281–300.

[112] A.I. Molev, *Finite-dimensional irreducible representations of twisted Yangians*, J. Math. Phys. **39** (1998), 5559–5600.

[113] A.I. Molev, *A basis for representations of symplectic Lie algebras*, Comm. Math. Phys. **201** (1999), 591–618.

[114] A.I. Molev, *A weight basis for representations of even orthogonal Lie algebras*, Combinatorial Methods in Representation Theory, Adv. Studies in Pure Math., Vol. 28 (2000), 223–242.

[115] A.I. Molev, *Weight bases of Gelfand–Tsetlin type for representations of classical Lie algebras*, J. Phys. A: Math. Gen. **33** (2000), 4143–4168.

[116] A.I. Molev, *Irreducibility criterion for tensor products of Yangian evaluation modules*, Duke Math. J., to appear.

[117] A.I. Molev, *Yangians and transvector algebras*, Discrete Math., to appear.

[118] A. Molev and M. Nazarov, *Capelli identities for classical Lie algebras*, Math. Ann. **313** (1999), 315–357.

[119] A. Molev, M. Nazarov and G. Olshanski, *Yangians and classical Lie algebras*, Russian Math. Surveys **51** (2) (1996), 205–282.

[120] A. Molev and G. Olshanski, *Centralizer construction for twisted Yangians*, Selecta Math. (N.S.) **6** (2000), 269–317.

[121] S. Murakami and F. Göhmann, *Yangian symmetry and quantum inverse scattering method for the one-dimensional Hubbard model*, Phys. Lett. A **227** (1997), 216–226.

[122] H. Nakajima, *Quiver varieties and finite-dimensional representations of quantum affine algebras*, J. Amer. Math. Soc. **14** (2001), 145–238.

[123] T. Nakanishi, *Fusion, mass, and representation theory of the Yangian algebra*, Nucl. Phys. B **439** (1995), 441–460.

[124] M.L. Nazarov, *Quantum Berezinian and the classical Capelli identity*, Lett. Math. Phys. **21** (1991), 123–131.

[125] M.L. Nazarov, *Yangians of the 'strange' Lie superalgebras*, Quantum Groups, P.P. Kulish, ed., Lecture Notes in Math., Vol. 1510, Springer, Berlin–Heidelberg (1992), 90–97.

[126] M. Nazarov, *Yangians and Capelli identities*, Kirillov's Seminar on Representation Theory, G.I. Olshanski, ed., Amer. Math. Soc. Transl., Vol. 181, Amer. Math. Soc., Providence, RI (1998), 139–163.

[127] M. Nazarov, *Yangian of the queer Lie superalgebra*, Comm. Math. Phys. **208** (1999), 195–223.

[128] M. Nazarov and G. Olshanski, *Bethe subalgebras in twisted Yangians*, Comm. Math. Phys. **178** (1996), 483–506.

[129] M. Nazarov and V. Tarasov, *Yangians and Gelfand–Zetlin bases*, Publ. RIMS, Kyoto Univ. **30** (1994), 459–478.

[130] M. Nazarov and V. Tarasov, *Representations of Yangians with Gelfand–Zetlin bases*, J. Reine Angew. Math. **496** (1998), 181–212.

[131] M. Nazarov and V. Tarasov, *On irreducibility of tensor products of Yangian modules*, Internat. Math. Research Notices (1998), 125–150.

[132] M. Nazarov and V. Tarasov, *On irreducibility of tensor products of Yangian modules associated with skew Young diagrams*, Duke Math. J., to appear.

[133] A. Okounkov, *Quantum immanants and higher Capelli identities*, Transform. Groups **1** (1996), 99–126.

[134] A. Okounkov and G. Olshanski, *Shifted Schur functions*, St. Petersburg Math. J. **9** (1998), 239–300.

[135] A. Okounkov and G. Olshanski, *Shifted Schur functions II. Binomial formula for characters of classical groups and applications*, Kirillov's Seminar on Representation Theory, G.I. Olshanski, ed., Amer. Math. Soc. Transl., Vol. 181, Amer. Math. Soc., Providence, RI (1998), 245–271.

[136] G.I. Olshanski, *Extension of the algebra $U(g)$ for infinite-dimensional classical Lie algebras g, and the Yangians $Y(gl(m))$*, Soviet Math. Dokl. **36** (1988), 569–573.

[137] G.I. Olshanski, *Yangians and universal enveloping algebras*, J. Soviet Math. **47** (1989), 2466–2473.

[138] G.I. Olshanski, *Representations of infinite-dimensional classical groups, limits of enveloping algebras, and Yangians*, Topics in Representation Theory, A.A. Kirillov, ed., Advances in Soviet Math., Vol. 2, Amer. Math. Soc., Providence, RI (1991), 1–66.

[139] G. Olshanski, *Twisted Yangians and infinite-dimensional classical Lie algebras*, Quantum Groups, P.P. Kulish, ed., Lecture Notes in Math., Vol. 1510, Springer, Berlin–Heidelberg (1992), 103–120.

[140] A.M. Perelomov and V.S. Popov, *Casimir operators for $U(n)$ and $SU(n)$*, Soviet J. Nucl. Phys. **3** (1966), 676–680.

[141] A.M. Perelomov and V.S. Popov, *Casimir operators for the orthogonal and symplectic groups*, Soviet J. Nucl. Phys. **3** (1966), 819–824.

[142] A.M. Perelomov and V.S. Popov, *Casimir operators for semisimple Lie algebras*, Izv. Akad. Nauk SSSR **32** (1968), 1368–1390.

[143] H. Pfeiffer, *Factorizing twists and the universal R-matrix of the Yangian* $\mathrm{Y}(\mathfrak{sl}_2)$, J. Phys. A: Math. Gen. **33** (2000), 8929–8951.

[144] E. Ragoucy, *Twisted Yangians and folded $\mathcal{W}$-algebras*, Preprint math.QA/0012182.

[145] E. Ragoucy and P. Sorba, *Yangians and finite $\mathcal{W}$-algebras*, Quantum groups and integrable systems (Prague, 1998), Czechoslovak J. Phys. **48** (1998), 1483–1487.

[146] E. Ragoucy and P. Sorba, *Yangian realisations from finite $\mathcal{W}$-algebras*, Comm. Math. Phys. **203** (1999), 551–572.

[147] N.Yu. Reshetikhin, *Integrable models of quantum one-dimensional magnets with $O(n)$ and $Sp(2k)$-symmetry*, Theor. Math. Phys. **63** (3) (1985), 347–366.

[148] N.Yu. Reshetikhin and M.A. Semenov-Tian-Shansky, *Central extensions of quantum current groups*, Lett. Math. Phys. **19** (1990), 133–142.

[149] N.Yu. Reshetikhin, L.A. Takhtajan and L.D. Faddeev, *Quantization of Lie groups and Lie algebras*, Leningrad Math. J. **1** (1990), 193–225.

[150] K. Schoutens, *Yangian symmetry in conformal field theory*, Phys. Lett. B **331** (1994), 335–341.

[151] E.K. Sklyanin, *Quantum version of the method of inverse scattering problem*, J. Soviet. Math. **19** (1982), 1546–1596.

[152] E.K. Sklyanin, *Boundary conditions for integrable quantum systems*, J. Phys. A **21** (1988), 2375–2389.

[153] E.K. Sklyanin, *Separation of variables in the quantum integrable models related to the Yangian $Y[sl(3)]$*, J. Math. Sci. **80** (1996), 1861–1871.

[154] A. Soloviev, *Cartan–Weyl basis for Yangian double $DY(sl_3)$*, Theoret. Math. Phys. **111** (3) (1997), 731–743.

[155] A. Stolin and P. Kulish, *New rational solutions of Yang–Baxter equation and deformed Yangians*, Czechoslovak J. Phys. **47** (1997), 123–129.

[156] K. Takemura, *The Yangian symmetry in the spin Calogero model and its applications*, J. Phys. A **30** (1997), 6185–6204.

[157] V.O. Tarasov, *Structure of quantum L-operators for the R-matrix of the XXZ-model*, Theor. Math. Phys. **61** (1984), 1065–1071.

[158] V.O. Tarasov, *Irreducible monodromy matrices for the R-matrix of the XXZ-model and lattice local quantum Hamiltonians*, Theor. Math. Phys. **63** (1985), 440–454.

[159] V.O. Tarasov, *Cyclic monodromy matrices for $sl(n)$ trigonometric R-matrices*, Commun. Math. Phys. **158** (1993), 459–484.

[160] K. Takemura and D. Uglov, *The orthogonal eigenbasis and norms of eigenvectors in the spin Calogero–Sutherland model*, J. Phys. A **30** (1997), 3685–3717.

[161] L.A. Takhtajan and L.D. Faddeev, *Quantum inverse scattering method and the Heisenberg XYZ-model*, Russian Math. Surv. **34** (5) (1979), 11–68.

[162] V.N. Tolstoy, *Connection between Yangians and quantum affine algebras*, New Symmetries in the Theories of Fundamental Interactions (Karpacz, 1996), PWN, Warsaw (1997), 99–117.

[163] V.N. Tolstoy, *q-deformation of Yangian* $\mathrm{Y}(sl_2)$, Lie Theory and its Applications in Physics (Clausthal, 1995), World Sci. Publishing, River Edge, NJ (1996), 179–186.

[164] V.N. Tolstoy, *Drinfeldians*, Lie Theory and Its Applications in Physics II, Clausthal, Germany (1997).

[165] D.B. Uglov, *Symmetric functions and the Yangian decomposition of the Fock and basic modules of the affine Lie algebra* $\widehat{\mathfrak{sl}}_N$, Quantum Many-Body Problems and Representation Theory, MSJ Mem., Vol. 1, Math. Soc. Japan (1998), 183–241.

[166] D.B. Uglov, *Yangian Gelfand–Zetlin bases, $\mathfrak{gl}_N$-Jack polynomials and computation of dynamical correlation functions in the spin Calogero–Sutherland model*, Comm. Math. Phys. **191** (1998), 663–696.

[167] D.B. Uglov, *Skew Schur functions and Yangian actions on irreducible integrable modules of* $\widehat{\mathfrak{gl}}_N$, Conference on Combinatorics and Physics (Los Alamos, NM, 1998), Ann. Comb. **4** (2000), 383–400.

[168] D.B. Uglov, *Yangian actions on higher level irreducible integrable modules of affine* $\mathfrak{gl}(N)$, Preprint math.QA/9802048.

[169] D.B. Uglov and V.E. Korepin, *The Yangian symmetry of the Hubbard model*, Phys. Lett. A **190** (1994), 238–242.

[170] T. Umeda, *Newton's formula for* $\mathfrak{gl}_n$, Proc. Amer. Math. Soc. **126** (1998), 3169–3175.

[171] M. Varagnolo, *Quiver varieties and Yangians*, Lett. Math. Phys. **53** (2000), 273–283.

[172] H. Weyl, *Classical Groups, their Invariants and Representations*, Princeton Univ. Press, Princeton, NJ (1946).

[173] D.P. Želobenko, *Compact Lie Groups and Their Representations*, Transl. Math. Monographs, Vol. 40, Amer. Math. Soc., Providence, RI (1973).

[174] R.B. Zhang, *Representations of super Yangian*, J. Math. Phys. **36** (1995), 3854–3865.

[175] R.B. Zhang, *The quantum super-Yangian and Casimir operators of $U_q(gl(M|N))$*, Lett. Math. Phys. **33** (1995), 263–272.

[176] R.B. Zhang, *The $gl(M|N)$ super Yangian and its finite-dimensional representations*, Lett. Math. Phys. **37** (1996), 419–434.

[177] Y.-Z. Zhang, *Super-Yangian double and its central extension*, Phys. Lett. A **234** (1997), 20–26.

This Page Intentionally Left Blank

Lambda-Rings

Frédéric Patras

CNRS UMR 6621, Université de Nice, Parc Valrose, 06108 Nice cedex 2, France
E-mail: patras@math.unice.fr

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

This Page Intentionally Left Blank

1. Introduction

The theory of lambda-rings (λ-rings) originates in the work of Grothendieck on the theory of Chern classes [55, (1958)]. The definition of a λ-ring is obtained by a suitable axiomatization of the algebraic properties of exterior powers operations when they act, for example, on vector spaces, vector bundles or linear representations of groups. Using the λ-ring formalism, Grothendieck was able to construct universal Chern classes as maps from λ-rings (such as the topological K-groups) to naturally associated graded rings. Recall that the Riemann–Roch theorems (such as the Hirzebruch–Riemann–Roch theorem [63]) are concerned with the functorial properties of Chern classes and with the naturality of the Chern character. A striking application of the theory of λ-rings is Grothendieck's revolutionary approach to Riemann–Roch theorems, developed systematically in SGA 6 (Séminaire de géométrie algébrique du Bois Marie 1966/67, [1]).

In the present state of the art, the theory of λ-rings appears as a fundamental tool in (at least) four fields of mathematics: K-theory (topological K-groups, Riemann–Roch theorems, higher K-theory, homological algebra...); representations of groups (structure of representationrings, characteristic classes...), general algebra and algebraic combinatorics (symmetric functions, Witt vectors, generalizations and deformations of symmetric functions), geometry of convex polytopes (polytope groups, toric varieties). These various applications of the theory are closely related. For example, one can write a dictionary between the properties of the λ-ring structures appearing in the study of convex polytopes and the corresponding properties for the K-groups of toric manifolds. However, since the general theory of λ-rings may interest mathematicians of various interests and backgrounds, we have tried to give a presentation which is not oriented toward a particular kind of applications. Comments and references regarding specific applications of the theory to K-theory, group representations, etc., are included in the Notes at the end of the article.

In Sections 2–3, we give the general definitions and prove or survey the fundamental results. This includes the definitions of λ-rings, γ-rings and Ψ-rings which are formulated in modern (monadic) terms. We feel that the monadic point of view enlightens the classical definitions and provides the right framework for proving the general properties of λ-rings. Also are included: Hirzebruch polynomials, the γ-filtration and the Chern classes formalism, that is, the foundations of the "classical" theory of λ-rings.

In Sections 4–5, we give an account of developments of the theory that took place during the 1990s. This account includes the noncommutative generalizations of the λ-ring formalism with, in particular, the algebra of noncommutative symmetric functions, that we view as the algebra of "natural operations on noncommutative λ-rings". The second main development, closely related to the first one, concerns the combinatorial λ-ring structures appearing in the symmetric group algebras and in the theory of convex polytopes and toric varieties.

In the Notes, we mention other related or complementary results, survey briefly the main applications of the theory of λ-rings and give references for them, together with further indications on recent work related to the subject.

2. Lambda-rings

2.1. *(Co)free lambda-rings*

The formalism of lambda-rings encodes the algebraic properties of operations such as exterior powers or symmetric powers. Its algebraic background is the theory of symmetric functions or, equivalently, the theory of the symmetric group representations. Unless otherwise stated, rings are commutative with unit. We call R-algebras the associative algebras with unit over a ring R. We call algebras the $\mathbf{Z}$-algebras.

A polynomial in the variables $X_1, \ldots, X_n$ is *symmetric* if $P(X_{\sigma(1)}, \ldots, X_{\sigma(n)}) = P(X_1, \ldots, X_n)$ for any permutation σ in the symmetric group of order n, S_n. Recall the fundamental theorem of symmetric functions (see, e.g., [92, I,2]).

THEOREM 1. *Let e_i, $i = 1, \ldots, n$, be the i-th elementary symmetric function on $X_1, \ldots, X_n$, that is*:

$$e_i := \sum_{1 \leqslant j_1 < \cdots < j_i \leqslant n} X_{j_1} \ldots X_{j_i}.$$

Then, the e_i are algebraically independent and generate the ring $\mathbf{Sym}(n)$ of symmetric polynomials with integer coefficients in the variables $X_1, \ldots, X_n$.

Notice that $\mathbf{Sym}(n)$ is a graded ring: the component of degree k is the set of homogeneous symmetric polynomials of degree k, together with the zero polynomial. The graded ring of symmetric functions $\mathbf{Sym}$ is the inverse limit of the graded rings $\mathbf{Sym}(n)$ relative to the morphisms $\mathbf{Sym}(n+1) \to \mathbf{Sym}(n)$ that are induced by the map sending X_{n+1} to 0 and the other X_i's to themselves. In particular, the elements of $\mathbf{Sym}$ are formal infinite sums of monomials. The elementary symmetric functions in $\mathbf{Sym}$ are defined by the same formulas as above, where n is replaced by ∞. It follows from the fundamental theorem of symmetric functions that the ring $\mathbf{Sym}$ is freely generated by the elementary symmetric functions.

Let now R be any ring with identity 1. Let $\Lambda(R) := 1 + tR[\![t]\!]$ be the set of power series in t over R with constant term equal to 1. The product of power series provides $\Lambda(R)$ with the structure of an Abelian group.

The group $\Lambda(R)$ is provided with the structure of a ring as follows. Write f_i for the i-th elementary symmetric function on a second countable set of variables $Y_1, \ldots, Y_n, \ldots$. The coefficient of t^n in the product: $\prod_{i,j}(1 + X_i Y_j t)$ is a symmetric polynomial both in the variables X_i and Y_j. It can therefore be written uniquely as a polynomial in the elementary symmetric functions e_i and f_j: $P_n(e_1, \ldots, e_n; f_1, \ldots, f_n)$. The multiplication $\circ$ in $\Lambda(R)$ is defined by:

$$\left(1 + \sum_{n \geqslant 1} a_n t^n\right) \circ \left(1 + \sum_{n \geqslant 1} b_n t^n\right) := 1 + \sum_{n \geqslant 1} P_n(a_1, \ldots, a_n; b_1, \ldots, b_n) t^n.$$

PROPOSITION 1. *The product $\circ$ provides the group $\Lambda(R)$ with the structure of a ring.*

In other terms, Λ is an endofunctor of the category of rings. In fact, there exists a natural isomorphism of ring-valued functors between the (big) Witt vectors functor W and Λ, see [20], [59, 17.2], [74, p. 55] and [31]. The commutativity of $\circ$ is obvious. The unit of the product is $1+t$. Let us check for example that the product is associative. In the sequel, the proofs of properties that follow by a direct verification will be omitted. Detailed proofs may be found, for example, in [59], [74] or [7].

The associativity of the product reads:

$$\left(\left(1+\sum_{n\geqslant 1}a_nt^n\right)\circ\left(1+\sum_{n\geqslant 1}b_nt^n\right)\right)\circ\left(1+\sum_{n\geqslant 1}c_nt^n\right)$$
$$=\left(1+\sum_{n\geqslant 1}a_nt^n\right)\circ\left(\left(1+\sum_{n\geqslant 1}b_nt^n\right)\circ\left(1+\sum_{n\geqslant 1}c_nt^n\right)\right).$$

Looking at the coefficient of t^n in the formula, we have to prove an identity between certain polynomials in the a_is, the b_js and the c_ks. It is enough to prove the formula when they are algebraically independent, and we may therefore choose $a_i=e_i$, $b_j=f_j$, $c_k=g_k$, where g_k is the k-th elementary symmetric function associated to a third countable set of variables $Z_1,\dots,Z_n,\dots$. Since the generating function for the elementary symmetric functions $1+\sum_{n\geqslant 1}e_nt^n$ is $\prod_{i\geqslant 1}(1+X_it)$, the associativity follows from the identity:

$$\prod_{i,j}(1+X_iY_jt)\circ\prod_k(1+Z_kt)=\prod_{i,j,k}(1+X_iY_jZ_kt)$$
$$=\prod_i(1+X_it)\circ\prod_{j,k}(1+Y_jZ_kt).$$

Let us construct now a natural map $\lambda_R=\sum_{n\geqslant 0}\lambda^ns^n$ from $\Lambda(R)$ to $\Lambda(\Lambda(R))=1+s\Lambda(R)[\![s]\!]$. We write $P_{n,m}(e_1,\dots,e_{nm})$ for the polynomial in the elementary symmetric functions given by the coefficient of s^n in $\prod_{i_1<\dots<i_m}(1+X_{i_1}\dots X_{i_m}s)$. Then:

$$\lambda_R\left(1+\sum_{n\geqslant 1}a_nt^n\right)=1+\sum_{m\geqslant 1}\left(1+\sum_{n\geqslant 1}P_{n,m}(a_1,\dots,a_{mn})t^n\right)s^m,$$

that is: $\lambda^m(1+\sum_{n\geqslant 1}a_nt^n)=1+\sum_{n\geqslant 1}P_{n,m}(a_1,\dots,a_{mn})t^n$. We write from now on ε for the natural transformation of ring-valued functors from Λ to Id defined by: $\varepsilon_R(1+\sum_{n\geqslant 1}r_nt^n):=r_1$. That ε_R is a ring morphism follows from the identity in $\Lambda(R)$: $(1+at)\circ(1+bt)=1+abt$.

PROPOSITION 2. *The natural transformation λ is a natural transformation of ring-valued functors from Λ to $\Lambda\Lambda$. Moreover, $(\Lambda,\lambda,\varepsilon)$ is a comonad (or cotriple) on the category of rings.*

Recall that a comonad on a category C is a comonoid in the category of endofunctors of C. That is, Λ, λ and ε satisfy the following identities:

$$\Lambda(\varepsilon_R) \circ \lambda_R = Id_{\Lambda(R)} = \varepsilon_{\Lambda(R)} \circ \lambda_R,$$
$$\Lambda(\lambda_R) \circ \lambda_R = \lambda_{\Lambda(R)} \circ \lambda_R.$$

The proposition can be proved by the same kind of techniques as above. We call from now on $\Lambda(R)$ the *cofree λ-ring* over R.

2.2. *Lambda-rings*

DEFINITION 1. A lambda-ring (or λ-ring) is a coalgebra over the comonad $(\Lambda, \lambda, \varepsilon)$.

That is, a lambda-ring is a pair (A, λ_t), where A is a ring and $\lambda_t = \sum_{n \geqslant 0} \lambda^n t^n$ a ring morphism from A to $\Lambda(A) = 1 + tA[\![t]\!]$ such that:

$$\varepsilon_A \circ \lambda_t = Id_A,$$
$$\lambda_A \circ \lambda_t = \Lambda(\lambda_t) \circ \lambda_t.$$

The maps λ^n are called the lambda operations. The definition can be rephrased in terms of a series of identities involving the lambda operations.

DEFINITION 2. Equivalently, a lambda-ring is a pair (A, λ_t) where A is a ring and $\lambda_t = \sum_{n \geqslant 0} \lambda^n t^n$ is a map from A to $\Lambda(A) = 1 + tA[\![t]\!]$ such that, for all a, b in A:
(1) $\lambda^1(a) = a$;
(2) $\lambda^n(a+b) = \sum_{i \leqslant n} \lambda^i(a)\lambda^{n-i}(b)$;
(3) $\lambda_t(1) = 1 + t$;
(4) $\lambda^n(ab) = P_n(\lambda^1(a), \ldots, \lambda^n(a); \lambda^1(b), \ldots, \lambda^n(b))$;
(5) $\lambda^m(\lambda^n(a)) = P_{m,n}(\lambda^1(a), \ldots, \lambda^{mn}(a))$.

Each of these identities expresses one of the conditions that the pair (A, λ_t) has to satisfy in order to be a coalgebra over $(\Lambda, \lambda, \varepsilon)$. The first identity is just $\varepsilon_A \circ \lambda_t = Id_A$. The second, the third and the fourth express the condition that λ_t is a ring morphism. The last one is $\lambda_A \circ \lambda_t = \Lambda(\lambda_t) \circ \lambda_t$.

A ring provided with a map λ_t from A to $\Lambda(A) = 1 + tA[\![t]\!]$ satisfying conditions (1)–(2) is called a pre-lambda-ring (pre-λ-ring). In the early articles on the subject, one finds the terminology λ-ring and special λ-ring instead of, respectively, pre-λ-ring and λ-ring.

A symmetric function $P = P(e_1, e_2, \ldots)$ defines a natural operation on λ-rings given by: $a \mapsto P(\lambda^1(a), \lambda^2(a), \ldots)$.

EXAMPLES. The ring $\mathbf{Z}$ is a λ-ring for the map λ_t defined by: $\lambda_t(m) = (1+t)^m$. Topological K-groups and complex representation rings are the most classical examples of λ-rings [7,5]. The lambda operations are induced by the exterior powers of fiber bundles, resp. of representations. For example, the λ-ring structure on the K-group of a point identifies with the above λ-ring structure on **Z**. The λ-ring structure on the complex representation ring of S_3 can be described explicitly as follows: $R(S_3)$ is a free Abelian group on three generators $1, x, y$, where 1 stands for the trivial representation. The ring structure is given by the relations: $x^2 = 1$, $y^2 = 1 + x + y$, $xy = y$. The lambda operations are given by: $\lambda_t(x) = 1 + xt$, $\lambda_t(y) = 1 + xt + yt^2$. The Burnside ring of finite G-sets over a finite group G is naturally a pre-λ-ring. The lambda operation λ^n is obtained by taking the set of subsets with n elements of a G-set. In general, this pre-λ-ring is not a λ-ring, see [133].

3. Gamma-filtration, Hirzebruch polynomials and Chern classes

The introduction of λ-rings by Grothendieck was motivated by the algebraic properties of Chern classes in topology and algebraic geometry. Let us recall briefly the associated Riemann–Roch formalism that motivated the definitions and constructions in this section. In topology, geometry and representation theory, one faces often the following situation. Let C be a category (e.g., the category of nonsingular quasi-projective algebraic varieties over an algebraically closed field; this is the category Grothendieck first studied when developing his λ-ring point of view on the theory of Chern classes, see [55]). Let K and A be two contravariant functors from C to the category of rings (e.g., respectively the K-groups and the Chow ring functors). Let ρ be a natural transformation from K to A (e.g., the Chern character). Assume furthermore there are push-forward homomorphisms for K and A viewed as functors to Abelian groups. Given a map $f : X \to Y$, we write respectively $f_K : K(X) \to K(Y)$ (resp. $f_A : A(X) \to A(Y)$) for the associated morphisms. Usually these maps do not commute with the Chern character: $f_A \circ \rho_X \neq \rho_Y \circ f_K$. The Grothendieck–Riemann–Roch theorems measures this defect of commutativity; in particular, one says that Riemann–Roch holds for f if there exist a (Todd) class τ_f in $A(X)$ such that $f_A \circ (\tau_f \cdot \rho_X) = \rho_Y \circ f_K$.

We cannot give here a complete survey of the Grothendieck's algebraic theory of Chern classes. However, we stress the main ideas of the theory and their technical background, namely: Hirzebruch polynomials, the gamma-filtration and the Adams operations. The theory of Hirzebruch polynomials is a simple algebraic device to construct characteristic classes such as Todd classes. The gamma-filtration techniques are meant to provide a universal algebraic framework for constructing Chern characters. When the γ-filtration converges, the eigenspaces of the Adams operations correspond, at least over the rationals, to the components of the associated graded ring. It is one of the nicest features of the λ-ring approach to Chern classes that many topologically and geometrically far-reaching computations on these classes can be handled by means of elementary computations on the eigenvalues and the eigenvectors of Adams operations.

3.1. *Hirzebruch polynomials*

The theory of Hirzebruch polynomials belongs to the general theory of symmetric functions. Let $\phi(t)$ be a power series with integer (or rational, if the ground field is **Q**) coefficients. Let $X = \{x_1, \ldots, x_n, \ldots\}$ be an countable alphabet. Then, we may form the power series (we assume that $\phi(0) = 0$ in the first case, resp. that $\phi(0) = 1$ in the second):

$$ch_\phi(t) = \sum_{i=1}^{\infty} \phi(x_i \cdot t), \qquad td_\phi(t) = \prod_{i=1}^{\infty} \phi(x_i \cdot t).$$

The coefficient of t^n in $ch_\phi(t)$ (resp. $td_\phi(t)$) is a symmetric function $H^+_{\phi,n}$ (resp. $H^\times_{\phi,n}$) that is called the (n-th) additive (resp. multiplicative) *Hirzebruch polynomial* associated to ϕ. These definitions generalize in a obvious way to the case of two (or more) alphabets and to symmetric functions of two (or more) sets of variables. By the same process as in Section 2.1, each Hirzebruch polynomial can be rewritten as a polynomial in the elementary symmetric functions and is therefore canonically associated to a natural operation on λ-rings.

EXAMPLES. If $\phi(t) = 1 + t$, the $H^\times_{\phi,n}$ are the elementary symmetric functions. If $\phi(t) = t/(1-t)$, the $H^+_{\phi,n}$ are the power sums: $H^+_{\phi,n} = \sum_{i=1}^{\infty} x_i^n$. If $\phi(t) = 1/(1-t)$, the $H^\times_{\phi,n}$ are called the complete symmetric functions. The product law $\circ$ on the cofree λ-rings is entirely determined by the polynomial in two variables x and y: $1 + xyt = (1+xt) \circ (1+yt)$. More generally, Hirzebruch polynomials can be used to check that two (naturally defined) morphisms between λ-rings coincide (see [11, Section 1], and in particular Theorem 1.10). For example, let $a \in R$. Since the product with $(1+at)$ in $\Lambda(R)$ satisfies the identity: $\forall x \in R$, $(1+at)(1+xt) = 1 + xat$, it identifies more generally with the map: $\tilde{a}(1 + \sum_{i=1}^{\infty} x_n t^n) = 1 + \sum_{i=1}^{\infty} x_n a^n t^n$ (this property can also be checked directly using the definition of the product in $\Lambda(R)$). The most important examples of Hirzebruch polynomials for the construction of Chern classes are the *Chern character*, $ch = ch_\phi$, where $\phi(t) = \mathrm{e}^t$ and the *Todd homomorphism* $td = td_\phi$, where $\phi(t) = t/(1 - \mathrm{e}^{-t})$.

3.2. *Gamma-rings, gamma-filtration*

Let us write g for the group automorphism of $\Lambda(R)$ (resp. the automorphism of the functor Λ viewed as a group-valued functor) defined by: $g(t) := t/(1-t)$. The inverse automorphism is the map l defined by: $l(t) = t/(1+t)$. Since g is an automorphism of Λ, we may use it to define a new comonad and a new category of coalgebras (canonically isomorphic to the category of λ-rings). For example, there is a new product on $\Lambda(R)$ defined by (from now on, we use systematically the Hirzebruch polynomial formalism and define operations on $\Lambda(R)$ by the corresponding formal power series):

$$\begin{aligned}
&(1+xt)\nabla(1+yt) \\
&\quad = g\big(l(1+xt) \circ l(1+yt)\big) \\
&\quad = g\left(\frac{1+(x+1)t}{1+t} \circ \frac{1+(y+1)t}{1+t}\right) \\
&\quad = g\big(\big\{\big(1+(x+1)t\big)\cdot(1+t)^{-1}\big\} \circ \big\{\big(1+(y+1)t\big)\cdot(1+t)^{-1}\big\}\big)
\end{aligned}$$

$$= g\{(1+(x+1)(y+1)t)\cdot(1+(x+1)t)^{-1}\cdot(1+(y+1)t)^{-1}\cdot(1+t)\}$$
$$= g\left\{\frac{(1+(x+1)(y+1)t)\cdot(1+t)}{(1+(x+1)t)\cdot(1+(y+1)t)}\right\} = \frac{1+(x+y+xy)t}{(1+xt)(1+yt)}.$$

The unit for ∇ is written 1_∇; we have: $1_\nabla = g(1+t) = 1/(1-t)$. We write $\hat{\Lambda}$ for the new ring-valued functor (that is, for the group-valued functor Λ provided with the new product ∇). A new natural transformation γ from $\hat{\Lambda}$ to $\hat{\Lambda}\circ\hat{\Lambda}$ is defined accordingly:

$$1 + tR[\![t]\!] \to 1 + s(1 + tR[\![t]\!])[\![s]\!],$$
$$\gamma := (g_s \circ \Lambda(g)) \circ \lambda \circ l,$$

where g_s stands for the group automorphism of $1 + s(1+tR[\![t]\!])[\![s]\!]$ defined by $g_s(s) := s/(1-s)$. It follows from the definitions that:

PROPOSITION 3. *The map γ is a natural transformation from $\hat{\Lambda}$ to $\hat{\Lambda}\hat{\Lambda}$. Moreover, $(\hat{\Lambda}, \gamma, \varepsilon)$ is a comonad on the category of rings.*

DEFINITION 3. A gamma-ring (or γ-ring) is a coalgebra over the comonad $(\hat{\Lambda}, \gamma, \varepsilon)$. If R is a γ-ring, we write γ_t for the structural map $\gamma_t : R \to \hat{\Lambda}(R)$.

PROPOSITION 4. *There is a natural isomorphism between the categories of λ-rings and γ-rings.*

The isomorphism is given by the identities: $\gamma_t = g \circ \lambda_t = \lambda_{t/(1-t)}$, $\lambda_t = l \circ \gamma = \gamma_{t/(1+t)}$.

DEFINITION 4. Let K be a λ-ring. An augmented λ-ring over K is, by definition, a λ-ring A provided with a surjective λ-ring morphism ε from A to K called the augmentation.

Augmented γ-rings are defined accordingly. Notice that many authors prefer to call γ-rings the augmentation ideals of augmented λ-rings (or equivalently, the augmentation ideals of our augmented γ-rings), see, e.g., [7].

We assume from now on in this section that K is fixed and write augmented λ-ring for augmented λ-ring over K. Morphisms of augmented λ-rings are λ-ring morphisms that are compatible with the augmentations. An augmented λ-ring A has an augmentation ideal $I = I_A = \mathrm{Ker}(\varepsilon)$.

Recall that a decreasing filtration by subgroups F^*: $\cdots \subset \cdots \subset F^{i+1} \subset F^i \subset \cdots \subset R = F^0$ of a ring R is called multiplicative if $F^i \cdot F^j \subset F^{i+j}$. If $S_* = (S_i)_{i\geqslant 1}$ is a sequence of subsets of R, the smallest multiplicative filtration F^* of R such that $S_i \subset F^i$ is called the ring filtration of R generated by S_*.

DEFINITION 5. Let A be an augmented λ-ring and set $G_i := \mathrm{Im}(\gamma^i \circ (Id_A - \varepsilon))$, $i \geqslant 1$. The ring filtration of A generated by G_*, written $F^i_\gamma(A)$, is called the gamma-filtration (γ-filtration) of A.

We write $Gr(A)$ for the graded ring associated to the gamma-filtration: $Gr_i(A) := F_\gamma^i(A)/F_\gamma^{i+1}(A)$. If $x \in F_\gamma^i(A)$, we write $cl_i(x)$ for the class of x in $Gr_i(A)$. Algebraic Chern classes can then be constructed as follows. Let $a \in A$, then: $\gamma^i(a-\varepsilon(a)) \in F_\gamma^i(A)$. The i-th Chern class of a is then, by definition: $c^i(a) := cl_i(\gamma^i(a-\varepsilon(a)))$. See [57,11, 39] for a comparison with the classical geometrical and topological definitions of Chern classes.

3.3. *Adams operations and Psi-rings*

We write W for the endofunctor of the category of rings $W(R) := R^{\mathbf{N}^*}$[1] and use a power series notation $\sum_{i\in\mathbf{N}^*} r_i t^i$ for the elements of $W(R)$. The ring structure on $W(R)$ is defined component-wise:

$$\left(\sum_{i\in\mathbf{N}^*} r_i t^i\right) + \left(\sum_{i\in\mathbf{N}^*} s_i t^i\right) = \sum_{i\in\mathbf{N}^*} (r_i + s_i) t^i,$$

$$\left(\sum_{i\in\mathbf{N}^*} r_i t^i\right) \times \left(\sum_{i\in\mathbf{N}^*} s_i t^i\right) = \sum_{i\in\mathbf{N}^*} (r_i \cdot s_i) t^i.$$

We also define a map Ψ from $W(R)$ to $WW(R)$ (resp. ν from $W(R)$ to R) as follows. We write $\sum_{i,j\in\mathbf{N}^*} s_{i,j} t^i u^j$ for an element of $WW(R)$. Then:

$$\Psi\left(\sum_{i\in\mathbf{N}^*} r_i t^i\right) := \sum_{i,j\in\mathbf{N}^*} r_{i\cdot j} t^i u^j,$$

$$\nu\left(\sum_{i\in\mathbf{N}^*} r_i t^i\right) := r_1.$$

PROPOSITION 5. *The maps Ψ and ν are natural transformations of ring-valued functors. Moreover, (W, Ψ, ν) is a comonad on the category of rings.*

DEFINITION 6. A coalgebra over this comonad is called a Psi-ring (or Ψ-ring). Equivalently, a Ψ-ring is a ring R provided with a family $(\Psi^n)_{n\in\mathbf{N}^*}$ of ring endomorphisms of R such that $\Psi^1 = Id$ and $\Psi^n \circ \Psi^m = \Psi^{nm}$.

The equivalence between the two definitions is immediate: if A is a coalgebra over (W, Ψ, ν) and Ψ_t is the structural map from A to $W(A)$, the Ψ^is are given by $\Psi_t = \sum_{i\in\mathbf{N}^*} \Psi^i t^i$, and conversely. Any ring is canonically provided with the structure of a Ψ-ring by setting: $\Psi^k := Id$. Such a Ψ-ring, that is a Ψ-ring for which the Adams operations are equal to the identity map, is called a *binomial* Ψ-ring.

[1] Often $W(R)$ denotes the ring of big Witt vectors over R, i.e. something isomorphic to $\Lambda(R)$; that is not the usage here (editor's note).

The Ψ-rings are related to λ-rings and γ-rings as follows. The logarithmic derivative is a natural transformation of group-valued functors:

$$\begin{aligned}
&t \cdot d\log\colon \Lambda(R) \to W(R),\\
&1+\sum_{i=1}^{\infty} r_i t^i \mapsto t \cdot \sum_{i=1}^{\infty} \rho_i t^{i-1} := t \cdot \frac{\sum_{i=1}^{\infty} i \cdot r_i t^{i-1}}{1+\sum_{i=1}^{\infty} r_i t^i}
\end{aligned}$$

and, therefore, so is the map:

$$\begin{aligned}
&L := \sum_{i=1}^{\infty} L_i t^i : \Lambda(R) \to W(R),\\
&1+\sum_{i=1}^{\infty} r_i t^i \mapsto \sum_{i=1}^{\infty} (-1)^{i-1} \rho_i t^i.
\end{aligned}$$

Notice that, in particular, we have:

$$L(1+r_1 t) = \sum_{i=1}^{\infty} r_1^i t^i = \frac{r_1 t}{1-r_1 t},$$

so that, from the point of view of additive Hirzebruch polynomials and symmetric functions, L_i identifies with the power sum $\sum_{k=1}^{\infty} x_k^i$.

PROPOSITION 6. *The natural transformation L is a natural transformation of ring-valued functors.*

Since L is additive, it is enough to prove that $L(a \circ b) = L(a) \times L(b)$ when $a = 1 + xt$, $b = 1 + yt$ (use either the method in the proof of Proposition 1 in Section 2.1 or, equivalently, the theory of Hirzebruch polynomials). We then have:

$$\begin{aligned}
L(a \circ b) = L(1+xyt) &= \sum_{i=1}^{\infty} (xy)^i t^i\\
&= \left(\sum_{i=1}^{\infty} x^i t^i\right) \times \left(\sum_{i=1}^{\infty} y^i t^i\right) = L(a) \times L(b),
\end{aligned}$$

and the proposition follows.

THEOREM 2. *Let A be a λ-ring. The Adams operations on A are the maps Ψ^k defined by*:

$$\Psi_{-t}(x) := -t \cdot d\log \lambda_t(x),$$

where: $\Psi_t = \sum_{k=1}^{\infty} \Psi^k t^k$. The Adams operations are λ-ring endomorphisms of A. They provide A with the structure of a Ψ-ring. Conversely, if B is a Ψ-ring such that $\mathbf{Q} \subset B$, then B is naturally provided with the structure of a λ-ring.

The proof follows from the properties of L, which is an isomorphism if $\mathbf{Q} \subset B$. See, e.g., [7, 5], [74, pp. 47–53] or [11, Proposition 7.5] for details.

Notice that, if A is a binomial Ψ-ring and $\mathbf{Q} \subset A$, then: $\lambda_t(x) = (1+t)^x$ or, equivalently, $\lambda^n(x) = \binom{x}{n}$, where $\binom{x}{n} := \frac{x(x-1)\dots(x-n+1)}{n(n-1)\dots 1}$; in particular we have: $n \cdot \lambda^n(x) = x \cdot \lambda^{n-1}(x-1)$. Indeed, since $\Psi_t(x) = \sum_n x \cdot t^n = x \cdot t/(1-t)$, we have: $d \log \lambda_t(x) = x/(1+t)$, that is:

$$\lambda_t(x) = \mathrm{e}^{\log(1+t)^x} = (1+t)^x.$$

A λ-ring A such that $n \cdot \lambda^n(x) = x \cdot \lambda^{n-1}(x-1)$ is called a binomial λ-ring.

LEMMA 1. *Let A be an augmented λ-ring over a binomial λ-ring R, then: for any $x \in F_\gamma^n(A)$, $\Psi^k(x) - k^n x \in F_\gamma^{n+1}(A)$.*

The proof relies, as usual, on computations on symmetric polynomials (choose a large enough alphabet of algebraic independent variables, etc.). However, since one has to take into account the existence of an augmentation, the techniques have to be adapted to that case, see, e.g., [7, 4 and Proposition 5.3].

COROLLARY 1. *The Adams operation Ψ^k acts as k^n on $Gr_n(A)$.*

In particular, if $\mathbf{Q} \subset A$ and if the γ-filtration is convergent ($F^n(A) = 0$ for $n \gg 0$), then the components of the graded ring $Gr(A)$ are canonically isomorphic to the eigenspaces of Ψ^k: $Gr_k(A) \cong V(k^n)$, where we write $V(k^n)$ for the eigenspace associated to the eigenvalue k^n of Ψ^k.

Let us give, as an indication of the practical use of these constructions in topology and geometry, a classical and elementary application. See [3] for details.

PROPOSITION 7. *Let X be a 2-cell complex formed by attaching a $2n$-cell to an n-sphere, where $n \neq 1, 2, 4$ or 8. Then, the square map $H^n(X, \frac{\mathbf{Z}}{2\mathbf{Z}}) \to H^{2n}(X, \frac{\mathbf{Z}}{2\mathbf{Z}})$ is zero.*

PROOF. Since the product in cohomology is commutative in the graded sense, $2x^2 = 0$ for $x \in H^n(X, \mathbf{Z})$ if n is odd. Moreover, $H^n(X, \mathbf{Z})$ is free if $n \neq 1$ since X is the attachment of a $2n$-cell to an n-sphere, so that the square map is always zero mod 2 for n odd, $n \neq 1$, and we may assume that $n = 2m$. Then, $H^*(X, \mathbf{Z})$ is the associated graded ring of $K(X)$ by the Atiyah–Hirzebruch spectral sequence. Therefore, the reduced K-theory $\widetilde{K}(X)$ is free on two generators a and b corresponding to the generators of the free Abelian groups $H^n(X, \mathbf{Z})$ and $H^{2n}(X, \mathbf{Z})$. To prove the theorem, it is enough to show that, if $m \neq 1, 2$ or 4, then: $a^2 = 0$ mod 2. Expanding the formula defining the Ψ^ks, we get: $\Psi^2(a) = a^2 - 2\lambda^2(a)$, so that we are reduced to prove that $\Psi^2(a)$ is equal to 0 mod 2.

Besides, the (topological) Chern character $ch\colon K^*(X)\otimes \mathbf{Q}\to H^*(X,\mathbf{Q})$ is compatible with the Adams operations in the sense that, if $ch(x)=\sum a_{2m}$, where $a_{2m}\in H^{2m}(X,\mathbf{Q})$, then: $ch\circ \Psi^k(x)=\sum k^m a_{2m}$ (this is the topological version of Corollary 1 above). One deduces, since $H^*(X,\mathbf{Z})$ is torsion-free, that:

$$\Psi^k(a)=k^m a \bmod K_{4m}(X),$$

so that for some integers μ and ν:

$$\Psi^2(a)=2^m a+\mu b,$$
$$\Psi^3(a)=3^m a+\nu b.$$

Since $\Psi^2\circ\Psi^3=\Psi^3\circ\Psi^2=\Psi^6$, we have:

$$3^m(2^m a+\mu b)+\nu 2^{2m} b=2^m(3^m a+\nu b)+\mu 3^{2m} b$$

so that, in particular:

$$3^m(3^m-1)\mu=2^m(2^m-1)\nu.$$

By elementary number theory, 2^m does not divide 3^m-1 if $m\neq 1,2,4$, so that μ is even, and the proposition follows. $\square$

4. Noncommutative generalizations of λ-rings

Many identities in λ-rings such as $\lambda^n(a+b)=\sum_{i\leqslant n}\lambda^i(a)\lambda^{n-i}(b)$ rely heavily on the commutativity of the product. On the other hand, in various algebraic or topological examples, one faces noncommutative rings that behave almost like λ-rings. That is, one can define λ-operations, a γ-filtration or Adams operations on them that share many properties with the corresponding operations in the commutative case. Classical examples are algebraic K-theory and the Hochschild and cyclic homologies of commutative algebras that are commutative algebras in the graded sense. That is, there is a sign rule $xy=\pm yx$ for x and y homogeneous, the sign depending on the degrees of x and y, see, e.g., [136,37,151]. It appears that it is possible to give sense in general to a purely noncommutative theory by introducing a suitable notion of noncommutative Ψ-rings [113]. Their most remarkable property is that, except for the product formulas, they behave exactly like the usual Ψ-rings (e.g., with respect to a suitably defined γ-filtration).

Another attempt to generalize the theory in the noncommutative setting is the theory of noncommutative symmetric functions [43]. Although it was not introduced in that way, this theory can be understood as the formal theory of natural operations on noncommutative Ψ-rings in the same way that the theory of symmetric functions is, up to a canonical isomorphism, the theory of natural operations on λ-rings. We summarize below the main features of the two theories. The technical details and further results can be found respectively in [113,43]. Let A be an algebra. We still write W for the endofunctor of the category of algebras $W(A) = A^{\mathbf{N}^*}$. The algebra structure is defined componentwise as in 3.3. There are still natural transformations Ψ and ν defined by the same formulas as in Section 3.

PROPOSITION 8. *The triple* (W, Ψ, ν) *is a comonad on the category of algebras.*

DEFINITION 7. A coalgebra over this comonad is called a Ψ-algebra. Equivalently, a Ψ-algebra is an algebra A provided with a family $(\Psi^n)_{n\in\mathbf{N}^*}$ of algebra endomorphisms such that $\Psi^1 = Id$ and $\Psi^n \circ \Psi^m = \Psi^{nm}$. The morphisms Ψ^k are called the Adams operations on A.

This definition also makes sense in categories of algebras over a ground ring R: replace algebra endomorphism by R-algebra endomorphism in the definition. Notice that the binomial Ψ-ring structure on the ground ring R induces the structure of a Ψ-algebra over R on R. An augmented Ψ-algebra over R is a Ψ-algebra over R together with a surjective morphism of Ψ-algebras over R to R.

Natural operations on Ψ-algebras (resp. Ψ-rings) are, by definition, natural transformations of the forgetful functor to the category of sets. The set $Op(\Psi\text{-}alg)$ (resp. $Op(\Psi\text{-}rg)$) of natural operations is naturally a Ψ-algebra (resp. Ψ-ring) and identifies as an algebra (resp. as a ring) with the free algebra (resp. the polynomial algebra) on the set $\{\Psi^k\}_{k\in\mathbf{N}^*}$.

Over a ground ring R containing the rationals, there is a symmetrization map Sym from $Op(\Psi\text{-}rg)$ to $Op(\Psi\text{-}alg)$ defined by sending a monomial in the Ψ^ks, $\Psi^{k_1}\dots\Psi^{k_n}$, to the corresponding symmetric noncommutative monomial in the Ψ^ks: $\sum_{\sigma\in S_n}\Psi^{k_{\sigma(1)}}\dots\Psi^{k_{\sigma(n)}}$. Besides, over a ground ring containing the rationals, the γ-filtration of an augmented Ψ-ring is well-defined since, in that case, an augmented Ψ-ring is canonically an augmented λ-ring. In particular, there is a γ-filtration on the Ψ-ring $Op(\Psi\text{-}rg)$ provided with the augmentation $\varepsilon(\Psi^k) := 0$. The γ-filtration of an augmented Ψ-algebra A over R, $\mathbf{Q} \subset R$, is then, by definition, the smallest decreasing algebra filtration F^* of A such that: (1) $F^1 = I$, the augmentation ideal of A; (2) for $P_\gamma \in F^n(Op(\Psi\text{-}rg))$ and $a \in I$, $Sym(P)(a) \in F^n$. The classical theory of Adams operations and their properties with respect to the γ-filtration go then over to the noncommutative case. In particular, we have:

PROPOSITION 9. *Let A be an augmented Ψ-algebra over R, $\mathbf{Q} \subset R$. Then, for any $x \in F^n$: $\Psi^k(x) - k^n x \in F^{n+1}$. That is, Ψ^k acts as k^n on the n-th component of the graded ring associated to the γ-filtration of A.*

4.1. *Noncommutative symmetric functions*

A symmetric function is, up to a canonical ring isomorphism, nothing else but a natural operation on the category of λ-rings (see Sections 2 and 4.1). By analogy, a noncommutative symmetric function with rational coefficients can be defined as a natural operation on the category of Ψ-algebras over $\mathbf{Q}$ (in fact, some properties of the algebra of noncommutative symmetric functions hold over the integers, see, e.g., [60] for relevant informations on the subject). As an algebra, the algebra of noncommutative symmetric functions is the free algebra over the family of generators $(\Psi^k)_{k\in\mathbf{N}^*}$. We prefer to write Φ^k for these generators to keep close to the original notation of the article [43] that writes Ψ^k for another family of noncommutative symmetric functions. The Φ^ks are called the power sum noncommutative symmetric functions of the second kind. By analogy with the commutative case, one can then construct, using formal series computations, families of noncommutative symmetric functions generalizing in the noncommutative setting the usual operations on λ-rings. For example, write Φ_t for $\sum_{k\geqslant 1}\Phi^k t^k$, then:

DEFINITION 8. The elementary noncommutative symmetric functions λ^k are defined by the identity:

$$\Phi_t = t\frac{d}{dt}\log(\lambda_{-t})^{-1},$$

where: $\lambda_t := \sum_{k\geqslant 0}\lambda^k t^k$.

The theory of exterior powers, as well as the theory of symmetric functions, appeared historically as a tool for understanding the algebraic properties of matrices. The same is true regarding the theory of noncommutative symmetric functions, which has various applications to noncommutative linear algebra. Examples are, besides the study of matrices with noncommutative entries and quasi-determinants: rational power series with coefficients in a skew field, noncommutative continued fractions, etc. These applications, as well as further references on these topics, can be found in the seminal and foundational article on the theory of noncommutative symmetric functions [43].

Another fundamental property of the algebra of noncommutative symmetric functions is that it is canonically isomorphic to the descent algebra of Solomon [122,112] and dual as a Hopf algebra to the Hopf algebra of quasi-symmetric functions [94]. These various algebras have been intensively studied in the 1990s and are still studied, appearing as one of the building blocks of algebraic combinatorics and of noncommutative algebra.

5. Symmetric group algebras and operations on convex polytopes

As we have just pointed out, the theory of noncommutative symmetric functions, that is, the theory of natural operations on Ψ-algebras, is algebraically equivalent to the Solomon's theory of the descent algebra. The latter algebra contains remarkable Adams and related operations that have various applications: Ψ-algebra structures in homological algebra, structure theorems for Hopf algebras, geometry of convex polytopes, ...

5.1. *Adams operations and the descent algebra*

Following the contemporary point of view on the descent algebra ([122,112], compare to [134]), this algebra can be defined as follows. Let $X = \{x_1, \dots, x_n, \dots\}$ be a countable alphabet. The free $\mathbf{Q}$-algebra over X, $T(X)$, has naturally the structure of a connected graded Hopf algebra (see, e.g., [137] for generalities on Hopf algebras). The graded connected Hopf algebra structure is entirely defined by the condition that the x_is are primitive elements of degree 1.

Recall that, for any Hopf algebra H with product Π and coproduct Δ, the set of linear endomorphisms $End(H)$ of H is naturally provided with the structure of an algebra by the convolution product $*$. For any linear endomorphisms f and g, we have:

$$f * g := \Pi \circ (f \otimes g) \circ \Delta.$$

The unit of $*$ is the composition of the unit and of the counit of H.

In the particular case of $T(X) = \bigoplus_{i=0}^{\infty} T_i(X)$, the opposite algebras $\mathbf{Q}[S_n]^{op}$ to the symmetric group algebras $\mathbf{Q}[S_n]$ embed into $End(T(X))$. The action is the action by permutation: for any $\sigma \in S_n$:

$$\begin{aligned} &\sigma(x_{i_1} \dots x_{i_n}) := x_{i_{\sigma(1)}} \dots x_{i_{\sigma(n)}}, \\ &\sigma(x_{i_1} \dots x_{i_k}) := 0, \quad k \neq n. \end{aligned}$$

It follows from the definition of the product and of the coproduct on $T(X)$ that $S_* := \bigoplus_n \mathbf{Q}[S_n]$ embeds into $End(T(X))$ as a convolution subalgebra.

DEFINITION 9. The descent algebra $\mathcal{D}$ is the convolution subalgebra of S_* generated by the canonical projections $p_i : T(X) \to T_i(X)$.

Notice that the descent algebra inherits from S_* the property of being a graded algebra. In particular, it splits linearly as a direct sum $\mathcal{D} = \bigoplus_n \mathcal{D}_n$, where $\mathcal{D}_n := \mathcal{D} \cap \mathbf{Q}[S_n]$. We write $\widehat{\mathcal{D}}$ for the completion of D with respect to the filtration induced by the grading: $\widehat{\mathcal{D}} := \prod_n \mathcal{D}_n$.

THEOREM 3. *The descent algebra is a free associative algebra canonically isomorphic to the algebra of noncommutative symmetric functions. It is dual, as a Hopf algebra, to the Hopf algebra of quasi-symmetric functions. It is a subalgebra of* $End(T(X))$ *for the composition product* $\circ$.

One may use these properties to construct new families of λ-operations, Adams operations, etc., *inside* $\mathbf{Q}[S_n]$. Be warned that these operations are *not* the same as the operations that can be defined as the inverse images in $\mathcal{D}$ of the Φ^ks or the λ^ks by means of the isomorphism between the descent algebra and the algebra of noncommutative symmetric functions. These new families of operations were first defined (using a different approach) by Feigin and Tsygan in their 1986 article on cyclic homology. They act on the Hochschild and the cyclic complexes of commutative algebras and induce, besides a Ψ-algebra structure, a splitting of these complexes. We shall see in the next section that these operations are also closely related to the Hilbert's 3rd problem on scissors congruences and connected topics (such as the Ehrhart polynomial or the K-theory of toric manifolds).

The following results generalize to all connected graded commutative or cocommutative Hopf algebras. They can be used to describe explicitly the isomorphisms appearing in classical theorems on the structure of Hopf algebras such as the Leray or the Cartier–Milnor–Moore theorem. See [111,112,118] for details.

DEFINITION 10. The Adams operations Ψ^k in $\widehat{\mathcal{D}}$ is the k-th convolution power of the identity map I of $T(X)$:

$$\Psi^k := I^{*k} = \prod_{n \geqslant 0} \Psi_n^k,$$

where $\Psi_n^k := p_n \circ \Psi^k \in \mathcal{D}_n$ is the component of degree n of Ψ^k.

An explicit formula for Ψ_n^k can be given in terms of shuffles (recall that the shuffles are the elements of the symmetric group algebras describing the product in the shuffle algebra, see, e.g., [122]). See [37, Chapter 5], [110, Corollary 1] or [46,112] for three different approaches to the formulas involving Adams and related operations in symmetric group algebras, respectively in terms of the combinatorics of shuffles, of convex polytopes and of Hopf algebras. In fact, two distinct sign conventions are possible when defining Adams operations: compare [37, Proposition 5.3.4] and [112, Corollary 2.4].

PROPOSITION 10. *We have*: $\Psi^k \circ \Psi^l = \Psi^{kl}$ *that is, equivalently, for all* n *in* $\mathbf{N}$: $\Psi_n^k \circ \Psi_n^l = \Psi_n^{kl}$.

The identity can be checked by a direct computation involving shuffles (as in the original proof of [37]) or deduced from the general properties of cocommutative Hopf algebras.

To construct λ-operations and γ-operations in $\mathbf{Q}[S_n]$, one can use a classical trick in the theory of λ-rings (see, e.g., [136] where it is used to define operations in algebraic K-theory; the constructions in the symmetric group algebras in [37] follow the same scheme). The trick is worthwhile mentioning because of its usefulness. Assume that there is a set of operations $(\Psi^k)_{k \in \mathbf{N}^*}$ acting additively on a Abelian group, or linearly on a R-module V and such that $\Psi^1 = Id_V$, $\Psi^k \circ \Psi^l = \Psi^{kl}$. Recall that the trivial extension $\widetilde{V}$ of $\mathbf{Z}$ (resp. R) by V is the ring $\mathbf{Z} \times V$ (resp. $R \times V$) with product $\times$ defined by: $(n, v) \times (m, w) := (nm, nw + mv)$. Then, $\widetilde{V}$ is an augmented Ψ-ring if one defines the Adams operation Ψ^k on $\widetilde{V}$ by: $\Psi^k(n, v) := (n, \Psi^k(v))$. The augmentation is the obvious one, in particular V identifies with the augmentation ideal of $\widetilde{V}$. If $\mathbf{Q} \subset R$, the restrictions to V of the corresponding λ- and γ-operations are then given by the formulas: $k\lambda^k = (-1)^k \Psi^k$ and $\gamma^k = \sum_{l=1}^k \binom{l-1}{k-1} \frac{\Psi^l}{l}$.

This construction applies to $\mathbf{Q}[S_n]$ and to the Adams operations Ψ_n^k (see [37, pp. 137–140]). It provides $\mathbf{Q}[S_n]$ with a convergent γ-filtration. Since the ground field is the field of rationals, this filtration is associated to a decomposition of $\mathbf{Q}[S_n]$ into eigenspaces under the action of the Ψ^ks. The associated projection from $\mathbf{Q}[S_n]$ to the eigenspace associated to the eigenvalue k^i, $i = 1, \ldots, n$, of Ψ^k is an idempotent of the symmetric group algebra. These idempotents have appeared in the work of Reutenauer on the tensor algebra and the Poincaré–Birkhoff–Witt theorem [121,122]. The dual idempotents were introduced in the homological algebra setting by Gerstenhaber and Schack [44]. They play a central role in the theory of noncommutative symmetric functions [43,79].

5.2. λ-rings, toric manifolds and the geometry of convex polytopes

Recall Hilbert's third problem. In Euclidean plane geometry, the area of polygons can be computed through a finite process of cutting and pasting. These operations of pasting and gluing define in any Euclidean space a notion of "scissors congruences" and "polytope groups". The third problem concerns the dimension 3: can a theory of volume for polyhedra be based on scissors congruences? More generally, given two polyhedra in an Euclidean space, decide whether one of them can be cut into polyhedral pieces and the pieces glued to obtain the other. Scissors congruences have produced an abundant literature, successive accounts of which are given in [58,12,128,34]. We survey briefly in this section the developments pertaining to the theory of λ-rings and to the theory of Chern classes for toric manifolds.

The relevance of the theory of λ-rings for scissors congruences relies on an elementary observation. Let $\mathcal{P}$ be the set of polytopes in $\mathbf{R}^n$ (here, we call polytope a convex polytope, that is the convex hull of a finite set of points). We write $\mathbf{P}$ for the free Abelian group over $\mathcal{P}$: $\mathbf{P} := \mathbf{Z}[\mathcal{P}]$. Then, $\mathbf{P}$ is provided with the structure of a ring by the Minkowski product:

$$\forall (P, Q) \in \mathcal{P}^2, \quad P \times Q := \{p + q,\ p \in P,\ q \in Q\}.$$

The dilation by k, written Ψ^k, acts in a obvious way on polytopes, and this action extends linearly to $\mathbf{P}$. Moreover, Ψ^k is a ring endomorphism of $\mathbf{P}$ and we have: $\Psi^1 = Id$ and $\Psi^k \circ \Psi^l = \Psi^{kl}$. That is, the family $(\Psi^k)_{k \in \mathbf{N}^*}$ of the dilations provides $\mathbf{P}$ with the structure of a Ψ-ring. Besides, the action of dilations commutes (up to direct isometries) with the operations of pasting and gluing. The dilations induce in fact a splitting of the various polytope groups associated to scissors congruence data into eigenspaces (see any of the above mentioned textbooks on the subject, e.g., [128, Proposition 5.1]). This splitting is called the weight decomposition of polytope groups. The highest weight corresponds to the volume.

These results can be understood using the λ-operations in the symmetric group algebras that have been introduced in the previous section: see [110]. For example, the dual of the Adams operation $\Psi_n^k \in \mathbf{Q}[S_n]$ can be naturally identified with the dilation by k acting on the polytope groups associated to $\mathbf{R}^n$.

Besides, it is known that the theory of lattice convex polytopes is a suitable framework for studying line bundles over toric varieties. Recall that a toric variety (or torus embedding) is a normal variety that contains a torus as a dense open subset, together with a global action of the torus on the variety extending the natural action of the torus on itself. There is a huge and continuously increasing literature on the subject. Some classical references are: [28,107,139,38]. The basic idea is the following. Let X be a smooth, compact toric variety and let $\mathcal{E}$ be an ample and equivariant line bundle on X. The weights associated to the action of the torus on the vector space of global sections of $\mathcal{E}$ are the points of a lattice. They lie inside a convex polyhedron whose vertices belong to the lattice. In this setting, the Minkowski sum of polytopes corresponds to the tensor product of line bundles whereas dilations correspond to the K-theoretical Adams operations. This dictionary between toric geometry and convex polyhedra (see, e.g., [139] for a list of entries of the dictionary) can be extended from line bundles to the whole K-theory of the toric variety. In particular, the λ-ring structure of the K-theory of a toric manifold corresponds exactly to the λ-ring structure appearing in the associated (translational) polytope groups, as shown by Morelli [98–100].

6. Notes

The theory of λ-rings and Chern classes in λ-rings dates from 1957/1958 with Grothendieck's article *La théorie des classes de Chern* [55]. The article is an appendix to the Borel–Serre account of Grothendieck's Riemann–Roch theorem [13]. Grothendieck points out his debt to Chern and Hirzebruch. Indeed, Hirzebruch's 1956 *Neue Topologische Methoden in der algebraischen Geometrie* [63] had already emphasized some of the main formal properties that lead to a purely algebraic theory of Chern classes.

Berthelot's account of Grothendieck's algebraic theory of Chern classes in SGA6 is still one of the very best references on the subject [11]. Generalities on λ-rings are included in various textbooks dealing with characteristic classes. The article by Atiyah and Tall [7] should certainly be quoted, since it contains a very synthetic account of the main properties of λ-rings. Knutson's Lecture Notes [74] develop systematically the links between the general algebraic properties of λ-rings and the representation theory of the symmetric group. The very close connexion between the two theories had been studied by Atiyah already in 1966 in a famous and seminal article on Adams and Steenrod operations in topology, *Power operations in K-theory* [4]. Another book dealing with the general properties of λ-rings is Hoffman's [65], also from the symmetric group representations point of view. The connexions between the theories of λ-rings and Witt vectors appear in [74]; further investigations are contained, e.g., in [59] or [31].

There is a huge literature about the Riemann–Roch theorems in algebraic geometry and many variants of the theorem. We quote, for example, Riemann–Roch for singular varieties (Baum–Fulton–MacPherson) [8] and the Gillet–Soulé arithmetic Riemann–Roch theorem [48]. Some general reference on the subject are, besides SGA6 *Théorie des intersections et théorème de Riemann–Roch*, Manin's *Lectures on the K-functor in algebraic geometry* [95] and the more recent and up-to-date *Riemann–Roch algebra* [39] by Fulton and Lang.

After its introduction, it was soon realized that the λ-ring formalism applies to the study of representation rings of groups and connected topics. Grothendieck's *Classes de Chern et représentations linéaires des groupes discrets* [56] and the article *Group representations, λ-rings and the J-homomorphism* by Atiyah and Tall [7] were both very influential. A general reference on the subject is Thomas' *Characteristic classes and the cohomology of finite groups* [143].

The λ-ring structures appearing in higher K-theory and related theories (such as Hochschild homology) have been intensively studied. The first definitions of λ-ring structures in algebraic K-theory appeared independently in Kratzer's [78] and Hiller's [61]. The definition was extended to the K-theory of schemes by Soulé [136]. Other approaches and further extensions appear in [52–54,49,87] or [88]. A general reference on the subject is Levine's *Mixed Motives* [89, Chapter 3]. The definition of a λ-ring structure on the cyclic homology of a commutative algebra is due to Feigin and Tsygan [37, Chapter 5], who also defined on that occasion λ-, γ- and Ψ-operations in the symmetric group algebras. Gerstenhaber and Schack introduced independently operations on the Hochschild cohomology of commutative algebras [44]. They discovered for example the Feigin–Tsygan's second Adams operation Ψ^2 (up to a constant) and investigated its algebraic properties (eigenvalues...), but did not notice that their results could be interpreted in terms of λ-ring operations. Their results also show that there is a deep connexion between λ-ring structures in homological algebra and Hodge structures [45]. Improvements and complements to these results are contained in [90,102,105,72,41,125]. Gerstenhaber and Schack also showed that the operation on the Hochschild (co)homology of commutative algebras can be deduced from the properties of the shuffle algebra viewed as a Hopf algebra [46]. For commutative differential graded algebras, an equivalence between the weight decompositions of Hochschild and cyclic homology associated to the γ-filtration and the decompositions obtained independently by Burghelea and Vigué [18] is established in [148,17]. There are generalized trace maps from algebraic K-theory to the homology theories for commutative algebras (Dennis, Karoubi). The compatibility of the λ-ring structures with these trace maps was investigated by Cathelineau, Weibel, Kantorowitz et al. [22,150,71]. A geometric interpretation of the Adams operations on the Hochschild homology of commutative algebras in terms of simplicial dilations (resp. simplicial subdivisions) was obtained in [110] (resp. [96]). The iterated bar constructions and the associated homology theories (the iterated Hochschild homology theories for commutative algebras) are natural generalizations of the bar construction. They compute, for example, the cohomology of iterated loop spaces. It is shown in [117] that they carry a natural Ψ-algebra structure.

Hilbert's third problem is concerned with the structure of polytope groups, which are the groups defined as the quotient of the free Abelian groups over the set of convex polytopes by the cutting and pasting relations. A group of isometries is underlying the pasting relations. The problem was first formulated for the Euclidean 3-space, but has been later on generalized to higher dimensions and other (i.e. non-Euclidean) geometries. Some classical or recent textbooks on the subject are [58,12,128,34]. A closely related problem that has been intensively studied recently is the problem of counting lattice points in convex polytopes. The dilations are the key ingredient of the theory in the Euclidean case, as clearly emphasized in Hadwiger's *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie*. In [110], it is shown that dilations, when they act on polytope groups, identify naturally with elements of the symmetric group algebras: in fact, they are equal to the Adams operations previously defined by Feigin and Tsygan [37]. The λ-ring point of view on polytope groups has been developed systematically by R. Morelli in a series of articles. He has also shown that the classical dictionary between convex polytopes and global sections of invertible sheaves over toric manifolds [107,28,139] extends to polytope groups and K-theory [98–100]. The λ-ring structure of polytopes groups can also be used to investigate the formal algebraic properties of iterated integrals in the sense of Chen. These integrals are defined as integrals of differential forms over simplices. The study of the action of the geometric Adams operations (i.e. of the dilations) on the simplices provides useful algebraic informations on the iterated integrals themselves and on the corresponding (usual or equivariant) free loop spaces cohomology groups: see [114,115].

The notion of Ψ-algebra, that is, the notion of a "noncommutative Ψ-ring", was introduced and investigated in [113]. More generally, in algebraic topology, many products on chain or cochain complexes are often associative and/or commutative only up to homotopy. One may wonder how to give a meaning to λ-ring structures and characteristic classes in this setting. A first attempt in this direction is contained in [116], where "Ψ-algebras up to homotopy" structures are introduced in order to study the periodic and negative cyclic homology theories. The noncommutative symmetric functions were introduced independently in [79]. Their properties were investigated in a series of articles [32,80,81,142]. The fact that the theory of noncommutative symmetric functions is equivalent to the more classical Solomon's theory of the descent algebra is already in [79]. A general reference on the descent algebra and its role in algebra and algebraic combinatorics is Reutenauer's *Free Lie algebras*.

The idea of investigating the power maps of the identity in the convolution algebra of the tensor algebra over a countable alphabet appears in [121] in relation to the Poincaré–Birkhoff–Witt theorem. Gerstenhaber and Schack noticed that the dual maps identify with the Adams operations appearing in cyclic and Hochschild homology and that the fundamental Ψ-ring identity: $\Psi^k \circ \Psi^l = \Psi^{kl}$ for these operations can be deduced formally from the Hopf algebraic properties of the shuffle Hopf algebra [46]. It is shown in [111,112,118] that the Ψ-ring-type properties of (graded connected, over a field of characteristic 0 and commutative) Hopf algebras lead to new techniques in the field and to new proofs of classical structure theorems for Hopf algebras such as the Cartier–Milnor–Moore or the Leray theorem.

References

[1] P. Berthelot, A. Grothendieck and L. Illusie, eds, *Théorie des intersections et théorème de Riemann–Roch*, SGA 6: Séminaire de géométrie algébrique du Bois Marie 1966/67, Lecture Notes in Math., Vol. 225, Springer, Berlin–Heidelberg–New York (1971).

[2] J.F. Adams, *The Selected Works of J. Frank Adams*, Vols. 1 and 2, Cambridge University Press, Cambridge (1992).

[3] J.F. Adams and M.F. Atiyah, *K-theory and the Hopf invariant*, Quart. J. Math. Oxford Ser. (2) **17** (1966), 31–38.

[4] M.F. Atiyah, *Power operations in K-theory*, Quart. J. Math. Oxford Ser. (2) **17** (1966), 165–193.

[5] M.F. Atiyah, *K-Theory*, Benjamin, New York (1967).

[6] M. Atiyah, *Collected Works, Vol. 2: K-Theory*, Clarendon Press, Oxford Science Publications, Oxford (1988).

[7] M.F. Atiyah and D.O. Tall, *Group representations, λ-rings and the J-homomorphism*, Topology **8** (1969), 253–297.

[8] P. Baum, W. Fulton and R. MacPherson, *Riemann–Roch for singular varieties*, Publ. Math., Inst. Hautes Étud. Sci. **45** (1975), 101–145.

[9] D.J. Benson, *Lambda and psi operations on Green rings*, J. Algebra **87** (2) (1984), 360–367.

[10] D. Benson, *Modular Representation Theory: New Trends and Methods*, Lecture Notes in Math., Vol. 1081, Springer, Berlin (1984).

[11] P. Berthelot, *Généralités sur les λ-anneaux*, Théorie des intersections et théorème de Riemann–Roch, Séminaire de Géométrie Algébrique du Bois-Marie 1966–1967 (SGA 6), Lecture Notes in Math., Vol. 225, Springer, Berlin–New York (1971), 297–364.

[12] V.G. Boltyanskii, *Hilbert's Third Problem*, Scripta Series in Mathematics, John Wiley & Sons, New York (1978).

[13] A. Borel and J.-P. Serre, *La théorie des classes de Chern*, Bull. Soc. Math. France **86** (1958), 137–154.
[14] A.K. Bousfield, *On λ-rings and the K-theory of infinite loop spaces*, K-Theory **10** (1) (1996), 1–30.
[15] A.K. Bousfield, *On p-adic λ-rings and the K-theory of H-spaces*, Math. Z. **223** (3) (1996), 483–519.
[16] J.M. Bryden, *Exterior power operations in the representation theory of the classical Weyl groups*, Comm. Algebra **27** (12) (1999), 6273–6296.
[17] D. Burghelea, Z. Fiedorowicz and W. Gajda, *Adams operations in Hochschild and cyclic homology of de Rham algebra and free loop spaces*, K-Theory **4** (3) (1991), 269–287.
[18] D. Burghelea and M. Vigué-Poirrier, *Cyclic homology of commutative algebras. I. Algebraic topology, rational homotopy*, Proc. Conf., Louvain-la-Neuve/Belg. 1986, Lecture Notes in Math., Vol. 1318 (1988), 51–72.
[19] J. Burroughs, *Operations in Grothendieck rings and the symmetric group*, Canad. J. Math. **26** (1974), 543–550.
[20] P. Cartier, *Groupes formels associés aux anneaux de Witt generalisés*, C. R. Acad. Sci. Paris, Sér. A **265** (1967), 49–52.
[21] Ph. Cassou-Noguès and M.J. Taylor, *Opérations d'Adams et théorèmes de points fixes en K-théorie des algèbres de groupes*, J. Algebra **216** (1) (1999), 334–365.
[22] J.-L. Cathelineau, *λ-structures in algebraic K-theory and cyclic homology*, K-Theory **4** (6) (1990/91), 591–606.
[23] T. Chinburg, B. Erez, G. Pappas and M.J. Taylor, *Riemann–Roch type theorems for arithmetic schemes with a finite group action*, J. Reine Angew. Math. **489** (1997), 151–187.
[24] T. Chinburg, G. Pappas and M.J. Taylor, *ε-constants and the Galois structure of de Rham cohomology. II*, J. Reine Angew. Math. **519** (2000), 201–230.
[25] F.J.B.J. Clauwens, *The K-groups of λ-rings. I. Construction of the logarithmic invariant*, Compositio Math. **61** (3) (1987), 295–328.
[26] F.J.B.J. Clauwens, *The K-groups of λ-rings. II. Invertibility of the logarithmic map*, Compositio Math. **92** (2) (1994), 205–225.
[27] F.J.B.J. Clauwens, *Commuting polynomials and λ-ring structures on* $\mathbf{Z}[x]$, J. Pure Appl. Algebra **95** (3) (1994), 261–269.
[28] V.I. Danilov, *Geometry of toric varieties*, Russian Math. Surveys **33** (2) (1978), 97–154; transl.: Uspekhi Mat. Nauk **33** (2) (1978), 85–134.
[29] A.A. Davydov, *Periodic λ-rings and periods of finite groups*, Mat. Sb. **188** (8) (1997), 75–82; transl.: Sb. Math. **188** (8) (1997), 1183–1190.
[30] J. Dieudonné, *Schur functions and group representations*, Young Tableaux and Schur Functors in Algebra and Geometry (Toruń, 1980), Astérisque **87–88** (1981), 7–19.
[31] A.W.M. Dress and C. Siebeneicher, *The Burnside ring of the infinite cyclic group and its relations to the necklace algebra, λ-rings, and the universal ring of Witt vectors*, Adv. Math. **78** (1) (1989), 1–41.
[32] G. Duchamp, A. Klyachko, D. Krob and J.-Y. Thibon, *Noncommutative symmetric functions. III. Deformations of Cauchy and convolution algebras*, Discrete Math. Theoret. Comput. Sci. **1** (1) (1997), 159–216.
[33] J. Dupont, *K-Theory*, Lecture Notes Ser., No. 11, Matematisk Institut, Aarhus Universitet, Aarhus (1968).
[34] J.L. Dupont, *Scissors Congruences, Group Homology and Characteristic Classes*, Nankai Tracts in Math., Vol. 1, World Scientific, Singapore (2001).
[35] B. Eckmann and G. Mislin, *Chern classes of group representations over a number field*, Compositio Math. **44** (1–3) (1981), 41–65.
[36] M. Elhamdadi, *A note on λ-operations in orthogonal K-theory*, Proc. Amer. Math. Soc. **128** (1) (2000), 1–4.
[37] B.L. Feigin and B.L. Tsygan, *Additive K-theory*, K-theory, Arithmetic and Geometry, Semin., Moscow Univ. 1984–1986, Lecture Notes in Math., Vol. 1289 (1987), 67–209.
[38] W. Fulton, *Introduction to Toric Varieties*, Ann. of Math. Stud., Vol. 131, Princeton University Press, Princeton, NJ (1993).
[39] W. Fulton and S. Lang, *Riemann–Roch Algebra*, Grundlehren Math. Wiss., Vol. 277, Springer, New York (1985).
[40] A.M. Garsia and M. Haiman, *A remarkable q,t-Catalan sequence and q-Lagrange inversion*, J. Algebraic Combin. **5** (3) (1996), 191–244.

[41] Ph. Gaucher, *Lambda-opérations sur l'homologie d'une algèbre de Lie de matrices*, K-Theory **13** (2) (1998), 151–167.

[42] C.D. Gay, G.C. Morris and I. Morris, *Computing Adams operations on the Burnside ring of a finite group*, J. Reine Angew. Math. **341** (1983), 87–97.

[43] I.M. Gelfand, D. Krob, A. Lascoux, B. Leclerc, V.S. Retakh and J.-Y. Thibon, *Noncommutative symmetric functions*, Adv. Math. **112** (2) (1995), 218–348.

[44] M. Gerstenhaber and S.D. Schack, *A Hodge-type decomposition for commutative algebra cohomology*, J. Pure Appl. Algebra **48** (1987), 229–247.

[45] M. Gerstenhaber and S.D. Schack, *Algebraic cohomology and deformation theory*, Deformation Theory of Algebras and Structures and Applications, NATO Adv. Study Inst., Castelvecchio-Pascoli/Italy 1986, NATO ASI Ser., Ser. C **247** (1988), 11–264.

[46] M. Gerstenhaber and S.D. Schack, *The shuffle bi-algebra and the cohomology of commutative algebras*, J. Pure Appl. Algebra **70** (3) (1991), 263–272.

[47] H. Gillet and C. Soulé, *Intersection theory using Adams operations*, Invent. Math. **90** (2) (1987), 243–277.

[48] H. Gillet and C. Soulé, *An arithmetic Riemann–Roch theorem*, Invent. Math. **110** (3) (1992), 473–543.

[49] H. Gillet and C. Soulé, *Filtrations on higher algebraic K-theory*, Algebraic K-Theory (Seattle, WA, 1997) Proc. Sympos. Pure Math., Vol. 67, Amer. Math. Soc., Providence, RI (1999), 89–148.

[50] Y.Gh. Gouda, *Adams and Steenrod operators in dihedral homology*, Internat. J. Math. Math. Sci. **21** (2) (1998), 341–346.

[51] A. Goupil, *Generating functions for irreducible characters of S_n indexed with multiple hooks*, Ann. Sci. Math. Québec **23** (2) (1999), 189–198.

[52] D.R. Grayson, *Exterior power operations on higher K-theory*, K-Theory **3** (3) (1989), 247–260.

[53] D.R. Grayson, *Adams operations on higher K-theory*, K-Theory **6** (2) (1992), 97–111.

[54] D.R. Grayson, *Weight filtrations in algebraic K-theory*, U. Jannsen et al., eds, Motives. Proceedings of the Summer Research Conference on Motives, held at the University of Washington, Seattle, WA, USA, July 20–August 2, 1991, Proc. Sympos. Pure Math., Vol. 55 (1) (1994), 207–237.

[55] A. Grothendieck, *La théorie des classes de Chern*, Bull. Soc. Math. France **86** (1958), 137–154.

[56] A. Grothendieck, *Classes de Chern et représentations linéaires des groupes discrets*, Dix Exposés sur la Cohomologie des Schémas, North-Holland, Amsterdam/Masson, Paris (1968), 215–305.

[57] A. Grothendieck, *Classes de faisceaux et théorème de Riemann–Roch*, Sém. Géom. algébrique Bois Marie 1966/67, SGA 6, Lecture Notes Math., Vol. 225 (1971), 20–77.

[58] H. Hadwiger, *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie*, Grundlehren Math. Wiss., Vol. 93, Springer, Berlin–Göttingen–Heidelberg (1957).

[59] M. Hazewinkel, *Formal Groups and Applications*, Pure and Applied Mathematics, Vol. 78, Academic Press, New York–London (1978).

[60] M. Hazewinkel, *The primitives of the Hopf algebra of noncommutative symmetric functions*, Preprint, CWI, Amsterdam (2001).

[61] H.L. Hiller, *λ-rings and algebraic K-theory*, J. Pure Appl. Algebra **20** (3) (1981), 241–266.

[62] K. Hirata and A. Kono, *On the Bott cannibalistic classes*, Publ. Res. Inst. Math. Sci. **18** (3) (1982), 1187–1191.

[63] F. Hirzebruch, *Neue topologische Methoden in der algebraischen Geometrie*, Ergebn. Math. Grenzgeb. 9. Heft, Springer (1956).

[64] F. Hirzebruch, *Gesammelte Abhandlungen*, Collected Papers. Band I: 1951–1962. Band II: 1963–1987, Springer, Berlin (1987).

[65] P. Hoffman, *τ-Rings and Wreath Product Representations*, Lecture Notes in Math., Vol. 746, Springer, Berlin (1979).

[66] P. Hoffman, *Exponential maps and λ-rings*, J. Pure Appl. Algebra **27** (2) (1983), 131–162.

[67] D. Husemoller, *Fibre Bundles*, McGraw-Hill, New York–London–Sydney (1966).

[68] A. Joyal, *δ-anneaux et λ-anneaux*, C. R. Math. Rep. Acad. Sci. Canada **7** (4) (1985), 227–232.

[69] A. Joyal, *Règle des signes en algèbre combinatoire*, C. R. Math. Rep. Acad. Sci. Canada **7** (5) (1985), 285–290.

[70] J.-M. Kantor and A. Khovanskij, *Une application du théorème de Riemann–Roch combinatoire au polynôme d'Ehrhart des polytopes entier de $\mathbf{R}^d$*, C. R. Acad. Sci., Paris, Sér. I **317** (5) (1993), 501–507.

[71] M.R. Kantorovitz, *Adams operations and the Dennis trace map*, J. Pure Appl. Algebra **144** (1) (1999), 21–27.

[72] C. Kassel, *Une formule de Künneth pour la décomposition de l'homologie cyclique des algèbres commutatives*, Math. Scand. **70** (1) (1992), 27–33.

[73] M. Kervaire, *Opérations d'Adams en théorie des représentations linéaires des groupes finis*, Enseign. Math. (2) **22** (1–2) (1976), 1–28.

[74] D. Knutson, *λ-Rings and the Representation Theory of the Symmetric Group*, Lecture Notes in Math., Vol. 308, Springer, Berlin (1973).

[75] B. Köck, *Shuffle products in higher K-theory*, J. Pure Appl. Algebra **92** (3) (1994), 269–307.

[76] B. Köck, *The Grothendieck–Riemann–Roch theorem for group scheme actions*, Ann. Sci. École Norm. Sup. (4) **31** (3) (1998), 415–458.

[77] C. Kratzer, *λ-structure en K-théorie algébrique*, Comment. Math. Helv. **55** (2) (1980), 233–254.

[78] C. Kratzer, *Opérations d'Adams et représentations de groupes*, Enseign. Math. (2) **26** (1–2) (1980), 141–154.

[79] D. Krob, B. Leclerc and J.-Y. Thibon, *Noncommutative symmetric functions. II. Transformations of alphabets*, Internat. J. Algebra Comput. **7** (2) (1997), 181–264.

[80] D. Krob and J.-Y. Thibon, *Noncommutative symmetric functions. IV. Quantum linear groups and Hecke algebras at $q = 0$*, J. Algebraic Combin. **6** (4) (1997), 339–376.

[81] D. Krob and J.-Y. Thibon, *Noncommutative symmetric functions. V. A degenerate version of $U_q(\mathrm{gl}_N)$*, Internat. J. Algebra Comput. **9** (3–4) (1999), 405–430.

[82] K. Kurano and P.C. Roberts, *Adams operations, localized Chern characters, and the positivity of Dutta multiplicity in characteristic* 0, Trans. Amer. Math. Soc. **352** (7) (2000), 3103–3116.

[83] L. Lapointe, A. Lascoux and J. Morse, *Determinantal expressions for Macdonald polynomials*, Internat. Math. Res. Notices **18** (1998), 957–978.

[84] A. Lascoux, *Calcul de Schur et extensions Grassmanniennes des λ-anneaux*, Combinatoire et représentation du groupe symétrique, Lecture Notes in Math., Vol. 579, Springer, Berlin (1977), 182–216.

[85] A. Lascoux and M. Lassalle, *Une identité remarquable en théorie des partitions*, Math. Ann. **318** (2) (2000), 299–313.

[86] F. Lecomte, *Opérations d'Adams en K-théorie algébrique*, K-Theory **13** (2) (1998), 179–207.

[87] F. Lecomte, *Simplicial schemes and Adams operations*, Algebraic K-Theory and its Applications (Trieste, 1997), World Sci. Publishing, River Edge, NJ (1999), 437–449.

[88] M. Levine, *Lambda-operations, K-theory and motivic cohomology*, V.P. Snaith, ed., Algebraic K-Theory, Papers from the 2nd Great Lakes conference, Canada, March 1996 (in memory of Robert Wayne Thomason), Providence, Fields Inst. Commun. **16** (1997), 131–184.

[89] M. Levine, *Mixed Motives*, Mathematical Surveys and Monographs, Vol. 57, Amer. Math. Soc., Providence, RI (1998).

[90] J.-L. Loday, *Opérations sur l'homologie cyclique des algèbres commutatives*, Invent. Math. **96** (1) (1989), 205–230.

[91] J.-L. Loday and C. Procesi, *Cyclic homology and lambda operations*, Algebraic K-Theory: Connections with Geometry and Topology, Proc. Meet., Lake Louise/Can. 1987, NATO ASI Ser., Ser. C, Vol. 279 (1989), 209–224.

[92] I.G. Macdonald, *Polynomial functors and wreath products*, J. Pure Appl. Algebra **18** (2) (1980), 173–204.

[93] I.G. Macdonald, *Symmetric Functions and Hall Polynomials*, 2nd edn, Clarendon Press, Oxford (1995).

[94] C. Malvenuto and C. Reutenauer, *Duality between quasi-symmetric functions and the Solomon descent algebra*, J. Algebra **177** (3) (1995), 967–982.

[95] Yu.I. Manin, *Lectures on the K-functor in algebraic geometry*, Russ. Math. Surv. **24** (5) (1969), 1–89.

[96] R. McCarthy, *On operations for Hochschild homology*, Commun. Algebra **21** (8) (1993), 2947–2965.

[97] R.V. Moody and A. Pianzola, *λ-mappings between representation rings of Lie algebras*, Canad. J. Math. **35** (5) (1983), 898–960.

[98] R. Morelli, *A theory of polyhedra*, Adv. Math. **97** (1) (1993), 1–73.

[99] R. Morelli, *The K-theory of a toric variety*, Adv. Math. **100** (2) (1993), 154–182.

[100] R. Morelli, *Pick's theorem and the Todd class of a toric variety*, Adv. Math. **100** (2) (1993), 183–231.

[101] I. Morris and C.B. Wensley, *Adams operations and λ-operations in β-rings*, Discrete Math. **50** (2–3) (1984), 253–270.

[102] T. Natsume and S.D. Schack, *A decomposition for the cyclic cohomology of a commutative algebra*, J. Pure Appl. Algebra **61** (3) (1989), 273–282.

[103] A. Nenashev, *Simplicial definition of λ-operations in higher K-theory*, Algebraic K-Theory, Pap. Semin., Leningrad/USSR, Adv. Sov. Math. **4** (1991), 9–20.

[104] A. Nenashev, *Comparison theorem for λ-operations in higher algebraic K-theory*, C. Kassel et al., eds, K-Theory. Contributions of the International Colloquium, Strasbourg, France, June 29–July 3, 1992, Astérisque **226** (1994), 335–369.

[105] Ph. Nuss, *Décomposition de la cohomologie cyclique bivariante des algèbres commutatives*, Math. Scand. **70** (1) (1992), 5–26.

[106] G. Ochoa, *λ-ring structure on the Burnside ring of Hamiltonian groups*, Arch. Math. **43** (1984), 312–321.

[107] T. Oda, *Convex Bodies and Algebraic Geometry. An Introduction to the Theory of Toric Varieties*, Ergebn. Math. Grenzgeb., 3 Folge, Bd. 15, Springer, Berlin (1988).

[108] A. Osse, *λ-structures and representation rings of compact connected Lie groups*, J. Pure Appl. Algebra **121** (1) (1997), 69–93.

[109] G. Pappas, *Galois module structure and the γ-filtration*, Compositio Math. **121** (1) (2000), 79–104.

[110] F. Patras, *Construction géométrique des idempotents Eulériens. Filtration des groupes de polytopes et des groupes d'homologie de Hochschild*, Bull. Soc. Math. France **119** (2) (1991), 173–198.

[111] F. Patras, *La décomposition en poids des algèbres de Hopf*, Ann. Inst. Fourier **43** (4) (1993), 1067–1087.

[112] F. Patras, *L'algèbre des descentes d'une bigèbre graduée*, J. Algebra **170** (2) (1994), 547–566.

[113] F. Patras, *Lambda-anneaux non commutatifs*, Commun. Algebra **23** (6) (1995), 2067–2078.

[114] F. Patras, *Adams operations, iterated integrals and free loop spaces cohomology*, J. Pure Appl. Algebra **114** (3) (1997), 273–286.

[115] F. Patras, *Adams operations and periodic cyclic cohomology*, Algebraic K-Theory and its Applications (Trieste, 1997), World Sci. Publishing, River Edge, NJ (1999), 567–588.

[116] F. Patras, *Adams operations, algebras up to homotopy and cyclic homology*, Topology **39** (6) (2000), 1089–1101.

[117] F. Patras, *Generic algebras and iterated Hochschild homology*, J. Pure Appl. Algebra **162** (2–3) (2001), 337–357.

[118] F. Patras and C. Reutenauer, *Higher Lie idempotents*, J. Algebra **222** (1) (1999), 51–64.

[119] C. Pedrini, *Riemann–Roch and Chow theories for singular varieties*, Semin. Geom., Univ. Studi Bologna 1987–1988 (1988), 87–113.

[120] D. Quillen, *On the cohomology and K-theory of the general linear groups over a finite field*, Ann. of Math. (2) **96** (1972), 552–586.

[121] C. Reutenauer, *Theorem of Poincaré–Birkhoff–Witt, logarithm and symmetric group representations of degrees equal to Stirling numbers*, Combinatoire Énumérative, Proc. Colloq., Montréal/Can. 1985, Lecture Notes in Math., Vol. 1234 (1986), 267–284.

[122] C. Reutenauer, *Free Lie Algebras*, London Mathematical Society Monographs, New Series, Vol. 7, Clarendon Press, Oxford (1993).

[123] D. Roessler, *Lambda structure en K-théorie des fibrés algébriques Hermitiens*, C. R. Acad. Sci. Paris Sér. I Math. **322** (3) (1996), 251–254.

[124] D. Roessler, *An Adams–Riemann–Roch theorem in Arakelov geometry*, Duke Math. J. **96** (1) (1999), 61–126.

[125] M. Ronco, *On the Hochschild homology decompositions*, Commun. Algebra **21** (12) (1993), 4699–4712.

[126] M. Ronco, *Free Lie algebra and lambda-ring structure*, Bull. Aust. Math. Soc. **50** (3) (1994), 373–382.

[127] N.W. Rymer, *Power operations on the Burnside ring*, J. London Math. Soc. (2) **15** (1) (1977), 75–80.

[128] C.-H. Sah, *Hilbert's Third Problem: Scissors Congruence*, Research Notes in Math., Vol. 33, Pitman Advanced Publishing Program, San Francisco–London–Melbourne (1979).

[129] T. Scharf and J.-Y. Thibon, *A Hopf-algebra approach to inner plethysm*, Adv. Math. **104** (1) (1994), 30–58.

[130] L. Schwartz, *Vecteurs propres de l'opération d'Adams dans* $K_0(C\mathrm{P}^\infty/C\mathrm{P}^k, Z/p)$, C. R. Acad. Sci. Paris Sér. I Math. **296** (23) (1983), 1005–1008.

[131] G.B. Segal, *Exponential isomorphisms for λ-rings*, Quart. J. Math. Oxford Ser. (2) **22** (1971), 371–378.

[132] W.K. Seiler, *λ-rings and Adams operations in algebraic K-theory. Beilinson's conjectures on special values of L-functions*, Meet. Oberwolfach/FRG 1986, Perspect. Math. **4** (1988), 93–102.

[133] C. Siebeneicher, *λ-Ringstrukturen auf dem Burnsidering der Permutationsdarstellungen einer endlichen Gruppe*, Math. Z. **146** (1976), 223–238.

[134] L. Solomon, *A Mackey formula in the group of a Coxeter group*, J. Algebra **41** (1976), 255–268.

[135] C. Soulé, *Operations on étale K-theory. Applications*, Algebraic K-Theory, Part I (Oberwolfach, 1980), Lecture Notes in Math., Vol. 966, Springer, Berlin (1982), 271–303.

[136] C. Soulé, *Opérations en K-théorie algébrique*, Canad. J. Math. **37** (1985), 488–550.

[137] M.E. Sweedler, *Hopf Algebras*, Benjamin, New York (1969).

[138] D.O. Tall and G.C. Wraith, *Representable functors and operations on rings*, Proc. London Math. Soc. (3) **20** (1970), 619–643.

[139] B. Teissier, *Variétés toriques et polytopes*, Semin. Bourbaki, 33e année, Vol. 1980/81, Exp. No. 565, Lecture Notes in Math., Vol. 901 (1981), 71–84.

[140] J.-Y. Thibon, *Fonctions symétriques et opérations d'Adams de l'anneau de Grothendieck d'un groupe symétrique*, C. R. Acad. Sci. Paris Sér. I Math. **313** (12) (1991), 829–832.

[141] J.-Y. Thibon, *The inner plethysm of symmetric functions and some of its applications*, Bayreuth. Math. Schr. No. 40 (1992), 177–201.

[142] J.-Y. Thibon, *Lectures on noncommutative symmetric functions*, Interaction of Combinatorics and Representation Theory, MSJ Mem., Vol. 11, Math. Soc. Japan, Tokyo (2001), 39–94.

[143] C.B. Thomas, *Characteristic Classes and the Cohomology of Finite Groups*, Cambridge Studies in Advanced Mathematics, Vol. 9, Cambridge University Press, Cambridge (1986).

[144] T. tom Dieck, *Transformation Groups and Representation Theory*, Lecture Notes in Math., Vol. 766, Springer, Berlin (1979).

[145] T. tom Dieck, *Über λ-Ringstrukturen auf dem Burnside-Ring*, Math. Z. **187** (4) (1984), 505–510.

[146] T. tom Dieck, *The Artin–Hasse logarithm for λ-rings*, Algebraic Topology (Arcata, CA, 1986), Lecture Notes in Math., Vol. 1370, Springer, Berlin (1989), 409–415.

[147] E. Vallejo, *The free β-ring on one generator*, J. Pure Appl. Algebra **86** (1) (1993), 95–108.

[148] M. Vigué-Poirrier, *Décompositions de l'homologie cyclique des algèbres différentielles graduées commutatives*, K-Theory **4** (5) (1991), 399–410.

[149] W. Wang, *Equivariant K-theory, wreath products, and Heisenberg algebra*, Duke Math. J. **103** (1) (2000), 1–23.

[150] C. Weibel, *Le caractre de Chern en homologie cyclique periodique*, C. R. Acad. Sci., Paris, Sér. I **317** (9) (1993), 867–871.

[151] C.A. Weibel, *An Introduction to Homological Algebra*, Cambridge Studies in Advanced Mathematics, Vol. 38, Cambridge University Press, Cambridge (1994).

[152] C. Wilkerson, *Lambda-rings, binomial domains, and vector bundles over $CP(\infty)$*, Comm. Algebra **10** (3) (1982), 311–328.

[153] M. Yang and J.B. Remmel, *Hook–Schur functions analogues of Littlewood's identities and their bijective proofs*, European J. Combin. **19** (2) (1998), 257–272.

Section 5A
Groups and Semigroups

This Page Intentionally Left Blank

Branch Groups

Laurent Bartholdi

Section de Mathématiques, Université de Genève, CP 240, 1211 Genève 24, Switzerland
E-mail: laurent@math.berkeley.edu

Rostislav I. Grigorchuk

Department of Ordinary Differential Equations, Steklov Mathematical Institute, Gubkina Street 8, Moscow 119991, Russia
E-mail: grigorch@mi.ras.ru

Zoran Šunić

Department of Mathematics, Cornell University, 431 Malott Hall, Ithaca, NY 14853, USA
E-mail: sunik@math.cornell.edu

Contents

HANDBOOK OF ALGEBRA, VOL. 3
Edited by M. Hazewinkel

0. Introduction

Branch groups were defined only recently although they make nonexplicit appearances in the literature in the past, starting with the article of J. Wilson [Wil71]. Moreover, such examples as infinitely iterated wreath products or the group of tree automorphisms $\mathsf{Aut}(\mathcal{T})$, where $\mathcal{T}$ is a regular rooted tree, go back to the work of L. Kaloujnine, B. Neumann, P. Hall and others.

Branch groups were explicitly defined for the first time at the 1997 St-Andrews conference in Bath in a talk by the second author. Immediately, this sparked a great interest among group theorists, who started investigating numerous properties of branch groups (see [Gri00,BG00b,BG02,GW00]) as well as J. Wilson's classification of just-infinite groups.

There are two new approaches to the definition of a branch group, given in [Gri00]. The first one is purely algebraic, defining branch groups as groups whose lattice of subnormal subgroups is similar to the structure of a spherically homogeneous rooted tree. The second one is based on a geometric point of view according to which branch groups are groups acting spherically transitively on a spherically homogeneous rooted tree and having structure of subnormal subgroups similar to the corresponding structure in the full group $\mathsf{Aut}(\mathcal{T})$ of automorphisms of the tree.

Until 1980 no examples of finitely generated branch groups were known and the first such examples were constructed in [Gri80]. These examples are usually referred to as the first and the second Grigorchuk groups, following Pierre de la Harpe [Har00]. Other examples soon appeared in [Gri83,Gri84,Gri85a,GS83a,GS83b,GS84,Neu86] and these examples are the basic examples of branch groups, the study of which continues at the present time. Let us mention that the examples of S. Alešin [Ale72] and V. Suščans'kiĭ [Suš79] that appeared earlier also belong to the class of finitely generated branch groups, but the methods used in the study of the groups from [Ale72] and [Suš79] did not allow the discovery of the branch structure and this was done much later.

Already in [Gri80] the main features of a general method which works for almost any finitely generated branch group had appeared: one considers the stabilizer of a vertex on the first level and projects it on the corresponding subtree. Then either this projection is equal to the initial group and thus one gets a self-similarity property, or otherwise one gets a finite or infinite chain of branch groups related by some homomorphisms. One of the essential properties of this chain is that these homomorphisms satisfy a "Lipschitz" property of norm reduction, which lends itself to arguments using direct induction on length in the case of a self-similar group, or simultaneous induction on length for all groups in a chain.

Before we give more information of a historical character and briefly describe the main directions of investigation and the main results in the area, let us explain why the class of branch groups is important. There is a lot of evidence that this is indeed the case. For example,

(1) The class of branch groups is one of the classes into which the class of just-infinite groups naturally splits (just-infinite groups are groups all of whose proper quotients are finite).

(2) The class contains groups with many extraordinary properties, like infinite finitely generated torsion groups, groups of intermediate growth, amenable but not elementary amenable groups, groups of finite width, etc.
(3) Branch groups have many applications and are related to analysis, geometry, combinatorics, probability, computer science, etc.
(4) They are relatively easy to handle and usually the proofs even of deep theorems are short and do not require special techniques. Therefore the branch groups constitute an easy-to-study class of groups, whose basic examples have already appeared in many textbooks and lecture notes, for example, in [KM82,Bau93,Rob96,Har00].

This survey article deals almost exclusively with abstract branch groups. The theory of profinite branch groups is also being actively developed at present (see [GHZ00,Gri00, Wil00]), but we hardly touch on this subject.

The survey does not pretend to be complete. There are several topics that we did not include in the text due to the lack of space and time. Among them, we mention the results of S. Sidki from [Sid97] on thin algebras associated to Gupta–Sidki groups, the results on automorphisms of branch groups of S. Sidki from [Sid87a] and the recent results of Lavreniuk and Nekrashevich from [LN02], the results of C. Röver on embeddings of Grigorchuk groups into finitely presented simple groups and on abstract commensurators from [Röv99, Röv00], the results of V. Sushchanskiĭ on factorizations based on the use of torsion branch groups [Sus89,Sus94], the results of B. Fine, A. Gaglione, A. Myasnikov and D. Spellman from [FGMS01] on discriminating groups, etc.

0.1. *Just-infinite groups*

Let $\mathcal{P}$ be any property which is preserved under homomorphic images (we call such a property an *$\mathcal{H}$-property*). Any infinite finitely generated group can be mapped onto a just-infinite group (see [Gri00,Har00]), so if there is an infinite finitely generated group with the $\mathcal{H}$-property $\mathcal{P}$ then there is a just-infinite finitely generated group with the same property. Among the $\mathcal{H}$-properties let us mention the property of being a torsion group, not containing the free group F_2 on two elements as a subgroup, having subexponential growth, being amenable, satisfying a given identity, having bounded generation, finite width, trivial space of pseudocharacters (for a relation to bounded cohomology see [Gri95]), only finite-index maximal subgroups, T-property of Kazhdan, etc.

The branch just-infinite groups are precisely the just-infinite groups whose structure lattice of subnormal subgroups (with some identifications) is isomorphic to the lattice of closed and open subsets of the Cantor set. This is the approach of J. Wilson from [Wil71].

In that paper, J. Wilson split the class of just-infinite groups into two subclasses – the groups with finite and the groups with infinite structure lattice. The dichotomy of J. Wilson can be reformulated (see [Gri00]) in the form of a trichotomy according to which any finitely generated just-infinite group is either a branch group or can easily be constructed from a simple group or from a hereditarily just-infinite group (i.e. a residually finite group all of whose subgroups of finite index are just-infinite).

Therefore the study of finitely generated just-infinite groups naturally splits into the study of branch groups, infinite simple groups and hereditarily just-infinite groups. Unfortunately, at the moment, none of these classes of groups are well understood, but we have several (classes of) examples.

There are several examples and constructions of finitely generated infinite simple groups, probably starting with the example of G. Higman in [Hig51], followed by the finitely presented example of R. Thompson, generalized by G. Higman in [Hig74] (see also the survey [CFP96] and [Bro87]), the constructions of different monsters by A. Ol'shanskiĭ (see [Ol'91]), as well as by S. Adyan and I. Lysionok in [AL91], and more recently some finitely presented examples by C. Röver in [Röv99]. The $\mathcal{H}$-properties that can be satisfied by such groups are, for instance, the Burnside identity x^p, for large prime p, and triviality of the space of pseudocharacters. The latter holds for the simple groups T and V of R. Thompson (this follows from the results on finiteness of commutator length, see [GS87, Bro87]).

All known hereditarily just-infinite groups (like the projective groups $\mathsf{PSL}(n, \mathbb{Z})$ for $n > 2$) are linear (in the profinite case there are extra examples like the Nottingham group), so by the alternative of Tits they contain F_2 as a subgroup and therefore cannot be amenable, of intermediate growth, torsion, etc. However, they can have bounded generation: it is shown in [CK83] that this holds for $\mathsf{SL}(n, \mathbb{Z})$, $n > 2$, and therefore also for $\mathsf{PSL}(n, \mathbb{Z})$.

It seems that there are fewer constraints in the class of branch groups and that they can have various $\mathcal{H}$-properties, some of which are listed below. It is conjectured that many of these properties do not hold for groups from the previous two classes. On the other hand, branch groups cannot satisfy nontrivial identities (see [Leo97b,Wil00] where the proof is given for the just-infinite case).

0.2. *Algorithmic aspects*

Branch groups have good algorithmic properties. In the branch groups of G or GGS type (or more generally spinal type groups) the word problem is solvable by an universal branch algorithm described in [Gri84]. This algorithm is very fast and requires a minimal amount of memory.

The conjugacy problem was unsettled for a long time, and it was solved for the basic examples of branch groups just recently. The article [WZ97] solves the problem for regular branch p-groups, where p is an odd prime, and the argument uses the property of "conjugacy separability" as well as profinite group machinery. In [Leo98a] and [Roz98] a different approach was used, which also works in case $p = 2$. These ideas were developed in [GW00] in different directions. For instance, it was shown that, under certain conditions, the conjugacy problem is solvable for all subgroups of finite index in a given branch group (we mention here that the property of solvability of conjugacy problem, in contrast with the word problem, is not preserved when one passes to subgroups of finite index). Still, we are far from understanding whether the conjugacy problem is solvable in all branch groups with solvable word problem.

The isomorphism problem was also considered in [Gri84] where it is proven that each of the uncountably many constructed groups G_ω is isomorphic to at most countably many of them, thus showing that the construction gives uncountably many non-isomorphic examples. It would be very interesting to distinguish all these examples.

Branch groups are related to groups of finite automata. A brief account is given in Section 1.5 (see also [GNS00]). Every group generated by finite automata has a solvable word problem. It is unclear if every such group has solvable conjugacy problem. On the other hand, it seems that the isomorphism problem cannot be solved in this particular case. Indeed, according to the results in [KBS91], the freeness of a matrix group with integer entries cannot be determined, and the general linear group $\mathsf{GL}(n,\mathbb{Z})$ can be embedded in the group of automata defined over an alphabet on 2^n letters as shown by A. Brunner and S. Sidki (see [BS98]).

0.3. *Group presentations*

In Section 4 we study presentations of branch groups by generators and relations. It seems probable that no branch group is finitely presented. However, the regular branch groups have nice recursive presentations called L-presentations. The first such presentation was found for the first Grigorchuk group by I. Lysionok in [Lys85]. Shortly afterwards, S. Sidki devised a general method yielding recursive definitions of such groups, and applied it to the Gupta–Sidki group [Sid87b]. In [Gri98,Gri99] the idea and the result of I. Lysionok were developed in different directions. In [Gri99] it was proven that the I. Lysionok system of relations is minimal and the Schur multiplier of the group was computed: it is $(\mathbb{Z}/2)^\infty$. Thus the second homology group of the first Grigorchuk group is infinite dimensional. In [Gri99] it was indicated that the Gupta–Sidki p-groups also have finite L-presentations. On the other hand, it was shown in [Gri98] how L-presentations can be used to embed some branch groups into finitely presented groups using just one HNN extension. The important feature of this embedding is that it preserves amenability. The first examples of finitely presented, amenable but not elementary amenable groups were constructed this way, thus providing new examples of good fundamental groups (in the terminology of Freedman and Teichler [FT95]).

In [Bar] the notion of an L-presentation was slightly extended to the notion of an endomorphic presentation in a way that allowed to show that a finitely generated, fractal, regular branch group satisfying some natural extra conditions has a finite endomorphic presentation. A number of concrete L and endomorphic presentations of branch groups appear in the article along with general facts on such presentations.

As was stated, no known branch group has finite presentation. For the first Grigorchuk group this was already mentioned in [Gri80] with a sketch of a proof that was given completely in [Gri99]. In [Gri84] two other proofs were presented. Yet another approach for proving the absence of finite presentations is used by N. Gupta in [Gup84]. More on the history of the presentation problem and related methods appears in Section 4.1.

0.4. *Burnside groups*

The third part of the survey is devoted to the algebraic properties of branch groups in general and of the most important examples.

The first examples of branch groups appeared in [Gri80] as examples of infinite finitely generated torsion groups. Thus the branch groups are related to the Burnside Problem on torsion groups. This difficult problem has three branches: the Unbounded Burnside Problem, the Bounded Burnside Problem and the Restricted Burnside problem (see [Ady79, Kos90]) and, in one way or another, all of them are solved. However, there is still a series of unsolved problems in the neighborhood of the Burnside problem, which are very important to the theory of groups – to name one, "is there a finitely presented infinite torsion group?". The first example of an infinite finitely generated torsion group, which provided a negative answer to the Unbounded Burnside Problem, was constructed by Golod in [Gol64] and it was based on the Golod–Shafarevich theorem. The actual problem of constructing simple examples which do not require the use of such deep results as the Golod–Shafarevich theorem remained open until such examples appeared in [Gri80]. Soon, more examples appeared in [Gri83,Gri84,Gri85a,GS83a,GS83b,GS84] and more recently in [BŠ01,Gri00, Bar00a,Šun00]. The early examples are finitely generated infinite p-groups, for p a prime, and the latter papers contain interesting examples that are not p-groups.

We already mentioned the idea of induction on word length, based on the fact that the projections on coordinates decrease the length. In conjunction, the idea of fixing larger and larger layers of the tree under taking powers was developing. The stabilization occurs in the first Grigorchuk group from [Gri80] after three steps, and for the second example from the same article it occurs after the second step of taking p-th powers. Using a slightly modified metric on the group [Bar98], the stabilization can be made to appear after just one step; this is extensively developed in [BŠ01]. Examples with strong stabilization properties for the standard word metric are constructed in [GS83a], where stabilization takes place after the first step. The notion of *depth* of an element, i.e. the number of decompositions one must perform to decrease the length down to 1 was introduced by S. Sidki in [Sid87a], and is very useful in some situations.

One of the important principles of modern group theory is to try to develop asymptotic methods related to growth, amenability and other asymptotic notions. In [Gri84] the torsion growth functions were introduced for finitely generated torsion groups and it was shown that some examples from [Gri80,Gri83,Gri84] have polynomial growth in this sense. These results were improved in many directions in [Lys98,Leo97a,Leo99,BŠ01] and some of them are described in Section 6.

Among the main consequences of the theory of E. Zelmanov (see [Zel90,Zel91]) is that if a finitely generated torsion residually finite group has finite exponent (i.e. there exists $n \neq 0$ such that $g^n = 1$ for every element g in the group) then the group is finite. Although the results of E. Zelmanov do not depend on the classification of finite simple groups, the above mentioned consequence does and it would be nice to produce a proof that is independent of the classification. A simple proof that finitely generated torsion branch groups always have infinite exponent is provided (Theorem 6.9).

The profinite completion of a group of finite exponent is a torsion profinite group. If it has a just-infinite quotient (and this is the case if the group G is a virtually pro-p-group) then one gets a profinite just-infinite torsion group of bounded exponent. By Wilson's alternative such a group is either just-infinite branch, or hereditarily just-infinite; but by the results from [GHZ00] if it were branch it would have unbounded exponent, so the search for profinite groups of bounded exponent can be narrowed to hereditarily just-infinite groups. We believe such groups do not exist.

In Section 6 we give a simple proof of the fact that a finitely generated torsion branch group has infinite exponent. An interesting question is to describe the type of torsion growth that distinguishes the finite and infinite groups (it follows from the results of E. Zelmanov that there exists a recursive, unbounded function $z_k(n)$, depending on the number of generators k, such that every torsion group on k generators whose torsion growth is bounded above by $z_k(n)$ is finite). It seems likely that the problem can be reduced to the case of branch torsion groups.

In Section 6 we also analyze carefully the idea of the first example in [Gri80] which is not related to the stabilization, but rather uses a covering of a group by kernels of homomorphisms. The torsion groups (called G groups) that generalize the constructions of [Gri84, Gri85a] are investigated in greater detail in [BŠ01,Šun00]. We exhibit the construction and some interesting examples, based on existence of finite groups with certain required properties (D. Holt's example, for instance).

Until recently, it was not known whether there exist torsion-free just-infinite branch groups. Such an example was constructed in [BG02].

0.5. *Subgroups of branch groups*

The study of the subgroup structure of any class of groups is an important part of the investigation. Branch groups have a rich and nice subgroup structure which has not yet been completely investigated. In the early works attention was paid to some particular subgroups of small index such as the stabilizers of the first few levels, the initial members of the lower central series, derived series, etc. A fundamental observation made by N. Gupta and S. Sidki in [GS84] is that many GGS groups contain a normal subgroup K of finite index with the property that K contains *geometrically* K^m (m is the degree of the tree and "geometrically" means that the product K^m acts on the subtrees on the first level) as a subgroup of finite index. It so happens that all the main examples of branch groups have such a subgroup and this fact lies at the base of the definition of regular branch groups.

Rigid stabilizers and stabilizers of the first Grigorchuk group are described in [Roz90, BG02]. For the Gupta–Sidki p-groups this is done in [Sid87a]. The structure of normal subgroups of P. Neumann's groups [Neu86] is very simple, since the normal subgroups coincide with the (rigid) stabilizers. For branch p-groups it is more difficult to obtain the structure of the lattice of normal subgroups (this is related to the fact that pro-p-groups usually have a rich structure of subgroups of finite index). The lattice of normal subgroups of the first Grigorchuk group was recently described by the first author in [Bar00c] (see also [CST01] where the normal subgroups are described up to the fourth level).

In the study of infinite finitely generated groups an important role is played by the maximal and weakly maximal subgroups (i.e. subgroups of infinite index maximal with respect to this property). It is strange that little attention was paid to the latter until recently. The main result of E. Pervova [Per00] claims that in the basic examples of branch groups every proper maximal subgroup has finite index. This is in contrast with lattices in semisimple Lie groups, as follows from results of Margulis and Soifer [MS81].

Important examples of weakly maximal subgroups are the parabolic subgroups, i.e. the stabilizers of infinite paths in the tree. The structure of parabolic subgroups is described in [BG02] for some particular examples. They are not finitely generated and have a tree-like structure. It would be interesting to obtain a complete description of weakly maximal subgroups in branch groups.

0.6. *Lie algebras*

Section 8 deals with central series and associated Lie algebras of branch groups. There is a canonical way, due to W. Magnus in which a central series corresponds to a graded Lie ring or Lie algebra. The most interesting central series are the lower central series and the series of dimension subgroups. It was proved in [Gri89] that the Cesarò averages of the ranks of the factors in the lower central series of the first Grigorchuk group (which are elementary 2-groups) are uniformly bounded and it was conjectured that the ranks themselves were uniformly bounded, i.e. that the first Grigorchuk group has finite width. An important step in proving this conjecture was made in [Roz96], and a complete proof appears in [BG00a], using ideas of L. Kaloujnine [Kal48] and the notion of uniserial module. Moreover, a negative answer to a problem of E. Zelmanov on the classification of just-infinite profinite groups of finite width is provided in [BG00a], a new example of a group of finite width was constructed and the structure of the Cayley graph of the associated Lie algebras was described. This is one of the few cases of a nontrivial computation of a Cayley (or Lie) graph of a graded Lie algebra.

The question of the finiteness of width of other basic branch groups (first of all the Gupta–Sidki groups) was open for a long time and recently answered negatively by the first author in [Bar00c]. These results are also presented in Section 8.

An important role in the study of profinite completions of branch groups is played by the "congruence subgroup property" with respect to the sequence of stabilizers, meaning that every finite-index subgroup contains a level stabilizer $\mathsf{St}_G(n)$ for some n, and which holds for many branch groups. Nevertheless, there are branch groups without this property and the complete solution of the congruence subgroup property problem for the class of all branch groups is not completely resolved.

0.7. *Growth*

The fourth part of the paper deals with some geometric and analytic properties of branch groups. The main notion in asymptotic group theory is the notion of *growth* of a finitely generated group. The growth function $\gamma(n)$ of a finitely generated group G with respect to a system of generators S counts the number of group elements of length at most n. The group's type of growth – exponential, intermediate, polynomial – does not depend on the choice of S. One can easy construct an example of a group of polynomial growth of any given degree d (for instance, $\mathbb{Z}^d$) or a group of exponential growth (for instance F_2, the free group on two generators) but it is a highly nontrivial task to construct a group of intermediate growth. The question of existence of such groups of intermediate growth was posed by J. Milnor [Mil68c] and solved fifteen years later in [Gri83,Gri84,Gri85a], were the second author shows that the first group in [Gri80] and all p-groups G_ω in [Gri84, Gri85a] have intermediate growth; the estimates are of the form

$$e^{\sqrt{n}} \precsim \gamma(n) \precsim e^{n^\beta}, \tag{1}$$

for some $\beta < 1$, where for two functions $f, g : \mathbb{N} \to \mathbb{N}$ we write $f \precsim g$ to mean that there exists a constant C with $f(n) \leqslant g(Cn)$ for all $n \in \mathbb{N}$.

Milnor's problem was therefore solved using branch groups. Up to the present time all known groups of intermediate growth are either branch groups or groups constructed using branch groups and we believe that all just-infinite groups of intermediate growth are branch groups.

By using branch groups, the second author showed in [Gri84] that there exist uncountable chains and anti-chains of intermediate growth functions of groups acting on trees.

The upper bound in (1) was improved in [Bar98], and a general improvement of the upper bounds for all groups G_ω was given in [BŠ01] and [MP01].

One of the main remaining question on growth is whether there exists a group with growth precisely $e^{\sqrt{n}}$. It is known that if a group is residually nilpotent and its growth is strictly less than $e^{\sqrt{n}}$, then the group is virtually nilpotent and therefore has polynomial growth [Gri89,LM89]. Using arguments given in [Gri89] (see also [BG00a]), it follows that if a group of growth $e^{\sqrt{n}}$ exists in the class of residually-p groups, then it must have finite width. For some time, among all known examples of groups of intermediate growth only the first Grigorchuk group was known to have finite width, and the second author conjectured that this group has precisely this type of growth. However, this conjecture was killed by Y. Leonov [Leo00] and the first author [Bar01]; indeed the growth of the first Grigorchuk group is bounded below by e^{n^α} for some $\alpha > \frac{1}{2}$.

The notion of growth can be defined for other algebraic and geometric objects as well: algebras, graphs, etc. A very interesting topic is the study of the growth of graded Lie algebras $\mathcal{L}(G)$ associated to groups. In case of GGS groups some progress the was achieved in [Bar00c], where the growth of $\mathcal{L}(G)$ for the Gupta–Sidki 3-group and some other groups is computed; in particular, it is shown that the Gupta–Sidki group does not have finite width. A connection between the Lie algebra structure and the tree structure is used in the majoration of the growth of the associated Lie algebra by the growth of any homogeneous space G/P, where P is a parabolic subgroup, i.e. the stabilizer of an infinite path in a tree [Bar00c]. As metric spaces, these homogeneous spaces are equivalent to Schreier graphs. These graphs have an interesting structure: they are substitutional graphs, and have a fractal behaviour in the case of many fractal branch groups. They have polynomial growth, usually of nonintegral degree. These results are presented in Section 10.3.

One of the promising directions of research is the study of spectral properties of the above graphs. This question is linked to several famous problems of operator K-theory and theory of C^*-algebras. One of the first works in this direction is [BG00b], where it is shown that the spectrum of the discrete Laplace operator on such graphs can be a Cantor set, optionally with extra isolated points. The computation of these spectra is related to operator recursions that hold for the Laplace- or Hecke-type operators associated to the dynamical system $(G, \partial\mathcal{T}, \mu)$, where $(\partial\mathcal{T}, \mu)$ is the boundary of the tree endowed with the uniform measure. The main results are presented in Section 11.

Finally, there are a great number of open questions on branch groups. Some of them are listed in the final part of the paper; we hope that they will stimulate the development of the subject.

0.8. *Some notation*

We include zero in the set of natural numbers $\mathbb{N} = \{0, 1, 2, \dots\}$. The set of positive integers is denoted by $\mathbb{N}_+ = \{1, 2, \dots\}$.

Expressions as $((1,2)(3,4,5))$ listing nontrivial cycles are used to describe permutations in the group of symmetries $\mathsf{Sym}(n)$ of n elements.

We want all the group actions to be on the right. Thus we conjugate as follows,

$$g^h = h^{-1}gh,$$

and we write

$$[g,h] = g^{-1}h^{-1}gh = g^{-1}g^h.$$

The commutator subgroup of G is denoted by $[G,G]$ and the Abelianization $G/[G,G]$ by G^{ab}.

Let the group A act on the right on the group H through $\alpha : A \to \mathsf{Aut}(H)$. We define the *semidirect product* $G = H \rtimes_\alpha A$ as the group whose elements are the ordered pairs from the set $H \times A$ and the operation is given by

$$(h,a)(g,b) = \bigl(hg^{(a^{-1})\alpha}, ab\bigr).$$

After the identification $(h,1) = h$ and $(1,a) = a$ we see that G is a group containing H as a normal subgroup and A as its complement, i.e. $HA = G$ and $H \cap A = 1$. Moreover, the conjugation of the elements in H by the elements in A is given by the action α, i.e. $h^a = h^{(a)\alpha}$.

If we start with a group G that has a normal subgroup H with a complement A in G, we say that G is the *internal semidirect product* of H and A. Indeed, $G = H \rtimes A$ where the action of A on H is through conjugation (note that $hagb = hg^{a^{-1}}ab$ for $h, g \in H$ and $a, b \in A$).

Let G and A be groups acting on the set X and the finite set Y, respectively. We define the *permutational wreath product* $G \wr_Y A$ that acts on the set $Y \times X$ (note the change in the order) as follows: let A act on the direct power G^Y on the right by permuting the coordinates of G^Y by

$$\bigl(h^a\bigr)_y = h_{y^{a^{-1}}}, \tag{2}$$

for $h \in G^Y$, $a \in A$, $y \in Y$; then define $G \wr_Y A$ as the semidirect product

$$G \wr_Y A = G^Y \rtimes A$$

obtained through the action of A on G^Y; finally let the wreath product act on the right on the set $Y \times X$ by

$$(y,x)^{ha} = \bigl(y^a, x^{h_y}\bigr),$$

for $y \in Y$, $x \in X$, $h \in G^Y$ and $a \in A$. Note that the equality (2) which represents the action of A on G^Y, also represents conjugation in the wreath product, exactly as we want, and that this wreath product is associative, modulo the necessary natural identifications.

All actions defined by now were right actions. However, we achieved this by introducing inversion at several crucial places, thus introducing left actions through the back door. Another possibility is to let the semidirect product of A and H as above be the group whose elements the ordered pairs in $A \times H$ and define $(a, h)(b, g) = (ab, h^{(b)\alpha} g)$. This works well, but we choose not to do it.

We introduce here the basic notation of growth series. Growth series will be used in Sections 8 and 10.

Let X be a set on which the group G acts, and fix a base point $* \in X$ and a set S that generates G as a monoid. The *growth function* of X is

$$\gamma_{*,G}(n) = \left|\left\{x \in X \mid x = *^{s_1 \cdots s_n} \text{ for some } s_i \in S\right\}\right|.$$

The *growth series* of X is

$$\mathsf{growth}(X) = \sum_{n \geqslant 0} \gamma_{*,G}(n) \hbar^n.$$

Let $V = \bigoplus_{n \geqslant 0} V_n$ be a graded vector space. The *Hilbert–Poincaré* series of V is the formal power series

$$\mathsf{growth}(V) = \sum_{n \geqslant 0} \dim V_n \hbar^n.$$

A preorder $\precsim$ is defined on the set of nondecreasing functions $\mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ by $f \precsim g$ if there exists a positive constant C such that $f(n) \leqslant g(Cn)$, for all n in $\mathbb{R}_{\geqslant 0}$. An equivalence relation $\sim$ is defined by $f \sim g$ if $f \precsim g$ and $g \precsim f$.

Several branch groups are distinguished enough to be given separate notation. They are the first Grigorchuk group $\mathfrak{G}$, the Grigorchuk supergroup $\widetilde{\mathfrak{G}}$, the Gupta–Sidki 3-group $\overline{\overline{\Gamma}}$, the Fabrykowski–Gupta group Γ and the Bartholdi–Grigorchuk group $\overline{\Gamma}$. See Section 1.6 for the definitions.

Acknowledgments

Thanks to D. Segal and P. Harpe for valuable suggestions and to C. Röver for careful reading of the text and numerous corrections and clarifications.

The second author would like to thank the Royal Society of the United Kingdom, University of Birmingham and John Wilson, and also acknowledge the support from a Russian grant "Leading Schools of the Russian Federation", project N 00-15-96107.

The third author thanks F. Guzmán, L. Saloff-Coste and C. Pittet, for their advice, help and valuable conversations.

Part 1. Basic definitions and examples

1. Branch groups and spherically homogeneous trees

1.1. *Algebraic definition of a branch group*

We start with the main definition of the survey, namely the definition of a branch group. The definition is given in purely algebraic terms, emphasizing the subgroup structure of the groups. We give a geometric version of the definition in Section 1.3 in terms of actions on rooted trees. The two definitions are not equivalent and we will say something about the difference later.

DEFINITION 1.1. Let G be a group. We say that G is a *branch group* if there exist two decreasing sequences of subgroups $(L_i)_{i\in\mathbb{N}}$ and $(H_i)_{i\in\mathbb{N}}$ and a sequence of integers $(k_i)_{i\in\mathbb{N}}$ such that $L_0 = H_0 = G$, $k_0 = 1$,

$$\bigcap_{i\in\mathbb{N}} H_i = 1$$

and, for each i,

(1) H_i is a normal subgroup of G of finite index.
(2) H_i is a direct product of k_i copies of the subgroup L_i, i.e. there are subgroups $L_i^{(1)}, \dots, L_i^{(k_i)}$ of G such that

$$H_i = L_i^{(1)} \times \cdots \times L_i^{(k_i)} \tag{3}$$

and each of the factors is isomorphic to L_i.
(3) k_i properly divides k_{i+1}, i.e. $m_{i+1} = k_{i+1}/k_i \geqslant 2$, and the product decomposition (3) of H_{i+1} refines the product decomposition (3) of H_i in the sense that each factor $L_i^{(j)}$ of H_i contains m_{i+1} of the factors of H_{i+1}, namely the factors $L_{i+1}^{(\ell)}$ for $\ell = (j-1)m_{i+1} + 1, \dots, jm_{i+1}$.
(4) Conjugation by the elements in G transitively permute the factors in the product decomposition (3).

The definition implies that branch groups are infinite, but residually finite groups. Note that the subgroups L_i are not normal, but they are subnormal of defect 2.

DEFINITION 1.2. Let G be a branch group. Keeping the notation from the previous definition, we call the sequence of pairs $(L_i, H_i)_{i\in\mathbb{N}}$ a *branch structure* on G.

The branch structure of a branch group is depicted in Figure 1. The branch structure on a group G is not unique, since any subsequence of pairs $(L_{i_j}, H_{i_j})_{j=1}^{\infty}$ is also a branch structure on G.

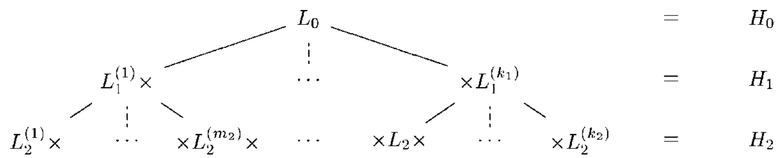


Fig. 1. Branch structure of a branch group.

One can quickly construct examples of branch groups, using infinitely iterated wreath products. For example, let p be a prime and $\mathbb{Z}/p\mathbb{Z}$ act on the set $Y = \{1, \dots, p\}$ by cyclic permutations. Define the permutational wreath product

$$G_n = \underbrace{((\mathbb{Z}/p\mathbb{Z} \wr_Y \cdots) \wr_Y \mathbb{Z}/p\mathbb{Z}) \wr_Y \mathbb{Z}/p\mathbb{Z}}_{n},$$

and let G be the inverse limit $\varprojlim G_n$, where the projections from G_n to G_{n-1} are just the natural restrictions. Since $G = G \wr_Y \mathbb{Z}/p\mathbb{Z}$, G is a branch group.

Similarly, for $m \geqslant 2$, let $\mathsf{Sym}(m)$ be the group of permutations of $Y = \{1, \dots, m\}$, define G_n as the permutational wreath product

$$G_n = \underbrace{((\mathsf{Sym}(m) \wr_Y \cdots) \wr_Y \mathsf{Sym}(m)) \wr_Y \mathsf{Sym}(m)}_{n},$$

and G as the inverse limit $\varprojlim G_n$. Since $G = G \wr_Y \mathsf{Sym}(m)$, G is a branch group.

In the next section we will look at the last group from a geometric point of view. We will also develop some terminology for groups acting on rooted trees that will be used for the second, more geometric, definition of branch groups.

1.2. *Spherically homogeneous rooted trees*

We will define the notion of a spherically homogeneous tree as a set of words ordered by the prefix relation and then make a connection to the graph-theoretical version of the same notion. We find it useful to live in both worlds and use their terminology and notation.

1.2.1. *The trees.* Let

$$\overline{m} = m_1, m_1, m_3, \dots$$

be a sequence of integers with $m_i \geqslant 2$ and let

$$\overline{Y} = Y_1, Y_2, Y_3, \dots$$

be a sequence of alphabets with $|Y_i| = m_i$. A *word of length n* over $\overline{Y}$ is any sequence of letters of the form $w = y_1 y_2 \dots y_n$ where $y_i \in Y_i$ for all i. The unique word of length 0,

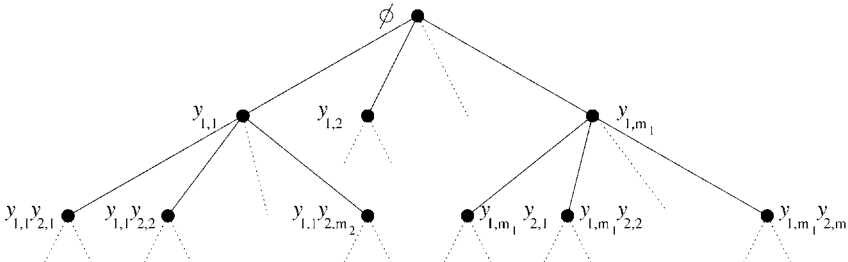


Fig. 2. The tree $\mathcal{T}$.

the *empty word*, is denoted by $\emptyset$. The length of the word u is denoted by $|u|$. Denote the set of words over $\overline{Y}$ by $\overline{Y}^*$. We introduce a partial order on the set of all words over $\overline{Y}$ by the *prefix relation* $\leqslant$. Namely, $u \leqslant v$ if u is an initial segment of the sequence v, i.e. if $u = u_1 \ldots u_n$, $v = v_1 \ldots v_k$, $n \leqslant k$, and $u_i = v_i$, for $i \in \{1, \ldots, n\}$. The partially ordered set of words over $\overline{Y}$, denoted by $\mathcal{T}^{(\overline{Y})}$, is called the *spherically homogeneous tree* over $\overline{Y}$. The sequence $\overline{m}$ is the *sequence of branching indices* of the tree $\mathcal{T}^{(\overline{Y})}$. If there is no room for confusion we denote $\mathcal{T}^{(\overline{Y})}$ by $\mathcal{T}$. For the remainder of the section (and later on) we think of $\overline{Y}$ as being fixed, and let $Y_i = \{y_{i,1}, \ldots, y_{i,m_i}\}$, for $i \in \mathbb{N}_+$. In case all the sets Y_i are equal, say to Y, the tree $\mathcal{T}^{(\overline{Y})}$ is said to be *regular*, and is denoted by $\mathcal{T}^{(Y)}$.

Let us give now the graph-theoretical interpretation of $\mathcal{T}$ and thus justify our terminology. Every word over $\overline{Y}$ represents a vertex in a rooted tree. Namely, the empty word $\emptyset$ represents the *root*, the m_1 one-letter words $y_{1,1}, \ldots, y_{1,m_1}$ represent the m_1 children of the root, the m_2 two-letter words $y_{1,1}y_{2,1}, \ldots, y_{1,1}y_{2,m_2}$ represent the m_2 children of the vertex $y_{1,1}$, etc. More generally, if u is a word over $\overline{Y}$, then the words uy, for y in $Y_{|u|+1}$, of length $|u| + 1$ represent the $m_{|u|+1}$ *children* (or *successors*) of u (see Figure 2).

The graph structure of $\mathcal{T}$ induces a distance function on the set of words by

$$d(u, v) = |u| + |v| - 2|u \wedge v|,$$

where $u \wedge v$ is the longest common prefix of u and v. In particular, the words of length n represent the vertices that are at distance n to the root. Such vertices constitute the *level n* of the tree, denoted by $\mathcal{L}_n^{(\overline{Y})}$ or, when $\overline{Y}$ is assumed, just by $\mathcal{L}_n$. In the terminology of metric spaces, the vertices on the level n are precisely the elements of the *sphere of radius n* with center at the root. In the sequel, we will rarely make any distinction between a word u over $\overline{Y}$, the vertex represented by u and the unique path from the root to the vertex u.

1.2.2. *Tree automorphisms.* A permutation of the words over $\overline{Y}$ that preserves the prefix relation is an *automorphism* of the tree $\mathcal{T}$. From the graph-theoretical point of view an automorphism of $\mathcal{T}$ is just a graph automorphism that fixes the root. We denote the group of automorphisms of $\mathcal{T}$ by $\mathsf{Aut}(\mathcal{T})$. Clearly, the orbits of the action of $\mathsf{Aut}(\mathcal{T})$ on $\mathcal{T}$ are precisely the levels of the tree. The fact that the automorphism group acts transitively on the spheres centered at the root is precisely the reason for which these trees are called spherically homogeneous.

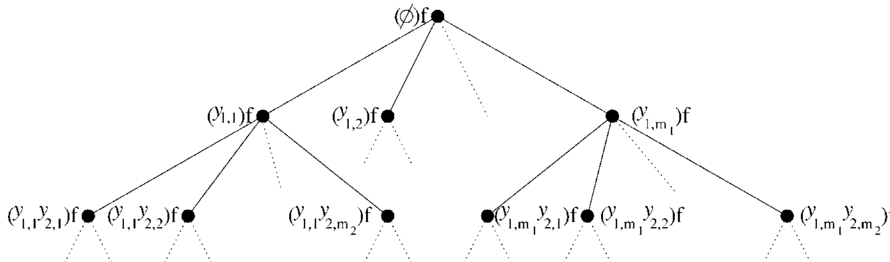


Fig. 3. The automorphism f of $\mathcal{T}$.

Consider an automorphism f of $\mathcal{T}$ and a word u over $\overline{Y}$. The image of u under f is denoted by u^f. For a letter y in $Y_{|u|+1}$ we have $(uy)^f = u^f y'$ where y' is a uniquely determined letter in $Y_{|u|+1}$. Clearly, the induced map $y \mapsto y'$ is a permutation of $Y_{|u|+1}$, we denote this permutation by $(u)f$ and we call it the *vertex permutation* of f at u. If we denote the image of y under $(u)f$ by $y^{(u)f}$, we have

$$(uy)^f = u^f y^{(u)f}, \tag{4}$$

and this easily extends to

$$(y_1 y_2 \dots y_n)^f = y_1^{(\emptyset)f} y_2^{(y_1)f} \dots y_n^{(y_1 y_2 \dots y_{n-1})f}. \tag{5}$$

Any tuple $((u)g)_{u\in\overline{Y}^*}$, indexed by the words u over $\overline{Y}$, where the entry $(u)g$ is a permutation of the alphabet $Y_{|u|+1}$, determines an automorphism g of $\mathcal{T}$ given by

$$(y_1 y_2 \dots y_n)^g = y_1^{(\emptyset)g} y_2^{(y_1)g} \dots y_n^{(y_1 y_2 \dots y_{n-1})g}.$$

Therefore, we can think of an automorphism f of $\mathcal{T}$ as the tuple of vertex permutations $((u)f)_{u\in\overline{Y}^*}$ and we can represent the automorphism f on the tree $\mathcal{T}$ by decorating each vertex u in $\mathcal{T}$ by its permutation $(u)f$ (see Figure 3). The decorated tree that represents f is called the *portrait* of f.

The portrait of f gives an intuitively clear picture of the action of f on $\mathcal{T}$, if we can imagine what happens when we perform all the vertex permutations at once. If only finitely many vertex permutations are nontrivial this is not difficult to do.

By using (4), we can easily see that

$$(u)fg = (u)f \circ \left(u^f\right)g \quad \text{and} \quad (u)f^{-1} = \left[\left(u^{f^{-1}}\right)f\right]^{-1}, \tag{6}$$

for all words u over $\overline{Y}$ and automorphisms f and g of $\mathcal{T}$.

We introduce the *shift operator* σ that acts on sequences as follows:

$$\sigma(s_1, s_2, s_3, \dots) = s_2, s_3, s_4, \dots.$$

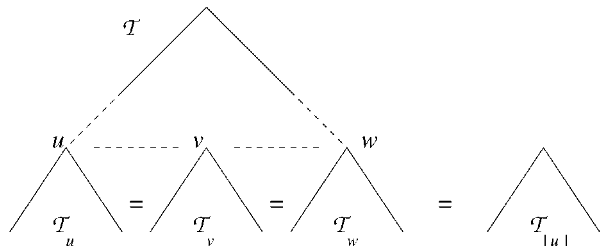


Fig. 4. Canonically isomorphic subtrees and shifted trees.

Acting n times on the sequence of alphabets $\overline{Y}$ gives the shifted sequence of alphabets

$$\sigma^n \overline{Y} = Y_{n+1}, Y_{n+2}, Y_{n+3}, \dots$$

which has the following shifted sequence of branching indices

$$\sigma^n \overline{m} = m_{n+1}, m_{n+2}, m_{n+3}, \dots,$$

and this new sequence of alphabets defines the spherically homogeneous tree $\mathcal{T}^{(\sigma^n \overline{Y})}$. Let u be a word over $\overline{Y}$ of length n and denote by $\mathcal{T}_u^{(\overline{Y})}$ the spherically homogeneous tree that consists of all words over $\overline{Y}$ with prefix u ordered by the prefix relation. It is the subtree of $\mathcal{T}^{(\overline{Y})}$ that is hanging below the vertex u. Clearly, the trees $\mathcal{T}_u^{(\overline{Y})}$ and $\mathcal{T}^{(\sigma^n \overline{Y})}$ are canonically isomorphic under the isomorphism δ_u that deletes the prefix u from the words in $\mathcal{T}_u^{(\overline{Y})}$, and any two trees $\mathcal{T}_u^{(\overline{Y})}$ and $\mathcal{T}_v^{(\overline{Y})}$, where u and v are words over $\overline{Y}$ of the same length, are canonically isomorphic under the isomorphism that deletes the prefix u and replaces it by the prefix v (this isomorphism is just the composition $\delta_u \delta_v^{-1}$). In order to avoid cumbersome notation we denote the tree $\mathcal{T}^{(\sigma^n \overline{Y})}$ by $\mathcal{T}_n$ and the tree $\mathcal{T}_u^{(\overline{Y})}$ by $\mathcal{T}_u$ when $\overline{Y}$ is assumed to be fixed. The previous observations then say that $\mathcal{T}_{|u|}$ and $\mathcal{T}_u$ are canonically isomorphic (see Figure 4).

Let f be an automorphism of $\mathcal{T}$ and u a word over Y. The *section* of f at u (other words in use are *component*, *projection* and *slice*), is the automorphism f_u of $\mathcal{T}_{|u|}$ defined by the vertex permutations

$$(w) f_u = (uw) f, \tag{7}$$

for all words w over $\sigma^{|u|}\overline{Y}$. Therefore, f_u uses the vertex permutations of f at and below the vertex u and assigns them to words over $\sigma^{|u|}\overline{Y}$ in a natural way.

The set $G_u = \{g_u \mid g \in G\}$ of sections at u of the elements in G is called the *section* of G at u. We mention that the section G_u is not necessarily a subgroup of G even if the tree is regular.

It is easy to show, using (5) and (7), that

$$(uv)^f = u^f v^{f_u} \tag{8}$$

for all automorphisms f of $\mathcal{T}$, words u over $\overline{Y}$ and words v over $\sigma^{|u|}\overline{Y}$. Using (6), (7) and (8) we obtain the equalities

$$(fg)_u = f_u g_{u^f} \quad \text{and} \quad \left(f^{-1}\right)_u = \left(f_{u^{f^{-1}}}\right)^{-1}, \tag{9}$$

that hold for all automorphisms f and g of $\mathcal{T}$ and words u over $\overline{Y}$.

Before we move on, let us look at trees and automorphisms from another point of view. The set of infinite paths (rays) from the root $\emptyset$ in $\mathcal{T}$ is called the *boundary* of $\mathcal{T}$ and is denoted by $\partial\mathcal{T}$. We define a metric d on $\partial\mathcal{T}$ by

$$d(r,s) = \begin{cases} \frac{1}{2^{|r\wedge s|}} & \text{if } r \neq s, \\ 0 & \text{if } r = s, \end{cases}$$

for all infinite rays r and s in $\partial\mathcal{T}$. Any automorphism f of $\mathcal{T}$ defines an isometry $\bar{f}$ of the space $\partial\mathcal{T}$ given by

$$(y_1 y_2 y_3 \ldots)^{\bar{f}} = y_1^{(\emptyset)f} y_2^{(y_1)f} y_3^{(y_1 y_2)f} \ldots.$$

Conversely, any isometry $\bar{g}$ of $\partial\mathcal{T}$ defines an automorphism g of $\mathcal{T}$ as follows: u^g is the prefix of $r^{\bar{g}}$ of length $|u|$, where r is any infinite path in $\partial\mathcal{T}$ with prefix u. Therefore, $\mathsf{Aut}(\mathcal{T}) = \mathsf{Isom}(\partial\mathcal{T})$.

Note that the definition of the metric d above was very arbitrary. Given a strictly decreasing sequence $\bar{d} = (d_i)_{i\in\mathbb{N}}$ of positive numbers with limit 0, we could define a metric on $\partial\mathcal{T}$ by $d(r,s) = d_{|r\wedge s|}$ if $r \neq s$, and it can be shown that the topology of the metric space $\partial\mathcal{T}$ is independent of the choice of the sequence $\bar{d}$.

The metric space $(\partial\mathcal{T}, d)$ is a universal model for ultrametric homogeneous spaces as is mentioned in [Gri00] and explained in more details in Proposition 6.2 in [GNS00].

1.2.3. *Level and rigid stabilizers.* We introduce the notions of (rigid) vertex and level stabilizers, as well as the congruence subgroup property.

DEFINITION 1.3. Let G be a group of automorphisms of $\mathcal{T}$. The subgroup $\mathsf{St}_G(u)$ of G, called the *vertex stabilizer* of u in G, consists of those automorphisms in G that fix the vertex u.

For any two automorphisms f and g in $\mathsf{St}_G(u)$, by using (9), we have

$$(fg)_u = f_u g_{u^f} = f_u g_u,$$

so that the map

$$\varphi_u^G : \mathsf{St}_G(u) \to \mathsf{Aut}(\mathcal{T}_{|u|})$$

given by

$$(f)\varphi_u^G = f_u$$

is a homomorphism. We call this homomorphism the *section homomorphism* at u, and we usually avoid the superscript. We denote the image of the section homomorphism φ_u by U_u^G, or just by U_u when G is assumed, and call it the *upper companion* of G at u. Note that the upper companion of G at u is a subgroup of $\mathsf{Aut}(\mathcal{T}_{|u|})$, and is not necessarily a subgroup of G, even in case of a regular tree. Nevertheless, in many important cases in which the tree $\mathcal{T}$ is regular the upper companion groups are equal to G after the canonical identification of the original tree $\mathcal{T}$ with its subtrees.

DEFINITION 1.4. Let G be a group of automorphisms of a regular tree $\mathcal{T}$. The group G is *fractal* if for every vertex u the upper companion group U_u is equal to G (after the tree identifications).

The vertex stabilizers lead to the notion of level stabilizers as follows:

DEFINITION 1.5. Let G be a group of automorphisms of $\mathcal{T}$ and let $\mathsf{St}_G(\mathcal{L}_n)$, called the n-th *level stabilizer* in G, denote the subgroup of G consisting of the automorphisms of $\mathcal{T}$ that fix all the vertices on the level n (and up of course), i.e.

$$\mathsf{St}_G(\mathcal{L}_n) = \bigcap_{u \in \mathcal{L}_n} \mathsf{St}_G(u).$$

The homomorphism

$$\psi_n^G : \mathsf{St}_G(\mathcal{L}_n) \rightarrowtail \prod_{u \in \mathcal{L}_n} U_u \leqslant \prod_{u \in \mathcal{L}_n} \mathsf{Aut}(\mathcal{T}_{|u|})$$

given by

$$(f)\psi_n^G = \big((f)\varphi_u^G\big)_{u \in \mathcal{L}_n} = (f_u)_{u \in \mathcal{L}_n}$$

is an embedding, since the only automorphism that fixes all the vertices at level n and acts trivially on all the subtrees hanging below the level n is the trivial one. In case $n = 1$ we almost always omit the index 1 in ψ_1, and we omit the superscript G for all n. We will see in a moment that the level stabilizers of G have finite index in G. It follows that the same is true for the vertex stabilizers.

We note that the current literature contains several versions of definitions of fractal branch groups. In some of them the sufficient condition from Lemma 1.7 below is used as a definition. One can impose even stronger conditions.

DEFINITION 1.6. Let G be a group of automorphisms of a regular tree $\mathcal{T}$. The group G is *strongly fractal* if it is fractal and the embedding

$$\psi : \mathsf{St}_G(\mathcal{L}_1) \to \prod_{i=1}^{m} G$$

is subdirect, i.e. surjective on each factor.

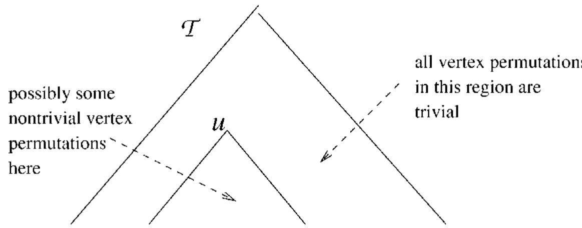


Fig. 5. An automorphism in the rigid stabilizer of u.

LEMMA 1.7. *Let G be a group of automorphisms of $\mathcal{T}^{(p)}$, p a prime, and let all vertex permutations of the automorphisms in G be powers of a fixed cyclic permutation of order p. Then G is fractal if and only if G is strongly fractal.*

DEFINITION 1.8. A group G of tree automorphisms satisfies the *congruence subgroup property* if every finite index subgroup of G contains a level stabilizer $\mathsf{St}_G(\mathcal{L}_n)$, for some n.

We now move to the rigid version of stabilizers:

DEFINITION 1.9. The *rigid vertex stabilizer* of u in G, denoted by $\mathsf{Rst}_G(u)$, is the subgroup of G that consists of those automorphisms of $\mathcal{T}$ that fix all vertices not having u as a prefix.

The automorphisms in $\mathsf{Rst}_G(u)$ must also fix u, and the only vertex permutations that are possibly nontrivial are those corresponding to the vertices in $\mathcal{T}_u$ (see Figure 5).

The rigid stabilizer $\mathsf{Rst}_G(u)$ is also known as the *lower companion* of G at u, denoted by L_u^G, or by L_u when G is assumed. Clearly, the lower companion group at u can be embedded in the upper companion group which is contained in the corresponding section, i.e.

$$L_u \rightarrowtail U_u \subseteq G_u.$$

DEFINITION 1.10. The subgroup of G generated by all the rigid stabilizers of vertices on the level n is the *rigid n-th level stabilizer* and it is denoted by $\mathsf{Rst}_G(\mathcal{L}_n)$.

Clearly, automorphisms in different rigid vertex stabilizers on the same level commute and

$$\mathsf{Rst}_G(\mathcal{L}_n) = \prod_{u\in\mathcal{L}_n} \mathsf{Rst}_G(u).$$

The level stabilizer $\mathsf{St}_G(\mathcal{L}_n)$ and the rigid level stabilizer $\mathsf{Rst}_G(\mathcal{L}_n)$ are normal in $\mathsf{Aut}(\mathcal{T})$. Further, the following relations hold:

$$\prod_{u\in\mathcal{L}_n} L_u = \mathsf{Rst}_G(\mathcal{L}_n) \leqslant \mathsf{St}_G(\mathcal{L}_n) \overset{\psi}{\rightarrowtail} \prod_{u\in\mathcal{L}_n} U_u.$$

In contrast to the level stabilizers, the rigid level stabilizers may have infinite index, and may even be trivial.

Let us restrict our attention, for a moment, to the case when G is the full automorphism group $\mathsf{Aut}(\mathcal{T})$. Clearly, every automorphism of $\mathcal{T}_{|u|}$ is a section of an automorphism of $\mathcal{T}$, since any choice of vertex permutations at and below u is possible for automorphisms of the tree $\mathcal{T}$ that fix u. Therefore, $\mathsf{Aut}(\mathcal{T})_u = \mathsf{Aut}(\mathcal{T}_{|u|}) = \mathsf{Aut}(\mathcal{T}_u)$, i.e. the section is equal to the full automorphism group of the corresponding subtree. Moreover, the section groups are equal to the corresponding upper companion groups. It is also clear that the rigid stabilizer $\mathsf{Rst}_{\mathsf{Aut}(\mathcal{T})}(u)$ is canonically isomorphic to $\mathsf{Aut}(\mathcal{T}_u)$, that the rigid and the level stabilizer of the same level are equal, and ψ is an isomorphism.

Consider the subgroup $\mathsf{Aut}_f(\mathcal{T})$ of automorphisms that have only finitely many nontrivial vertex permutations. The automorphisms in this group are called *finitary*. The group of finitary automorphisms is the union of the chain of subgroups $\mathsf{Aut}_{[n]}(\mathcal{T})$ for $n \in \mathbb{N}$, where $\mathsf{Aut}_{[n]}(\mathcal{T})$ denotes the group of tree automorphisms whose vertex permutations at level n and below are trivial. The group $\mathsf{Aut}_{[n]}(\mathcal{T})$ is canonically isomorphic to the automorphism group $\mathsf{Aut}(\mathcal{T}_{[n]})$ of the finite tree $\mathcal{T}_{[n]}$ that consists of the vertices of $\mathcal{T}$ represented by words no longer than n (level n and above). The group $\mathsf{Aut}(\mathcal{T}_{[n]})$ is isomorphic to the iterated permutational wreath product

$$\mathsf{Aut}(\mathcal{T}_{[n]}) \cong \Big(\big(\cdots\big(\mathsf{Sym}(Y_n) \wr \mathsf{Sym}(Y_{n-1})\big) \wr \cdots\big)\Big) \wr \mathsf{Sym}(Y_1),$$

and its cardinality is $m_1!(m_2!)^{m_1}(m_3!)^{m_1m_2}\dots(m_n!)^{m_1m_2\dots m_{n-1}}$. Also, the equality

$$\mathsf{Aut}(\mathcal{T}) = \mathsf{St}_{\mathsf{Aut}(\mathcal{T})}(\mathcal{L}_n) \rtimes \mathsf{Aut}_{[n]}(\mathcal{T}),$$

holds. As the intersection of all level stabilizers is trivial we see that $\mathsf{Aut}(\mathcal{T})$ is residually finite and, as a corollary, every subgroup of $\mathsf{Aut}(\mathcal{T})$ is residually finite.

We organize some of the remarks we already made in the following diagram:

$$\begin{tikzcd}
\mathsf{Aut}(\mathcal{T}) \arrow[d, phantom, "||"] & \mathsf{Aut}(\mathcal{T}) \arrow[d, phantom, "||"] & \mathsf{Aut}(\mathcal{T}) \arrow[d, phantom, "||"] & \\
\mathsf{St}_{\mathsf{Aut}(\mathcal{T})}(\mathcal{L}_0) \arrow[d, phantom, "\rtimes"] & \mathsf{St}_{\mathsf{Aut}(\mathcal{T})}(\mathcal{L}_1) \arrow[l, hook'] \arrow[d, phantom, "\rtimes"] & \mathsf{St}_{\mathsf{Aut}(\mathcal{T})}(\mathcal{L}_2) \arrow[l, hook'] \arrow[d, phantom, "\rtimes"] & \cdots \arrow[l, hook'] \\
\mathsf{Aut}_{[0]}(\mathcal{T}) & \mathsf{Aut}_{[1]}(\mathcal{T}) \arrow[l, two heads] & \mathsf{Aut}_{[2]}(\mathcal{T}) \arrow[l, two heads] & \cdots \arrow[l, two heads]
\end{tikzcd}$$

The homomorphisms in the bottom row are the natural restrictions, and $\mathsf{Aut}(\mathcal{T})$ is the inverse limit of the inverse system represented by this row. Thus $\mathsf{Aut}(\mathcal{T})$ is a profinite group with topology that coincides with the Tychonoff product topology.

In this topological setting, we recall the *Hausdorff dimension* of a subgroup of $\mathsf{Aut}(\mathcal{T})$:

DEFINITION 1.11 [BS97]. Let $G < \mathsf{Aut}(\mathcal{T})$ be a closed subgroup. Its *Hausdorff dimension* is

$$\limsup_{n\to\infty} \frac{\log |G/\mathsf{St}_G(\mathcal{L}_n)|}{\log |\mathsf{Aut}(\mathcal{T})/\mathsf{St}_{\mathsf{Aut}(\mathcal{T})}(\mathcal{L}_n)|},$$

a real number in [0, 1].

Note that, according to our agreements, the iterated permutational wreath product

$$\prod_{i=1}^{n}(\wr)\,\mathsf{Sym}(m_i) = \Big(\big(\cdots\big(\mathsf{Sym}(m_n)\wr\mathsf{Sym}(m_{n-1})\big)\wr\cdots\big)\Big)\wr\mathsf{Sym}(m_1),$$

naturally acts on $Y_1 \times Y_2 \times \cdots \times Y_n$ which is exactly the set of words of length n. The action is by permutations f that respect prefixes in the sense that

$$|u \wedge v| = \left|u^f \wedge v^f\right|,$$

for all words u and v of length n. This allows us to define the action on the set of all words of length at most n, which is exactly why we may think of $\prod_{i=1}^{n}(\wr)\,\mathsf{Sym}(m_i)$ as being the automorphism group $\mathsf{Aut}(\mathcal{T}_{[n]})$ of the finite tree $\mathcal{T}_{[n]}$.

Since $\prod_{i=1}^{n}(\wr)\,\mathsf{Sym}(m_i)$ acts on the words of length n, the inverse limit $\varprojlim_n \prod_{i=1}^{n}(\wr)\,\mathsf{Sym}(m_i)$ acts on the set of infinite words by isometries, which is one of the interpretations of $\mathsf{Aut}(\mathcal{T})$ we already mentioned.

We agree on a simplified notation concerning the word $u = y_{1,j_1}y_{2,j_2}\dots y_{n,j_n}$ over $\overline{Y}$, the section f_u of the automorphism f and the homomorphism φ_u. We will write sometimes just $u = j_1j_2\dots j_n$ since the sequence of indices $j_1j_2\dots j_n$ uniquely determines and is uniquely determined by the word u. Also, we will write $f_{j_1j_2\dots j_n}$ and $\varphi_{j_1j_2\dots j_n}$ for the appropriate section f_u and section homomorphism φ_u. Actually, we could agree that $Y_i = \{1,2,\dots,m_i\}$, for $i \in \mathbb{N}_+$, in which case the original and the simplified notation are the same.

1.3. *Geometric definition of a branch group*

For the duration of this section we make the important assumption that G is a group of automorphisms of $\mathcal{T}$ that acts transitively on each level of the tree. In this case we say that G acts *spherically transitively*.

It follows easily from (9), that all the vertex stabilizers $\mathsf{St}_G(u)$ corresponding to vertices u on the same level are conjugate in G. Indeed, if h fixes u then h^g fixes u^g and $(h_u)^{g_u} = (h^g)_{u^g}$. This also shows that, in case of a spherically transitive action, the upper companion groups U_u, u a vertex in the level $\mathcal{L}_n$, are conjugate in $\mathsf{Aut}(\mathcal{T}_n)$, we denote by U_n^G, or just by U_n, their isomorphism type, and we call it the *upper companion group* of G at level n or the n-th upper companion:

$$\begin{tikzcd}
L_u = \mathsf{Rst}_G(u) \arrow[r, hook, "\trianglelefteq"] \arrow[d, tail, twoheadrightarrow, "(.)^g"'] & \mathsf{St}_G(u) \arrow[rr, twoheadrightarrow, "\varphi_u"] \arrow[d, tail, twoheadrightarrow, "(.)^g"'] \arrow[drr, phantom, "(h_u)^{g_u} = (h^g)_{u^g}"] & & U_u \arrow[d, tail, twoheadrightarrow, "(.)^{g_u}"] \\
L_{u^g} = \mathsf{Rst}_G(u^g) \arrow[r, hook, "\trianglelefteq"] & \mathsf{St}_G(u^g) \arrow[rr, twoheadrightarrow, "\varphi_{u^g}"'] & & U_{u^g}
\end{tikzcd}$$

Moreover, we note that if the section g_u is trivial then the upper companion groups U_u and U_{u^g} are not only conjugate, but they are equal.

Similarly, the rigid vertex stabilizers of vertices on the same level are also conjugate in G, we denote by L_n^G, or just by L_n, their isomorphism type, and we call it the *lower companion group* of G at level n or the n-th lower companion (see the above diagram). We note that the rigid vertex stabilizer $\mathsf{Rst}_G(u)$ is a normal subgroup in the corresponding vertex stabilizer $\mathsf{St}_G(u)$. Moreover, the lower companion group L_u naturally embeds via the section homomorphisms φ_u in the upper companion group U_u as a normal subgroup. In case of the full automorphism group, we already remarked that this embedding is an isomorphism. In general, this is not true, and we will study more closely the "next best case" when the embedded subgroup has a finite index.

PROPOSITION 1.12. *Let G be a group of automorphisms of $\mathcal{T}$ acting spherically transitively. If $\mathsf{Rst}_G(\mathcal{L}_n)$ has finite index in G for all n, then G is branch group with branch structure $(L_n^G, \mathsf{Rst}_G(\mathcal{L}_n))_{i=1}^{\infty}$.*

DEFINITION 1.13. Let G be a group of automorphisms of $\mathcal{T}$ acting spherically transitively. We say that G is a

(1) *branch group acting on a tree* if all rigid stabilizers of G have finite index in G,
(2) *weakly branch group acting on a tree* if all rigid stabilizers of G are nontrivial (which implies that they are infinite),
(3) *rough group acting on a tree* if all rigid stabilizers of G are trivial.

The branch structure from the previous proposition is not unique, as usual, and we see that for G to be branch group it is enough if we require that each rigid vertex stabilizer group $\mathsf{Rst}_G(u)$, u a vertex in $\mathcal{T}$, has a subgroup $L(u)$ such that $H_n = \prod\{L(u) \mid u \in \mathcal{L}_n\}$ is normal of finite index in G, for all n.

A particularly important type of branch groups is introduced by the following definitions.

DEFINITION 1.14. A fractal branch group G acting on the regular tree $\mathcal{T}^{(m)}$ is a *regular branch group* if there exists a finite index subgroup K of G such that K^m is contained in $(K)\psi$ as a subgroup of finite index. In such a case, we say that G is *branching over K*. We also say that K *geometrically contains* K^m. In case K contains K^m but the index is infinite we say that G is *weakly regular branch* over K.

DEFINITION 1.15. Let G be a regular branch group generated as a monoid by a finite set S, and consider the induced word metric on G. We say G is *contracting* if there exist positive constants $\lambda < 1$ and C such that for every word $w \in S^*$ representing an element of $\mathsf{St}_G(\mathcal{L}_1)$, writing $(w)\psi = (w_1, \ldots, w_m)$, we have

$$|w_i| < \lambda |w| \quad \text{for all } i \in Y, \text{ as soon as } |w| > C. \tag{10}$$

The constant λ is called a *contracting constant*.

In a loose sense, the abstract branch groups are groups that remind us of the full automorphism groups of a spherically homogeneous rooted trees. Any branch group has a natural action on a rooted tree. Indeed, let G be a branch group with branch structure $(L_i, H_i)_{i\in\mathbb{N}}$.

The set of subgroups $\{L_i^{(j)} \mid i \in \mathbb{N},\ j = 1, \dots, k_i\}$ ordered by inclusion forms a spherically homogeneous tree with branching sequence $\overline{m} = m_1, m_2, \dots$, where $m_i = k_i/k_{i-1}$. The group G acts on this set by conjugation and, because of the refinement conditions, the resulting permutation is a tree automorphism (see Figure 1).

The action of the branch group G on the tree determined by the branch structure is not faithful in general. Indeed, it is known that a branch group that satisfies the conditions in Proposition 1.12 is centerless (see [Gri00]). On the other hand, a direct product of a branch group G in the sense of our algebraic definition with a finite group H is still a branch group. If H has nontrivial center, then $G \times H$ is a branch group with nontrivial center.

It would be interesting to understand the nature of the kernel of the action in the passing from an abstract branch group to a group that acts on a tree. In particular, is it correct that this kernel is always in the center (Question 2)? We note that the kernel is trivial in case of a just-infinite branch group (see Section 5).

1.4. *Portraits and branch portraits*

A tree automorphism can be described by its *portrait*, already defined before, and repeated here in the following form:

DEFINITION 1.16. Let f be an automorphism of $\mathcal{T}$. The *portrait* of g is a decoration of the tree $\mathcal{T}$, where the decoration of the vertex u belongs to $\mathsf{Sym}(Y_{|u|+1})$, and is defined inductively as follows: first, there is $\pi_\emptyset \in \mathsf{Sym}(Y_1)$ such that $g = h\pi_\emptyset$ and h stabilizes the first level. This $\pi_\emptyset$ is the label of the root vertex. Then, for all $y \in Y_1$, label the tree below y with the portrait of the section g_y.

The following notion of a branch portrait based on the branch structure of the group in question is useful in some considerations:

DEFINITION 1.17. Let G be a branch group, with branch structure $(L_i, H_i)_{i\in\mathbb{N}}$. The *branch portrait* of $g \in G$ is a decoration of the tree $\mathcal{T}^{(\overline{Y})}$, where the decoration of the root vertex belongs to G/H_1 and the decoration of the vertex $y_1 \dots y_i$ belongs to $L^{(y_1\dots y_i)}/\prod_{y\in Y_{i+1}} L^{(y_1\dots y_i y)}$. Fix once and for all transversals for the above coset spaces. The branch portrait of g is defined inductively as follows: the decoration of the root vertex is $H_1 g$, and the choice of the transversal gives us an element $(g_{y_1})_{y_1\in Y_1}$ of H_1. Decorate then $y_1 \in Y_1$ by $\prod_{y_2\in Y_2} L^{(y_1 y_2)} g_{y_1}$; again the choice of transversals gives us elements $g_{y_1 y_2} \in L^{(y_1 y_2)}$; etc.

There are uncountably many possible branch portraits that use the chosen transversals, even when G is a countable branch group. We therefore introduce the following notion:

DEFINITION 1.18. Let G be a branch group. Its *tree completion* $\overline{G}$ is the inverse limit

$$\varprojlim G/\operatorname{St}_G(n).$$

This is also the closure in $\operatorname{Aut}\mathcal{T}$ of G in the topology given by its action on the tree $\mathcal{T}$.

Note that since $\overline{G}$ is closed in $\mathsf{Aut}(\mathcal{T})$ it is a profinite group, and thus is compact, and totally disconnected. If G has the congruence subgroup property [Gri00], then $\overline{G}$ is also the profinite completion of G.

LEMMA 1.19. *Let G be a branch group and $\overline{G}$ its tree completion. Then Definition* 1.17 *yields a bijection between the set of branch portraits over G and $\overline{G}$.*

Branch portraits are very useful to express, for instance, the lower central series. They appear also, in more or less hidden manner, in most results on growth and torsion.

1.5. *Groups of finite automata and recursively defined automorphisms*

We introduce two more ways to think about tree automorphisms in the case of a regular tree. It is not impossible to extend the definitions to more general cases, but we choose not to do so. Thus for the duration of this section we set $Y = \{1, \ldots, m\}$ and we work with the regular tree $\mathcal{T} = \mathcal{T}^{(Y)} = \mathcal{T}^{(m)}$.

1.5.1. *Recursively defined automorphisms.* Let $X = \{x^{(1)}, \ldots, x^{(n)}\}$ be a set of symbols and F a finite set of finitary automorphisms of $\mathcal{T}$. The following equations

$$x^{(i)} = (w_{i,1}, \ldots, w_{i,m}) f^{(i)}, \quad i \in \{1, \ldots, n\}, \tag{11}$$

where $w_{i,j}$ are words over $X \cup F$ and $f^{(i)}$ are elements in F, define an automorphism of $\mathcal{T}$, still written $x^{(i)}$, for each symbol $x^{(i)}$ in X. The way in which Eqs. (11) define automorphisms recursively is as follows: we interpret the m-tuple $(w_{i,1}, \ldots, w_{i,m})$ as an automorphism fixing the first level and the $w_{i,j}$ are just the sections at j, $j \in \{1, \ldots, m\}$; Eqs. (11) clearly define the vertex permutation at the root for all $x^{(i)}$; if the vertex permutations of all the $x^{(i)}$ are defined for the first k levels then the vertex permutations of all the $w_{i,j}$ are defined for the first k levels, which in turn defines all the vertex permutations of all the $x^{(i)}$ for the first $k+1$ levels.

Every automorphism of $\mathcal{T}$ that can be defined as a member of some set X of recursively defined automorphisms as above is called a *recursively definable* automorphism of $\mathcal{T}$. The set of recursively definable automorphisms of $\mathcal{T}$ forms a subgroup $\mathsf{Aut}_r(\mathcal{T})$ of $\mathsf{Aut}(\mathcal{T})$. The group $\mathsf{Aut}_r(\mathcal{T})$ is a regular branch group which properly contains $\mathsf{Aut}_f(\mathcal{T})$.

When one defines automorphisms recursively it is customary to choose all finitary automorphisms $f^{(i)}$ to be rooted automorphisms (see Definition 1.23). The advantage in that case is that $w_{i,j}$ is exactly the section of $x^{(i)}$ at j. As an example of a recursively defined automorphisms consider

$$b = (a, b)$$

acting on the binary tree $\mathcal{T}^{(2)}$, where $a = ((1,2))$ is the nontrivial rooted automorphism of $\mathcal{T}^{(2)}$. Clearly, the diagram in Figure 6 represents b through its vertex permutations.

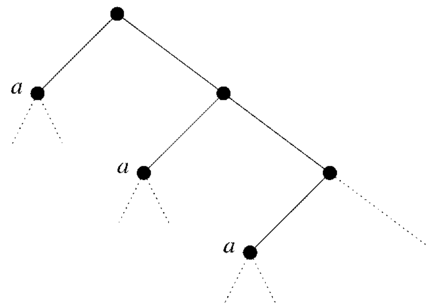


Fig. 6. The recursively defined automorphism b.

It is easy to see that all tree automorphisms are recursively definable if we extend our definition to allow infinite sets X. Indeed

$$g_u = (g_{u1}, \ldots, g_{um})(u)g, \quad u \in Y^*,$$

defines recursively g and all of its sections.

1.5.2. *Groups of finite automata.* Since we want to define automata that behave like tree automorphisms we need automata that transform words rather then recognize them, i.e. we will be working with transducers. The fact that we want our automata to preserve lengths and permute words while preserving prefixes strongly suggests the choices made in the following definition.

DEFINITION 1.20. A *synchronous invertible finite transducer* is a quadruple $T = (Q, Y, \tau, \lambda)$ where

(1) Q is a finite set (set of *states* of T),
(2) Y is a finite set (the *alphabet* of T),
(3) τ is a map $\tau : Q \times Y \to Q$ (the *transition function* of T), and
(4) λ is a map $\lambda : Q \times Y \to Y$ (the *output function* of T) such that the induced map $\lambda_q : Y \to Y$ obtained by fixing a state q is a permutation of Y, for all states $q \in Q$.

If T is a synchronous invertible finite transducer and q a state in Q we sometimes give q a distinguished status, call it the *initial state* and define the *initial synchronous invertible finite transducer* T_q as the transducer T with initial state q.

We say just (initial) transducer in the sequel rather then (initial) synchronous invertible finite transducer.

It is customary to represent transducers by directed labelled graphs where Q is the set of vertices and there exists an edge from q_0 to q_1 if and only if $\tau(q_0, y) = q_1$, for some $y \in Y$, in which case the edge is labelled by $y|\lambda(q_0, y)$. The diagram in Figure 7 gives an example.

Informally speaking, given an initial transducer T_q and an input word w over Y we start at the vertex q and we travel through the graph by reading w one letter at a time and following the values of the transition function. Thus if we find ourselves at the state q'

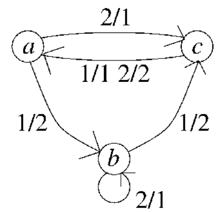


Fig. 7. An example of a transducer.

and we read the letter y we move to the state $\tau(q', y)$ by following the edge labelled by $y|\lambda(q', y)$. In the same time, we write down an output word, one letter at a time, simply by writing down the letters after the vertical bar in the labels of the edges we used in our journey.

More formally, given an initial transducer T_q we define recursively the maps $\tau_q : Y^* \to Q$ and $\lambda_q : Y^* \to Y^*$ as follows:

$$\begin{aligned}
\tau_q(\emptyset) &= q, \\
\tau_q(wy) &= \tau\big(\tau_q(w), y\big) && \text{for } w \in Y^*, \\
\lambda_q(\emptyset) &= \emptyset, \\
\lambda_q(wy) &= \lambda_q(w)\lambda\big(\tau_q(w), y\big) && \text{for } w \in Y^*.
\end{aligned}$$

It is not difficult to see that λ_q, the *output function* of the initial transducer T_q, represents an automorphism of $\mathcal{T}$. The set of all tree automorphisms $\mathsf{Aut}_{ft}(\mathcal{T})$ that can be realized as output functions of some initial transducer forms a subgroup of $\mathsf{Aut}(\mathcal{T})$ and this subgroup is a regular branch group sitting properly between the group of finitary and the group of recursively definable automorphisms of $\mathcal{T}$.

If we allow infinitely many states, then every automorphism g of $\mathcal{T}$ can be realized by an initial transducer. Indeed, we may define the set of states Q to be the set of sections of g, i.e. $Q = \{g_u \mid u \in Y^*\}$ and

$$\tau(g_u, y) = g_{uy} \quad \text{and} \quad \lambda(g_u, y) = y^{(u)g}.$$

We could index the states by the vertices in $\mathcal{T}$, but by indexing them by the sections of g we see that an automorphism g can be defined by a finite initial transducer if it has only finitely many distinct sections. The converse is also true.

PROPOSITION 1.21. *An automorphism of $\mathcal{T}$ is the output function of some initial transducer if and only if it has finitely many distinct sections.*

Every transducer T defines a group G_T of tree automorphisms generated by the initial transducers of T (one for each state). Groups that are defined by transducers are known as *groups of automata*.

Note that the notion of a group of automata is different from the notion of automatic group in the sense of J. Cannon. For more information on groups of automata we refer the reader to [GNS00].

1.6. *Examples of branch groups*

We have already seen a couple of examples of branch groups acting on a regular tree $\mathcal{T}$. Namely, the groups $\mathsf{Aut}_f(\mathcal{T})$, $\mathsf{Aut}_{ft}(\mathcal{T})$, $\mathsf{Aut}_r(\mathcal{T})$ and the full automorphism group $\mathsf{Aut}(\mathcal{T})$.

PROPOSITION 1.22. *Let $\mathcal{T}$ be regular tree and G be any of the groups $\mathsf{Aut}_f(\mathcal{T})$, $\mathsf{Aut}_{ft}(\mathcal{T})$, $\mathsf{Aut}_r(\mathcal{T})$ or $\mathsf{Aut}(\mathcal{T})$. Then G is a regular branch group with $G = G \wr_{Y_1} \mathsf{Sym}(Y_1)$.*

None of the groups in the previous proposition is finitely generated, but the first three are countable. Another example of a regular branch group is, for a permutation group A of Y, the group $\mathsf{Aut}_A(\mathcal{T}^{(Y)})$ that consists of those automorphisms of the regular tree $\mathcal{T}^{(Y)}$ whose vertex permutations come from A. A special case of the last example was mentioned before as the infinitely iterated wreath product of copies of the cyclic group $\mathbb{Z}/p\mathbb{Z}$ and the full automorphism group is another special case.

In the sequel we give some examples of finitely generated branch groups. We make use of rooted and directed automorphisms.

DEFINITION 1.23. An automorphism of $\mathcal{T}$ is *rooted* if all of its vertex permutations that correspond to nonempty words are trivial.

Clearly, the rooted automorphisms are precisely the finitary automorphisms from $\mathsf{Aut}_{[1]}$. A rooted automorphism f just permutes rigidly the m_1 trees $\mathcal{T}_1, \dots, \mathcal{T}_{m_1}$ as prescribed by the root permutation $(\emptyset)f$. It is convenient not to make too much difference between the root vertex permutation $(\emptyset)f$ and the rooted automorphism f defined by it. Therefore, if a is a permutation of Y_1 we also say that a is a rooted automorphism of $\mathcal{T}$. More generally, if a is the vertex permutation of f at u and all the vertex permutations below u are trivial, then we do not distinguish a from the section f_u defined by it, i.e. we write $(f)\varphi_u = f_u = a = (u)f$.

DEFINITION 1.24. Let $\ell = y_1y_2y_3\dots$ be an infinite ray in $\mathcal{T}$. We say that the automorphism f of $\mathcal{T}$ is *directed* along ℓ and we call ℓ the *spine* of f if all vertex permutations along the ray ℓ and all vertex permutations corresponding to vertices whose distance to the ray ℓ is at least 2 are trivial.

In the sequel, we define many directed automorphisms that use the rightmost infinite ray in $\mathcal{T}$ as a spine, i.e. the spine is $m_1m_2m_3\dots$. Therefore, the only vertices that can have a nontrivial permutation are the vertices of the form $m_1m_2\dots m_nj$ where $j \neq m_{n+1}$. Note that directed automorphisms fix the first level, i.e. their root vertex permutation is trivial.

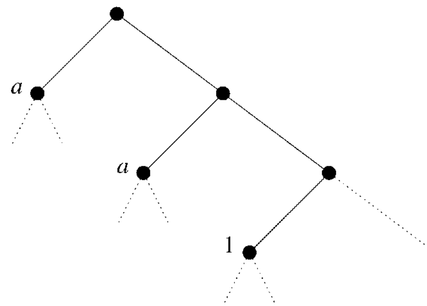


Fig. 8. The directed automorphism b.

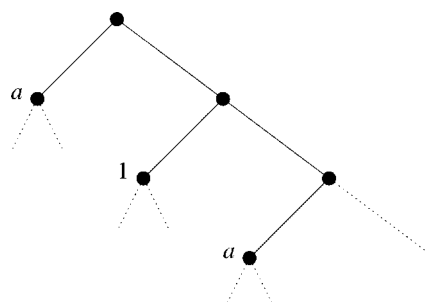


Fig. 9. The directed automorphism c.

1.6.1. *The first Grigorchuk group* $\mathfrak{G}$. A description of the first Grigorchuk group, denoted $\mathfrak{G}$, appeared for the first time in 1980 in [Gri80]. Since then, the group $\mathfrak{G}$ has been used as an example or counter-example in many nontrivial situations.

The group $\mathfrak{G}$ acts on the rooted binary tree $\mathcal{T}^{(2)}$ and it is generated by the four automorphisms a, b, c and d defined below. The automorphism a is the only possible rooted automorphism $a = ((1,2))$ that permutes rigidly the two subtrees below the root. Parts of the portraits along the spine of the generators b, c, and d are depicted in Figures 8–10. We implicitly assume that the patterns that are visible in the diagrams repeat indefinitely along the spine, i.e. along the rightmost ray.

Another way to define the directed generators of $\mathfrak{G}$ is by the following recursive definition:

$$b = (a, c), \quad c = (a, d), \quad d = (1, b).$$

It is clear from this recursive definition that $\mathfrak{G}$ can also be defined as a group of automata.

The group $\mathfrak{G}$ is a 2-group, has a solvable word problem and intermediate growth (see [Gri84]). The best known estimates of the growth of the first Grigorchuk group are given by the first author in [Bar98,Bar01] (see also [Leo98b,Leo00]). The subgroup structure of $\mathfrak{G}$ is a subject of many articles (see [Roz96,BG02] and Section 7) and it turns out that $\mathfrak{G}$ has finite width. An infinite set of defining relations is given by I. Lysionok

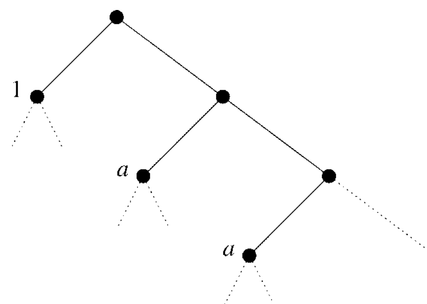


Fig. 10. The directed automorphism d.

in [Lys85], and the second author shows that this system is minimal in [Gri99]. The conjugacy problem is solved by Y. Leonov in [Leo98a] and A. Rozhkov in [Roz98]. A detailed exposition of many of the known properties of $\mathfrak{G}$ is included in the book [Har00] by P. de la Harpe. Another exposition (in Italian) appears in [CMS01].

Most properties of $\mathfrak{G}$, in one way or another, follow from the following

PROPOSITION 1.25. *The group $\mathfrak{G}$ is a regular branch group over the subgroup $K = \langle [a, b] \rangle^{\mathfrak{G}}$.*

PROOF. The first step is to prove that K has finite index in $\mathfrak{G}$. We check that a^2, b^2, c^2, d^2, bcd, $(ad)^4$ are relators in $\mathfrak{G}$. It follows that $B = \langle b \rangle^{\mathfrak{G}}$ has index 8, since $\mathfrak{G}/B$ is $\langle aB, dB \rangle$, a dihedral group of order 8. Consequently $B = \langle b, b^a, b^{ad}, b^{ada} \rangle$. Now B/K is $\langle bK \rangle$, of order 2, so K has index 2 in B and thus 16 in $\mathfrak{G}$.

Then we consider $L = \langle [b, d^a] \rangle^{\mathfrak{G}}$ in K. A simple computation gives $([b, d^a])\psi = ([a, b], 1)$, so $(L)\psi = K \times 1$, and we get $(K)\psi \geqslant (L \times L^a)\psi = K \times K$. $\square$

The index of K in $\mathfrak{G}$ is 16 and the index of $K \times K$ in $(K)\psi$ is 4.

1.6.2. *The second Grigorchuk group.* The second Grigorchuk group was introduced in the same paper as the first one [Gri80]. It acts on the 4-regular tree $\mathcal{T}^{(4)}$ and it is generated by the cyclic rooted automorphism $a = ((1, 2, 3, 4))$ and the directed automorphism b whose portrait is given in Figure 11.

A recursive definition of b is

$$b = (a, 1, a, b).$$

The second Grigorchuk group is not investigated nearly as thoroughly as the first one. It is finitely generated, infinite, residually finite and centerless. In addition, it is not finitely presented, and is a torsion group with a solvable word problem. More on this group can be found in [Vov00].

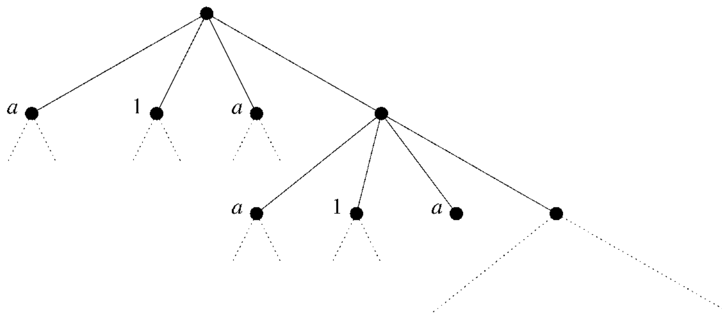


Fig. 11. The directed automorphism b in the second Grigorchuk group.

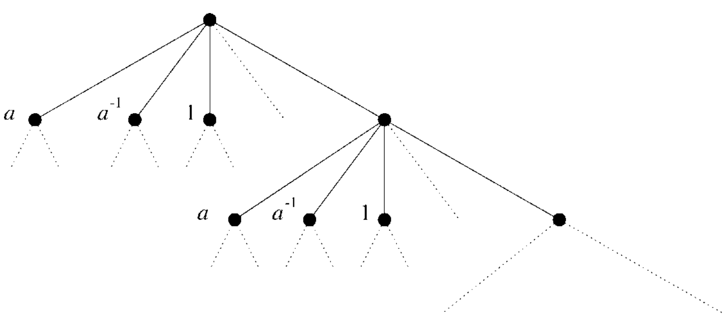


Fig. 12. The directed automorphism b in the Gupta–Sidki groups.

1.6.3. *Gupta–Sidki p-groups.* The first Gupta–Sidki p-groups were introduced in [GS83a]. For odd prime p, we define the rooted automorphism $a = ((1, 2, \ldots, p))$ and the directed automorphism b of $\mathcal{T}^{(p)}$ whose portrait is given in Figure 12.

The group $G = \langle a, b\rangle$ is a 2-generated p-group with solvable word problem and no finite presentation (consider [Sid87b]). In case $p = 3$ the automorphism group, centralizers and derived groups were calculated by Sidki in [Sid87a].

More examples of 2-generated p-groups along the same lines were constructed by N. Gupta and S. Sidki in [GS83b]. In this paper the directed automorphism b is defined recursively by

$$b = \left(a, a^{-1}, a, a^{-1}, \ldots, a, a^{-1}, b\right).$$

It is shown in [GS83b] that these groups are just-infinite p-groups. Also, every finite p-group is contained in the corresponding Gupta–Sidki infinite p-group.

1.6.4. *Three groups acting on the ternary tree.* We define three groups acting on the ternary tree $\mathcal{T}^{(3)}$. Each of them is 2-generated, with generators a and b, where a is the

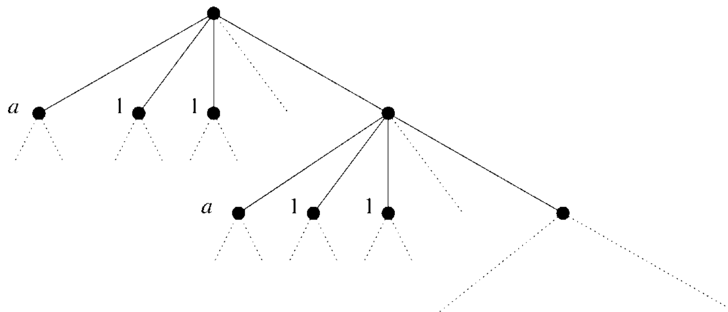


Fig. 13. The directed automorphism b.

rooted automorphism $a = ((1, 2, 3))$ and b is one of the following three directed automorphisms

$$b = (a, 1, b) \quad \text{or} \quad b = (a, a, b) \quad \text{or} \quad b = \big(a, a^2, b\big).$$

The corresponding group $G = \langle a, b\rangle$ is denoted by Γ, $\overline{\Gamma}$ and $\overline{\overline{\Gamma}}$, respectively. The group Γ is called the Fabrykowski–Gupta group and is the first example of a group of intermediate growth that was not constructed by the second author. The construction appears in [FG85], with an incorrect proof, and in [FG91]. The group $\overline{\Gamma}$ is called the Bartholdi–Grigorchuk group and is studied in [BG02]. The article shows that both Γ and $\overline{\Gamma}$ are virtually torsion-free with a torsion-free subgroup of index 3. The group $\overline{\overline{\Gamma}}$ is known as the Gupta–Sidki group, it is the first one of the three to appear in print in [GS83a]. All three groups have intermediate growth. The first author has calculated the central lower series for Γ and $\overline{\overline{\Gamma}}$ (see [Bar00a,Bar00b]). We note here that $\overline{\Gamma}$ is not branch group but only weakly branch group, and the other two are regular branch groups over their commutator subgroups.

1.6.5. *Generalization of the Fabrykowski–Gupta example.* The following examples of branch groups acting on the regular tree $\mathcal{T}^{(m)}$, for $m \geqslant 3$, are studied in [Gri00]. The group $G = \langle a, b\rangle$ is generated by the rooted automorphism $a = ((1, 2, \ldots, m))$ and the directed automorphism b whose portrait is given in Figure 13.

The group G is a regular branch group over its commutator. Moreover, the rigid vertex stabilizers are isomorphic to the commutator subgroup. Clearly, for $m = 3$ we obtain the Fabrykowski–Gupta group Γ. For $m \geqslant 5$, G is just-infinite, and for a prime $m \geqslant 7$ the group G has the congruence subgroup property. The last two results can probably be extended to other branching indices.

1.6.6. *Examples of Peter Neumann.* The following example is constructed in [Neu86]. Let $A = \mathsf{Alt}(6)$ be the alternating group acting on the alphabet $Y = \{1, \ldots, 6\}$. For each pair (a, y) with $y \in Y$ and $a \in \mathsf{St}_A(y)$, define an automorphism $b_{(a,y)}$ of the regular tree $\mathcal{T}^{(6)}$ recursively by

$$b_{(a,y)} = (1, \ldots, 1, b_{(a,y)}, 1, \ldots, 1)a,$$

where the only nontrivial section appears at the vertex y. Let G be the group generated by all these tree automorphisms, i.e.

$$G = \langle \{ b_{(a,y)} \mid y \in Y,\ a \in \mathsf{St}_A(y) \} \rangle.$$

Since G is generated by 6 perfect subgroups (one copy of $\mathsf{Alt}(5)$ for each $y \in Y$) we see that G is perfect. It is also easy to see that if a and a' fix both x and y, with $x \neq y$, then $[b_{(a,x)}, b_{(a',y)}] = [a, a']$. Since

$$\Big\langle \bigcup_{x \neq y \in Y} [\mathsf{St}_A(\{x,y\}), \mathsf{St}_A(\{x,y\})] \Big\rangle = A,$$

we see that G contains A as a rooted subgroup. It follows that $G \cong G \wr_Y A$.

THEOREM 1.26 (P. Neumann [Neu86]). *Let A be a non-Abelian finite simple group acting faithfully and transitively on the set Y. If G is a perfect, residually finite group such that $G \cong G \wr_Y A$ then*

(1) *All nontrivial normal subgroups of G have finite index, i.e. G is just-infinite* (*see Section* 5).
(2) *Every subnormal subgroup of G is isomorphic to a finite direct power of G, but G does not satisfy the ascending chain condition on subnormal subgroups.*
(3) *G is minimal* (*see Section* 5 *again*).

A group that satisfies the conditions of the previous theorem is a regular branch group over itself, acting on the regular tree $\mathcal{T}^{(Y)}$. Furthermore, the only normal subgroups of G are the (rigid) level stabilizers $\mathsf{St}_G(\mathcal{L}_n) = \mathsf{Rst}_G(\mathcal{L}_n) \cong G^{m^n}$.

1.6.7. *The examples of Dan Segal.* For more details on the following examples check [Seg01].

For $i \in \mathbb{N}$, let A_i be a finite, perfect, transitive permutation subgroup of $\mathsf{Sym}(Y_{i+1})$, where Y_{i+1} is a set of m_{i+1} elements. We assume that all stabilizers $\mathsf{St}_{A_i}(y)$ are distinct, for fixed i and $y \in Y_{i+1}$. Let $A_i = \langle a_i^{(1)}, \dots, a_i^{(k)} \rangle$. For $j \in \{1, \dots, k\}$, the diagram in Figure 14 represents the directed automorphism $b_0^{(j)}$ of $\mathcal{T}^{(\overline{Y})}$, which is recursively defined through

$$b_i^{(j)} = (a_{i+1}^j, 1, \dots, 1, b_{i+1}^{(j)}), \quad i \in \mathbb{N},\ j \in \{1, \dots, k\}.$$

We define a group $G_i = \langle A_i \cup B_i \rangle$ where $B_i = \langle b_i^{(1)}, \dots, b_i^{(k)} \rangle$, for $i \in \mathbb{N}$.

Let x be an element in A_0 such that x fixes 1 but not m_1. Then

$$[b_0^{(i)}, (b_0^{(j)})^x] = ([a_1^{(i)}, a_1^{(j)}], 1, \dots, 1),$$

which, by the perfectness of A_1, shows that the rigid stabilizers of the vertices on the first level contain the rooted subgroup A_1. Since $((a_1^{(j)})^{-1}, 1, \dots, 1)b_0^j = (1, \dots, b_1^{(j)})$ we see that the rigid vertex stabilizers of the vertices on the first level are exactly the upper

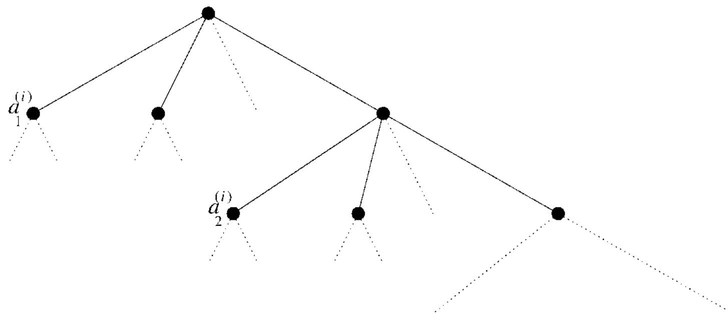


Fig. 14. The directed automorphism in the examples of Segal.

companion groups. Similar claims hold for the other levels and we see that $G = G_0$ is a branch group with

$$\mathsf{St}_G(\mathcal{L}_n) = \mathsf{Rst}_G(\mathcal{L}_n) \cong \prod^{m_1 m_2 \ldots m_n} G_n.$$

In a similar fashion, in case $k = 2$ we can construct a branch group on only three generators as follows. For each $i \in \mathbb{N}$ choose a permutation $\mu_i \in A_i$ of Y_{i+1} that fixes m_{i+1} but μ_i^2 does not fix the symbol $r_{i+1} \in Y_{i+1}$. Then define the directed automorphism b_0 recursively through

$$b_i = \bigl(1, \ldots, 1, a_{i+1}^{(1)}, 1, \ldots, 1, a_{i+1}^{(2)}, 1, \ldots, 1, b_{i+1}\bigr), \quad i \in \mathbb{N},$$

where $a_{i+1}^{(1)}$ is on position r_{i+1} and $a_{i+1}^{(2)}$ is on position $r_{i+1}^{\mu_i}$, for $i \in \mathbb{N}$. Define $G_i = \langle A_i \cup \{b_i\}\rangle$, for $i \in \mathbb{N}$. Then we also get a branch group $G = G_0$ with

$$\mathsf{St}_G(\mathcal{L}_n) = \mathsf{Rst}_G(\mathcal{L}_n) \cong \prod^{m_1 m_2 \ldots m_n} G_n.$$

Different choices of groups A_i together with appropriate actions give various groups with various interesting properties. We list two interesting results from [Seg01], one below and another in Theorem 8.21 that use the above examples:

THEOREM 1.27. *For every collection $\mathcal{S}$ of finite non-Abelian simple groups, there exists a* 63*-generated just-infinite group G whose upper composition factors* (*composition factors of the finite quotients*) *are precisely the members of $\mathcal{S}$. In addition, there exists a* 3*-generated just-infinite group $\overline{G}$ whose noncyclic upper composition factors are precisely the members of $\mathcal{S}$.*

2. Spinal groups

In this section we introduce the class of spinal groups of tree automorphisms. This class is rich in examples of finitely generated branch groups with various exceptional prop-

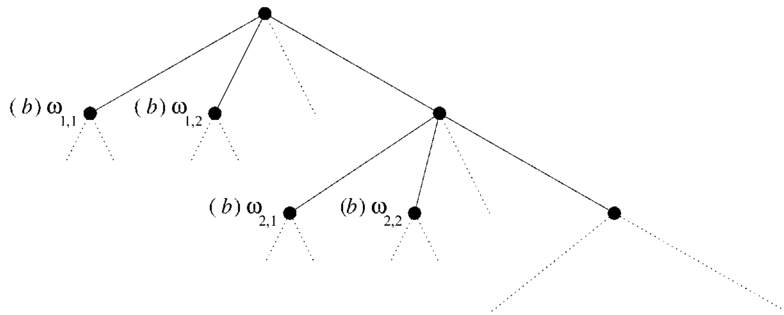


Fig. 15. The directed automorphism b_ω.

erties, constructed by the second author in [Gri80,Gri84,Gri85a], N. Gupta and S. Sidki in [GS83a,GS83b], A. Rozhkov in [Roz86], J. Fabrykowski and N. Gupta in [FG85,FG91], and more recently the first and the third author in [BŠ01,Bar00a,Bar00b,Šun00] and D. Segal in [Seg00]. We will discuss some of these examples in the sequel, after we give a definition that covers all of them. Many of these examples were already presented in Section 1.6.

All examples of finitely generated branch groups that we mentioned by now are spinal groups, except for the examples of P. Neumann from [Neu86] (see Section 1.6). It is not known at present if the groups of P. Neumann are conjugate to spinal groups (Question 4).

2.1. *Construction, basic tools and properties*

2.1.1. *Definition of spinal groups.* Let ω be a triple consisting of a group of rooted automorphisms A_ω, a group B and a doubly indexed family $\overline{\omega}$ of homomorphisms:

$$\omega_{ij} : B \to \mathsf{Sym}(Y_{i+1}), \quad i \in \mathbb{N},\ j \in \{1, \dots, m_i - 1\}.$$

Such a triple is called a *defining triple*. Each $b \in B$ defines a directed automorphism b_ω whose portrait is depicted in the diagram in Figure 15.

Therefore, $B_\omega = \{b_\omega \mid b \in B\}$ is a set of directed automorphisms. We can think of B as acting on the tree $\mathcal{T}$ by automorphisms. We define now the group G_ω, where ω is a defining triple, as the group of tree automorphisms generated by A_ω and B_ω, namely $G_\omega = \langle A_\omega \cup B_\omega \rangle$. We call A_ω the *rooted part*, or the *root group*, and B the *directed part* of G_ω. The family $\overline{\omega}$ is sometimes referred to as the *defining family of homomorphisms*.

Let us define the *shifted triple* $\sigma^r\omega$, for $r \in \mathbb{N}_+$. The triple $\sigma^r\omega$ consists of the group $A_{\sigma^r\omega}$ of rooted automorphisms of $\mathcal{T}_r$ defined by

$$A_{\sigma^r\omega} = \left\langle \bigcup_{j=1}^{m_r-1} (B)\omega_{rj} \right\rangle,$$

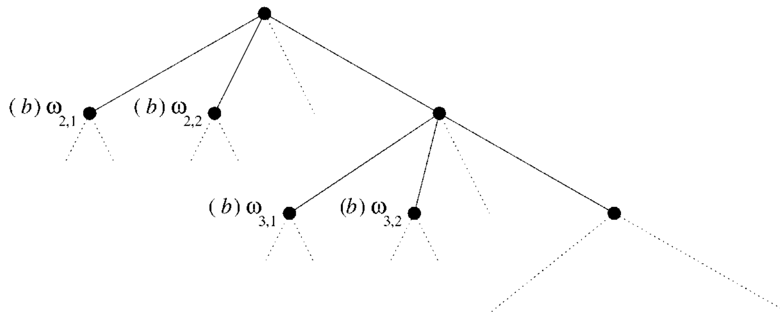


Fig. 16. The directed automorphism $b_{\sigma\omega}$.

the same group B as in ω, and the shifted family $\sigma^r\overline{\omega}$ of homomorphisms

$$\omega_{i+r,j} : B \to \mathsf{Sym}(Y_{i+r+1}), \quad i \in \mathbb{N},\ j \in \{1, \dots, m_{i+r} - 1\}.$$

With the natural agreement that $\sigma^0\omega = \omega$, we see that $\sigma^r\omega$ defines a group $G_{\sigma^r\omega}$ of tree automorphisms of $\mathcal{T}_r$ for each $r \in \mathbb{N}$. Note that the diagram in Figure 16 describes the action of $b_{\sigma\omega}$ on the shifted tree $\mathcal{T}^{(\sigma\overline{Y})} = \mathcal{T}_1$, and that $b_{\sigma\omega}$ is just the section of b_ω at m_1.

We are ready now for the

DEFINITION 2.1. Let $\omega = (A_\omega, B, \overline{\omega})$ be a defining triple. The group

$$G_\omega = \langle A_\omega \cup B_\omega \rangle$$

is called the *spinal group* defined by ω if the following two conditions are satisfied:

(1) *Spherical transitivity condition*: $A_{\sigma^r\omega}$ acts transitively on the corresponding alphabet Y_{r+1}, for all $r \in \mathbb{N}$.

(2) *Strong kernel intersection condition*:

$$\bigcap_{i \geqslant r} \bigcap_{j=1}^{m_i - 1} \mathrm{Ker}(\omega_{ij}) = 1, \quad \text{for all } r.$$

The spherical transitivity condition guarantees that G_ω acts spherically transitively on $\mathcal{T}$, as well as that the same is true for the actions of the shifted groups $G_{\sigma^r\omega}$ on the corresponding shifted trees. Similarly, the strong intersection condition guarantees that the action of B on $\mathcal{T}$ is faithful, and that the same is true for the actions of the shifted groups $G_{\sigma^r\omega}$ on the corresponding shifted trees.

The class of defining triples ω that satisfy the above two conditions will be denoted by Ω. The above considerations indicate that Ω is closed under the shift, i.e. if ω is in Ω then so is any shift $\sigma^r\omega$. This fact is crucial in many arguments involving spinal groups, but we will rarely mention it explicitly.

In the following subsections we introduce the tools and constructions we use in the investigation of the spinal groups along with some basic properties that follow quickly from the given considerations.

2.1.2. *Simple reductions.* The abstract group B is canonically isomorphic to the group of tree automorphisms $B_{\sigma^r\omega}$, for any r, so that we will not make too much difference between them and will frequently omit the index in the notation. Letters like $b, b_1, b', \dots$ are exclusively reserved for the nontrivial elements in B and are called B-letters. Letters like $a, a_1, a', \dots$ are exclusively reserved for the nontrivial elements in $A_{\sigma^r\omega}$, $r \in \mathbb{N}$, and are called A-letters. Note that the groups $A_{\sigma^r\omega}$, $r \in \mathbb{N}$ are not necessarily isomorphic but we omit the index sometimes anyway.

The set $S_\omega = (A_\omega \cup B_\omega) \setminus \{1\}$ is the *canonical generating set* of G_ω. The generators in $A_\omega \setminus \{1\}$ are called *A-generators* and the generators in $B_\omega \setminus \{1\}$ are called *B-generators*. Note that S_ω does not contain the identity and generates G_ω as a monoid, since it is closed under inversion.

We will not be very careful to distinguish the group elements in G_ω from the words in the canonical generators that represent them. At first, it is a sacrifice to the clarity of presentation, but our opinion is that in the long run we only gain by avoiding useless distinctions.

Define the *length of an element g of G_ω* to be the shortest length of a word over S that represents g, and denote this length by $|g|$. There may be more than one word of shortest length representing the same element.

Clearly, G_ω is a homomorphic image of the free product $A_\omega * B_\omega$. Therefore, every g in G_ω can be written in the form

$$[a_0]b_1a_1b_2a_2\dots a_{k-1}b_k[a_k], \tag{12}$$

where the appearances of a_0 and a_k are optional. Relations of the following 4 types:

$$a_1a_2 \to 1, \quad a_3a_4 \to a_5, \quad b_1b_2 \to 1, \quad b_3b_4 \to b_5,$$

that follow from the corresponding relations in A and B are called *simple relations*. A *simple reduction* is any single application of a simple relation from left to right (indicated above by the arrows). Any word of the form (12) is called a *reduced word* and any word can be rewritten in unique reduced form using simple reductions. Among all the words that represent an element g, the ones of shortest length are necessarily reduced, but those that are reduced do not necessarily have the shortest length.

Note that the system of reductions described above is complete, i.e. it always terminates with a word in reduced form and the order in which we apply the reductions does not change the final reduced word obtained by the reduction.

In some considerations one needs to perform cyclic reductions. A reduced word F over S of the form $F = s_1us_2$ for some $s_1, s_2 \in S$, $u \in S^*$, is *cyclically reduced* if the word us_2s_1 obtained from F by a cyclic shift is also reduced. If $F = s_1us_2$ is reduced but not cyclically reduced, the word obtained form us_2s_1 after one application of a simple reduction is said to be obtained from F by one *cyclic reduction*. On the group level the simple cyclic reduction described above corresponds to conjugation by s_1.

Table 1
The maps φ_i associated with the permutation $a = (m_1, \dots, 2, 1)$

	φ_1	φ_2	φ_3	$\cdots$	φ_{m_1-1}	φ_{m_1}
b_ω	$(b)\omega_{1,1}$	$(b)\omega_{1,2}$	$(b)\omega_{1,3}$	$\cdots$	$(b)\omega_{1,m_1-1}$	$b_{\sigma\omega}$
b_ω^a	$(b)\omega_{1,2}$	$(b)\omega_{1,3}$	$(b)\omega_{1,4}$	$\cdots$	$b_{\sigma\omega}$	$(b)\omega_{1,1}$
$b_\omega^{a^2}$	$(b)\omega_{1,3}$	$(b)\omega_{1,4}$	$(b)\omega_{1,5}$	$\cdots$	$(b)\omega_{1,1}$	$(b)\omega_{1,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$b_\omega^{a^{m_1-2}}$	$(b)\omega_{1,m_1-1}$	$b_{\sigma\omega}$	$(b)\omega_{1,1}$	$\cdots$	$(b)\omega_{1,m_1-3}$	$(b)\omega_{1,m_1-2}$
$b_\omega^{a^{m_1-1}}$	$b_{\sigma\omega}$	$(b)\omega_{1,1}$	$(b)\omega_{1,2}$	$\cdots$	$(b)\omega_{1,m_1-2}$	$(b)\omega_{1,m_1-1}$

2.1.3. *Level stabilizers.* In order to simplify the notation we denote the level stabilizer $\mathsf{St}_{G_\omega}(\mathcal{L}_n)$ by $\mathsf{St}_\omega(\mathcal{L}_n)$. In the sequel we often simplify notation by replacing G_ω as a superscript or subscript just by ω, and we do this without warning. Since each element in B fixes the first level, a word u over S represents an element in $\mathsf{St}_\omega(\mathcal{L}_1)$ if and only if the word in A-letters obtained after deleting all B-letters in u represents the identity element.

Further, $\mathsf{St}_\omega(\mathcal{L}_1)$ is the normal closure of B_ω in G_ω, $G_\omega = \mathsf{St}_\omega(\mathcal{L}_1) \rtimes A_\omega$, and $\mathsf{St}_\omega(\mathcal{L}_1)$ is generated by the elements $b_\omega^g = g^{-1} b_\omega g$, for b_ω in B_ω and g in A_ω.

Clearly, $(b_\omega)\psi = ((b)\omega_{1,1}, (b)\omega_{1,2}, \dots, (b)\omega_{1,m_1-1}, b_{\sigma\omega})$. For any a in A_ω, $(b_\omega^a)\psi$ has the same components as $(b_\omega)\psi$ does, but in different positions depending on a. More precisely:

LEMMA 2.2. *For any h in $\mathsf{St}_\omega(\mathcal{L}_1)$, g in A_ω, b in B and $i \in \{1, \dots, m_1\}$, we have*

(1) $(h^g)\varphi_i = (h)\varphi_{i^{g^{-1}}}$.
(2) *The coordinates of $(b^g)\psi$ are: $(b)\omega_{1,j}$ at the coordinate j^g, for $j \in \{1, \dots, m_1 - 1\}$, and b at m_1^g.*

For example, if a is the cyclic permutation $a = ((m_1, \dots, 2, 1))$, the images of $b_\omega^{a^j}$ under various φ_i^ω are given in Table 1.

Since the root group A_ω acts transitively on Y_1 we get all the elements from $A_{\sigma\omega}$ and all the elements from $B_{\sigma\omega}$ in the image of $\mathsf{St}_\omega(\mathcal{L}_1)$ under φ_i, for any i. Therefore, the shifted group $G_{\sigma\omega}$ is precisely the image of the first level stabilizer $\mathsf{St}_\omega(\mathcal{L}_1)$ under any of the section homomorphisms φ_i, $i \in \{1, \dots, m_1\}$.

Let g be an element in G_ω. There are unique elements h in $\mathsf{St}_\omega(\mathcal{L}_n)$ and a in A_ω such that $g = ha$. Clearly, a is the vertex permutation of g at the root, i.e. $a = (\emptyset)g$, and $(h)\psi = (g_1, \dots, g_{m_1})$ where g_i is the section automorphisms of g at the vertex y_i, $i \in \{1, \dots, m_1\}$. Therefore, the section G_{y_i} is contained in the image of $\mathsf{St}_\omega(\mathcal{L}_1)$ under φ_i, and we already established that this image is the shifted group $G_{\sigma\omega}$. Therefore, for all $i = 1, \dots, m_1$,

$$G_{y_i} = U_{y_i} = G_{\sigma\omega},$$

i.e. the sections on the first level, upper companion group and the shifted $G_{\sigma\omega}$ group coincide for spinal groups.

Further, the n-th upper companion group of G_ω is the shifted group $G_{\sigma^n\omega}$ and the homomorphism

$$\psi_n : \mathsf{St}_\omega(\mathcal{L}_n) \to \prod_{i=1}^{m_1 m_2 \ldots m_n} G_{\sigma^n\omega}$$

given by

$$(g)\psi_n = \big((g)\varphi_{1\ldots1}, \ldots, (g)\varphi_{m_1\ldots m_n}\big) = (g_{1\ldots1}, \ldots, g_{m_1\ldots m_n})$$

is an embedding.

We end this section with an easy proposition:

PROPOSITION 2.3. *Every spinal group G_ω is infinite.*

PROOF. The proper subgroup $\mathsf{St}_\omega(\mathcal{L}_1)$ of G_ω maps under φ_1 onto $G_{\sigma\omega}$, which is also spinal. □

The above proof is very simple, so let us take the opportunity here to point out an important feature that is shared by many of the proofs involving spinal groups. The proof does not work with a fixed sequence ω, but rather involves arguments and facts about all the shifts of ω (note the importance of the fact that Ω is closed for shifts). In other words, the group G_ω is always considered together with all of its companions, and the only way we extract some information about some spinal group G_ω is through such a synergic cooperation between the group and its companions. Unavoidably, we also make observations about the companion groups.

2.1.4. *Portraits.* The following construction corresponds to the constructions exhibited in [Gri84,Gri85a,Bar98], but it is more general and allows different modifications and applications.

A *profile* is a sequence $\overline{P} = (P_t)_{t\in\mathbb{N}}$ of sets of automorphisms, called *profile sets*, where $P_t \subseteq \mathsf{Aut}(\mathcal{T}_t)$. We define now a *portrait* of an element f of G_ω with respect to the profile $\mathcal{P}$. The portrait is defined inductively as follows: if f belongs to the profile set P_0 then the portrait of f is the tree that consists of one vertex decorated by f; otherwise the portrait of f is the tree that is decorated by $a = (\emptyset)f$ at the root and has the portraits of the sections $f_1, \ldots, f_{m_1}$, with respect to the shifted sequence $\sigma\overline{P}$ of profile sets, hanging on the m_1 labelled vertices below the root. Therefore, the portrait of f is a subtree (finite or infinite) of the tree on which f acts, its interior vertices are decorated by the corresponding vertex permutations of f, and its leaves are decorated by elements in the chosen profile sets and are equal to the corresponding sections.

For example, if all the profile sets are empty, we obtain the portrait representation of f through its vertex permutations that we already defined (see Figure 3). We sometimes refer to this profile as the *full profile* of f. If P_t is empty for $t = 0, \ldots, r-1$ and equal to $G_{\sigma^r\omega}$ for $t = r$, then the portrait of f is the subtree of $\mathcal{T}$ that consists of the first r levels, the

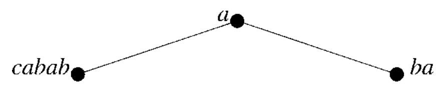


Fig. 17. Depth-1 decomposition of g.

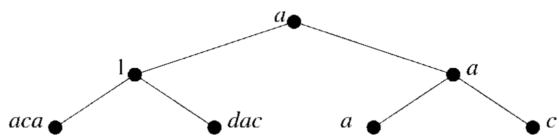


Fig. 18. Depth-2 decomposition of g.

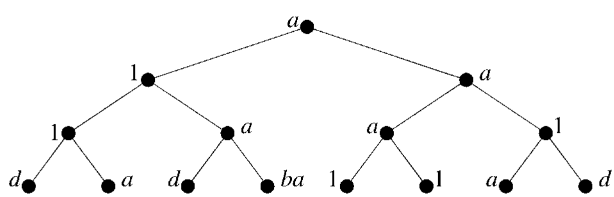


Fig. 19. Depth-3 decomposition of g.

vertices at the levels 0 through $r-1$ are decorated by their vertex permutations and the vertices at level r are decorated by their corresponding sections. Such a portrait is called the *depth-r decomposition* of f. The depth-1, depth-2 and depth-3 decompositions of the element $g = abacadacabadac$ in $\mathfrak{G}$ are given in Figures 17–19. The decompositions can be easily calculated from Table 2 that describes φ_1 and φ_2 (recall the definition of $\mathfrak{G}$ from Section 1.6.1). The calculations follow, in which we identified the elements and their images under the decomposition map ψ:

$$\begin{aligned}
&g = abacadacabadac = b^a cd^a cb^a dc^a a = (cabac1d, ad1daba)a \\
&\quad = (cabab, ba)a, \\
&cabab = cb^a b = (aca, dac), \\
&ba = (a, c)a, \\
&aca = c^a = (d, a), \\
&dac = dc^a a = (d, ba)a, \\
&a = (1, 1)a.
\end{aligned}$$

Let us note that, in general, the leaves of a portrait do not have to be all at the same level, and it is possible that some paths from the root end with a leaf and some are infinite. The various types of portraits carry important information about the elements they represent, which is not surprising, since different elements have different portraits.

Table 2
The table for φ_1 and φ_2 in $\mathfrak{G}$

	φ_1	φ_2
b	a	c
c	a	d
d	1	b
b^a	c	a
c^a	d	a
d^a	b	1

Consider the construction of a portrait more closely. Let f be represented by a reduced word

$$F = [a_0]b_1a_1b_2a_2\dots a_{k-1}b_k[a_k]$$

over S. We rewrite $f = F$ in the form

$$\begin{aligned} f &= b_1^{[a_0^{-1}]} b_2^{([a_0]a_1)^{-1}} \dots b_k^{([a_0]a_1\dots a_{k-1})^{-1}} [a_0]a_1\dots a_{k-1}[a_k] \\ &= b_1^{g_1} b_2^{g_2} \dots b_k^{g_k} g, \end{aligned} \tag{13}$$

where $g_i = ([a_0]a_1\dots a_{i-1})^{-1}$ and $[a_0]a_1\dots a_{k-1}[a_k] = g$ are in A_ω. Next, using the homomorphisms $\omega_{1,j}$, $j \in \{1, \dots, m_1 - 1\}$, and a table similar to Table 1 (but for all possible a) we compute the (not necessarily reduced) words $\overline{F_1}, \dots, \overline{F_{m_1}}$ representing the first level sections $f_1, \dots, f_{m_1}$, respectively. Then we reduce these m_1 words using simple reductions and obtain the reduced words $F_1, \dots, F_{m_1}$. Note that, on the word level, the order in which we perform the reductions is unimportant since the system of simple reductions is complete.

Let us consider any of the words $\overline{F_1}, \dots, \overline{F_{m_1}}$, say $\overline{F_j}$, and study its possible content. Each factor b^g from (13) contributes to the word $\overline{F_j}$ either

- one appearance of the B-letter b, in case $j = m_1^g$ or
- one appearance of the A-letter $(b)\omega_{1,j^{g^{-1}}}$, in case $j \neq m_1^g$ and $b \notin \mathrm{Ker}(\omega_{1,j^{g^{-1}}})$ or
- the empty word, in case $j \neq m_1^g$ and $b \in \mathrm{Ker}(\omega_{1,j^{g^{-1}}})$.

Therefore, the length of any of the reduced words $F_1, \dots, F_{m_1}$ does not exceed k, i.e. does not exceed $(n+1)/2$ where n is the length of F. As a consequence we obtain the following:

LEMMA 2.4. *For any $f \in G_\omega$,*

$$|f_i| \leqslant \big(|f| + 1\big)/2.$$

The *canonical profile* is the profile $\overline{P} = (S_{\sigma^i\omega} \cup \{1\})_{i\in\mathbb{N}}$, whose profile set, on each level, consists of the canonical generators together with the identity. For example, the canonical portrait of $g = abacadacabadac$ is given in Figure 20.

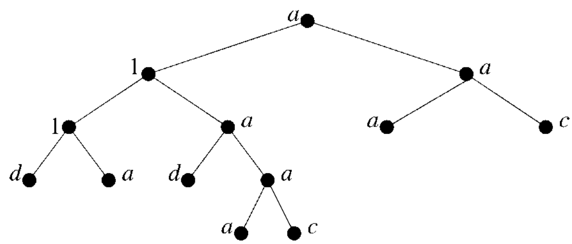


Table 20. The canonical portrait of g.

COROLLARY 2.5. *The depth of the canonical portrait of a word w over S_ω is no larger than* $\lceil \log_2(|w|) \rceil + 1$.

2.2. G *groups*

The G groups are natural generalizations of the first Grigorchuk group $\mathfrak{G}$ from [Gri80] and of the groups G_ω from [Gri84,Gri85a]. The idea of these examples is based on the strong covering property (see Definition 2.6). This section presents an uncountable family of spinal groups of G type. All of them are finitely generated, just-infinite, residually finite, centerless, amenable, not elementary amenable, recursively but not finitely presented torsion groups of intermediate growth (a definition of intermediate growth will come later in Section 10). For proofs see [Gri84,Gri85a], as well as the later sections in this text.

These examples were generalized in [BŠ01,Šun00]. The generalized examples share many of the properties of the Grigorchuk groups mentioned above.

In all the Grigorchuk examples the homomorphisms $\omega_{i,j}$ are trivial for all $j \neq 1$, and we denote $\omega_i = \omega_{i,1}$, for all $i \geqslant 1$. Therefore, the only homomorphisms in the defining family of homomorphisms $\overline{\omega}$ that we need to specify are the homomorphisms in the sequence $\omega_1\omega_2\omega_3\ldots$, which we call the *defining sequence of the triple* ω, and we avoid complications in our notation by simply writing $\overline{\omega} = \omega_1\omega_2\omega_3\ldots$.

2.2.1. *Grigorchuk p-groups.* The Grigorchuk 2-groups, which are a natural generalization of $\mathfrak{G}$, were introduced in [Gri84]. They act on the rooted binary tree $\mathcal{T}^{(2)}$. The rooted group $A = \{1, a\}$ and the group $B = \{1, b, c, d\} = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ are still the same as for $\mathfrak{G}$. There are three nontrivial homomorphisms from B to $\mathsf{Sym}(2) = \{1, a\}$, and we denote them as follows:

$$0 = \begin{pmatrix} 1 & b & c & d \\ 1 & a & a & 1 \end{pmatrix},$$
$$1 = \begin{pmatrix} 1 & b & c & d \\ 1 & a & 1 & a \end{pmatrix},$$
$$2 = \begin{pmatrix} 1 & b & c & d \\ 1 & 1 & a & a \end{pmatrix}.$$

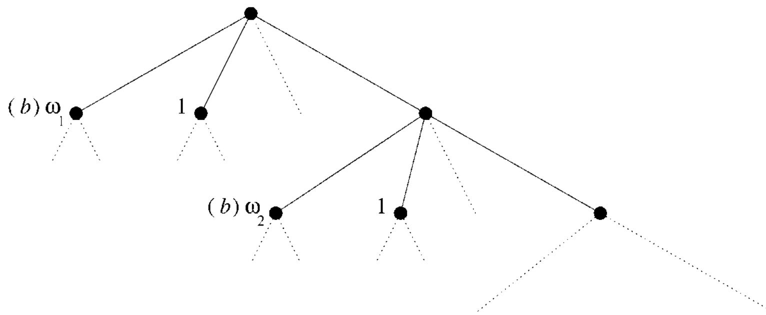


Table 21. The directed automorphism b_ω in G groups.

Note that the vectors representing the elements of P also correspond to the lines in the projective 2-dimensional space over the finite Galois field $\mathbb{F}_p$. A Grigorchuk 2-group is defined by any infinite sequence of homomorphisms $\overline{\omega} = \omega_1\omega_2\omega_3\ldots$, where the homomorphisms $\omega_1, \omega_2, \omega_3, \ldots$ come from the set of homomorphisms $H = \{0, 1, 2\}$ and each homomorphism from H occurs infinitely many times in $\overline{\omega}$. In this setting, the first Grigorchuk group $\mathfrak{G}$ is defined by the periodic sequence of homomorphisms $\overline{\omega} = 012012\ldots$.

Grigorchuk p-groups, introduced in 1985 in [Gri85a], act on the rooted p-ary tree $\mathcal{T}^{(p)}$, for p a prime. The rooted group $A = \langle a\rangle$ is the cyclic group of order p generated by the cyclic permutation $a = ((1, 2, \ldots, p))$ and the group B is the group $\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$. Denote by $\left[\begin{smallmatrix} u \\ v \end{smallmatrix}\right]$ the homomorphism from B to $\mathsf{Sym}(p)$ sending (x, y) to a^{ux+vy} and let P be the set of homomorphisms

$$P = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} p-1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}.$$

A Grigorchuk p-group is defined by any infinite sequence of homomorphisms $\overline{\omega} = \omega_1\omega_2\omega_3\ldots$, where the homomorphisms $\omega_1, \omega_2, \omega_3, \ldots$ come from the set P and each homomorphism from P occurs infinitely many times in $\overline{\omega}$.

Only those Grigorchuk p-groups that are defined by a recursive sequence $\overline{\omega}$ have solvable word problem (see [Gri84]). The conjugacy problem is solvable under the same condition (see Section 3).

2.2.2. G *groups.* The following examples generalize the class of Grigorchuk groups from the previous subsection. A subclass of the class of groups we are about to define is the subject of [BŠ01] and the general case is considered in [Šun00].

We are now going to specify the triples ω that define the groups in the class of G groups. As before, the homomorphisms $\omega_{i,j}$ are trivial for all $j \neq 1$, and we denote $\omega_i = \omega_{i,1}$, for all $i \geqslant 1$. For $b \in B$, the corresponding directed automorphism b_ω is given by the diagram in Figure 21.

For all positive r, we have

$$A_{\sigma^r\omega} = (B)\omega_r$$

and we denote

$$K_i = \mathrm{Ker}(\omega_i).$$

Therefore, for positive r, each of the rooted groups $A_{\sigma^r\omega}$ is a homomorphic image of B.

DEFINITION 2.6. Let G_ω be a spinal group defined by the triple $\omega = (A_\omega, B, \overline{\omega})$ with $\omega_{i,j} = 1$ whenever $j \neq 1$. The group G_ω is a G group if the following *strong covering condition* is satisfied:

$$\bigcup_{i=r}^{\infty} K_i = B, \quad \text{for all } r.$$

The strong kernel intersection condition is equivalent to a statement that $(b)\omega_i$ is non-trivial for infinitely many indices, while the strong covering condition is equivalent to a statement that $(b)\omega_i$ is trivial for infinitely many indices. The class of triples ω that satisfy the above conditions and thus define G groups will be denoted by $\widehat{\Omega}$. Clearly, $\widehat{\Omega}$ is closed for shifting under σ, and the use of this fact is essential but hardly ever emphasized.

Let us show an easy way to build examples of G groups. Start with a group B that has a covering by a family of proper normal subgroups $\{N_\alpha \mid \alpha \in I\}$ of finite index and with trivial intersection, i.e.

$$\bigcup_{\alpha\in I} N_\alpha = B \quad \text{and} \quad \bigcap_{\alpha\in I} N_\alpha = 1.$$

Choose a sequence of normal subgroups $N_1 N_2 N_3 \ldots$, where N_i come from the above family of normal subgroups, such that the strong intersection and covering conditions hold, i.e.

$$\bigcup_{i=r}^{\infty} N_i = B, \quad \text{and} \quad \bigcap_{i=r}^{\infty} N_i = 1, \quad \text{for all } r.$$

For each i, let the factor group B/N_i act transitively and faithfully as a permutation group on some alphabet Y_{i+1}, and let ω_i be the natural homomorphism from B to the permutation group B/N_i followed by the embedding of B/N_i in the symmetric group $\mathsf{Sym}(Y_{i+1})$. Choose a group A_ω and an alphabet Y_1 on which A_ω acts transitively and faithfully. The triple ω that consist of A_ω, B and the sequence of homomorphisms $\overline{\omega} = \omega_1\omega_2\omega_3\ldots$ defines a spinal group G_ω which is a G group acting on the tree $\mathcal{T}^{(\overline{Y})}$. Clearly, the strong intersection and covering conditions are satisfied since $K_i = N_i$, and the spherical transitivity condition is satisfied since $A_{\sigma^i\omega} = B/N_i$.

It is of special interest to consider the case when B is finite. The family of all groups B that have a covering by a finite family of proper normal subgroups can be characterized, according to a theorem of M. Brodie, R. Chamberlain and L.-C. Kappe from [BCK88], as the family of those finite groups that have $\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$ as a factor group, for some prime p (actually the theorem holds in general and characterizes both finite and infinite groups B that have coverings with a finite family of proper normal subgroups). In addition, to make sure that there exist a covering with trivial intersection, we must add the condition that B is not subdirectly irreducible, i.e. the intersection of all nontrivial normal subgroups of B must be trivial.

However, for some reason (aesthetics or something deeper) we might want to restrict our attention only to G groups that act on regular trees. For example, such a case appears in [BŠ01] where the added restriction is that all the factors $B/N_i = A_{\sigma^i\omega}$ are isomorphic to a fixed group A, and they all act in the same way on the same alphabet $\{1, 2, \dots, m\}$. The G groups defined in this way act on the regular tree $\mathcal{T}^{(m)}$. It is fairly easy to construct examples of G groups with this added restrictions in case of an Abelian group B. Any group that is a direct product of proper powers of cyclic groups can be used as B in the above construction, i.e. any group of the form

$$B = (\mathbb{Z}/n_1\mathbb{Z})^{k_1} \times (\mathbb{Z}/n_2\mathbb{Z})^{k_2} \times \cdots \times (\mathbb{Z}/n_s\mathbb{Z})^{k_s}$$

with $k_1, \dots, k_s \geqslant 2$ has a family of normal subgroups of the required type such that all the factors are isomorphic to

$$A = (\mathbb{Z}/n_1\mathbb{Z})^{k_1-1} \times (\mathbb{Z}/n_2\mathbb{Z})^{k_2-1} \times \cdots \times (\mathbb{Z}/n_s\mathbb{Z})^{k_s-1}.$$

In particular, we see that any finite Abelian group can be used in the role of the rooted group A.

Characterizing the family of finite groups B that have a covering with a family of proper normal subgroups with trivial intersection and such that all factors are isomorphic is an interesting problem. The smallest known non-Abelian example so far was communicated to the authors by D. Holt through the Group Pub Forum (see http://www.bath.ac.uk/~masgcs/gpf.html).

Let $B = \langle b_1, b_2, b_3, b_4, b_5, b_6, x_{12}, x_{34}\rangle$ where $b_1, b_2, b_3, b_4, b_5, b_6$ all have order 3 and commute with each other, x_{12} and x_{34} have order 2, commute and

$$b_i^{x_{jk}} = \begin{cases} b_i & \text{if } i \in \{j, k\}, \\ b_i^{-1} & \text{otherwise.} \end{cases}$$

In other words, B is the semidirect product $(\mathbb{Z}/3\mathbb{Z})^6 \rtimes (\mathbb{Z}/2\mathbb{Z})^2$ where $(\mathbb{Z}/2\mathbb{Z})^2 = \langle x_{12}, x_{34}\rangle$, x_{12} fixes the first two and acts by inversion on the last four coordinates of $(\mathbb{Z}/3\mathbb{Z})^6$, x_{34} fixes the middle two and acts by inversion on the other four coordinates and, consequently, $x_{56} = x_{12}x_{34}$ fixes the last two and inverts the first four coordinates.

The following 12 subgroups are normal in B, their intersection is trivial, their union is B, and each factor is isomorphic to the symmetric group $\mathbb{Z}/3\mathbb{Z} \rtimes \mathbb{Z}/2\mathbb{Z} = \mathsf{Sym}(3)$:

$$\begin{array}{ll}
\langle b_1, b_3, b_4, b_5, b_6, x_{12}\rangle, & \langle b_1, b_2, b_3b_4, b_5, b_6, x_{34}\rangle, \\
\langle b_2, b_3, b_4, b_5, b_6, x_{12}\rangle, & \langle b_1, b_2, b_3b_4^2, b_5, b_6, x_{34}\rangle, \\
\langle b_1b_2, b_3, b_4, b_5, b_6, x_{12}\rangle, & \langle b_1, b_2, b_3, b_4, b_5, x_{56}\rangle, \\
\langle b_1b_2^2, b_3, b_4, b_5, b_6, x_{12}\rangle, & \langle b_1, b_2, b_3, b_4, b_6, x_{56}\rangle,
\end{array}$$

$$\langle b_1, b_2, b_3, b_5, b_6, x_{34}\rangle, \qquad \langle b_1, b_2, b_3, b_4, b_5b_6, x_{56}\rangle,$$
$$\langle b_1, b_2, b_4, b_5, b_6, x_{34}\rangle, \qquad \langle b_1, b_2, b_3, b_4, b_5b_6^2, x_{56}\rangle.$$

The smallest example of a non-Abelian group B that has a covering by normal subgroups with trivial intersection is the dihedral group D_6 on 12 elements. Indeed, D_6 is covered by its 3 normal subgroups, call them N_1, N_2 and N_3, of index 2. The intersection of these 3 groups is not trivial, but if we include the center $Z(D_6)$ in the covering, we do get trivial intersection. The corresponding factors are $\mathbb{Z}/2\mathbb{Z} = D_6/N_1 = D_6/N_2 = D_6/N_3$ and the dihedral group on six elements $D_3 = D_6/Z(D_6)$. The first three factors can act as the symmetric group on $\{1, 2\}$, and the last one can act either regularly on $\{1, 2, 3, 4, 5, 6\}$ by the right regular representation or as the symmetric group of $\{1, 2, 3\}$.

We can define even more involved examples of G groups. For example, we can let B itself be the first Grigorchuk group $\mathfrak{G}$. It has 7 subgroups of index 2, and 3 of them cover the group. Since we need trivial intersection we can use the level stabilizers to accomplish this. Therefore, we can easily define a G in which the directed part itself is a G group, for example, $\mathfrak{G}$.

It is clear that in the examples when B is infinite the branching indices of the trees on which the group acts is not bounded. On the other hand, in case of a finite B, the branching indices are bounded by the order of B and they have to be divisors of the order of B, except for the first branching index m_1, which can be arbitrarily large.

2.3. GGS *groups*

The GGS groups (Grigorchuk–Gupta–Sidki groups, the terminology comes from [Bau93]) are natural generalizations of the second Grigorchuk group from [Gri80] and the Gupta–Sidki examples from [GS83a]. They act on a regular tree $\mathcal{T}^{(m)}$, where in most of the examples we present m is prime or a prime power, and they have a special *stabilization* property, namely if an element $g \in G$ is not in the level stabilizer $\mathsf{St}_G(\mathcal{L}_1)$ then the first level sections of the power g^m are either in the stabilizer or are closer to be in the stabilizer than the original element g.

In all examples that we give here, except the general case considered by Bartholdi in [Bar00b], the rooted part $A = \langle a\rangle$ is the cyclic group of order m generated by the permutation $a = ((1, 2, \ldots, m))$. The group $B = \langle b\rangle$ is also a cyclic group of order m. All homomorphisms $\omega_{i,j}$ map the elements from B to powers of a and $\omega_{i,j} = \omega_{i',j}$, for all indices. Therefore, in order to define a spinal group we only need to specify a vector $E = (\varepsilon_1, \ldots, \varepsilon_{m-1})$, where ε_j are integers, and let $(b)\omega_{i,j} = a^{\varepsilon_j}$, for all indices. The group defined by the vector E is denoted by G_E.

Therefore, $G_E = \langle a, b\rangle$ where a is the rooted automorphism defined by the cyclic permutation $a = ((1, 2, 3, \ldots, m))$, and the directed automorphism $b = b_E$ is defined by the diagram in Figure 22.

In order for E to define a spinal group it is necessary and sufficient that $\gcd(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m-1}, m) = 1$. Note that in this situation $\omega = \sigma\omega$, $G_\omega = G_{\sigma\omega}$, etc., and we have fractal groups.

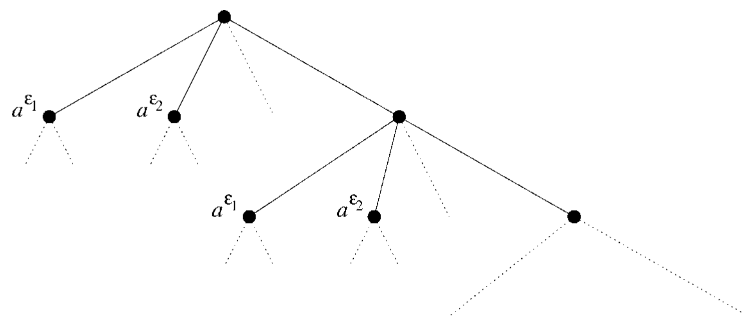


Table 22. The directed automorphism b_E.

2.3.1. GGS *groups with small branching index.* In the case $m = 2$, the only possible nontrivial vector $E = (1)$ defines the infinite dihedral group. This is clear since this spinal group is infinite and generated by two elements of order 2.

There are only three essentially different vectors in case $m = 3$, and they are $(1, 0)$, $(1, 1)$ and $(1, 2)$. The corresponding groups were already introduced as the Fabrykowski–Gupta group Γ, Bartholdi–Grigorchuk group $\overline{\Gamma}$ and Gupta–Sidki group $\overline{\overline{\Gamma}}$.

We have already mentioned one example of GGS group in case $m = 4$ and it is the second Grigorchuk group. The defining vector for the second Grigorchuk group is $E = (1, 0, 1)$.

2.3.2. *Gupta–Sidki examples.* The Gupta–Sidki groups from [GS83a] act on the regular tree $\mathcal{T}^{(p)}$ where p is an odd prime and are defined by the vector $E = (1, -1, 0, 0, \ldots, 0)$. The p-groups introduced in [GS83b] are defined by the vector $E = (1, -1, \ldots, 1, -1)$.

2.3.3. *More examples of* GGS *groups.* The defining vector $E = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m-1})$, where $m = p^n$ is a prime power, defines an infinite 2-generated p-group if and only if

$$\sum_{s \in O_k(m)} \varepsilon_s \equiv 0 \left(\operatorname{mod} p^{k+1}\right),$$

for $k = 0, \ldots, n-1$, where

$$O_k(m) = \left\{p^k, 2p^k, \ldots, \left(p^{n-k} - 1\right)p^k\right\}.$$

The sufficiency in the above claim (in a more general setting) is proved by Gupta and Sidki in [GS84] and the necessity by Vovkivsky in [Vov00]. The latter article also shows that in case the defining vector E does satisfy the condition above, the obtained p-group is just-infinite, not finitely presented branch group.

2.3.4. *General version of* GGS *groups.* One chapter of the Ph.D. dissertation of L. Bartholdi [Bar00b] is devoted to a class of groups that comes as a natural generalization of all of the previous examples of GGS groups. The groups act on the tree $\mathcal{T}^{(m)}$, where m is arbitrary. A defining vector $E = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m-1})$ is a vector of permutations of the alphabet $Y = \{1, 2, \ldots, m\}$ of the tree such that the group of permutations

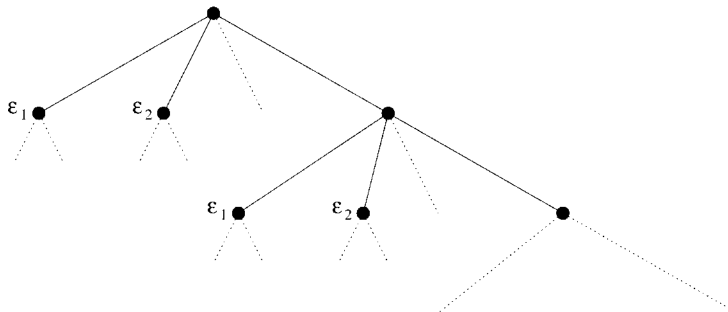


Table 23. The directed automorphism b_E in GGS groups.

$A = \langle \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m-1} \rangle$ acts transitively on Y, and the spinal group G_E is generated by the rooted automorphisms from A together with the directed automorphism b_E (simply written b) defined by the diagram in Figure 23.

To see how these examples fit in our general scheme, note that the group $B = \langle b \rangle$ is a cyclic group of order equal to the least common multiple of the orders of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m-1}$, and the homomorphisms in the triple ω are defined by $(b)\omega_{i,j} = \varepsilon_j$ for all indices.

The first author shows in his dissertation [Bar00b] that unless $m = 2$ (in which case the group is the infinite dihedral group) the GGS groups are finitely generated, residually finite, centerless and not finitely presented groups. He also presents a sufficient and necessary condition for a GGS group to be torsion group (see Section 6).

THEOREM 2.7. *Let G be a* GGS *group. Then G is weakly regular branch on the nontrivial subgroup*

$$K = \left(\mho_{|A|}(G')\right)',$$

where $\mho_{|A|}(G')$ denotes the subgroup of G' generated by the $|A|$-th powers.

SKETCH OF A PROOF. G' contains all elements of the form $[b, b^a]$, which can be written through ψ as

$$\left(a_1, \ldots, [a_i, b], a_{i+1}, \ldots, [b, a_m]\right).$$

Write $n = |A|$, then $\mho_n(G')$ contains all elements of the form $(1, \ldots, [a_i, b]^n, 1, \ldots, [b, a_m]^n)$, and its derived subgroup contains all elements of the form $(1, \ldots, 1, [[b, a]^n, [b, a']^n])$. We therefore have $1 \times \cdots \times 1 \times K \leqslant \psi(K)$.

To show that K is not trivial, we first argue that G' is nontrivial, since it has finite index in an infinite group. Then, since the torsion has unbounded order in G, $\mho_n(G')$ is nontrivial; now this last group has a subgroup mapping onto G (namely $\mathsf{St}_G(\mathcal{L}_N)$ for large N), and therefore cannot be Abelian. $\square$

Part 2. Algorithmic aspects

3. Word and conjugacy problem

Branch groups have good algorithmic properties. A universal algorithm solving the word problem for G and GGS groups and many other branch groups was mentioned in [Gri80] and described in [Gri84] (see also [Gri98,Gri99]). This algorithm is very fast and needs a minimal amount of space and we will describe it below.

However, there are branch groups with unsolvable word problem. For instance, the following claim is proved in [Gri84]

THEOREM 3.1. *The group G_ω has solvable word problem if and only if $\overline{\omega}$ is a recursive sequence.*

Note that the proof in [Gri84] is given for the considered case of 2-groups (see Section 2.2.1), but it can be easily extended to the other cases.

This result inspired the second author to use Kolmogorov complexity to study word problems in branch groups and other classes of groups in [Gri85b].

The generalized word problem (is there an algorithm which for any element $g \in G$ and a finitely generated subgroup $H \leqslant G$ given by a generating set decides if g belongs to H) was considered only recently in [GW01] for the first Grigorchuk group $\mathfrak{G}$ and it was shown that $\mathfrak{G}$ has a solvable generalized word problem.

The solution of the conjugacy problem for the basic examples of branch groups came much later. First, Wilson and Zaleskiĭ solved the conjugacy problem in GGS p-groups, for p an odd prime, by using the notion of Maltcev's conjugacy separability and pro-p methods (see [WZ97]). Slightly later, simultaneously and independently, Leonov in [Leo98a] and Rozhkov in [Roz98] solved the conjugacy problem for $p = 2$. The paper of Rozhkov deals only with the first Grigorchuk group $\mathfrak{G}$, while Leonov considers all 2-groups G_ω from [Gri84]. Also, the results of Leonov are stronger, since upper bounds on the length of conjugating elements are given in terms of depth.

The ideas of Leonov and Rozhkov were further developed in [GW00] in different directions. Still, there are some GGS groups with unsettled conjugacy problem. For example, the results from [GW00] do not apply to the group $\overline{\Gamma}$ (see 1.6.4), which is not branch but only weakly branch.

One of the improvements reached in [GW00] is that the conjugacy problem is solvable not only for the considered groups, but for their subgroups of finite index (note that there are groups with solvable conjugacy problem that contain subgroups of index 2 with unsolvable conjugacy problem, see [CM77]).

Theorem C from [GW00] is the strongest result on the conjugacy problem. The corresponding algorithm in Corollary C in the same article uses the principle of Dirichlet and is quite different from the Leonov–Rozhkov algorithm.

There is also a result on the isomorphism problem for the spinal 2-groups G_ω defined in [Gri84]. Namely, for every sequence $\overline{\omega}$ there are only countably many sequences $\overline{\omega}'$ with $G_\omega \cong G_{\omega'}$. Thus, there are uncountably many finitely generated branch groups.

3.1. *The word problem*

We describe an algorithm that solves the word problem in the case of spinal groups modulo an oracle that has effective knowledge of the defining triple ω in a sense that we make clear below. In fact, the way in which we answer the question "does the word F over S represent the identity in G_ω?" is by constructing the canonical portrait of F and we need tables similar to Table 1, for all possible a, in order to do that.

Let us assume that we have an oracle $\mathcal{O}_\omega(n)$ that has *effective knowledge* of the first n levels of the defining triple ω, meaning that

(1) the multiplication table of B is known,
(2) the permutation groups $A_\omega, \dots, A_{\sigma^n\omega}$ are known, in the sense that we can actually perform the permutations,
(3) the homomorphisms in the first n levels in $\overline{\omega}$ are known, i.e. for all $b \in B$, $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m_i - 1\}$, it is known exactly what permutation in $A_{\sigma^i\omega}$ is equal to $(b)\omega_{i,j}$.

Note that if we have an oracle $\mathcal{O}_\omega(n)$, then we also have oracles $\mathcal{O}_{\sigma^t\omega}(n-t)$, for $t \in \{0, \dots, n\}$.

PROPOSITION 3.2. *Let G_ω be a spinal group and let the oracle $\mathcal{O}_\omega(n)$ have effective knowledge of the first n levels of ω. The word problem in G_ω is solvable, by using the oracle, for all words of length at most $2^{n+1}+1$.*

PROOF. If F is not reduced we can reduce it since we have the multiplication tables for B and A_ω (even in case $n=0$). So assume that F is reduced and let

$$F = [a_0]b_1a_1 \dots a_{k-1}b_k[a_k].$$

Note that if $[a_0]a_1 \dots a_{k-1}[a_k] \neq 1$ in A_ω then F does not stabilize the first level of $\mathcal{T}$ and does not represent the identity.

We prove the claim by induction on n, for all spinal groups simultaneously.

No reduced nonempty word F of length at most 3 represents the identity. This is true because either F does not fix the first level or it represents a conjugate of b, for some $b \in B$. This completes the base case $n=0$.

Let F be reduced, $4 \leqslant |F| \leqslant 2^{n+1}+1$ and $[a_0]a_1 \dots a_{k-1}[a_k] = 1$ in A_ω. Rewrite F as

$$F = b_1^{g_1} \dots b_k^{g_k},$$

where $g_i = ([a_0]a_1 \dots a_{i-1})^{-1}$ and then use the oracle's knowledge of the first level of ω to construct tables similar to Table 1 and calculate the possibly unreduced words $\overline{F_i}$, $i \in \{1, \dots, m_1\}$, that represent the first level sections of F. Each of these words has length no greater than $(|F|+1)/2 \leqslant 2^n + 1$ and we may apply the inductive hypothesis to solve the word problem for each of them by using the oracle $\mathcal{O}_{\sigma\omega}(n-1)$. The word F represents the identity in G_ω if and only if each of the words $F_1, \dots, F_{m_1}$ represents the identity in $G_{\sigma\omega}$ which is clear since $(F)\psi = (F_1, \dots, F_{m_1})$ and ψ is an embedding. $\square$

COROLLARY 3.3. *If the oracle $\mathcal{O}_\omega$ has effective knowledge of all the levels of G_ω, the word problem in G_ω is solvable.*

For many spinal groups the converse is also true. Namely, if the word problem is solvable, one can use the strategy from [Gri84] and construct test-words that would help to recover the kernels of the homomorphisms used in the definition of ω, which is then enough to recover the actual homomorphisms. For example, if we know that G_ω is one of the Grigorchuk 2-groups from [Gri84] and we have an oracle that solves the word problem, we can check the test-words $(ab)^4$, $(ac)^4$ and $(ad)^4$ and find out which one represents the identity, which then tells us that ω_1 was 2, 1 or 0, respectively. Depending on this result, we construct longer test-words that give us information about ω_2, and then longer for ω_3, etc.

3.2. *The conjugacy problem in $\mathfrak{G}$*

Two algorithms which solve the conjugacy problem for regular branch groups satisfying some natural conditions involving length functions are described in [GW00]. The first one accumulates the ideas from [Leo98a,Roz98] and is presented here in the simplest form for the first Grigorchuk group $\mathfrak{G}$. The second one, based on the Dirichlet principle, is quite different and has more potential for applications.

We describe now the first algorithm for the case of $\mathfrak{G}$. Recall that $\mathfrak{G}$ is fractal regular branch group with $K \times K \preceq K \trianglelefteq G$, where K is the normal closure of $[a,b]$. For $g, h \in \mathfrak{G}$ we define

$$Q(g,h) = \{Kf \mid g^f = h\}.$$

Clearly, $Q(g,h) = \emptyset$ if and only if g and h are not conjugate in $\mathfrak{G}$.

THEOREM 3.4. *The conjugacy problem is solvable for $\mathfrak{G}$.*

PROOF.

LEMMA 3.5. *Let $f, g, h \in \mathfrak{G}$ and let*

$$g^f = h. \tag{14}$$

(1) *Let $g = (g_1, g_2)$, $h = (h_1, h_2) \in \mathsf{St}_{\mathfrak{G}}(\mathcal{L}_1)$.*
  (a) *If $f \in \mathsf{St}_{\mathfrak{G}}(\mathcal{L}_1)$ and $f = (f_1, f_2)$ then* (14) *is equivalent to*

$$\begin{cases} g_1^{f_1} = h_1, \\ g_2^{f_2} = h_2. \end{cases}$$

  (b) *If $f \notin \mathsf{St}_{\mathfrak{G}}(\mathcal{L}_1)$ and $f = (f_1, f_2)a$ then* (14) *is equivalent to*

$$\begin{cases} g_1^{f_1} = h_2, \\ g_2^{f_2} = h_1. \end{cases}$$

(2) *Let* $g = (g_1, g_2)a$, $h = (h_1, h_2)a \notin \mathsf{St}_{\mathfrak{G}}(\mathcal{L}_1)$.
(a) *If* $f \in \mathsf{St}_{\mathfrak{G}}(\mathcal{L}_1)$ *and* $f = (f_1, f_2)$ *then* (14) *is equivalent to*

$$\begin{cases} (g_1 g_2)^{f_1} = h_1 h_2, \\ f_2 = g_2 f_1 h_2^{-1}. \end{cases}$$

(b) *If* $f \notin \mathsf{St}_{\mathfrak{G}}(\mathcal{L}_1)$ *and* $f = (f_1, f_2)a$ *then* (14) *is equivalent to*

$$\begin{cases} (g_1 g_2)^{f_1} = h_2 h_1, \\ f_2 = g_1^{-1} f_1 h_2. \end{cases}$$

LEMMA 3.6. *Let* $x = (x_1, y_1)$ *and* $y = (y_1, y_2)$ *be elements of* $\mathfrak{G}$. *Let*

$$\begin{aligned} Q(x, y) &= \{K z_i \mid i \in I\}, \\ Q(x_1, y_1) &= \{K z_j \mid j \in J\}, \\ Q(x_2, y_2) &= \{K z_{j'} \mid j' \in J'\}. \end{aligned}$$

For every $i \in I$ *there exists* $j \in J$ *and* $j' \in J'$ *such that the element* $z_{jj'} = (z_j, z_{j'})$ *is in* $\mathfrak{G}$ *and* $K z_i = K z_{jj'}$.

The meaning of Lemma 3.6 is that if we have $Q(x_1, y_1)$ and $Q(x_2, y_2)$ already calculated, then we can calculate $Q(x, y)$. We just need to check all pairs $(z_j, z_{j'})$ that are in $\mathfrak{G}$ and chose one in each coset of K.

We describe an algorithm that solves the conjugacy problem in $\mathfrak{G}$ by calculating $Q(g, h)$, given g and h.

Split $Q(g, h)$ as

$$Q(g, h) = Q_{1a}(g, h) \cup Q_{1b}(g, h)$$

or as

$$Q(g, h) = Q_{2a}(g, h) \cup Q_{2b}(g, h)$$

according to the cases described in Lemma 3.5.

If we know how to calculate $Q(g_1, h_1)$ and $Q(g_2, h_2)$ for the case 1(a), we could compute $Q_{1a}(g, h)$.

Thus to prove the theorem we need only to show that there is a reduction in length in each of the cases. This is obvious for the cases 1(a) and 1(b) as

$$|g_i| + |h_j| < |g| + |h|,$$

for $i, j \in \{1, 2\}$, except when $|g| = |h| = 1$, which case can be handled directly. For the cases 2(a) and 2(b) we only have

$$|g_1 g_2| + |h_1 h_2| \leqslant |g| + |h|,$$

and equality is possible only if the letter d does not appear in the word representing g nor in the one representing f. In case of an equality we repeat the argument with (g_1g_2, h_1h_2) or (g_1g_2, h_2h_1), depending on the case.

After at most three steps there must be some reduction in the length and we may apply induction on the length. □

4. Presentations and endomorphic presentations of branch groups

Branch groups are probably never finitely presented and this is established for the known examples. There are many approaches one can use to prove that a given branch group cannot have a finite presentation and we will say more about them below, along with some historical remarks.

Every branch group that has a solvable word problem, as all basic examples do, has a recursive presentation.

An important discovery was made by I. Lysionok in 1985, who showed in [Lys85] that the first Grigorchuk group can be given by the presentation (15). No due attention was given to this fact for a long time, until the second author used this presentation in [Gri98] to construct an example of finitely presented amenable but not elementary amenable group, thus answering a question of M. Day from [Day57].

The idea of I. Lysionok was developed in a different direction in [Gri99], where the presentations of the form (15), which are finite presentations modulo the iteration of a single endomorphism, were called L-presentations. It was shown in [Gri99] that some GGS groups (for example, the Gupta–Sidki p-groups, for $p > 3$) have such presentations.

The study of L-presentations was continued, and some aspects were completely resolved in [Bar]. First of all, the notion of L-presentations was extended to the notion of endomorphic presentation in a way that allows several endomorphisms to be included in the presentation. On the basis of this extended definition, a general result was obtained, claiming that all regular branch groups satisfying certain natural requirements have finite endomorphic presentations (see Theorem 4.7 below). It would be interesting to answer the question what branch groups have L-presentations in the sense of [Gri99] and, in particular, resolve the status of the Gupta–Sidki group $\overline{\overline{\Gamma}}$ (Question 7).

We finish the section by providing several examples of groups with finite endomorphic presentations, mostly taken from [Bar].

4.1. *Nonfinite presentability*

In his early work, the second author proved in various ways that $\mathfrak{G}$ is not finitely presented. We review briefly these ideas, since they generalize differently to various other examples.

THEOREM 4.1. *The first Grigorchuk group $\mathfrak{G}$ is not finitely presented.*

FIRST PROOF (from [Gri80]). More details can be found in [Gri99]. Assume in contradiction that $\mathfrak{G}$ is finitely presented, say as $\langle S|R\rangle$. The Reidemeister–Schreier method then gives a presentation of $\mathrm{St}_{\mathfrak{G}}(1)$ with relators $R \cup R^a$; writing for each relator $(r)\psi = (r_1, r_2)$ we obtain a presentation of $\mathfrak{G}$ with relators $\{r_1, r_2\}_{r\in R}$. Now since $|r_i| \leqslant |r|/2$ for all $r \in R$ of length at least 2 (using cyclic reductions), we obtain after enough applications of the above process a presentation of $\mathfrak{G}$ with relations of length 1, i.e. a free group.

This contradicts almost every property of $\mathfrak{G}$: that it is torsion, of subexponential growth, just-infinite, or contains elements $(x, 1)$ and $(1, y)$ that commute. □

SECOND PROOF (from [Gri84]). The set of groups with given generator set $\{a_1, \dots, a_m\}$ is a topological space, with the "weak topology": a sequence (G_i) of groups converges to G if for all radii R the sequence of balls $B_{G_i}(R)$ in the Cayley graphs of the corresponding groups stabilize to the ball $B_G(R)$.

Assume for contradiction that $\mathfrak{G}$ is finitely presented, say as $\langle S|R\rangle$. Then for any sequence $G_i \to \mathfrak{G}$ the groups G_i are quotients of $\mathfrak{G}$ for i large enough. However, if one considers all Grigorchuk groups $\mathfrak{G}_\omega$ from [Gri84], defined through infinite sequences $\overline{\omega} \in \{0, 1, 2\}^{\mathbb{N}}$, one notes that the map $\omega \to \mathfrak{G}_\omega$ is continuous, with the Tychonoff topology on $\{0, 1, 2\}^{\mathbb{N}}$. There are therefore infinite groups converging to $\mathfrak{G}$, which contradicts $\mathfrak{G}$'s just-infiniteness. □

A third proof involves a complete determination of the presentation, and of its Schur multiplier. The results are:

THEOREM 4.2 (Lysionok [Lys85]). *The Grigorchuk group $\mathfrak{G}$ admits the following presentation*:

$$\mathfrak{G} = \big\langle a, c, d \mid \phi^i(a^2), \phi^i(ad)^4, \phi^i(adacac)^4 \ (i \geqslant 0)\big\rangle, \tag{15}$$

where $\phi : \{a, c, d\}^* \to \{a, c, d\}^*$ is defined by $\phi(a) = aca$, $\phi(c) = cd$, $\phi(d) = c$.

THEOREM 4.3 (Grigorchuk [Gri99]). *The Schur multiplier $H_2(\mathfrak{G}, \mathbb{Z})$ of the first Grigorchuk group is $(\mathbb{Z}/2)^\infty$, with basis $\{\phi^i[d, d^a], \phi^i[d^{ac}, d^{aca}]\}_{i\in\mathbb{N}}$.*

Given a presentation $G = F/R$, where F is a free group, the Schur multiplier may be computed as $H_2(G, \mathbb{Z}) = (R \cap [F, F])/[F, R]$ (see [Kar87] or [Bro94] for details). This implies instantly that $\mathfrak{G}$ may not be finitely presented, and moreover that no relation can be omitted.

Another approach that deserves attention is demonstrated in [Gup84]. Namely, certain recursively presented groups are constructed there and the strategy is to build an increasing sequence of normal subgroups $(R_n)_{n\in\mathbb{N}}$ of the free group F whose union is the kernel R of the presentation of the constructed group G as F/R.

The most general result, at least for spinal groups, is given in the third author's dissertation [Šun00], and it follows the "topological" approach from [Gri84], but without the explicit use of the topology.

THEOREM 4.4. *Let $\mathcal{C}$ be a class of groups that is closed under homomorphic images and subgroups* (*of finite index*) *and $\omega = (A_\omega, B, \overline{\omega})$ be a sequence that defines a spinal* *group in $\mathcal{C}$. Further, assume that, for every r, there exists a triple $\eta^{(r)}$ of the form $\eta^{(r)} = (A_{\sigma^r\omega}, B, \overline{\eta})$, where $\overline{\eta}^{(r)}$ is a doubly indexed family of homomorphisms*

$$\eta_{i,j} : B \to \mathsf{Sym}(Y_{j+1}), \quad i \in \{r+1, r+2, \ldots\},\ j \in \{1, \ldots, m_i - 1\},$$

defining a group of tree automorphisms (*not necessarily spinal*) $G_{\eta^{(r)}}$ *that acts on the shifted tree $\mathcal{T}^{(\sigma^r Y)}$ and is not in $\mathcal{C}$. Then, the spinal group G_ω is not finitely presented.*

PROOF. Assume, on the contrary, that G_ω is finitely presented.

Further, assume that the length of the longest relator in the finite presentation of G_ω is no greater than $2^{n+1} + 1$. If ω' is any triple (not necessarily defining a spinal group) that agrees with ω on the first n levels then any word of length no greater than $2^{n+1} + 1$ representing the identity in G_ω represents the identity in $G_{\omega'}$. Thus all relators from the finite presentation of G_ω represent the identity in $G_{\omega'}$, so that $G_{\omega'}$ is a homomorphic image of G_ω, and, therefore, a member of $\mathcal{C}$.

Define ω' so that it agrees with ω on the first n levels and it uses the definition of $\eta^{(n)}$ to define the rest of the levels (just concatenate the definition of $\eta^{(n)}$ to the definition of the first n levels of ω). Since $G_{\omega'}$ is a member of $\mathcal{C}$, so is each of its upper companion groups, including $G_{\eta^{(n)}}$, a contradiction. $\square$

Since the class of torsion groups is closed for subgroups and images and since it is fairly easy to construct triples $\eta^{(r)}$ that define groups containing elements of infinite order, we obtain the following:

COROLLARY 4.5. *No torsion spinal group is finitely presented.*

4.2. *Endomorphic presentations of branch groups*

The recursive structure of branch groups appears explicitly in their presentations by generators and relators, and such presentations have been described since the mid-80's for the first example, the Grigorchuk group.

In this section, we will mainly consider finitely generated, regular branch groups, the reason being that the regularity of presentations becomes much more apparent in these cases. The main result is best formulated in terms of "endomorphic presentations" [Gri99, Bar]:

DEFINITION 4.6. An *endomorphic presentation* is an expression of the form

$$L = \langle S | Q | \Phi | R \rangle,$$

where S is an alphabet (i.e. a set of symbols), $Q, R \subset F_S$ are sets of reduced words (where F_S is the group on S), and Φ is a set of group homomorphisms $\phi : F_S \to F_S$.

L is *finite* if Q, R, S, Φ are finite. It is *ascending* if Q is empty.

L gives rise to a group G_L defined as

$$G_L = F_S \Big/ \Big\langle Q \cup \bigcup_{\phi \in \Phi^*} (R)\phi \Big\rangle^{\#},$$

where $\langle\cdot\rangle^{\#}$ denotes normal closure and Φ^* is the monoid generated by Φ, i.e. the closure of $\{1\} \cup \Phi$ under composition.

As is customary, we identify the endomorphic presentation L and the group G_L it defines.

An endomorphic presentation that has exactly one homomorphism in Φ is called *L-presentation*.

The geometric interpretation of endomorphic presentations in the context of branch groups is the following: one has a finite generating set (S), a finite collection of relations, some of which (R) are related to the branching and therefore can be "moved from one tree level down to the next" by endomorphisms (Φ).

The main result of this section is:

THEOREM 4.7 [Bar]. *Let G be a finitely generated, contracting, regular branch group. Then G has a finite endomorphic presentation. However, G is not finitely presented.*

The motivations in studying group presentations of branch groups are the following:
- They exhibit a regularity that closely parallels the branching structure;
- They allow explicit embeddings of branch groups in finitely presented groups (see Theorem 4.9);
- They may give an explicit basis for the Schur multiplier of branch groups (see Theorem 4.3).

This section gives an outline of the proof of Theorem 4.7. Details may be found in [Bar].

Let G be regular branch on its subgroup K, and fix generating sets S for G and T for K. Without loss of generality, assume $K \leqslant \mathrm{St}_G(\mathcal{L}_1)$, since one may always replace K by $K \cap \mathrm{St}_G(\mathcal{L}_1)$.

First, there exists a finitely presented group $\Gamma = \langle S | Q \rangle$ with subgroups Δ and $\Upsilon = \langle T \rangle$ corresponding to $\mathrm{St}_G(\mathcal{L}_1)$ and K, such that the map $\psi : \mathrm{St}_G(\mathcal{L}_1) \to G^m$ lifts to a map $\Delta \to \Gamma^m$.

The data are summarized in the following diagram:

$$\begin{tikzcd}
 & \Gamma \arrow[r, two heads] \arrow[d, dash] & G \arrow[d, dash] & \\
\Gamma^m \arrow[d, dash] & \Delta \arrow[l, "\tilde{\psi}"'] \arrow[r, two heads] \arrow[d, dash] & \mathrm{St}_G(\mathcal{L}_1) \arrow[r, hook, "\psi"] \arrow[d, dash] & G^m \arrow[d, dash] \\
\Upsilon^m & \Upsilon \arrow[r, two heads] & K & K^m \arrow[l, hook']
\end{tikzcd}$$

Since $\operatorname{Im}\tilde{\psi}$ contains Υ^m, it has finite index in Γ^m. Since Γ^m is finitely presented, $\operatorname{Im}\tilde{\psi}$ too is finitely presented. Similarly, Δ is finitely presented, and we may express $\operatorname{Ker}\tilde{\psi}$ as the normal closure $\langle R_1\rangle^{\#}$ in Δ of those relators in $\operatorname{Im}\tilde{\psi}$ that are not relators in Δ. Clearly R_1 may be chosen to be finite.

We now use the assumption that G is contracting, with constant C. Let R_2 be the set of words over S of length at most C that represent the identity in G. Set $R = R_1 \cup R_2$, which clearly is finite.

We consider T as a set distinct from S, and not as a subset of S^*. We extend each φ_y to a monoid homomorphism $\tilde{\varphi}_y : (S \cup T)^* \to (S \cup T)^*$ by defining it arbitrarily on S.

Assume $\Gamma = \langle S | Q\rangle$, and let $w_t \in S^*$ be a representation of $t \in T$ as a word in S. We claim that the following is an endomorphic presentation of G:

$$G = \big\langle S \cup T \,|\, Q \cup \{t^{-1} w_t\}_{t\in T} \,|\, \{\tilde{\varphi}_y\}_{y\in Y} \,|\, R_1 \cup R_2 \big\rangle. \tag{16}$$

For this purpose, consider the following subgroups Ξ_n of Γ: first $\Xi_0 = \{1\}$, and by induction

$$\Xi_{n+1} = \big\{\gamma \in \Delta \mid (\gamma)\tilde{\psi} \in \Xi_n^m\big\}.$$

We computed $\Xi_1 = \langle R\rangle^{\#}$. Since G acts transitively on the n-th level of the tree, a set of normal generators for Ξ_n is given by $\bigcup_{y\in Y^n}(R)\varphi_{y_1}\cdots\varphi_{y_n}$. We also note that $(\Xi_{n+1})\tilde{\psi} = \Xi_n^m$.

We will have proven the claim if we show $G = \Gamma / \bigcup_{n\geqslant 0}\Xi_n$. Let $w \in \Gamma$ represent the identity in G. After ψ is applied $|w|$ times, we obtain $m^{|w|}$ words that are all of length at most C, that is, they belong to Ξ_1. Then since $(\Xi_{n+1})\tilde{\psi} = \Xi_n^m$, we get $w \in \Xi_{|w|+1}$, and (16) is a presentation of G.

As a bonus, the presentation (16) expresses K as the subgroup of G generated by T.

4.3. *Examples*

We describe here a few examples of branch group presentations. As a first example, let us consider the group $\mathsf{Aut}_f(\mathcal{T}^{(2)})$ of finitary automorphisms of the binary tree.

THEOREM 4.8. *A presentation of* $\mathsf{Aut}_f(\mathcal{T}^{(2)})$ *is given by*

$$T = \big\langle x_0, x_1, \dots \mid x_i^2,\ [x_j, x_k^{x_i}],\ \forall j,k > i\big\rangle,$$

and these relators are independent.

PROOF. The generator x_i is interpreted as the element whose portrait has a single nontrivial label, at level i. The relations are easily checked, and they yield a presentation because they are sufficient to put words in the x_i in wreath product normal form.

Finally, if G_n is the quotient of $\mathsf{Aut}_f(\mathcal{T})$ acting on the n-th level, $H_2(G_n, \mathbb{Z}) = (\mathbb{Z}/2)^{\binom{n+1}{3}}$ by [Bla72] or [Kar87]. $\square$

We now stick to the L-presentation notation, and give presentations for the following examples:

The "first Grigorchuk group". The group $\mathfrak{G}$ admits the ascending L-presentation

$$\mathfrak{G} = \big\langle a, c, d \,\|\, \phi \,|\, a^2, [d, d^a], \big[d^{ac}, (d^{ac})^a\big]\big\rangle,$$

where $\phi : \{a, c, d\}^* \to \{a, c, d\}^*$ is defined by

$$\phi(a) = aca, \quad \phi(c) = cd, \quad \phi(d) = c.$$

These relators are independent, and $H_2(\mathfrak{G}, \mathbb{Z}) = (\mathbb{Z}/2)^\infty$.

The "Grigorchuk supergroup" [BG02]. The group $\widetilde{\mathfrak{G}} = \langle a, \tilde{b}, \tilde{c}, \tilde{d}\rangle$ acting on the binary tree, where a is the rooted automorphism $a = ((1, 2))$ and the other three generators are the directed automorphisms defined recursively by

$$\tilde{b} = (a, \tilde{c}), \quad \tilde{c} = (1, \tilde{d}), \quad \tilde{d} = (1, \tilde{b}),$$

admits the ascending L-presentation

$$\widetilde{\mathfrak{G}} = \big\langle a, \tilde{b}, \tilde{c}, \tilde{d} \,\|\, \tilde{\phi} \,|\, a^2, [\tilde{b}, \tilde{c}], [\tilde{c}, \tilde{c}^a], [\tilde{c}, \tilde{d}^a], [\tilde{d}, \tilde{d}^a], \big[\tilde{c}^{a\tilde{b}}, (\tilde{c}^{a\tilde{b}})^a\big],$$
$$\big[\tilde{c}^{a\tilde{b}}, (\tilde{d}^{a\tilde{b}})^a\big], \big[\tilde{d}^{a\tilde{b}}, (\tilde{d}^{a\tilde{b}})^a\big]\big\rangle,$$

where $\tilde{\phi} : \{a, \tilde{b}, \tilde{c}, \tilde{d}\}^* \to \{a, \tilde{b}, \tilde{c}, \tilde{d}\}^*$ is defined by

$$a \mapsto a\tilde{b}a, \quad \tilde{b} \mapsto \tilde{d}, \quad \tilde{c} \mapsto \tilde{b}, \quad \tilde{d} \mapsto \tilde{c}.$$

These relators are independent, and $H_2(\widetilde{\mathfrak{G}}, \mathbb{Z}) = (\mathbb{Z}/2)^\infty$.

The "Fabrykowski–Gupta group" [FG91,BG02]. The group Γ admits the ascending endomorphic presentation

$$\big\langle a, r \,\|\, \phi, \chi_1, \chi_2 \,|\, a^3, [r^{1+a^{-1}-1+a+1}, a], [a^{-1}, r^{1+a+a^{-1}}][r^{a+1+a^{-1}}, a]\big\rangle,$$

where $\sigma, \chi_1, \chi_2 : \{a, r\}^* \to \{a, r\}^*$ are given by

$$\begin{aligned} \phi(a) &= r^{a^{-1}}, & \phi(r) &= r,\\ \chi_1(a) &= a, & \chi_1(r) &= r^{-1},\\ \chi_2(a) &= a^{-1}, & \chi_2(r) &= r. \end{aligned}$$

These relators are independent, and $H_2(\Gamma, \mathbb{Z}) = (\mathbb{Z}/3)^\infty$.

The "Gupta–Sidki group" [Sid87b]. The Gupta–Sidki group $\overline{\overline{\Gamma}}$ admits the endomorphic presentation

$$\big\langle a, t, u, v \,|\, a^3, t^3, u^{-1}t^a, v^{-1}t^{a^{-1}} \,|\, \phi, \chi \,|\, (tuv)^3, [v, t][vt, u^{-1}tv^{-1}u],$$
$$[t, u]^3[u, v]^3[t, v]^3\big\rangle,$$

where $\phi, \chi : \{t, u, v\}^* \to \{t, u, v\}^*$ are given by

$$\phi : \begin{cases} t \mapsto t, \\ u \mapsto [u^{-1}t^{-1}, t^{-1}v^{-1}]t = u^{-1}tv^{-1}tuvt^{-1}, \\ v \mapsto t[tv, ut] = t^{-1}vutv^{-1}tu^{-1}, \end{cases} \qquad \chi : \begin{cases} t \mapsto t^{-1}, \\ u \mapsto u^{-1}, \\ v \mapsto v^{-1}. \end{cases}$$

These relators are independent, and $H_2(\overline{\overline{\Gamma}}, \mathbb{Z}) = (\mathbb{Z}/3)^\infty$. Note that χ is induced by the automorphism of $\overline{\overline{\Gamma}}$ defined by $a \mapsto a$, $t \mapsto t^{-1}$; however, ϕ does not extend to an endomorphism of $\overline{\overline{\Gamma}}$.

It is precisely for that reason that no ascending endomorphic presentation of $\overline{\overline{\Gamma}}$ is known.

The "Brunner–Sidki–Vieira group" [BSV99]. Consider the group $G = \langle \mu, \tau \rangle$ acting on the binary tree, where μ and τ are defined recursively by

$$\mu = (1, \mu^{-1})a, \quad \tau = (1, \tau)a.$$

Note that G is not branch, but it is weakly branch. Writing $\lambda = \tau\mu^{-1}$, G admits the ascending L-presentation

$$G = \langle \lambda, \tau \| \phi | [\lambda, \lambda^\tau], [\lambda, \lambda^{\tau^3}] \rangle,$$

where ϕ is defined by $\tau \mapsto \tau^2$ and $\lambda \mapsto \tau^2\lambda^{-1}\tau^2$.

The above L-presentation for $\mathfrak{G}$ allowed the second author to answer a question of M. Day [Day57] for the class of finitely presented groups (the question is formulated in [CFP96] in this special setting):

THEOREM 4.9 (Grigorchuk [Gri98]). *There exists a finitely presented amenable group that is not elementarily amenable.*

Recall that a group G is amenable if it admits a left-invariant finitely-additive measure. Examples include the finite groups, the Abelian groups and all groups obtained from previous examples by short exact sequences and direct limits. The smallest class containing the finite and Abelian groups and closed for the mentioned basic constructions is known as the class of *elementarily amenable* groups.

PROOF. Consider the presentation of the first Grigorchuk group given above, and form the HNN extension H amalgamating $\mathfrak{G}$ with $\phi(\mathfrak{G})$. It is an ascending HNN extension, so H is amenable; and H admits the (ordinary) finite presentation

$$H = \langle a, c, d, t | a^2, [d, d^a], [d^{ac}, d^{aca}], a^t aca, c^t cd, d^t c \rangle. \qquad \square$$

Another presentation of the group H from the previous proof, due to the first author, is given in [CGH99]

$$H = \langle a, t | a^2, TaTatataTatataTataT, (Tata)^8, (T^2ataTat^2aTata)^4 \rangle,$$

where T denotes the inverse of t.

Part 3. Algebraic aspects

5. Just-infinite branch groups

DEFINITION 5.1. A group G is *just-infinite* if it is infinite but all of its proper quotients are finite, i.e. if all of its nontrivial normal subgroups have finite index.

The following simple criterion from [Gri00] characterizes the just-infinite branch groups acting on a tree.

THEOREM 5.2. *Let G be a branch group acting on a tree and let $(L_i, H_i)_{i\in\mathbb{N}}$ be a corresponding branch structure. The following three conditions are equivalent*

(1) *G is just-infinite.*
(2) *The Abelianization H_i^{ab} of H_i is finite, for each $i \in \mathbb{N}$.*
(3) *The commutator subgroup H_i' of H_i has finite index in G, for each $i \in \mathbb{N}$.*

The statement is a corollary of the fact that H_i', being characteristic in the normal subgroup H_n, is a normal subgroup of G, for each $i \in \mathbb{N}$, and the following useful lemma that says that weakly branch groups satisfy the following property:

LEMMA 5.3. *Let G be a weakly branch group acting on a tree and let $(L_i, H_i)_{i\in\mathbb{N}}$ be a corresponding branch structure. Then every nontrivial normal subgroup N of G contains the commutator subgroup H_n', for some n depending on N.*

PROOF. Let g be an element in $G \setminus \mathsf{St}_G(\mathcal{L}_1)$ and let $N = \langle g\rangle^G$ be its normal closure in G. Then $g = ha$ for some $h \in \mathsf{St}_{\mathsf{Aut}(\mathcal{T})}(\mathcal{L}_1)$ with decomposition $h = (h_1, \dots, h_{m_1})$ and a a nontrivial rooted automorphism of $\mathcal{T}$. Without loss of generality we may assume that $1^a = m_1$.

For arbitrary elements $\xi, \nu \in L_1$, we define $f, t \in H_1$ by $f = (\xi, 1, \dots, 1)$ and $t = (\nu, 1, \dots, 1)$ and calculate

$$[g, f] = \bigl(\xi, 1, \dots, 1, (\xi^{-1})^{h_1}\bigr),$$
$$\bigl[[g, f], t\bigr] = \bigl([\xi, \nu], 1, \dots, 1\bigr).$$

Since $[[g, f], t]$ is always in $N = \langle g\rangle^G$, we obtain $L_1' \times 1 \times \cdots \times 1 \preceq N$ and, by the spherical transitivity of G, it follows that

$$L_1' \times L_1' \times \cdots \times L_1' = H_1' \preceq N.$$

Thus any normal subgroup of G that contains g also contains H_1'.

Similarly, if g is an element in $\mathsf{St}_G(\mathcal{L}_n) \setminus \mathsf{St}_G(\mathcal{L}_{n+1})$ and N is the normal closure $N = \langle g^G\rangle$, then N contains H_{n+1}'. $\square$

In particular, the above results imply that all finitely generated torsion weakly branch groups are just-infinite branch groups.

The study of just-infinite groups is motivated by their minimality. More precisely, we have the following

THEOREM 5.4 [Gri00]. *Every finitely generated infinite group has a just-infinite quotient.*

Therefore, if $\mathcal{C}$ is a class of groups closed for taking quotients and it contains a finitely generated infinite group, then it contains a finitely generated just-infinite group.

Note that there are non-finitely generated groups that do not have just-infinite quotients, for example, the additive group of rational numbers $\mathbb{Q}$.

It is known (see [Wil71]) that a just-infinite group with nontrivial Baer radical is a finite extension of a free Abelian group of finite rank (recall that the *Baer radical* of the group G is the subgroup of G generated by the cyclic subnormal subgroups of G). Moreover, the only just-infinite group with nontrivial center is the infinite cyclic group $\mathbb{Z}$. Therefore, an Abelian group has a just-infinite quotient if and only if it can be mapped onto $\mathbb{Z}$. In particular, no Abelian torsion group has just-infinite quotients. The last fact is in a sharp contrast with the fact that there are large classes of centerless, torsion, just-infinite, branch groups, for instance, G groups with finite directed part B (see Section 6) and many GGS groups.

DEFINITION 5.5. A group G is *hereditarily just-infinite* if it is residually finite and all of its nontrivial normal subgroups are just-infinite.

We mention that our definition of hereditarily just-infinite group differs from the one given in [Wil00] in that we require residual finiteness. Note that all nontrivial normal subgroups of a group G are just infinite if and only if all subgroups of finite index in G are just infinite. This is true since every subgroup of finite index in G contains a normal subgroup of G of finite index.

Examples of hereditarily just-infinite groups are the infinite cyclic group $\mathbb{Z}$, the infinite dihedral group D_∞ and the projective special linear groups $\mathsf{PSL}(n, \mathbb{Z})$, for $n \geqslant 3$. However, the whole class is far from well understood and described.

The following result from [Gri00], which modifies the result of J. Wilson from [Wil71] (see also [Wil00]), strongly motivates the study of the branch groups.

THEOREM 5.6 (Trichotomy of just-infinite groups). *Let G be a finitely generated just-infinite group. Then exactly one of the following holds*:

(1) *G is a branch group.*

(2) *G has a normal subgroup H of finite index of the form*

$$H = L^{(1)} \times \cdots \times L^{(k)} = L^k,$$

where the factors are copies of a group L, the conjugations by the elements in G transitively permute the factors of H and L has exactly one of the following two properties:

(a) *L is hereditary just-infinite* (*in case G is residually finite*).

(b) *L is simple* (*in case G is not residually finite*).

The proof of this theorem presented in [Gri00] uses only the statement from [Wil71] that every subnormal subgroup in a just-infinite group with trivial Bear radical has a near complement (but this is probably one of the most important facts from Wilson's theory). The proof actually works for any (not necessarily finitely generated) just-infinite group with trivial Baer radical, for instance, just-infinite groups which are not virtually cyclic.

The results of Wilson in [Wil71] (see also [Wil00]) combined with the above trichotomy result show that the following characterization of just-infinite branch groups is possible. Define an equivalence relation on the set of subnormal subgroups of a group G by, $H \sim K$ if the intersection $H \cap K$ has a finite index both in H and in K. The set of equivalence classes of subnormal subgroups, ordered by the order induced by inclusion, forms a Boolean lattice, which, following J. Wilson, is called the *structure lattice* of G.

THEOREM 5.7. *Let G be a just-infinite group. Then G is branch group if and only if it has infinite structure lattice.*

Moreover, in such a case, the structure lattice is isomorphic to the lattice of closed and open subsets of the Cantor set.

On the intuitive level, the just-infinite groups should be considered as "small" groups in contrast to, say, the free or nonelementary hyperbolic groups, which are "large" groups.

There is a rigorous approach to the concept of largeness in groups. Namely, following Pride [Pri80], we say that a group G is *larger* than a group H, and we denote $G \succeq H$, if H has a subgroup of finite index that is a homomorphic image of a subgroup of G of finite index. The groups G and H are *equally large* (or *Pride equivalent*) if $G \succeq H$ and $H \succeq G$. The set of equivalence classes of equally large groups is partially ordered by $\succeq$ and the class of finite groups is the obvious smallest element.

We denote the class of groups equally large to G by $[G]$. A group G is called *minimal* if the only class below $[G]$ is the class of finite groups $[1]$. The *height* of a group G is the height of the class $[G]$ in the ordering, i.e. the length of a maximal chain between $[1]$ and $[G]$. Therefore, the minimal groups are the groups of height 1. Such groups are called *atomic* in [Neu86].

THEOREM 5.8 [GW]. *The first Grigorchuk group $\mathfrak{G}$ and the Gupta–Sidki p-groups are minimal.*

A number of questions about just-infinite groups was asked in [Pri80,EP84]. Positive answer to Problem 5 from [Pri80] (Problem 4′ in [EP84]) that asks if there exist finitely generated just-infinite groups that do not satisfy the ascending chain condition on subnormal subgroups was provided in [Gri84]. Later, P. Neumann constructed in [Neu86] more examples of finitely generated just-infinite regular branch groups answering the same question (and also some other questions) raised by M. Edjvet and S. Pride in [EP84]. In particular, P. Neumann provided negative answer to the question if every finitely generated minimal group is finite-by-D_2-by-finite (here D_2 denotes the class of groups in which every nontrivial subnormal subgroup has finite index). Negative answer to this last question also follows from Theorem 5.8 above.

The question of possible heights of finitely generated just-infinite groups is an interesting one. All hereditarily just-infinite and all infinite simple groups are minimal. It is plausible that the Grigorchuk 2-groups from [Gri84] that are defined by non-periodic sequences have infinite height (see Question 13).

6. Torsion branch groups

As mentioned in the introduction, the most elegant examples of finitely generated infinite torsion groups were constructed within the class of branch groups [Gri80,GS83a]. Therefore, branch groups play important role in problems of Burnside type. At the moment, there is a number of constructions of torsion branch groups. Besides the early works [Ale72,Suš79,Gri80,Gri83,Gri84,GS83a,GS83b,GS84] there are more modern and general constructions [BŠ01,Gri00,Šun00]. Nevertheless, all these constructions follow the same idea of stabilization and length reduction.

Recall that finitely generated torsion branch groups acting on a tree are just-infinite (Theorem 5.2).

We provide a proof here that all G groups with torsion directed group B are themselves torsion groups, and the argument follows the mentioned general scheme. This is an improvement over the result in [BŠ01] that deals only with regular trees and all the root actions (the actions of $A_{\sigma^t\omega}$, for all t) are isomorphic and regular. Thus whenever B is a finitely generated torsion group the corresponding G group is a Burnside group and there are uncountably many non-isomorphic examples. We know that torsion spinal groups cannot have finite presentation (see Corollary 4.5). The results in Theorem 6.9 below show that these groups cannot have finite exponent. Therefore, in the finitely generated case, one of our goals is to give upper bounds on the order of an element depending on its length. This leads to the notion of torsion growth.

Let G be finitely generated infinite torsion group and let S be a finite generating set that generates G as a monoid. For any non-negative real number n, the maximal order of an element of length at most n is finite, and we denote it by $\pi_G^S(n)$. The function π_G^S, defined on the non-negative real numbers, is called the *torsion growth* function of G with respect to S.

The group order of an element g is denoted by $\pi(g)$. In case F is a word $\pi(F)$ denotes the order of the element represented by the word F.

We describe a step in a procedure that successfully implements the ideas and constructions introduced in [Gri84]. The final result of the procedure is a tree that helps us to show that all G groups whose directed part B is torsion are themselves torsion groups and also to determine some upper bounds on the torsion growth of the constructed groups. The construction presented here is slightly more complicated than the construction from [Gri84] because of the fact that we need to take into account the possibility of noncyclic (and even nonregular) actions of the groups $A_{\sigma^t\omega}$ on their corresponding alphabets.

Let G_ω, be a G group defined by the triple ω (recall Definitions 2.1 and 2.6) and let F be a reduced word of even length of the form

$$F = b_1 a_1 \dots b_k a_k, \tag{17}$$

where a_i represent nontrivial rooted automorphisms in A_ω and b_j represent nontrivial directed automorphisms in B_ω. Rewrite F in the form $F = b_1 b_2^{g_2} \dots b_k^{g_k} a_1 \dots a_k$, where $g_i = (a_1 \dots a_{i-1})^{-1}$, $i = 2, \dots, k$. Set $g = a_1 \dots a_k \in A_\omega$ and let its order be s. Note that $g = 1$ corresponds to $F \in \mathsf{St}_\omega(\mathcal{L}_1)$. Put

$$H = b_1 b_2^{g_2} \dots b_k^{g_k},$$

consider the word $F^s = (Hg)^s \in \mathsf{St}_\omega(\mathcal{L}_1)$ and rewrite it in the form $F^s = H H^{g^{-1}} H^{g^{-2}} \dots H^{g^{-(s-1)}}$. Next, by using tables similar to Table 1, but for all possible a, we calculate the possibly unreduced words $\overline{F_1}, \dots, \overline{F_{m_1}}$ representing the first level sections $(F^s)\varphi_1, \dots, (F^s)\varphi_{m_1}$, respectively. We have

$$\overline{F_i} = H_i H_{i^g} \dots H_{i^{g^{s-1}}},$$

for $i = 1, \dots, m_1$, where H_j represents the corresponding section of H. Note that any two words $\overline{F_i}$ and $\overline{F_j}$ that correspond to two indices in the same orbit of g represent conjugate elements of G_ω. This is clear since

$$\overline{F_{i^g}} = H_{i^g} H_{i^{g^2}} \dots H_{i^{g^{s-1}}} H_{i^g} = \overline{F_i}^{H_i}.$$

For $i = 1, \dots, m_1$, let the length of the cycle of i in g be t_i. Then

$$\overline{F_i} = (H_i H_{i^g} \dots H_{i^{g^{t_i - 1}}})^{s/t_i}$$

and let F_i be a cyclically reduced word obtained after applying simple reductions (including the cyclic ones) to the word $\widetilde{F}_i = H_i H_{i^g} \dots H_{i^{g^{t_i-1}}}$. Clearly

$$(F^s)\psi = \big(\widetilde{F}_1^{s/t_1}, \dots, \widetilde{F}_{m_1}^{s/t_{m_1}}\big)$$

and F has finite order if and only if $F_1, \dots, F_{m_1}$ all have finite order. In the case of a finite order, the order $\pi(F)$ of F is a divisor of $s \cdot \mathrm{lcm}(\pi(F_1), \dots, \pi(F_{m_1}))$, since the order $\pi(F_i^{s/t_i})$ divides the order $\pi(F_i)$.

Let us make a couple of simple observations on the structure of the possibly unreduced words $\widetilde{F}_1, \dots, \widetilde{F}_{m_1}$ used to obtain the reduced words $F_1, \dots, F_{m_1}$. We have

$$\begin{aligned}
\widetilde{F}_i &= H_i H_{i^g} \dots H_{i^{g^{t_i-1}}} \\
&= (b_1)\varphi_i (b_2^{g_2})\varphi_i \dots (b_k^{g_k})\varphi_i (b_1)\varphi_{i^g} (b_2^{g_2})\varphi_{i^g} \dots (b_k^{g_k})\varphi_{i^g} \dots \\
&\quad (b_1)\varphi_{i^{g^{t_i-1}}} (b_2^{g_2})\varphi_{i^{g^{t_i-1}}} \dots (b_k^{g_k})\phi_{i^{g^{t_i-1}}}.
\end{aligned} \tag{18}$$

It is important to note that the indices $i, i^g, \dots, i^{g^{t_i-1}}$ are all distinct, since the length of the cycle of i in g is t_i. This means that at most one of $(b_1)\varphi_i, (b_1)\varphi_{(i)^g}, \dots, (b_1)\varphi_{(i)g^{t_i-1}}$ can be equal to b_1, at most one can be equal to $(b_1)\omega_1$, and all others are empty. Thus we conclude that the word $\widetilde{F_i}$ consists of some of the letters $b_1, \dots, b_k$, possibly not in that order, and no more than k A-letters. In particular, it is possible to get k A-letters only if none of the letters $b_1, \dots, b_k$ is in the kernel K_1.

We have already observed that any two, possibly unreduced, words $\widetilde{F_i}$ and $\widetilde{F_j}$ such that the indices i and j are in the same cycle of g are conjugate. Choose a representative for each cycle of g, and let $\widetilde{F}_{j_1}, \dots, \widetilde{F}_{j_c}$ be such representatives. The important feature of our construction is that each letter b_i from the representation of F as in (17), as well as each $(b_i)\omega_1$, will appear in exactly one of the representatives $\widetilde{F}_{j_1}, \dots, \widetilde{F}_{j_c}$ before reduction.

The tree that has F at its root and the cyclically reduced cycle representatives $F_{j_1}, \dots, F_{j_c}$ at the vertices below the root is called the *pruned period decomposition* of F.

THEOREM 6.1. *A* G *group is a torsion group if and only of its directed part B is a torsion group.*

PROOF. We prove that the order of any element g in G_ω is finite in case B is a torsion group. The proof is by induction on the length n of g, for all G groups simultaneously.

The statement is clear for $n = 0$ and $n = 1$. Assume that it is true for all words of length less than n, where $n \geqslant 2$, and consider an element g of length n.

If n is odd the element g is conjugate to an element of smaller length and we are done by the inductive hypothesis. Assume then that n is even. Clearly, g is conjugate to an element that can be represented by a word of the form

$$F = b_1 a_1 \dots b_k a_k.$$

If all the cycle representatives F_i from the pruned period decomposition of F have length shorter than n we are done by the inductive hypothesis.

Assume that some of the cycle representatives F_i have length n. This is possible only when F does not have any B-letters from K_1. Also, the words $\widetilde{F_i}$ corresponding to the words F_i of length n must be reduced, so that the words F_i that have length n must have the same B-letters as F does. For each of these finitely many words we repeat the discussion above. Either all of the constructed words F_{ij} are strictly shorter than n, and we get the result by induction; or some have length n, but the B-letters appearing in them do not come from $K_1 \cup K_2$.

This procedure cannot go on forever since $K_1 \cup K_2 \cup \dots \cup K_r = B$ holds for some $r \in \mathbb{N}$. Therefore at some stage we get a shortening in all the words and we conclude that the order of F is finite. $\square$

In the sequel we just list some estimates on the period growth in case the directed part B is a finite group. The proofs can be found in [BŠ01].

A finite subsequence $\omega_{i+1}\omega_{i+2}\dots\omega_{i+r}$ of the defining sequence $\overline{\omega} = \omega_1\omega_2\dots$ is *complete* if each element of B is sent to the identity by at least one homomorphism from the sequence $\omega_{i+1}\omega_{i+2}\dots\omega_{i+r}$, i.e. if $\bigcup_{j=1}^{r} K_{i+j} = B$. We note that the complete sequence $\omega_{i+1}\omega_{i+2}\dots\omega_{i+r}$ must have length at least $m+1$, where m is the minimal branching index in the branching sequence, since each kernel K_{i+j} has index $|A_{\sigma^{i+j}\omega}| \geqslant m_{i+j+1} \geqslant m$ in B, for all $j = 1, \dots, r$. In particular, the length of a complete sequence is never shorter than 3. By the definition of a G group, all sequences that define a G group can be factored into finite complete subsequences.

A defining sequence $\overline{\omega}$ is *r-homogeneous*, for $r \geqslant 3$, if all of its finite subsequences of length r are complete. A defining sequence $\overline{\omega}$ is *r-factorable*, for $r \geqslant 3$, if it can be factored in complete subsequences of length at most r.

THEOREM 6.2 (Period η-estimate). *If $\overline{\omega}$ is an r-homogeneous sequence and B has exponent q, then there exist a positive constant C such that the torsion growth function of the group G_ω satisfies*

$$\pi_\omega(n) \leqslant Cn^{\log_{1/\eta_r}(q)}$$

where η_r is the positive root of the polynomial $x^r + x^{r-1} + x^{r-2} - 2$.

THEOREM 6.3 (Period 3/4-estimate). *If $\overline{\omega}$ is an r-factorable sequence and B has exponent q, then there exists a positive constant C such that the torsion growth function of the group G_ω satisfies*

$$\pi_\omega(n) \leqslant Cn^{r\log_{4/3}(q)}.$$

THEOREM 6.4 (Period 2/3-estimate). *If $\overline{\omega}$ is an r-factorable sequence such that each factor contains three letters whose kernels cover B and B has exponent q, then there exists a positive constant C such that the torsion growth function of the group G_ω satisfies*

$$\pi_\omega(n) \leqslant Cn^{r\log_{3/2}(q)}.$$

Let us assume now that all the branching indices are prime numbers and the groups $A_{\sigma^t\omega}$ are cyclic of prime order, for all t. There is no loss in generality if we assume that $A_{\sigma^t\omega}$ is generated by the cyclic permutation $a = (1, 2, \dots, m_{t+1})$. Note that our assumptions force B to be Abelian group, since B is always a subdirect product of several copies of the root groups $A_{\sigma^t\omega}$.

THEOREM 6.5. *Let the branching sequence consists only of primes, B have exponent q and $\overline{\omega}$ be an r-homogeneous word. There exists a positive constant C such that the torsion growth function of G_ω satisfies*

$$\pi_\omega(n) \leqslant Cn^{(r-1)\log_2(q)}.$$

Finally we give a tighter upper bound on the period growth of the Grigorchuk 2-groups (as defined in [Gri84]).

THEOREM 6.6. *Let G_ω be a Grigorchuk 2-group. If $\overline{\omega}$ is an r-homogeneous word, then there exists a positive constant C such that the torsion growth function of the group G_ω satisfies*

$$\pi_\omega(n) \leqslant Cn^{r/2}.$$

In addition to the above estimates [BŠ01] provides lower bounds on the torsion growth function in some cases and, in particular, shows that some Grigorchuk 2-groups have torsion growth functions $\pi(n)$ that are at least linear in n (for example, the group defined by the sequence $\overline{\omega} = 01020102\ldots$). Previous results of I. Lysionok and Yu.G. Leonov [Lys98, Leo97a] already established a lower bound of $C_1 n^{1/2}$ for the torsion growth function of the first Grigorchuk group, while Theorem 6.6 establishes an upper bound of $C_2 n^{3/2}$.

The result following the definition below finds its predecessor in the work of N. Gupta and S. Sidki [GS83b], where they prove that the Gupta–Sidki p-groups contain arbitrary long iterated wreath products of cyclic groups of order p and therefore contain a copy of each finite p-group. The general argument below follows the approach from [GHZ00] and applies well to more broad settings.

DEFINITION 6.7. Let $\mathcal{P}$ be a nonempty set of primes. A group G of automorphisms of $\mathcal{T}$ has *omnipresent $\mathcal{P}$-torsion* if, for every vertex $u \in \mathcal{T}$, the lower companion group L_u^G (the rigid stabilizer at u) has nontrivial elements of $\mathcal{P}$-order.

In case $\mathcal{P}$ consists of all primes we just say that G has *omnipresent torsion* and in case $\mathcal{P} = \{p\}$ we say that G has omnipresent p-torsion. We note that every spherically transitive group with omnipresent $\mathcal{P}$-torsion must be a weakly branch group. We also note that if G has omnipresent $\mathcal{P}$-torsion then so does each of its lower companion groups L_u considered as a group of automorphisms of $\mathcal{T}_u$.

LEMMA 6.8. *Let $\mathcal{P}$ be a nonempty set of primes and G a group of tree automorphisms with omnipresent $\mathcal{P}$-torsion. Then G contains arbitrary long iterated wreath products of cyclic groups of the form*

$$\big((\mathbb{Z}/p_1\mathbb{Z} \wr \cdots) \wr \mathbb{Z}/p_{n-1}\mathbb{Z}\big) \wr \mathbb{Z}/p_n\mathbb{Z},$$

where $n \in \mathbb{N}$ and each p_i is a $\mathcal{P}$-prime, for $i \in \{1, \ldots, n\}$.

SKETCH OF A PROOF. We prove the claim by induction on n, simultaneously for all groups of tree automorphisms with omnipresent $\mathcal{P}$-torsion. The claim is obvious for $n = 1$. Assume that $n \geqslant 2$ and that the claim holds for all positive natural numbers $< n$.

Choose an arbitrary nontrivial element g of G of finite $\mathcal{P}$-prime order p_n. Let g fix the level $\mathcal{L}_k$ but not $\mathcal{L}_{k+1}$ and let u be a vertex on level k with nontrivial vertex permutation $(u)g = a$. All nontrivial cycles of a have length p and, without loss of generality, we may assume that one such cycle is $(1, 2, \ldots, p)$. Without loss of generality we may also assume that the sections $g_{u_1}, \ldots, g_{u_p}$ are trivial (we may accomplish this by conjugation if necessary). By the inductive hypothesis, the lower companion group L_{uy} contains an iterated wreath product of length $n - 1$

$$Q = \big((\mathbb{Z}/p_1\mathbb{Z} \wr \cdots) \wr \mathbb{Z}/p_{n-2}\mathbb{Z}\big) \wr \mathbb{Z}/p_{n-1}\mathbb{Z}$$

of the required form. But then

$$\langle Q, g\rangle \cong Q \wr \langle g\rangle = \big((\mathbb{Z}/p_1\mathbb{Z} \wr \cdots) \wr \mathbb{Z}/p_{n-1}\mathbb{Z}\big) \wr \mathbb{Z}/p_n\mathbb{Z}. \qquad \square$$

The above lemma has many corollaries, some of which are summed up in the following theorem:

THEOREM 6.9. *Let G be a group of tree automorphisms. If, for each vertex u, the lower companion group L_u of G has an element of finite order, then G has elements of unbounded finite order. Further,*

(1) *Every weakly branch torsion group has infinite exponent.*
(2) *Every weakly branch p-group contains a copy of every finite p-group.*
(3) *Every weakly regular branch group is weakly branched over a torsion free group or contains a copy of every finite p-group, for some prime p.*
(4) *A regular branch group is either virtually torsion free or it contains a copy of every finite p-group, for some prime p.*

Finally, we note that the class of GGS groups is also rich with examples of torsion groups, starting with the second Grigorchuk group from [Gri80], Gupta–Sidki p-groups from [GS83a,GS83b] and certain Gupta–Sidki extensions from [GS84]. For a wide class of torsion branch GGS groups see Section 2.3.3.

7. Subgroup structure

We study in this section some subgroup series (derived series, powers series) and general facts about branch groups. We then describe important small-index subgroups in the examples $\mathfrak{G}$, Γ, $\overline{\Gamma}$, $\overline{\overline{\Gamma}}$ (recall the definitions form Section 1.6). The lower central series is treated in the next section. Most of the results come from [BG02].

7.1. *The derived series*

Let G be a group. The *derived series* $(G^{(n)})_{n\in\mathbb{N}}$ of G is defined by $G^{(0)} = G$ and $G^{(n)} = [G^{(n-1)}, G^{(n-1)}]$. A group is *solvable* if $G^{(n)} = \{1\}$ for some $n \in \mathbb{N}$. It is *residually solvable* if $\bigcap_{n\in\mathbb{N}} G^{(n)} = \{1\}$. Note that if $(\gamma_n(G))_{n\in\mathbb{N}}$ is the lower central series of G, then a general result states that $G^{(n)} \leqslant \gamma_{2^n}(G)$ holds for all n.

7.1.1. *The derived series of* $\mathfrak{G}$. Since $\mathfrak{G}$ is regular branch over $K = \langle x\rangle^G$, where $x = [a, b]$, we consider the finite-index subgroups $\mathsf{Rst}_{\mathfrak{G}}(\mathcal{L}_n) = K \times \cdots \times K$ with 2^n factors, for $n \geqslant 2$.

THEOREM 7.1. $\mathfrak{G}^{(n)} = \mathsf{Rst}_{\mathfrak{G}}(\mathcal{L}_{2n-3})$ *for all* $n \geqslant 3$*, and* $K^{(n)} = \mathsf{Rst}_{\mathfrak{G}}(\mathcal{L}_{2n})$ *for all* $n \geqslant 1$.

PROOF. First, one may check by elementary means that $\mathfrak{G}^{(3)} = \mathsf{Rst}_{\mathfrak{G}}(\mathcal{L}_3) = K^{\times 8}$. Then K' is the normal closure in $\mathfrak{G}$ of $[x^d, x]$, and $[x^d, x]\psi = ([(ca)^b, ca], 1)$, and $[(ca)^b, ca]\psi = (x, 1)$. Therefore $K' = K^{\times 4}$. □

7.1.2. *The derived series of* $\overline{\overline{\Gamma}}$. The result is even slightly simpler for $\overline{\overline{\Gamma}}$, which is regular branch over $\overline{\overline{\Gamma}}'$:

THEOREM 7.2. $\overline{\overline{\Gamma}}^{(n)} = (\overline{\overline{\Gamma}}'')^{\times 3^{n-2}}$ *for all* $n \geqslant 2$.

PROOF. The core of the argument is to show that $\overline{\overline{\Gamma}}^{(3)} = \overline{\overline{\Gamma}}'' \times \overline{\overline{\Gamma}}'' \times \overline{\overline{\Gamma}}''$. This follows from $\overline{\overline{\Gamma}}^{(3)} = \gamma_8(\overline{\overline{\Gamma}})$ and $\overline{\overline{\Gamma}}'' = \gamma_5(\overline{\overline{\Gamma}})$, but the computations are tricky – see Section 8.2.3 for details. $\square$

7.2. *The powers series*

Let G be a group and d an integer. The *powers series* $(\mho_d^n(G))_{n\in\mathbb{N}}$ is defined by $\mho_d^0(G) = G$ and

$$\mho_d^n(G) = \langle x^d \text{ for all } x \in \mho_d^{n-1}(G)\rangle.$$

THEOREM 7.3. *The 2-powers series of* $\mathfrak{G}$ *is as follows* $\mho_2(\mathfrak{G}) = \mathfrak{G}'$ *and*

$$\mho_2^n(\mathfrak{G}) = \langle \Delta\big((\mho_2 K)^{\times 2^{n-2}}\big), \Delta\big(K^{\times 2^{n-1}}\big)\rangle,$$

where $\Delta(G^j) = \{(g, \ldots, g) \mid g \in G\}$ *is the diagonal subgroup.*

7.3. *Parabolic subgroups*

In the context of groups acting on a hyperbolic space, a parabolic subgroup is the stabilizer of a point on the boundary. We give here a few general facts concerning parabolic subgroups of branch groups, and recall some results on growth of groups and sets on which they act.

More information and uses of parabolic subgroups appear in the context of representations (Section 9.0.1), Schreier graphs (Section 10.3) and spectrum (Section 11).

DEFINITION 7.4. A *ray* e in $\mathcal{T}$ is an infinite geodesic starting at the root of $\mathcal{T}$, or equivalently an element of $\partial\mathcal{T} = Y^{\mathbb{N}}$.

Let $G \leqslant \mathsf{Aut}(\mathcal{T})$ be any subgroup acting spherically transitively and e be a ray. The associated *parabolic subgroup* is $P_e = \mathsf{St}_G(e)$.

The following important facts are easy to prove:

- For any $e \in \partial\mathcal{T}$, we have $\bigcap_{f\in\partial\mathcal{T}} P_f = \bigcap_{g\in G} P_e^g = 1$.
- Let $e = e_1e_2\ldots$ be an infinite ray and define the subgroups $P_n = \mathsf{St}_G(e_1\ldots e_n)$. Then P_n has index $m_1m_2\cdots m_n$ in G (since G acts transitively) and

$$P_e = \bigcap_{n\in\mathbb{N}} P_n.$$

- P has infinite index in G, and has the same image as P_n in the quotient $G_n = G/\mathsf{St}_G(\mathcal{L}_n)$.

DEFINITION 7.5. Two infinite sequences $\sigma, \tau : \mathbb{N} \to Y$ are *confinal* if there is an $N \in \mathbb{N}$ such that $\sigma_n = \tau_n$ for all $n \geqslant N$.

Confinality is an equivalence relation, and equivalence classes are called *confinality classes*.

The following result is due to V. Nekrashevich and V. Sushchanskiĭ.

PROPOSITION 7.6. *Let G be a group acting on a regular rooted tree $\mathcal{T}^{(m)}$, and assume that for any generator $g \in G$ and infinite sequence τ, the sequences τ and τ^g differ only in finitely many places. Then the confinality classes are unions of orbits of the action of G on $\partial\mathcal{T}$. If moreover for all $u \in \mathcal{T}$ and $v \in \mathcal{T} \setminus u$ there is some $a \in \mathsf{St}_G(u) \cap \mathsf{St}_G(v)$ transitive on the m subtrees below v, then the orbits of the action are the confinality classes.*

DEFINITION 7.7. The subgroup H of G is *weakly maximal* if H is of infinite index in G, but all subgroups of G strictly containing H are of finite index in G.

Note that every infinite finitely generated group admits maximal subgroups, by Zorn's lemma.

However, some branch groups may not contain any infinite-index maximal subgroups; this is the case for $\mathfrak{G}$, as was shown by E. Pervova (see [Per00]).

PROPOSITION 7.8. *Let P be a parabolic subgroup of a branch group G with branch structure $(L_i, H_i)_{i \in \mathbb{N}}$. Then P is weakly maximal.*

PROOF. Let $P = \mathsf{St}_G(e)$ where $e = e_1 e_2 \ldots$. Recall that G contains a product of k_n copies of L_n at level n, and clearly P contains a product of $k_n - 1$ copies of L_n at level n, namely all but the one indexed by the vertex $e_1 \ldots e_n$.

Take $g \in G \setminus P$. There is then an $n \in \mathbb{N}$ such that $(e_1 \ldots e_n)^g \neq e_1 \ldots e_n$, so $\langle P, P^g \rangle$ contains the product $H_n = L_n^1 \times \cdots \times L_n^{k_n}$ of k_n copies of L_n at level n, hence is of finite index in G. $\square$

7.4. *The structure of* $\mathfrak{G}$

Recall that $\mathfrak{G}$, introduced in Section 1.6.1, is the group acting on the binary tree, generated by the rooted automorphism a and the directed automorphisms b, c, d satisfying $\psi(b) = (a, c)$, $\psi(c) = (a, d)$ and $\psi(d) = (1, b)$.

$\mathfrak{G}$ has 7 subgroups of index 2:

$$\begin{array}{ll}
\langle b, ac \rangle, & \langle c, ad \rangle,\ \langle d, ab \rangle, \\
\langle b, a, a^c \rangle, & \langle c, a, a^d \rangle,\ \langle d, a, a^b \rangle, \\
\mathsf{St}_G(1) = \langle b, c, b^a, c^a \rangle. &
\end{array}$$

As can be computed from its presentation [Lys85] and a computer algebra system [S+93], $\mathfrak{G}$ has the following subgroup count (see Table 3).

See [Bar00c] and [CST01] for more information.

Table 3

Index	Subgroups	Normal	In $\mathsf{St}_{\mathfrak{G}}(\mathcal{L}_1)$	Normal
1	1	1	0	0
2	7	7	1	1
4	19	7	9	4
8	61	7	41	7
16	237	5	169	5
32	843	3	609	3

7.4.1. *Normal closures of generators.* They are as follows:

$$\begin{aligned}
A &= \langle a\rangle^{\mathfrak{G}} = \langle a, a^b, a^c, a^d\rangle, && \mathfrak{G}/A \cong \mathbb{Z}/2\mathbb{Z}\times\mathbb{Z}/2\mathbb{Z},\\
B &= \langle b\rangle^{\mathfrak{G}} = \langle b, b^a, b^{ad}, b^{ada}\rangle, && \mathfrak{G}/B \cong D_8,\\
C &= \langle c\rangle^{\mathfrak{G}} = \langle c, c^a, c^{ad}, c^{ada}\rangle, && \mathfrak{G}/C \cong D_8,\\
D &= \langle d\rangle^{\mathfrak{G}} = \langle d, d^a, d^{ac}, d^{aca}\rangle, && \mathfrak{G}/D \cong D_{16}.
\end{aligned}$$

7.4.2. *Some other subgroups.* To complete the picture, we introduce the following subgroups of $\mathfrak{G}$:

$$\begin{array}{lll}
K = \big\langle (ab)^2\big\rangle^{\mathfrak{G}}, & L = \big\langle (ac)^2\big\rangle^{\mathfrak{G}}, & M = \big\langle (ad)^2\big\rangle^{\mathfrak{G}},\\
\overline{B} = \langle B, L\rangle, & \overline{C} = \langle C, K\rangle, & \overline{D} = \langle D, K\rangle,\\
 & T = K^2 = \big\langle (ab)^4\big\rangle^{\mathfrak{G}}, & \\
T_{(m)} = \underbrace{T\times\cdots\times T}_{2^m}, & K_{(m)} = \underbrace{K\times\cdots\times K}_{2^m}, & N_{(m)} = T_{(m-1)}K_{(m)}.
\end{array}$$

THEOREM 7.9.
- *In the Lower Central Series,* $\gamma_{2^m+1}(\mathfrak{G}) = N_{(m)}$ *for all* $m \geqslant 1$.
- *In the Derived Series,* $K^{(n)} = \mathsf{Rst}_{\mathfrak{G}}(2n)$ *for all* $n \geqslant 2$ *and* $\mathfrak{G}^{(n)} = \mathsf{Rst}_{\mathfrak{G}}(2n-3)$ *for all* $n \geqslant 3$.
- *The rigid stabilizers satisfy*

$$\mathsf{Rst}_{\mathfrak{G}}(n) = \begin{cases} D & \text{if } n = 1,\\ K_{(n)} & \text{if } n \geqslant 2.\end{cases}$$

- *The level stabilizers satisfy*

$$\mathsf{St}_{\mathfrak{G}}(\mathcal{L}_n) = \begin{cases} \langle b, c, d\rangle^{\mathfrak{G}} & \text{if } n = 1,\\ \langle D, T\rangle & \text{if } n = 2,\\ \langle N_{(2)}, (ab)^4(adabac)^2\rangle & \text{if } n = 3,\\ \underbrace{\mathsf{St}_{\mathfrak{G}}(\mathcal{L}_3)\times\cdots\times\mathsf{St}_{\mathfrak{G}}(\mathcal{L}_3)}_{2^{n-3}} & \text{if } n \geqslant 4.\end{cases}$$

Consequently, the index of $\mathsf{St}_{\mathfrak{G}}(\mathcal{L}_n)$ *is*

$$\left|\mathfrak{G}/\mathsf{St}_{\mathfrak{G}}(\mathcal{L}_n)\right| = 2^{5\cdot 2^{n-3}+2}. \tag{19}$$

- *There is for all* $\sigma \in Y^n$ *a surjection* $\cdot|_\sigma : \mathsf{St}_{\mathfrak{G}}(\mathcal{L}_n) \twoheadrightarrow \mathfrak{G}$ *given by projection on the factor indexed by* σ.

The top of the lattice of normal subgroups of $\mathfrak{G}$ *below* $\mathsf{St}_{\mathfrak{G}}(\mathcal{L}_1)$ *is given in Figure* 24.

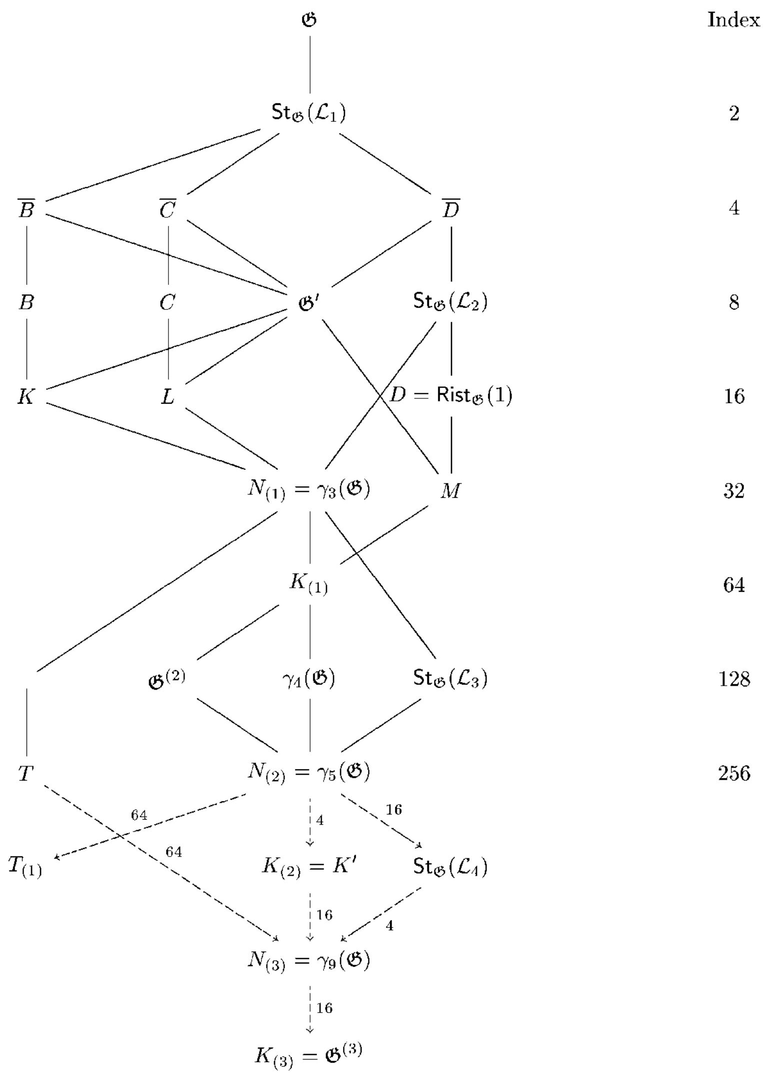


Table 24. The top of the lattice of normal subgroups of $\mathfrak{G}$ below $\mathsf{St}_{\mathfrak{G}}(\mathcal{L}_1)$. The index of the inclusions are indicated next to the edges.

COROLLARY 7.10. *The closure of $\mathfrak{G}$ in $\mathsf{Aut}(\mathcal{T})$ has Hausdorff dimension $5/8$.*

7.4.3. *The subgroup P.* Let e be the ray 2^∞ and let P be the corresponding parabolic subgroup. We describe its structure completely as follows:

THEOREM 7.11. *P/P' is an infinite elementary 2-group generated by the images of c, $d = (1, b)$ and of all elements of the form $(1, \dots, 1, (ac)^4)$ in $\mathsf{Rst}_{\mathfrak{G}}(n)$ for $n \in \mathbb{N}$. The following decomposition holds*:

$$P = \big(B \times \big(\big(K \times \big((K \times \cdots) \rtimes \langle (ac)^4 \rangle\big)\big) \rtimes \langle b, (ac)^4 \rangle\big)\big) \rtimes \langle c, (ac)^4 \rangle,$$

where each factor (of nesting n) in the decomposition acts on the subtree just below some e_n but not containing e_{n+1}.

Note that we use the same notation for a subgroup B or K acting on a subtree, keeping in mind the identification of a subtree with the original tree. Note also that ψ is omitted when it would make the notations too heavy.

PROOF. Define the following subgroups of $\mathfrak{G}_n$:

$$H_n = \langle b, c \rangle^{\mathfrak{G}_n}; \qquad B_n = \langle b \rangle^{\mathfrak{G}_n}; \qquad K_{(n)} = \big\langle (ab)^2 \big\rangle^{\mathfrak{G}_n};$$
$$Q_n = B_n \cap P_n; \qquad R_n = K_{(n)} \cap P_n.$$

Then the theorem follows from the following proposition. □

PROPOSITION 7.12. *These subgroups have the following structure*:

$$P_n = (B_{n-1} \times Q_{n-1}) \rtimes \langle c, (ac)^4 \rangle;$$
$$Q_n = (K_{n-1} \times R_{n-1}) \rtimes \langle b, (ac)^4 \rangle;$$
$$R_n = (K_{n-1} \times R_{n-1}) \rtimes \langle (ac)^4 \rangle.$$

PROOF. A priori, P_n, as a subgroup of H_n, maps in $(B_{n-1} \times B_{n-1}) \rtimes \langle (a, d), (d, a) \rangle$. Restricting to those pairs that fix e_n gives the result. Similarly, Q_n, as a subgroup of B_n, maps in $(K_{n-1} \times K_{n-1}) \rtimes \langle (a, c), (c, a) \rangle$, and R_n, as a subgroup of K_n, maps in $(K_{n-1} \times K_{n-1}) \rtimes \langle (ac, ca), (ca, ac) \rangle$. □

COROLLARY 7.13. *The group $\mathfrak{G}_n$ and its subgroups H_n, B_n, K_n, P_n, R_n, Q_n are arranged in a lattice depicted below, where the quotients or the indices are represented* *next to the edges.*

$$
\begin{tikzcd}[column sep=small]
 & & \mathfrak{G}_n \arrow[d, dash, "\langle a \rangle"'] & \\
 & & H_n \arrow[dl, dash, "{\langle d, d^a \rangle}"'] \arrow[dr, dash, "2^{n-1}"] & \\
 & B_n \arrow[dl, dash, "\langle b \rangle"'] \arrow[dr, dash, "2^{n-1}"] & & P_n \arrow[dl, dash, "2^2"] \\
K_n \arrow[dr, dash, "2^{n-1}"'] & & Q_n \arrow[dl, dash, "2"] & \\
 & R_n & &
\end{tikzcd}
$$

7.5. *The structure of Γ*

Recall that Γ is the group acting on the ternary tree, generated by the rooted automorphism $a = ((1, 2, 3))$ and the directed automorphism t satisfying $(t)\psi = (a, 1, t)$.

Define the elements $x = at$, $y = ta$ of Γ. Let K be the subgroup of Γ generated by x and y, and let L be the subgroup of K generated by K' and cubes in K. Write $H = \mathsf{St}_\Gamma(\mathcal{L}_1)$.

PROPOSITION 7.14. *We have the following diagram of normal subgroups*:

$$
\begin{tikzcd}[column sep=small]
 & \Gamma \arrow[dl, dash, "\langle a | a^3 \rangle"'] \arrow[dr, dash, "\langle a | a^3 \rangle"] & \\
K \arrow[dr, dash, "\langle x | x^3 \rangle"'] & & \mathsf{St}_\Gamma(\mathcal{L}_1) \arrow[dl, dash, "\langle t | t^3 \rangle"] \\
 & \Gamma' = K \cap H = [K, H] \arrow[d, dash] & \\
 & L = \langle K', K^3 \rangle = \gamma_3(\Gamma) \arrow[dl, dash] \arrow[dr, dash] & \\
K' \arrow[d, dash] & & H' = (\Gamma' \times \Gamma' \times \Gamma')\psi^{-1} = \mathsf{St}_\Gamma(\mathcal{L}_2) \\
\langle L \times L \times L, x^3 y^{-3}, [x, y^3] \rangle = \gamma_4(\Gamma) \arrow[d, dash] & & \\
\langle L \times L \times L, [x, y^3] \rangle = \gamma_5(\Gamma) & &
\end{tikzcd}
$$

where the quotients are represented next to the edges; *all edges represent normal inclusions of index* 3. *Furthermore* $L = K \cap (K \times K \times K)\psi^{-1}$.

PROOF. First we prove K is normal in Γ, of index 3, by writing $y^t = x^{-1}y^{-1}$, $y^{a^{-1}} = y^{-1}x^{-1}$, $y^{t^{-1}} = y^a = x$; similar relations hold for conjugates of x. A transversal of K in Γ is $\langle a\rangle$. All subgroups in the diagram are then normal.

Since $[a,t] = y^{-1}x = t^a t^{-1}$, we clearly have $\Gamma' < K \cap H$. Now as $\Gamma' \neq K$ and $\Gamma' \neq H$ and Γ' has index 3^2, we must have $\Gamma' = K \cap H$. Finally $[a,t] = [x,t]^{t^{-1}}$, so $\Gamma' = [K, H]$.

Next $x^3 = [a,t][t,a^{-1}][a^{-1},t^{-1}]$ and similarly for y, so $K^3 < \Gamma'$ and $L < \Gamma'$. Also, $[x,y]\psi = (y^{-1}, y^{-1}, x^{-1})$ and $(x^3)\psi = (y,x,y)$ both belong to $K \times K \times K$, while $[a,t]$ does not; so L is a proper subgroup of Γ', of index 3 (since K/L is the elementary Abelian group $(\mathbb{Z}/3\mathbb{Z})^2$ on x and y).

Consider now H'. It is in $\mathsf{St}_\Gamma(\mathcal{L}_2)$ since $H = \mathsf{St}_\Gamma(\mathcal{L}_1)$. Also, $[t, t^a] = y^3[y^{-1}, x]$ and similarly for other conjugates of t, so $H' < L$, and $[t,t^a]\psi = ([a,t],1,1)$, so $(H')\psi = \Gamma' \times \Gamma' \times \Gamma'$. Finally H' it is of index 3 in L (since $H/H' = (\mathbb{Z}/3\mathbb{Z})^3$ on $t, t^a, t^{a^{-1}}$), and since $\mathsf{St}_\Gamma(\mathcal{L}_2)$ is of index 3^4 in Γ (with quotient $\mathbb{Z}/3\mathbb{Z} \wr \mathbb{Z}/3\mathbb{Z}$) we have all the claimed equalities. $\square$

PROPOSITION 7.15. *Γ is a just-infinite fractal group, is regular branch over Γ', and has the congruence property.*

PROOF. Γ is fractal by Lemma 1.7 and the nature of the map ψ. By direct computation, $[\Gamma : \Gamma'] = [\Gamma' : (\Gamma' \times \Gamma' \times \Gamma')\psi^{-1}] = [(\Gamma' \times \Gamma' \times \Gamma')\psi^{-1} : \Gamma''] = 3^2$, so Γ is branched on Γ'. Then $\Gamma'' = \gamma_5(\Gamma)$, as is shown in [Bar00c], so Γ'' has finite index and Γ is just-infinite by Theorem 5.2.

$\Gamma' \geqslant \mathsf{St}_\Gamma(\mathcal{L}_2)$, so Γ has the congruence property. $\square$

PROPOSITION 7.16. *Writing $\langle S\rangle$ for the 3-Abelian quotient of $\langle S\rangle$, we have exact sequences*

$$1 \to \Gamma' \times \Gamma' \times \Gamma' \to (H)\psi \to \langle t, t^a, t^{a^2}\rangle_{3-ab},$$
$$1 \to \Gamma' \times \Gamma' \times \Gamma' \to (\Gamma')\psi \to \langle [a,t], [a^2,t]\rangle_{3-ab}.$$

THEOREM 7.17. *The subgroup K of Γ is torsion-free*; *thus Γ is virtually torsion-free.*

PROPOSITION 7.18. *The finite quotients $\Gamma_n = \Gamma/\mathsf{St}_\Gamma(\mathcal{L}_n)$ of Γ have order $3^{3^{n-1}+1}$ for $n \geqslant 2$, and 3 for $n = 1$.*

PROOF. Follows immediately from $[\Gamma : \Gamma'] = 3^2$ and $[\Gamma' : (\Gamma' \times \Gamma' \times \Gamma')\psi^{-1}] = 3^2$. $\square$

COROLLARY 7.19. *The closure of Γ in $\mathsf{Aut}(\mathcal{T})$ is isomorphic to the profinite completion $\widehat{\Gamma}$ and is a pro-3-group. It has Hausdorff dimension $1/3$.*

7.6. *The structure of $\overline{\Gamma}$*

Recall that $\overline{\Gamma}$ is the group acting on the ternary tree, generated by the rooted automorphism $a = ((1,2,3))$ and the directed automorphism t defined by $(t)\psi = (a,a,t)$.

Define the elements $x = ta^{-1}$, $y = a^{-1}t$ of $\overline{\Gamma}$, and let K be the subgroup of $\overline{\Gamma}$ generated by x and y. Then K is normal in $\overline{\Gamma}$, because $x^t = y^{-1}x^{-1}$, $x^a = x^{-1}y^{-1}$, $x^{t^{-1}} = x^{a^{-1}} = y$, and similar relations hold for conjugates of y. Moreover K is of index 3 in $\overline{\Gamma}$, with transversal $\langle a\rangle$. Write $H = \mathsf{St}_{\overline{\Gamma}}(\mathcal{L}_1)$.

LEMMA 7.1. *H and K are normal subgroups of index* 3 *in $\overline{\Gamma}$, and $\overline{\Gamma}' = \mathsf{St}_K(\mathcal{L}_1) = H \cap K$ is of index* 9; *furthermore $(H \cap K)\psi \lhd K \times K \times K$. For any element $g = (u, v, w) \in (H \cap K)\psi$ one has $wvu \in H \cap K$.*

PROOF. First note that $\mathsf{St}_K(\mathcal{L}_1) = \langle x^3, y^3, xy^{-1}, y^{-1}x\rangle$, for every word in x and y whose number of a's is divisible by 3 can be written in these generators. Then compute

$$
\begin{aligned}
(x^3)\psi &= (y, x^{-1}y^{-1}, x), \qquad & (y^3)\psi &= (x^{-1}y^{-1}, x, y),\\
(xy^{-1})\psi &= (1, x^{-1}, x), \qquad & (y^{-1}x)\psi &= (y, 1, y^{-1}).
\end{aligned}
$$

The last assertion is also checked by this computation. □

PROPOSITION 7.20. *Writing $c = [a, t] = x^{-1}y^{-1}x^{-1}$ and $d = [x, y]$, we have the following diagram of normal subgroups*:

$$
\begin{tikzcd}
& \overline{\Gamma} \arrow[dl, dash, "\langle a|a^3\rangle"'] \arrow[dr, dash, "\langle a|a^3\rangle"] & \\
K \arrow[dr, dash, "\langle x,y|x^3,y^3,x=y\rangle"'] & & H \arrow[dl, dash, "\langle t_0,t_1,t_2|t_0^3,t_0=t_1=t_2\rangle"] \\
& \overline{\Gamma}' = \langle c, c^t, c^{a^{-1}}, c^{at}\rangle = K \cap H = [K, H] \arrow[dl, dash, "\mathbb{Z}^2"'] \arrow[dr, dash, "(\mathbb{Z}/3\mathbb{Z})^2"] & \\
K' = \langle d, d^t, d^{a^{-1}}, d^{at}\rangle \arrow[dr, dash, "\mathbb{Z}^2"'] & & H' \arrow[dl, dash] \\
& \overline{\Gamma}'' = (K' \times K' \times K')\psi^{-1} \arrow[d, dash, "(\mathbb{Z}/3\mathbb{Z})^2"] & \\
& K'' &
\end{tikzcd}
$$

where the quotients are represented next to the edges; *additionally,*

$$
\begin{aligned}
&K/K' = \big\langle x, y \big| [x, y]\big\rangle \cong \mathbb{Z}^2,\\
&\overline{\Gamma}'/\overline{\Gamma}'' = \big\langle c, c^t, c^{a^{-1}}, c^{at} \big| [c, c^t], \dots\big\rangle \cong \mathbb{Z}^4,\\
&K'/K'' = \big\langle d, d^t, d^{a^{-1}}, d^{at} \big| [d, d^t], \dots, (d/d^{at})^3, (d^{a^{-1}}/d^t)^3\big\rangle \cong \mathbb{Z}^2 \times (\mathbb{Z}/3\mathbb{Z})^2.
\end{aligned}
$$

Writing each subgroup in the generators of the groups above it, we have

$$
\begin{aligned}
K &= \langle x = at^{-1}, y = a^{-1}t \rangle, \\
H &= \langle t, t_1 = t^a, t_2 = t^{a^{-1}} \rangle, \\
\overline{\Gamma}' &= \langle b_1 = xy^{-1}, b_2 = y^{-1}x, b_3 = x^3, b_4 = y^3 \rangle \\
&= \langle c_1 = tt_1^{-1}, c_2 = tt_1t, c_3 = tt_2^{-1}, c_4 = tt_2t \rangle.
\end{aligned}
$$

COROLLARY 7.21. *The congruence property does not hold for $\overline{\Gamma}$; nor is it regular branch.*

PROPOSITION 7.22. *$\overline{\Gamma}$ is a fractal group, is weakly branch, and just-nonsolvable.*

PROOF. $\overline{\Gamma}$ is fractal by Lemma 1.7 and the nature of the map ψ. The subgroup K described above has an infinite-index derived subgroup K' (with infinite cyclic quotient), from which we conclude that $\overline{\Gamma}$ is not just-infinite; indeed K' is normal in $\overline{\Gamma}$ and $\overline{\Gamma}/K' \cong \mathbb{Z}^2 \rtimes \left(\begin{smallmatrix}-1 & 1\\ -1 & 0\end{smallmatrix}\right)$ is infinite. $\square$

PROPOSITION 7.23. *The subgroup K of $\overline{\Gamma}$ is torsion-free; thus $\overline{\Gamma}$ is virtually torsion-free.*

PROOF. For $1 \neq g \in K$, let $|g|_t$, the t-length of g, denote the minimal number of $t^{\pm 1}$'s required to write g as a word over the alphabet $\{a^{\pm 1}, t^{\pm 1}\}$. We will show by induction on $|g|_t$ that g is of infinite order.

First, if $|g|_t = 1$, i.e. $g \in \{x^{\pm 1}, y^{\pm 1}\}$, we conclude from $(x^3)\psi = (*, *, x)$ and $(y^3)\psi = (*, *, y)$ that g is of infinite order.

Suppose now that $|g|_t > 1$, and $g \in \mathsf{St}_{\overline{\Gamma}}(\mathcal{L}_n) \setminus \mathsf{St}_{\overline{\Gamma}}(\mathcal{L}_{n+1})$. Then there is some sequence σ of length n that is fixed by g and such that $g|_\sigma \notin H$. By Lemma 7.1, $g|_\sigma \in K$, so it suffices to show that all $g \in K \setminus H$ are of infinite order.

Such a g can be written as $(u, v, w)\psi^{-1}z$ for some $(u, v, w) \in (K \cap H)\psi$ and $z \in \{x^{\pm 1}, y^{\pm 1}\}$; by symmetry let us suppose $z = x$. Then $g^3 = (uavawt, vawtua, wtuava)\psi^{-1} = (g_0, g_1, g_2)\psi^{-1}$, say. For any i, we have $|g_i|_t \leqslant |g|_t$, because all the components of $(x)\psi$ and $(y)\psi$ have t-length $\leqslant 1$. We distinguish three cases:

(1) $g_i = 1$ for some i. Then consider the image $\overline{g_i}$ of g_i in $\overline{\Gamma}/\overline{\Gamma}'$. By Lemma 7.1, $wvu \in G'$, so $\overline{g_i} = 1 = \overline{a^2t}$. But this is a contradiction, because $\overline{\Gamma}/\overline{\Gamma}'$ is elementary Abelian of order 9, generated by the independent images $\bar{a}$ and $\bar{t}$.

(2) $0 < |g_i|_t < |g|_t$ for some i. Then by induction g_i is of infinite order, so g^3 too, and g too.

(3) $|g_i|_t = |g|_t$ for all i. We repeat the argument with g_i substituted for g. As there are finitely many elements h with $|h|_t = |g|_t$, we will eventually reach either an element of shorter length or an element already considered. In the latter case we obtain a relation of the form $(g^{3^n})\psi^n = (\dots, g, \dots)$ from which g is seen to be of infinite order. $\square$

PROPOSITION 7.24. *The finite quotients $\overline{\Gamma}_n = \overline{\Gamma}/\mathsf{St}_{\overline{\Gamma}}(\mathcal{L}_n)$ of $\overline{\Gamma}$ have order $3^{\frac{1}{4}(3^n+2n+3)}$ for $n \geqslant 2$, and $3^{\frac{1}{2}(3^n-1)}$ for $n \leqslant 2$.*

PROOF. Define the following family of two-generated finite Abelian groups:

$$A_n = \begin{cases} \langle x, y \mid x^{3^{n/2}}, y^{3^{n/2}}, [x, y]\rangle & \text{if } n \equiv 0[2], \\ \langle x, y \mid x^{3^{(n+1)/2}}, y^{3^{(n+1)/2}}, (xy^{-1})^{3^{(n-1)/2}}, [x, y]\rangle & \text{if } n \equiv 1[2]. \end{cases}$$

First suppose $n \geqslant 2$; consider the diagram of groups described above, and quotient all the groups by $\mathsf{St}_{\overline{\Gamma}}(\mathcal{L}_n)$. Then the quotient K/K' is isomorphic to A_n, generated by x and y, and the quotient $K'/\overline{\Gamma}''$ is isomorphic to A_{n-1}, generated by $[x, y]$ and $[x, y]^t$. As $|A_n| = 3^n$, the index of K'_n in $\overline{\Gamma}_n$ is 3^{n+1} and the index of $\overline{\Gamma}''_n$ is 3^{2n}. Then as $\overline{\Gamma}''_n \cong K^3_{n-1}$ and $|\overline{\Gamma}''_2| = 1$ we deduce by induction that $|\overline{\Gamma}''_n| = 3^{\frac{1}{4}(3^n - 6n+3)}$ and $|K'_n| = 3^{\frac{1}{4}(3^n - 2n-1)}$, from which $|\overline{\Gamma}_n| = 3^{2n} + |\overline{\Gamma}''_n| = 3^{\frac{1}{4}(3^n + 2n+3)}$ follows.

For $n \leqslant 2$ we have $\overline{\Gamma}_n = \mathsf{Aut}(\mathcal{T})_n = \mathbb{Z}/3 \wr \cdots \wr \mathbb{Z}/3$. □

COROLLARY 7.25. *The closure $\overline{\overline{\Gamma}}$ of $\overline{\Gamma}$ in $\mathsf{Aut}(\mathcal{T})$ has Hausdorff dimension* $1/2$.

PROPOSITION 7.26. *We have exact sequences*

$$1 \to K' \times K' \times K' \to (H)\psi \to \mathbb{Z}^4 \rtimes \mathbb{Z}/3\mathbb{Z} \to 1,$$
$$1 \to K' \times K' \times K' \to (K')\psi \to \mathbb{Z}^2 \to 1.$$

7.7. *The structure of $\overline{\overline{\Gamma}}$*

Recall that $\overline{\overline{\Gamma}}$ is the group acting on the ternary tree, generated by the rooted automorphism $a = ((1, 2, 3))$ and the directed automorphism t satisfying $(t)\psi = (a, a^{-1}, t)$. Write $H = \mathsf{St}_{\overline{\overline{\Gamma}}}(\mathcal{L}_1)$.

PROPOSITION 7.27. *We have the following diagram of normal subgroups*:

$$\begin{tikzcd}
\overline{\overline{\Gamma}} \arrow[d, dash, "\langle a \mid a^3\rangle"] \\
H = \mathsf{St}_{\overline{\overline{\Gamma}}}(\mathcal{L}_1) \arrow[d, dash, "\langle t \mid t^3\rangle"] \\
\overline{\overline{\Gamma}}' = [G, H] \arrow[d, dash, "{[a,t]}"] \\
\gamma_3(\overline{\overline{\Gamma}}) = \overline{\overline{\Gamma}}^3 = \mathsf{St}_{\overline{\overline{\Gamma}}}(\mathcal{L}_2) \arrow[d, dash, "(at)^3"] \\
H' = (\overline{\overline{\Gamma}}' \times \overline{\overline{\Gamma}}' \times \overline{\overline{\Gamma}}')\psi^{-1}
\end{tikzcd}$$

where the quotients are represented next to the arrows; *all edges represent normal inclusions of index* 3.

PROPOSITION 7.28. *$\overline{\overline{\Gamma}}$ is a just-infinite fractal group, and is a regular branch group over $\overline{\overline{\Gamma}}'$.*

PROPOSITION 7.29. *$\overline{\overline{\Gamma}}' \geqslant \mathsf{St}_{\overline{\overline{\Gamma}}}(\mathcal{L}_2)$, so $\overline{\overline{\Gamma}}$ has the congruence property.*

8. Central series, finiteness of width and associated Lie algebras

In this section we study the lower central, lower p-central, and dimension series of basic examples of branch groups, and describe the associated Lie algebras. This section is very much connected to the previous one.

We exhibit two branch groups of finite width: $\mathfrak{G}$ and Γ, and describe the "Lie graph" of their associated Lie algebras. We show that the Gupta–Sidki 3-group has unbounded width, and its Lie algebra has growth of degree $n^{\log 3/\log(1+\sqrt{2})-1}$; we also describe its Lie graph.

For all regular branch groups, the corresponding Lie algebras have polynomial growth (usually of nonintegral degree), see Theorem 8.9. The technique used below, described in [Bar00c], is a far-reaching development of the methods of [BG00a].

We start by recalling the main construction, due to W. Magnus [Mag40].

8.1. *N-series*

DEFINITION 8.1. Let G be a group. An *N-series* is a series $\{G_n\}$ of normal subgroups with $G_1 = G$, $G_{n+1} \leqslant G_n$ and $[G_m, G_n] \leqslant G_{m+n}$ for all $m, n \geqslant 1$. The associated Lie ring is

$$\mathcal{L}(G) = \bigoplus_{n=1}^{\infty} \mathcal{L}_n,$$

with $\mathcal{L}_n = G_n/G_{n+1}$ and the bracket operation $\mathcal{L}_n \otimes \mathcal{L}_m \to \mathcal{L}_{m+n}$ induced by commutation in G.

For p a prime, an *N_p-series* is an N-series $\{G_n\}$ such that $\mho_p(G_n) \leqslant G_{pn}$, and the associated Lie ring is a restricted Lie algebra over $\mathbb{F}_p$ [Jac41],

$$\mathcal{L}_{\mathbb{F}_p}(G) = \bigoplus_{n=1}^{\infty} \mathcal{L}_n,$$

with the Frobenius operation $\mathcal{L}_n \to \mathcal{L}_{pn}$ induced by raising to the power p in G. (Recall the definition of $\mho_d(G)$ from Section 7.2.)

The standard example of N-series is the *lower central series*, $\{\gamma_n(G)\}_{n=1}^{\infty}$, given by $\gamma_1(G) = G$ and $\gamma_n(G) = [G, \gamma_{n-1}(G)]$, or the *lower p-central series* or *Frattini series* given by $P_1(G) = G$ and $P_n(G) = [G, P_{n-1}(G)]\mho_p(P_{n-1}(G))$. It differs from the lower central series in that its successive quotients are all elementary p-groups.

The standard example of N_p-series is the *dimension series*, also known as the Zassenhaus [Zas40], Jennings [Jen41], Lazard [Laz53] or Brauer series, given by $G_1 = G$ and $G_n = [G, G_{n-1}]\mho_p(G_{\lceil n/p\rceil})$, where $\lceil n/p\rceil$ is the least integer greater than or equal to n/p. It can alternately be described, by a result of Lazard [Laz53], as

$$G_n = \prod_{i\cdot p^j \geqslant n} \mho_{p^j}\big(\gamma_i(G)\big),$$

or as

$$G_n = \{g \in G \mid g - 1 \in \Delta^n\},$$

where Δ is the augmentation (or fundamental) ideal of the group algebra $\mathbb{F}_p G$. Note that this last definition extends to characteristic 0, giving a graded Lie algebra $\mathcal{L}_{\mathbb{Q}}(G)$ over $\mathbb{Q}$. In that case, the subgroup G_n is the isolator $\gamma_n(G)$:

$$G_n = \sqrt{\gamma_n(G)} = \big\{g \in G \mid \langle g\rangle \cap \gamma_n(G) \neq \{1\}\big\}.$$

We mention finally for completeness another N_p-series, the *Lie dimension series* $L_n(G)$ defined by

$$L_n(G) = \big\{g \in G \mid g - 1 \in \Delta^{(n)}\big\},$$

where $\Delta^{(n)}$ is the n-th Lie power of $\Delta < \Bbbk G$, given by $\Delta^{(1)} = \Delta$ and $\Delta^{(n)} = [\Delta^{(n)}, \Delta] = \{xy - yx \mid x \in \Delta^{(n)},\ y \in \Delta\}$. It is then known [PS75] that

$$L_n(G) = \prod_{(i-1)\cdot p^j \geqslant n} \mho_{p^j}\big(\gamma_i(G)\big)$$

if $\Bbbk$ is of characteristic p, and

$$L_n(G) = \sqrt{\gamma_n(G)} \cap [G, G]$$

if $\Bbbk$ is of characteristic 0.

DEFINITION 8.2. An N-series $\{G_n\}$ has *finite width* if there is a uniform constant W such that $l_n := \operatorname{rank} G_n/G_{n+1} \leqslant W$ holds for all n, where $\operatorname{rank} A$ is the minimal number of generators of the Abelian group A. A group has *finite width* if its lower central series has finite width – this definition comes from [KLP97].

The following result is well known, and shows that sometimes the Lie ring $\mathcal{L}(G)$ is actually a Lie algebra over $\mathbb{F}_p$.

LEMMA 8.3. *Let G be a group generated by a set S. Let $\mathcal{L}(G)$ be the Lie ring associated to the lower central series.*

(1) *If S is finite, then $\mathcal{L}_n$ is a finite-rank $\mathbb{Z}$-module for all n.*

(2) *If there is a prime p such that all generators $s \in S$ have order p, then the Lie algebra associated to the lower p-central series coincides with $\mathcal{L}$. As a consequence, $\mathcal{L}_n$ is a vector space over $\mathbb{F}_p$ for all n.*

We return to the lower p-central series of G. Consider the graded algebra

$$\overline{\mathbb{F}_p G} = \bigoplus_{n \in \mathbb{N}} \Delta^n / \Delta^{n+1}.$$

A fundamental result connecting $\mathcal{L}_{\mathbb{F}_p}(G)$ and $\overline{\mathbb{F}_p G}$ is the

PROPOSITION 8.4 (Quillen [Qui68]). *$\overline{\mathbb{F}_p G}$ is the enveloping p-algebra of $\mathcal{L}_{\mathbb{F}_p}(G)$.*

The Poincaré–Birkhoff–Witt theorem then gives a basis of $\overline{\mathbb{F}_p G}$ consisting of monomials over a basis of $\mathcal{L}_{\mathbb{F}_p}(G)$, with exponents at most $p-1$. As a consequence, we have the

PROPOSITION 8.5 (Jennings [Jen41]). *Let G be a group, and let $\sum_{n \geqslant 1} l_n \hbar^n$ be the Hilbert–Poincaré series of $\mathcal{L}_{\mathbb{F}_p}(G)$. Then*

$$\mathsf{growth}(\overline{\mathbb{F}_p G}) = \prod_{n=1}^{\infty} \left(\frac{1 - \hbar^{pn}}{1 - \hbar^n} \right)^{l_n}.$$

As a consequence, we have the following proposition, firstly observed by Bereznii (for a proof see [Pet99] and [BG00a]):

PROPOSITION 8.6. *Let G be a group and expand the power series* $\mathsf{growth}(\mathcal{L}_{\mathbb{F}_p}(G)) = \sum_{n \geqslant 1} l_n \hbar^n$ *and* $\mathsf{growth}(\overline{\mathbb{F}_p G}) = \sum_{n \geqslant 0} f_n \hbar^n$. *Then*

(1) *$\{f_n\}$ grows exponentially if and only if $\{l_n\}$ does, and we have*

$$\limsup_{n \to \infty} \frac{\ln l_n}{n} = \limsup_{n \to \infty} \frac{\ln f_n}{n}.$$

(2) *If $l_n \sim n^d$, then $f_n \sim \mathrm{e}^{n^{(d+1)/(d+2)}}$.*

Finally, we recall a connection between the growth of G and that of $\overline{\mathbb{F}_p G}$:

PROPOSITION 8.7 ([BG00a], Lemma 2.5). *Let G be a group generated by a finite set S. Then*

$$\frac{\mathsf{growth}(G)}{1 - \hbar} \geqslant \frac{\mathsf{growth}(\overline{\mathbb{F}_p G})}{1 - \hbar},$$

the inequality being valid coefficient-wise.

The following result exhibits a "gap in the spectrum" of growth, for residually-p groups:

COROLLARY 8.8 [Gri89,BG00a]. *Let G be a residually-p group for some prime p. Then the growth of G is either polynomial, in case G is virtually nilpotent, or is at least* $\mathrm{e}^{\sqrt{n}}$.

8.2. *Lie algebras of branch groups*

Our main purpose, in this section, is to illustrate the following result by examples:

THEOREM 8.9. *Let G be a finitely generated regular branch group and $\mathcal{L}G$ the Lie ring associated to its lower central series. Then* $\mathsf{growth}(\mathcal{L}G)$ *has polynomial growth* (*not necessarily of integer degree*).

Its proof relies on branch portraits, introduced in Section 1.4.

SKETCH OF PROOF. Let G be regular branch over K. The Lie algebra of G is isomorphic to that of $\overline{G}$, so we consider the latter. For each $n \in \mathbb{N}$, consider the set of branch portraits associated to $\gamma_n(G)$. Since K has finite index, it suffices to consider only the $\gamma_n(G) \leqslant K$. Let $n(\ell)$ be minimal such that the branch portraits of $\gamma_{n(\ell)}(G)$ are trivial in their first ℓ levels. It suffices to show that this function is exponential. Consider the portraits of $\gamma_n(G)$, for $n(\ell) \leqslant n \leqslant n(\ell+1)$. For n close to $n(\ell)$, there will be all portraits that are trivial except in a subtree at level ℓ. Then for larger n there will be, using commutation with a generator that is nontrivial at the root vertex, portraits trivial except in two subtrees, where they have labels P and P^{-1}, respectively. As n becomes larger and larger, the only remaining portraits will be those whose labels in all subtrees at level ℓ are identical. This passage from one level to the next is exponential. □

We obtain an explicit description of the lower central series in several cases, and show:

- For the first Grigorchuk group $\mathfrak{G}$, the Grigorchuk supergroup $\widetilde{\mathfrak{G}}$ and the Fabrykowski–Gupta group Γ, the Lie algebras L and $\mathcal{L}_{\mathbb{F}_p}$ have finite width.
- For the Gupta–Sidki group $\overline{\overline{\Gamma}}$, the Lie algebras L and $\mathcal{L}_{\mathbb{F}_p}$ have polynomial growth of degree $d = \log 3/\log(1+\sqrt{2}) - 1$.

The first result obtained in that direction is due to A. Rozhkov. He proved in [Roz96] that for the first Grigorchuk group $\mathfrak{G}$ one has

$$\operatorname{rank} \gamma_n(\mathfrak{G})/\gamma_{n+1}(\mathfrak{G}) = \begin{cases} 3 & \text{if } n = 1, \\ 2 & \text{if } n = 2^m + 1 + r, \text{ with } 0 \leqslant r < 2^{m-1}, \\ 1 & \text{if } n = 2^m + 1 + r, \text{ with } 2^{m-1} \leqslant r < 2^m. \end{cases}$$

However, the Lie algebra structure contained in an N-series $\{G_n\}$ is much richer than the series $\{\operatorname{rank} G_n/G_{n+1}\}$, and we will give a fuller description of the $\gamma_n(\mathfrak{G})$ below.

All our examples will satisfy the following conditions:

(1) G is finitely generated by a set S;

(2) there is a prime p such that all $s \in S$ have order p.

It then follows from Lemma 8.3 that $\gamma_n(G)/\gamma_{n+1}(G)$ is a finite-dimensional vector space over $\mathbb{F}_p$, and therefore that $\mathcal{L}(G)$ is a Lie algebra over $\mathbb{F}_p$ that is finite in each dimension. Clearly the same property holds for the restricted algebra $\mathcal{L}_{\mathbb{F}_p}(G)$.

We describe such Lie algebras in terms of oriented labelled graphs, in the following notation:

DEFINITION 8.10. Let $\mathcal{L} = \bigoplus_{n \geqslant 1} \mathcal{L}_n$ be a graded Lie algebra over $\mathbb{F}_p$, and choose a basis B_n and a scalar product $\langle | \rangle$ of $\mathcal{L}_n$ for all $n \geqslant 1$.

The *Lie graph* associated to these choices is an abstract graph. Its vertex set is $\bigcup_{n \geqslant 1} B_n$, and each vertex $x \in B_n$ has a degree, $n = \deg x$. Its edges are labelled as αx, with $x \in B_1$ and $\alpha \in \mathbb{F}_p$, and may only connect a vertex of degree n to a vertex of degree $n+1$. For all $x \in B_1$, $y \in B_n$ and $z \in B_{n+1}$, there is an edge labelled $\langle [x, y] | z \rangle x$ from y to z.

If $\mathcal{L}$ is a restricted algebra over $\mathbb{F}_p$, there are additional edges from vertices of degree n to vertices of degree pn. For all $x \in B_n$ and $y \in B_{pn}$, there is an edge labelled $\langle x^p | y \rangle \cdot p$ from x to y.

Edges labelled $0x$ are naturally omitted, and edges labelled $1x$ are simply written x.

There is some analogy between this definition and that of a Cayley graph – this topic will be developed in Section 10.3. The generators (in the Cayley sense) are simply chosen to be the $\mathsf{ad}(x)$ with x running through B_1, a basis of $G/[G, G]$.

As an example of Lie graph, let G be the infinite dihedral group $D_\infty = \langle a, b | a^2, b^2 \rangle$. Then $\gamma_n(G) = \langle (ab)^{2^{n-1}} \rangle$ for all $n \geqslant 2$, and its Lie ring is again a Lie algebra over $\mathbb{F}_2$, with Lie graph

$$\begin{tikzcd}[row sep=small]
a \arrow[dr, "b"] & & & & \\
& (ab)^2 \arrow[r, "{a,b}"] & (ab)^4 \arrow[r, "{a,b}"] & (ab)^8 \arrow[r, dashed, dash, "{a,b}"] & {} \\
b \arrow[ur, "a"'] & & & &
\end{tikzcd}$$

Note that the lower 2-central series of G is different: we have $G_{2^n} = G_{2^n+1} = \cdots = G_{2^{n+1}-1} = \gamma_{n+1}(G)$, so the Lie graph of $\mathcal{L}_{\mathbb{F}_2}(G)$ is

$$\begin{tikzcd}[row sep=small]
a \arrow[dr, "b"] & & & & \\
& (ab)^2 \arrow[r, "\cdot 2"] & (ab)^4 \arrow[r, "\cdot 2"] & (ab)^8 \arrow[r, dashed, dash, "\cdot 2"] & {} \\
b \arrow[ur, "a"'] & & & &
\end{tikzcd}$$

We shall also need the following notation: let G be a regular branch group over K, embedded in $G \wr (\mathbb{Z}/m\mathbb{Z})$. For all $i \in \mathbb{N}$ and all $g \in G$ define the maps

$$i(g) = \mathsf{ad}\big((1, \ldots, m)\big)^i (g, 1, \ldots, 1) = \big(g^{\binom{i}{0}}, g^{-\binom{i}{1}}, \ldots, g^{(-1)^{m-1}\binom{i}{m-1}}\big);$$

concretely, for $m = 2$ one has

$$0(g) = (g, 1), \qquad 1(g) = (g, g)$$

and for $m = 3$ one has

$$0(g) = (g, 1, 1), \quad 1(g) = (g, g^{-1}, 1),$$

$$2(g) = (g, g^{-2}, g) \equiv (g, g, g) \bmod \mho_3(G).$$

When m is prime, one clearly has $i(g) = 0$ for all $i \geqslant m$, and if $g \in K$ then $i(g) \in K$ for all $i \in \{0, 1, \ldots, d-1\}$.

8.2.1. *The group* $\mathfrak{G}$. We give an explicit description of the Lie algebra of $\mathfrak{G}$, and compute its Hilbert–Poincaré series. These results were obtained in [BG00a].

Set $x = (ab)^2$. Then $\mathfrak{G}$ is branch over $K = \langle x \rangle^{\mathfrak{G}}$, and $K/(K \times K)$ is cyclic of order 4, generated by x.

Extend the generating set of $\mathfrak{G}$ to a formal set $S = \left\{a, b, c, d, \left\{\begin{smallmatrix} b \\ c \end{smallmatrix}\right\}, \left\{\begin{smallmatrix} c \\ d \end{smallmatrix}\right\}, \left\{\begin{smallmatrix} d \\ b \end{smallmatrix}\right\}\right\}$, whose meaning shall be made clear later. Define the transformation σ on words in S^* by

$$\sigma(a) = a \begin{Bmatrix} b \\ c \end{Bmatrix} a, \quad \sigma(b) = d, \quad \sigma(c) = b, \quad \sigma(d) = c,$$

extended to subsets by $\sigma\left\{\begin{smallmatrix} x \\ y \end{smallmatrix}\right\} = \left\{\begin{smallmatrix} \sigma x \\ \sigma y \end{smallmatrix}\right\}$. Note that for any fixed $g \in G$, all elements $h \in \mathsf{St}_{\mathfrak{G}}(1)$ such that $\psi(h) = (g, *)$ are obtained by picking a letter from each set in $\sigma(g)$. This motivates the definition of S.

THEOREM 8.11. *Consider the following Lie graph*: *its vertices are the symbols* $X(x)$ *and* $X(x^2)$, *for words* $X \in \{0, 1\}^*$. *Their degrees are given by*

$$\deg X_1 \ldots X_n(x) = 1 + \sum_{i=1}^{n} X_i 2^{i-1} + 2^n,$$

$$\deg X_1 \ldots X_n(x^2) = 1 + \sum_{i=1}^{n} X_i 2^{i-1} + 2^{n+1}.$$

There are four additional vertices: a, b, d *of degree* 1, *and* $[a, d]$ *of degree* 2.

Define the arrows as follows: *an arrow labelled* $\left\{\begin{smallmatrix} x \\ y \end{smallmatrix}\right\}$ *stands for two arrows, labelled* x *and* y, *and the arrows labelled* c *are there to expose the symmetry of the graph* (*indeed* $c = bd$ *is not in our chosen basis of* $G/[G, G]$):

$$a \xrightarrow{b,c} x \qquad\qquad a \xrightarrow{c,d} [a, d]$$

$$b \xrightarrow{a} x \qquad\qquad d \xrightarrow{a} [a, d]$$

$$x \xrightarrow{a,b,c} x^2 \qquad\qquad x \xrightarrow{c,d} 0(x)$$

$$[a, d] \xrightarrow{b,c} 0(x) \qquad\qquad 0* \xrightarrow{a} 1*$$

$$1^n(x) \xrightarrow{\sigma^n\left\{\begin{smallmatrix} c \\ d \end{smallmatrix}\right\}} 0^{n+1}(x) \qquad\qquad 1^n(x) \xrightarrow{\sigma^n\left\{\begin{smallmatrix} b \\ d \end{smallmatrix}\right\}} 0^n(x^2)$$

$$1^n 0* \xrightarrow{\sigma^n\left\{\begin{smallmatrix} c \\ d \end{smallmatrix}\right\}} 0^n 1* \quad \textit{if } n \geqslant 1.$$

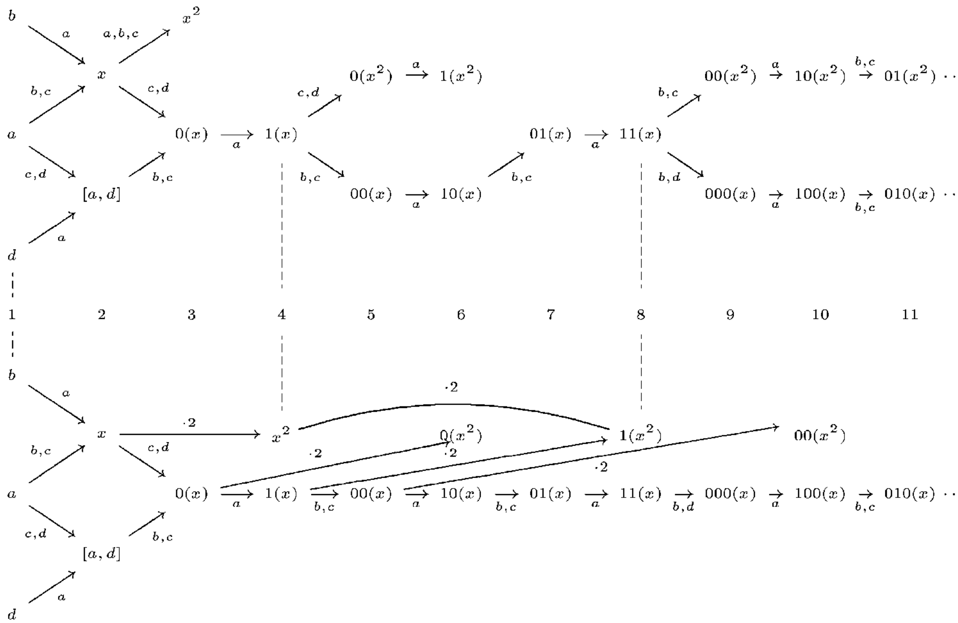


Table 25. The beginning of the Lie graphs of $\mathcal{L}(\mathfrak{G})$ (top) and $\mathcal{L}_{\mathbb{F}_2}(\mathfrak{G})$ (below).

Then the resulting graph is the Lie graph of $\mathcal{L}(\mathfrak{G})$. A slight modification gives the Lie graph of $\mathcal{L}_{\mathbb{F}_2}(\mathfrak{G})$: the degree of $X_1 \dots X_n(x^2)$ is $2 \deg X_1 \dots X_n(x)$; and the square maps are given by

$$X(x) \xrightarrow{\cdot 2} X(x^2),$$

$$1^n(x^2) \xrightarrow{\cdot 2} 1^{n+1}(x^2).$$

The subgraph spanned by a, t, the $X_1 \dots X_i(x)$ for $i \leqslant n-2$ and the $X_1 \dots X_i(x^2)$ for $i \leqslant n-4$ is the Lie graph associated to the finite quotient $\mathfrak{G}/\operatorname{St}_{\mathfrak{G}}(n)$.

PROOF. The proof proceeds by induction on length of words, or, what amounts to the same, on depth in the lower central series.

First, the assertion is checked "manually" up to degree 3. The details of the computations are the same as in [BG00a].

We claim that for all words X, Y with $\deg Y(x) > \deg X(x)$ we have $Y(x) \in \langle X(x) \rangle^{\mathfrak{G}}$, and similarly $Y(x^2) \in \langle X(x^2) \rangle^{\mathfrak{G}}$. The claim is verified by induction on $\deg X$.

We then claim that for any nonempty word X, either $\operatorname{ad}(a)X(*) = 0$ (if X starts with "1") or $\operatorname{ad}(v)X(*) = 0$ for $v \in \{b, c, d\}$ (if X starts with "0"). Again this holds by induction.

We then prove that the arrows are as described above. For instance, for the last one,

$$\mathsf{ad}\big(\sigma^n\{\begin{smallmatrix}c\\d\end{smallmatrix}\}\big)1^n0* = \begin{cases} \big(\mathsf{ad}\big(\sigma^n\{\begin{smallmatrix}d\\b\end{smallmatrix}\}\big)1^{n-1}0*, \mathsf{ad}\big(\{\begin{smallmatrix}a\\1\end{smallmatrix}\}\big)1^{n-1}0*\big) & \\ \quad = 0\,\mathsf{ad}\big(\sigma^{n-1}\{\begin{smallmatrix}c\\d\end{smallmatrix}\}\big)1^{n-1}0* = 0^n1* & \text{if } n \geqslant 2, \\ \big(\mathsf{ad}\big(\{\begin{smallmatrix}b\\c\end{smallmatrix}\}\big)0*, \mathsf{ad}(a)0*\big) = 01* & \text{if } n = 1. \end{cases}$$

Finally we check that the degrees of all basis elements are as claimed. Fix a word $X(*)$, and consider the largest n such that $X(*) \in \gamma_n(\mathfrak{G})$. Thus there is a sequence of $n-1$ arrows leading from the left of the Lie graph of $\mathcal{L}(\mathfrak{G})$ to $X(*)$, and no longer sequence, so $\deg X(*) = n$.

The modification giving the Lie graph of $\mathcal{L}_{\mathbb{F}_2}(\mathfrak{G})$ is justified by the fact that in $\mathcal{L}(\mathfrak{G})$ we always have $\deg X(x^2) \leqslant 2\deg X(x)$, so the element $X(x^2)$ appears always last as the image of $X(x)$ through the square map. The degrees are modified accordingly. Now $X(x^2) = X1(x^2)$, and $2\deg X1(x) \geqslant 4\deg X(x)$, with equality only when $X = 1^n$. This gives an additional square map from $1^n(x^2)$ to $1^{n+1}(x^2)$, and requires no adjustment of the degrees. $\square$

COROLLARY 8.12. *Define the polynomials*

$$Q_2 = -1 - \hbar,$$

$$Q_3 = \hbar + \hbar^2 + \hbar^3,$$

$$Q_n(\hbar) = (1+\hbar)Q_{n-1}(\hbar^2) + \hbar + \hbar^2 \quad \textit{for } n \geqslant 4.$$

Then Q_n is a polynomial of degree $2^{n-1}-1$, and the first $2^{n-3}-1$ coefficients of Q_n and Q_{n+1} coincide. The term-wise limit $Q_\infty = \lim_{n\to\infty} Q_n$ therefore exists.

The Hilbert–Poincaré series of $\mathcal{L}(\mathfrak{G}/\operatorname{St}_{\mathfrak{G}}(n))$ is $3\hbar + \hbar^2 + \hbar Q_n$, and the Hilbert–Poincaré series of $\mathcal{L}(\mathfrak{G})$ is $3\hbar + \hbar^2 + \hbar Q_\infty$.

The Hilbert–Poincaré series of $\mathcal{L}_{\mathbb{F}_2}(\mathfrak{G})$ is $3 + (2\hbar + \hbar^2)/(1-\hbar^2)$.

As a consequence, $\mathfrak{G}/\operatorname{St}_{\mathfrak{G}}(n)$ is nilpotent of class 2^{n-1}, and $\mathfrak{G}$ has finite width.

8.2.2. *The group Γ.* We give here an explicit description of the Lie algebra of Γ, and compute its Hilbert–Poincaré series.

THEOREM 8.13 [Bar00c]. *In Γ write $c = [a,t]$ and $u = [a,c] \equiv 2(at)$. For words $X = X_1 \dots X_n$ with $X_i \in \{0,1,2\}$ define symbols $\overline{X_1 \dots X_n}(c)$ (representing elements of Γ) by*

$$\overline{i0}(c) = i0(c)/i(u),$$

$$\overline{i2^{m+1}1^n}(c) = i\big(\overline{2^{m+1}1^n}(c) \cdot 01^m0^n(u)^{(-1)^n}\big), \quad \textit{and}$$

$$\overline{iX}(c) = i\overline{X}(c) \quad \textit{for all other } X.$$

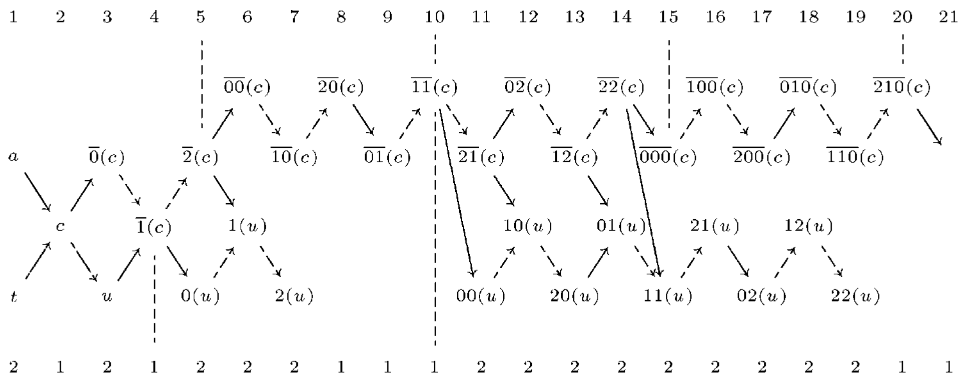


Table 26. The beginning of the Lie graph of $\mathcal{L}(\Gamma)$. The generator $\mathsf{ad}(t)$ is shown by plain/blue arrows, and the generator $\mathsf{ad}(a)$ is shown by dotted/red arrows.

Consider the following Lie graph: its vertices are the symbols $\overline{X}(c)$ and $X(u)$. Their degrees are given by

$$\deg \overline{X_1 \dots X_n}(c) = 1 + \sum_{i=1}^{n} X_i 3^{i-1} + \tfrac{1}{2}(3^n + 1),$$

$$\deg X_1 \dots X_n(u) = 1 + \sum_{i=1}^{n} X_i 3^{i-1} + (3^n + 1).$$

There are two additional vertices, labelled a and t, of degree 1.

Define the arrows as follows, for all $n \geqslant 1$:

$$\begin{array}{ll}
a \xrightarrow{-t} c & t \overset{a}{\dashrightarrow} c \\[1em]
c \xrightarrow{-t} 0(c) & c \xrightarrow{a} u \\[1em]
u \xrightarrow{-t} 1(c) & \overline{2^n}(c) \xrightarrow{-t} \overline{0^{n+1}}(c) \\[1em]
0{*} \overset{a}{\dashrightarrow} 1{*} & 1{*} \overset{a}{\dashrightarrow} 2{*} \\[1em]
2^n 0{*} \xrightarrow{t} 0^n 1{*} & 2^n 1{*} \xrightarrow{t} 0^n 2{*} \\[1em]
\multicolumn{2}{l}{\overline{X_1 \dots X_n}(c) \xrightarrow{-(-1)^{\sum X_i} t} (X_1 - 1) \dots (X_n - 1)(u)}
\end{array}$$

Then the resulting graph is the Lie graph of $\mathcal{L}(\Gamma)$.

The subgraph spanned by a, t, the $\overline{X_1 \dots X_i}(c)$ for $i \leqslant n-2$ and the $X_1 \dots X_i(u)$ for $i \leqslant n-3$ is the Lie graph associated to the finite quotient $\Gamma / \mathsf{St}_\Gamma(n)$.

COROLLARY 8.14. *Define the integers* $\alpha_n = \frac{1}{2}(5 \cdot 3^{n-2} + 1)$, *and the polynomials*

$$\begin{aligned} Q_2 &= 1, \\ Q_3 &= 1 + 2\hbar + \hbar^2 + \hbar^3 + \hbar^4 + \hbar^5 + \hbar^6, \\ Q_n(\hbar) &= (1 + \hbar + \hbar^2) Q_{n-1}(\hbar^3) + \hbar + \hbar^{\alpha_n - 2} \quad \textit{for } n \geqslant 4. \end{aligned}$$

Then Q_n *is a polynomial of degree* $\alpha_n - 2$, *and the first* $3^{n-2} + 1$ *coefficients of* Q_n *and* Q_{n+1} *coincide. The term-wise limit* $Q_\infty = \lim_{n\to\infty} Q_n$ *therefore exists.*

The Hilbert–Poincaré series of $\mathcal{L}(\Gamma / \mathsf{St}_\Gamma(n))$ *is* $2\hbar + \hbar^2 Q_n$, *and the Hilbert–Poincaré series of* $\mathcal{L}(\Gamma)$ *is* $2\hbar + \hbar^2 Q_\infty$.

As a consequence, $\Gamma / \mathsf{St}_\Gamma(n)$ *is nilpotent of class* α_n, *and* Γ *has finite width.*

In quite the same way as for $\overline{\overline{\Gamma}}$, we may improve the general result $\Gamma^{(k)} \leqslant \gamma_{2^k}(\Gamma)$:

THEOREM 8.15. *The derived series of* Γ *satisfies* $\Gamma' = \gamma_2(\Gamma)$ *and* $\Gamma^{(k)} = \gamma_5(\Gamma)^{\times 3^{k-2}}$ *for* $k \geqslant 2$. *We have for all* $k \in \mathbb{N}$

$$\Gamma^{(k)} \leqslant \gamma_{2+3^{k-1}}(\Gamma).$$

THEOREM 8.16. *Keep the notations of Theorem* 8.13. *Define now furthermore symbols* $\overline{X_1 \ldots X_n}(u)$ (*representing elements of* Γ) *by*

$$\overline{2^n}(u) = 2^n(u) \cdot 2^{n-1}0(c) \cdot 2^{n-2}01(c) \cdots 201^{n-2}(c),$$

and

$$\overline{X}(u) = X(u) \quad \textit{for all other } X.$$

Consider the following Lie graph: *its vertices are the symbols* $\overline{X}(c)$ *and* $\overline{X}(u)$. *Their degrees are given by*

$$\begin{aligned} &\deg \overline{X_1 \ldots X_n}(c) = 1 + \sum_{i=1}^{n} X_i 3^{i-1} + \tfrac{1}{2}(3^n + 1), \\ &\deg 2^n(u) = 3^{n+1}, \\ &\deg X_1 \ldots X_n(u) \\ &\quad = \max\left\{ 1 + \sum_{i=1}^{n} X_i 3^{i-1} + (3^n + 1), \tfrac{1}{2}(9 - 3^n) + 3 \sum_{i=1}^{n} X_i 3^{i-1} \right\}. \end{aligned}$$

There are two additional vertices, labelled a *and* t, *of degree* 1.

Define the arrows as follows, for all $n \geqslant 1$:

$$a \xrightarrow{-t} c \qquad t \overset{a}{\dashrightarrow} c$$

$$c \xrightarrow{-t} 0(c) \qquad c \xrightarrow{a} u$$

$$u \xrightarrow{-t} 1(c) \qquad \overline{2^n}(c) \xrightarrow{-t} \overline{0^{n+1}}(c)$$

$$0{*} \overset{a}{\dashrightarrow} 1{*} \qquad 1{*} \overset{a}{\dashrightarrow} 2{*}$$

$$2^n 0{*} \xrightarrow{t} 0^n 1{*} \qquad 2^n 1{*} \xrightarrow{t} 0^n 2{*}$$

$$\overline{X_1 \dots X_n}(c) \xrightarrow{-(-1)^{\sum X_i} t} (X_1 - 1)\dots(X_n - 1)(u)$$

$$c \xrightarrow{\cdot 3} \overline{00}(c) \qquad \overline{2^n}(u) \xrightarrow{\cdot 3} \overline{2^{n+1}}(u)$$

$$*0(c) \xrightarrow{\cdot 3} *2(u) \quad \textit{if } 3 \deg *0(c) = \deg *2(u)$$

Then the resulting graph is the Lie graph of $\mathcal{L}_{\mathbb{F}_3}(\Gamma)$.

The subgraph spanned by a, t, the $\overline{X_1 \dots X_i}(c)$ for $i \leqslant n-2$ and the $X_1 \dots X_i(u)$ for $i \leqslant n-3$ is the Lie graph of the Lie algebra $\mathcal{L}_{\mathbb{F}_3}(\Gamma/\mathsf{St}_\Gamma(n))$.

As a consequence, the dimension series of $\Gamma/\mathsf{St}_\Gamma(n)$ has length 3^{n-1} (the degree of $\overline{2^n}(u)$), and Γ has finite width.

8.2.3. *The group $\overline{\overline{\Gamma}}$.* We give here an explicit description of the Lie algebra of $\overline{\overline{\Gamma}}$, and compute its Hilbert–Poincaré series.

Introduce the following sequence of integers:

$$\alpha_1 = 1, \quad \alpha_2 = 2, \quad \alpha_n = 2\alpha_{n-1} + \alpha_{n-2} \quad \text{for } n \geqslant 3,$$

and $\beta_n = \sum_{i=1}^n \alpha_i$. One has

$$\alpha_n = \frac{1}{2\sqrt{2}}\big((1+\sqrt{2})^n - (1-\sqrt{2})^n\big),$$

$$\beta_n = \frac{1}{4}\big((1+\sqrt{2})^{n+1} + (1-\sqrt{2})^{n+1} - 2\big).$$

The first few values are

n	1	2	3	4	5	6	7	8
α_n	1	2	5	12	29	70	169	398
β_n	1	3	8	20	49	119	288	686

THEOREM 8.17 [Bar00c]. *In $\overline{\overline{\Gamma}}$ write $c = [a, t]$ and $u = [a, c] = 2(t)$. Consider the following Lie graph: its vertices are the symbols $X_1 \dots X_n(x)$ with $X_i \in \{0, 1, 2\}$ and $x \in \{c, u\}$. Their degrees are given by*

$$\deg X_1 \dots X_n(c) = 1 + \sum_{i=1}^{n} X_i \alpha_i + \alpha_{n+1},$$

$$\deg X_1 \dots X_n(u) = 1 + \sum_{i=1}^{n} X_i \alpha_i + 2\alpha_{n+1}.$$

There are two additional vertices, labelled a and t, of degree 1.

Define the arrows as follows:

$$a \xrightarrow{-t} c \qquad c \xrightarrow{t} 0(c)$$

$$t \stackrel{a}{\dashrightarrow} c \qquad c \stackrel{a}{\dashrightarrow} u$$

$$u \xrightarrow{t} 1(c)$$

$$0{*} \stackrel{a}{\dashrightarrow} 1{*} \qquad 1{*} \stackrel{a}{\dashrightarrow} 2{*}$$

$$2{*} \xrightarrow{t} 0\# \quad \textit{whenever } {*} \xrightarrow{t} \#$$

$$2(c) \xrightarrow{t} 1(u) \qquad 1(c) \xrightarrow{-t} 0(u)$$

$$10{*} \xrightarrow{-t} 01{*} \qquad 11{*} \xrightarrow{-t} 02{*}$$

$$20{*} \xrightarrow{t} 11{*} \qquad 21{*} \xrightarrow{t} 12{*}$$

(*Note that these last* 3 *lines can be replaced by the rules* $2{*} \xrightarrow{t} 1\#$ *and* $1{*} \xrightarrow{-t} 0\#$ *for all arrows* ${*} \xrightarrow{a} \#$.)

Then the resulting graph is the Lie graph of $\mathcal{L}(\overline{\overline{\Gamma}})$. It is also the Lie graph of $\mathcal{L}_{\mathbb{F}_3}(\overline{\overline{\Gamma}})$, with the only nontrivial cube maps given by

$$2^n(c) \xrightarrow{\cdot 3} 2^n 00(c), \qquad 2^n(c) \xrightarrow{\cdot 3} 2^n 1(u).$$

The subgraph spanned by a, t, the $X_1 \dots X_i(c)$ for $i \leqslant n - 2$ and the $X_1 \dots X_i(u)$ for $i \leqslant n - 3$ is the Lie graph associated to the finite quotient $\overline{\overline{\Gamma}} / \mathsf{St}_{\overline{\overline{\Gamma}}}(n)$.

COROLLARY 8.18. *Define the following polynomials*:

$$Q_1 = 0,$$
$$Q_2 = \hbar + \hbar^2,$$

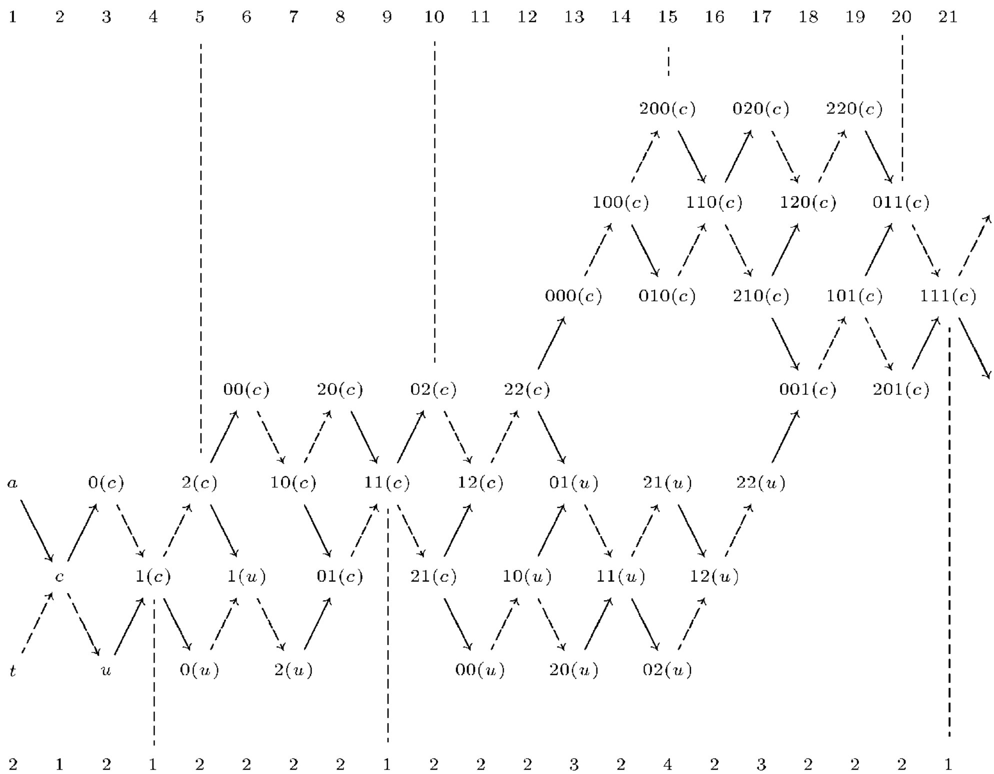


Table 27. The beginning of the Lie graph of $\mathcal{L}(\overline{\overline{\Gamma}})$. The generator $\mathsf{ad}(t)$ is shown by plain/blue arrows, and the generator $\mathsf{ad}(a)$ is shown by dotted/red arrows.

$$Q_3 = \hbar + \hbar^2 + 2\hbar^3 + \hbar^4 + \hbar^5,$$
$$Q_n = (1 + \hbar^{\alpha_n - \alpha_{n-1}})Q_{n-1} + \hbar^{\alpha_{n-1}}(\hbar^{-\alpha_{n-2}} + 1 + \hbar^{\alpha_{n-2}})Q_{n-2} \quad \textit{for } n \geqslant 3.$$

Then Q_n is a polynomial of degree α_n, and the polynomials Q_n and Q_{n+1} coincide on their first $2\alpha_{n-1}$ terms. The coefficient-wise limit $Q_\infty = \lim_{n\to\infty} Q_n$ therefore exists.

The largest coefficient in Q_{2n+1} is 2^n, at position $\frac{1}{2}(\alpha_{2n+1}+1)$, so the coefficients of Q_∞ are unbounded. The integers k such that $\hbar^k$ has coefficient 1 in Q_∞ are precisely the $\beta_n + 1$.

The Hilbert–Poincaré series of $\mathcal{L}(\overline{\overline{\Gamma}}/\operatorname{St}_{\overline{\overline{\Gamma}}}(n))$ is $\hbar + Q_n$, and the Hilbert–Poincaré series of $\mathcal{L}(\overline{\overline{\Gamma}})$ is $\hbar + Q_\infty$. The same holds for the Lie algebra $\mathcal{L}_{\mathbb{F}_3}(\overline{\overline{\Gamma}}/\operatorname{St}_{\overline{\overline{\Gamma}}}(n))$ and $\mathcal{L}_{\mathbb{F}_3}(\overline{\overline{\Gamma}})$.

As a consequence, $\overline{\overline{\Gamma}}/\operatorname{St}_{\overline{\overline{\Gamma}}}(n)$ is nilpotent of class α_n, and $\overline{\overline{\Gamma}}$ does not have finite width.

We note as an immediate consequence that

$$\left[\overline{\overline{\Gamma}} : \gamma_{\beta_n+1}(\overline{\overline{\Gamma}})\right] = 3^{\frac{1}{2}(3^n+1)},$$

so that the asymptotic growth of $l_n = \dim(\gamma_n(\overline{\overline{\Gamma}})/\gamma_{n+1}(\overline{\overline{\Gamma}}))$ is polynomial of degree $d = \log 3/\log(1+\sqrt{2}) - 1$, meaning that d is minimal such that

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} l_i}{\sum_{i=1}^{n} i^d} < \infty.$$

We then have by Proposition 8.6 the

COROLLARY 8.19. *The growth of* $\overline{\overline{\Gamma}}$ *is at least* $\mathrm{e}^{n^{(\log 3)/(\log(1+\sqrt{2})+\log 3)}} \cong \mathrm{e}^{n^{0.554}}$.

We may also improve the general result $\overline{\overline{\Gamma}}^{(k)} \leqslant \gamma_{2^k}(\overline{\overline{\Gamma}})$ to the following

THEOREM 8.20. *For all* $k \in \mathbb{N}$ *we have*

$$\overline{\overline{\Gamma}}^{(k)} \leqslant \gamma_{\alpha_{k+1}}(\overline{\overline{\Gamma}}).$$

8.3. *Subgroup growth*

For a finitely generated group G, its *subgroup growth function* is $a_n(G) = |\{H \leqslant G \mid [G:H] = n\}|$, and its *normal subgroup growth* is the function $b_n(G) = |\{N \triangleleft G \mid [G:N] = n\}|$. Building on their earlier papers [Seg86a,Seg86b,GSS88,MS90,LM91] A. Lubotzky, A. Mann and D. Segal obtained in [LMS93] a characterization of finitely generated groups with polynomial subgroup growth. Namely, the finitely generated groups of polynomial subgroup growth are precisely the virtually solvable groups of finite rank. For a well written survey refer to [Lub95b].

Since a_n and b_n only count finite-index subgroups, it is especially interesting to estimate the subgroup growth of just-infinite groups, and branch groups appear naturally in this context. A recent result lends support to that view:

THEOREM 8.21 [Seg01]. *Let* $f : \mathbb{N} \to \mathbb{R}_{\geqslant 0}$ *be nondecreasing and such that* $\log(f(n))/\log(n)$ *is unbounded as* $n \to \infty$. *Then there exists a* 4*-generator branch* (*spinal*) *group G whose subgroup growth is not polynomial, but satisfies*

$$a_n(G) \precsim n^{f(n)}.$$

In order to prove the above theorem D. Segal uses the construction described in Section 1.6.7 with $A_i = \mathsf{PSL}(2, p_i)$, $i \in \mathbb{N}$, as rooted subgroups and the action is the natural doubly transitive action of $\mathsf{PSL}(2, p_i)$ on a set of $p_i + 1$ elements. The subgroup growth function can be made slow by a choice of a sequence of primes $(p_i)_{i\in\mathbb{N}}$ that grows quickly enough. In addition, D. Segal shows that there exists a continuous range of possible "slow" subgroup growths.

9. Representation theory of branch groups

This section deals with the representation theory of branch groups and of their finite quotients over vector spaces of finite dimension and over Hilbert spaces. For infinite discrete groups the theory of infinite-dimensional unitary representations is a quite difficult subject. Fortunately, for branch groups the situation is easier to handle, and we produce several results, all taken from [BG00b], confirming this.

The study of unitary representations of profinite branch groups is of great importance and is motivated from several directions. For just-infinite branch groups it is equivalent to the study of irreducible representations of finite quotients $G/\operatorname{St}_G(\mathcal{L}_n)$. The first step in this direction is to consider the quasi-regular representations ρ_{G/P_n}, where P_n is the stabilizer of a point of level n, and this will be done below, again following [BG00b]. Another direction is the study of representations of all 3 types of groups (discrete branch p-groups, branch pro-p-groups, and their finite quotients) in vector spaces over the finite field $\mathbb{F}_p$. This last study was initiated by D. Passman and W. Temple in [PT96].

The spectral properties of the quasi-regular representations $\rho_{G/P}$ will be described in Section 11.

We introduce the following notion:

DEFINITION 9.1. Let G be a group, and $\Bbbk$ an algebraically closed field. We define $F_G(n) \in \mathbb{N} \cup \{\infty\}$ as the number of irreducible representations of degree at most n of G over $\Bbbk$. Similarly, $f_G(n)$ denotes the number of such representations of degree exactly n.

Therefore, $f_G(n)$ is the growth function of the representation ring of G over $\Bbbk$, whose degree-n component is generated by $\Bbbk G$-modules of dimension n, and whose addition and multiplication are $\oplus$ and $\otimes$.

First, we remark that if G is finitely generated, $\Bbbk$ is algebraically closed, and G does not have any char($\Bbbk$)-torsion, then $F_G(n)$ is finite for all n. This follows from a theorem of Weil (see [Far87]). We assume these conditions are satisfied by G.

If H is a finite-index subgroup of G, we have $F_G \sim F_H$, as shown in [PT96].

The first lower bound on $F(n)$ appears in a paper by D. Passman and W. Temple [PT96], where it was stated for the Gupta–Sidki p-groups. We improve slightly the result:

THEOREM 9.2 [PT96]. *Let G be a finitely-generated m-regular branch group over K, and consider its representations over any field $\Bbbk$. Then*

$$F_G \succsim n^{(m-1)\log_{[\psi(K):K^m]}[K:K']-1}.$$

The proof given in [PT96] extends easily to all regular branch groups. We note that this result is obtained by considering all possible inductions of degree-1 representations from K up to G; it may well be that the function F_G grows significantly faster than claimed, and the whole representation theory of $(K)\psi/K^m$ should be taken into account.

As a consequence, we obtain:

- $F_{\mathfrak{G}} \succsim n^2$, since $[K : K'] = 64$ and $[(K)\psi : K^2] = 4$;
- $F_{\Gamma} \succsim n^3$, since $[\Gamma' : \Gamma''] = 81$ and $[(\Gamma')\psi : (\Gamma')^2] = 9$;

- $F_{\overline{\overline{\Gamma}}} \succsim n^3$, for the same reason;
- for the Gupta–Sidki p-groups G_p of Section 1.6.3, the general result $F_{G_p} \succsim n^{p-2}$.

Note that since these groups are just-infinite, nonfaithful representations in vector spaces must factor through a finite quotient; and since these groups are of intermediate growth, they cannot be linear (by the Tits alternative), so in fact all finite-dimensional representations factor through a finite quotient of G, which may even be taken to be of the form $G_n = G/\operatorname{St}_G(\mathcal{L}_n)$ if G has the congruence subgroup property.

For concrete cases, like $\mathfrak{G}$ and Γ, we may exhibit some unitary irreducible representations as follows:

9.0.1. *Quasi-regular representations.* The representations we consider here are associated to parabolic subgroups, i.e. stabilizers of an infinite ray in the tree (see Section 7.3).

For G a group acting on a tree and P a parabolic subgroup, we let $\rho_{G/P}$ denote the quasi-regular representation of G, acting by right-multiplication on the space $\ell^2(G/P)$. This representation is infinite-dimensional, and a criterion for irreducibility, due to George Mackey, follows:

DEFINITION 9.3. The *commensurator* of a subgroup H of G is

$$\operatorname{comm}_G(H) = \{g \in G \mid H \cap H^g \text{ is of finite index in } H \text{ and } H^g\}.$$

Equivalently, letting H act on the left on the right cosets $\{gH\}$,

$$\operatorname{comm}_G(H) = \{g \in G \mid H \cdot (gH) \text{ and } H \cdot (g^{-1}H) \text{ are finite orbits}\}.$$

THEOREM 9.4 (Mackey [Mac76,BH97]). *Let G be an infinite group and let P be any subgroup of G. Then the quasi-regular representation $\rho_{G/P}$ is irreducible if and only if* $\operatorname{comm}_G(P) = P$.

The following results appear in [BG02]:

THEOREM 9.5. *If G is weakly branch, then* $\operatorname{comm}_G(P) = P$*, and therefore $\rho_{G/P}$ is irreducible.*

Note that the quasi-regular representations we consider are good approximants of the regular representation, in the sense that ρ_G is a subrepresentation of $\bigotimes_{P \text{ parabolic}} \rho_{G/P}$. We have a continuum of parabolic subgroups $P_e = \operatorname{St}_G(e)$, where e runs through the boundary of a tree, so we also have a continuum of quasi-regular representations. If G is countable, there are uncountably many non-equivalent representations, because among the uncountably many P_e only countably many are conjugate. As a consequence,

THEOREM 9.6. *There are uncountably many non-equivalent representations of the form $\rho_{G/P}$, where P is a parabolic subgroup.*

We now consider the finite-dimensional representations ρ_{G/P_n}, where P_n is the stabilizer of the vertex at level n in the ray defining P. These are permutational representations on the sets G/P_n of cardinality $m_1 \dots m_n$. The ρ_{G/P_n} are factors of the representation $\rho_{G/P}$. Noting that $P = \bigcap_{n \geqslant 0} P_n$, it follows that

$$\rho_{G/P_n} \Rightarrow \rho_{G/P},$$

in the sense that for any nontrivial $g \in G$ there is an $n \in \mathbb{N}$ with $\rho_{G/P_n}(g) \neq 1$.

We describe now the decomposition of the finite quasi-regular representations ρ_{G/P_n}. It turns out that it is closely related to the orbit structure of P_n on G/P_n. We state the result for the examples $\mathfrak{G}$, $\widetilde{\mathfrak{G}}$, Γ, $\overline{\Gamma}$, $\overline{\overline{\Gamma}}$:

THEOREM 9.7 [BG02]. *$\rho_{\mathfrak{G}/P_n}$ and $\rho_{\widetilde{\mathfrak{G}}/P_n}$ decompose as a direct sum of $n+1$ irreducible components, one of degree 2^i for each $i \in \{1, \dots, n-1\}$ and two of degree 1.*

ρ_{Γ/P_n}, $\rho_{\overline{\Gamma}/P_n}$ and $\rho_{\overline{\overline{\Gamma}}/P_n}$ decompose as a direct sum of $2n+1$ irreducible components, two of degree 2^i for each $i \in \{1, \dots, n-1\}$ and three of degree 1.

Part 4. Geometric and analytic aspects

10. Growth

The notion of growth in finitely generated groups was introduced by Efremovich in [Efr53] and Shvarts in [Šva55] in their study of Riemannian manifolds. The work of J. Milnor in the late sixties [Mil68b,Mil68a] contributed to current reinforced interest in the topic. Before we make brief historical remarks on the research made in connection to word growth in finitely generated groups, let us introduce the necessary definitions. We concentrate solely on finitely generated infinite groups.

Let $S = \{s_1, \dots, s_k\}$ be a nonempty set of symbols. A *weight function* on S is any function $\tau : S \to \mathbb{R}_{>0}$. Therefore, each symbol in S is assigned a positive weight. The *weight* of any word over S is then defined by the extension of τ to a homomorphism, still written $\tau : S^* \to \mathbb{R}_{\geqslant 0}$, defined on the free monoid S^* of words over S. Therefore, for any word over S we have

$$\tau(s_{i_1} s_{i_2} \dots s_{i_\ell}) = \sum_{j=1}^{\ell} \tau(s_{i_j}).$$

Note that the empty word is the only word of weight 0. For any non-negative real number n there are only finitely many words in S^* of weight at most n.

Let G be an infinite group and $\rho : S^* \twoheadrightarrow G$ a surjective monoid homomorphism. Therefore, G is finitely generated and $\rho(S) = \{\rho(s_1), \dots, \rho(s_k)\}$ generates G as a monoid. The *weight* of an element g in G with respect to the triple (S, τ, ρ) is, by definition, the smallest weight of a word u in S^* that represents g, i.e. the smallest weight of a word in $\rho^{-1}(g)$. The weight of g with respect to (S, τ, ρ) is denoted by $\partial_G^{(S,\tau,\rho)}(g)$.

For n a nonnegative real number, the elements in G that have weight at most n with respect to (S, τ, ρ) constitute the *ball* of radius n in G with respect to (S, τ, ρ), denoted by $B_G^{(S,\tau,\rho)}(n)$.

Let G acts transitively on a set X on the right, and let x be an element of X. We define $B_{x,G}^{(S,\tau,\rho)} = \{x^g \mid g \in B_G^{(S,\tau,\rho)}\}$ to be the *ball* in X of radius n with center at x.

The number of elements in $B_{x,G}^{(S,\tau,\rho)}(n)$ is finite and we denote it by $\gamma_{x,G}^{(S,\tau,\rho)}(n)$. The function $\gamma_{x,G}^{(S,\tau,\rho)}$, defined on the non-negative real numbers, is called the *growth function* of X at x as a G-set with respect to (S,τ,ρ).

The equivalence class of $\gamma_{x,G}^{(S,\tau,\rho)}$ under $\sim$ (recall the notation from the introduction) is called the *degree of growth* of G and it does not depend on the (finite) set S, the weight function τ defined on S, the homomorphism ρ, nor the choice of x.

PROPOSITION 10.1 (invariance of the growth function). *If $\gamma_{x,G}^{(S,\tau,\rho)}$ and $\gamma_{x',G}^{(S',\tau',\rho')}$ are two growth functions of the group G, then they are equivalent with respect to $\sim$.*

When we define a weight function on a group G we usually pick a finite generating subset of G closed for inversion and not containing the identity, assign a weight function to those generating elements and extend the weight function to the whole group G in a natural way, thus blurring the distinction between a word over the generating set and the element in G represented by that word, and completely avoiding the discussion of ρ.

A standard way to assign a weight function is to assign the weight 1 to each generator. In that case we use the standard notation and terminology, i.e. we denote the weight of a word u by $|u|$ and call it the *length* of u. In this setting, the length of the group element g is the distance from g to the identity in the Cayley graph of the group with respect to the generating set S.

If we let G act on itself by right multiplication we see that the growth function of G just counts the number of elements in the corresponding ball in G, i.e. $\gamma_{1,G}^{(S,\tau,\rho)}(n) = |B_G^{(S,\tau,\rho)}(n)|$. Since the degree of growth is an invariant of the group we are more interested in it than in the actual growth function for a given generating set.

Consider a couple of examples. In the next few examples the weight 1 is assigned to the generating elements. If $G = \mathbb{Z}$ and $S = \{1, -1\}$ then $\gamma_G(n) = 2n + 1 \sim n$. More generally, if G is a free Abelian group of rank k, we have $\gamma_G \sim n^k$. If G is the free group of rank $k \geqslant 2$ with the standard generating set together with the inverses, then $\gamma_G(n) = (k(2k-1)^n - 1)/(k-1) \sim \mathrm{e}^n$. It is clear that the growth functions of groups on k generators are bounded above by the growth function of the free group on k generators. Therefore, an exponential degree of growth is the largest possible degree of growth.

For any finitely generated infinite group G, the following trichotomy exists: G is of

- *polynomial growth* if $\gamma_G(n) \precsim n^d$, for some $d \in \mathbb{N}$;
- *intermediate growth* if $n^d \precnsim \gamma_G(n) \precnsim \mathrm{e}^n$, for all $d \in \mathbb{N}$;
- *exponential growth* if $\mathrm{e}^n \sim \gamma_G(n)$.

We say G is of *subexponential growth* if $\gamma_G(n) \precnsim \mathrm{e}^n$ and of *superpolynomial growth* if $n^d \precnsim \gamma_G(n)$ for all $d \in \mathbb{N}$.

By the results of J. Milnor, J. Wolf and B. Hartley (see [Mil68a] and [Wol68]), solvable groups have exponential growth unless they are virtually nilpotent in which case the growth is polynomial. There is a formula giving the degree of polynomial growth in terms of the lower central series of G, due to Y. Guivarc'h and H. Bass [Gui70,Bas72]. Namely, if d_j

represents the torsion-free rank of the j-th factor in the lower central series of the finitely generated nilpotent group G, then the degree of growth of G is polynomial of degree $\sum jd_j$.

By the Tits alternative [Tit72], a finitely generated linear group is either virtually solvable or it contains the free group of rank two. Therefore, the growth of a linear group must be exponential or polynomial. In the same spirit, Ching Chou showed in [Cho80] that the growth of elementary amenable groups must be either polynomial or exponential (recall that the class of elementary groups is the smallest class containing all finite and all Abelian groups that is closed under subgroups, homomorphic images, extensions and directed unions). Any nonelementary hyperbolic group must have exponential growth [GH90].

A fundamental result of M. Gromov [Gro81] states that every group of polynomial growth is virtually nilpotent. The results of J. Milnor, J. Wolf, Y. Guivarc'h, H. Bass and M. Gromov together imply that a group has a polynomial growth if and only if it is virtually nilpotent, and in this case the growth function is equivalent to n^d where d is the integer $\sum jd_j$, as stated above.

We recall the definition of amenable group:

DEFINITION 10.2. Let G be a group acting on a set X. This action is *amenable* in the sense of von Neumann [vN29] if there exists a finitely additive measure μ on X, invariant under the action of G, with $\mu(X) = 1$.

A group G is *amenable* if its action on itself by right-multiplication is amenable.

The following criterion, due to Følner, can sometimes be used to show that a certain action is amenable.

THEOREM 10.3. *Let G act transitively on a set X and let S be a generating set for G. The action is amenable if and only if, for every positive real number λ there exists a finite set F such that $|F \Delta Fs| < \lambda|F|$, for all $s \in S$, where Δ denotes the symmetric difference.*

Using the Følner's criterion one can show that groups of subexponential growth are always amenable.

In [Mil68c] J. Milnor asked the following question: "Is the function $\gamma(n)$ necessarily equivalent either to a power of n or to the exponential function e^n?". In other words, Milnor asked if the growth is always polynomial or exponential.

The first examples of groups of intermediate growth were constructed by the second author in [Gri80], and they are known as the first and the second Grigorchuk group. Both examples are 2-groups of intermediate growth and they are both amenable but not elementary. These examples show that the answer to Milnor's question is "no". In the same time, these examples show that there exist amenable but not elementary amenable groups, thus answering a question of M. Day from [Day57].

Other examples of groups of intermediate growth are Γ, see Section 7.5 and the first torsion-free example from [Gri85a], which is based on the first Grigorchuk group $\mathfrak{G}$.

It was shown in [Gri84,Gri85a] that the Grigorchuk p-groups are of intermediate growth and that there are uncountably large chains and antichains of growth functions associated with these examples, which are all branch groups.

An important unanswered question in the theory of groups of intermediate growth is the existence of a group whose degree of growth is $e^{\sqrt{n}}$. We remarked before that residually-p groups that have degree of growth strictly below $e^{\sqrt{n}}$ must be virtually nilpotent, and therefore of polynomial growth (see 8.8). The same conclusion holds for residually nilpotent groups (see [LM91]).

The historical remarks on the research on growth above are biased towards the existence and development of examples of groups of intermediate growth. There is plenty of great research on growth in group theory that is not concerned with this aspect. An excellent review of significant results and a good bibliography can be found in [Har00] and [GH97].

Before we move on to more specific examples, we make an easy observation.

PROPOSITION 10.4. *No weakly branch group has polynomial growth, i.e. no weakly branch group is virtually nilpotent.*

10.1. *Growth of* G *groups with finite directed part*

We will concentrate on the case of G groups with finite directed part B. Note that all branching indices, except maybe the first one, are bounded above by $|B|$ since each homomorphic image $A_{\sigma^r\omega} = (B)\omega_r$ acts transitively on a set of m_{r+1} elements. Therefore, $|B| \geqslant |A_{\sigma^r\omega}| \geqslant m_{r+1}$, for all r. Let us denote, once and for all, the largest branching index by M and the smallest one by m.

All the estimates of word growth that we give in this and the following sections are done with respect to the canonical generating set $S_\omega = (A_\omega \cup B_\omega) - 1$. As a shorthand, we use $\gamma_\omega(n)$ instead of $\gamma_{G_\omega}(n)$.

In order to express some of the results we use the notions of complete subsequence and r-homogeneous and r-factorable sequence (see the remarks before Theorem 6.2). We recall that all sequences in $\widehat{\Omega}$, i.e. sequences that define G groups (see Definition 2.6), can be factored into finite complete subsequences.

THEOREM 10.5. *All* G *groups with finite directed part have subexponential growth.*

The following lemma is a direct generalization of [Gri85a, Lemma 1]. The proof is similar, but adapted to our more general setting.

LEMMA 10.6 (3/4-shortening). *Let $\overline{\omega}$ be a sequence that starts with a complete sequence of length r. Then the following inequality holds for every reduced word F representing an element in $\mathsf{St}_\omega(\mathcal{L}_r)$:*

$$\big|L_r(F)\big| \leqslant \tfrac{3}{4}|F| + M^r,$$

where $|L_r(F)|$ represents the total length of the words on the level r of the r-level decomposition of F.

PROOF. Define ξ_i to be the number of B-letters from $K_i \setminus (K_{i-1} \cup \cdots \cup K_1)$ appearing in the words at the level $i-1$, and ν_i to be the number of simple reductions performed to get the words $F_{j_1 \ldots j_i}$ on the level i from their unreduced versions $\overline{F_{j_1 \ldots j_i}}$.

A reduced word F of length n has at most $(n+1)/2$ B-letters. Every B-letter in F that is in K_1 contributes one B-letter and no A-letters to the unreduced words $\overline{F_1}, \ldots, \overline{F_{m_1}}$. The B-letters in F that are not in K_1, and there are at most $(n+1)/2 - \xi_1$ such letters, contribute one B-letter and one A-letter. Finally, the ν_1 simple reductions reduce the number of letters on level 1 by at least ν_1. Therefore,

$$\left|L_1(F)\right| \leqslant 2\bigl((n+1)/2 - \xi_1\bigr) + \xi_1 - \nu_1 = n + 1 - \xi_1 - \nu_1.$$

In the same manner, each of the ξ_2 B-letters on level 1 that is in $K_2 \setminus K_1$ contributes one B-letter to the unreduced words on level 2. The other B-letters, and there are at most $(|L_1(F)| + M)/2 - \xi_2$ of them, contribute at most 2 letters, so, after simple reductions, we have

$$\left|L_2(F)\right| \leqslant n + 1 + M - \xi_1 - \xi_2 - \nu_1 - \nu_2.$$

Proceeding in the same manner, we obtain the estimate

$$\begin{aligned} \left|L_r(F)\right| \leqslant n + 1 + M + \cdots + M^{r-1} - \xi_1 - \xi_2 - \cdots - \xi_r \\ - \nu_1 - \nu_2 - \cdots - \nu_r. \end{aligned} \tag{20}$$

If $\nu_1 + \nu_2 + \cdots + \nu_r \geqslant n/4$, the claim of the lemma immediately follows.

Let us therefore consider the case when

$$\nu_1 + \nu_2 + \cdots + \nu_r < n/4. \tag{21}$$

For $i = 0, \ldots, r-1$, define $|L_i(F)|^+$ to be the number of B-letters from $B \setminus (K_1 \cup \cdots \cup K_i)$ appearing in the words at the level i. Clearly, $|L_0(F)|^+$ is the number of B-letters in F and

$$\left|L_0(F)\right|^+ \geqslant \frac{n-1}{2}.$$

Going from the level 0 to the level 1, each B-letter contributes one B letter of the same type. Therefore, the words $\overline{F_1}, \ldots, \overline{F_{m_1}}$ from the first level before the reduction takes place have exactly $|L_0(F)|^+ - \xi_1$ letters that come from $B - K_1$. Since we lose at most $2\nu_1$ letters due to the simple reductions, we obtain

$$\left|L_1(F)\right|^+ \geqslant \frac{n-1}{2} - \xi_1 - 2\nu_1.$$

Next, we go from level 1 to level 2. There are $|L_1(F)|^+$ B-letters on level 1 that come from $B \setminus K_1$, so there are exactly $|L_1(F)|^+ - \xi_2$ B-letters from $B \setminus (K_1 \cup K_2)$ in the words

$\overline{F_{11}}, \ldots, \overline{F_{m_1 m_2}}$, and then we lose at most $2\nu_2$ B-letters due to the simple reductions. We get

$$\left|L_2(F)\right|^+ \geqslant \frac{n-1}{2} - \xi_1 - \xi_2 - 2\nu_1 - 2\nu_2,$$

and, by proceeding in a similar manner,

$$\left|L_{r-1}(F)\right|^+ \geqslant \frac{n-1}{2} - \xi_1 - \cdots - \xi_{r-1} - 2\nu_1 - \cdots - 2\nu_{r-1}. \tag{22}$$

Since $\omega_1 \ldots \omega_r$ is complete, we have $K_r \setminus (K_1 \cup \cdots \cup K_{r-1}) = B \setminus (K_1 \cup \cdots \cup K_{r-1})$ and $\xi_r = |L_{r-1}(F)|^+$ so that the inequalities (20), (21) and (22) yield

$$\left|L_r(F)\right| \leqslant \frac{n}{2} + \frac{1}{2} + 1 + M + \cdots + M^{r-1} + \nu_1 + \cdots + \nu_{r-1} - \nu_r,$$

which implies the claim. □

Now we can finish the proof of Theorem 10.5 using the approach used by the second author in [Gri85a] (see also [Har00, Theorem VIII.61]). We use the following easy lemma, which follows from the fact every subgroup of index L has a transversal whose representatives have length at most $L-1$ (use a Schreier transversal).

LEMMA 10.7. *Let G be a group and H be a subgroup of finite index L in G. Let $\gamma(n)$ denote the growth function of G with respect to some finite generating set S and the standard length function on S, and let $\beta(n)$ denote the number of words of length at most n that are in H, i.e.*

$$\beta(n) = \left|\left\{g \mid g \in H,\ |g| \leqslant n\right\}\right| = \left|B_G^S(n) \cap H\right|.$$

Then

$$\gamma(n) \leqslant L\beta(n + L - 1).$$

Let

$$e_\omega = \lim_{n\to\infty} \sqrt[n]{\gamma_\omega(n)}$$

denote the *exponential growth rate* of G_ω. It is known that this rate is 1 if and only if the group in question has subexponential growth. Therefore, all we need to show is that this rate is 1.

PROOF OF THEOREM 10.5. For any $\varepsilon > 0$ there exists n_0 such that

$$\gamma_\omega(n) \leqslant (e_\omega + \varepsilon)^n,$$

for all $n \geqslant n_0$. If we denote $C_\omega = \gamma_\omega(n_0)$, we obtain

$$\gamma_\omega(n) \leqslant C_\omega(e_\omega + \varepsilon)^n,$$

for all n. Note that C_ω depends on ε.

Let ω start with a complete sequence of length r. Denote

$$\beta_\omega(n) = \left|\left\{g \mid g \in \mathsf{St}_\omega(\mathcal{L}_r),\ |g| \leqslant n\right\}\right|.$$

By Lemma 10.6 and the fact that ψ_r is an embedding we have

$$\beta_\omega(n) \leqslant \sum \gamma_{\sigma^r\omega}(n_1)\gamma_{\sigma^r\omega}(n_2)\ldots\gamma_{\sigma^r\omega}(n_s),$$

where $s = m_1 \cdot m_2 \cdots m_r$, and the summation is over all tuples $(n_1, \ldots, n_s)$ of non-negative integers with $n_1 + n_2 + \cdots + n_s \leqslant \frac{3}{4}n + M^r$. Let L be the index of $\mathsf{St}_\omega(\mathcal{L}_r)$ in G_ω. By the previous lemma and the above discussions we have

$$\gamma_\omega(n) \leqslant L\beta_\omega(n + L - 1) \leqslant LC^s_{\sigma^r\omega} \sum (e_{\sigma^r\omega} + \varepsilon)^{n_1+\cdots+n_s},$$

where the summation is over all tuples $(n_1, \ldots, n_s)$ of non-negative integers with $n_1 + \cdots + n_s \leqslant \frac{3}{4}(n + L - 1) + M^r$. The number of such tuples is polynomial $P(n)$ in n (depending also on the constants L, M and r). Therefore,

$$\gamma_\omega(n) \leqslant LC^s_{\sigma^r\omega}P(n)(e_{\sigma^r\omega} + \varepsilon)^{\frac{3}{4}(n+L-1)+M^r}.$$

Taking the n-th root on both sides and the limit as n tends to infinity gives

$$e_\omega \leqslant (e_{\sigma^r\omega} + \varepsilon)^{3/4},$$

and since this inequality holds for all positive ε, we obtain

$$e_\omega \leqslant (e_{\sigma^r\omega})^{3/4}.$$

Since the exponential growth rate $e_{\sigma^t\omega}$ is bounded above by $|A_{\sigma^t\omega}| + |B_{\sigma^t\omega}| - 1 \leqslant 2|B| - 1$, for all $t > 0$, it follows that $e_\omega = 1$ for all $\omega \in \widehat{\Omega}$. □

A general lower bound, tending to e^n when $m \to \infty$, exists on the word growth, and holds for all G groups:

THEOREM 10.8. *All* G *groups with finite directed part have superpolynomial growth. Moreover, the growth of* G_ω *satisfies*

$$\mathrm{e}^{n^\alpha} \precsim \gamma_\omega(n),$$

where $\alpha = \log(m)/(\log(m) - \log\frac{1}{2})$.

A proof can be found in [BŠ01] for the more special case that is considered there. Note that $\alpha > \frac{1}{2}$, as long as $m \neq 2$. Putting the last several results together gives

THEOREM 10.9. *All* G *groups with finite directed part have intermediate growth.*

For the special case of $\mathfrak{G}$ slightly better lower bounds exist, due to Y. Leonov [Leo98b] who obtained $\mathrm{e}^{n^{0.5041}} \preceq \gamma(n)$, and to the first author [Bar01] who obtained $\mathrm{e}^{n^{0.5157}} \preceq \gamma(n)$.

10.2. *Growth of* G *groups defined by homogeneous sequences*

Implicitly, all defining triples ω have r-homogeneous defining sequence $\overline{\omega}$ for some fixed r. The following estimate holds:

THEOREM 10.10 (η-estimate). *If $\overline{\omega}$ is an r-homogeneous sequence, then the growth function of the group G_ω satisfies*

$$\gamma_\omega(n) \precsim \mathrm{e}^{n^\alpha},$$

where $\alpha = \log(M)/(\log(M) - \log(\eta_r)) < 1$ *and η_r is the positive root of the polynomial* $x^r + x^{r-1} + x^{r-2} - 2$.

We mentioned already that the weight assignment is irrelevant as far as the degree of growth is concerned. But, appropriately chosen weight assignments can make calculations easier, and this is precisely how the above estimates are obtained.

Let G be a group that is generated as a monoid by the set of generators S that does not contain the identity. A weight function τ on S is called triangular if $\tau(s_1) + \tau(s_2) \geqslant \tau(s_3)$ whenever $s_1, s_2, s_3 \in S$ and $s_1 s_2 = s_3$.

In the case of spinal groups, in order to define a triangular weight we must have

$$\tau(a_1) + \tau(a_2) \geqslant \tau(a_1 a_2) \quad \text{and} \quad \tau(b_1) + \tau(b_2) \geqslant \tau(b_1 b_2),$$

for all $a_1, a_2 \in A$ and $b_1, b_2 \in B$ such that $a_1 a_2 \neq 1$ and $b_1 b_2 \neq 1$.

Every g in G_ω admits a minimal form with respect to a triangular weight τ

$$[a_0] b_1 a_1 b_2 a_2 \ldots a_{k-1} b_k [a_k],$$

where all a_i are in $A - 1$, all b_i are in $B - 1$, and the appearances of a_0 and a_k are optional.

The following weight assignment generalizes the approach taken in [Bar98] by the first author in order to estimate the growth of $\mathfrak{G}$ (see also [BŠ01]).

The linear system of equations in the variables $\tau_0, \ldots, \tau_r$:

$$\begin{cases} \eta_r(\tau_0 + \tau_i) = \tau_0 + \tau_{i-1} & \text{for } i = r, \ldots, 2, \\ \eta_r(\tau_0 + \tau_1) = \tau_r, \end{cases} \tag{23}$$

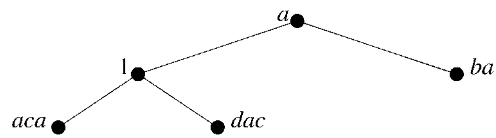


Table 28. Portrait of g of size 3 (same as the portrait of size 4).

has a solution, up to a constant multiple, given by

$$\begin{cases} \tau_i = \eta_r^r + \eta_r^{r-i} - 1 & \text{for } i = r, \dots, 1, \\ \tau_0 = 1 - \eta_r^r. \end{cases} \tag{24}$$

We also require $\tau_1 + \tau_2 = \tau_r$ and we get that η_r must be a root of the polynomial $x^r + x^{r-1} + x^{r-2} - 2$. We choose η_r to be the root of this polynomial that is between 0 and 1 obtain that the solution (24) of the system (23) satisfies the additional properties

$$0 < \tau_1 < \cdots < \tau_r < 1, \quad 0 < \tau_0 < 1, \tag{25}$$

$$\tau_i + \tau_j \geqslant \tau_k \quad \text{for all } 1 \leqslant i, j, k \leqslant r \text{ with } i \neq j. \tag{26}$$

The index r in η_r will be omitted without warning.

Now, given $\omega \in \Omega^{(r)}$, we define the weight of the generating elements in S_ω as follows: $\tau(a) = \tau_0$, for a in $A - 1$ and $\tau(b_\omega) = \tau_i$, where i is the smallest index with $(b)\omega_i = 1$, i.e. the smallest index with $b \in K_i = \mathrm{Ker}(\omega_i)$.

Clearly, τ is a triangular weight function. The only point worth mentioning is that if b and c are two B-letters of the same weight and $bc = d \neq 1$ then d has no greater weight than b or c (this holds because $b_\omega, c_\omega \in K_i$ implies $d_\omega \in K_i$).

For obvious reasons, the weight $\partial_{G_\omega}^{(S_\omega,\tau,\rho)}(g)$, for $g \in G_\omega$, is denoted by $\partial^\tau(g)$ and, more often, just by $\partial(g)$.

We define the *portrait of an element of G_ω of size n* with respect to the weight τ as the portrait with respect to the sequence of profile sets $B^\tau_{G_\omega}(n)$, $B^\tau_{G_{\sigma\omega}}(n), \dots,$ which is the sequence of balls of radius n in the corresponding companion groups. Therefore, in the process of building a portrait only those vertices that would be decorated by an element that has weight larger than n are decomposed further at least one more level and become interior vertices decorated by vertex permutations.

Just as an easy example, we give the portraits of size 3 and size 2 with respect to the standard word length of the element $g = abacadacabadac$ in $\mathfrak{G}$ in Figures 28 and 29. Other portraits of the same element are given in Section 2.1.4.

The following lemma says that the total sum of the weights of the sections of an element f is significantly shorter (by a factor less than 1) than the weight of f. This observation leads to upper bounds on the word growth in G groups.

LEMMA 10.11 (η-shortening). *Let $f \in G_\omega$ and $\overline{\omega}$ be an r-homogeneous sequence. Then*

$$\sum_{i=1}^{m_1} \partial^\tau(f_i) \leqslant \eta_r\big(\partial^\tau(f) + \tau_0\big).$$

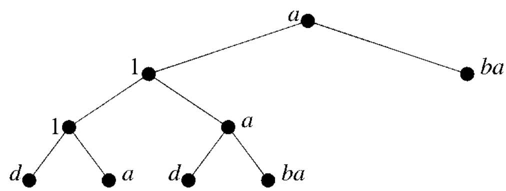


Table 29. Portrait of g of size 2.

PROOF. Let a minimal form of f be

$$f = [a_0]b_1a_1 \dots b_{k-1}a_{k-1}b_k[a_k].$$

Further let $f = hg$ where $g \in A_\omega$ and $h \in \mathsf{St}_\omega(\mathcal{L}_1)$. Then h can be written in the form $h = [a_0]b_1a_1 \dots a_{k-1}b_k[a_k]g^{-1}$ and rewritten in the form

$$h = b_1^{g_1} \dots b_k^{g_k}, \tag{27}$$

where $g_i = ([a_0]a_1 \dots a_{i-1})^{-1} \in A_\omega$. Clearly, $\partial(f) \geqslant (k-1)\tau_0 + \sum_{j=1}^{k} \tau(b_j)$, which yields

$$\sum_{i=1}^{k} \eta\bigl(\tau_0 + \tau(b_j)\bigr) \leqslant \eta\bigl(\partial(f) + \tau_0\bigr). \tag{28}$$

Now, observe that if the B-generator b is of weight τ_i with $i > 1$ then $(b^g)\psi$ has as components one B-generator of weight τ_{i-1} and one A-generator (of weight τ_0 of course) with the rest of the components trivial. Therefore, such a b^g (from (27)) contributes at most $\tau_0 + \tau_{i-1} = \eta(\tau_0 + \tau(b))$ to the sum $\sum \partial(f_i)$. On the other hand, if b is a B-generator of weight τ_1 then $(b^g)\psi$ has as components one B-generator of weight at most τ_r, and the rest of the components are trivial. Such a b^g contributes at most $\tau_r = \eta(\tau_0 + \tau(b))$ to the sum $\sum \partial(f_i)$. Therefore

$$\sum_{i=1}^{m_1} \partial(f_i) \leqslant \sum_{j=1}^{k} \eta\bigl(\tau_0 + \tau(b_j)\bigr) \tag{29}$$

and the claim of the lemma follows by combining (28) and (29). □

For a chosen C we can now construct the portraits of size C of the elements of G_ω. In case C is large enough, the previous lemma guarantees that these portraits are finite.

LEMMA 10.12. *There exists a positive constant C such that*

$$L(n) \precsim n^\alpha,$$

with $\alpha = \log(M)/(\log(M) - \log(\eta_r))$, *where* $L(n)$ *is the maximal possible number of leaves in the portrait of size* C *of an element of weight at most* n.

We give a proof that does not work exactly, but gives the right idea. One can find a complete proof in [BŠ01].

SKETCHY PROOF. We present how the proof would work if

$$\sum_{i=1}^{m_1} \partial(f_i) \leqslant \eta\partial(f) \tag{30}$$

holds rather than the inequality in Lemma 10.11.

We choose C big enough so that the portraits of size C are finite. Define a function $L'(n)$ on $\mathbb{R}_{\geqslant 0}$ by

$$L'(n) = \begin{cases} 1 & \text{if } n \leqslant C, \\ n^\alpha & \text{if } n > C. \end{cases}$$

We prove, by induction on n, that $L(n) \leqslant L'(n)$. If the weight n of g is $\leqslant C$, the portrait has 1 leaf and $L'(n) = 1$. Otherwise, the portrait of g is made up of those of $g_1, \dots, g_{m_1}$. Let the weights of these m_1 elements be $n_1, \dots, n_{m_1}$. By induction, the number of leaves in the portrait of g_i is at most $L'(n_i)$, $i = 1, \dots, m_1$, and the number of leaves in the portrait of g is, therefore, at most $\sum_{i=1}^{m_1} L'(n_i)$.

There are several cases, but let us just consider the case when all of the numbers $n_1, \dots, n_{m_1}$ are greater than C. Using Jensen's inequality, the inequality (30), the facts that $\eta^\alpha = M^{\alpha-1}$, $0 < \alpha < 1$, and direct calculation, we see that

$$\begin{aligned} \sum_{i=1}^{m_1} L'(n_i) &= \sum_{i=1}^{m_1} n_i^\alpha \leqslant m_1 \left(\frac{1}{m_1} \sum_{i=1}^{m_1} n_i \right)^\alpha \\ &\leqslant \frac{1}{(m_1)^{\alpha-1}} (\eta n)^\alpha = \left(\frac{M}{m_1} \right)^{\alpha-1} n^\alpha \leqslant n^\alpha = L'(n). \end{aligned}$$ □

PROOF OF THEOREM 10.10. The number of labelled, rooted trees with at most $L(n)$ leaves, whose branching indices do not exceed M, is $\precsim e^{n^\alpha}$. A tree with $L(n)$ leaves has $\sim L(n)$ interior vertices, so there are $\precsim e^{n^\alpha}$ ways to decorate the interior vertices. The decoration of the leaves can also be chosen in $\precsim e^{n^\alpha}$ ways. Therefore $\gamma_\omega(n) \precsim e^{n^\alpha}$. □

10.2.1. *Growth in the case of factorable sequences.* An upper bound on the degree of word growth in case of r-factorable sequences can thus be obtained from Theorem 10.10, since every r-factorable sequence is $(2r-1)$-homogeneous, but we can do slightly better if we combine Lemma 10.6 with the idea of portrait of an element. We omit the proof because of its similarity to the other proofs in this section.

THEOREM 10.13 (3/4-estimate). *If $\overline{\omega}$ is an r-factorable sequence, then the growth function of the group G_ω satisfies*

$$\gamma_\omega(n) \precsim \mathrm{e}^{n^\alpha}$$

where $\alpha = \log(M^r)/(\log(M^r) - \log(3/4)) = \log(M)/(\log(M) - \log(\sqrt[r]{3/4})) < 1$.

The 3/4-estimate was obtained only for the class of Grigorchuk p-groups defined by r-homogeneous (not r-factorable as above) sequences by R. Muchnik and I. Pak in [MP01] by different means. The same article contains some sharper considerations in case $p = 2$.

We can provide a small improvement in a special case that includes all Grigorchuk 2-groups. Namely, we are going to assume that $\overline{\omega}$ is an r-factorable sequence such that each factor contains three homomorphisms whose kernels cover B.

LEMMA 10.14 (2/3-shortening). *Let $\omega \in \widehat{\Omega}$ defines a group acting on the rooted binary tree. In addition, let the defining sequence $\overline{\omega}$ be such that there exist 3 terms ω_k, ω_ℓ and ω_s, $1 \leqslant k < \ell < m \leqslant r$, with the property that $K_k \cup K_\ell \cup K_s = B$. Then the following inequality holds for every reduced word F representing an element in $\mathsf{St}_\omega(\mathcal{L}_r)$:*

$$\left|L_r(F)\right| < \tfrac{2}{3}|F| + 3 \cdot M^r.$$

We note that the above shortening lemma cannot be improved, unless one starts paying attention to reductions beyond the simple ones. Indeed, in the first Grigorchuk group $\mathfrak{G}$ the word $F = (abadac)^{4k}$ has length $24k$, while $|L_3(F)| = 16k = \frac{2}{3}|F|$. On the other hand $F = (abadac)^{16} = 1$ in $\mathfrak{G}$, so by taking into account other relations the multiplicative constant of Lemma 10.14 could possibly be sharpened.

As a corollary to the shortening lemma above, we obtain:

THEOREM 10.15 (2/3-estimate). *If $\overline{\omega}$ is an r-factorable sequence such that each factor contains three letters whose kernels cover B, then the growth function of the group G_ω satisfies*

$$\gamma_\omega(n) \precsim \mathrm{e}^{n^\alpha},$$

where $\alpha = \log(M^r)/(\log(M^r) - \log(2/3)) = \log(M)/(\log(M) - \log(\sqrt[r]{2/3})) < 1$.

10.3. *Parabolic space and Schreier graphs*

We describe here the aspects of the parabolic subgroups (introduced in Section 7.3) related to growth. The G-space we study is defined as follows:

DEFINITION 10.16. Let G be a branch group, and let $P = \mathsf{St}_G(e)$ be a parabolic subgroup of G. The associated *parabolic space* is the G-set G/P.

We consider in this section only finitely-generated, contracting groups. We assume a branch group G, with fixed generating set S, has been chosen.

The proof of the following result appears in [BG00b].

PROPOSITION 10.17. *Let G, a group of automorphisms of the regular tree of branching index m, satisfy the conditions of Proposition* 7.6, *and suppose it is contracting* (*see Definition* 1.15). *Let P be a parabolic subgroup. Then G/P, as a G-set, is of polynomial growth of degree at most $\log_{1/\lambda}(m)$, where λ is the contracting constant. If moreover G is spherically transitive, then G/P's asymptotical growth is polynomial of degree $\log_{1/\lambda'}(m)$, with λ' the infimum of the λ as above.*

The growth of G/P is also connected to the growth of the associated Lie algebra (see Section 8). The following appears in [Bar00c]:

THEOREM 10.18. *Let G be a branch group, with parabolic space G/P. Then there exists a constant C such that*

$$\frac{C\,\mathsf{growth}(G/P)}{1-\hbar} \geqslant \frac{\mathsf{growth}\,\mathcal{L}(G)}{1-\hbar},$$

where $\mathsf{growth}(X)$ denotes the growth formal power series of X.

Anticipating, we note that in the above theorem equality holds for $\mathfrak{G}$, but does not hold for Γ nor $\overline{\overline{\Gamma}}$ (see Corollaries 8.8 and 8.19).

The most convenient way to describe the parabolic space G/P by giving it a graph structure:

DEFINITION 10.19. Let G be a group generated by a set S and H a subgroup of G. The *Schreier graph* $\mathcal{S}(G, H, S)$ of G/H is the directed graph on the vertex set G/H, with for every $s \in S$ and every $gH \in G/H$ an edge from gH to sgH. The base point of $\mathcal{S}(G, H, S)$ is the coset H.

Note that $\mathcal{S}(G, 1, S)$ is the Cayley graph of G relative to S. It may happen that $\mathcal{S}(G, P, S)$ have loops and multiple edges even if S is disjoint from H. Schreier graphs are $|S|$-regular graphs, and any degree-regular graph $\mathcal{G}$ containing a 1-factor (i.e. a regular subgraph of degree 1; there is always one if $\mathcal{G}$ has even degree) is a Schreier graph [Lub95a, Theorem 5.4].

For all $n \in \mathbb{N}$ consider the finite quotient $G_n = G/\operatorname{St}_G(n)$, acting on the n-th level of the tree. We first consider the finite graphs $\mathcal{G}_n = \mathcal{S}(G, P_n \operatorname{St}_G(n), S) = \mathcal{S}(G_n, P/(P \cap \operatorname{St}_G(n), S)$. Due to the fractal (or recursive) nature of branch groups, there are simple local rules producing $\mathcal{G}_{n+1}$ from $\mathcal{G}_n$, the limit of this process being the Schreier graph of G/P. Before stating a general result, we start by describing these rules for two of our examples: $\mathfrak{G}$ and Γ.

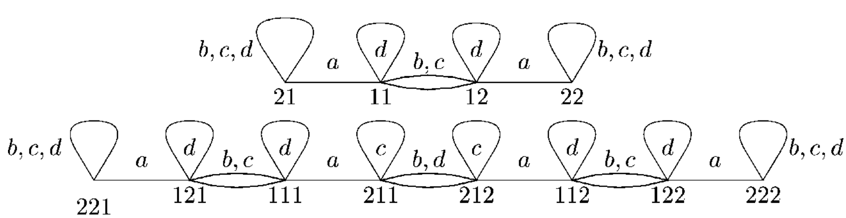
10.3.1. $\mathcal{S}(\mathfrak{G}, P, S)$. Assume the notation of Section 7.4. The graphs $\mathcal{G}_n = \mathcal{S}(\mathfrak{G}_n, P_n, S)$ will have edges labelled by $S = \{a, b, c, d\}$ (and not oriented, because all $s \in S$ are involutions) and vertices labelled by Y^n, where $Y = \{1, 2\}$.

First, it is clear that $\mathcal{G}_0$ is a graph on one vertex, labelled by the empty sequence $\emptyset$, and four loops at this vertex, labelled by a, b, c, d. Next, $\mathcal{G}_1$ has two vertices, labelled by 1 and 2; an edge labelled a between them; and three loops at 1 and 2 labelled by b, c, d.

Now given $\mathcal{G}_n$, for some $n \geqslant 1$, perform on it the following transformation: replace the edge-labels b by d, d by c, c by b; replace the vertex-labels σ by 2σ; and replace all edges labelled by a connecting σ and τ by: edges from 2σ to 1σ and from 2τ to 1τ, labelled a; two edges from 1σ to 1τ labelled b and c; and loops at 1σ and 1τ labelled d. We claim the resulting graph is $\mathcal{G}_{n+1}$.

To prove the claim, it suffices to check that the letters on the edge-labels act as described on the vertex-labels. If $b(\sigma) = \tau$, then $d(2\sigma) = 2\tau$, and similarly for c and d; this explains the cyclic permutation of the labels b, c, d. The other substitutions are verified similarly.

As an illustration, here are $\mathcal{G}_2$ and $\mathcal{G}_3$ for $\mathfrak{G}$. Note that the sequences in Y^* that appear correspond to "Gray enumeration", i.e. enumeration of integers in base 2 where only one bit is changed from a number to the next:



10.3.2. $\mathcal{S}(\Gamma, P, S)$. Assume the notation of Section 7.5. First, $\mathcal{G}_0$ has one vertex, labelled by the empty sequence $\emptyset$, and four loops, labelled a, a^{-1}, t, t^{-1}. Next, $\mathcal{G}_1$ has three vertices, labelled 1, 2, 3, cyclically connected by a triangle labelled a, a^{-1}, and with two loops at each vertex, labelled t, t^{-1}. In the pictures only geometrical edges, in pairs $\{a, a^{-1}\}$ and $\{t, t^{-1}\}$, are represented.

Now given $\mathcal{G}_n$, for some $n \geqslant 1$, perform on it the following transformation: replace the vertex-labels σ by 2σ; replace all triangles labelled by a, a^{-1} connecting ρ, σ, τ by: three triangles labelled by a, a^{-1} connecting respectively $1\rho, 2\rho, 3\rho$ and $1\sigma, 2\sigma, 3\sigma$ and $1\tau, 2\tau, 3\tau$; a triangle labelled by t, t^{-1} connecting $1\rho, 1\sigma, 1\tau$; and loops labelled by t, t^{-1} at $2\rho, 2\sigma, 2\tau$. We claim the resulting graph is $\mathcal{G}_{n+1}$.

As above, it suffices to check that the letters on the edge-labels act as described on the vertex-labels. If $a(\rho) = \sigma$ and $t(\rho) = \tau$, then $t(1\rho) = 1\sigma$, $t(2\rho) = 2\rho$ and $t(3\sigma) = 3\tau$. The verification for a edges is even simpler.

10.3.3. *Substitutional graphs.* The two Schreier graphs presented in the previous section are special cases of *substitutional graphs*, which we define below.

Substitutional graphs were introduced in the late 70's to describe growth of multicellular organisms. They bear a strong similarity to L-systems [RS80], as was noted by M. Gromov [Gro84]. Another notion of graph substitution has been studied by [Pre98], where he had the same convergence preoccupations as us.

Let us make a convention for this section: all graphs $\mathcal{G} = (V(\mathcal{G}), E(\mathcal{G}))$ shall be *labelled*, i.e. endowed with a map $E(\mathcal{G}) \to C$ for a fixed set C of colors, and *pointed*, i.e. shall have a distinguished vertex $* \in V(\mathcal{G})$. A *graph embedding* $\mathcal{G}' \hookrightarrow \mathcal{G}$ is just an injective map $E(\mathcal{G}') \to E(\mathcal{G})$ preserving the adjacency operations.

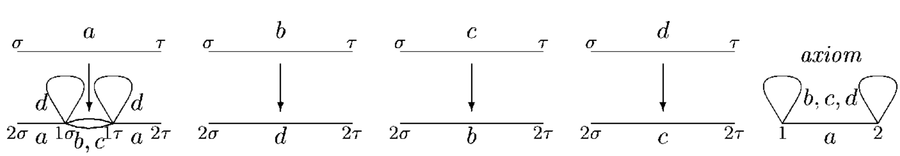
DEFINITION 10.20. A *substitutional rule* is a tuple $(U, R_1, \dots, R_n)$, where U is a finite m-regular edge-labelled graph, called the *axiom*, and each R_i, $i = 1, \dots, n$, is a rule of the form $X_i \to Y_i$, where X_i and Y_i are finite edge-labelled graphs. The graphs X_i are required to have no common edge. Furthermore, for each $i = 1, \dots, n$, there is an inclusion, written ι_i, of the vertices of X_i in the vertices of Y_i; the degree of $\iota_i(x)$ is the same as the degree of x for all $x \in V(X_i)$, and all vertices of Y_i not in the image of ι_i have degree m.

Given a substitutional rule, one sets $\mathcal{G}_0 = U$ and constructs iteratively $\mathcal{G}_{n+1}$ from $\mathcal{G}_n$ by listing all embeddings of all X_i in $\mathcal{G}_n$ (they are disjoint), and replacing them by the corresponding Y_i. If the base point $*$ of $\mathcal{G}_n$ is in a graph X_i, the base point of $\mathcal{G}_{n+1}$ will be $\iota_i(*)$.

Note that this expansion operation preserves the degree, so $\mathcal{G}_n$ is a m-regular finite graph for all n. We are interested in fixed points of this iterative process.

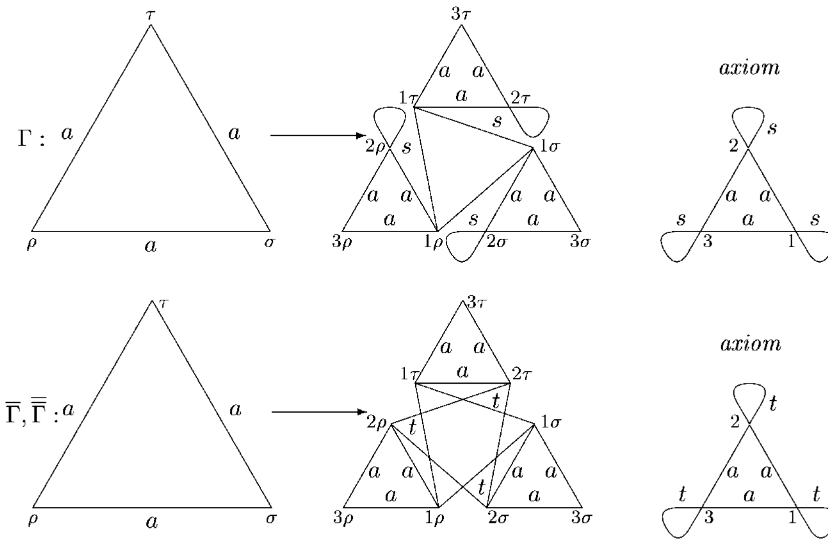
For any $R \in \mathbb{N}$, consider the balls $B_{*,n}(R)$ of radius R at the base point $*$ in $\mathcal{G}_n$. Since there is only a finite number of rules, there is an infinite sequence $n_0 < n_1 < \cdots$ such that the balls $B_{*,n_i}(R) \subseteq \mathcal{G}_{n_i}$ are all isomorphic. We consider $\mathcal{G}$, a limit graph in the sense of [GŻ97] (the limit exists), and call it a *substitutional graph*.

THEOREM 10.21 [BG00b]. *The following four substitutional rules describe the Schreier graph of* $\mathfrak{G}$:



where the vertex inclusions are given by $\sigma \mapsto 2\sigma$ *and* $\tau \mapsto 2\tau$. *The base point is the vertex* $222\dots$.

THEOREM 10.22 [BG00b]. *The substitutional rules producing the Schreier graphs of* Γ, $\overline{\Gamma}$ *and* $\overline{\overline{\Gamma}}$ *are given below. Note that the Schreier graphs of* $\overline{\Gamma}$ *and* $\overline{\overline{\Gamma}}$ *are isomorphic*:



where the vertex inclusions are given by $\rho \mapsto 3\rho$, $\sigma \mapsto 3\sigma$ *and* $\tau \mapsto 3\tau$. *The base point is the vertex* $333\ldots$.

By Proposition 10.17, these two limit graphs have asymptotically polynomial growth of degree no higher than $\log_2(3)$.

Note that there are maps $\pi_n : V(\mathcal{G}_{n+1}) \to V(\mathcal{G}_n)$ that locally (i.e. in each copy of some right-hand rule Y_i) are the inverse of the embedding ι_i. In case these π_n are graph morphisms one can consider the projective system $\{\mathcal{G}_n, \pi_n\}$ and its inverse limit $\widehat{\mathcal{G}} = \varprojlim \mathcal{G}_n$, which is a profinite graph [RZ00]. We devote our attention to the discrete graph $\mathcal{G} = \varinjlim \mathcal{G}_n$.

The growth series of $\mathcal{G}$ can often be described as an infinite product. We give such an expression for the graph in Figure 30, making use of the fact that $\mathcal{G}$ "looks like a tree" (even though it is amenable).

Consider the finite graphs $\mathcal{G}_n$; recall that $\mathcal{G}_n$ has 3^n vertices. Let D_n be the diameter of $\mathcal{G}_n$ (maximal distance between two vertices), and let $\gamma_n = \sum_{i \in \mathbb{N}} \gamma_n(i) X^i$ be the growth series of $\mathcal{G}_n$ (here $\gamma_n(i)$ denotes the number of vertices in $\mathcal{G}_n$ at distance i from the base point $*$).

The construction rule for $\mathcal{G}$ implies that $\mathcal{G}_{n+1}$ can be constructed as follows: take three copies of $\mathcal{G}_n$, and in each of them mark a vertex V at distance D_n from $*$. At each V delete the loop labelled s, and connect the three copies by a triangle labelled s at the three V's. It then follows that $D_{n+1} = 2D_n + 1$, and $\gamma_{n+1} = (1 + 2X^{D_n+1})\gamma_n$. Using the initial values $\gamma_0 = 1$ and $D_0 = 0$, we obtain by induction

$$D_n = 2^n - 1, \qquad \gamma_n = \prod_{i=0}^{n-1} \bigl(1 + 2X^{2^i}\bigr).$$

We have also shown that the ball of radius 2^n around $*$ contains 3^n points, so the growth of $\mathcal{G}$ is at least $n^{\log_2(3)}$. But Proposition 10.17 shows that it is also an upper bound, and we conclude:

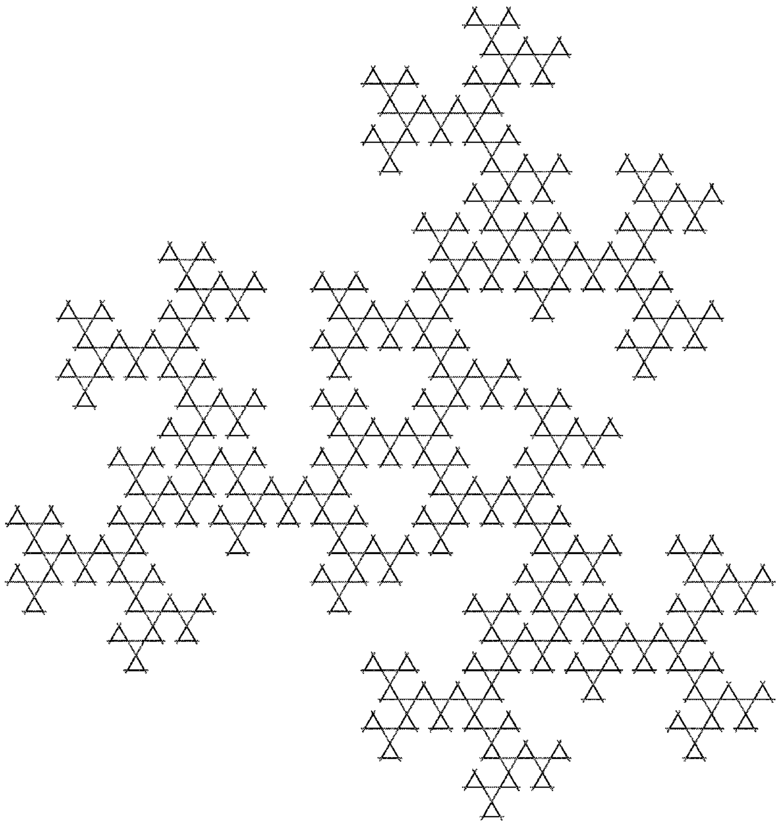


Table 30. The Schreier graph of Γ_6. The edges represent the generators s and a.

PROPOSITION 10.23. *Γ is an amenable 4-regular graph whose growth function is transcendental, and admits the product decomposition*

$$\gamma(X) = \prod_{i \in \mathbb{N}} \bigl(1 + 2X^{2^i}\bigr).$$

It is planar, and has polynomial growth of degree $\log_2(3)$.

Any graph is a metric space when one identifies each edge with a disjoint copy of an interval $[0, L]$ for some $L > 0$. We turn $\mathcal{G}_n$ in a diameter-1 metric space by giving to each edge in $\mathcal{G}_n$ the length $L = \operatorname{diam}(\mathcal{G}_n)^{-1}$. The family $\{\mathcal{G}_n\}$ then converges, in the following sense:

Let A, B be closed subsets of the metric space (X, d). For any ε, let $A_\varepsilon = \{x \in X \mid d(x, A) \leqslant \varepsilon\}$, and define the *Hausdorff distance*

$$d_X(A, B) = \inf\{\varepsilon \mid A \subseteq B_\varepsilon, B \subseteq A_\varepsilon\}.$$

This defines a metric on closed subsets of X. For general metric spaces (A, d) and (B, d), define their *Gromov–Hausdorff distance*

$$d^{GH}(A, B) = \inf_{X,i,j} d_X\big(i(A), j(B)\big),$$

where i and j are isometric embeddings of A and B in a metric space X.

We may now rephrase the considerations above as follows: the sequence $\{\mathcal{G}_n\}$ is convergent in the Gromov–Hausdorff metric. The limit set $\mathcal{G}_\infty$ is a compact metric space.

The limit spaces are then: for $\mathfrak{G}$ and $\widetilde{\mathfrak{G}}$, the limit $\mathcal{G}_\infty$ is the interval [0, 1] (in accordance with its linear growth, see Proposition 10.17). The limit spaces for Γ, $\overline{\Gamma}$ and $\overline{\overline{\Gamma}}$ are fractal sets of dimension $\log_2(3)$.

11. Spectral properties of unitary representations

We describe here some explicit computations of the spectrum of the Schreier graphs defined in Section 10.3. The natural viewpoint is that of spectra of representations, which we define now:

DEFINITION 11.1. Let G be a group generated by a finite symmetric set S. The *spectrum* $\operatorname{spec}(\tau)$ of a representation $\tau : G \to \mathcal{U}(\mathcal{H})$ with respect to the given set of generators is the spectrum of $\Delta_\tau = \sum_{s\in S} \tau(s)$ seen as an bounded operator on $\mathcal{H}$.

(The condition that S be symmetric ensures that $\operatorname{spec}(\tau)$ is a subset of $\mathbb{R}$.)

Let H be a subgroup of G. Then the spectrum of the (ℓ^2-adjacency operator of the) Schreier graph $\mathcal{S}(G, H, S)$ is the spectrum of the quasi-regular representation $\rho_{G/H}$ of G in $\ell^2(G/H)$. This establishes the connection with the previous section.

11.1. *Unitary representations*

Let G act on the rooted tree τ. Then G also acts on the boundary $\partial\mathcal{T}$ of the tree. Since G preserves the uniform measure on this boundary, we have a unitary representation π of G in $L^2(\partial\mathcal{T}, \nu)$, where ν is the Bernoulli measure; equivalently, we have a representation in $L^2([0, 1], \text{Lebesgue})$. Let $\mathcal{H}_n$ be the subspace of $L^2(\partial\mathcal{T}, \nu)$ spanned by the characteristic functions χ_σ of the rays e starting by σ, for all $\sigma \in Y^n$. It is of dimension k_n, and can equivalently be seen as spanned by the characteristic functions in $L^2([0, 1], \text{Lebesgue})$ of intervals of the form $[(i-1)/k_n, i/k_n]$, $1 \leqslant i \leqslant k_n$. These $\mathcal{H}_n$ are invariant subspaces, and afford representations $\pi_n = \pi|_{\mathcal{H}_n}$. As clearly π_{n-1} is a subrepresentation of π_n, we set $\pi_n^\perp = \pi_n \ominus \pi_{n-1}$, so that $\pi = \bigoplus_{n=0}^\infty \pi_n^\perp$.

Let P be a parabolic subgroup (see Section 7.3), and set $P_n = P\,\mathsf{St}_G(n)$. Denote by $\rho_{G/P}$ the quasi-regular representation of G in $\ell^2(G/P)$ and by ρ_{G/P_n} the finite-dimensional representations of G in $\ell^2(G/P_n)$, of degree k_n. Since G is level-transitive, the representations π_n and ρ_{G/P_n} are unitary equivalent.

The G-spaces G/P are of polynomial growth when the conditions of Proposition 10.17 are fulfilled, and therefore, according to the criterion of Følner given in Theorem 10.3, they are amenable.

The following result belongs to the common lore:

PROPOSITION 11.2. *Let H be a subgroup of G. Then the quasi-regular representation $\rho_{G/H}$ is weakly contained in ρ_G if and only if H is amenable.*

("Weakly contained" is a topological extension of "contained"; it implies for instance inclusion of spectra.)

THEOREM 11.3. *Let G be a group acting on a regular rooted tree, and let π, π_n and $\pi_n^\perp$ be as above.*

(1) *If G is weakly branch, then $\rho_{G/P}$ is an irreducible representation of infinite dimension.*
(2) *π is a reducible representation of infinite dimension whose irreducible components are precisely those of the $\pi_n^\perp$ (and thus are all finite-dimensional). Moreover*

$$\operatorname{spec}(\pi) = \overline{\bigcup_{n\geqslant 0} \operatorname{spec}(\pi_n)} = \overline{\bigcup_{n\geqslant 0} \operatorname{spec}(\pi_n^\perp)}.$$

(3) *The spectrum of $\rho_{G/P}$ is contained in $\overline{\bigcup_{n\geqslant 0} \operatorname{spec}(\rho_{G/P_n})} = \overline{\bigcup_{n\geqslant 0} \operatorname{spec}(\pi_n)}$, and thus is contained in the spectrum of π. If moreover either P or G/P are amenable, these spectra coincide, and if P is amenable, they are contained in the spectrum of ρ_G:*

$$\operatorname{spec}(\rho_{G/P}) = \operatorname{spec}(\pi) = \overline{\bigcup_{n\geqslant 0} \operatorname{spec}(\pi_n)} \subseteq \operatorname{spec}(\rho_G).$$

(4) *Δ_π has a pure-point spectrum, and its spectral radius $r(\Delta_\pi) = s \in \mathbb{R}$ is an eigenvalue, while the spectral radius $r(\Delta_{\rho_{G/P}})$ is not an eigenvalue of $\Delta_{\rho_{G/P}}$. Thus $\Delta_{\rho_{G/P}}$ and Δ_π are different operators having the same spectrum.*

11.1.1. *Can one hear a representation?* We end by turning to a question of M. Kać [Kac66]: "Can one hear the shape of a drum?". This question was answered in the negative in [GWW92], and we here give an answer in the negative to a related question: "Can one hear a representation?". Indeed $\rho_{G/P}$ and π have same spectrum (i.e. cannot be distinguished by hearing), but are not equivalent. Furthermore, if G is a branch group, there are uncountably many nonequivalent representations within $\{\rho_{G/\operatorname{St}_G(e)} \mid e \in \partial\mathcal{T}\}$, as is shown in [BG02].

The same question may be asked for graphs: "are there two non-isomorphic graphs with same spectrum?". There are finite examples, obtained through the notion of *Sunada pair* [Lub95a]. C. Béguin, A. Valette and A. Żuk produced the following example in [BVZ97]: let Γ be the integer Heisenberg group (free 2-step nilpotent on 2 generators x, y). Then $\Delta = x + x^{-1} + y + y^{-1}$ has spectrum $[-2, 2]$, which is also the spectrum of $\mathbb{Z}^2$ for an independent generating set. As a consequence, their Cayley graphs have same spectrum, but are not quasi-isometric (they do not have the same growth).

Using a result of N. Higson and G. Kasparov [HK97] (giving a partial positive answer to the Baum–Connes conjecture), we may infer the following

PROPOSITION 11.4. *Let Γ be a torsion-free amenable group with finite generating set $S = S^{-1}$ such that there is a map $\phi : \Gamma \to \mathbb{Z}/2\mathbb{Z}$ with $\phi(S) = \{1\}$. Then*

$$\operatorname{spec}\left(\sum_{s\in S} \rho(s)\right) = \big[-|S|, |S|\big].$$

In particular, there are countably many non-quasi-isometric graphs with the same spectrum, including the graphs of $\mathbb{Z}^d$, of free nilpotent groups and of suitable torsion-free groups of intermediate growth (for the first examples, see [Gri85a]).

In contrast to Proposition 11.4 stating that the spectrum of the regular representation is an interval, the spectra of the representation π may be totally disconnected. The first two authors prove in [BG00b] the following

THEOREM 11.5. *For $\lambda \in \mathbb{R}$, define*

$$J(\lambda) = \overline{\overline{\pm\sqrt{\lambda \pm \sqrt{\lambda \pm \sqrt{\lambda \pm \sqrt{\cdots}}}}}}$$

(*note the closure operator written in bar notation on the top*). *Then we have the following spectra*:

$\mathbf{G}$	$\operatorname{spec}(\pi)$
$\mathfrak{G}$	$[-2, 0] \cup [2, 4]$
$\widetilde{\mathfrak{G}}$	$[0, 4]$
Γ	$\{4, 1\} \cup 1 + J(6)$
$\overline{\Gamma}, \overline{\overline{\Gamma}}$	$\{4, -2, 1\} \cup 1 \pm \sqrt{\frac{9}{2} \pm 2J(\frac{45}{16})}$

In particular, the spectrum of the graph in Figure 30 is the totally disconnected set $\{4, 1\} \cup 1 + J(6)$, which is the union of a Cantor set of null Lebesgue measure and a countable collection of isolated points.

11.2. *Operator recursions*

We now describe the computations of the spectra for the examples in Theorem 11.5, which share the property of acting on an m-regular tree $\mathcal{T}$. Let $\mathcal{H}$ be an infinite-dimensional Hilbert space, and suppose $\Phi : \mathcal{H} \to \mathcal{H} \oplus \cdots \oplus \mathcal{H}$ is an isomorphism, where the domain of Φ is a sum of $m \geqslant 2$ copies of $\mathcal{H}$. Let S be a finite subset of $\mathcal{U}(\mathcal{H})$, and suppose that for all $s \in S$, if we write $\Phi^{-1} s \Phi$ as an operator matrix $(s_{i,j})_{i,j\in\{1,\ldots,d\}}$ where the $s_{i,j}$ are operators in $\mathcal{H}$, then $s_{i,j} \in S \cup \{0, 1\}$.

This is precisely the setting in which we will compute the spectra of our five example groups: for $\mathfrak{G}$, we have $m = 2$ and $S = \{a, b, c, d\}$ with

$$a = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix},$$

$$c = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}, \qquad d = \begin{pmatrix} 1 & 0 \\ 0 & b \end{pmatrix}.$$

For $\widetilde{\mathfrak{G}}$, we also have $m = 2$, and $S = \{a, \tilde{b}, \tilde{c}, \tilde{d}\}$ given by

$$b = \begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix}, \quad c = \begin{pmatrix} 1 & 0 \\ 0 & d \end{pmatrix}, \qquad d = \begin{pmatrix} 1 & 0 \\ 0 & b \end{pmatrix}.$$

For $\Gamma = \langle a, s\rangle$, $\overline{\Gamma} = \langle a, t\rangle$ and $\overline{\overline{\Gamma}} = \langle a, r\rangle$, we have $m = 3$ and

$$a = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \qquad s = \begin{pmatrix} a & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & s \end{pmatrix},$$

$$t = \begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & t \end{pmatrix}, \qquad r = \begin{pmatrix} a & 0 & 0 \\ 0 & a^2 & 0 \\ 0 & 0 & r \end{pmatrix}.$$

Each of these operators is unitary. The families $S = \{a, b, c, d\}, \ldots$ generate subgroups $G(S)$ of $\mathcal{U}(\mathcal{H})$. The choice of the isomorphism Φ defines a unitary representation of $\langle S\rangle$.

We note, however, that the expression of each operator as a matrix of operators does not uniquely determine the operator, in the sense that different isomorphisms Φ can yield non-isomorphic operators satisfying the same recursions. Even if Φ is fixed, it may happen that different operators satisfy the same recursions. We considered two types of isomorphisms: $\mathcal{H} = \ell^2(G/P)$, where Φ is derived from the decomposition map ψ; and $\mathcal{H} = L^2(\partial\mathcal{T})$, where $\Phi : \mathcal{H} \to \mathcal{H}^Y$ is defined by $\Phi(f)(\sigma) = (f(0\sigma), \ldots, f((m-1)\sigma))$, for $f \in L^2(\partial\mathcal{T})$ and $\sigma \in \partial\mathcal{T}$. There are actually uncountably many non-equivalent isomorphisms, giving uncountably many non-equivalent representations of the same group, as indicated in Section 11.1.1.

11.2.1. *The spectrum of* $\mathfrak{G}$. For brevity we shall only describe the spectrum of π for the first Grigorchuk group. Details and computations for the other examples appear in [BG00b].

Denote by a_n, b_n, c_n, d_n the permutation matrices of the representation $\pi_n = \rho_{G/P_n}$. We have

$$a_0 = b_0 = c_0 = d_0 = (1),$$

$$a_n = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad b_n = \begin{pmatrix} a_{n-1} & 0 \\ 0 & c_{n-1} \end{pmatrix},$$

$$c_n = \begin{pmatrix} a_{n-1} & 0 \\ 0 & d_{n-1} \end{pmatrix}, \qquad d_n = \begin{pmatrix} 1 & 0 \\ 0 & b_{n-1} \end{pmatrix}.$$

The Hecke–Laplace operator of π_n is

$$\Delta_n = a_n + b_n + c_n + d_n = \begin{pmatrix} 2a_{n-1}+1 & 1 \\ 1 & \Delta_{n-1} - a_{n-1} \end{pmatrix},$$

and we wish to compute its spectrum. We start by proving a slightly stronger result: define

$$Q_n(\lambda, \mu) = \Delta_n - (\lambda+1)a_n - (\mu+1)$$

and

$$\begin{aligned} \Phi_0 &= 2 - \mu - \lambda, \\ \Phi_1 &= 2 - \mu + \lambda, \\ \Phi_2 &= \mu^2 - 4 - \lambda^2, \\ \Phi_n &= \Phi_{n-1}^2 - 2(2\lambda)^{2^{n-2}} \quad (n \geqslant 3). \end{aligned}$$

Then the following steps compute the spectrum of π_n: first, for $n \geqslant 2$, we have

$$\left|Q_n(\lambda, \mu)\right| = (4-\mu^2)^{2^{n-2}} \left| Q_{n-1}\left(\frac{2\lambda^2}{4-\mu^2}, \mu + \frac{\mu\lambda^2}{4-\mu^2}\right)\right| \quad (n \geqslant 2).$$

Therefore, for all n we have

$$|Q_n| = \Phi_0 \Phi_1 \cdots \Phi_n.$$

Then, for all n we have

$$\begin{aligned} &\big\{(\lambda, \mu)\colon\ Q_n(\lambda,\mu) \text{ noninvertible}\big\} \\ &\quad = \big\{(\lambda,\mu)\colon\ \Phi_0(\lambda,\mu) = 0\big\} \cup \big\{(\lambda,\mu)\colon\ \Phi_1(\lambda,\mu)=0\big\} \\ &\qquad \cup \left\{(\lambda,\mu)\colon\ 4-\mu^2+\lambda^2+4\lambda\cos\left(\frac{2\pi j}{2^n}\right) = 0 \text{ for some } j = 1, \ldots, 2^{n-1}-1\right\}. \end{aligned}$$

In the (λ, μ) system, the spectrum of Q_n is thus a collection of 2 lines and $2^{n-1}-1$ hyperbolæ.

The spectrum of Δ_n is precisely the set of θ such that $|Q_n(-1, \theta-1)| = 0$. From the computations above we obtain

PROPOSITION 11.6.

$$\operatorname{spec}(\Delta_n) = \big\{1 \pm \sqrt{5 - 4\cos\phi}\colon\ \phi \in 2\pi\mathbb{Z}/2^n\big\} \setminus \{0, -2\}.$$

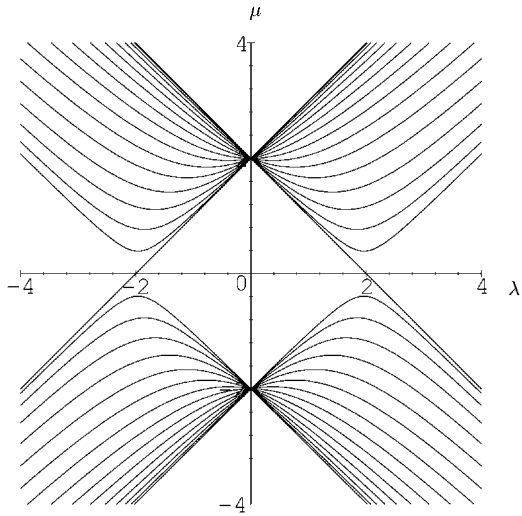


Table 31. The spectrum of $Q_n(\lambda, \mu)$ for $\mathfrak{G}$.

Therefore the spectrum of π, for the group $\mathfrak{G}$, is

$$\mathrm{spec}(\Delta) = [-2, 0] \cup [2, 4].$$

The first eigenvalues of Δ_n are 4; 2; $1 \pm \sqrt{5}$; $1 \pm \sqrt{5 \pm 2\sqrt{2}}$; $1 \pm \sqrt{5 \pm 2\sqrt{2 \pm \sqrt{2}}}$; etc.

The numbers of the form $\pm\sqrt{\lambda \pm \sqrt{\lambda \pm \sqrt{\cdots}}}$ appear as preimages of the quadratic map $z^2 - \lambda$, and after closure produce a Julia set for this map (see [Bar88]). In the given example this Julia set is just an interval. For the groups Γ, $\overline{\Gamma}$, $\overline{\overline{\Gamma}}$, however, the spectra are simple transformations of Julia sets which are totally disconnected, as similar computations show.

12. Open problems

This section collects questions for which no answer is yet known. Here we use the geometric definition of branch group given in Definition 1.13, except in Question 2.

The theory of branch group is a recent development and questions arise and die almost every day, so there are no longstanding problems in the area and there is no easy way to say which questions are difficult and which are not. The list below just serves as a list of problems that one should naturally ask at this given moment. We kindly invite the interested readers to solve as many of the problems below as they can.

ADDED IN THE PROOF. The authors had a chance to proofread the article 18 months after the initial submission. Some of the proposed problems have been solved in the meantime and we include appropriate comments to that effect.

QUESTION 1. Is there a finitely generated fractal regular branch group G, branched over K, acting on the binary tree, such that the index of the geometric embedding of $K \times K$ into K is two?

QUESTION 2. Every branch group from Definition 1.1 acts canonically on the tree determined by its branch structure as a group of tree automorphisms. Does the center have a finite index in the kernel?

QUESTION 3. Does every finitely generated branch p-group, where p is a prime, satisfy the congruence subgroup property?

QUESTION 4. Is every finitely generated branch group isomorphic to a spinal group?

QUESTION 5. Is the conjugacy problem solvable in all branch groups with solvable word problem?

QUESTION 6. When do the defining triples ω and ω' define non-isomorphic examples of branch groups in [Gri84,Gri85a] and more generally in G and GGS groups?

QUESTION 7. What branch groups have finite L-presentations? Finite ascending L-presentations? In particular, what is the status of the Gupta–Sidki 3-group?

All that is known at present is that the Gupta–Sidki 3-group has a finite endomorphic presentation.

QUESTION 8. Do there exist finitely presented branch groups?

QUESTION 9. Is it correct that there are no finitely generated hereditary just-infinite torsion groups?

QUESTION 10. Is every finitely generated just-infinite group of intermediate growth necessarily a branch group?

QUESTION 11. Is every finitely generated hereditarily just-infinite group necessarily linear?

QUESTION 12. Recall that G has bounded generation if there exist elements $g_1, \ldots, g_k$ in G such that every element in G can be written as $g_1^{n_1} \ldots g_k^{n_k}$ for some $n_1, \ldots, n_k \in \mathbb{Z}$. Can a just-infinite branch group have bounded generation? Can infinite simple group have bounded generation?

QUESTION 13. What is the height (in the sense of [Pri80], see Section 5) of a Grigorchuk 2-group G_ω when the defining sequence $\overline{\omega}$ is not periodic? Same question for arbitrary G groups. In particular, can the height be infinite?

QUESTION 14. Is every maximal subgroup in a finitely generated branch group necessarily of finite index?

QUESTION 15. Is there is a finitely generated branch group containing the free group F_2 on two generators?

Positive answer is provided by S. Sidki and J. Wilson in [SW02].

QUESTION 16. Are there finitely generated branch groups with exponential growth that do not contain the free group F_2?

QUESTION 17. Is there a finitely generated branch group whose degree of growth is $e^{\sqrt{n}}$? Is there such a group in the whole class of finitely generated groups?

QUESTION 18. What is the exact degree of growth of any of the basic examples of regular branch groups (for example $\mathfrak{G}$, Γ, $\overline{\Gamma}$, ...)?

QUESTION 19. What is the growth of the Brunner–Sidki–Vieira group (see [BSV99] and Section 4.3)?

It is known that the Brunner–Sidki–Vieira group does not contain any non-Abelian free groups, but it is not known whether it contains a non-Abelian free monoid. Note that this group is not a branch group, but it is a weakly branch group.

QUESTION 20. Are there finitely generated non-amenable branch groups not containing the free group F_2?

QUESTION 21. Is it correct that in each finitely generated fractal branch group G every finitely generated subgroup is either finite or Pride equivalent with G?

A stronger property holds for $\mathfrak{G}$, namely, J. Wilson and the second author have proved in [GW01] that every finitely generated subgroup of $\mathfrak{G}$ is either finite or commensurable with $\mathfrak{G}$. C. Röver has announced that the answer is also positive for the Gupta–Sidki group $\overline{\overline{\Gamma}}$. However, the answer is negative in general.

QUESTION 22. Do there exist branch groups with the property (T)?

References

[Ady79] S.I. Adyan, *The Burnside Problem and Identities in Groups*, Springer, Berlin (1979). (Translated from Russian by J. Lennox and J. Wiegold.)

[AL91] S.I. Adyan and I.G. Lysënok, *Groups, all of whose proper subgroups are finite cyclic*, Izv. Akad. Nauk SSSR Ser. Mat. **55** (5) (1991), 933–990.

[Ale72] S.V. Alešin, *Finite automata and the Burnside problem for periodic groups*, Mat. Zametki **11** (1972), 319–328.

[Bar] L. Bartholdi, *L-presentations and branch groups*, to appear in J. Algebra.

[Bar88] M. Barnsley, *Fractals Everywhere*, Academic Press Inc., Boston, MA (1988).

[Bar98] L. Bartholdi, *The growth of Grigorchuk's torsion group*, Internat. Math. Res. Notices **20** (1998), 1049–1054.

[Bar00a] L. Bartholdi, *A class of groups acting on rooted trees*, Unpublished (2000).

[Bar00b] L. Bartholdi, *Croissance des groupes agissant sur des arbres*, Ph.D. thesis, Université de Genève (2000).

[Bar00c] L. Bartholdi, *Lie algebras and growth in branch groups*, Preprint (2000).

[Bar01] L. Bartholdi, *Lower bounds on the growth of a group acting on the binary rooted tree* (2001).

[Bas72] H. Bass, *The degree of polynomial growth of finitely generated nilpotent groups*, Proc. London Math. Soc. (3) **25** (1972), 603–614.

[Bau93] G. Baumslag, *Topics in Combinatorial Group Theory*, Lectures in Mathematics, ETH, Zürich/Birkhäuser, Basel (1993).

[BCK88] M.A. Brodie, R.F. Chamberlain and L.-C. Kappe, *Finite coverings by normal subgroups*, Proc. Amer. Math. Soc. **104** (3) (1988), 669–674.

[BG00a] L. Bartholdi and R.I. Grigorchuk, *Lie methods in growth of groups and groups of finite width*, Computational and Geometric Aspects of Modern Algebra, M. Atkinson et al., eds, London Math. Soc. Lecture Note Ser., Vol. 275, Cambridge Univ. Press, Cambridge (2000), 1–27.

[BG00b] L. Bartholdi and R.I. Grigorchuk, *On the spectrum of Hecke type operators related to some fractal groups*, Trudy Mat. Inst. Steklov. **231** (2000), 5–45, math.GR/9910102.

[BG02] L. Bartholdi and R.I. Grigorchuk, *On parabolic subgroups and Hecke algebras of some fractal groups*, Serdica Math. J. **28** (1) (2002), 47–90.

[BH97] M. Burger and P. de la Harpe, *Constructing irreducible representations of discrete groups*, Proc. Indian Acad. Sci. Math. Sci. **107** (3) (1997), 223–235.

[Bla72] N. Blackburn, *Some homology groups of wreath products*, Illinois J. Math. **16** (1972), 116–129.

[Bro87] K.S. Brown, *Finiteness properties of groups*, Proceedings of the Northwestern Conference on Cohomology of Groups (Evanston, IL, 1985), J. Pure Appl. Algebra **44** (1987), 45–75.

[Bro94] K.S. Brown, *Cohomology of Groups*, Springer, New York (1994). (Corrected reprint of the 1982 original.)

[BS97] Y. Barnea and A. Shalev, *Hausdorff dimension, pro-p groups, and Kac–Moody algebras*, Trans. Amer. Math. Soc. **349** (12) (1997), 5073–5091.

[BS98] A.M. Brunner and S.N. Sidki, *The generation of* $\mathrm{gl}(n,\mathbb{Z})$ *by finite state automata*, Internat. J. Algebra Comput. **8** (1) (1998), 127–139.

[BŠ01] L. Bartholdi and Z. Šunić, *On the word and period growth of some groups of tree automorphisms*, Comm. Algebra (2001).

[BSV99] A.M. Brunner, S.N. Sidki and A.C. Vieira, *A just nonsolvable torsion-free group defined on the binary tree*, J. Algebra **211** (1) (1999), 99–114.

[BVZ97] C. Béguin, A. Valette and A. Zuk, *On the spectrum of a random walk on the discrete Heisenberg group and the norm of Harper's operator*, J. Geom. Phys. **21** (4) (1997), 337–356.

[CFP96] J.W. Cannon, W.J. Floyd and W.R. Parry, *Introductory notes on Richard Thompson's groups*, Enseign. Math. (2) **42** (2–4) (1996), 215–256.

[CGH99] T.G. Ceccherini-Silberstein, R.I. Grigorchuk and P. de la Harpe, *Amenability and paradoxical decompositions for pseudogroups and discrete metric spaces*, Trudy Mat. Inst. Steklov. **224** (1999), 68–111. (Dedicated to Academician Lev Semenovich Pontryagin on the occasion of his 90th birthday (in Russian).)

[Cho80] C. Chou, *Elementary amenable groups*, Illinois J. Math. **24** (3) (1980), 396–407.

[CK83] D. Carter and G. Keller, *Bounded elementary generation of* $\mathsf{SL}_n(\mathcal{O})$, Amer. J. Math. **105** (3) (1983), 673–687.

[CM77] D.J. Collins and C.F. Miller, III, *The conjugacy problem and subgroups of finite index*, Proc. London Math. Soc. (3) **34** (3) (1977), 535–556.

[CMS01] T.G. Ceccherini-Silberstein, A. Machì and F. Scarabotti, *The Grigorchuk group of intermediate growth*, Rend. Circ. Mat. Palermo (2) **50** (1) (2001), 67–102.

[CST01] T.G. Ceccherini-Silberstein, F. Scarabotti and F. Tolli, *The top of the lattice of normal subgroups of the Grigorchuk group*, J. Algebra **246** (1) (2001), 292–310.

[Day57] M.M. Day, *Amenable semigroups*, Illinois J. Math. **1** (1957), 509–544.

[Efr53] V.A. Efremovich, *The proximity geometry of Riemannian manifolds*, Uspekhi Mat. Nauk **8** (1953), 189.

[EP84] M. Edjvet and S.J. Pride, *The concept of "largeness" in group theory. II*, Groups – Korea, 1983 (Kyoungju, 1983), Springer, Berlin (1984), 29–54.

[Far87] D.R. Farkas, *Semisimple representations and affine rings*, Proc. Amer. Math. Soc. **101** (2) (1987), 237–238.

[FG85] J. Fabrykowski and N.D. Gupta, *On groups with sub-exponential growth functions*, J. Indian Math. Soc. (N.S.) **49** (3–4) (1985), 249–256.

[FG91] J. Fabrykowski and N.D. Gupta, *On groups with sub-exponential growth functions. II*, J. Indian Math. Soc. (N.S.) **56** (1–4) (1991), 217–228.

[FGMS01] B.L. Fine, A.M. Gaglione, A.G. Myasnikov and D. Spellman, *Discriminating groups* (2001).

[FT95] M.H. Freedman and P. Teichner, 4-*manifold topology. I. Subexponential groups*, Invent. Math. **122** (3) (1995), 509–529.

[Føl57] E. Følner, *Note on groups with and without full Banach mean value*, Math. Scand. **5** (1957), 5–11.

[GH90] É. Ghys and P. de la Harpe, *Sur les groupes hyperboliques d'après Mikhael Gromov*, Progress in Mathematics, Vol. 83, Birkhäuser Boston Inc., Boston, MA (1990). (Papers from the Swiss Seminar on Hyperbolic Groups held in Bern, 1988.)

[GH97] R.I. Grigorchuk and P. de la Harpe, *On problems related to growth, entropy, and spectrum in group theory*, J. Dynam. Control Systems **3** (1) (1997), 51–89.

[GHZ00] R.I. Grigorchuk, W.N. Herfort and P.A. Zaleskiĭ, *The profinite completion of certain torsion p-groups*, Algebra (Moscow, 1998), de Gruyter, Berlin (2000), 113–123.

[GNS00] R.I. Grigorchuk, V.V. Nekrashevich and V.I. Sushchanskiĭ, *Automata, dynamical systems, and groups*, Trudy Mat. Inst. Steklov. **231** (4) (2000), 134–214.

[Gol64] E.S. Golod, *On nil-algebras and finitely approximable p-groups*, Izv. Akad. Nauk SSSR Ser. Mat. **28** (1964), 273–276.

[Gri80] R.I. Grigorchuk, *On Burnside's problem on periodic groups*, Funktsional. Anal. i Prilozhen. **14** (1) (1980), 53–54; English transl.: Functional Anal. Appl. **14** (1980), 41–43.

[Gri83] R.I. Grigorchuk, *On the Milnor problem of group growth*, Dokl. Akad. Nauk SSSR **271** (1) (1983), 30–33.

[Gri84] R.I. Grigorchuk, *Degrees of growth of finitely generated groups and the theory of invariant means*, Izv. Akad. Nauk SSSR Ser. Mat. **48** (5) (1984), 939–985; English transl.: Math. USSR-Izv. **25** (2) (1985), 259–300.

[Gri85a] R.I. Grigorchuk, *Degrees of growth of p-groups and torsion-free groups*, Mat. Sb. (N.S.) **126 (168)** (2) (1985), 194–214, 286.

[Gri85b] R.I. Grigorchuk, *A relationship between algorithmic problems and entropy characteristics of groups*, Dokl. Akad. Nauk SSSR **284** (1) (1985), 24–29.

[Gri89] R.I. Grigorchuk, *On the Hilbert–Poincaré series of graded algebras that are associated with groups*, Mat. Sb. **180** (2) (1989), 207–225, 304; English transl.: Math. USSR-Sb. **66** (1) (1990), 211–229.

[Gri95] R.I. Grigorchuk, *Some results on bounded cohomology*, Combinatorial and Geometric Group Theory (Edinburgh, 1993), Cambridge Univ. Press, Cambridge (1995), 111–163.

[Gri98] R.I. Grigorchuk, *An example of a finitely presented amenable group that does not belong to the class EG*, Mat. Sb. **189** (1) (1998), 79–100.

[Gri99] R.I. Grigorchuk, *On the system of defining relations and the Schur multiplier of periodic groups generated by finite automata*, Groups St. Andrews 1997 in Bath, I, N. Ruskuc, C.M. Campbell, E.F. Robertson and G.C. Smith, eds, Cambridge Univ. Press, Cambridge (1999), 290–317.

[Gri00] R.I. Grigorchuk, *Just infinite branch groups*, New Horizons in pro-p Groups, M.P.F. du Sautoy, D. Segal and A. Shalev, eds, Birkhäuser Boston, Boston, MA (2000), 121–179.

[Gri01] R.I. Grigorchuk, *On branch groups*, Mat. Zametki **67** (6) (2001), 852–858.

[Gro81] M. Gromov, *Groups of polynomial growth and expanding maps*, Inst. Hautes Études Sci. Publ. Math. (53) (1981), 53–73.

[Gro84] M. Gromov, *Infinite groups as geometric objects*, Proceedings of the International Congress of Mathematicians (Warsaw, 1983), Vols. 1, 2, PWN (1984), 385–392.

[GS83a] N.D. Gupta and S.N. Sidki, *On the Burnside problem for periodic groups*, Math. Z. **182** (1983), 385–388.

[GS83b] N.D. Gupta and S.N. Sidki, *Some infinite p-groups*, Algebra i Logika **22** (5) (1983), 584–589.

[GS84] N.D. Gupta and S.N. Sidki, *Extension of groups by tree automorphisms*, Contributions to Group Theory, Amer. Math. Soc., Providence, RI (1984), 232–246.

[GS87] É. Ghys and V. Sergiescu, *Sur un groupe remarquable de difféomorphismes du cercle*, Comment. Math. Helv. **62** (2) (1987), 185–239.

[GSS88] F.J. Grunewald, D. Segal and G.C. Smith, *Subgroups of finite index in nilpotent groups*, Invent. Math. **93** (1) (1988), 185–223.

[Gui70] Y. Guivarc'h, *Groupes de Lie à croissance polynomiale*, C. R. Acad. Sci. Paris Sér. A–B **271** (1970), A237–A239.

[Gup84] N.D. Gupta, *Recursively presented two generated infinite p-groups*, Math. Z. **188** (1) (1984), 89–90.

[GW] R.I. Grigorchuk and J. Wilson, *A minimality property of certain branch groups*, to appear.

[GW00] R.I. Grigorchuk and J.S. Wilson, *The conjugacy problem for certain branch groups*, Trudy Mat. Inst. Steklov. **231** (4) (2000), 215–230.

[GW01] R.I. Grigorchuk and J. Wilson, *A rigidity property concerning abstract commensurability of subgroups*, Preprint of University of Birmingham (2001).

[GWW92] C. Gordon, D.L. Webb and S. Wolpert, *One cannot hear the shape of a drum*, Bull. Amer. Math. Soc. (N.S.) **27** (1) (1992), 134–138.

[GŻ97] R.I. Grigorchuk and A. Żuk, *On the asymptotic spectrum of random walks on infinite families of graphs*, Proceedings of the Conference "Random Walks and Discrete Potential Theory" (Cortona), M. Picardello and W. Woess, eds, Symposia Mathematica, Cambridge University Press (1997) 134–150.

[Har00] P. de la Harpe, *Topics in Geometric Group Theory*, University of Chicago Press (2000).

[Hig51] G. Higman, *A finitely generated infinite simple group*, J. London Math. Soc. **26** (1951), 61–64.

[Hig74] G. Higman, *Finitely presented infinite simple groups*, Notes on Pure Mathematics, No. 8, Department of Pure Mathematics, Department of Mathematics, I.A.S. Australian National University, Canberra (1974).

[HK97] N. Higson and G.G. Kasparov, *Operator K-theory for groups which act properly and isometrically on Hilbert space*, Electron. Res. Announc. Amer. Math. Soc. **3** (1997), 131–142 (electronic).

[Jac41] N. Jacobson, *Restricted Lie algebras of characteristic p*, Trans. Amer. Math. Soc. **50** (1941), 15–25.

[Jen41] S.A. Jennings, *The structure of the group ring of a p-group over a modular field*, Trans. Amer. Math. Soc. **50** (1941), 175–185.

[Kac66] M. Kac, *Can one hear the shape of a drum?*, Amer. Math. Monthly **73** (4) (1966), 1–23.

[Kal48] L.A. Kaloujnine, *La structure des p-groupes de Sylow des groupes symétriques finis*, Ann. École Norm. Sup. (3) **65** (1948), 239–276.

[Kar87] G. Karpilovsky, *The Schur Multiplier*, Clarendon Press, Oxford University Press, New York (1987).

[KBS91] D.A. Klarner, J.-C. Birget and W. Satterfield, *On the undecidability of the freeness of integer matrix semigroups*, Internat. J. Algebra Comput. **1** (2) (1991), 223–226.

[KLP97] G. Klaas, C.R. Leedham-Green and W. Plesken, *Linear Pro-p-groups of Finite Width*, Lecture Notes in Math., Vol. 1674, Springer, Berlin (1997).

[KM82] M.I. Kargapolov and Y.I. Merzlyakov, *Fundamentals of Group Theory*, 3rd edn, Nauka, Moscow (1982).

[Kos90] A.I. Kostrikin, *Around Burnside*, Springer, Berlin (1990). (Translated from Russian and with a preface by J. Wiegold.)

[Laz53] M. Lazard, *Sur les groupes nilpotents et les anneaux de Lie*, Ann. École Norm. Sup. (3) **71** (1953), 101–190.

[Leo97a] Yu.G. Leonov, *A lower bound for the growth function of periods in Grigorchuk groups*, Mat. Stud. **8** (2) (1997), 192–197, 237.

[Leo97b] Yu.G. Leonov, *On identities in groups of automorphisms of trees*, Visnyk of Kyiv State University of T.G. Shevchenko (3) (1997), 37–44.

[Leo98a] Yu.G. Leonov, *The conjugacy problem in a class of* 2*-groups*, Mat. Zametki **64** (4) (1998), 573–583.

[Leo98b] Yu.G. Leonov, *On growth function for some torsion residually finite groups*, International Conference dedicated to the 90th Anniversary of L.S. Pontryagin (Moscow), Algebra, Steklov Mathematical Institute (September 1998), 36–38.

[Leo99] Yu.G. Leonov, *On precisement of estimation of periods' growth for Grigorchuk's 2-groups*, Unpublished (1999).

[Leo00] Yu.G. Leonov, *On a lower bound for the growth function of the Grigorchuk group*, Mat. Zametki **67** (3) (2000), 475–477.

[LM89] A. Lubotzky and A. Mann, *Residually finite groups of finite rank*, Math. Proc. Cambridge Philos. Soc. **106** (3) (1989), 385–388.

[LM91] A. Lubotzky and A. Mann, *On groups of polynomial subgroup growth*, Invent. Math. **104** (3) (1991), 521–533.

[LMS93] A. Lubotzky, A. Mann and D. Segal, *Finitely generated groups of polynomial subgroup growth*, Israel J. Math. **82** (1–3) (1993), 363–371.

[LN02] Y. Lavreniuk and V. Nekrashevich, *Rigidity of branch groups acting on rooted trees* (2002).

[Lub95a] A. Lubotzky, *Cayley graphs: Eigenvalues, expanders and random walks*, Surveys in Combinatorics, 1995 (Stirling), Cambridge Univ. Press, Cambridge (1995), 155–189.

[Lub95b] A. Lubotzky, *Subgroup growth*, Proceedings of the International Congress of Mathematicians, Vols. 1, 2 (Zürich, 1994) (Basel), Birkhäuser (1995), 309–317.

[Lys85] I.G. Lysionok, *A system of defining relations for the Grigorchuk group*, Mat. Zametki **38** (1985), 503–511.

[Lys98] I.G. Lysionok, *Growth of orders of elements in the Grigorchuk 2-group*, Unpublished (1998).

[Mac76] G.W. Mackey, *The Theory of Unitary Group Representations*, University of Chicago Press, Chicago, IL (1976). (Based on notes by J.M.G. Fell and D.B. Lowdenslager of lectures given at the University of Chicago, Chicago, IL, Chicago Lectures in Mathematics (1955).)

[Mag40] W. Magnus, *Über Gruppen und zugeordnete Liesche Ringe*, J. Reine Angew. Math. **182** (1940), 142–149.

[Mil68a] J.W. Milnor, *Growth of finitely generated solvable groups*, J. Differential Geom. **2** (1968), 447–449.

[Mil68b] J.W. Milnor, *A note on curvature and fundamental group*, J. Differential Geom. **2** (1968), 1–7.

[Mil68c] J.W. Milnor, *Problem 5603*, Amer. Math. Monthly **75** (1968), 685–686.

[MP01] R. Muchnik and I. Pak, *On growth of Grigorchuk groups*, Preprint (2001).

[MS81] G.A. Margulis and G.A. Soĭfer, *Maximal subgroups of infinite index in finitely generated linear groups*, J. Algebra **69** (1) (1981), 1–23.

[MS90] A. Mann and D. Segal, *Uniform finiteness conditions in residually finite groups*, Proc. London Math. Soc. (3) **61** (3) (1990), 529–545.

[Neu86] P.M. Neumann, *Some questions of Edjvet and Pride about infinite groups*, Illinois J. Math. **30** (2) (1986), 301–316.

[Ol'91] A.Yu. Ol'shanskiĭ, *Geometry of Defining Relations in Groups*, Kluwer Academic Publishers Group, Dordrecht (1991). (Translated from the 1989 Russian original by Yu.A. Bakhturin.)

[Per00] E.L. Pervova, *Everywhere dense subgroups of one group of tree automorphisms*, Proc. Steklov Inst. Math. **231** (2000), 339–350.

[Pet99] V.M. Petrogradskiĭ, *Growth of finitely generated polynilpotent Lie algebras and groups, generalized partitions, and functions analytic in the unit circle*, Internat. J. Algebra Comput. **9** (2) (1999), 179–212.

[Pre98] J.P. Previte, *Graph substitutions*, Ergodic Theory Dynam. Systems **18** (3) (1998), 661–685.

[Pri80] S.J. Pride, *The concept of "largeness" in group theory*, Word Problems, II, Conf. on Decision Problems in Algebra (Oxford, 1976), North-Holland, Amsterdam (1980), 299–335.

[PS75] I. Bir, S. Passi and S.K. Sehgal, *Lie dimension subgroups*, Comm. Algebra **3** (1975), 59–73.

[PT96] D.S. Passman and W.V. Temple, *Representations of the Gupta–Sidki group*, Proc. Amer. Math. Soc. **124** (5) (1996), 1403–1410.

[Qui68] D.G. Quillen, *On the associated graded ring of a group ring*, J. Algebra **10** (1968), 411–418.

[Rob96] D.J.S. Robinson, *A Course in the Theory of Groups*, 2nd edn, Springer, New York (1996).

[Röv99] C.E. Röver, *Constructing finitely presented simple groups that contain Grigorchuk groups*, J. Algebra **220** (1) (1999), 284–313.

[Röv00] C.E. Röver, *Abstract commensurators of groups acting on rooted trees*, Preprint (2000).

[Roz86] A.V. Rozhkov, *On the theory of Alyoshin-type groups*, Mat. Zametki **40** (5) (1986), 572–589.

[Roz90] A.V. Rozhkov, *Stabilizers of sequences in Alëshin-type groups*, Mat. Zametki **47** (3) (1990), 121–128, 143.

[Roz96] A.V. Rozhkov, *Lower central series of a group of tree automorphisms*, Mat. Zametki **60** (2) (1996), 225–237, 319.

[Roz98] A.V. Rozhkov, *The conjugacy problem in an automorphism group of an infinite tree*, Mat. Zametki **64** (4) (1998), 592–597.

[RS80] G. Rozenberg and A. Salomaa, *The Mathematical Theory of L Systems*, Academic Press Inc., New York (1980).

[RZ00] L. Ribes and P. Zaleskiĭ, *Pro-p Trees and Applications*, Progr. Math., Vol. 184, Birkhäuser Boston, Boston, MA (2000).

[S$^+$93] M. Schönert et al., *GAP: Groups, Algorithms and Programming*, RWTH, Aachen (1993).

[Seg86a] D. Segal, *Subgroups of finite index in soluble groups. I*, Proceedings of Groups – St. Andrews 1985, Cambridge Univ. Press, Cambridge (1986), 307–314.

[Seg86b] D. Segal, *Subgroups of finite index in soluble groups. II*, Proceedings of groups – St. Andrews 1985, Cambridge Univ. Press, Cambridge (1986), 315–319.

[Seg00] D. Segal, *The finite images of finitely generated groups*, Proc. Conf. Bielefeld (2000).

[Seg01] D. Segal, *The finite images of finitely generated groups*, Proc. London Math. Soc. (3) **82** (3) (2001), 597–613.

[Sid87a] S.N. Sidki, *On a* 2*-generated infinite* 3*-group: subgroups and automorphisms*, J. Algebra **110** (1) (1987), 24–55.

[Sid87b] S.N. Sidki, *On a* 2*-generated infinite* 3*-group: the presentation problem*, J. Algebra **110** (1) (1987), 13–23.

[Sid97] S.N. Sidki, *A primitive ring associated to a Burnside* 3*-group*, J. London Math. Soc. (2) **55** (1) (1997), 55–64.

[Šun00] Z. Šunić, *On a class of periodic spinal groups of intermediate growth*, Ph.D. thesis, State University of New York, Binghamton (2000).

[Suš79] V.I. Suščans'kiĭ, *Periodic p-groups of permutations and the unrestricted Burnside problem*, Dokl. Akad. Nauk SSSR **247** (3) (1979), 557–561.

[Sus89] V.I. Sushchanskiĭ, *Wreath products and periodic factorizable groups*, Mat. Sb. **180** (8) (1989), 1073–1091, 1151.

[Sus94] V.I. Sushchanskiĭ, *Wreath products and factorizable groups*, Algebra i Analiz **6** (1) (1994), 203–238.

[Šva55] A.S. Švarc, *A volume invariant of coverings*, Dokl. Akad. Nauk SSSR (105) (1955), 32–34 (in Russian).

[SW02] S.N. Sidki and J.S. Wilson, *Free subgroups of branch groups*, Preprint (2002).

[Tit72] J. Tits, *Free subgroups in linear groups*, J. Algebra **20** (1972), 250–270.

[vN29] J. von Neumann, *Zur allgemeinen Theorie des Masses*, Fund. Math. **13** (1929), 73–116, 333; Collected Works, Vol. I (1961), 599–643.

[Vov00] T. Vovkivsky, *Infinite torsion groups arising as generalizations of the second Grigorchuk group*, Algebra (Moscow, 1998), de Gruyter, Berlin (2000), 357–377.

[Wil71] J.S. Wilson, *Groups with every proper quotient finite*, Math. Proc. Cambridge Philos. Soc. **69** (1971), 373–391.

[Wil00] J.S. Wilson, *On just infinite abstract and profinite groups*, New Horizons in Pro-p Groups, Birkhäuser Boston, Boston, MA (2000), 181–203.

[Wol68] J.A. Wolf, *Growth of finitely generated solvable groups and curvature of Riemanniann manifolds*, J. Differential Geom. **2** (1968), 421–446.

[WZ97] J.S. Wilson and P.A. Zaleskiĭ, *Conjugacy separability of certain torsion groups*, Arch. Math. (Basel) **68** (6) (1997), 441–449.

[Zas40] H. Zassenhaus, *Ein Verfahren, jeder endlichen p-Gruppe einen Lie-Ring mit der Characteristik p zuzuordnen*, Abh. Math. Sem. Univ. Hamburg **13** (1940), 200–207.

[Zel90] E.I. Zel'manov, *Solution of the restricted Burnside problem for groups of odd exponent*, Izv. Akad. Nauk SSSR Ser. Mat. **54** (1) (1990), 42–59, 221.

[Zel91] E.I. Zel'manov, *Solution of the restricted Burnside problem for* 2*-groups*, Mat. Sb. **182** (4) (1991), 568–592.

Subject Index

This Page Intentionally Left Blank