Algebra: Chapter 0

Paolo Aluffi

Graduate Studies
in Mathematics
Volume 104



American Mathematical Society

Algebra: Chapter 0

Algebra: Chapter 0

Paolo Aluffi

Graduate Studies
in Mathematics
Volume 104



American Mathematical Society
Providence, Rhode Island

Editorial Board

David Cox (Chair)
Steven G. Krantz
Rafe Mazzeo
Martin Scharlemann

2000 *Mathematics Subject Classification.* Primary 00–01; Secondary 12–01, 13–01, 15–01, 18–01, 20–01.

For additional information and updates on this book, visit
www.ams.org/bookpages/gsm-104

Library of Congress Cataloging-in-Publication Data

Aluffi, Paolo, 1960–
Algebra: chapter 0 /Paolo Aluffi.
p. cm. — (Graduate studies in mathematics ; v. 104)
Includes index.
ISBN 978-0-8218-4781-7 (alk. paper)
1. Algebra—Textbooks. I. Title.

QA154.3.A527 2009
512—dc22 2009004043

Copying and reprinting. Individual readers of this publication, and nonprofit libraries acting for them, are permitted to make fair use of the material, such as to copy a chapter for use in teaching or research. Permission is granted to quote brief passages from this publication in reviews, provided the customary acknowledgment of the source is given.

Republication, systematic copying, or multiple reproduction of any material in this publication is permitted only under license from the American Mathematical Society. Requests for such permission should be addressed to the Acquisitions Department, American Mathematical Society, 201 Charles Street, Providence, Rhode Island 02904-2294, USA. Requests can also be made by e-mail to `reprint-permission@ams.org`.

© 2009 by the American Mathematical Society. All rights reserved.
The American Mathematical Society retains all rights
except those granted to the United States Government.
Printed in the United States of America.

⊗ The paper used in this book is acid-free and falls within the guidelines
established to ensure permanence and durability.
Visit the AMS home page at `http://www.ams.org/`

10 9 8 7 6 5 4 3 2 1 14 13 12 11 10 09

Contents

Introduction xv

Chapter I. Preliminaries: Set theory and categories 1

§1. Naive set theory 1
1.1. Sets 1
1.2. Inclusion of sets 3
1.3. Operations between sets 4
1.4. Disjoint unions, products 5
1.5. Equivalence relations, partitions, quotients 6

Exercises 8

§2. Functions between sets 8
2.1. Definition 8
2.2. Examples: Multisets, indexed sets 10
2.3. Composition of functions 10
2.4. Injections, surjections, bijections 11
2.5. Injections, surjections, bijections: Second viewpoint 12
2.6. Monomorphisms and epimorphisms 13
2.7. Basic examples 15
2.8. Canonical decomposition 15
2.9. Clarification 16

Exercises 17

§3. Categories 18
3.1. Definition 18
3.2. Examples 20

Exercises 26

§4. Morphisms 27
4.1. Isomorphisms 27
4.2. Monomorphisms and epimorphisms 29

Exercises 30
§5. Universal properties 31
5.1. Initial and final objects 31
5.2. Universal properties 33
5.3. Quotients 33
5.4. Products 35
5.5. Coproducts 36
Exercises 38

Chapter II. Groups, first encounter 41
§1. Definition of group 41
1.1. Groups and groupoids 41
1.2. Definition 42
1.3. Basic properties 43
1.4. Cancellation 45
1.5. Commutative groups 45
1.6. Order 46
Exercises 48
§2. Examples of groups 49
2.1. Symmetric groups 49
2.2. Dihedral groups 52
2.3. Cyclic groups and modular arithmetic 54
Exercises 56
§3. The category Grp 58
3.1. Group homomorphisms 58
3.2. Grp: Definition 59
3.3. Pause for reflection 60
3.4. Products et al. 61
3.5. Abelian groups 62
Exercises 63
§4. Group homomorphisms 64
4.1. Examples 64
4.2. Homomorphisms and order 66
4.3. Isomorphisms 66
4.4. Homomorphisms of abelian groups 68
Exercises 69
§5. Free groups 70
5.1. Motivation 70
5.2. Universal property 71
5.3. Concrete construction 72
5.4. Free *abelian* groups 75
Exercises 78
§6. Subgroups 79
6.1. Definition 79

6.2. Examples: Kernel and image 80
6.3. Example: Subgroup generated by a subset 81
6.4. Example: Subgroups of cyclic groups 82
6.5. Monomorphisms 84
Exercises 85
§7. Quotient groups 88
7.1. Normal subgroups 88
7.2. Quotient group 89
7.3. Cosets 90
7.4. Quotient by normal subgroups 92
7.5. Example 94
7.6. kernel $\iff$ normal 95
Exercises 95
§8. Canonical decomposition and Lagrange's theorem 96
8.1. Canonical decomposition 97
8.2. Presentations 99
8.3. Subgroups of quotients 100
8.4. HK/H vs. $K/(H \cap K)$ 101
8.5. The index and Lagrange's theorem 102
8.6. Epimorphisms and cokernels 104
Exercises 105
§9. Group actions 108
9.1. Actions 108
9.2. Actions on sets 109
9.3. Transitive actions and the category G-Set 110
Exercises 113
§10. Group objects in categories 115
10.1. Categorical viewpoint 115
Exercises 117
Chapter III. Rings and modules 119
§1. Definition of ring 119
1.1. Definition 119
1.2. First examples and special classes of rings 121
1.3. Polynomial rings 124
1.4. Monoid rings 126
Exercises 127
§2. The category Ring 129
2.1. Ring homomorphisms 129
2.2. Universal property of polynomial rings 130
2.3. Monomorphisms and epimorphisms 132
2.4. Products 133
2.5. $\mathrm{End}_{\mathsf{Ab}}(G)$ 134
Exercises 136

§3. Ideals and quotient rings 138
3.1. Ideals 138
3.2. Quotients 139
3.3. Canonical decomposition and consequences 141
Exercises 143
§4. Ideals and quotients: Remarks and examples. Prime and maximal ideals 144
4.1. Basic operations 144
4.2. Quotients of polynomial rings 146
4.3. Prime and maximal ideals 150
Exercises 153
§5. Modules over a ring 156
5.1. Definition of (left-)R-module 156
5.2. The category R-Mod 158
5.3. Submodules and quotients 160
5.4. Canonical decomposition and isomorphism theorems 162
Exercises 163
§6. Products, coproducts, etc., in R-Mod 164
6.1. Products and coproducts 164
6.2. Kernels and cokernels 166
6.3. Free modules and free algebras 167
6.4. Submodule generated by a subset; Noetherian modules 169
6.5. Finitely generated vs. finite type 171
Exercises 172
§7. Complexes and homology 174
7.1. Complexes and exact sequences 174
7.2. Split exact sequences 177
7.3. Homology and the snake lemma 178
Exercises 183
Chapter IV. Groups, second encounter 187
§1. The conjugation action 187
1.1. Actions of groups on sets, reminder 187
1.2. Center, centralizer, conjugacy classes 189
1.3. The Class Formula 190
1.4. Conjugation of subsets and subgroups 191
Exercises 193
§2. The Sylow theorems 194
2.1. Cauchy's theorem 194
2.2. Sylow I 196
2.3. Sylow II 197
2.4. Sylow III 199
2.5. Applications 200
Exercises 202

§3. Composition series and solvability 205
3.1. The Jordan-Hölder theorem 205
3.2. Composition factors; Schreier's theorem 207
3.3. The commutator subgroup, derived series, and solvability 210
Exercises 213
§4. The symmetric group 214
4.1. Cycle notation 214
4.2. Type and conjugacy classes in S_n 216
4.3. Transpositions, parity, and the alternating group 219
4.4. Conjugacy in A_n; simplicity of A_n and solvability of S_n 220
Exercises 224
§5. Products of groups 226
5.1. The direct product 226
5.2. Exact sequences of groups; extension problem 228
5.3. Internal/semidirect products 230
Exercises 233
§6. Finite abelian groups 234
6.1. Classification of finite abelian groups 234
6.2. Invariant factors and elementary divisors 237
6.3. Application: Finite subgroups of multiplicative groups of fields 239
Exercises 240

Chapter V. Irreducibility and factorization in integral domains 243
§1. Chain conditions and existence of factorizations 244
1.1. Noetherian rings revisited 244
1.2. Prime and irreducible elements 246
1.3. Factorization into irreducibles; domains with factorizations 248
Exercises 249
§2. UFDs, PIDs, Euclidean domains 251
2.1. Irreducible factors and greatest common divisor 251
2.2. Characterization of UFDs 253
2.3. PID $\Longrightarrow$ UFD 254
2.4. Euclidean domain $\Longrightarrow$ PID 255
Exercises 258
§3. Intermezzo: Zorn's lemma 261
3.1. Set theory, reprise 261
3.2. Application: Existence of maximal ideals 264
Exercises 265
§4. Unique factorization in polynomial rings 267
4.1. Primitivity and content; Gauss's lemma 268
4.2. The field of fractions of an integral domain 270
4.3. R UFD $\Longrightarrow$ $R[x]$ UFD 273
Exercises 276

§5. Irreducibility of polynomials 280
5.1. Roots and reducibility 281
5.2. Adding roots; algebraically closed fields 283
5.3. Irreducibility in $\mathbb{C}[x]$, $\mathbb{R}[x]$, $\mathbb{Q}[x]$ 285
5.4. Eisenstein's criterion 288
Exercises 289
§6. Further remarks and examples 291
6.1. Chinese remainder theorem 291
6.2. Gaussian integers 294
6.3. Fermat's theorem on sums of squares 298
Exercises 300

Chapter VI. Linear algebra 305
§1. Free modules revisited 305
1.1. R-Mod 305
1.2. Linear independence and bases 306
1.3. Vector spaces 308
1.4. Recovering B from $F^R(B)$ 309
Exercises 312
§2. Homomorphisms of free modules, I 314
2.1. Matrices 314
2.2. Change of basis 318
2.3. Elementary operations and Gaussian elimination 320
2.4. Gaussian elimination over Euclidean domains 323
Exercises 324
§3. Homomorphisms of free modules, II 327
3.1. Solving systems of linear equations 327
3.2. The determinant 328
3.3. Rank and nullity 333
3.4. Euler characteristic and the Grothendieck group 334
Exercises 338
§4. Presentations and resolutions 340
4.1. Torsion 340
4.2. Finitely presented modules and free resolutions 341
4.3. Reading a presentation 344
Exercises 347
§5. Classification of finitely generated modules over PIDs 349
5.1. Submodules of free modules 350
5.2. PIDs and resolutions 353
5.3. The classification theorem 354
Exercises 357
§6. Linear transformations of a free module 359
6.1. Endomorphisms and similarity 359
6.2. The characteristic and minimal polynomials of an endomorphism 361

6.3. Eigenvalues, eigenvectors, eigenspaces 365
Exercises 368
§7. Canonical forms 371
7.1. Linear transformations of free modules; actions of polynomial rings 371
7.2. $k[t]$-modules and the rational canonical form 373
7.3. Jordan canonical form 377
7.4. Diagonalizability 380
Exercises 381
Chapter VII. Fields 385
§1. Field extensions, I 385
1.1. Basic definitions 385
1.2. Simple extensions 387
1.3. Finite and algebraic extensions 391
Exercises 397
§2. Algebraic closure, Nullstellensatz, and a little algebraic geometry 400
2.1. Algebraic closure 400
2.2. The Nullstellensatz 404
2.3. A little affine algebraic geometry 406
Exercises 414
§3. Geometric impossibilities 417
3.1. Constructions by straightedge and compass 417
3.2. Constructible numbers and quadratic extensions 422
3.3. Famous impossibilities 425
Exercises 427
§4. Field extensions, II 428
4.1. Splitting fields and normal extensions 429
4.2. Separable polynomials 433
4.3. Separable extensions and embeddings in algebraic closures 436
Exercises 438
§5. Field extensions, III 440
5.1. Finite fields 441
5.2. Cyclotomic polynomials and fields 445
5.3. Separability and simple extensions 449
Exercises 452
§6. A little Galois theory 454
6.1. The Galois correspondence and Galois extensions 454
6.2. The fundamental theorem of Galois theory, I 459
6.3. The fundamental theorem of Galois theory, II 461
6.4. Further remarks and examples 464
Exercises 466
§7. Short march through applications of Galois theory 468
7.1. Fundamental theorem of algebra 468

7.2. Constructibility of regular n-gons 469
7.3. Fundamental theorem on symmetric functions 471
7.4. Solvability of polynomial equations by radicals 474
7.5. Galois groups of polynomials 478
7.6. Abelian groups as Galois groups over $\mathbb{Q}$ 479
Exercises 480

Chapter VIII. Linear algebra, reprise 483
§1. Preliminaries, reprise 483
1.1. Functors 483
1.2. Examples of functors 485
1.3. When are two categories 'equivalent'? 487
1.4. Limits and colimits 489
1.5. Comparing functors 492
Exercises 496
§2. Tensor products and the Tor functors 500
2.1. Bilinear maps and the definition of tensor product 501
2.2. Adjunction with Hom and explicit computations 504
2.3. Exactness properties of tensor; flatness 507
2.4. The Tor functors. 509
Exercises 511
§3. Base change 515
3.1. Balanced maps 515
3.2. Bimodules; adjunction again 517
3.3. Restriction and extension of scalars 518
Exercises 520
§4. Multilinear algebra 522
4.1. Multilinear, symmetric, alternating maps 522
4.2. Symmetric and exterior powers 524
4.3. Very small detour: Graded algebra 527
4.4. Tensor algebras 529
Exercises 532
§5. Hom and duals 535
5.1. Adjunction again 536
5.2. Dual modules 537
5.3. Duals of free modules 538
5.4. Duality and exactness 539
5.5. Duals and matrices; biduality 541
5.6. Duality on vector spaces 542
Exercises 543
§6. Projective and injective modules and the Ext functors 545
6.1. Projectives and injectives 546
6.2. Projective modules 547
6.3. Injective modules 548

6.4. The Ext functors 551
6.5. $\mathrm{Ext}^*_{\mathbb{Z}}(\mathbf{G}, \mathbb{Z})$ 554
Exercises 555

Chapter IX. Homological algebra 559
§1. (Un)necessary categorical preliminaries 560
1.1. Undesirable features of otherwise reasonable categories 560
1.2. Additive categories 561
1.3. Abelian categories 564
1.4. Products, coproducts, and direct sums 567
1.5. Images; canonical decomposition of morphisms 570
Exercises 574
§2. Working in abelian categories 576
2.1. Exactness in abelian categories 576
2.2. The snake lemma, again 578
2.3. Working with 'elements' in a small abelian category 581
2.4. What is missing? 587
Exercises 589
§3. Complexes and homology, again 591
3.1. Reminder of basic definitions; general strategy 591
3.2. The category of complexes 594
3.3. The long exact cohomology sequence 597
3.4. Triangles 600
Exercises 602
§4. Cones and homotopies 605
4.1. The mapping cone of a morphism 605
4.2. Quasi-isomorphisms and derived categories 607
4.3. Homotopy 611
Exercises 614
§5. The homotopic category. Complexes of projectives and injectives 616
5.1. Homotopic maps are identified in the derived category 616
5.2. Definition of the homotopic category of complexes 618
5.3. Complexes of projective and injective objects 619
5.4. Homotopy equivalences vs. quasi-isomorphisms in $\mathsf{K}(\mathsf{A})$ 620
5.5. Proof of Theorem 5.9 624
Exercises 626
§6. Projective and injective resolutions and the derived category 628
6.1. Recovering A 629
6.2. From objects to complexes 631
6.3. Poor man's derived category 635
Exercises 638
§7. Derived functors 641
7.1. Viewpoint shift 641
7.2. Universal property of the derived functor 643

7.3. Taking cohomology 645
7.4. Long exact sequence of derived functors 647
7.5. Relating $\mathscr{F}$, $\mathsf{L}_i\mathscr{F}$, $\mathsf{R}^i\mathscr{F}$ 653
7.6. Example: A little group cohomology 655
Exercises 658
§8. Double complexes 661
8.1. Resolution by acyclic objects 662
8.2. Complexes of complexes 665
8.3. Exactness of the total complex 670
8.4. Total complexes and resolutions 672
8.5. Acyclic resolutions again and balancing Tor and Ext 675
Exercises 677
§9. Further topics 680
9.1. Derived categories 681
9.2. Triangulated categories 683
9.3. Spectral sequences 686
Exercises 695
Index 699

Introduction

This text presents an introduction to algebra suitable for upper-level undergraduate or beginning graduate courses. While there is a very extensive offering of textbooks at this level, in my experience teaching this material I have invariably felt the need for a self-contained text that would start 'from zero' (in the sense of not assuming that the reader has had substantial previous exposure to the subject) but that would impart from the very beginning a rather modern, categorically minded viewpoint and aim at reaching a good level of depth. Many textbooks in algebra brilliantly satisfy some, but not all, of these requirements. This book is my attempt at providing a working alternative.

There is a widespread perception that categories should be avoided at first blush, that the abstract language of categories should not be introduced until a student has toiled for a few semesters through example-driven illustrations of the nature of a subject like algebra. According to this viewpoint, categories are only tangentially relevant to the main topics covered in a beginning course, so they can simply be mentioned occasionally for the general edification of the reader, who will in time learn about them (by osmosis?). Paraphrasing a reviewer of a draft of the present text, 'Discussions of categories at this level are the reason why God created appendices.'

It will be clear from a cursory glance at the table of contents that I think otherwise. In this text, categories are introduced on page 18, after a scant reminder of the basic language of naive set theory, for the main purpose of providing a context for universal properties. These are in turn evoked constantly as basic definitions are introduced. The word 'universal' appears at least 100 times in the first three chapters.

I believe that awareness of the categorical language, and especially some appreciation of universal properties, is particularly helpful in approaching a subject such as algebra 'from the beginning'. The reader I have in mind is someone who has reached a certain level of mathematical maturity—for example, who needs no

special assistance in grasping an induction argument—but may have only been exposed to algebra in a very cursory manner. My experience is that many upper-level undergraduates or beginning graduate students at Florida State University and at comparable institutions fit this description. For these students, seeing the many introductory concepts in algebra as instances of a few powerful ideas (encapsulated in suitable universal properties) helps to build a comforting unifying context for these notions. The amount of categorical language needed for this catalyzing function is very limited; for example, functors are not really necessary in this acclimatizing stage.

Thus, in my mind the benefit of this approach is precisely that it helps a true beginner, if it is applied with due care. This is my experience in the classroom, and it is the main characteristic feature of this text. The very little categorical language introduced at the outset informs the first part of the book, introducing in general terms groups, rings, and modules. This is followed by a (rather traditional) treatment of standard topics such as Sylow theorems, unique factorization, elementary linear algebra, and field theory. The last third of the book wades into somewhat deeper waters, dealing with tensor products and Hom (including a first introduction to Tor and Ext) and including a final chapter devoted to homological algebra. Some familiarity with categorical language appears indispensable to me in order to appreciate this latter material, and this is hopefully uncontroversial. Having developed a feel for this language in the earlier parts of the book, students find the transition into these more advanced topics particularly smooth.

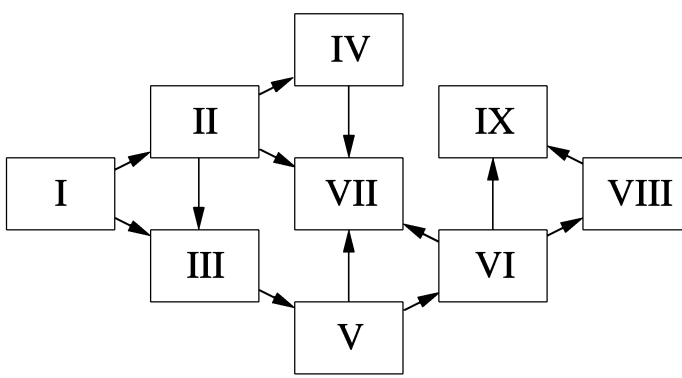
A first version of this book was essentially a careful transcript of my lectures in a run of the (three-semester) algebra sequence at FSU. The chapter on homological algebra was added at the instigation of Ed Dunne, as were a very substantial number of the exercises. The main body of the text has remained very close to the original 'transcript' version: I have resisted the temptation of expanding the material when revising it for publication. I believe that an effective introductory textbook (this is Chapter 0, after all...) should be realistic: it must be possible to cover in class what is covered in the book. Otherwise, the book veers into the 'reference' category; I never meant to write a reference book in algebra, and it would be futile (of me) to try to ameliorate excellent available references such as Lang's 'Algebra'.

The problem sets will give an opportunity to a teacher, or any motivated reader, to get quite a bit beyond what is covered in the main text. To guide in the choice of exercises, I have marked with a $\triangleright$ those problems that are directly referenced from the text, and with a $\neg$ those problems that are referenced from other problems. A minimalist teacher may simply assign all and only the $\triangleright$ problems; these do nothing more than anchor the understanding by practice and may be all that a student can realistically be expected to work out while juggling TA duties and two or three other courses of similar intensity as this one. The main body of the text, together with these exercises, forms a self-contained presentation of essential material. The other exercises, and especially the threads traced by those marked with $\neg$, will offer the opportunity to cover other topics, which some may well consider just as essential: the modular group, quaternions, nilpotent groups, Artinian rings, the Jacobson radical, localization, Lagrange's theorem on four squares, projective space and

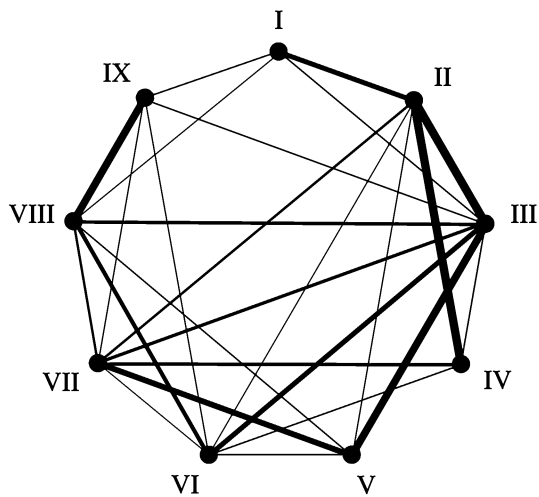
Grassmannians, Nakayama's lemma, associated primes, the spectral theorem for normal operators, etc., are some examples of topics that make their appearance in the exercises. Often a topic is presented over the course of several exercises, placed in appropriate sections of the book. For example, 'Wedderburn's little theorem' is mentioned in Remark III.1.16 (that is: Remark 1.16 in Chapter III); particular cases are presented in Exercises III.2.11 and IV.2.17, and the reader eventually obtains a proof in Exercise VII.5.14, following preliminaries given in Exercises VII.5.12 and VII.5.13. The ¬ label and perusal of the index should facilitate the navigation of such topics. To help further in this process, I have decorated every exercise with a list (added in square brackets) of the places in the book that refer to it. For example, an instructor evaluating whether to assign Exercise V.2.25 will be immediately aware that this exercise is quoted in Exercise VII.5.18, proving a particular case of Dirchlet's theorem on primes in arithmetic progressions, and that this will in turn be quoted in §VII.7.6, discussing the realization of abelian groups as Galois groups over $\mathbb{Q}$.

I have put a high priority on the requirement that this should be a self-contained text which essentially crosses all t's and dots all i's, and does not require that the reader have access to other texts while working through it. I have therefore made a conscious effort to *not* quote other references: I have avoided as much as possible the exquisitely tempting escape route 'For a proof, see ....' This is the main reason why this book is as thick as it is, even if so many topics are *not* covered in it. Among these, commutative algebra and representation theory are perhaps the most glaring omissions. The first is represented to the extent of the standard basic definitions, which allow me to sprinkle a little algebraic geometry here and there (for example, see §VII.2), and of a few slightly more advanced topics in the exercises, but I stopped short of covering, e.g., primary decompositions. The second is missing altogether. It is my hope to complement this book with a 'Chapter 1' in an undetermined future, where I will make amends for these and other shortcomings.

By its nature, this book should be quite suitable for self-study: readers working on their own will find here a self-contained starting point which should work well as a prelude to future, more intensive, explorations. Such readers may be helped by the following '9-fold way' diagram of logical interdependence of the chapters:



This may however better reflect my original intention than the final product. For a more objective gauge, this alternative diagram captures the web of references from a chapter to earlier chapters, with the thickness of the lines representing (roughly) the number of references:



With the self-studying reader especially in mind, I have put extra effort into providing an extensive index. It is not realistic to make a fanfare for each and every new term introduced in a text of this size by an official 'definition'; the index should help a lone traveler find the way back to the source of unfamiliar terminology.

Internal references are handled in a hopefully transparent way. For example, Remark III.1.16 refers to Remark 1.16 in Chapter III; if the reference is made from within Chapter III, the same item is called Remark 1.16. The list in brackets following an exercise indicates other exercises or sections in the book referring to that exercise. For example, Exercise 3.1 in Chapter I is followed by [5.1, §VIII.1.1, §IX.1.2, IX.1.10]: this alerts the reader that there are references to this problem in Exercise 5.1 in Chapter I, section 1.1 in Chapter VIII, section 1.2 in Chapter IX, and Exercise 1.10 in Chapter IX (and nowhere else).

Acknowledgments. My debt to Lang's book, to David Dummit and Richard Foote's 'Abstract Algebra,' or to Artin's 'Algebra' will be evident to anyone who is familiar with these sources. The chapter on homological algebra owes much to David Eisenbud's appendix on the topic in his 'Commutative Algebra', to Gelfand-Manin's 'Methods of homological algebra', and to Weibel's 'An introduction to homological algebra'. But in most cases it would simply be impossible for me to retrace the original source of an expository idea, of a proof, of an exercise, or of a specific pedagogical emphasis: these are all likely offsprings of ideas from any one of these and other influential references and often of associations triggered by following the manifold strands of the World Wide Web. This is another reason why, in a spirit of equanimity, I resolved to essentially avoid references altogether. In any case, I believe all the material I have presented here is standard, and I only retain absolute ownership of every error left in the end product.

I am very grateful to my students for the constant feedback that led me to write this book in this particular way and who contributed essentially to its success in my classes. Some of the students provided me with extensive lists of typos and outright mistakes, and I would especially like to thank Kevin Meek, Jay Stryker, and Yong Jae Cha for their particularly helpful comments. I had the opportunity to try out the material on homological algebra in a course given at Caltech in the fall of 2008, while on a sabbatical from FSU, and I would like to thank Caltech and the audience of the course for their hospitality and the friendly atmosphere. Thanks are also due to MSRI for hospitality during the winter of 2009, when the last fine-tuning of the text was performed.

A few people spotted big and small mistakes in preliminary versions of this book, and I will mention Georges Elencwajg, Xia Liao, and Mirroslav Yotov for particularly precious contributions. I also commend Arlene O'Sean and the staff at the AMS for the excellent copyediting and production work.

Special thanks go to Ettore Aldrovandi for expert advice, to Matilde Marcolli for her encouragement and indispensable help, and to Ed Dunne for suggestions that had a great impact in shaping the final version of this book.

Support from the Max-Planck-Institut in Bonn, from the NSA, and from Caltech, at different stages of the preparation of this book, is gratefully acknowledged.

Preliminaries: Set theory and categories

Set theory is a mathematical field in itself, and its proper treatment (say via the famous 'Zermelo-Fränkel' axioms) goes well beyond the scope of this book and the competence of this writer. We will only deal with so-called 'naive' set theory, which is little more than a system of notation and terminology enabling us to express precisely mathematical definitions, statements, and their proofs.

Familiarity with this language is essential in approaching a subject such as algebra, and indeed the reader is assumed to have been previously exposed to it. In this chapter we first review some of the language of naive set theory, mainly in order to establish the notation we will use in the rest of the book. We will then get a small taste of the language of *categories,* which plays a powerful unifying role in algebra and many other fields. Our main objective is to convey the notion of 'universal property', which will be a constant refrain throughout this book.

1. Naive set theory

1.1. Sets. The notion of *set* formalizes the intuitive idea of 'collection of objects'. A set is determined by the *elements* it contains: two sets A, B are equal (written $A = B$) if and only if they contain precisely the same elements. 'What is an *element*?' is a forbidden question in naive set theory: the buck must stop somewhere. We can conveniently pretend that a 'universe' of elements is available to us, and we draw from this universe to construct the elements and sets we need, implicitly assuming that all the operations we will explore can be performed within this universe. (This is the tricky point!) In any case, we specify a *set* by giving a precise recipe determining which elements are in it. This definition is usually put between braces and may consist of a simple, complete, list of elements:

$$A := \{1, 2, 3\}$$

is[1] the set consisting of the integers 1, 2, and 3. By convention, the order[2] in which the elements are listed, or repetitions in the list, are immaterial to the definition. Thus, the same set may be written out in many ways:

$$\{1,2,3\} = \{1,3,2\} = \{1,2,1,3,3,2,3,1,1,2,1,3\}.$$

This way of denoting sets may be quite cumbersome and in any case will only really work for *finite* sets. For infinite sets, a popular way around this problem is to write a list in which some of the elements are understood as being part of a pattern—for example, the set of even integers may be written

$$E = \{\cdots, -2, 0, 2, 4, 6, \cdots\},$$

but such a definition is inherently ambiguous, so this leaves room for misinterpretation. Further, some sets are simply 'too big' to be listed, even in principle: for example (as one hopefully learns in advanced calculus) there are simply too many *real numbers* to be able to 'list' them as one may 'list' the integers.

It is often better to adopt definitions that express the elements of a set as elements s of some larger (and already known) set S, satisfying some property P. One may then write

$$A = \{s \in S \mid s \text{ satisfies } P\}$$

($\in$ means *element of...*) and this is in general precise and unambiguous[3].

We will occasionally encounter a variation on the notion of set, called 'multiset'. A multiset is a set in which the elements are allowed to appear 'with multiplicity': that is, a notion for which $\{2,2\}$ would be *distinct* from $\{2\}$. The correct way to define a multiset is by means of *functions,* which we will encounter soon (see Example 2.2).

A few famous sets are

- $\emptyset$: the *empty set,* containing no elements;
- $\mathbb{N}$: the set of *natural numbers* (that is, nonnegative integers);
- $\mathbb{Z}$: the set of *integers;*
- $\mathbb{Q}$: the set of *rational numbers;*
- $\mathbb{R}$: the set of *real numbers;*
- $\mathbb{C}$: the set of *complex numbers.*

Also, the term *singleton* is used to refer to any set consisting of precisely one element. Thus $\{1\}$, $\{2\}$, $\{3\}$ are different sets, but they are all *singletons.*

Here are a few useful symbols (called *quantifiers*):

- $\exists$ means *there exists...* (the *existential* quantifier);

[1] $:=$ is a notation often used to mean that the symbol on the left-hand side is defined by whatever is on the right-hand side. Logically, this is just expressing the equality of the two sides and could just as well be written '='; the extra : is a psychologically convenient decoration inherited from computer science.

[2] *Ordered* lists are denoted with round parentheses: $(1,2,3)$ is not the same as $(1,3,2)$.

[3] But note that there exist pathologies such as *Russell's paradox,* showing that even this style of definitions can lead to nonsense. All is well so long as S is indeed known to be a *set* to begin with.

- $\forall$ means *for all...* (the *universal* quantifier).

Also, $\exists!$ is used to mean *there exists a unique...*

For example, the set of even integers may be written as

$$E = \{a \in \mathbb{Z} \,|\, (\exists n \in \mathbb{Z})\, a = 2n\} :$$

in words, "all integers a such that there exists an integer n for which $a = 2n$". In this case we could replace $\exists$ by $\exists!$ without changing the set—but that has to do with properties of $\mathbb{Z}$, not with mathematical syntax. Also, it is common to adopt the shorthand

$$E = \{2n \,|\, n \in \mathbb{Z}\},$$

in which the existential quantifier is understood.

Being able to parse such strings of symbols effortlessly, and being able to write them out fluently, is extremely important. The reader of this book is assumed to have already acquired this skill.

Note that the *order* in which things are written may make a big difference. For example, the statement

$$(\forall a \in \mathbb{Z})\, (\exists b \in \mathbb{Z}) \quad b = 2a$$

is *true:* it says that the result of doubling an arbitrary integer yields an integer; but

$$(\exists b \in \mathbb{Z})\, (\forall a \in \mathbb{Z}) \quad b = 2a$$

is *false:* it says that there exists a fixed integer b which is 'simultaneously' twice as much as *every* integer—there is no such thing.

Note also that writing simply

$$b = 2a$$

by itself does not convey enough information, unless the context makes it completely clear what quantifiers are attached to a and b: indeed, as we have just seen, different quantifiers may make this into a true or a false statement.

1.2. Inclusion of sets. As mentioned above, two sets are equal if and only if they contain the same elements. We say that a set S is a *subset* of a set T if every element of S is an element of T, in symbols,

$$S \subseteq T.$$

By convention, $S \subset T$ means the same thing: that is (unlike $<$ vs. $\leq$), it does *not* exclude the possibility that S and T may be equal. To avoid any confusion, we will consistently use $\subseteq$ in this book. One adopts $S \subsetneq T$ to mean that S is 'properly' contained in T: that is, $S \subseteq T$ *and* $S \neq T$.

We can think of 'inclusion of sets' in terms of logic: $S \subseteq T$ means that

$$s \in S \implies s \in T$$

(the quantifier $\forall s$ is understood); that is, 'if s is an element of S, then s is an element of T'; that is, all elements of S are elements of T; that is, $S \subseteq T$ as promised.

Note that for all sets S, $\emptyset \subseteq S$ and $S \subseteq S$.

If $S \subseteq T$ and $T \subseteq S$, then $S = T$.

The symbol $|S|$ denotes the *number of elements* of S, if this number is finite; otherwise, one writes $|S| = \infty$. If S and T are finite, then

$$S \subseteq T \implies |S| \leq |T|.$$

The subsets of a set S form a set, called the *power set,* or the *set of parts* of S. For example, the power set of the empty set $\emptyset$ consists of one element: $\{\emptyset\}$. The power set of S is denoted $\mathscr{P}(S)$; a popular alternative is 2^S, and indeed $|\mathscr{P}(S)| = 2^{|S|}$ if S is finite (cf. Exercise 2.11).

1.3. Operations between sets. Once we have a few sets to play with, we can obtain more by applying certain standard operations. Here are a few:

- $\cup$: the *union;*
- $\cap$: the *intersection;*
- $\smallsetminus$: the *difference;*
- $\amalg$: the *disjoint union;*
- $\times$: the (Cartesian) *product;*
- and the important notion of 'quotient by an equivalence relation'.

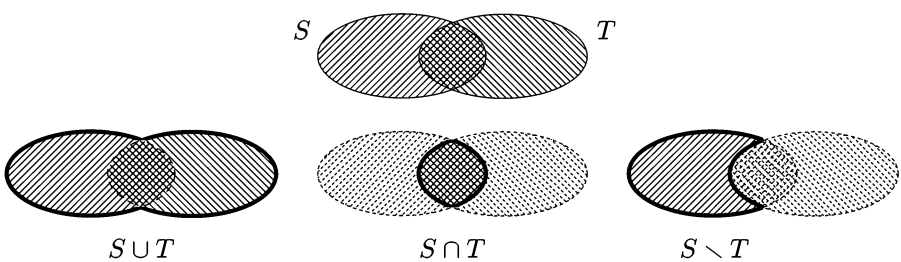
Most of these operations should be familiar to the reader: for example,

$$\{1,2,4\} \cup \{3,4,5\} = \{1,2,3,4,5\}$$

while

$$\{1,2,4\} \smallsetminus \{3,4,5\} = \{1,2\}.$$

In terms of *Venn diagrams* of infamous 'new math' memory:



(the solid black contour indicates the set included in the operation).

Several of these operations may be written out in a transparent way in terms of logic: thus, for example,

$$s \in S \cap T \iff (s \in S \text{ and } s \in T).$$

Two sets S and T are *disjoint* if $S \cap T = \emptyset$, that is, if *no* element is 'simultaneously' in both of them.

The *complement* of a subset T in a set S is the difference set $S \smallsetminus T$ consisting of all elements of S which are *not* in T. Thus, for example, the complement of the set of even integers in $\mathbb{Z}$ is the set of odd integers.

The operations $\amalg$, $\times$, and quotients by equivalence relations are slightly more mysterious, and it is very instructive to contemplate them carefully. We will look

at them in a particularly naive way first and come back to them in a short while when we have acquired more language and can view them from a more sophisticated viewpoint.

1.4. Disjoint unions, products. One problem with these operations is that their output may not be defined *as a set,* but rather as a set *up to isomorphisms of sets,* that is, up to bijections. To make sense out of this, we have to talk about *functions,* and we will do that in a moment.

Roughly speaking, the *disjoint union* of two sets S and T is a set $S \amalg T$ obtained by first producing 'copies' S' and T' of the sets S and T, with the property that $S' \cap T' = \emptyset$, and then taking the (ordinary) union of S' and T'. The careful reader will feel uneasy, since this 'recipe' does not *define* one set: whatever it means to produce a 'copy' of a set, surely there are many ways to do so. This ambiguity will be clarified below.

Nevertheless, note that we can say something about $S \amalg T$ even on these very shaky grounds: for example, if S consists of 3 elements and T consists of 4 elements, the reader should expect (correctly) that $S \amalg T$ consists of 7 elements.

Products are marred by the same kind of ambiguity, but fortunately there is a convenient convention that allows us to write down 'one' set representing the product of two sets S and T: given S and T, we let $S \times T$ be the set whose elements are the *ordered pairs*[4] (s,t) of elements of S and T:

$$S \times T := \{(s,t) \text{ such that } s \in S,\, t \in T\}.$$

Thus, if $S = \{1,2,3\}$ and $T = \{3,4\}$, then

$$S \times T = \{(1,3),(1,4),(2,3),(2,4),(3,3),(3,4)\}.$$

For a more sophisticated example, $\mathbb{R} \times \mathbb{R}$ is the set of pairs of real numbers, which (as we learn in calculus) is a good way to represent a *plane.* The set $\mathbb{Z} \times \mathbb{Z}$ could be represented by considering the points in this plane that happen to have integer coordinates. Incidentally, it is common to denote these sets $\mathbb{R}^2$, $\mathbb{Z}^2$; and similarly, the product $A \times A$ of a set by itself is often denoted A^2.

If S and T are finite sets, clearly $|S \times T| = |S|\,|T|$.

Also note that we can use products to obtain explicit 'copies' of sets as needed for the disjoint union: for example, we could let $S' = \{0\} \times S$, $T' = \{1\} \times T$, guaranteeing that S' and T' are disjoint (why?); and there is an evident way to 'identify' S and S', T and T'. Again, making this precise requires a little more vocabulary.

The operations $\cup$, $\cap$, $\amalg$, $\times$ extend to operations on whole 'families' of sets: for example, if $S_1, \ldots, S_n$ are sets, we write

$$\bigcap_{i=1}^{n} S_i = S_1 \cap S_2 \cap \cdots \cap S_n$$

[4]One can define the ordered pair (s,t) *as a set* by setting $(s,t) = \{s, \{s,t\}\}$: this carries the information of the elements s, t, as well as conveying the fact that s is special (= the first element of the pair).

for the set whose elements are those elements which are simultaneously elements of all sets $S_1, \ldots, S_n$; and similarly for the other operations. But note that while it is clear from the definitions that, for example,

$$S_1 \cup S_2 \cup S_3 = (S_1 \cup S_2) \cup S_3 = S_1 \cup (S_2 \cup S_3),$$

it is not so clear in what sense the sets

$$S_1 \times S_2 \times S_3, \quad (S_1 \times S_2) \times S_3, \quad S_1 \times (S_2 \times S_3)$$

should be 'identified' (where we can define the leftmost set as the set of 'ordered triples' of elements of S_1, S_2, S_3, by analogy with the definition for two sets). In fact, again, we can really make sense of such statements only after we acquire the language of functions. However, all such statements do turn out to be true, as the reader probably expects; by virtue of this fortunate circumstance, we can be somewhat cavalier and gloss over such subtleties.

More generally, if $\mathscr{S}$ is a *set of sets*, we may consider sets

$$\bigcup_{S \in \mathscr{S}} S, \quad \bigcap_{S \in \mathscr{S}} S, \quad \coprod_{S \in \mathscr{S}} S, \quad \prod_{S \in \mathscr{S}} S,$$

for the union, intersection, disjoint union, product of all sets in $\mathscr{S}$. There are important subtleties concerning these definitions: for example, if all $S \in \mathscr{S}$ are nonempty, does it follow that $\prod_{S \in \mathscr{S}} S$ is nonempty? The reader probably thinks so, but (if $\mathscr{S}$ is infinite) this is a rather thorny issue, amounting to the *axiom of choice*.

By and large, such subtleties do not affect the material in this course; we will partly come to terms with them in due time[5], when they become more relevant to the issues at hand (cf. §V.3).

1.5. Equivalence relations, partitions, quotients. Intuitively, a *relation* on elements of a set S is some special affinity among selections of elements of S. For example, the relation $<$ on the set $\mathbb{Z}$ is a way to compare the size of two integers: since $2 < 5$, 2 'is related to' 5 in this sense, while 5 is not related to 2 in the same sense.

For all practical purposes, what a relation 'means' is completely captured by which elements are related to which elements in the set. We would really know all there is to know about $<$ on $\mathbb{Z}$ if we had a complete list of all pairs (a, b) of integers such that $a < b$. For example, $(2, 5)$ is such a pair, while $(5, 2)$ is not.

This leads to a completely straightforward definition of the notion of relation: a *relation* on a set S is simply a *subset R of the product $S \times S$*. If $(a, b) \in R$, we say that a and b are 'related by R' and write

$$a \, R \, b.$$

Often we use fancier symbols for relations, such as $<$, $\leq$, $=$, $\sim$, ...

[5]The reader will have to employ the axiom of choice in some exercises every now and then, even before we come back to these issues, but this will probably happen below the awareness level, and so it should.

The prototype of a well-behaved relation is '=', which corresponds to the 'diagonal'

$$\{(a,b) \in S \times S \mid a = b\} = \{(a,a) \mid a \in S\} \subseteq S \times S.$$

Three properties of this very special relation turn out to be particularly important: if $\sim$ denotes for a moment the relation $=$ of equality, then $\sim$ satisfies

- *reflexivity:* $(\forall a \in S)\ a \sim a$;
- *symmetry:* $(\forall a \in S)\ (\forall b \in S)\ a \sim b \implies b \sim a$;
- *transitivity:* $(\forall a \in S)\ (\forall b \in S)\ (\forall c \in S)$, $(a \sim b$ and $b \sim c) \implies a \sim c$.

That is, every a is equal to itself; if a is equal to b, then b is equal to a; etc.

Definition 1.1. An *equivalence relation* on a set S is any relation $\sim$ satisfying these three properties. ⌟

In terms of the corresponding subset R of $S \times S$, 'reflexivity' says that the diagonal is contained in R; 'symmetry' says that R is unchanged if flipped about the diagonal (that is, if every (a,b) is interchanged with (b,a)); while unfortunately 'transitivity' does not have a similarly nice pictorial translation.

The datum of an equivalence relation on S turns out to be equivalent to a type of information which looks a little different at first, that is, a *partition* of S. A partition of S is a family of *disjoint* nonempty subsets of S, *whose union is S:* for example,

$$\mathscr{P} = \{\{1,4,7\},\{2,5,8\},\{3,6\},\{9\}\}$$

is a partition of the set

$$\{1,2,3,4,5,6,7,8,9\}.$$

Here is how to get a partition of S from a relation $\sim$ on S: for every element $a \in S$, the *equivalence class* of a (w.r.t. $\sim$) is the subset of S defined by

$$[a]_\sim := \{b \in S \mid b \sim a\};$$

then the equivalence classes form a partition $\mathscr{P}_\sim$ of S (Exercise 1.2).

Conversely (Exercise 1.3) every partition $\mathscr{P}$ is the partition corresponding in this fashion to an equivalence relation. Therefore, the notions of 'equivalence relation on S' and 'partition of S' are really equivalent.

Now we can view $\mathscr{P}_\sim$ as a *set* (whose elements are the equivalence classes with respect to $\sim$). This is the quotient operation mentioned in §1.3.

Definition 1.2. The *quotient* of the set S *with respect to* the equivalence relation $\sim$ is the set

$$S/\sim := \mathscr{P}_\sim$$

of equivalence classes of elements of S with respect to $\sim$. ⌟

Example 1.3. Take $S = \mathbb{Z}$, and let $\sim$ be the relation defined by

$$a \sim b \iff a - b \text{ is even.}$$

Then $\mathbb{Z}/\sim$ consists of two equivalence classes:

$$\mathbb{Z}/\sim = \{[0]_\sim, [1]_\sim\}.$$

Indeed, every integer b is either even (and hence $b-0$ is even, so $b \sim 0$, and $b \in [0]_\sim$) or odd (and hence $b-1$ is even, so $b \sim 1$, and $b \in [1]_\sim$). This is of course the starting point of *modular arithmetic,* which we will cover in due detail later on (§II.2.3). ⌟

One way to think about this operation is that the equivalence relation 'becomes equality in the quotient': that is, two elements of the quotient $S/\sim$ are equal if and only if the corresponding elements in S are related by $\sim$. In other words, taking a quotient is a way to turn any equivalence relation into an equality. This observation will be further formalized in 'categorical terms' in a short while (§5.3).

Exercises

Exercises marked with a ▷ are referred to from the text; exercises marked with a ¬ are referred to from other exercises. These referring exercises and sections are listed in brackets following the current exercise; see the introduction for further clarifications, if necessary.

1.1. Locate a discussion of Russell's paradox, and understand it.

1.2. ▷ Prove that if $\sim$ is a relation on a set S, then the corresponding family $\mathscr{P}_\sim$ defined in §1.5 is indeed a partition of S: that is, its elements are nonempty, disjoint, and their union is S. [§1.5]

1.3. ▷ Given a partition $\mathscr{P}$ on a set S, show how to define a relation $\sim$ on S such that $\mathscr{P}$ is the corresponding partition. [§1.5]

1.4. How many different equivalence relations may be defined on the set $\{1, 2, 3\}$?

1.5. Give an example of a relation that is reflexive and symmetric but not transitive. What happens if you attempt to use this relation to define a partition on the set? (Hint: Thinking about the second question will help you answer the first one.)

1.6. ▷ Define a relation $\sim$ on the set $\mathbb{R}$ of real numbers by setting $a \sim b \iff b-a \in \mathbb{Z}$. Prove that this is an equivalence relation, and find a 'compelling' description for $\mathbb{R}/\sim$. Do the same for the relation $\approx$ on the plane $\mathbb{R} \times \mathbb{R}$ defined by declaring $(a_1, a_2) \approx (b_1, b_2) \iff b_1 - a_1 \in \mathbb{Z}$ and $b_2 - a_2 \in \mathbb{Z}$. [§II.8.1, II.8.10]

2. Functions between sets

2.1. Definition. A common thread we will follow for just about every structure introduced in this book will be to try to understand both the type of structures *and* the ways in which different instances of a given structure may interact.

Sets interact with each other through *functions.* It is tempting to think of a function f from a set A to a set B in 'dynamic' terms, as a way to 'go from A to B'. Similarly to the business with relations, it is straightforward to formalize this notion in ways that do not need to invoke any deep 'meaning' of any given f: everything that can be known about a function f is captured by the information of

which element b of B is the image of any given element a of A. This information is nothing but a subset of $A \times B$:

$$\Gamma_f := \{(a,b) \in A \times B \mid b = f(a)\} \subseteq A \times B.$$

This set Γ_f is the *graph* of f; officially, a function really 'is' its graph[6].

Not all subsets $\Gamma \subseteq A \times B$ correspond to ('are') functions: we need to put one requirement on the graphs of *functions*, which can be expressed as follows:

$$(\forall a \in A)\,(\exists! b \in B) \quad (a,b) \in \Gamma_f,$$

or ('in functional notation')

$$(\forall a \in A)\,(\exists! b \in B) \quad f(a) = b.$$

That is, a function must send each element a of A to exactly one element of B, depending on a. 'Multivalued functions' such as $\pm\sqrt{x}$ (which are very important in, e.g., the study of Riemann surfaces) are *not* functions in this sense.

To announce that f is a function from a set A to a set B, one writes $f : A \to B$ or draws the following picture ('diagram'):

$$A \xrightarrow{\ f\ } B\ .$$

The action of a function $f : A \to B$ on an element $a \in A$ is sometime indicated by a 'decorated' arrow, as in

$$a \mapsto f(a).$$

The collection of all functions from a set A to a set B is itself a set[7], denoted B^A. If we take seriously the notion that a function is really the same thing as its graph, then we can view B^A as a (special) subset of the power set of $A \times B$.

Every set A comes equipped with a very special function, whose graph is the diagonal in $A \times A$: the *identity function* on A

$$\mathrm{id}_A : A \to A$$

defined by $(\forall a \in A)\ \mathrm{id}_A(a) = a$. More generally, the inclusion of any subset S of a set A determines a function $S \to A$, simply sending every element s of S to 'itself' in A.

If S is a subset of A, we denote by $f(S)$ the subset of B defined by

$$f(S) := \{b \in B \mid (\exists a \in A)\, b = f(a)\}.$$

That is, $f(S)$ is the subset of B consisting of all elements that are images of elements of S by the function f. The largest such subset, that is, $f(A)$, is called the *image of f*, denoted 'im f'.

Also, $f|_S$ denotes the 'restriction' of f to the subset S: this is the function $S \to B$ defined by

$$(\forall s \in S): \quad f|_S(s) = f(s).$$

---

[6] To be precise, it is the graph Γ_f *together with* the information of the source A and the target B of f. These are part of the data of the function.

[7] This is another 'operation among sets', not listed in §1.3. Can you see why we use B^A for this set? (Cf. Exercise 2.10.)

That is, $f|_S$ is the composition (in the sense explained in the next subsection) $f \circ i$, where $i : S \to A$ is the inclusion. Note that $f(S) = \mathrm{im}(f|_S)$.

2.2. Examples: Multisets, indexed sets. The 'multisets' mentioned briefly in §1.1 are a simple example of a notion easily formalized by means of functions. A multiset may be defined by giving a function from a (regular) set A to the set $\mathbb{N}^*$ of positive[8] integers; if $m : A \to \mathbb{N}^*$ is such a function, the corresponding multiset consists of the elements $a \in A$, each taken $m(a)$ times. Thus, the *multiset* $\{a, a, a, b, b, b, b, b, c\}$ is really the function $m : \{a, b, c\} \to \mathbb{N}^*$ for which $m(a) = 3$, $m(b) = 5$, $m(c) = 1$. As with ordinary sets, the order in which the elements are listed is not part of the information carried by a multiset. Simple set-theoretic notions such as inclusion, union, etc., extend in a straightforward way to multisets. For another viewpoint on multisets, see Exercise 3.9.

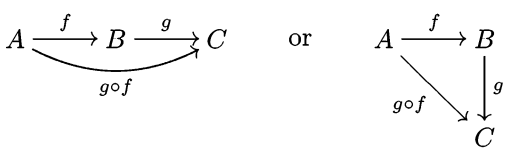
Another example is given by the use of 'indices'. If we write *let* $a_1, \ldots, a_n$ *be integers...*, we really mean *consider a function* $a : \{1, \ldots, n\} \to \mathbb{Z}...$, with the understanding that a_i is shorthand for the value $a(i)$ (for $i = 1, \ldots, n$). It is tempting to think of an *indexed set* $\{a_i\}_{i \in I}$ simply as a set whose elements happen to be denoted a_i, for i ranging over some 'set of indices' I; but such an indexed set is more properly a *function* $I \to A$, where A is some set from which we draw the elements a_i. For example, this allows us to consider a_0 and a_1 as distinct elements of $\{a_i\}_{i \in \mathbb{N}}$, even if by coincidence $a_0 = a_1$ as elements of the target set A.

It is easy to miss such subtleties, and some abuse of notation is common and usually harmless. These distinctions play a role in (for example) discussions of *linear independence* of sets of vectors; cf. §VI.1.2.

2.3. Composition of functions. Functions may be *composed:* if $f : A \to B$ and $g : B \to C$ are functions, then so is the operation $g \circ f$ defined by

$$(*) \qquad (\forall a \in A) \quad (g \circ f)(a) := g(f(a)) :$$

that is, we use f to go from A to B, then apply g to reach C. Graphically we may draw pictures such as

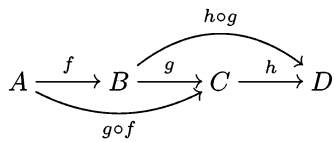


Such graphical representations of collections of (for example) sets connected by functions are called *diagrams*. We will draw many diagrams, and in contexts substantially more general than the one at hand right now.

We say that the diagrams drawn above 'commute', or 'are commutative', meaning that if we start from A and travel to C in either of the two possible ways prescribed by the diagram, the result of applying the functions one encounters is the same. This is precisely the content of the statement (*).

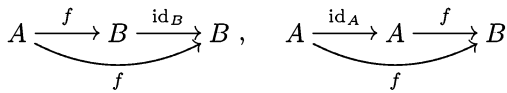
[8] Some references allow 0 as a possible multiplicity.

Composition is associative: that is, if $f : A \to B$, $g : B \to C$, and $h : C \to D$ are functions, then $h \circ (g \circ f) = (h \circ g) \circ f$. Graphically, the diagram



commutes. This important observation should be completely evident from the definition of composition.

The identity function is very special with respect to compositions: if $f : A \to B$ is any function, then $\mathrm{id}_B \circ f = f$ and $f \circ \mathrm{id}_A = f$. Graphically, the diagrams



commute.

2.4. Injections, surjections, bijections. Special kinds of functions deserve highlighting:

- A function $f : A \to B$ is *injective* (or *an injection* or *one-to-one*) if
$$a' \neq a'' \implies f(a') \neq f(a'') :$$
that is, if f sends different elements to different elements[9].
- A function $f : A \to B$ is *surjective* (or *a surjection* or *onto*) if
$$(\forall b \in B)\,(\exists a \in A) \quad b = f(a) :$$
that is, if f 'covers the whole of B'; more precisely, if $\operatorname{im} f = B$.

Injections are often drawn $\hookrightarrow$; surjections are often drawn $\twoheadrightarrow$.

If f is *both* injective and surjective, we say it is *bijective* (or *a bijection* or *a one-to-one correspondence* or *an isomorphism of sets.*) In this case we often write $f : A \xrightarrow{\sim} B$, or
$$A \cong B,$$
and we say that A and B are 'isomorphic' sets.

Of course the identity function $\mathrm{id}_A : A \to A$ is a bijection.

If $A \cong B$, that is, if there is a bijection $f : A \to B$, then the sets A and B may be 'identified' through f, in the sense that we can match precisely the elements a of A with the corresponding elements $f(a)$ of B. For example, if A is a finite set and $A \cong B$, then B is necessarily also a finite set and $|A| = |B|$.

This terminology allows us to make better sense of the considerations on 'disjoint union' given in §1.3: the 'copies' A', B' of the given sets A, B should simply

[9]Often one checks this definition in the contrapositive (hence equivalent) formulation, that is,
$$(\forall a' \in A)\,(\forall a'' \in A) \quad f(a') = f(a'') \implies a' = a''.$$

be *isomorphic* sets to A, B, respectively. The proposal given at the end of §1.4 to produce such disjoint 'copies' works, because (for example) the function

$$f : A \to \{0\} \times A$$

defined by

$$(\forall a \in A) \quad f(a) = (0, a)$$

is manifestly a bijection.

2.5. Injections, surjections, bijections: Second viewpoint. There is an alternative and instructive way to think about these notions.

If $f : A \to B$ is a bijection, then we can 'flip its graph' and define a function

$$g : B \to A :$$

that is, we can let $a = g(b)$ precisely when $b = f(a)$. (The fact that f is both injective and surjective guarantees that the flip of Γ_f is the graph of a *function* according to the definition given in §2.1. Check this!)

This function g has a very interesting property: graphically,

$$A \xrightarrow{f} B \xrightarrow{g} A \ , \qquad B \xrightarrow{g} A \xrightarrow{f} B$$

commute; that is, $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$. The first identity tells us that g is a 'left-inverse'[10] of f; the second tells us that g is a 'right-inverse' of f. We simply say that it is *the inverse of* f, denoted f^{-1}. Thus, 'bijections have inverses'.

What about the converse? If a function has an inverse, is it a bijection? This is true, but in fact we can be much more precise.

Proposition 2.1. *Assume $A \neq \emptyset$, and let $f : A \to B$ be a function. Then*

(1) f has a left-inverse if and only if it is injective.

(2) f has a right-inverse if and only if it is surjective.

Proof. Let's prove (1).

($\Longrightarrow$) If $f : A \to B$ has a left-inverse, then there exists a $g : B \to A$ such that $g \circ f = \mathrm{id}_A$. Now assume that $a' \neq a''$ are arbitrary different elements in A; then

$$g(f(a')) = \mathrm{id}_A(a') = a' \neq a'' = \mathrm{id}_A(a'') = g(f(a''));$$

that is, g sends $f(a')$ and $f(a'')$ to different elements. This forces $f(a')$ and $f(a'')$ to be different, showing that f is injective.

($\Longleftarrow$) Now assume $f : A \to B$ is injective. In order to construct a function $g : B \to A$, we have to assign a unique value $g(b) \in A$ for each element $b \in B$. For this, choose any fixed element $s \in A$ (which we can do because $A \neq \emptyset$); then set

$$g(b) := \begin{cases} a & \text{if } b = f(a) \text{ for some } a \in A, \\ s & \text{if } b \notin \operatorname{im} f. \end{cases}$$

[10] Never mind that g is *drawn* to the right of f in the diagram—we say that g is a *left*-inverse of f because it is *written* to the left of f: $g \circ f = \mathrm{id}_A$.

In words, if b is the image of an element a of A, send it back to a; otherwise, send it to your fixed element s.

The given assignment defines a function, precisely because f is injective: indeed, this guarantees that every b that is the image of some $a \in A$ by f is the image of a *unique* a (two distinct elements of A cannot be simultaneously sent to b by f, since f is injective). Thus every $b \in B$ is sent to a unique well-defined element of A, as is required of functions.

Finally, the function $g : B \to A$ is a left-inverse of f. Indeed, if $a \in A$, then $b = f(a)$ is of the first type, so it is sent back to a by g; that is, $g \circ f(a) = a = \mathrm{id}_A(a)$ for all $a \in A$, as needed.

The proof of (2) is left as an exercise (Exercise 2.2). □

Corollary 2.2. *A function $f : A \to B$ is a bijection if and only if it has a (two-sided) inverse.*

If a function is injective *but not surjective,* then it will not have a right-inverse, and it will necessarily have more than one left-inverse (this should be clear from the argument given in the proof of Proposition 2.1). Similarly, a surjective function will in general have many right-inverses; they are often called *sections.*

Proposition 2.1 hints that something deep is going on here. The definition of injective and surjective maps given in §2.4 relied crucially on working directly with the *elements* of our sets; Proposition 2.1 shows that in fact these properties are detected by the way *functions* are 'organized' among sets. Even if we did not know what 'elements' means, still we could make sense of the notions of injectivity and surjectivity (and hence of isomorphisms of sets) by exclusively referring to properties of functions.

This is a more 'mature' point of view and one that will be championed when we talk about categories. To some extent, it should cure the reader from the discomfort of talking about 'elements', as we did in our informal introduction to sets, without defining what these mysterious entities are supposed to be.

The standard notation for *the* inverse of a bijection f is f^{-1}. This symbol is also used for functions that are not bijections, but in a slightly different context: if $f : A \to B$ is any function and $T \subseteq B$ is a subset of B, then $f^{-1}(T)$ denotes the subset of A of 'all elements that map to T'; that is,

$$f^{-1}(T) = \{a \in A \mid f(a) \in T\}.$$

If $T = \{q\}$ consists of a single element of B, $f^{-1}(T)$ (abbreviated $f^{-1}(q)$) is called the *fiber* of f over q. Thus a function $f : A \to B$ is a bijection if it has nonempty fibers over all elements of B (that is, f is surjective), and these fibers are in fact singletons (that is, f is injective). In this case, this notation f^{-1} matches nicely with the notation of 'inverse' mentioned above.

2.6. Monomorphisms and epimorphisms. There is *yet another* way to express injectivity and surjectivity, which appears at first more complicated than what we have seen so far but which is in fact even more basic.

A function $f : A \to B$ is a *monomorphism* (or *monic*) if the following holds:

$$\textit{for all sets } Z \textit{ and all functions } \alpha', \alpha'' : Z \to A$$
$$f \circ \alpha' = f \circ \alpha'' \implies \alpha' = \alpha''.$$

Proposition 2.3. *A function is injective if and only if it is a monomorphism.*

Proof. ($\implies$) By Proposition 2.1, if a function $f : A \to B$ is injective, then it has a left-inverse $g : B \to A$. Now assume that α', α'' are arbitrary functions from another set Z to A and that

$$f \circ \alpha' = f \circ \alpha'';$$

compose on the left by g, and use associativity of composition:

$$(g \circ f) \circ \alpha' = g \circ (f \circ \alpha') = g \circ (f \circ \alpha'') = (g \circ f) \circ \alpha'';$$

since g is a left-inverse of f, this says

$$\mathrm{id}_A \circ \alpha' = \mathrm{id}_A \circ \alpha'',$$

and therefore

$$\alpha' = \alpha'',$$

as needed to conclude that f is a monomorphism.

($\impliedby$) Now assume that f is a monomorphism. This says something about arbitrary sets Z and arbitrary functions $Z \to A$; we are going to use a microscopic portion of this information, choosing Z to be any singleton $\{p\}$. Then assigning functions $\alpha', \alpha'' : Z \to A$ amounts to choosing to which elements $a' = \alpha'(p)$, $a'' = \alpha''(p)$ we should send the single element p of Z. For this particular choice of Z, the property defining monomorphisms,

$$f \circ \alpha' = f \circ \alpha'' \implies \alpha' = \alpha''$$

becomes

$$f \circ \alpha'(p) = f \circ \alpha''(p) \implies \alpha' = \alpha'';$$

that is

$$f(a') = f(a'') \implies \alpha' = \alpha''.$$

Now two functions from $Z = \{p\}$ to A are equal if and only if they send p to the same element, so this says

$$f(a') = f(a'') \implies a' = a''.$$

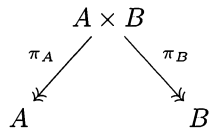
This has to be true for all α', α'', that is, for all choices of distinct a', a'' in A. In other words, f has to be injective, as was to be shown. □

The reader should now expect that there be a definition in the style of the one given for monomorphisms and which will turn out to be equivalent to 'surjective'. This is the case: such a notion is called *epimorphism*. Finding it, and proving the equivalence with the ordinary definition of 'surjective', is left to the reader[11] (Exercise 2.5).

[11] This is a particularly important exercise, and we recommend that the reader write out all the gory details carefully.

2.7. Basic examples. The basic operations on sets provided us with several important examples of injective and surjective functions.

Example 2.4. Let A, B be sets. Then there are *natural projections* π_A, π_B:

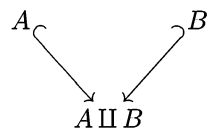


defined by

$$\pi_A((a,b)) := a, \quad \pi_B((a,b)) := b$$

for all $(a,b) \in A \times B$. Both of these maps are (clearly) surjective. ⌟

Example 2.5. Similarly, there are natural injections from A and B to the disjoint union:



obtained by sending $a \in A$ (resp., $b \in B$) to the corresponding element in the isomorphic copy A' of A (resp., B' of B) in $A \amalg B$. ⌟

Example 2.6. If $\sim$ is an equivalence relation on a set A, there is a (clearly surjective) canonical projection

$$A \twoheadrightarrow A/\!\sim$$

obtained by sending every $a \in A$ to its equivalence class $[a]_\sim$. ⌟

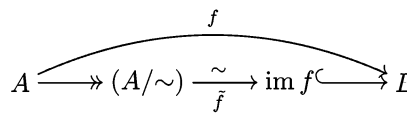
2.8. Canonical decomposition. The reason why we focus our attention on injective and surjective maps is that they provide the basic 'bricks' out of which *any* function may be constructed.

To see this, we observe that every function $f : A \to B$ determines an equivalence relation $\sim$ on A as follows: for all $a', a'' \in A$,

$$a' \sim a'' \iff f(a') = f(a'').$$

(The reader should check that this is indeed an *equivalence* relation.)

Theorem 2.7. *Let $f : A \to B$ be any function, and define $\sim$ as above. Then f decomposes as follows:*



where the first function is the canonical projection $A \to A/\!\sim$ (as in Example 2.6), the third function is the inclusion $\operatorname{im} f \subseteq B$, and the bijection $\tilde{f}$ in the middle is defined by

$$\tilde{f}([a]_\sim) := f(a)$$

for all $a \in A$.

The formula defining $\tilde{f}$ shows immediately that the diagram commutes; so all we have to verify in order to prove this theorem is that

- that formula *does* define a function;
- that function is in fact a bijection.

The first item is an instance of a class of verifications of the utmost importance. The formula given for $\tilde{f}$ has a colossal built-in ambiguity: *the same* element in $A/\sim$ may be the equivalence class of *many* elements of A; applying the formula for $\tilde{f}$ requires *choosing* one of these elements and applying f to it. We have to prove that the result of this operation is *independent* from this choice: that is, that all possible choices of representatives for that equivalence class lead to the same result.

We encode this type of situation by saying that we have to verify that $\tilde{f}$ is *well-defined.* We will often have to check that the operations we consider are well-defined, in contexts very similar to the one epitomized here.

Proof. Spelling out the first item discussed above, we have to verify that, for all a', a'' in A,

$$[a']_\sim = [a'']_\sim \implies f(a') = f(a'').$$

Now $[a']_\sim = [a'']_\sim$ means that $a' \sim a''$, and the definition of $\sim$ has been engineered precisely so that this would mean $f(a') = f(a'')$ as required here. So $\tilde{f}$ is indeed well-defined.

To verify the second item, that is, that $\tilde{f} : A/\sim \to \operatorname{im} f$ is a bijection, we check explicitly that $\tilde{f}$ is injective and surjective.

Injective: If $\tilde{f}([a']_\sim) = \tilde{f}([a'']_\sim)$, then $f(a') = f(a'')$ by definition of $\tilde{f}$; hence $a' \sim a''$ by definition of $\sim$, and then $[a']_\sim = [a'']_\sim$. Therefore

$$\tilde{f}([a']_\sim) = f([a'']_\sim) \implies [a']_\sim = [a'']_\sim$$

proving injectivity.

Surjective: Given any $b \in \operatorname{im} f$, there is an element $a \in A$ such that $f(a) = b$. Then

$$\tilde{f}([a]_\sim) = f(a) = b$$

by definition of $\tilde{f}$. Since b was arbitrary in $\operatorname{im} f$, this shows that $\tilde{f}$ is surjective, as needed. □

Theorem 2.7 shows that *every* function is the composition of a surjection, followed by an isomorphism, followed by an injection. While its proof is trivial, this is a result of some importance, since it is the prototype of a situation that will occur several times in this book. It will resurface every now and then, with names such as 'the first isomorphism theorem'.

2.9. Clarification. Finally, we can begin to clarify one comment about *disjoint unions*, products, and quotients, made in §1.4. Our definition of $A \amalg B$ was the (conventional) union of two disjoint sets A', B' isomorphic to A, B, respectively. It is easy to provide a way to effectively produce such isomorphic copies (as we did in §1.4); but it is in fact a little too easy—many other choices are possible, and one does not look any better than any other. It is in fact more sensible *not* to make a

fixed choice once and for all and simply accept the fact that all of them produce acceptable candidates for $A \amalg B$. From this egalitarian standpoint, the result of the operation $A \amalg B$ is *not* 'well-defined' *as a set* in the sense specified above. However, it is easy to see (Exercise 2.9) that $A \amalg B$ *is* well-defined *up to isomorphism:* that is, that any two choices for the copies A', B' lead to *isomorphic* candidates for $A \amalg B$. The same considerations apply to products and quotients.

The main feature of sets obtained by taking disjoint unions, products, or quotients is not really 'what elements they contain' but rather 'their relationship with all other sets'. This will be (even) clearer when we revisit these operations and others[12] in the context of categories.

Exercises

2.1. ▷ How many different bijections are there between a set S with n elements and itself? [§II.2.1]

2.2. ▷ Prove statement (2) in Proposition 2.1. You may assume that given a family of disjoint subsets of a set, there is a way to choose one element in each member of the family[13]. [§2.5, V.3.3]

2.3. Prove that the inverse of a bijection is a bijection and that the composition of two bijections is a bijection.

2.4. ▷ Prove that 'isomorphism' is an equivalence relation (on any set of sets). [§4.1]

2.5. ▷ Formulate a notion of *epimorphism,* in the style of the notion of *monomorphism* seen in §2.6, and prove a result analogous to Proposition 2.3, for epimorphisms and surjections. [§2.6, §4.2]

2.6. With notation as in Example 2.4, explain how any function $f : A \to B$ determines a section of π_A.

2.7. Let $f : A \to B$ be any function. Prove that the graph Γ_f of f is isomorphic to A.

2.8. Describe as explicitly as you can all terms in the canonical decomposition (cf. §2.8) of the function $\mathbb{R} \to \mathbb{C}$ defined by $r \mapsto e^{2\pi i r}$. (This exercise matches one assigned previously. Which one?)

2.9. ▷ Show that if $A' \cong A''$ and $B' \cong B''$, and further $A' \cap B' = \emptyset$ and $A'' \cap B'' = \emptyset$, then $A' \cup B' \cong A'' \cup B''$. Conclude that the operation $A \amalg B$ (as described in §1.4) is well-defined *up to isomorphism* (cf. §2.9). [§2.9, 5.7]

2.10. ▷ Show that if A and B are finite sets, then $|B^A| = |B|^{|A|}$. [§2.1, 2.11, §II.4.1]

[12]The reader should also be aware that there are important variations on the operations we have seen so far—particularly important are the *fibered* flavors of products and disjoint unions.

[13]This (reasonable) statement is the *axiom of choice;* cf. §V.3.

2.11. ▷ In view of Exercise 2.10, it is not unreasonable to use 2^A to denote the set of functions from an arbitrary set A to a set with 2 elements (say $\{0,1\}$). Prove that there is a bijection between 2^A and the *power set* of A (cf. §1.2). [§1.2, III.2.3]

3. Categories

The language of categories is affectionately known as *abstract nonsense,* so named by Norman Steenrod. This term is essentially accurate and not necessarily derogatory: categories refer to *nonsense* in the sense that they are all about the 'structure', and not about the 'meaning', of what they represent. The emphasis is less on how you run into a specific set you are looking at and more on how that set may sit in relationship with all other sets. Worse (or better) still, the emphasis is less on studying sets, and functions between sets, than on studying 'things, and things that go from things to things' without necessarily being explicit about what these things are: they may be sets, or groups, or rings, or vector spaces, or modules, or other objects that are so exotic that the reader has no right whatsoever to know about them (yet).

'Categories' will intuitively look like sets at first, and in multiple ways. Categories may make you think of sets, in that they are 'collections of objects', and further there will be notions of 'functions from categories to categories' (called *functors*[14]). At the same time, every category may make you think of the collection of all sets, since there will be analogs of 'functions' among the things it contains.

3.1. Definition. The definition of a category looks complicated at first, but the gist of it may be summarized quickly: a category consists of a collection of 'objects', and of 'morphisms' between these objects, satisfying a list of natural conditions.

The reader will note that I refrained from writing a *set* of objects, opting for the more generic 'collection'. This is an annoying, but unavoidable, difficulty: for example, we want to have a 'category of sets', in which the 'objects' are sets and the 'morphisms' are functions between sets, and the problem is that there simply is not a *set of all sets*[15]. In a sense, the collection of all sets is 'too big' to be a set. There are however ways to deal with such 'collections', and the technical name for them is *class*. There *is* a 'class' of all sets (and there will be classes taking care of groups, rings, etc.).

An alternative would be to define a large enough set (called a *universe*) and then agree that all objects of all categories will be chosen from this gigantic entity.

In any case, all the reader needs to know about this is that there is a way to make it work. We will use the term 'class' in the definition, but this will not affect any proof or any other definition in this book. Further, in some of the examples considered below the class in question *is* a set (we say that the category is *small* in this case), so the reader will feel perfectly at home when contemplating these examples.

[14]However, we will not consider functors until later chapters: our first formal encounter with functors will be in Chapter VIII.

[15]That is one thing we learn from Russell's paradox.

Definition 3.1. A *category* C consists of

- a class $\mathrm{Obj}(\mathsf{C})$ of *objects* of the category; and
- for every two objects A, B of C, a set $\mathrm{Hom}_{\mathsf{C}}(A, B)$ of *morphisms,* with the properties listed below. ⌟

As a prototype to keep in mind, think of the objects as 'sets' and of morphisms as 'functions'. This one example should make the defining properties of morphisms look natural and easy to remember:

- For every object A of C, there exists (at least) one morphism $1_A \in \mathrm{Hom}_{\mathsf{C}}(A, A)$, the 'identity' on A.
- One can compose morphisms: two morphisms $f \in \mathrm{Hom}_{\mathsf{C}}(A, B)$ and $g \in \mathrm{Hom}_{\mathsf{C}}(B, C)$ determine a morphism $gf \in \mathrm{Hom}_{\mathsf{C}}(A, C)$. That is, for every triple of objects A, B, C of C there is a function (of sets)
$$\mathrm{Hom}_{\mathsf{C}}(A, B) \times \mathrm{Hom}_{\mathsf{C}}(B, C) \to \mathrm{Hom}_{\mathsf{C}}(A, C),$$
and the image of the pair (f, g) is denoted gf.
- This 'composition law' is associative: if $f \in \mathrm{Hom}_{\mathsf{C}}(A, B)$, $g \in \mathrm{Hom}_{\mathsf{C}}(B, C)$, and $h \in \mathrm{Hom}_{\mathsf{C}}(C, D)$, then
$$(hg)f = h(gf).$$
- The identity morphisms are identities with respect to composition: that is, for all $f \in \mathrm{Hom}_{\mathsf{C}}(A, B)$ we have
$$f1_A = f, \quad 1_B f = f.$$

This is really a mouthful, but again, to remember all this, just think of functions of sets. One further requirement is that the sets
$$\mathrm{Hom}_{\mathsf{C}}(A, B), \quad \mathrm{Hom}_{\mathsf{C}}(C, D)$$
be *disjoint* unless $A = C$, $B = D$; this is something you do not usually think about, but again it holds for ordinary set-functions[16]. That is, if two functions are one and the same, then necessarily they have the same source and the same target: source and target are part of the datum of a set-function.

A morphism of an object A of a category C to itself is called an *endomorphism;* $\mathrm{Hom}_{\mathsf{C}}(A, A)$ is denoted $\mathrm{End}_{\mathsf{C}}(A)$. One of the axioms of a category tells us that this is a 'pointed' set, as $1_A \in \mathrm{End}_{\mathsf{C}}(A)$. The reader should note that composition defines an 'operation' on $\mathrm{End}_{\mathsf{C}}(A)$: if f, g are elements of $\mathrm{End}_{\mathsf{C}}(A)$, so is their composition gf.

Writing '$f \in \mathrm{Hom}_{\mathsf{C}}(A, B)$' gets tiresome in the long run. If the category is understood, one may safely drop the index C, or even use arrows as we do with set-functions: $f : A \to B$. This also allows us to draw *diagrams* of morphisms in any category; a diagram is said to 'commute' (or to be a 'commutative' diagram) if all ways to traverse it lead to the same results of composing morphisms along the way, just as explained for diagrams of functions of sets in §2.3.

[16] We will often use the term 'set-function' to emphasize that we are dealing with a function in the context of sets.

In fact, we will now feel free to use *diagrams* as possible objects of categories. The official definition of a diagram in this context would be a set of objects of a category C, along with prescribed morphisms between these objects; the diagram commutes if it does in the sense specified above. The specifics of the visual representation of a diagram are of course irrelevant.

3.2. Examples. The reader should note that 90% of the definition of the notion of category goes into explaining the properties of its *morphisms;* it is fair to say that the morphisms *are* the important constituents of a category. Nevertheless, it is psychologically irresistible to think of a category in terms of its *objects:* for example, one talks about the 'category of sets'. The point is that usually the kind of 'morphisms' one may consider are (psychologically at least) determined by the objects: if one is talking about sets, what can one possibly mean for 'morphism' other than a function of sets? In other situations (cf. Example 3.5 below or Exercise 3.9) it is a little less clear what the morphisms should be, and looking for the 'right' notion may be an interesting project.

Example 3.2. It is hopefully crystal clear by now that sets (as objects), together with set-functions (as morphisms), form a category; if not, the reader must stop here and go no further until this assertion sheds any residual mystery[17].

There is no universally accepted, official notation for this important category. It is customary to write the word 'Set' or 'Sets', with some fancy decoration for emphasis. For example, in the literature one may encounter SET, **Sets**, $\mathfrak{Set}$, $(Sets)$, and many amusing variations on these themes. We will use 'sans-serif' fonts to denote categories; thus, Set will denote the category of sets. Thus

- $\mathrm{Obj}(\mathsf{Set})$ = the class of all sets;
- for A, B in $\mathrm{Obj}(\mathsf{Set})$ (that is, for A, B sets) $\mathrm{Hom}_{\mathsf{Set}}(A, B) = B^A$.

Note that the presence of the operations recalled in §§1.3–1.5 is *not* part of the definition of category: these operations highlight interesting features of Set, which may or may not be shared by other categories. We will soon come back to some of these operations and understand more precisely what they say about Set. ⌟

Example 3.3. Here is a completely different example.

Suppose S is a set and $\sim$ is a relation on S satisfying the reflexive and transitive properties. Then we can encode this data into a category:

- *objects:* the elements of S;
- *morphisms:* if a, b are objects (that is, if $a, b \in S$), then let $\mathrm{Hom}(a, b)$ be the set consisting of the element $(a, b) \in S \times S$ if $a \sim b$, and let $\mathrm{Hom}(a, b) = \emptyset$ otherwise.

[17] We will give the reader such prompts every now and then: at key times, it is more useful to take stock of what one knows than blindly march forward hoping for the best. A difficulty at this time signals the need to reread the previous material carefully. If the mystery persists, that's what office hours are there for. But typically you should be able to find your way out on your own, based on the information we have given you, and you will most likely learn more this way. You should give it your best try before seeking professional help.

Note that (unlike in Set) there are *very few* morphisms: at most one for any pair of objects, and no morphisms at all between 'unrelated' objects.

We have to define 'composition of morphisms' and verify that the conditions specified in §3.1 are satisfied. First of all, do we have 'identities'? If a is an object (that is, if $a \in S$), we need to find an element

$$1_a \in \mathrm{Hom}(a, a).$$

This is precisely why we are assuming that $\sim$ is reflexive: this tells us that $\forall a$, $a \sim a$; that is, $\mathrm{Hom}(a, a)$ consists of the single element (a, a). So we have no choice: we must let

$$1_a = (a, a) \in \mathrm{Hom}(a, a).$$

As for composition, let a, b, c be objects (that is, elements of S) and

$$f \in \mathrm{Hom}(a, b), \quad g \in \mathrm{Hom}(b, c);$$

we have to define a corresponding morphism $gf \in \mathrm{Hom}(a, c)$. Now,

$$f \in \mathrm{Hom}(a, b)$$

tells us that $\mathrm{Hom}(a, b)$ is nonempty, and according to the definition of morphisms in this category that means that $a \sim b$, and f is in fact the element (a, b) of $A \times B$. Similarly, $g \in \mathrm{Hom}(b, c)$ tells us $b \sim c$ and $g = (b, c)$. Now

$$a \sim b \text{ and } b \sim c \implies a \sim c$$

since we are assuming that $\sim$ is transitive. This tells us that $\mathrm{Hom}(a, c)$ consists of the single element (a, c). Thus we again have no choice: we must let

$$gf := (a, c) \in \mathrm{Hom}(A, C).$$

Is this operation associative? If $f \in \mathrm{Hom}(a, b)$, $g \in \mathrm{Hom}(b, c)$, and $h \in \mathrm{Hom}(c, d)$, then necessarily

$$f = (a, b), \quad g = (b, c), \quad h = (c, d)$$

and

$$gf = (a, c), \quad hg = (b, d)$$

and hence

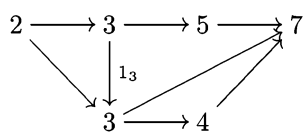
$$h(gf) = (a, d) = (hg)f,$$

proving associativity.

The reader will have no difficulties checking that 1_a is an identity with respect to this composition, as needed (Exercise 3.3).

The most trivial instance of this construction is the category obtained from a set S taken with the equivalence relation '='; that is, the only morphisms are the identity morphisms. These categories are called *discrete.*

As another example, consider the category corresponding to endowing $\mathbb{Z}$ with the relation $\leq$. For example,



is a (randomly chosen) commutative diagram in this category. It would still be a (commutative) diagram in this category if we reversed the vertical arrow $3 \to 3$ or if we added an arrow from 3 to 4, while we are not allowed to draw an arrow *from* 4 *to* 3, since $4 \not\leq 3$.

These categories are very special—for example, every diagram drawn in them is necessarily commutative, and this is *very far* from being the case in, e.g., Set. Also note that these categories are all *small.* ⌟

Example 3.4. Here is another[18] example of a small category.

Let S again be a set. Define a category $\hat{\mathsf{S}}$ by setting

- $\mathrm{Obj}(\hat{\mathsf{S}}) = \mathscr{P}(S)$, the power set S (cf. §1.2 and Exercise 2.11);
- for A, B objects of $\hat{\mathsf{S}}$ (that is, $A \subseteq S$ and $B \subseteq S$) let $\mathrm{Hom}_{\hat{\mathsf{S}}}(A, B)$ be the pair (A, B) if $A \subseteq B$, and let $\mathrm{Hom}_{\hat{\mathsf{S}}}(A, B) = \emptyset$ otherwise.

The identity 1_A consists of the pair (A, A) (which is one, and in fact the only one, morphism from A to A, since $A \subseteq A$). Composition is obtained by stringing inclusions: if there are morphisms

$$A \to B, \quad B \to C$$

in $\hat{\mathsf{S}}$, then $A \subseteq B$ and $B \subseteq C$; hence $A \subseteq C$ and there is a morphism $A \to C$. Checking the axioms specified in §3.1 should be routine (make sure this is the case!).

Examples in this style (but employing more sophisticated structures, such as the family of *open* subsets of a topological space) are hugely important in well-established fields such as algebraic geometry. ⌟

Example 3.5. The next example is very abstract, but thinking about it will make you rather comfortable with everything we have seen so far; and it is a very common construction, variations of which will abound in the course.

Let C be a category, and let A be an object of C. We are going to define a category C_A whose objects are certain *morphisms* in C and whose morphisms are certain *diagrams* of C (surprise!).

- $\mathrm{Obj}(\mathsf{C}_A)$ = all morphisms from any object of C to A; thus, an object of C_A is a morphism $f \in \mathrm{Hom}_{\mathsf{C}}(Z, A)$ for some object Z of C. Pictorially, an object of C_A is an arrow $Z \xrightarrow{f} A$ *in* C; these are often drawn 'top-down', as in

$$\begin{array}{c} Z \\ f \big\downarrow \\ A \end{array}$$

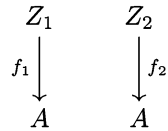
What are morphisms in C_A going to be? There really is only one sensible way to assign morphisms to a category with objects as above. The brave reader will want to stop reading here and continue only after having come up with the definition independently. There will be many similar examples lurking behind constructions

[18] Actually, this is again an instance of the categories considered in Example 3.3. Do you see why? (Exercise 3.5.)

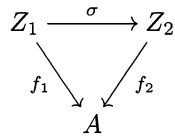
we will encounter in this book, and ideally speaking, they should appear completely *natural* when the time comes. A bit of effort devoted now to understanding this prototype situation will have ample reward in the future. Spoiler follows, so put these notes away *now* and jot down the definition of morphism in C_A.

Welcome back.

- Let f_1, f_2 be objects of C_A, that is, two arrows



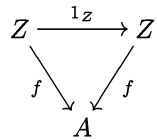
  in C. Morphisms $f_1 \to f_2$ are defined to be *commutative diagrams*



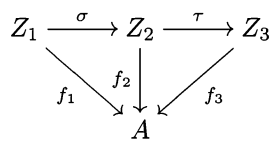
  in the 'ambient' category C.

That is, morphisms $f \to g$ correspond precisely to those morphisms $\sigma : Z_1 \to Z_2$ in C *such that* $f_1 = f_2\sigma$.

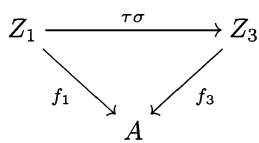
Once you understand what morphisms have to be, checking that they satisfy the axioms spelled out in §3.1 is straightforward. The identities are inherited from the identities in C: for $f : Z \to A$ in C_A, the identity 1_f corresponds to the diagram



which commutes by virtue of the fact that C is a category. Composition is also a subproduct of composition in C. Two morphisms $f_1 \to f_2 \to f_3$ in C_A correspond to putting two commutative diagrams side-by-side:



and then it follows (again because C is a category!) that the diagram obtained by removing the central arrow, i.e.,



also commutes. Check all this(!), and verify that composition in C_A is associative (again, this follows immediately from the fact that composition is associative in C).

Categories constructed in this fashion are called *slice categories* in the literature; they are particular cases of *comma categories.* ⌟

Example 3.6. For the sake of concreteness, let's apply the construction given in Example 3.5 to the category constructed in Example 3.3, say for $S = \mathbb{Z}$ and $\sim$ the relation $\leq$. Call C this category, and choose an object A of C—that is, an integer, for example, $A = 3$. Then the objects of C_A are morphisms in C with target 3, that is, pairs $(n, 3) \in \mathbb{Z} \times \mathbb{Z}$ with $n \leq 3$. There is a morphism

$$(m, 3) \to (n, 3)$$

if and only if $m \leq n$. In this case C_A may be harmlessly identified with the 'subcategory' of integers ≤ 3, with 'the same' morphisms as in C. ⌟

Example 3.7. An entirely similar example to the one explored in Example 3.5 may be obtained by considering morphisms in a category C *from* a fixed object A to all objects in C, again with morphisms defined by suitable commutative diagrams. This leads to *coslice categories.* The reader should provide details of this construction (Exercise 3.7). ⌟

Example 3.8. As a 'concrete' instance of a category as in Example 3.7, let $\mathsf{C} = \mathsf{Set}$ and $A =$ a fixed singleton $\{*\}$. Call the resulting category Set^*.

An object in Set^* is then a morphism $f : \{*\} \to S$ in Set, where S is any set. The information of an object in Set^* consists therefore of the choice of a nonempty set S *and* of an element $s \in S$—that is, the element $f(*)$: this element determines, and is determined by, f.

Thus, we may denote objects of Set^* as pairs (S, s), where S is any set and $s \in S$ is any element of S.

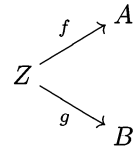
A morphism between two such objects, $(S, s) \to (T, t)$, corresponds then (check this!) to a set-function $\sigma : S \to T$ *such that* $\sigma(s) = t$.

Objects of Set^* are called 'pointed sets'. Many of the structures we will study in this book will be pointed sets. For example (as we will see) a 'group' is a set G with, among other requirements, a distinguished element e_G (its 'identity'); 'group homomorphisms' will be functions which, among other properties, send identities to identities; thus, they are morphisms of pointed sets in the sense considered above. ⌟

Example 3.9. It is useful to contemplate a few more 'abstract' examples in the style of Examples 3.5 and 3.7. These will be essential ingredients in the promised revisitation of some of the operations mentioned in §1.3. Their definition will appear disappointedly simple-minded to the reader who has mastered Examples 3.5 and 3.7.

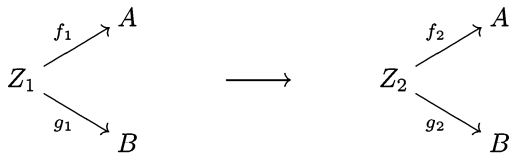
This time we start from a given category C and *two* objects A, B of C. We can define a new category $\mathsf{C}_{A,B}$ by essentially the same procedure that we used in order to define C_A:

- $\mathrm{Obj}(\mathsf{C}_{A,B})$ = diagrams

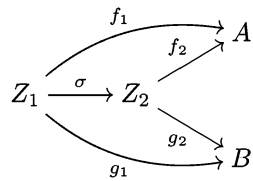


in C; and

- morphisms



are *commutative* diagrams



We will leave to the reader the task of formalizing this rough description. This example is really nothing more than a mixture of C_A and C_B, where the two structures interact because of the stringent requirement that the same σ must make both sides of the diagram commute:

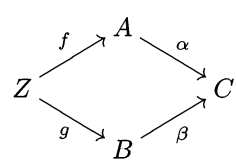
$$f_1 = f_2\sigma \quad \underline{\text{and}} \quad g_1 = g_2\sigma$$

'simultaneously'.

Flipping most of the arrows gives an analogous variation of Example 3.7, producing a category which we may[19] denote $\mathsf{C}^{A,B}$; details are left to the reader. ⌟

Example 3.10. As a final variation on these examples, we conclude by considering the *fibered* version of $\mathsf{C}_{A,B}$ (and $\mathsf{C}^{A,B}$). Take this as a test to see if you have really understood $\mathsf{C}_{A,B}$—experts would tell you that this looks fairly sophisticated for students just learning categories, so don't get disheartened if it does not flow too well at first (but pat yourself on the shoulder if it does!). Start with a given category C, and this time choose two fixed *morphisms* $\alpha : A \to C$, $\beta : B \to C$ in C, with the same target C. We can then consider a category $\mathsf{C}_{\alpha,\beta}$ as follows:

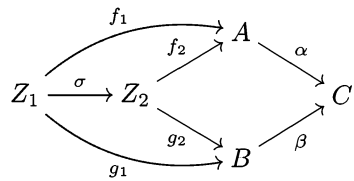
- $\mathrm{Obj}(\mathsf{C}_{\alpha,\beta})$ = *commutative diagrams*



in C, and

[19] There does not seem to be an established notation for these commonplace categories.

- morphisms correspond to commutative diagrams



A solid understanding of Example 3.9 will make this example look just as tame; at this point the reader should have no difficulties formalizing it (that is, explaining how composition works, what identities are, etc.).

Also left to the reader is the construction of the 'mirror' example $\mathsf{C}^{\alpha,\beta}$, starting from two morphisms $\alpha : C \to A$, $\beta : C \to B$ with common *source.* ⌟

Exercises

3.1. ▷ Let C be a category. Consider a structure C^{op} with

- $\mathrm{Obj}(\mathsf{C}^{op}) := \mathrm{Obj}(\mathsf{C})$;
- for A, B objects of C^{op} (hence objects of C), $\mathrm{Hom}_{\mathsf{C}^{op}}(A, B) := \mathrm{Hom}_{\mathsf{C}}(B, A)$.

Show how to make this into a category (that is, define composition of morphisms in C^{op} and verify the properties listed in §3.1).

Intuitively, the 'opposite' category C^{op} is simply obtained by 'reversing all the arrows' in C. [5.1, §VIII.1.1, §IX.1.2, IX.1.10]

3.2. If A is a finite set, how large is $\mathrm{End}_{\mathsf{Set}}(A)$?

3.3. ▷ Formulate precisely what it means to say that 1_a is an identity with respect to composition in Example 3.3, and prove this assertion. [§3.2]

3.4. Can we define a category in the style of Example 3.3 using the relation $<$ on the set $\mathbb{Z}$?

3.5. ▷ Explain in what sense Example 3.4 is an instance of the categories considered in Example 3.3. [§3.2]

3.6. ▷ (Assuming some familiarity with linear algebra.) Define a category V by taking $\mathrm{Obj}(\mathsf{V}) = \mathbb{N}$ and letting $\mathrm{Hom}_{\mathsf{V}}(n, m) =$ the set of $m \times n$ matrices with real entries, for all $n, m \in \mathbb{N}$. (We will leave the reader the task of making sense of a matrix with 0 rows or columns.) Use product of matrices to define composition. Does this category 'feel' familiar? [§VI.2.1, §VIII.1.3]

3.7. ▷ Define carefully objects and morphisms in Example 3.7, and draw the diagram corresponding to composition. [§3.2]

3.8. ▷ A *subcategory* C' of a category C consists of a collection of objects of C, with morphisms $\mathrm{Hom}_{\mathsf{C}'}(A, B) \subseteq \mathrm{Hom}_{\mathsf{C}}(A, B)$ for all objects A, B in $\mathrm{Obj}(\mathsf{C}')$, such that identities and compositions in C make C' into a category. A subcategory C' is *full*

if $\mathrm{Hom}_{\mathsf{C}'}(A,B) = \mathrm{Hom}_{\mathsf{C}}(A,B)$ for all A, B in $\mathrm{Obj}(\mathsf{C}')$. Construct a category of *infinite sets* and explain how it may be viewed as a full subcategory of Set. [4.4, §VI.1.1, §VIII.1.3]

3.9. ▷ An alternative to the notion of *multiset* introduced in §2.2 is obtained by considering sets endowed with equivalence relations; equivalent elements are taken to be multiple instances of elements 'of the same kind'. Define a notion of morphism between such enhanced sets, obtaining a category MSet containing (a 'copy' of) Set as a full subcategory. (There may be more than one reasonable way to do this! This is intentionally an open-ended exercise.) Which objects in MSet determine ordinary multisets as defined in §2.2 and how? Spell out what a morphism of multisets would be from this point of view. (There are several natural notions of morphisms of multisets. Try to define morphisms in MSet so that the notion you obtain for ordinary multisets captures your intuitive understanding of these objects.) [§2.2, §3.2, 4.5]

3.10. Since the objects of a category C are not (necessarily) sets, it is not clear how to make sense of a notion of 'subobject' in general. In some situations it *does* make sense to talk about subobjects, and the subobjects of any given object A in C are in one-to-one correspondence with the morphisms $A \to \Omega$ for a fixed, special object Ω of C, called a *subobject classifier*. Show that Set has a subobject classifier.

3.11. ▷ Draw the relevant diagrams and define composition and identities for the category $\mathsf{C}^{A,B}$ mentioned in Example 3.9. Do the same for the category $\mathsf{C}^{\alpha,\beta}$ mentioned in Example 3.10. [§5.5, 5.12]

---

4. Morphisms

Just as in Set we highlight certain types of functions (injective, surjective, bijective), it is useful to try to do the same for morphisms in an arbitrary category. The reader should note that defining qualities of morphisms by their actions on 'elements' is not an option in the general setting, because objects of an arbitrary category do not (in general) have 'elements'.

This is why we spent some time analyzing injectivity, etc., from different viewpoints in §§2.4-2.6. It turns out that the other viewpoints on these notions do transfer nicely into the categorical setting.

4.1. Isomorphisms.

Let C be a category.

Definition 4.1. A morphism $f \in \mathrm{Hom}_{\mathsf{C}}(A,B)$ is an *isomorphism* if it has a (two-sided) inverse under composition: that is, if $\exists g \in \mathrm{Hom}_{\mathsf{C}}(B,A)$ such that

$$gf = 1_A, \quad fg = 1_B. \qquad \lrcorner$$

Recall that in §2.5 the inverse of a bijection of sets f was defined 'elementwise'; in particular, there was no ambiguity in its definition, and we introduced the notation f^{-1} for this function. By contrast, the 'inverse' g produced in Definition 4.1 does not appear to have this uniqueness explicitly built into its definition. Luckily, its defining property *does* guarantee its uniqueness, but this requires a verification:

Proposition 4.2. *The inverse of an isomorphism is unique.*

Proof. We have to verify that if both g_1 and $g_2 : B \to A$ act as inverses of a given isomorphism $f : A \to B$, then $g_1 = g_2$. The standard trick for this kind of verification is to compose f on the left by one of the morphisms, and on the right by the other one; then apply associativity. The whole argument can be compressed into one line:

$$g_1 = g_1 1_B = g_1(fg_2) = (g_1 f)g_2 = 1_A g_2 = g_2$$

as needed. □

Note that the argument really proves that if f is a morphism with a left-inverse g_1 and a right-inverse g_2, then necessarily f is an isomorphism, $g_1 = g_2$, and this morphism is the (unique) inverse of f.

Since the inverse of f is uniquely determined by f, there is no ambiguity in denoting it by f^{-1}.

Proposition 4.3. *With notation as above:*

- *Each identity 1_A is an isomorphism and is its own inverse.*
- *If f is an isomorphism, then f^{-1} is an isomorphism and further $(f^{-1})^{-1} = f$.*
- *If $f \in \mathrm{Hom}_{\mathsf{C}}(A, B)$, $g \in \mathrm{Hom}_{\mathsf{C}}(B, C)$ are isomorphisms, then the composition gf is an isomorphism and $(gf)^{-1} = f^{-1}g^{-1}$.*

Proof. These all 'prove themselves'. For example, it is immediate to verify that $f^{-1}g^{-1}$ is a left-inverse of gf: indeed[20],

$$(f^{-1}g^{-1})(gf) = f^{-1}((g^{-1}g)f) = f^{-1}(1_B f) = f^{-1}f = 1_A.$$

The verification that $f^{-1}g^{-1}$ is also a right-inverse of gf is analogous. □

Note that taking the inverse reverses the order of composition: $(gf)^{-1} = f^{-1}g^{-1}$.

Two objects A, B of a category are *isomorphic* if there is an isomorphism $f : A \to B$. An immediate corollary of Proposition 4.3 is that 'isomorphism' is an equivalence relation[21]. If two objects A, B are isomorphic, one writes $A \cong B$.

Example 4.4. Of course, the isomorphisms in the category Set are precisely the bijections; this was observed at the beginning of §2.5. ⌟

Example 4.5. As noted in Proposition 4.3, identities are isomorphisms. They may be the *only* isomorphisms in a category: for example, this is the case in the category C obtained from the relation $\leq$ on $\mathbb{Z}$, as in Example 3.3. Indeed, for a, b objects of C (that is, $a, b \in \mathbb{Z}$), there is a morphism $f : a \to b$ *and* a morphism $g : b \to a$ only if $a \leq b$ and $b \leq a$, that is, if $a = b$. So an isomorphism in C necessarily acts from an object a to itself; but in C there is only one such morphism, that is, 1_a. ⌟

[20] Associativity of composition implies that parentheses may be shuffled at will in longer expressions, as done here (cf. Exercise 4.1).

[21] The reader should have checked this in Exercise 2.4, for Set; the same proof will work in any category.

Example 4.6. On the other hand, there are categories in which *every* morphism is an isomorphism; such categories are called *groupoids*. The reader 'already knows' many examples of groupoids; cf. Exercise 4.2. ⌟

An *automorphism* of an object A of a category C is an isomorphism from A to itself. The set of automorphisms of A is denoted $\mathrm{Aut}_{\mathsf{C}}(A)$; it is a subset of $\mathrm{End}_{\mathsf{C}}(A)$. By Proposition 4.3, composition confers on $\mathrm{Aut}_{\mathsf{C}}(A)$ a remarkable structure:

- the composition of two elements $f, g \in \mathrm{Aut}_{\mathsf{C}}(A)$ is an element $gf \in \mathrm{Aut}_{\mathsf{C}}(A)$;
- composition is associative;
- $\mathrm{Aut}_{\mathsf{C}}(A)$ contains the element 1_A, which is an identity for composition (that is, $f1_A = 1_A f = f$);
- every element $f \in \mathrm{Aut}_{\mathsf{C}}(A)$ has an inverse $f^{-1} \in \mathrm{Aut}_{\mathsf{C}}(A)$.

In other words, $\mathrm{Aut}_{\mathsf{C}}(A)$ is a *group,* for all objects A of all categories C.

We will soon devote all our attention to groups!

4.2. Monomorphisms and epimorphisms. As pointed out above, we do not have the option of defining for morphisms of an arbitrary category a notion such as 'injective' in the same way as we do for set-functions in §2.4: that definition requires a notion of 'element', and in general no such notion is available for objects of a category. But nothing prevents us from defining *monomorphisms* as we did in §2.6, in an arbitrary category:

Definition 4.7. Let C be a category. A morphism $f \in \mathrm{Hom}_{\mathsf{C}}(A, B)$ is a *monomorphism* if the following holds:

$$\text{for all objects } Z \text{ of } \mathsf{C} \text{ and all morphisms } \alpha', \alpha'' \in \mathrm{Hom}_{\mathsf{C}}(Z, A),$$
$$f \circ \alpha' = f \circ \alpha'' \implies \alpha' = \alpha''.$$ ⌟

Similarly, *epimorphisms* are defined as follows:

Definition 4.8. Let C be a category. A morphism $f \in \mathrm{Hom}_{\mathsf{C}}(A, B)$ is an *epimorphism* if the following holds:

$$\text{for all objects } Z \text{ of } \mathsf{C} \text{ and all morphisms } \beta', \beta'' \in \mathrm{Hom}_{\mathsf{C}}(B, Z),$$
$$\beta' \circ f = \beta'' \circ f \implies \beta' = \beta''.$$ ⌟

Example 4.9. As proven in Proposition 2.3, in the category Set the monomorphisms are precisely the injective functions. The reader should have by now checked that, likewise, in Set the epimorphisms are precisely the *surjective* functions (cf. Exercise 2.5). Thus, while the definitions given in §2.6 may have looked counterintuitive at first, they work as natural 'categorical counterparts' of the ordinary notions of injective/surjective functions. ⌟

Example 4.10. In the categories of Example 3.3, *every* morphism is both a monomorphism and an epimorphism. Indeed, recall that there is *at most one* morphism between any two objects in these categories; hence the conditions defining monomorphisms and epimorphisms are vacuous. ⌟

Contemplating Example 4.10 reveals a few unexpected twists in these definitions, which defy our intuition as set-theorists. For instance, in Set, a function is an isomorphism if and only if it is both injective and surjective, hence if and only if it is both a monomorphism and an epimorphism. But in the category defined by $\leq$ on $\mathbb{Z}$, *every* morphism is both a monomorphism and an epimorphism, while the only isomorphisms are the identities (Example 4.5). Thus this property is a special feature of Set, and we should not expect it to hold automatically in every category; it will not hold in the category Ring of *rings* (cf. §III.2.3). It *will* hold in every *abelian category* (of which Set is *not* an example!), but that is a story for a very distant future (Lemma IX.1.9).

Similarly, in Set a function is an epimorphism, that is, surjective, if and only if it has a right-inverse (Proposition 2.1); this may fail in general, even in respectable categories such as the category Grp of groups (cf. Exercise II.8.24).

Exercises

4.1. ▷ Composition is defined for *two* morphisms. If more than two morphisms are given, e.g.,

$$A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D \xrightarrow{i} E,$$

then one may compose them in several ways, for example:

$$(ih)(gf), \quad (i(hg))f, \quad i((hg)f), \quad \text{etc.}$$

so that at every step one is only composing two morphisms. Prove that the result of any such nested composition is independent of the placement of the parentheses. (Hint: Use induction on n to show that any such choice for $f_n f_{n-1} \cdots f_1$ equals

$$((\cdots((f_n f_{n-1})f_{n-2})\cdots)f_1).$$

Carefully working out the case $n = 5$ is helpful.) [§4.1, §II.1.3]

4.2. ▷ In Example 3.3 we have seen how to construct a category from a set endowed with a relation, provided this latter is reflexive and transitive. For what types of relations is the corresponding category a groupoid (cf. Example 4.6)? [§4.1]

4.3. Let A, B be objects of a category C, and let $f \in \mathrm{Hom}_{\mathsf{C}}(A, B)$ be a morphism.

- Prove that if f has a right-inverse, then f is an epimorphism.
- Show that the converse does not hold, by giving an explicit example of a category and an epimorphism without a right-inverse.

4.4. Prove that the composition of two monomorphisms is a monomorphism. Deduce that one can define a subcategory $\mathsf{C}_{\mathrm{mono}}$ of a category C by taking the same objects as in C and defining $\mathrm{Hom}_{\mathsf{C}_{\mathrm{mono}}}(A, B)$ to be the subset of $\mathrm{Hom}_{\mathsf{C}}(A, B)$ consisting of monomorphisms, for all objects A, B. (Cf. Exercise 3.8; of course, in general $\mathsf{C}_{\mathrm{mono}}$ is not full in C.) Do the same for epimorphisms. Can you define a subcategory $\mathsf{C}_{\mathrm{nonmono}}$ of C by restricting to morphisms that are *not* monomorphisms?

4.5. Give a concrete description of monomorphisms and epimorphisms in the category MSet you constructed in Exercise 3.9. (Your answer will depend on the notion of morphism you defined in that exercise!)

---

5. Universal properties

The 'abstract' examples in §3 may have left the reader with the impression that one can produce at will a large number of minute variations of the same basic ideas, without really breaking any new ground. This may be fun in itself, but *why* do we really want to explore this territory?

Categories offer a rich unifying language, giving us a bird's eye view of many constructions in algebra (and other fields). In this course, this will be most apparent in the steady appearance of constructions satisfying suitable *universal properties*. For instance, we will see in a moment that products and disjoint unions (as reviewed in §1.3 and following) are characterized by certain universal properties having to do with the categories $\mathsf{C}_{A,B}$ and $\mathsf{C}^{A,B}$ considered in Example 3.9.

Many of the concepts introduced in this course will have an explicit description (such as the definition of product of sets given in §1.4) and an accompanying description in terms of a universal property (such as the one we will see in §5.4). The 'explicit' description may be very useful in concrete computations or arguments, but as a rule it is the universal property that clarifies the true nature of the construction. In some cases (such as for the disjoint union) the explicit description may turn out to depend on a seemingly arbitrary choice, while the universal property will have no element of arbitrariness. In fact, viewing the construction in terms of its corresponding universal property clarifies why one can only expect it to be defined 'up to isomorphism'.

Also, deeper relationships become apparent when the constructions are viewed in terms of their universal properties. For example, we will see that products of sets and disjoint unions of sets are really 'mirror' constructions (in the sense that reversing arrows transforms the universal property for one into that for the other). This is not so clear (to this writer, anyway) from the explicit descriptions in §1.4.

5.1. Initial and final objects.

Definition 5.1. Let C be a category. We say that an object I of C is *initial* in C if for every object A of C there exists *exactly one* morphism $I \to A$ in C:

$$\forall A \in \mathrm{Obj}(\mathsf{C}): \quad \mathrm{Hom}_{\mathsf{C}}(I, A) \text{ is a singleton.}$$

We say that an object F of C is *final* in C if for every object A of C there exists *exactly one* morphism $A \to F$ in C:

$$\forall A \in \mathrm{Obj}(\mathsf{C}): \quad \mathrm{Hom}_{\mathsf{C}}(A, F) \text{ is a singleton.}$$ ⌟

One may use *terminal* to denote either possibility, but in general we would advise the reader to be explicit about which 'end' of C one is considering.

A category need not have initial or final objects, as the following example shows.

Example 5.2. The category obtained by endowing $\mathbb{Z}$ with the relation $\leq$ (see Example 3.3) has no initial or final object. Indeed, an initial object in this category would be an integer i such that $i \leq a$ for all integers a; there is no such integer. Similarly, a final object would be an integer f larger than every integer, and there is no such thing.

By contrast, the category considered in Example 3.6 *does* have a final object, namely the pair $(3, 3)$; it still has no initial object. ⌟

Also, initial and final objects, when they exist, may or may not be unique:

Example 5.3. In Set, the empty set $\emptyset$ is initial (the 'empty graph' defines the unique function from $\emptyset$ to any given object!), and clearly it is the unique set that fits this requirement (Exercise 5.2).

Set also has final objects: for every set A, there is a unique function from A to a singleton $\{p\}$ (that is, the 'constant' function). *Every* singleton is final in Set; thus, final objects are not unique in this category. ⌟

However, we claim that if initial/final objects exist, then they are unique *up to a unique isomorphism.* We will invoke this fact frequently, so here is its official statement and its (immediate) proof:

Proposition 5.4. *Let* C *be a category.*

- *If* I_1, I_2 *are both initial objects in* C, *then* $I_1 \cong I_2$.
- *If* F_1, F_2 *are both final objects in* C, *then* $F_1 \cong F_2$.

Further, these isomorphisms are uniquely determined.

Proof. Recall that (by definition of category!) for every object A of C there is at least one element in $\mathrm{Hom}_{\mathsf{C}}(A, A)$, namely the identity 1_A. If I is initial, then there is a *unique* morphism $I \to I$, which therefore must be the identity 1_I.

Now assume I_1 and I_2 are both initial in C. Since I_1 is initial, there is a *unique* morphism $f : I_1 \to I_2$ in C; we have to show that f is an isomorphism. Since I_2 is initial, there is a unique morphism $g : I_2 \to I_1$ in C. Consider $gf : I_1 \to I_1$; as observed, necessarily

$$gf = 1_{I_1}$$

since I_1 is initial. By the same token

$$fg = 1_{I_2}$$

since I_2 is initial. This proves that $f : I_1 \to I_2$ is an isomorphism, as needed.

The proof for final objects is entirely analogous (Exercise 5.3). □

Proposition 5.4 "explains" why, while not unique, the final objects in Set are all isomorphic: no singleton is more 'special' than any other singleton; this is the typical situation. There may be psychological reasons why *one* initial or final object looks more compelling than others (for example, the singleton $\{\emptyset\} = 2^{\emptyset}$ may look to some like the most 'natural' choice among all singletons), but this plays no role in how these objects sit in their category.

5.2. Universal properties. The most natural context in which to introduce *universal properties* requires a good familiarity with the language of *functors,* which we will only introduce at a later stage (cf. §VIII.1.1). For the purpose of the examples we will run across in (most of) this book, the following 'working definition' should suffice.

We say that a construction *satisfies a universal property* (or 'is the solution to a universal problem') when it may be viewed as a terminal object of a category. The category depends on the context and is usually explained 'in words' (and often without even mentioning the word *category*).

In particularly simple cases this may take the form of a statement such as $\emptyset$ *is universal with respect to the property of mapping to sets;* this is synonymous with the assertion that $\emptyset$ is initial in the category of Set.

More often, the situation is more complex. Since being initial/final amounts to the existence and uniqueness of certain morphisms, the 'explanation' of a universal property may follow the pattern, "object X is universal with respect to the following property: for any Y such that..., there exists a unique morphism $Y \to X$ such that...."

The not-so-naive reader will recognize that this explanation hides the definition of an accessory category and the statement that X is terminal (probably final in this case) in this new category. It is useful to learn how to translate such wordy explanations into what they really mean. Also, the reader should keep in mind that it is not uncommon to sweep under the rug part of the essential information about the solution to a universal problem (usually some key morphism): this information is presumably implicit in any given set-up. This will be apparent from the examples that follow.

5.3. Quotients. Let $\sim$ be an equivalence relation defined on a set A. Let's parse the assertion:

"The quotient $A/\sim$ is universal with respect to the property of mapping A to a set in such a way that equivalent elements have the same image."

What can this possibly mean, and is it true?

The assertion is talking about functions

$$A \xrightarrow{\varphi} Z$$

with Z any set, satisfying the property

$$a' \sim a'' \implies \varphi(a') = \varphi(a'').$$

These morphisms are objects of a category (very similar to the category defined in Example 3.7); for convenience, let's denote such an object by (φ, Z). The only reasonable way to define morphisms $(\varphi_1, Z_1) \to (\varphi_2, Z_2)$ is as commutative diagrams

$$\begin{array}{ccc} Z_1 & \xrightarrow{\sigma} & Z_2 \\ & {\scriptstyle\varphi_1}\nwarrow \quad \nearrow{\scriptstyle\varphi_2} & \\ & A & \end{array}$$

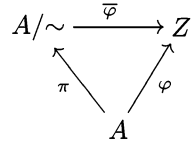
This is the same definition considered in Example 3.7.

Does this category have initial objects?

Claim 5.5. *Denoting by π the 'canonical projection' defined in Example 2.6, the pair $(\pi, A/\!\sim)$ is an initial object of this category.*

This is what our writer meant by the mysterious assertion copied above. Once this is understood, it is very easy to prove that the assertion is indeed correct.

Proof. Consider any (φ, Z) as above. We have to prove that there exists a unique morphism $(\pi, A/\!\sim) \to (\varphi, Z)$, that is, a unique commutative diagram



that is, a unique function $\overline{\varphi}$ making this diagram commute.

Let $[a]_\sim$ be an arbitrary element of $A/\!\sim$. If the diagram is indeed going to commute, then necessarily

$$\overline{\varphi}([a]_\sim) = \varphi(a);$$

this tells us that $\overline{\varphi}$ is indeed *unique,* if it exists at all—that is, *if* this prescription does define a function $A/\!\sim \to Z$.

Hence, all we have to check is that $\overline{\varphi}$ is well-defined, that is, that if $[a_1]_\sim = [a_2]_\sim$, then $\varphi(a_1) = \varphi(a_2)$; and indeed

$$[a_1]_\sim = [a_2]_\sim \implies a_1 \sim a_2 \implies \varphi(a_1) = \varphi(a_2).$$

This is precisely the condition that morphisms in our category satisfy. □

Note the several levels of sloppiness in the assertion considered above: it does not tell us very explicitly what category to consider; it does not tell us that we should especially pay attention to *initial* objects in this category. Worst of all, the solution to the universal problem is *not really* $A/\!\sim$, but rather *the morphism* $\pi : A \to A/\!\sim$.

The reader should practice the skill of translating loose assertions such as the one given above into precise statements; it is not at all uncommon to run into examples at the same level of 'abuse of language' as this one.

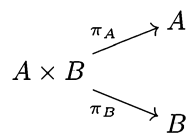
The reason why we get away with writing such assertions is that the context really allows the experienced reader to parse them effectively, and they are substantially more concise than their spelled-out version. After all, there is in general no conceivable choice for a morphism $A \to A/\!\sim$ other than the canonical projection; hence, neglecting to mention it is forgivable. Also, the *final* object in the category considered above is supremely uninteresting (what is it? cf. Exercise 5.5), so surely we must have meant the *initial* one.

What do we learn by viewing quotients in terms of their universal property? For example, suppose $\sim$ is the equivalence relation defined starting from a function $f : A \to B$, as in §2.8. Then the reader will realize easily that $\operatorname{im} f$ *also* satisfies

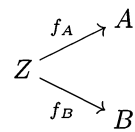
the universal property given above for $A/\sim$; therefore (by Proposition 5.4) $\operatorname{im} f$ and $A/\sim$ must be isomorphic. This is precisely the content of Theorem 2.7; thus, the universal property sheds some light on the 'canonical decomposition' studied in §2.8.

5.4. Products. It is also a very good exercise to stare at a familiar construction and try to see the universal property which may be behind it. We will now encourage you, dear reader, to contemplate the notion of the *product of two sets* given in §1.4 and to see if the universal property it satisfies jumps out at you. Spoiler follows, so this is a good time to stop reading these notes and to try on your own.

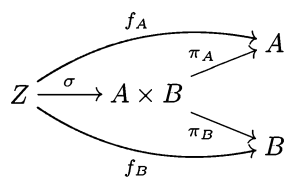
Here is the universal property. Let A, B be sets, and consider the product $A \times B$, with the two natural projections:



(see Example 2.4). Then *for every set Z and morphisms*



there exists a unique morphism $\sigma : Z \to A \times B$ such that the diagram



commutes.

In this situation, σ is usually denoted $f_A \times f_B$.

Proof. Define $\forall z \in Z$

$$\sigma(z) = (f_A(z), f_B(z)).$$

This function[22] manifestly makes the diagram commute: $\forall z \in Z$

$$\pi_A \sigma(z) = \pi_A(f_A(z), f_B(z)) = f_A(z),$$

showing that $\pi_A \sigma = f_A$ and similarly $\pi_B \sigma = f_B$.

Further, the definition is forced by the commutativity of the diagram; so σ is unique, as claimed. □

[22]Note that there is no 'well-definedness' issue this time.

In other words, products of sets (or, more precisely, products of sets together with the information of their natural projections to the factors) are *final* objects in the category $\mathsf{C}_{A,B}$ considered in Example 3.9, for $\mathsf{C} = \mathsf{Set}$.

What is the advantage of viewing products this way? The main advantage is that the universal property may be stated in *any* category, while the definition of products given in §1.4 only makes sense in Set (and possibly in other categories where one has a notion of 'elements'). We say that a category C *has (finite) products,* or is a category 'with (finite) products', if for all objects A, B in C the category $\mathsf{C}_{A,B}$ considered in Example 3.9 has final objects. Such a final object consists of the data of an object of C, usually denoted $A \times B$, and of two morphisms $A \times B \to A$, $A \times B \to B$.

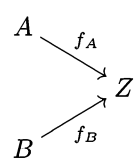
Note that a 'product' from this perspective does not need to 'look like a product'. Consider our recurring example of the category obtained from $\leq$ on $\mathbb{Z}$, as in Example 3.3. Does this category have products? Objects of this category are simply integers $a, b \in \mathbb{Z}$; call $a \times b$ for a moment the 'categorical' product of a and b. The universal property written out above becomes, in this case, *for all* $z \in \mathbb{Z}$ *such that* $z \leq a$ *and* $z \leq b$, *we have* $z \leq a \times b$.

This universal problem *does* have a solution $\forall a, b$: it is conventionally not called $a \times b$, but rather $\min(a, b)$. It is immediate to see that $\min(a, b)$ satisfies the property. Thus this category *has products,* and in fact we see that the product in this category amounts to the familiar operation of taking the minimum of two integers.

Thus there is an unexpected connection between 'the Cartesian product of two sets' and 'the minimum of two integers': both are examples of products, taken in different categories; they both satisfy 'the same' universal property, in different contexts.

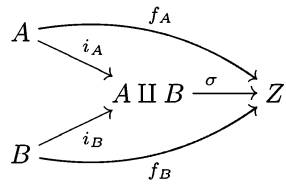
5.5. Coproducts. The prefix *co-* usually indicates that one is 'reversing all arrows'. Just as *products* are final objects in the categories $\mathsf{C}_{A,B}$ obtained by considering morphisms in C with common *source,* whose *targets* are A and B, co*products* will be *initial* objects in the categories[23] $\mathsf{C}^{A,B}$ of morphisms with common *target,* whose *sources* are A and B. Dear reader, look away and spell this universal property out before we do.

Here it is. Let A, B be objects of a category C. A *coproduct* $A \amalg B$ of A and B will be an object of C, endowed with two morphisms $i_A : A \to A \amalg B$, $i_B : B \to A \amalg B$ and satisfying the following universal property: *for all objects* Z *and morphisms*



[23]These categories were also considered in Example 3.9; cf. Exercise 3.11.

there exists a unique morphism $\sigma : A \amalg B \to Z$ *such that the diagram*



commutes.

The symmetry with the universal property of products is hopefully completely apparent. We say that a category C *has coproducts* if this universal problem has a solution for all pairs of objects A and B.

Is the reader familiar with any coproduct? Yes!

Proposition 5.6. *The* disjoint union *is a coproduct in* Set.

Proof. Recall (§1.4) that the disjoint union $A \amalg B$ is defined as the union of two disjoint isomorphic copies A', B' of A, B, respectively; for example, we may let $A' = \{0\} \times A$, $B' = \{1\} \times B$. The functions i_A, i_B are defined by

$$i_A(a) = (0, a), \quad i_B(b) = (1, b),$$

where we view these elements as elements of $(\{0\} \times A) \cup (\{1\} \times B)$.

Now let $f_A : A \to Z$, $f_B : B \to Z$ be arbitrary morphisms to a common target. Define

$$\sigma : A \amalg B = (\{0\} \times A) \cup (\{1\} \times B) \to Z$$

by

$$\sigma(c) = \begin{cases} f_A(a) & \text{if } c = (0, a) \in \{0\} \times A, \\ f_B(b) & \text{if } c = (1, b) \in \{1\} \times B. \end{cases}$$

This definition makes the relevant diagram commute and is in fact forced upon us by this commutativity, proving that σ exists and is unique. □

This observation tells us that the category Set has coproducts and further sheds considerable light on the mysteries of disjoint unions. For example, there was an element of arbitrariness in our choice of 'a' disjoint union, although different choices led to isomorphic notions. Now we see why: terminal objects of a category are not unique in general, although they are unique up to isomorphism (Proposition 5.4); there is not a 'most beautiful' disjoint union of two sets just as there is not a 'most beautiful' singleton in Set (cf. §5.1).

Also, an unexpected 'symmetry' between products and disjoint unions becomes suddenly apparent from the point of view of universal properties.

The reader is invited to contemplate the notion of coproduct in the other categories we have encountered. For example (and probably not surprisingly at this point) the category obtained from $\leq$ on $\mathbb{Z}$ *does* have coproducts: the coproduct of two objects (i.e., integers) a, b is simply the *maximum* of a and b.

Exercises

5.1. Prove that a final object in a category C is initial in the opposite category C^{op} (cf. Exercise 3.1).

5.2. ▷ Prove that $\emptyset$ is the *unique* initial object in Set. [§5.1]

5.3. ▷ Prove that final objects are unique up to isomorphism. [§5.1]

5.4. What are initial and final objects in the category of 'pointed sets' (Example 3.8)? Are they unique?

5.5. ▷ What are the final objects in the category considered in §5.3? [§5.3]

5.6. ▷ Consider the category corresponding to endowing (as in Example 3.3) the set $\mathbb{Z}^+$ of positive integers with the *divisibility* relation. Thus there is exactly one morphism $d \to m$ in this category if and only if d divides m without remainder; there is no morphism between d and m otherwise. Show that this category has products and coproducts. What are their 'conventional' names? [§VII.5.1]

5.7. Redo Exercise 2.9, this time using Proposition 5.4.

5.8. Show that in every category C the products $A \times B$ and $B \times A$ are isomorphic, if they exist. (Hint: Observe that they both satisfy the universal property for the product of A and B; then use Proposition 5.4.)

5.9. Let C be a category with products. Find a reasonable candidate for the universal property that the product $A \times B \times C$ of *three* objects of C ought to satisfy, and prove that both $(A \times B) \times C$ and $A \times (B \times C)$ satisfy this universal property. Deduce that $(A \times B) \times C$ and $A \times (B \times C)$ are necessarily isomorphic.

5.10. Push the envelope a little further still, and define products and coproducts for *families* (i.e., indexed sets) of objects of a category.

Do these exist in Set?

It is common to denote the product $\underbrace{A \times \cdots \times A}_{n \text{ times}}$ by A^n.

5.11. Let A, resp. B, be a set, endowed with an equivalence relation $\sim_A$, resp. $\sim_B$. Define a relation $\sim$ on $A \times B$ by setting

$$(a_1, b_1) \sim (a_2, b_2) \iff a_1 \sim_A a_2 \text{ and } b_1 \sim_B b_2.$$

(This is immediately seen to be an equivalence relation.)

- Use the universal property for quotients (§5.3) to establish that there are functions $(A \times B)/\!\sim \to A/\!\sim_A$, $(A \times B)/\!\sim \to B/\!\sim_B$.
- Prove that $(A \times B)/\!\sim$, with these two functions, satisfies the universal property for the product of $A/\!\sim_A$ and $B/\!\sim_B$.
- Conclude (without further work) that $(A \times B)/\!\sim \cong (A/\!\sim_A) \times (B/\!\sim_B)$.

5.12. ¬ Define the notions of *fibered products* and *fibered coproducts,* as terminal objects of the categories $\mathsf{C}_{\alpha,\beta}$, $\mathsf{C}^{\alpha,\beta}$ considered in Example 3.10 (cf. also Exercise 3.11), by stating carefully the corresponding universal properties.

As it happens, Set has both fibered products and coproducts. Define these objects 'concretely', in terms of naive set theory. [II.3.9, III.6.10, III.6.11]

Groups, first encounter

In this chapter we introduce groups, we observe they form a category (called Grp), and we study 'general' features of this category: what are the monomorphisms and the epimorphisms in this category? what is the appropriate notion of 'equivalence relation' and 'quotients' for a group? does a 'decomposition theorem' hold in Grp? and other analogous questions.

In Chapter III we will acquire a similar degree of familiarity with *rings* and *modules.* A more object-oriented analysis of Grp (for example, a treatment of the famous *Sylow theorems,* 'composition series', or the classification of finite abelian groups) is deferred to Chapter IV.

1. Definition of group

1.1. Groups and groupoids.

Joke 1.1. *Definition: A* group *is a groupoid with a single object.* ⌟

This is actually a perfectly viable definition, since groupoids have been defined already (in Example I.4.6); but most mathematicians would find it ludicrous to introduce groups in this fashion, or they will at the very least politely express doubts on the pedagogical effectiveness of doing so. In order to redeem himself, the author will parse this definition right away to show what it really says. If $*$ is the lone object of such a groupoid G,

$$\mathrm{Hom}_{\mathsf{G}}(*,*) = \mathrm{Aut}_{\mathsf{G}}(*)$$

(because G is a groupoid!), and this set carries all the information about G. Call this set G. Then (by definition of category) there is an associative operation on G, with an identity 1_*, and (by definition of groupoid, which says that every morphism in G is an isomorphism) every $g \in G$ has an inverse $g^{-1} \in G$.

That is what a group is[1]: a set G with a composition law satifsying a few key axioms, i.e., associativity, existence of identity, and existence of inverses.

1.2. Definition. Now for the official definition. Let G be a nonempty set, endowed with a *binary operation*, that is, a 'multiplication' map

$$\bullet : G \times G \to G.$$

Our notation will be

$$\bullet(g, h) =: g \bullet h$$

or simply gh if the name of the operation can be understood. The careful reader may have expected that we should write $h \bullet g$, in the style of what we have done for categories, but this is what common conventions dictate[2].

Definition 1.2. The set G, endowed with the binary operation $\bullet$ (briefly, $(G, \bullet)$, or simply G if the operation can be understood) is a *group* if

(i) the operation $\bullet$ is *associative,* that is,

$$(\forall g, h, k \in G): \quad (g \bullet h) \bullet k = g \bullet (h \bullet k);$$

(ii) there exists an *identity element* e_G for $\bullet$, that is,

$$(\exists e_G \in G)\,(\forall g \in G): \quad g \bullet e_G = g = e_G \bullet g;$$

(iii) every element in G has an *inverse* with respect to $\bullet$, that is,

$$(\forall g \in G)\,(\exists h \in G): \quad g \bullet h = e_G = h \bullet g. \qquad \lrcorner$$

Example 1.3. Since we explicitly require G to be nonempty, the most economical way to concoct a group is by letting $G = \{e\}$ be a singleton. There is only one function $G \times G \to G$ in this case, so there is only one possible binary operation on G, defined by

$$e \bullet e := e.$$

The three axioms trivially hold for this example, so $\{e\}$ is equipped with a unique group structure.

This is usually called the *trivial group;* purists should call any such group *a* trivial group, since every singleton gives rise to one. $\lrcorner$

Example 1.4. The reader should check carefully (cf. Exercise 1.2) that $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, $(\mathbb{C}, +)$, and several variations using $\cdot$ (for example, the subset $\{+1, -1\}$ of $\mathbb{Z}$, with ordinary multiplication) all give example of groups. While very interesting in themselves, these examples do not really capture at any intuitive level 'what' a group really is, because they are too special. For example, all these examples are *commutative* (see §1.5). $\lrcorner$

[1]From this perspective, Joke 1.1 is a little imprecise: the group is not the groupoid G, but rather the set of isomorphisms in G, endowed with the operation of composition of morphisms.

[2]Not without exceptions; see for example permutation groups, discussed in §2.

Example 1.5. My readers are likely familiar with an extremely important *non-commutative* example, namely the group of *invertible, $n \times n$ matrices* with (say) real entries, $n \geq 2$. We will generally shy away from this class of examples in these early chapters, since we will have ample opportunities to think about matrices when we approach linear algebra (starting in Chapter VI). But we may occasionally borrow a matrix or two before then. The reader should now check that 2×2 matrices

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with real entries, and such that $ad - bc \neq 0$, form a group under the ordinary matrix multiplication:

$$\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \cdot \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} = \begin{pmatrix} a_1a_2 + b_1c_2 & a_1b_2 + b_1d_2 \\ c_1a_2 + d_1c_2 & c_1b_2 + d_1d_2 \end{pmatrix}.$$

(The condition $ad - bc \neq 0$ guarantees that the matrix is *invertible*. What is its inverse?) Since, for example,

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \neq \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

this group is indeed *not* commutative.

The group of invertible $n \times n$ matrices with real entries is denoted $\mathrm{GL}_n(\mathbb{R})$. ⌟

We will encounter more representative examples in §2.

1.3. Basic properties. From the 'groupoid' point of view, the identity e_G would be denoted 1_*. It is not uncommon to omit G from the notation (when the group is understood) and to use different symbols rather than e to denote this element: 1 and 0 are popular alternatives, depending on the context. In any case, for any given group G this element is unique. That is, no other element of G can work as an identity:

Proposition 1.6. *If $h \in G$ is an identity of G, then $h = e_G$.*

Incidentally, this makes groups *pointed sets* in the sense of Example I.3.8: every group has a well-defined distinguished element.

Proof. Using first that e_G is an identity and then that h is an identity, one gets[3]

$$h = e_G h = e_G.$$

(Amusingly, this argument only uses that e_G is a 'left' identity and h is a 'right' identity.) □

Proposition 1.7. *The inverse is also unique: if h_1, h_2 are both inverses of g in G, then $h_1 = h_2$.*

[3]As previously announced, we may omit the symbol • for the operation, since for the time being we are only considering *one* operation. This will be done without warning in the future.

Proof. This actually follows from Proposition I.4.2 (by viewing G as the set of isomorphisms of a groupoid with a single object). The reader should construct a stand-alone proof, using the same trick, but carefully hiding any reference to morphims. □

Proposition 1.7 authorizes us to give a name to *the* inverse of g: this is usually[4] denoted g^{-1}.

One more notational item is in order. The definition of a group only contemplates the 'product' of *two* elements; in multiplying a string of elements, one may in principle have a choice as to the order in which products are executed. For example,

$$(g_1 \bullet g_2) \bullet g_3$$

stands for: apply the operation $\bullet$ to g_1 and g_2, and then apply $\bullet$ again to the result of this operation and g_3; while

$$g_1 \bullet (g_2 \bullet g_3)$$

stands for: apply $\bullet$ to g_2 and g_3, and then apply it again to g_1 and to the result of this operation.

Associativity tells us precisely that the result of the operation on three elements does not depend on the way in which we perform it. With this in mind, we *are* authorized to write

$$g_1 \bullet g_2 \bullet g_3;$$

this expression is not ambiguous, *by associativity*. What about four or more elements? The reader should have checked in Exercise I.4.1 that all ways to associate any number of elements leads to the same result. So we are also authorized to write things like

$$g_1 \bullet g_2 \bullet g_3 \bullet \cdots \bullet g_{17};$$

this is also unambiguous. The reader should keep in mind, however, that of course the order in which the *elements* are listed *is* important: in general,

$$g_1 \bullet g_2 \bullet g_3 \bullet \cdots \bullet g_{17} \neq g_2 \bullet g_1 \bullet g_3 \bullet \cdots \bullet g_{17}.$$

Of course no such care is necessary if all g_i coincide; the conventional 'power' notation can then be used[5]: $g^0 = e_G$, and for a positive integer n

$$g^n = \underbrace{g \cdot \cdots \cdot g}_{n \text{ times}}, \quad g^{-n} = \underbrace{g^{-1} \cdot \cdots \cdot g^{-1}}_{n \text{ times}}.$$

It is easy to check that then $\forall g \in G$ and $\forall m, n \in \mathbb{Z}$

$$g^{m+n} = g^m g^n.$$

[4] But note the 'abelian' case, discussed in §1.5.

[5] In the abelian case one uses 'multiples'; cf. §1.5.

1.4. Cancellation. 'Cancellation' holds in groups. That is,

Proposition 1.8. *Let G be a group. Then $\forall a, g, h \in G$*

$$ga = ha \implies g = h, \quad ag = ah \implies g = h.$$

Proof. Both statements are proven by multiplying (on the appropriate side) by a^{-1} and applying associativity. For example,

$$ga = ha \implies (ga)a^{-1} = (ha)a^{-1} \implies g(aa^{-1}) = h(aa^{-1}) \implies ge_G = he_G$$
$$\implies g = h.$$

The proof for the other implication follows the same pattern. □

Examples of operations which do *not* satisfy cancellation, and hence do *not* define groups, abound. For instance, the operation of ordinary multiplication does *not* make the set $\mathbb{R}$ of real numbers into a group: indeed, 0 'cannot be cancelled' since $1 \cdot 0 = 2 \cdot 0$ even if $1 \neq 2$. Of course, the problem here is that 0 does not have an inverse in $\mathbb{R}$, with respect to multiplication. As it happens, this is the *only* problem with this example: ordinary multiplication *does* make

$$\mathbb{R}^* := \mathbb{R} \smallsetminus \{0\}$$

into a group, as the reader should check immediately.

1.5. Commutative groups. One axiom *not* appearing in the definition of group is *commutativity:* we say that the operation $\bullet$ is 'commutative' if

(iv) $(\forall g, h \in G)$: $\quad g \bullet h = h \bullet g.$

We say that two elements g, h 'commute' if $gh = hg$. Thus, in a commutative group any two elements commute.

'Commutative groups' are important objects: they arise naturally in several contexts, especially as 'modules over the ring $\mathbb{Z}$' (about which we will have a lot to say in Chapter III and beyond). When they do so, they are usually called *abelian groups.*

The notation used in treating abelian groups differs somewhat from the standard notation for groups. This is to emphasize the '$\mathbb{Z}$-module structure', and it is helpful when an abelian group coexists with other operations—a situation which we will encounter frequently.

Thus, the *operation* in an abelian group A is, as a rule, denoted by $+$ and is called 'addition'; the *identity* is then called 0_A; and the *inverse* of an element $a \in A$ is denoted $-a$ (and maybe should be called the 'opposite'?). The 'power' notation is of course replaced by 'multiple': $0a = 0$, and for a positive integer n

$$na = \underbrace{a + \cdots + a}_{n \text{ times}}, \quad (-n)a = \underbrace{(-a) + \cdots (-a)}_{n \text{ times}}.$$

The reader should keep in mind that at this stage 'na' is a *notation,* not the result of applying a binary operation to two elements n, a of A. Indeed, $n \in \mathbb{Z}$ may very well not be an element of A in any reasonable sense. Moreover, it may very

well be that $na = ma$ even if $n \neq m$, in spite of the fact that 'cancellation' works in groups.

The qualifier 'abelian' and the notation 0_A, $-a$, etc., are mostly used for commutative groups arising in certain standard situations: for example, the notions of rings, modules, vector spaces are defined by suitably enriching a commutative group, which is then promoted to 'abelian' for notational convenience.

There are other situations in which commutative groups arise naturally, without triggering the 'abelian' notation. For example, the group $(\mathbb{R}^*, \cdot)$ mentioned at the end of §1.4 *is* commutative, but its operation is indicated by $\cdot$ (if at all), and its identity element is written 1. History, rather than logic, is often the main factor determining notation.

1.6. Order.

Definition 1.9. An element g of a group G has *finite order* if[6] $g^n = e$ for some positive integer n. In this case, the *order* $|g|$ is the *smallest* positive n such that $g^n = e$. One writes $|g| = \infty$ if g does not have finite order. ⌟

By definition, if $g^n = e$ for some positive integer n, then $|g| \leq n$. One can be more precise:

Lemma 1.10. *If $g^n = e$ for some positive integer n, then $|g|$ is a divisor of n.*

Proof. As observed, $n \geq |g|$ by definition of order, that is, $n - |g| \geq 0$. There must then exist[7] a positive integer m such that

$$r = n - |g| \cdot m \geq 0 \quad \text{and} \quad n - |g| \cdot (m+1) < 0,$$

that is, $r < |g|$. Note that

$$g^r = g^{n-|g|\cdot m} = g^n \cdot (g^{|g|})^{-m} = e \cdot e^{-m} = e.$$

By definition of order, $|g|$ is the smallest *positive* integer such that $g^{|g|} = e$. Since r is smaller than $|g|$ and $g^r = e$, r cannot be positive; hence $r = 0$ necessarily. This says

$$0 = n - |g| \cdot m,$$

that is, $n = |g| \cdot m$, proving that n is indeed an integer multiple of $|g|$ as claimed. □

This lemma has the following immediate and useful consequence, which we encourage the reader to keep firmly in mind:

Corollary 1.11. *Let g be an element of finite order, and let $N \in \mathbb{Z}$. Then*

$$g^N = e \iff N \text{ is a multiple of } |g|.$$

---

[6] Of course in an abelian group we would write the following prescription as $ng = 0$.

[7] Purists may object that here we are surreptitiously using fairly sophisticated information about $\mathbb{Z}$, namely the 'division algorithm', hence essentially the fact that $\mathbb{Z}$ is a Euclidean domain! This is material that will have to wait until Chapter V to be given some justice. We may as well be open about it and admit that yes, we are assuming that the readers have already acquired a thorough familiarity with the operations of addition and multiplication among integers. Shame on us!

Definition 1.12. If G is finite as a set, its *order* $|G|$ is the number of its elements; we write $|G| = \infty$ if G is infinite. ⌟

Cancellation implies that $|g| \leq |G|$ for all $g \in G$. Indeed, this is vacuously true if $|G| = \infty$; if G is finite, consider the $|G| + 1$ powers

$$g^0 = e, g, g^2, g^3, \ldots, g^{|G|}$$

of g. These cannot all be distinct; hence

$$(\exists i, j) \quad 0 \leq i < j \leq |G| \quad \text{such that } g^i = g^j.$$

By cancellation (that is, multiplying on the right by g^{-i})

$$g^{j-i} = e,$$

showing $|g| \leq (j - i) \leq |G|$.

We will soon be able to formulate a much more precise statement concerning the relation between the order of a group and the order of its elements: if $g \in G$ and $|G|$ is finite, then the order of g *divides* the order of G. This will be an immediate consequence of *Lagrange's theorem;* cf. Example 8.15.

Another general remark concerning orders is that their behavior with respect to the operation of the group is not always predictable: it may very well happen that g, h have finite order in a group G, and yet $|gh| = \infty$, or $|gh| =$ your favorite positive integer: work out Exercise 1.12 and Exercise 2.6 if you don't believe it.

On the other hand, the situation is more constrained if g and h commute. In the extreme case in which $g = h$, it is easy to obtain a very precise statement:

Proposition 1.13. *Let $g \in G$ be an element of finite order. Then g^m has finite order $\forall m \geq 0$, and in fact*[8]

$$|g^m| = \frac{\operatorname{lcm}(m, |g|)}{m} = \frac{|g|}{\gcd(m, |g|)}.$$

Proof. The equality of the two numbers $\frac{\operatorname{lcm}(m,|g|)}{m}$ and $\frac{|g|}{\gcd(m,|g|)}$ follows from elementary properties of gcd and lcm: $\operatorname{lcm}(a, b) = ab/\gcd(a, b)$ for all a and b. So we only need to prove that $|g^m| = \frac{\operatorname{lcm}(m,|g|)}{m}$.

The order of g^m is the least positive d for which

$$g^{md} = e,$$

that is (by Corollary 1.11) for which md is a multiple of $|g|$. In other words, $m|g^m|$ is the smallest multiple of m which is also a multiple of $|g|$:

$$m|g^m| = \operatorname{lcm}(m, |g|).$$

The stated formula follows immediately from this. □

In general, for commuting elements,

Proposition 1.14. *If $gh = hg$, then $|gh|$ divides* $\operatorname{lcm}(|g|, |h|)$.

[8]The notation lcm stands for 'least common multiple'. We are also assuming that the reader is familiar with simple properties of gcd and lcm.

Proof. Let $|g| = m$, $|h| = n$. If N is any common multiple of m and n, then $g^N = h^N = e$ by Corollary 1.11. Since g and h commute,

$$(gh)^N = \underbrace{(gh)(gh)\cdot\cdots\cdot(gh)}_{N \text{ times}} = \underbrace{gg\cdot\cdots\cdot g}_{N \text{ times}} \cdot \underbrace{hh\cdot\cdots\cdot h}_{N \text{ times}} = g^N h^N = e.$$

As this holds for every common multiple N of m and n, in particular

$$(gh)^{\mathrm{lcm}(m,n)} = e.$$

The statement then follows from Lemma 1.10. □

One cannot say more about $|gh|$ in general, even if g and h commute (Exercise 1.13). But see Exercise 1.14 for an important special case.

Exercises

1.1. ▷ Write a careful proof that every group is the group of isomorphisms of a groupoid. In particular, every group is the group of automorphisms of some object in some category. [§2.1]

1.2. ▷ Consider the 'sets of numbers' listed in §1.1, and decide which are made into groups by conventional operations such as + and ·. Even if the answer is negative (for example, $(\mathbb{R}, \cdot)$ is not a group), see if variations on the definition of these sets lead to groups (for example, $(\mathbb{R}^*, \cdot)$ *is* a group; cf. §1.4). [§1.2]

1.3. Prove that $(gh)^{-1} = h^{-1}g^{-1}$ for all elements g, h of a group G.

1.4. Suppose that $g^2 = e$ for all elements g of a group G; prove that G is commutative.

1.5. The 'multiplication table' of a group is an array compiling the results of all multiplications $g \bullet h$:

$\bullet$	e	$\cdots$	h	$\cdots$
e	e	$\cdots$	h	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
g	g	$\cdots$	$g \bullet h$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

(Here e is the identity element. Of course the table depends on the order in which the elements are listed in the top row and leftmost column.) Prove that every row and every column of the multiplication table of a group contains all elements of the group exactly once (like Sudoku diagrams!).

1.6. ¬ Prove that there is only *one* possible multiplication table for G if G has exactly 1, 2, or 3 elements. Analyze the possible multiplication tables for groups with exactly 4 elements, and show that there are *two* distinct tables, up to reordering the elements of G. Use these tables to prove that all groups with ≤ 4 elements are commutative.

(You are welcome to analyze groups with 5 elements using the same technique, but you will soon know enough about groups to be able to avoid such brute-force approaches.) [2.19]

1.7. Prove Corollary 1.11.

1.8. ¬ Let G be a finite group, with exactly one element f of order 2. Prove that $\prod_{g\in G} g = f$. [4.16]

1.9. Let G be a finite group, of order n, and let m be the number of elements $g \in G$ of order exactly 2. Prove that $n - m$ is odd. Deduce that if n is even, then G necessarily contains elements of order 2.

1.10. Suppose the order of g is odd. What can you say about the order of g^2?

1.11. Prove that for all g, h in a group G, $|gh| = |hg|$. (Hint: Prove that $|aga^{-1}| = |g|$ for all a, g in G.)

1.12. ▷ In the group of invertible 2×2 matrices, consider

$$g = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad h = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}.$$

Verify that $|g| = 4$, $|h| = 3$, and $|gh| = \infty$. [§1.6]

1.13. ▷ Give an example showing that $|gh|$ is not necessarily equal to $\operatorname{lcm}(|g|, |h|)$, even if g and h commute. [§1.6, 1.14]

1.14. ▷ As a counterpoint to Exercise 1.13, prove that if g and h commute *and* $\gcd(|g|, |h|) = 1$, then $|gh| = |g|\,|h|$. (Hint: Let $N = |gh|$; then $g^N = (h^{-1})^N$. What can you say about this element?) [§1.6, 1.15, §IV.2.5]

1.15. ¬ Let G be a commutative group, and let $g \in G$ be an element of maximal *finite* order, that is, such that if $h \in G$ has finite order, then $|h| \leq |g|$. Prove that in fact if h has finite order in G, then $|h|$ *divides* $|g|$. (Hint: Argue by contradiction. If $|h|$ is finite but does not divide $|g|$, then there is a prime integer p such that $|g| = p^m r$, $|h| = p^n s$, with r and s relatively prime to p and $m < n$. Use Exercise 1.14 to compute the order of $g^{p^m} h^s$.) [§2.1, 4.11, IV.6.15]

---

2. Examples of groups

2.1. Symmetric groups. In §I.4.1 we have already observed that every object A of every category C determines a group, called $\operatorname{Aut}_{\mathsf{C}}(A)$, namely the group of *automorphisms* of A. In a somewhat artificial sense it is clear that every group arises in this fashion (cf. Exercise 1.1); this fact is true in more 'meaningful' ways, which will become apparent when we discuss *group actions* (§9): cf. especially Theorem 9.5 and Exercise 9.17.

In any case, this observation provides the reader with an infinite class of very important examples:

Definition 2.1. Let A be a set. The *symmetric group*, or *group of permutations* of A, denoted S_A, is the group $\mathrm{Aut}_{\mathsf{Set}}(A)$. The group of permutations of the set $\{\mathbf{1}, \ldots, \mathbf{n}\}$ is denoted by S_n. ⌟

The terminology is easily justified: the automorphisms of a set A are the set-isomorphisms, that is, the bijections, from A to itself; applying such a bijection amounts precisely to permuting ('scrambling') the elements of A. This operation may be viewed as a transformation of A which does not change it (as a set), hence a 'symmetry'.

The groups S_A are famously large: as the reader checked in Exercise I.2.1, $|S_n| = n!$. For example, $|S_{70}| > 10^{100}$, which is substantially larger than the estimated number of elementary particles in the observable universe.

Potentially confusing point: The various conventions clash in the way the operation in S_A should be written. From the 'automorphism' point of view, elements of S_A are functions and should be composed as such; thus, if $f, g \in S_A = \mathrm{Aut}_{\mathsf{Set}}(A)$, then the 'product' of f and g should be written $g \circ f$ and should act as follows:

$$(\forall p \in A): \quad g \circ f(p) = g(f(p)).$$

But the prevailing style of notation in group theory would write this element as fg, apparently reversing the order in which the operation is performed.

Everything would fall back into agreement if we adopted the convention of writing functions *after* the elements on which they act rather than before: $(p)f$ rather than $f(p)$. But one cannot change century-old habits, so we have no alternative but to live with both conventions and to state carefully which one we are using at any given time.

Contemplating the groups S_n for small values of n is an exercise of inestimable value. Of course S_1 is a trivial group; S_2 consists of the two possible permutations:

$$\begin{cases} \mathbf{1} \mapsto \mathbf{1} \\ \mathbf{2} \mapsto \mathbf{2} \end{cases} \quad \text{and} \quad \begin{cases} \mathbf{1} \mapsto \mathbf{2} \\ \mathbf{2} \mapsto \mathbf{1} \end{cases}$$

which we could call e (identity) and f (flip), with operation

$$ee = ff = e, \quad ef = fe = f.$$

In practice we cannot give a new name to every different element of every permutation group, so we have to develop a more flexible notation. There are in fact several possible choices for this; for the time being, we will indicate an element $\sigma \in S_n$ by listing the effect of applying σ underneath the list $1, \ldots, n$, as a matrix[9]. Thus the elements e, f in S_2 may be denoted by

$$e = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix} \quad , \quad f = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

[9]This is only a notational device—these matrices should not be confused with the matrices appearing in linear algebra.

In the same notational style, S_3 consists of

$$\left\{\begin{pmatrix}1&2&3\\1&2&3\end{pmatrix},\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix},\begin{pmatrix}1&2&3\\3&2&1\end{pmatrix},\begin{pmatrix}1&2&3\\1&3&2\end{pmatrix},\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix},\begin{pmatrix}1&2&3\\2&3&1\end{pmatrix}\right\}.$$

For the multiplication, we will adopt the sensible (but not very standard) convention mentioned above and have permutations act 'on the right': thus, for example,

$$\mathbf{1}\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}=\mathbf{2}\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}=\mathbf{1}$$

and similarly

$$\mathbf{2}\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}=\mathbf{3},\quad \mathbf{3}\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}=\mathbf{2}.$$

That is,

$$\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}=\begin{pmatrix}1&2&3\\1&3&2\end{pmatrix}$$

since the permutations on both sides of the equal sign act in the same way on **1**, **2**, **3**. The reader should now check that

$$\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}=\begin{pmatrix}1&2&3\\3&2&1\end{pmatrix}.$$

That is, letting

$$x=\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix},\quad y=\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix},$$

then

$$yx\neq xy,$$

showing that the operation in S_3 does not satisfy the commutative axiom. Thus, S_3 is a noncommutative group; the reader will immediately realize that in fact S_n is noncommutative for all $n\geq 3$.

While the commutation relation does not hold, other interesting relations do hold in S_3. For example,

$$x^2=e,\quad y^3=e,$$

showing that S_3 contains elements of order 1 (the identity e), 2 (the element x), and 3 (the element y); cf. Exercise 2.2. (Incidentally, this shows that the result of Exercise 1.15 does require the commutativity hypothesis.) Also,

$$yx=\begin{pmatrix}1&2&3\\3&2&1\end{pmatrix}=xy^2$$

as the reader may check. Using these relations, we see that every product of any assortment of x and y, $x^{i_1}y^{i_2}x^{i_3}y^{i_4}\cdots$, may be reduced to a product x^iy^j with $0\leq i\leq 1$, $0\leq j\leq 2$, that is, to one of the six elements

$$e,\quad y,\quad y^2,\quad x,\quad xy,\quad xy^2:$$

for example,

$$y^7x^{13}y^5=(y^3)^2y(x^2)^6xy^3y^2=(yx)y^2=(xy^2)y^2=xy^3y=xy.$$

On the other hand, these six elements are all distinct—this may be checked by cancellation and order considerations[10]. For example, if we had $xy^2 = y$, then we would get $x = y^{-1}$ by cancellation, and this cannot be since the relations tell us that x has order 2 and y^{-1} has order 3.

The conclusion is that the six products displayed above must be *all* six elements of S_3:

$$S_3 = \{e, x, y, xy, y^2, xy^2\}.$$

In the process we have verified that S_3 may also be described as the group 'generated' by two elements x and y, with the 'relations' $x^2 = e$, $y^3 = e$, $yx = xy^2$.

More generally, a subset A of a group G 'generates' G if every element of G may be written as a product of elements of A and of inverses of elements of A. We will deal with this notion more formally in §6.3 and with descriptions of groups in terms of generators and relations in §8.2.

2.2. Dihedral groups. A 'symmetry' is a transformation which preserves a structure. This is of course just a loose way to talk about automorphisms, when we may be too lazy to define rigorously the relevant category. As automorphisms of objects of a category, symmetries will naturally form groups.

One context in which this notion may be visualized vividly is that of 'geometric figures' such as polygons in the plane or polyhedra in space. The relevant category could be defined as follows: let the objects be subsets of an ordinary plane $\mathbb{R}^2$ and let morphisms between two subsets A, B consist of the 'rigid motions' of the plane (such as translations, rotations, or reflections about a line) which map A to a subset of B. A rigorous treatment of these notions would be too distracting at this point, so we will appeal to the intuition of the reader, as we do every now and then.

From this perspective, the 'symmetries' of a subset of the plane are the rigid motions which map it onto itself; they clearly form a group.

The *dihedral* groups may be defined as these groups of symmetries for the *regular polygons*. Placing the polygon so that it is centered at the origin (thereby excluding translations as possible symmetries), we see that the dihedral group for a regular n-sided polygon consists of the n rotations by $2\pi/n$ radians about the origin and the n reflections about lines through the origin and one of the vertices.

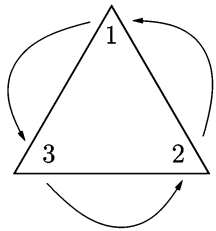
Thus, the dihedral group for a regular n-sided polygon consists of $2n$ elements; we will denote[11] this group by the symbol D_{2n}.

Again, contemplating these groups, at least for small values of n, is a wonderful exercise. There is a simple way to relate the dihedral groups to the symmetric groups of §2.1, capturing the fact that a symmetry of a regular polygon P is determined by the fate of the *vertices* of P. For example, label the vertices of an equilateral triangle clockwise by **1**, **2**, **3**; then a counterclockwise rotation by an angle of $2\pi/3$ sends vertex **1** to vertex **3**, **3** to **2**, and **2** to **1**, and no other symmetry of the triangle does the same.

[10] It may of course also be checked by explicit computation of the corresponding permutations, but we are trying to illustrate the fact that the relations are 'all we need to know'.

[11] Unfortunately there does not seem to be universal agreement on this notation: some references use the symbol D_n for what we call D_{2n} here.

Visually, this looks like



In other words, we can associate with the counterclockwise rotation of an equilateral triangle by $2\pi/3$ the permutation

$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \in S_3.$$

Such a labeling defines a *function*

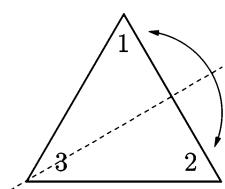
$$D_6 \to S_3;$$

further, this function is injective (since a symmetry is determined by the permutation of vertices it induces). In fancier language, which we will master in due time, we say that D_6 *acts (faithfully) on the set* $\{\mathbf{1}, \mathbf{2}, \mathbf{3}\}$.

It is clear that this can be done in several ways (for example, we could label the vertices in different ways). However, any such assignment will have the property that composition of symmetries in D_{2n} corresponds to composition of permutations in S_n; the reader should carefully work this out for several examples involving $D_6 \to S_3$ and $D_8 \to S_4$.

A concise way to describe the situation is that these functions are *(group) homomorphisms* (cf. §4). Since both D_6 and S_3 have 6 elements and the function $D_6 \to S_3$ given above is injective, it must also be surjective. Thus there are bijective homomorphisms between D_6 and S_3; we say that these groups are *isomorphic* (cf. §4.3). We will study these concepts very carefully in the next several sections.

As an alternative (and more abstract) way to draw the same conclusion, denote by y the counterclockwise rotation considered above and by x the reflection about the line through the center and vertex 3 of our equilateral triangle:



Reflecting twice gives the identity, as does rotating three times; thus

$$x^2 = e, \quad y^3 = e.$$

Further, yx (rotating counterclockwise by $2\pi/3$, then flipping about the slanted line) is the same symmetry as xy^2 (flipping first, then rotating *clockwise* by $2\pi/3$). That is,

$$yx = xy^2.$$

In other words, the group D_6 is *also* generated by two elements x, y, subject to the relations $x^2 = e$, $y^3 = e$, $yx = xy^2$—precisely as we found for S_3. Since D_6 and S_3 admit matching descriptions in terms of generators and relations (that is, matching *presentations;* cf. §8.2), 'of course' they are isomorphic.

2.3. Cyclic groups and modular arithmetic. Let n be a positive integer. Consider the equivalence relation on $\mathbb{Z}$ defined by[12]

$$(\forall a, b \in \mathbb{Z}): \quad a \equiv b \mod n \iff n \mid (b - a).$$

This is called *congruence modulo* n. We have encountered this relation already, for $n = 2$, in Example I.1.3. It is very easy to check that it is an equivalence relation, for all n; the set of equivalence classes is often denoted by $\mathbb{Z}_n$, $\mathbb{Z}/(n)$, $\mathbb{Z}/n\mathbb{Z}$, or $\mathbb{F}_n$. We will opt for $\mathbb{Z}/n\mathbb{Z}$, which is not preempted by other notions[13]. We will denote by $[a]_n$ the equivalence class of the integer a modulo n, or simply $[a]$ if no ambiguity arises.

The reader should check carefully that $\mathbb{Z}/n\mathbb{Z}$ consists of exactly n elements, namely

$$[0]_n, \quad [1]_n \quad , \cdots , \quad [n].$$

We can use the group structure on $\mathbb{Z}$ to induce an (abelian) group structure on $\mathbb{Z}/n\mathbb{Z}$. In order to do this, we define an operation $+$ on $\mathbb{Z}/n\mathbb{Z}$, by setting $\forall a, b \in \mathbb{Z}$

$$[a] + [b] := [a + b].$$

Of course we have to check that this prescription is well-defined; luckily, this is very easy: the following small lemma does the job, as it shows that the result of the operation does not depend on the representatives chosen for the classes.

Lemma 2.2. *If* $a \equiv a' \bmod n$ *and* $b \equiv b' \bmod n$, *then*

$$(a + b) \equiv (a' + b') \mod n.$$

Proof. By hypothesis $n \mid (a' - a)$ and $n \mid (b' - b)$; therefore $\exists k, \ell \in \mathbb{Z}$ such that

$$(a' - a) = kn, \quad (b' - b) = \ell n.$$

Then

$$(a' + b') - (a + b) = (a' - a) + (b' - b) = kn + \ell n = (k + \ell)n,$$

proving that n divides $(a' + b') - (a + b)$, as needed. □

Therefore, we have a binary operation $+$ on $\mathbb{Z}/n\mathbb{Z}$. It is immediately checked that the resulting structure is a group. Associativity is inherited from $\mathbb{Z}$:

$$([a] + [b]) + [c] = [a + b] + [c] = [(a + b) + c] = [a + (b + c)] = [a] + [b + c] = [a] + ([b] + [c]);$$

and so are the identity $[0]$ and 'inverse' $-[a] = [-a]$.

It is also immediately checked that the resulting groups $\mathbb{Z}/n\mathbb{Z}$ are commutative, as the abelian-style notation suggests:

$$[a] + [b] = [a + b] = [b + a] = [b] + [a].$$

[12] The notation $n \mid m$ stands for n *is a divisor of* m; that is, $m = nk$ for some integer k.

[13] When $n = p$ is prime, $\mathbb{Z}_p$ is the official notation for 'p-adic integers', which are a completely different concept; see Exercise VIII.1.19.

We trust that this material is not new to the reader, who should in any case check all these assertions carefully.

The abelian groups thus obtained, together with $\mathbb{Z}$, are called *cyclic groups;* a popular alternative notation for the group $(\mathbb{Z}/n\mathbb{Z}, +)$ is C_n. This is adopted especially when one wants to use the 'multiplicative' rather than 'additive' notation; thus we can say that C_n *is generated by one element* x*, with the relation* $x^n = e$.

Cyclic groups are tremendously important, and we will come back to them in later sections. For the time being we record the fact that the element

$$[1]_n \in \mathbb{Z}/n\mathbb{Z}$$

generates the group, in the sense that every other element may be obtained as a multiple of this element. For example, if $m \geq 0$ is an integer, then

$$[m]_n = \underbrace{[1 + \cdots + 1]_n}_{m \text{ times}} = \underbrace{[1]_n + \cdots + [1]_n}_{m \text{ times}} = m \cdot [1]_n.$$

Equivalently, we may phrase this fact by observing that the order of $[1]_n$ in $\mathbb{Z}/n\mathbb{Z}$ is n: this implies that the n multiples $0 \cdot [1]_n$, $1 \cdot [1]_n$, ..., $(n-1) \cdot [1]_n$ must all be distinct, and hence they must fill up $\mathbb{Z}/n\mathbb{Z}$.

Proposition 2.3. *The order of* $[m]_n$ *in* $\mathbb{Z}/n\mathbb{Z}$ *is* 1 *if* $n \mid m$*, and more generally*

$$|[m]_n| = \frac{n}{\gcd(m, n)}.$$

Proof. If $n \mid m$, then $[m]_n = [0]_n$. If n does not divide m, observe again that $[m]_n = m\,[1]_n$ and apply Proposition 1.13. □

Remark 2.4. As a consequence, the order of every element of $\mathbb{Z}/n\mathbb{Z}$ divides $n = |\mathbb{Z}/n\mathbb{Z}|$, the order of the group. We will see later (Example 8.15) that this is a general features of the order of elements in any finite group. ⌟

Corollary 2.5. *The class* $[m]_n$ *generates* $\mathbb{Z}/n\mathbb{Z}$ *if and only if* $\gcd(m, n) = 1$.

This simple result is quite important. For example, if $n = p$ is a prime integer, it shows that every *nonzero* class in the group $\mathbb{Z}/p\mathbb{Z}$ generates it. In any case, it allows us to construct more examples of interesting groups. The reader should check (or recall; cf. Exercise 2.14) that there also is a well-defined *multiplication* on $\mathbb{Z}/n\mathbb{Z}$, given by

$$[a]_n \cdot [b]_n := [ab]_n.$$

This operation does *not* define a group structure on $\mathbb{Z}/n\mathbb{Z}$: indeed, the class $[0]_n$ does *not* have a multiplicative inverse. On the other hand, for any positive n denote by $(\mathbb{Z}/n\mathbb{Z})^*$ the subset of $\mathbb{Z}/n\mathbb{Z}$ consisting of classes $[m]_n$ such that $\gcd(m, n) = 1$:

$$(\mathbb{Z}/n\mathbb{Z})^* := \{[m]_n \in \mathbb{Z}/n\mathbb{Z} \mid \gcd(m, n) = 1\}.$$

This subset is clearly well-defined: if $m \equiv m' \bmod n$, then $\gcd(m, n) = 1 \iff \gcd(m', n) = 1$ (Exercise 2.17), so the defining property of classes in $(\mathbb{Z}/n\mathbb{Z})^*$ is independent of representatives.

Proposition 2.6. *Multiplication makes* $(\mathbb{Z}/n\mathbb{Z})^*$ *into a group.*

Proof. Simple properties of gcd's show that if $\gcd(m_1, n) = \gcd(m_2, n) = 1$, then $\gcd(m_1 m_2, n) = 1$. (For example, if a prime integer divided both n and $m_1 m_2$, then it would necessarily divide m_1 or m_2, and one of the two gcd's would not be 1.) Therefore, the product of two elements in $(\mathbb{Z}/n\mathbb{Z})^*$ is an element of $(\mathbb{Z}/n\mathbb{Z})^*$, and $\cdot$ does define a binary operation

$$(\mathbb{Z}/n\mathbb{Z})^* \times (\mathbb{Z}/n\mathbb{Z})^* \to (\mathbb{Z}/n\mathbb{Z})^*.$$

It is clear that this operation is associative (because multiplication is associative in $\mathbb{Z}$); $[1]_n$ is an element of $(\mathbb{Z}/n\mathbb{Z})^*$ and is an identity with respect to multiplication; so all we have to check is that elements of $(\mathbb{Z}/n\mathbb{Z})^*$ have multiplicative inverses in $(\mathbb{Z}/n\mathbb{Z})^*$.

This follows from Corollary 2.5. If $\gcd(m, n) = 1$, then $[m]_n$ generates the *additive* group $\mathbb{Z}/n\mathbb{Z}$, and hence some multiple of $[m]_n$ must equal $[1]_n$:

$$(\exists a \in \mathbb{Z}): \quad a \cdot [m]_n = [1]_n;$$

this implies

$$[a]_n [m]_n = [1]_n.$$

Therefore $[m]_n$ does have a multiplicative inverse in $\mathbb{Z}/n\mathbb{Z}$, namely $[a]_n$. The reader will verify that $\gcd(a, n) = 1$, completing the proof. □

For instance, $[8]_{15}$ has a multiplicative inverse in $(\mathbb{Z}/15\mathbb{Z})^*$. Tracing the argument given in the proof,

$$2 \cdot 8 + (-1) \cdot 15 = 1,$$

and hence $[2]_{15} \cdot [8]_{15} = [1]_{15}$: the multiplicative inverse of $[8]_{15}$ is $[2]_{15}$.

For $n = p$ a positive prime integer, the group $((\mathbb{Z}/p\mathbb{Z})^*, \cdot)$ has order $(p - 1)$. We will have more to say about these groups in later sections (cf. Example 4.6).

Exercises

2.1. ¬ One can associate an $n \times n$ matrix M_σ with a permutation $\sigma \in S_n$ by letting the entry at $(i, \sigma(i))$ be 1 and letting all other entries be 0. For example, the matrix corresponding to the permutation

$$\sigma = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \in S_3$$

would be

$$M_\sigma = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Prove that, with this notation,

$$M_{\sigma\tau} = M_\sigma M_\tau$$

for all $\sigma, \tau \in S_n$, where the product on the right is the ordinary product of matrices. [IV.4.13]

2.2. ▷ Prove that if $d \leq n$, then S_n contains elements of order d. [§2.1]

2.3. For every positive integer n find an element of order n in $S_{\mathbb{N}}$.

2.4. Define a homomorphism $D_8 \to S_4$ by labeling vertices of a square, as we did for a triangle in §2.2. List the 8 permutations in the image of this homomorphism.

2.5. ▷ Describe generators and relations for all dihedral groups D_{2n}. (Hint: Let x be the reflection about a line through the center of a regular n-gon and a vertex, and let y be the counterclockwise rotation by $2\pi/n$. The group D_{2n} will be generated by x and y, subject to three relations[14]. To see that these relations really determine D_{2n}, use them to show that any product $x^{i_1}y^{i_2}x^{i_3}y^{i_4}\cdots$ equals $x^i y^j$ for some i, j with $0 \le i \le 1$, $0 \le j < n$.) [8.4, §IV.2.5]

2.6. ▷ For every positive integer n construct a group containing two elements g, h such that $|g| = 2$, $|h| = 2$, and $|gh| = n$. (Hint: For $n > 1$, D_{2n} will do.) [§1.6]

2.7. Find all elements of D_{2n} that commute with every other element. (The parity of n plays a role.)

2.8. Find the orders of the groups of symmetries of the five 'platonic solids'.

2.9. Verify carefully that 'congruence mod n' is an equivalence relation.

2.10. Prove that $\mathbb{Z}/n\mathbb{Z}$ consists of precisely n elements.

2.11. ▷ Prove that the square of every odd integer is congruent to 1 modulo 8. [§VII.5.1]

2.12. Prove that there are no integers a, b, c such that $a^2 + b^2 = 3c^2$. (Hint: By studying the equation $[a]_4^2 + [b]_4^2 = 3[c]_4^2$ in $\mathbb{Z}/4\mathbb{Z}$, show that a, b, c would all have to be even. Letting $a = 2k$, $b = 2\ell$, $c = 2m$, you would have $k^2 + \ell^2 = 3m^2$. What's wrong with that?)

2.13. ▷ Prove that if $\gcd(m, n) = 1$, then there exist integers a and b such that

$$am + bn = 1.$$

(Use Corollary 2.5.) Conversely, prove that if $am + bn = 1$ for some integers a and b, then $\gcd(m, n) = 1$. [2.15, §V.2.1, V.2.4]

2.14. ▷ State and prove an analog of Lemma 2.2, showing that the multiplication on $\mathbb{Z}/n\mathbb{Z}$ is a well-defined operation. [§2.3, §III.1.2]

2.15. ¬ Let $n > 0$ be an odd integer.

- Prove that if $\gcd(m, n) = 1$, then $\gcd(2m + n, 2n) = 1$. (Use Exercise 2.13.)
- Prove that if $\gcd(r, 2n) = 1$, then $\gcd(\frac{r+n}{2}, n) = 1$. (Ditto.)
- Conclude that the function $[m]_n \to [2m + n]_{2n}$ is a bijection between $(\mathbb{Z}/n\mathbb{Z})^*$ and $(\mathbb{Z}/2n\mathbb{Z})^*$.

[14] Two relations are evident. To 'see' the third one, hold your right hand in front of and away from you, pointing your fingers at the vertices of an imaginary regular pentagon. Flip the pentagon by turning the hand toward you; rotate it counterclockwise w.r.t. the line of sight by 72°; flip it again by pointing it away from you; and rotate it counterclockwise a second time. This returns the hand to the initial position. What does this tell you?

The number $\phi(n)$ of elements of $(\mathbb{Z}/n\mathbb{Z})^*$ is *Euler's ϕ-function.* The reader has just proved that if n is odd, then $\Phi(2n) = \Phi(n)$. Much more general formulas will be given later on (cf. Exercise V.6.8). [VII.5.11]

2.16. Find the last digit of $1238237^{18238456}$. (Work in $\mathbb{Z}/10\mathbb{Z}$.)

2.17. ▷ Show that if $m \equiv m' \bmod n$, then $\gcd(m,n) = 1$ if and only if $\gcd(m',n) = 1$. [§2.3]

2.18. For $d \leq n$, define an injective function $\mathbb{Z}/d\mathbb{Z} \to S_n$ preserving the operation, that is, such that the sum of equivalence classes in $\mathbb{Z}/n\mathbb{Z}$ corresponds to the product of the corresponding permutations.

2.19. ▷ Both $(\mathbb{Z}/5\mathbb{Z})^*$ and $(\mathbb{Z}/12\mathbb{Z})^*$ consist of 4 elements. Write their multiplication tables, and prove that no re-ordering of the elements will make them match. (Cf. Exercise 1.6.) [§4.3]

---

3. The category Grp

Groups will be the objects of the category Grp. In this section we define the *morphisms* in the category and deal with simple properties of these morphisms.

3.1. Group homomorphisms. As we know, a group consists of two distinct types of information: a set G and an operation[15]

$$m_G : G \times G \to G$$

satisfying certain properties. For two groups (G, m_G) and (H, m_H), a *group homomorphism*

$$\varphi : (G, m_G) \to (H, m_H)$$

is first of all a function (usually given the same name, φ in this case) between the underlying sets; but this function must 'know about' the operations m_G on G, m_H on H. What is the most natural requirement of this sort?

Note that the set-function $\varphi : G \to H$ determines a function

$$(\varphi \times \varphi) : G \times G \to H \times H :$$

we could invoke the universal property of products to obtain this function (cf. Exercise 3.1), but since we are dealing with sets, there is no need for fancy language here—just define the function by

$$(\forall (a,b) \in G \times G) : \quad (\varphi \times \varphi)(a,b) = (\varphi(a), \varphi(b)).$$

There is an diagram combining all these maps:

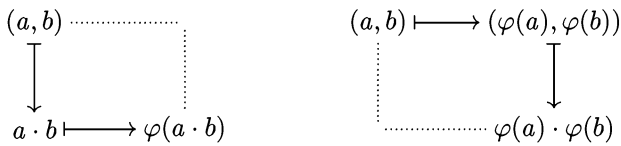
$$\begin{array}{ccc} G \times G & \xrightarrow{\varphi \times \varphi} & H \times H \\ {\scriptstyle m_G}\downarrow & & \downarrow{\scriptstyle m_H} \\ G & \xrightarrow[\varphi]{} & H \end{array}$$

[15] In §1, m_G was denoted $\bullet$; here we need to keep track of operations on different groups, so for a moment we will use a symbol recording the group (and evoking 'multiplication').

What requirement could be more natural than asking that this diagram *commute?*

Definition 3.1. The set-function $\varphi : G \to H$ defines a *group homomorphism* if the diagram displayed above commutes. ⌟

This is a seemingly complicated way of saying something simple: since φ and m_G, m_H are functions of sets, commutativity means the following. For all $a, b \in G$, the two ways to travel through the diagram give



where we now write $\cdot$ for both operations: in G on the left, in H on the right. Commutativity of the diagram means that we must get the same result in both cases; therefore, Definition 3.1 can be rephrased as

the set-function $\varphi : G \to H$ *is a* group homomorphism *if* $\forall a, b \in G$

$$\varphi(a \cdot b) = \varphi(a) \cdot \varphi(b).$$

In other words, φ is a homomorphism if it 'preserves the structure'. This may sound more familiar to our reader. As usual, the reason to bring in diagrams (as in Definition 3.1) is that this would make it easy to transfer part of the discussion to other categories.

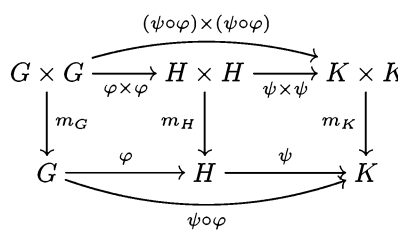
If the context is clear, one may simply write 'homomorphism', omitting the qualifier 'group'.

3.2. Grp: Definition. For G, H groups[16] we define

$$\mathrm{Hom}_{\mathsf{Grp}}(G, H)$$

to be the set of group homomorphisms $G \to H$.

If G, H, K are groups and $\varphi : G \to H$, $\psi : H \to K$ are two group homomorphisms, it is easy to check that the composition $\psi \circ \varphi : G \to K$ is a group homomorphism: from the diagram point of view, this amounts to observing that the 'outer rectangle' in



[16] We are yielding to the usual abuse of language that lets us omit explicit mention of the operation.

must commute if the two 'inner rectangles' commute. For arbitrary $a, b \in G$, this means,

$$\begin{aligned}(\psi \circ \varphi)(a \cdot b) &= \psi(\varphi(a \cdot b)) \overset{(1)}{=} \psi(\varphi(a) \cdot \varphi(b)) \overset{(2)}{=} \psi(\varphi(a)) \cdot \psi(\varphi(b)) \\ &= (\psi \circ \varphi)(a) \cdot (\psi \circ \varphi)(b)\end{aligned}$$

where (1) holds because φ is a homomorphism and (2) holds because ψ is a homomorphism.

Further, it is clear that composition is associative (because it is for set-functions) and that the identity function $\mathrm{id}_G : G \to G$ is a group homomorphism. Therefore, Grp is indeed a category.

3.3. Pause for reflection. The careful reader might raise an objection: the group axioms prescribe the existence of an identity element e_G and of an 'inverse', that is, a specific function

$$\iota_G : G \to G, \quad \iota(g) := g^{-1}.$$

Shouldn't the definition of morphism in Grp keep track of this type of data? The definition we have given only keeps track of the multiplication map m_G.

The reason why we can get away with this is that preserving m automatically preserves e and ι:

Proposition 3.2. *Let $\varphi : G \to H$ be a group homomorphism. Then*

- $\varphi(e_G) = e_H$;
- $\forall g \in G$, $\varphi(g^{-1}) = \varphi(g)^{-1}$.

In terms of diagrams, the second assertion amounts to saying that

$$\begin{array}{ccc} G & \xrightarrow{\varphi} & H \\ {\scriptstyle \iota_G}\downarrow & & \downarrow{\scriptstyle \iota_H} \\ G & \xrightarrow[\varphi]{} & H \end{array}$$

must commute.

Proof. The first item follows from the definition of homomorphism and cancellation: since $e_H = e_H \cdot e_H$,

$$e_H \cdot \varphi(e_G) = \varphi(e_G) = \varphi(e_G \cdot e_G) = \varphi(e_G) \cdot \varphi(e_G),$$

which implies $e_H = \varphi(e_G)$ by 'cancelling $\varphi(e_G)$'.

The proof of the second assertion is similar: $\forall g \in G$,

$$\varphi(g^{-1}) \cdot \varphi(g) = \varphi(g^{-1} \cdot g) = \varphi(e_G) = e_H = \varphi(g)^{-1} \cdot \varphi(g),$$

implying $\varphi(g^{-1}) = \varphi(g)^{-1}$ by cancellation. □

3.4. Products et al. The categories Grp and Set look rather alike at first: given a group, we can 'forget' the information of multiplication and we are left with a set; given a group homomorphism, we can forget that it preserves the multiplication and we are left with a set-function. A concise way to express this fact is that there is a 'functor' Grp $\rightsquigarrow$ Set (called in fact a 'forgetful' functor); we will deal with functors more extensively much later on, starting in Chapter VIII.

However, there are important differences between these two categories. For example, recall that Set has a unique initial object (that is, $\emptyset$) and this is not the same as the final objects (that is, the singletons). Also recall that a *trivial* group is a group consisting of a single element (Example 1.3).

Proposition 3.3. *Trivial groups are both* initial *and* final *in* Grp.

This makes trivial groups 'zero-objects' of the category Grp.

Proof. It should be clear that trivial groups are final: there is only one function from a set to a singleton, that is, the constant function; this is vacuously a group homomorphism.

To see that trivial groups are initial, let $T = \{e\}$ be a trivial group; for any group G, define $\varphi : T \to G$ by $T(e) = e_G$. This is clearly a group homomorphism, and it is the only possible one since every group homomorphism must send the identity to the identity (Proposition 3.2). □

Here is a similarity: Grp has products; in fact, the product of two groups G, H is supported on the product $G \times H$ of the underlying sets.

To see this, we need to define a multiplication on $G \times H$; the catchword here is *componentwise:* define the operation in $G \times H$ by performing the operation on each component separately. Explicitly, define $\forall g_1, g_2 \in G$, $\forall h_1, h_2 \in H$

$$(g_1, h_1) \cdot (g_2, h_2) := (g_1 g_2, h_1 h_2).$$

This operation defines a group structure on $G \times H$: the operation is associative, the identity is (e_G, e_H), and the inverse of (g, h) is (g^{-1}, h^{-1}). All needed verifications are left to the reader. The group $G \times H$ is called the *direct product* of the groups G and H.

Also note that the natural projections

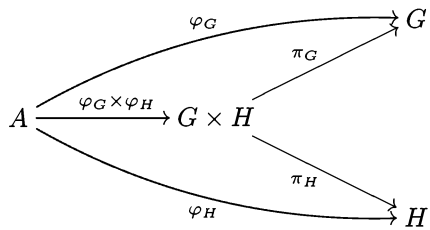
$$\begin{array}{ccc} & G \times H & \\ {\scriptstyle \pi_G}\swarrow & & \searrow{\scriptstyle \pi_H} \\ G & & H \end{array}$$

(defined as set-functions as in §I.2.4) are group homomorphisms: again, this follows immediately from the definitions.

Proposition 3.4. *With operation defined componentwise, $G \times H$ is a product in* Grp.

Proof. Recall (§I.5.4) that this means that $G \times H$ satisfies the following universal property: for any group A and any choice of group homomorphisms $\varphi_G : A \to G$,

$\varphi_H : A \to H$, there exists a unique group homomorphism $\varphi_G \times \varphi_H$ making the diagram



commute.

Now, a unique *set-function* $\varphi_G \times \varphi_H$ exists making the diagram commute, because the *set* $G \times H$ is a product of G and H in Set. So we only need to check that $\varphi_G \times \varphi_H$ is a group homomorphism, and this is immediate (if cumbersome): $\forall a, b \in A$,

$$\begin{aligned}\varphi_G \times \varphi_H(ab) &= (\varphi_G(ab), \varphi_H(ab)) = (\varphi_G(a)\varphi_G(b), \varphi_H(a)\varphi_H(b)) \\ &= (\varphi_G(a), \varphi_H(a))(\varphi_G(b), \varphi_H(b)) = (\varphi_G \times \varphi_H(a))(\varphi_G \times \varphi_H(b)).\end{aligned}$$

□

What about *co*products? They *do* exist in Grp, but their construction requires handling presentations more proficiently than we do right now, and general coproducts of groups will not be used in the rest of the book; so the reader will have to deal with them on his or her own. For an important example, see Exercise 3.8; more will show up in Exercises 5.6 and 5.7, since *free groups* are themselves particular cases of coproducts. The reader will finally produce the coproduct of any two groups explicitly in Exercise 8.7. For now, just realize that the *disjoint union*, which works as a coproduct in Set (Proposition I.5.6), is not an option in Grp: there is no reasonable group structure on the disjoint union. The coproduct of G and H in Grp is denoted $G * H$ and is called the *free product* of G and H.

3.5. Abelian groups. The category Ab whose objects are *abelian* groups, and whose morphisms are group homomorphisms, will in a sense be more important for us than the category Grp. In many ways, as we will see, Ab is a 'nicer' category[17] than Grp. Again the trivial groups are both initial and final (that is, 'zero') objects; products exist and coincide with products in Grp. But here is a difference: unlike in Grp, coproducts in Ab coincide with products. That is, if G and H are abelian groups, then the product $G \times H$ (with the two natural homomorphisms $G \to G \times H$, $H \to G \times H$) satisfies the universal property for coproducts in Ab (cf. Exercise 3.3). When working as a coproduct, the product $G \times H$ of two abelian groups is often called their *direct sum* and is denoted $G \oplus H$.

There is a pretty subtlety here, which may highlight the power of the language: even if G and H are commutative, the product $G \times H$ does *not* (necessarily) satisfy

[17] As we will see in due time (Proposition III.5.3), Ab is one instance of a general class of categories of 'modules over a commutative ring R' (for $R = \mathbb{Z}$). Unlike Grp, these categories are *abelian*, which makes them very well-behaved. We will learn two or three things about abelian categories in Chapter IX.

the universal property for coproducts *in* Grp, even if it does *in* Ab. For an explicit example, see Exercise 3.6.

Exercises

3.1. ▷ Let $\varphi : G \to H$ be a morphism in a category C with products. Explain why there is a unique morphism

$$(\varphi \times \varphi) : G \times G \to H \times H.$$

(This morphism is defined explicitly for C = Set in §3.1.) [§3.1, 3.2]

3.2. Let $\varphi : G \to H$, $\psi : H \to K$ be morphisms in a category with products, and consider morphisms between the products $G \times G$, $H \times H$, $K \times K$ as in Exercise 3.1. Prove that

$$(\psi\varphi) \times (\psi\varphi) = (\psi \times \psi)(\varphi \times \varphi).$$

(This is part of the commutativity of the diagram displayed in §3.2.)

3.3. ▷ Show that if G, H are *abelian* groups, then $G \times H$ satisfies the universal property for coproducts in Ab (cf. §I.5.5). [§3.5, 3.6, §III.6.1]

3.4. Let G, H be groups, and assume that $G \cong H \times G$. Can you conclude that H is trivial? (Hint: No. Can you construct a counterexample?)

3.5. Prove that $\mathbb{Q}$ is not the direct product of two nontrivial groups.

3.6. ▷ Consider the product of the cyclic groups C_2, C_3 (cf. §2.3): $C_2 \times C_3$. By Exercise 3.3, this group is a coproduct of C_2 and C_3 in Ab. Show that it is *not* a coproduct of C_2 and C_3 in Grp, as follows:

- find injective homomorphisms $C_2 \to S_3$, $C_3 \to S_3$;
- arguing by contradiction, assume that $C_2 \times C_3$ is a coproduct of C_2, C_3, and deduce that there would be a group homomorphism $C_2 \times C_3 \to S_3$ with certain properties;
- show that there is no such homomorphism.

[§3.5]

3.7. Show that there is a *surjective* homomorphism $\mathbb{Z} * \mathbb{Z} \to C_2 * C_3$. ($*$ denotes coproduct in Grp; cf. §3.4.)

One can think of $\mathbb{Z} * \mathbb{Z}$ as a group with two generators x, y, subject to no relations whatsoever. (We will study a general version of such groups in §5; see Exercise 5.6.)

3.8. ▷ Define a group G with two generators x, y, subject (only) to the relations $x^2 = e_G$, $y^3 = e_G$. Prove that G is a coproduct of C_2 and C_3 *in* Grp. (The reader will obtain an even more concrete description for $C_2 * C_3$ in Exercise 9.14; it is called the *modular group*.) [§3.4, 9.14]

3.9. Show that *fiber* products and coproducts exist in Ab. (Cf. Exercise I.5.12.)

4. Group homomorphisms

4.1. Examples. For any two groups G, H, the set $\mathrm{Hom}_{\mathsf{Grp}}(G,H)$ is certainly *nonempty:* at the very least we can define a homomorphism $G \to H$ by sending every element of G to the identity e_H of H. Thus, $\mathrm{Hom}_{\mathsf{Grp}}(G,H)$ is a 'pointed set' (cf. Example I.3.8).

There is a purely categorical way to think about this: since Grp has zero-objects $\{*\}$ (recall that trivial groups are both initial and final in Grp; cf. Proposition 3.3), there are unique morphisms

$$G \to \{*\}, \quad \{*\} \to H$$

in Grp; their composition is the distinguished element of $\mathrm{Hom}_{\mathsf{Grp}}(G,H)$ mentioned above; we will call this element the *trivial* morphism.

More 'meaningful' examples can be constructed by considering the groups encountered in §2: for instance, the function $D_6 \to S_3$ defined in §2.2 is a homomorphism. Such examples will likely all be instances of the notion of *group action.* In general, an 'action' of a group G on an object A of a category C is a homomorphism

$$G \to \mathrm{Aut}_{\mathsf{C}}(A);$$

that is, a group G 'acts' on an object if the elements of G determine isomorphisms of that object to itself, in a way compatible with compositions. If $\mathsf{C} = \mathsf{Set}$, this means that elements of G determine specific permutations of the set A. For example, the symmetries of an equilateral triangle (that is, elements of D_6) determine permutations of the vertices of that triangle (that is, permutations of a set with three elements), and they do so compatibly with composition; this is what gives us homomorphisms $D_6 \to S_3$. We say that D_6 *'acts on the set of vertices'* of the triangle.

The reader can (and should) construct many more examples of this kind. We will have much more to say about actions of groups and other algebraic entities in later sections.

Here is an example with a different flavor: the exponential function is a homomorphism from $(\mathbb{R},+)$ to the group $(\mathbb{R}^{>0},\cdot)$ of positive real numbers, with ordinary multiplication as operation. Indeed, $e^{a+b} = e^a e^b$. A similar (and very important) class of examples may be obtained as follows: let G be any group and $g \in G$ any element of G; define an 'exponential map' $\epsilon_g : \mathbb{Z} \to G$ by

$$(\forall a \in \mathbb{Z}): \quad \epsilon_g(a) := g^a.$$

Then ϵ_g is (clearly) a group homomorphism. The element g generates G if and only if ϵ_g is surjective.

One concrete instance of this homomorphism (in the abelian environment, thus using multiples rather than powers) is the 'quotient' function $\pi_n : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$,

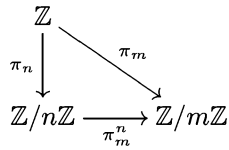
$$a \mapsto a \cdot [1]_n = [a]_n :$$

with the notation introduced above, this is $\epsilon_{[1]_n}$. This function is surjective; hence $[1]_n$ generates $\mathbb{Z}/n\mathbb{Z}$. In fact, as observed in §2.3 (Corollary 2.5), $[m]_n$ generates $\mathbb{Z}/n\mathbb{Z}$ if and only if $\gcd(m,n) = 1$.

If $m \mid n$, there is a homomorphism

$$\pi_m^n : \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z}$$

making the diagram



commute: that is,

$$\pi_m^n([a]_n) = [a]_m;$$

the reader should check carefully that this function is well-defined (Exercise 4.1).

If m_1 and m_2 are both divisors of n, we have homomorphisms $\pi_{m_1}^n$, $\pi_{m_2}^n$ from $\mathbb{Z}/n\mathbb{Z}$ to both $\mathbb{Z}/m_1\mathbb{Z}$ and $\mathbb{Z}/m_2\mathbb{Z}$ and hence to their direct product. For instance, since $6 = 2 \cdot 3$, there is a homomorphism

$$\mathbb{Z}/6\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$$

(or, in 'multiplicative notation', $C_6 \to C_2 \times C_3$). Explicitly,

$$[0]_6 \mapsto ([0]_2, [0]_3), \quad [1]_6 \mapsto ([1]_2, [1]_3), \quad [2]_6 \mapsto ([0]_2, [2]_3),$$
$$[3]_6 \mapsto ([1]_2, [0]_3), \quad [4]_6 \mapsto ([0]_2, [1]_3), \quad [5]_6 \mapsto ([1]_2, [2]_3).$$

Note that this homomorphism is a *bijection;* as we will see in a moment (§4.3), this makes it an *isomorphism;* in particular, C_6 is *also* a product of C_2 and C_3 in Grp.

One can concoct a homomorphism $\mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z}$ also if $n \mid m$: for example, the function $\mathbb{Z}/2\mathbb{Z} \to \mathbb{Z}/4\mathbb{Z}$ defined by

$$[0]_2 \to [0]_4, \quad [1]_2 \to [2]_4$$

is clearly a group homomorphism. Unlike π_m^n, this homomorphism is not nicely compatible[18] with the homomorphisms π_n.

On the other hand, is there a nontrivial group homomorphism (for example) $C_4 \to C_7$? Note that there are $7^4 = 2{,}401$ *set-functions* from C_4 to C_7 (cf. Exercise I.2.10); the question is whether any of these functions (besides the trivial homomorphism sending everything to e) preserves the operation. We already know that a homomorphism must send the identity to the identity (Proposition 3.2), and that already rules out all but 216 functions (why?); still, it is unrealistic to write all of them out explicitly to see if any is a homomorphism.

The reader should think about this before we spill the beans in the next subsection.

[18] Also, note that while π_n^m preserves *multiplication* as well as sum, this new homomorphism does not; that is, it is not a 'ring homomorphism'. This is immediately visible in the given example: $[1]_2 \cdot [1]_2 = [1]_2$, but $[2]_4 \cdot [2]_4 = [0]_4$.

4.2. Homomorphisms and order. Group homomorphisms are set-functions preserving the group structure; as such, they must preserve many features of the theory. Proposition 3.2 is an instance of this principle: group homomorphisms must preserve identities and inverses. It is also clear that if $\varphi : G \to H$ is a group homomorphism and g is an element *of finite order* in G, then $\varphi(g)$ must be an element of finite order in H: indeed, if $g^n = e_G$ for some $n > 0$, then

$$\varphi(g)^n = \varphi(g^n) = \varphi(e_G) = e_H.$$

In fact, this observation establishes a more precise statement:

Proposition 4.1. *Let $\varphi : G \to H$ be a group homomorphism, and let $g \in G$ be an element of finite order. Then $|\varphi(g)|$ divides $|g|$.*

Proof. As observed, $\varphi(g)^{|g|} = e_H$; applying Lemma 1.10 gives the statement. □

Example 4.2. There are no nontrivial homomorphisms $\mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}$: indeed, the image of every element of $\mathbb{Z}/n\mathbb{Z}$ must have finite order, and the only element with finite order in $(\mathbb{Z}, +)$ is 0.

There are no nontrivial homomorphisms $\varphi : C_4 \to C_7$. Indeed, the orders of elements in C_4 divide 4 (cf. Proposition 2.3), and the orders of elements in C_7 divides 7. Thus, the order of each $\varphi(g)$ must divide *both* 4 and 7; this forces $|\varphi(g)| = 1$ for all g, that is, $\varphi(g) = e$ for all $g \in C_4$. ⌟

Of course the order itself is not preserved: for example, $1 \in \mathbb{Z}$ has infinite order, while $[1]_n = \pi_n(1) \in \mathbb{Z}/n\mathbb{Z}$ has order n (with notation as in §4.1). Order *is* preserved through isomorphisms, as we will see in a moment.

4.3. Isomorphisms. An isomorphism of groups $\varphi : G \to H$ is (of course) an isomorphism in Grp, that is, a group homomorphism admitting an inverse

$$\varphi^{-1} : H \to G$$

which is also a group homomorphism. Taking our cue from Set, if a homomorphism of groups is an isomorphism, then it must in particular be a bijection between the underlying sets. Luckily, the converse also holds:

Proposition 4.3. *Let $\varphi : G \to H$ be a group homomorphism. Then φ is an isomorphism of groups if and only if it is a bijection.*

Proof. One implication is immediate, as pointed out above. For the other implication, assume $\varphi : G \to H$ is a bijective group homomorphism. As a bijection, φ has an inverse in Set:

$$\varphi^{-1} : H \to G;$$

we simply need to check that this is a group homomorphism. Let h_1, h_2 be elements of H, and let $g_1 = \varphi^{-1}(h_1)$, $g_2 = \varphi^{-1}(h_2)$ be the corresponding elements of G. Then

$$\varphi^{-1}(h_1 \cdot h_2) = \varphi^{-1}(\varphi(g_1) \cdot \varphi(g_2)) = \varphi^{-1}(\varphi(g_1 \cdot g_2)) = g_1 \cdot g_2 = \varphi^{-1}(h_1) \cdot \varphi^{-1}(h_2)$$

as needed. □

Example 4.4. The function $D_6 \to S_3$ defined in §2.2 is an isomorphism of groups, since it is a bijective group homomorphism. So is the exponential function $(\mathbb{R}, +) \to (\mathbb{R}^{>0}, \cdot)$ mentioned in §4.1. If the exponential function $\epsilon_g : \mathbb{Z} \to G$ determined by an element $g \in G$ (as in §4.1) is an isomorphism, we say that G is an 'infinite cyclic' group.

The function $\pi_2^6 \times \pi_3^6 : C_6 \to C_2 \times C_3$ studied ('additively') in §4.1 is an isomorphism. ⌟

Definition 4.5. Two groups G, H are *isomorphic* if they are isomorphic in Grp in the sense of §I.4.1, that is (by Proposition 4.3), if there is a bijective group homomorphism $G \to H$. ⌟

We have observed once and for all in §I.4.1 that 'isomorphic' is automatically an equivalence relation. We write $G \cong H$ if G and H are isomorphic.

Automorphisms of a group G are isomorphisms $G \to G$; these form a group $\mathrm{Aut}_{\mathsf{Grp}}(G)$ (cf. §I.4.1), usually denoted $\mathrm{Aut}(G)$.

Example 4.6. We have introduced our template of *cyclic groups* in §2.3. The notion of isomorphism allows us to give a formal definition:

Definition 4.7. A group G is *cyclic* if it is isomorphic to $\mathbb{Z}$ or to $C_n = \mathbb{Z}/n\mathbb{Z}$ for some[19] n. ⌟

Thus, $C_2 \times C_3$ is cyclic, of order 6, since $C_2 \times C_3 \cong C_6$. More generally (Exercise 4.9) $C_m \times C_n$ is cyclic if $\gcd(m, n) = 1$.

The reader will check easily (Exercise 4.3) that a group of order n is cyclic if and only if it contains an element of order n.

There is a somewhat surprising source of cyclic groups: *if p is prime, the group $((\mathbb{Z}/p\mathbb{Z})^*, \cdot)$ is cyclic.* We will prove a more general statement when we have accumulated more machinery (Theorem IV.6.10), but the adventurous reader can already enjoy a proof by working out Exercise 4.11. This is a relatively deep fact; note that, for example, $(\mathbb{Z}/12\mathbb{Z})^*$ is *not* cyclic (cf. Exercise 2.19 and Exercise 4.10). The fact that $(\mathbb{Z}/p\mathbb{Z})^*$ is cyclic for p prime means that there must be integers a such that *every* nonmultiple of p is congruent to a power of a; the usual proofs of this fact are not *constructive*, that is, they do not explicitly produce an integer with this property. There is a very pretty connection between the order of an element of the cyclic group $(\mathbb{Z}/p\mathbb{Z})^*$ and the so-called 'cyclotomic polynomials'; but that will have to wait for a little field theory (cf. Exercise VII.5.15).

As we have seen, the groups D_6 and S_3 are isomorphic. Are C_6 and S_3 isomorphic? There are *46,656* functions between the sets C_6 and S_3, of which 720 are bijections and 120 are bijections preserving the identity. The reader is welcome to list all 120 and attempt to verify by hand if any of them is a homomorphism. But maybe there is a better strategy to answer such questions.... ⌟

Isomorphic objects of a category are essentially indistinguishable in that category. Thus, isomorphic *groups* share every group-theoretic structure. In particular,

[19] This includes the possibility that $n = 1$, that is, trivial groups are cyclic.

Proposition 4.8. *Let $\varphi : G \to H$ be an isomorphism.*

- $(\forall g \in G) : |\varphi(g)| = |g|$;
- *G is commutative if and only if H is commutative.*

Proof. The first assertion follows from Proposition 4.1: the order of $\varphi(g)$ divides the order of g, and on the other hand the order of $g = \varphi^{-1}(\varphi(g))$ must divide the order of $\varphi(g)$; thus the two orders must be equal.

The proof of the second assertion is left to the reader. □

Further instances of this principle will be assumed without explicit mention.

Example 4.9. $C_6 \ncong S_3$, since one is commutative and the other is not. Here is another reason: in C_6 there is 1 element of order one, 1 of order two, 2 of order three, and 2 of order six; in S_3 the situation is different: 1 element of order one, 3 of order two, 2 of order three. Thus, *none* of the 120 bijections $C_6 \to S_3$ preserving the identity is a group homomorphism.

Note: Two finite *commutative* groups are isomorphic *if and only if* they have the same number of elements of any given order, but we are not yet in a position to prove this; the reader will verify this fact in due time (Exercise IV.6.13). The commutativity hypothesis is necessary: there *do* exist pairs of nonisomorphic finite groups with the same number of elements of any given order (same exercise). ⌟

4.4. Homomorphisms of abelian groups. We have already mentioned that Ab is in some ways 'better behaved' than Grp, and we are ready to highlight another instance of this observation. As we have seen, $\mathrm{Hom}_{\mathsf{Grp}}(G, H)$ is a *pointed set* for any two groups G, H. In Ab, we can say much more: $\mathrm{Hom}_{\mathsf{Ab}}$ *is a* group *(in fact, an abelian group) for any two abelian groups G, H.*

The operation in $\mathrm{Hom}_{\mathsf{Ab}}(G, H)$ is 'inherited' from the operation in H: if $\varphi, \psi : G \to H$ are two group homomorphisms, let $\varphi + \psi$ be the function defined by

$$(\forall a \in G) : \quad (\varphi + \psi)(a) := \varphi(a) + \psi(a).$$

Is $\varphi + \psi$ a group homomorphism? Yes, because $\forall a, b \in G$

$$\begin{aligned}(\varphi + \psi)(a + b) &= \varphi(a + b) + \psi(a + b) = (\varphi(a) + \varphi(b)) + (\psi(a) + \psi(b)) \\ &\overset{!}{=} (\varphi(a) + \psi(a)) + (\varphi(b) + \psi(b)) = (\varphi + \psi)(a) + (\varphi + \psi)(b).\end{aligned}$$

Note that the equality marked by ! uses crucially the fact that H is commutative.

With this operation, $\mathrm{Hom}_{\mathsf{Ab}}(G, H)$ is clearly a group: the associativity of $+$ is inherited from that of the operation in H; the trivial homomorphism is the identity element, and the inverse[20] of $\varphi : G \to H$ is defined (not surprisingly) by

$$(\forall a \in G) : \quad (-\varphi)(a) = -\varphi(a).$$

In fact, note that these conclusions may be drawn as soon as H is commutative: $\mathrm{Hom}_{\mathsf{Grp}}(G, H)$ is a group if H is commutative (even if G is not). In fact, if H is

[20] Unfortunate clash of terminology! We mean the 'inverse' as in 'group inverse', not as a possible function $H \to G$.

a commutative group, then $H^A = \mathrm{Hom}_{\mathsf{Set}}(A, H)$ is a group for all sets A; we will come back to this group in §5.4.

Exercises

4.1. ▷ Check that the function π_m^n defined in §4.1 is well-defined and makes the diagram commute. Verify that it is a group homomorphism. Why is the hypothesis $m \mid n$ necessary? [§4.1]

4.2. Show that the homomorphism $\pi_2^4 \times \pi_2^4 : C_4 \to C_2 \times C_2$ is *not* an isomorphism. In fact, is there *any* nontrivial isomorphism $C_4 \to C_2 \times C_2$?

4.3. ▷ Prove that a group of order n is isomorphic to $\mathbb{Z}/n\mathbb{Z}$ if and only if it contains an element of order n. [§4.3]

4.4. Prove that no two of the groups $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$ are isomorphic to one another. Can you decide whether $(\mathbb{R}, +)$, $(\mathbb{C}, +)$ *are* isomorphic to one another? (Cf. Exercise VI.1.1.)

4.5. Prove that the groups $(\mathbb{R} \smallsetminus \{0\}, \cdot)$ and $(\mathbb{C} \smallsetminus \{0\}, \cdot)$ are not isomorphic.

4.6. We have seen that $(\mathbb{R}, +)$ and $(\mathbb{R}^{>0}, \cdot)$ are isomorphic (Example 4.4). Are the groups $(\mathbb{Q}, +)$ and $(\mathbb{Q}^{>0}, \cdot)$ isomorphic?

4.7. Let G be a group. Prove that the function $G \to G$ defined by $g \mapsto g^{-1}$ is a homomorphism if and only if G is abelian. Prove that $g \mapsto g^2$ is a homomorphism if and only if G is abelian.

4.8. ¬ Let G be a group, and let $g \in G$. Prove that the function $\gamma_g : G \to G$ defined by $(\forall a \in G) : \gamma_g(a) = gag^{-1}$ is an automorphism of G. (The automorphisms γ_g are called 'inner' automorphisms of G.) Prove that the function $G \to \mathrm{Aut}(G)$ defined by $g \mapsto \gamma_g$ is a homomorphism. Prove that this homomorphism is trivial if and only if G is abelian. [6.7, 7.11, IV.1.5]

4.9. ▷ Prove that if m, n are positive integers such that $\gcd(m, n) = 1$, then $C_{mn} \cong C_m \times C_n$. [§4.3, 4.10, §IV.6.1, V.6.8]

4.10. ▷ Let $p \neq q$ be odd prime integers; show that $(\mathbb{Z}/pq\mathbb{Z})^*$ is not cyclic. (Hint: Use Exercise 4.9 to compute the order N of $(\mathbb{Z}/pq\mathbb{Z})^*$, and show that no element can have order N.) [§4.3]

4.11. ▷ In due time we will prove the easy fact that if p is a prime integer, then the equation $x^d = 1$ can have at most d solutions in $\mathbb{Z}/p\mathbb{Z}$. Assume this fact, and prove that the multiplicative group $G = (\mathbb{Z}/p\mathbb{Z})^*$ is cyclic. (Hint: Let $g \in G$ be an element of maximal order; use Exercise 1.15 to show that $h^{|g|} = 1$ for all $h \in G$. Therefore....) [§4.3, 4.15, 4.16, §IV.6.3]

4.12. ¬ • Compute the order of $[9]_{31}$ in the group $(\mathbb{Z}/31\mathbb{Z})^*$.

• Does the equation $x^3 - 9 = 0$ have solutions in $\mathbb{Z}/31\mathbb{Z}$? (Hint: Plugging in all 31 elements of $\mathbb{Z}/31\mathbb{Z}$ is too laborious and will not teach you much. Instead,

use the result of the first part: if c is a solution of the equation, what can you say about $|c|$?) [VII.5.15]

4.13. ¬ Prove that $\mathrm{Aut}_{\mathsf{Grp}}(\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}) \cong S_3$. [IV.5.14]

4.14. ▷ Prove that the order of the group of automorphisms of a cyclic group C_n is the number of positive integers $r < n$ that are *relatively prime* to n. (This is called *Euler's ϕ-function;* cf. Exercise 6.14.) [§IV.1.4, IV.1.22, §IV.2.5]

4.15. ¬ Compute the group of automorphisms of $(\mathbb{Z}, +)$. Prove that if p is prime, then $\mathrm{Aut}_{\mathsf{Grp}}(C_p) \cong C_{p-1}$. (Use Exercise 4.11.) [IV.5.12]

4.16. ¬ Prove *Wilson's theorem: a positive integer p is prime if and only if*

$$(p-1)! \equiv -1 \mod p.$$

(For one direction, use Exercises 1.8 and 4.11. For the other, assume d is a proper divisor of p, and note that d divides $(p-1)!$; therefore....) [IV.4.11]

4.17. For a few small (but not too small) primes p, find a generator of $(\mathbb{Z}/p\mathbb{Z})^*$.

4.18. Prove the second part of Proposition 4.8.

---

5. Free groups

5.1. Motivation. Having become more familiar with homomorphisms, we can now contemplate one fancier example of a group. The motivation underlying this new construction may be summarized as follows: given a set A, whose elements have no special 'group-theoretic' property, we want to construct a group $F(A)$ containing A 'in the most efficient way'.

For example, if $A = \emptyset$, then a trivial group will do. If $A = \{a\}$ is a singleton, then a trivial group will *not* do: because although a trivial group $\{a\}$ would itself be a singleton, that one element a in it would have to be the identity, and that is certainly a very special group-theoretic property. Instead, we construct an infinite cyclic group $\langle a \rangle$ whose elements are 'formal powers' a^n, $n \in \mathbb{Z}$, and we identify a with the power a^1:

$$\langle a \rangle := \{\cdots, a^{-2}, a^{-1}, a^0 = e, a^1 = a, a^2, a^3, \cdots\};$$

we take all these powers to be distinct and define multiplication in the evident way—so that the exponential map

$$\epsilon_a : \mathbb{Z} \to \langle a \rangle, \quad \epsilon_a(n) := a^n$$

is an isomorphism. The fact that 'all powers are distinct' is the formal way to implement the fact that there is nothing special about a: in the group $F(\{a\}) = \langle a \rangle$, a obeys no condition other than the inevitable $a^0 = e$.

Summarizing: if A is a singleton, then we may take $F(A)$ to be an infinite cyclic group.

The task is to formalize the heuristic motivation given above and construct a group $F(A)$ for *every* set A. As we often do, we will now ask the reader to put away this book and to try to figure out on his or her own what this may mean and how it may be accomplished.

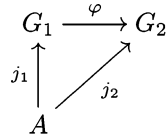
5.2. Universal property. Hoping that the reader has now acquired an individual viewpoint on the issue, here is the standard answer: the heuristic motivation is formalized by means of a suitable universal property. Given a set A, our group $F(A)$ will have to 'contain' A; therefore it is natural to consider the category $\mathscr{F}^A$ whose objects are pairs (j, G), where G is a group and

$$j : A \to G$$

is a set-function[21] from A to G and morphisms

$$(j_1, G_1) \to (j_2, G_2)$$

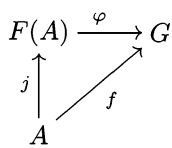
are *commutative* diagrams of set-functions



in which φ is required to be a *group homomorphism.*

The reader will be reminded of the categories we considered in Example I.3.7: the only difference here is that we are mixing objects and morphisms of one category (that is, Grp) with objects and morphisms of *another* (related) category (that is, Set). The fact that we are considering all possible functions $A \to G$ is a way to implement the fact that we have no *a priori* group-theoretic information about A: we do not want to put any restriction on what may happen to the elements of A once they are mapped to a group G; hence we consider all possibilities at once.

A *free group* $F(A)$ on A will be (the group component of) an *initial* object in $\mathscr{F}^A$. This choice implements the fact that A should map to $F(A)$ in the 'most efficient way': any other way to map A to a group can be reconstructed from this one, by composing with a group homomorphism. In the language of universal properties, we can state this as follows: $F(A)$ is a free group on the set A if there is a set-function $j : A \to F(A)$ such that, for all groups G and set-functions $f : A \to G$, *there exists a unique* group homomorphism $\varphi : F(A) \to G$ such that the diagram

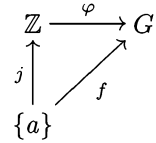


commutes. By general nonsense (Proposition I.5.4), this universal property defines $F(A)$ up to isomorphism, *if this group exists.* But does $F(A)$ exist?

Before giving a 'concrete' construction of $F(A)$, let's check that if $A = \{a\}$ is a singleton, then $F(A) \cong \mathbb{Z}$, as proposed in §5.1. The function $j : A \to \mathbb{Z}$ will send a to $1 \in \mathbb{Z}$. For any group G, giving a set-function $f : A \to G$ amounts to choosing

[21] We could assume that j is *injective,* identifying A with a subset of G; the construction would be completely analogous, and the resulting group would be the same. However, considering arbitrary functions leads to a stronger, more useful, universal property.

one element $g = f(a) \in G$. Now, if $a \in G$, then *there is a unique* homomorphism $\varphi : \mathbb{Z} \to G$ making the diagram



commute: because this forces $\varphi(1) = \varphi \circ j(a) = f(a) = g$, and then the homomorphism condition forces $\varphi(n) = g^n$. That is, φ is necessarily the exponential map ϵ_g considered in §4.1. Therefore, infinite cyclic groups do satisfy the universal property for free groups over a singleton.

5.3. Concrete construction. As we know, terminal objects of a category need not exist. So we have to convince the reader that free groups $F(A)$ exist, for every set A.

Given any set A, we are going to think of A as an 'alphabet' and construct 'words' whose letters are elements of A or 'inverses' of elements of A. To formalize this, consider a set A' isomorphic to A and disjoint from it; call a^{-1} the element in A' corresponding to $a \in A$. A *word* on the set A is an ordered list

$$(a_1, a_2, \cdots, a_n),$$

which we denote by the juxtaposition

$$w = a_1 a_2 \cdots a_n,$$

where each 'letter' a_i is either an element $a \in A$ or an element $a^{-1} \in A'$. We will denote the set of words on A by $W(A)$; the number n of letters is the 'length' of w; we include in $W(A)$ the 'empty word' $w = ( )$, consisting of *no* letters.

For example, if $A = \{a\}$ is a singleton, then an element of $W(A)$ may look like

$$a^{-1}a^{-1}aaaa^{-1}aa^{-1}.$$

An element of $W(\{x, y\})$ may look like

$$xxx^{-1}yy^{-1}xxy^{-1}x^{-1}yy^{-1}xy^{-1}x.$$

Now the notation we have chosen hints that elements in $W(A)$ may be redundant: for example,

$$xyy^{-1}x \quad \text{and} \quad xx$$

are distinct words, but they ought to end up being the same element of a group having to do with words. Therefore, we want to have a process of 'reduction' which takes a word and cleans it up by performing all cancellations. Note that we have to do this 'by hand', since we have not come close yet to defining an operation or making formal sense of considering a^{-1} to be the 'inverse' of a.

Describing the reduction process is invariably awkward—it is a completely evident procedure, but writing it down precisely and elegantly is a challenge. We will settle for the following.

- Define an 'elementary' reduction $r : W(A) \to W(A)$: given $w \in W(A)$, search for the first occurrence (from left to right) of a pair aa^{-1} or $a^{-1}a$, and let $r(w)$ be the word obtained by removing such a pair. In the two examples given above,

$$r(a^{-1}\underline{a^{-1}a}aaa^{-1}aa^{-1}) = a^{-1}aaa^{-1}aa^{-1},$$
$$r(x\underline{xx^{-1}}yy^{-1}xxy^{-1}x^{-1}yy^{-1}xy^{-1}x) = xyy^{-1}xxy^{-1}x^{-1}yy^{-1}xy^{-1}x.$$

- Note that $r(w) = w$ precisely when 'no cancellation is possible'. We say that w is a 'reduced word' in this case.

Lemma 5.1. *If $w \in W(A)$ has length n, then*[22] $r^{\lfloor \frac{n}{2} \rfloor}(w)$ *is a reduced word.*

Proof. Indeed, either $r(w) = w$ or the length of $r(w)$ is less than the length of w; but one cannot decrease the length of w more than $n/2$ times, since each non-identity application of r decreases the length by two. □

- Now define the 'reduction' $R : W(A) \to W(A)$ by setting $R(w) = r^{\lfloor \frac{n}{2} \rfloor}(w)$, where n is the length of w. By the lemma, $R(w)$ is always a reduced word. For example, $R(a^{-1}a^{-1}aaaa^{-1}aa^{-1})$ is the empty word, since

$$r^4(a^{-1}a^{-1}aaaa^{-1}aa^{-1}) = r^3(a^{-1}aaa^{-1}aa^{-1}) = r^2(aa^{-1}aa^{-1}) = r(aa^{-1}) = ();$$

and $R(xxx^{-1}yy^{-1}xxy^{-1}x^{-1}yy^{-1}xy^{-1}x) = xxxy^{-1}y^{-1}x$, as the reader may check.

Let $F(A)$ be the set of reduced words on A, that is, the image of the reduction map R we have just defined.

We are ready to (finally) define free groups 'concretely'. Define a binary operation on $F(A)$ by *juxtaposition & reduction:* for reduced words w, w', define $w \cdot w'$ as the reduction of the juxtaposition of w and w',

$$w \cdot w' := R(ww').$$

It is essentially evident that $F(A)$ is a group under this operation:

- The operation is associative.
- The empty word $e = ()$ is the identity in $F(A)$, since $ew = we = w$ (no reduction is necessary).
- If w is a reduced word, the inverse of w is obtained by reversing the order of the letters of w and replacing each $a \in A$ by $a^{-1} \in A'$ and each a^{-1} by a.

The most cumbersome of these statements to prove formally is associativity; it follows easily from (for example) Exercise 5.4.

There is a function $j : A \to F(A)$, defined by sending the element $a \in A$ to the word consisting of the single 'letter' a.

Proposition 5.2. *The pair $(j, F(A))$ satisfies the universal property for free groups on A.*

[22] $\lfloor q \rfloor$ denotes the largest integer $\leq q$.

Proof. This is also essentially evident, once one has absorbed all the notation. Any function $f : A \to G$ to a group extends uniquely to a map $\varphi : F(A) \to G$, determined by the homomorphism condition and by the requirement that the diagram commutes, which fixes its value on one-letter words $a \in A$ (as well as on $a^{-1} \in A'$).

To check more formally that φ exists as a homomorphism, one can proceed as follows. If $f : A \to G$ is any function, we can extend f to a set-function

$$\tilde{\varphi} : W(A) \to G$$

by insisting that on one-letter words a or a^{-1} (for $a \in A$),

$$\tilde{\varphi}(a) = f(a), \quad \tilde{\varphi}(a^{-1}) = f(a)^{-1},$$

and that $\tilde{\varphi}$ is compatible with juxtaposition:

$$\tilde{\varphi}(ww') = \tilde{\varphi}(w)\tilde{\varphi}(w')$$

for any two words w, w'. The key point now is that reduction is invisible for $\tilde{\varphi}$:

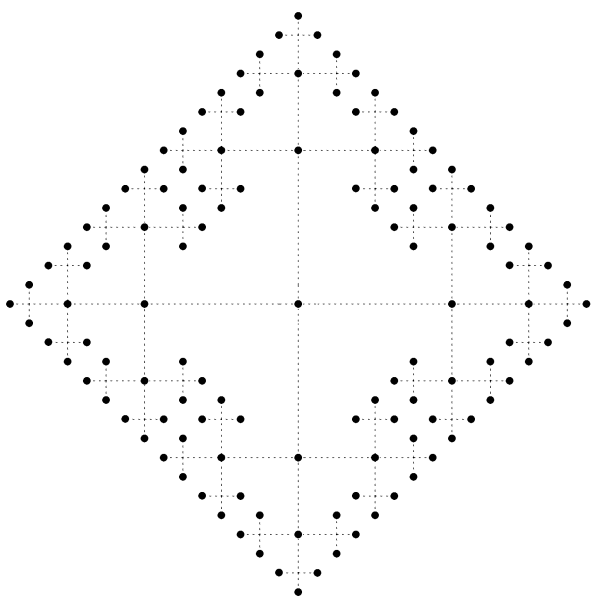
$$\tilde{\varphi}(R(w)) = \tilde{\varphi}(w),$$

since this is clearly the case for *elementary* reductions; therefore, since $\varphi : F(A) \to G$ agrees with $\tilde{\varphi}$ on reduced words, we have for $w, w' \in F(A)$

$$\varphi(w \cdot w') = \tilde{\varphi}(w \cdot w') = \tilde{\varphi}(R(ww')) = \tilde{\varphi}(ww') = \tilde{\varphi}(w)\tilde{\varphi}(w') = \varphi(w)\varphi(w') :$$

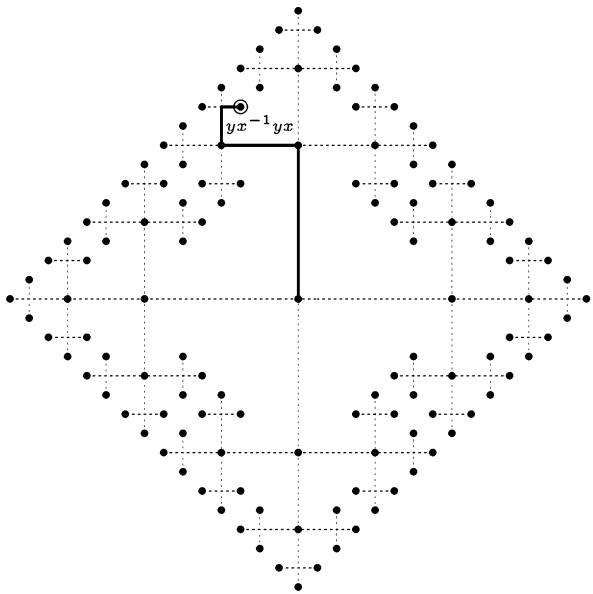
that is, φ is a homomorphism, as needed. □

Example 5.3. It is easy to 'visualize' $F(\{a\}) \cong \mathbb{Z}$; but it is already somewhat challenging for the free group on *two* generators, $F(\{x, y\})$. The best we can do is the following: behold the infinite graph[23]



[23]This is an example of the *Cayley graph* of a group (cf. Exercise 8.6): a graph whose vertices correspond to the elements of the group and whose edges connect vertices according to the action of generators.

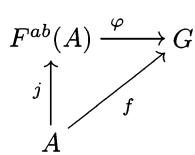
obtained by starting at a point (the center of the picture), then branching out in four directions by a length of 1, then branching out similarly by a length of 1/2, then by 1/4, then by 1/8, then... (and we stopped there, to avoid cluttering the picture too much). Then every element of $F(\{x, y\})$ corresponds in a rather natural way to exactly one dot in this diagram. Indeed, we can place the empty word at the center; and we can agree that every x in a word takes us one step to the right, every x^{-1} to the left, every y up, and every y^{-1} down. For example, the word $yx^{-1}yx$ takes us here:



The reader will surely encounter this group elsewhere: it is the *fundamental group* of the 'figure 8'. ⌟

5.4. Free *abelian* groups. We can pose in Ab the same question answered above for Grp: that is, ask for the *abelian* group $F^{ab}(A)$ which most efficiently contains the set A, provided that we do not have any additional information on the elements of A. Of course we *do* know something about the elements of A this time: they will have to commute with each other in $F^{ab}(A)$. This plays no role if $A = \{a\}$ is a singleton, and therefore $F^{ab}(\{a\}) = F(\{a\}) \cong \mathbb{Z}$; but the requirement is different for larger sets, so we should expect a different answer in general.

The formalization of the heuristic requirement is precisely the same universal property that gave us free groups, but (of course) stated in Ab: $F^{ab}(A)$ is a *free abelian group* on the set A if there is a set-function $j : A \to F^{ab}(A)$ such that, for all *abelian* groups G and set-functions $f : A \to G$, *there exists a unique* group homomorphism $\varphi : F^{ab}(A) \to G$ such that the following diagram commutes:



Again, Proposition I.5.4 guarantees that $F^{ab}(A)$ is unique up to isomorphism, *if* it exists; but we have to prove it exists! This is in some way simpler than for Grp, in the sense that $F^{ab}(A)$ is easier to understand, at least for finite sets A.

To fix ideas, we will first describe the answer for a finite set, say $A = \{1, \cdots, n\}$. We will denote by $\mathbb{Z}^{\oplus n}$ the direct sum

$$\underbrace{\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}}_{n\text{-times}};$$

recall (§3.5) that this group 'is the same as' the product[24] $\mathbb{Z}^n$ (but we view it as a *co*product). There is a function $j : A \to \mathbb{Z}^{\oplus n}$, defined by

$$j(i) := (0, \cdots, 0, \underset{i\text{-th place}}{1}, 0, \cdots, 0) \in \mathbb{Z}^{\oplus n}.$$

Claim 5.4. *For $A = \{1, \cdots, n\}$, $\mathbb{Z}^{\oplus n}$ is a free abelian group on A.*

Proof. Note that every element of $\mathbb{Z}^{\oplus n}$ can be written uniquely in the form $\sum_{i=1}^{n} m_i\, j(i)$: indeed,

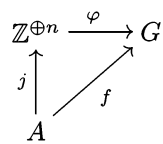
$$\begin{aligned}(m_1, \cdots, m_n) &= (m_1, 0, \cdots, 0) + (0, m_2, 0, \cdots, 0) + \cdots + (0, \cdots, 0, m_n)\\ &= m_1(1, 0, \cdots, 0) + m_2(0, 1, 0, \cdots, 0) + \cdots + m_n(0, \cdots, 0, 1)\\ &= m_1\, j(1) + \cdots + m_n\, j(n),\end{aligned}$$

and $(m_1, \cdots, m_n) = (0, \cdots, 0)$ if and only if all m_i are 0.

Now let $f : A \to G$ be any function from $A = \{1, \cdots, n\}$ to an abelian group G. We define $\varphi : \mathbb{Z}^{\oplus n} \to G$ by

$$\varphi\left(\sum_{i=1}^{n} m_i\, j(i)\right) := \sum_{i=1}^{n} m_i\, f(i) :$$

indeed, we have no choice—this definition is forced by the needed commutativity of the diagram



and by the homomorphism condition. Thus φ is certainly uniquely determined, and we just have to check that it is a homomorphism. This is where the commutativity of G enters:

$$\varphi\left(\sum_{i=1}^{n} m_i'\, j(i)\right) + \varphi\left(\sum_{i=1}^{n} m_i''\, j(i)\right) = \sum_{i=1}^{n} m_i'\, f(i) + \sum_{i=1}^{n} m_i''\, f(i) \overset{!}{=} \sum_{i=1}^{n} (m_i' + m_i'') f(i)$$

because G is commutative,

$$= \varphi\left(\sum_{i=1}^{n} (m_i' + m_i'')\, j(i)\right) = \varphi\left(\sum_{i=1}^{n} m_i'\, j(i) + \sum_{i=1}^{n} m_i''\, j(i)\right)$$

as needed. □

[24] Indeed, it is common to denote this group by $\mathbb{Z}^n$, omitting the $\oplus$. No confusion is likely, but we will try to distinguish the two to emphasize that they play different categorical roles.

Remark 5.5. A less hands-on, more high-brow argument can be given by contemplating the universal property defining free abelian groups vis-à-vis the universal property for coproducts; cf. Exercise 5.7. ⌟

Now for the general case: let A be any set. As we have seen, $H^A = \mathrm{Hom}_{\mathsf{Set}}(A, H)$ has a natural abelian group structure if H is an abelian group (§4.4); elements of H^A are arbitrary set-functions $\alpha : A \to H$. We can define a subset $H^{\oplus A}$ of H^A as follows:

$$H^{\oplus A} := \{\alpha : A \to H \mid \alpha(a) \neq e_H \text{ for only finitely many elements } a \in A\}.$$

The operation in H^A induces an operation in $H^{\oplus A}$, which makes $H^{\oplus A}$ into a group[25].

The reader should note that $H^{\oplus A}$ is the whole of H^A if A is a finite set; and that $\mathbb{Z}^{\oplus A} \cong \mathbb{Z}^{\oplus n}$ if $A = \{1, \cdots, n\}$: indeed, $(m_1, \cdots, m_n) \in \mathbb{Z}^{\oplus n}$ may be identified with the function $\{1, \cdots, n\} \to \mathbb{Z}$ sending i to m_i.

For $H = \mathbb{Z}$ there is a natural function $j : A \to \mathbb{Z}^{\oplus A}$, obtained by sending $a \in A$ to the function $j_a : A \to \mathbb{Z}$ defined by

$$(\forall x \in A): \quad j_a(x) := \begin{cases} 1 & \text{if } x = a, \\ 0 & \text{if } x \neq a. \end{cases}$$

Note that for $A = \{1, \cdots, n\}$ and identifying $\mathbb{Z}^{\oplus A} \cong \mathbb{Z}^{\oplus n}$, this function j is *the same function* denoted j earlier.

Proposition 5.6. *For every set A, $F^{ab}(A) \cong \mathbb{Z}^{\oplus A}$.*

Proof. The key point is again that every element of $\mathbb{Z}^{\oplus A}$ may be written uniquely as a *finite* sum

$$\sum_{a \in A} m_a\, j(a), \quad m_a \neq 0 \text{ for only finitely many } a;$$

once this is understood, the argument is precisely the same as for Claim 5.4. □

[25] Thus $H^{\oplus A}$ is a *subgroup* of H^A; cf. §6.

Exercises

5.1. Does the category $\mathscr{F}^A$ defined in §5.2 have final objects? If so, what are they?

5.2. Since trivial groups T are initial in Grp, one may be led to think that (e, T) should be initial in $\mathscr{F}^A$, for every A: e would be defined by sending every element of A to the (only) element in T; and for any other group G, *there is a unique* homomorphism $T \to G$. Explain why (e, T) is *not* initial in $\mathscr{F}^A$ (unless $A = \emptyset$).

5.3. ▷ Use the universal property of free groups to prove that the map $j : A \to F(A)$ is injective, for all sets A. (Hint: It suffices to show that for every two elements a, b of A there is a group G and a set-function $f : A \to G$ such that $f(a) \neq f(b)$. Why? How do you construct f and G?) [§III.6.3]

5.4. ▷ In the 'concrete' construction of free groups, one can try to reduce words by performing cancellations in any order; the process of 'elementary reductions' used in the text (that is, from left to right) is only one possibility. Prove that the result of iterating cancellations on a word is independent of the order in which the cancellations are performed. Deduce the associativity of the product in $F(A)$ from this. [§5.3]

5.5. Verify explicitly that $H^{\oplus A}$ is a group.

5.6. ▷ Prove that the group $F(\{x, y\})$ (visualized in Example 5.3) is a coproduct $\mathbb{Z} * \mathbb{Z}$ of $\mathbb{Z}$ by itself in the category Grp. (Hint: With due care, the universal property for one turns into the universal property for the other.) [§3.4, 3.7, 5.7]

5.7. ▷ Extend the result of Exercise 5.6 to free groups $F(\{x_1, \ldots, x_n\})$ and to free *abelian* groups $F^{ab}(\{x_1, \ldots, x_n\})$. [§3.4, §5.4]

5.8. Still more generally, prove that $F(A \amalg B) = F(A) * F(B)$ and that $F^{ab}(A \amalg B) = F^{ab}(A) \oplus F^{ab}(B)$ for all sets A, B. (That is, the constructions F, F^{ab} 'preserve coproducts'.)

5.9. Let $G = \mathbb{Z}^{\oplus \mathbb{N}}$. Prove that $G \times G \cong G$.

5.10. ¬ Let $F = F^{ab}(A)$.

- Define an equivalence relation $\sim$ on F by setting $f' \sim f$ if and only if $f - f' = 2g$ for some $g \in F$. Prove that $F/\!\sim$ is a finite set if and only if A is finite, and in that case $|F/\!\sim| = 2^{|A|}$.
- Assume $F^{ab}(B) \cong F^{ab}(A)$. If A is finite, prove that B is also, and that $A \cong B$ as sets. (This result holds for free groups as well, and without any finiteness hypothesis. See Exercises 7.13 and VI.1.20.)

[7.4, 7.13]

6. Subgroups

6.1. Definition. Let $(G, \cdot)$ be a group, and let $(H, \bullet)$ be another group, whose underlying set H is a subset of G.

Definition 6.1. $(H, \bullet)$ is a *subgroup* of G if the inclusion function $i : H \hookrightarrow G$ is a group homomorphism. ⌟

For example, the trivial group consisting of the single element e_G is a subgroup of G.

If $(H, \bullet)$ is a subgroup of $(G, \cdot)$, then $\forall h_1, h_2 \in H$:

$$(*) \qquad i(h_1 \bullet h_2) = i(h_1) \cdot i(h_2).$$

We say that the operation $\bullet$ on H is 'induced' from the operation $\cdot$ on G; in practice one omits explicit mention of i and of the operations, and (*) guarantees that no ambiguity will arise from this.

The subgroup condition may be streamlined. A subset H of a group G determines a subgroup if the operation $\cdot$ in G induces (by (*)) a binary operation in H (we say that H is *closed* with respect to the operation in G), satisfying the group axioms. Since the identity and inverses are preserved through homomorphisms (Proposition 3.2), the identity e_H of H will have to coincide with the identity e_G of G and the inverse of an element $h \in H$ has to be the same as the inverse of that element in G. The most economical way to say all this is

Proposition 6.2. *A* nonempty *subset H of a group G is a subgroup if and only if*

$$(\forall a, b \in H): \quad ab^{-1} \in H.$$

Proof. It is clear that if H is a subgroup, then the stated condition holds: indeed, if $b \in H$, then the inverse of b must also be in H and H is closed under the operation of G.

Conversely, assume the stated condition holds; we have to check that H is closed under the operation of G, the induced operation on H is associative, and it admits an identity element and inverses (that is, it contains e_G and is closed under taking inverses in G). Since H is nonempty, we can find an element $h \in H$. Choosing $a = b = h$, we see that

$$e_G = hh^{-1} = ab^{-1} \in H;$$

thus H contains the identity. Given any $h \in H$, choosing $a = e_G$ and $b = h$ shows that

$$h^{-1} = e_G h^{-1} = ab^{-1} \in H;$$

thus H contains the inverse of any of its elements. Given any $h_1, h_2 \in H$, choose $a = h_1$, $b = h_2^{-1}$; the stated condition says that

$$h_1 h_2 = h_1 (h_2)^{-1} = ab^{-1} \in H,$$

proving that H is closed under the operation.

Finally, the fact that the operation is associative in G implies immediately that the induced operation is associative in H, concluding the proof that H, with the induced operation, is a group. □

This criterion makes it particularly straightforward to check simple facts concerning subgroups. For example,

Lemma 6.3. *If $\{H_\alpha\}_{\alpha\in A}$ is any family of subgroups of a group G, then*

$$H = \bigcap_{\alpha\in A} H_i$$

is a subgroup of G.

Proof. This follows right away from Proposition 6.2: H is nonempty, because $e \in H_\alpha$ for all α, so $e \in H$; and

$$a, b \in H \implies (\forall \alpha \in A) : a, b \in H_\alpha \implies (\forall \alpha \in A) : ab^{-1} \in H_\alpha \implies ab^{-1} \in H,$$

proving that H is a subgroup of G. □

Similarly,

Lemma 6.4. *Let $\varphi : G \to G'$ be a group homomorphism, and let H' be a subgroup of G'. Then $\varphi^{-1}(H')$ is a subgroup of G.*

Proof. Recall (end of §I.2.5) that $\varphi^{-1}(H')$ consists of all $g \in G$ such that $\varphi(g) \in H'$. Since $\varphi(e_G) = e_{G'} \in H'$, this set is nonempty. If $a, b \in \varphi^{-1}(H')$, then $\varphi(a)$ and $\varphi(b)$ are in H', and hence

$$\varphi(ab^{-1}) = \varphi(a)\varphi(b)^{-1} \in H' :$$

thus, $ab^{-1} \in \varphi^{-1}(H)$. This implies that $\varphi^{-1}(H')$ is a subgroup of G, by Proposition 6.2, □

6.2. Examples: Kernel and image. Every group homomorphism $\varphi : G \to G'$ determines two interesting subgroups:

- the *kernel* of φ, $\ker\varphi \subseteq G$; and
- the *image* of φ, $\operatorname{im}\varphi \subseteq G'$.

Definition 6.5. The *kernel* of $\varphi : G \to G'$ is the subset of G consisting of elements mapping to the identity in G':

$$\ker\varphi := \{g \in G \,|\, \varphi(g) = e_{G'}\} = \varphi^{-1}(e_{G'}).$$ ⌟

Since $\{e_{G'}\}$ is a subgroup of G', Lemma 6.4 shows that $\ker\varphi$ is indeed a subgroup of G. For an (even) more explicit argument, note that $\ker\varphi$ is nonempty, since $e_G \in \ker\varphi$; and if a, b are in $\ker\varphi$, then

$$\varphi(ab^{-1}) = \varphi(a)\varphi(b)^{-1} = e_{G'}e_{G'}^{-1} = e_{G'},$$

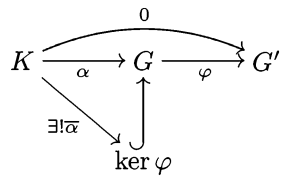
proving that $ab^{-1} \in \ker\varphi$. This shows that $\ker\varphi$ is a subgroup of G, by Proposition 6.2.

The verification that $\operatorname{im}\varphi$ is a subgroup is left to the reader. In fact, the reader should check that the image of *any* subgroup of G is a subgroup of G'.

We will soon (§7.1) see that kernels are 'special' subgroups. As with most constructions of importance in algebra, they satisfy a universal property, which may be expressed as follows.

Proposition 6.6. *Let $\varphi : G \to G'$ be a homomorphism. Then the inclusion $i : \ker\varphi \hookrightarrow G$ is final in the category*[26] *of group homomorphisms $\alpha : K \to G$ such that $\varphi \circ \alpha$ is the trivial map.*

In other words, every group homomorphism $\alpha : K \to G$ such that $\varphi \circ \alpha$ is the trivial homomorphism (denoted '0' in the diagram) factors uniquely through $\ker\varphi$:



Proof. If $\alpha : K \to G$ is such that $\varphi \circ \alpha$ is the trivial map, then $\forall k \in K$

$$\varphi \circ \alpha(k) = \varphi(\alpha(k)) = e_G,$$

that is, $\alpha(k) \in \ker\varphi$. We can (and must) then let $\overline{\alpha} : K \to \ker\alpha$ simply be α itself, with restricted target. □

Proposition 6.6 indicates how one might define a notion analogous to 'kernel' in very general settings. This viewpoint will be championed much later in this book, especially in Chapter IX.

Remark 6.7. The argument shows that in fact kernels of group homomorphisms satisfy a somewhat stronger universal property: any *set-function* $\alpha : K \to G$ such that the image of $\varphi \circ \alpha$ is the identity in H must factor (as a set-function) through $\ker\varphi$. ⌟

6.3. Example: Subgroup generated by a subset. If $A \subseteq G$ is *any* subset, we have a unique group homomorphism

$$\varphi_A : F(A) \to G,$$

by the universal property of free groups. The image of this homomorphism is a subgroup of G, the *subgroup generated by* A in G, often denoted[27] $\langle A \rangle$.

Of course, if G is abelian, then φ_A factors through $F^{ab}(A)$, so we may replace $F(A)$ by $F^{ab}(A)$ in this case.

The 'concrete' description of free groups (§5.3) leads to the following description of $\langle A \rangle$: it consists of all products in G of the form

$$a_1 a_2 a_3 \cdots a_n$$

where each a_i is either an element of A, the inverse of an element of A, or the identity. This is clearly the most 'economical' way to manufacture a subgroup of G, given the elements of A.

---

[26] The reader should specify what the morphisms are in this category.

[27] If $A = \{g_1, \ldots, g_r\}$ is a finite set, one writes $\langle g_1, \ldots, g_r \rangle$.

The reader who has not (yet) developed a taste for free groups may prefer the following alternative description: $\langle A\rangle$ is the intersection of all subgroups of G containing A,

$$\langle A\rangle = \bigcap_{H \text{ subgroup of } G,\ H \supseteq A} H.$$

Indeed, the intersection on the right-hand side is a subgroup of G by Lemma 6.3, it contains A, and it is clearly the smallest subgroup satisfying this condition.

If $A = \{g\}$ consists of a single element, then $F(A) = \mathbb{Z}$ and $\varphi_A : \mathbb{Z} \to G$ is nothing but the 'exponential map' ϵ_g (cf. §4.1); $\langle A\rangle = \langle g\rangle$ is then the image of this map:

$$\langle g\rangle = \mathrm{im}(\epsilon_g) = \{\dots, g^{-2}, g^{-1}, e, g, g^2, \dots\}.$$

The subgroup $\langle g\rangle$ is the 'cyclic subgroup generated by g': indeed, $\langle g\rangle$ is *cyclic* in the sense of Definition 4.7; the reader can easily check this fact already (Exercise 6.4); it will also be recovered as an immediate consequence of the construction of quotients (cf. §7.5).

Definition 6.8. A group G is *finitely generated* if there exists a *finite* subset $A \subseteq G$ such that $G = \langle A\rangle$. ⌟

For examples, cyclic groups are finitely generated (in fact, they are generated by a singleton). By definition, a group is finitely generated if and only if there is a surjective homomorphism

$$F(\{1, \dots, n\}) \twoheadrightarrow G$$

for some n. One of the most memorable results proven in this book will give a *classification of finitely generated abelian groups:* we will be able to prove that every such group is a *direct sum of cyclic groups* (Theorem IV.6.6, Exercise VI.2.19, and the generalization given in Theorem VI.5.6). The situation for general groups is considerably more complex. The classification of *finite* (simple) groups is one of the major achievements of twentieth-century mathematics, and it is spread over at least 10,000 pages of research articles. To appreciate the difference in complexity, note that there are *42 abelian* groups of order 1024 up to isomorphism (as the reader will be able to establish in due time: Exercise IV.6.6); allegedly, there are *49,487,365,402* if we count noncommutative ones as well[28].

6.4. Example: Subgroups of cyclic groups. We are ready to determine *all* subgroups of *all* cyclic groups, that is, all subgroups of $\mathbb{Z}$ and of $\mathbb{Z}/n\mathbb{Z}$, for all $n > 0$ (because every cyclic group is isomorphic to one of these; cf. Definition 4.7). The result is easy to remember: *subgroups of cyclic groups are themselves cyclic groups.*

It is convenient to start from $\mathbb{Z}$. For $d \in \mathbb{Z}$ we let

$$d\mathbb{Z} := \langle d\rangle = \{m \in \mathbb{Z} \mid \exists q \in \mathbb{Z}, m = dq\};$$

that is, $d\mathbb{Z}$ denotes the set of *integer multiples* of d. Of course this is nothing but the 'cyclic subgroup of $\mathbb{Z}$ generated by d'.

Proposition 6.9. *Let $G \subseteq \mathbb{Z}$ be a subgroup. Then $G = d\mathbb{Z}$ for some $d \geq 0$.*

[28] This comparison is a little unfair, however, since it so happens that more than 99% of all groups of order < 2000 have order 1024.

The proof will actually show that if $G \subseteq \mathbb{Z}$ is nontrivial, then d is the *smallest positive* element of G, and the reader is invited to remember this useful fact.

Remark 6.10. By Proposition 6.9, every nontrivial subgroup of $\mathbb{Z}$ is in fact isomorphic to $\mathbb{Z}$. Putting this a little strangely, it says that every subgroup of the *free* group on one generator is free. It is in fact true that *every subgroup of a* (finitely generated) *free group is free;* we will not prove this fact, although the diligent reader will get a taste of the argument in Exercise 9.16. In any case, beware that free groups on *two* generators already contain subgroups isomorphic to free groups on *arbitrarily many* generators. Indeed, the commutator subgroup (cf. Exercise 7.12) $[F, F]$ for $F = F(\{x, y\})$ is isomorphic to a free group on *infinitely many* generators (unfortunately, we will not prove this beautiful statement either). ⌟

Proof of Proposition 6.9. If $G = \{0\}$, then $G = 0\mathbb{Z}$. If not, note that G must contain *positive* integers: indeed, if $a \in G$ and $a < 0$, then $-a \in G$ and $-a > 0$. We can then let d be the *smallest positive integer*[29] in G, and we claim $G = d\mathbb{Z}$.

The inclusion $d\mathbb{Z} \subseteq G$ is clear. To verify the inclusion $G \subseteq d\mathbb{Z}$, let $m \in G$, and apply 'division with remainder' to write

$$m = dq + r,$$

with $0 \leq r < d$. Since $n \in G$ and $d\mathbb{Z} \subseteq G$ and since G is a subgroup, we see that

$$r = m - dq \in G.$$

But d is the smallest *positive* integer in G, and $r \in G$ is smaller than d; so r cannot be positive. This shows $r = 0$, that is, $m = qd \in d\mathbb{Z}$; $G \subseteq d\mathbb{Z}$ follows, and we are done. □

The 'quotient' homomorphism $\pi_n : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ (cf. §4.1) allows us to establish the analogous result for *finite* cyclic groups:

Proposition 6.11. *Let $n > 0$ be an integer and let $G \subseteq \mathbb{Z}/n\mathbb{Z}$ be a subgroup. Then G is the cyclic subgroup of $\mathbb{Z}/n\mathbb{Z}$ generated by $[d]_n$, for some divisor d of n.*

Proof. Let $\pi_n : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ be the quotient map, and consider $G' := \pi_n^{-1}(G)$. By Lemma 6.4, G' is a subgroup of $\mathbb{Z}$; by Proposition 6.9, G' is a *cyclic* subgroup of $\mathbb{Z}$, generated by a nonnegative integer d. It follows that

$$G = \pi_n(G') = \pi_n(\langle d \rangle) = \langle [d]_n \rangle;$$

thus G is indeed a cyclic subgroup of $\mathbb{Z}/n\mathbb{Z}$, generated by a class $[d]_n$. Further, since $n \in G'$ (because $\pi_n(n) = [n]_n = [0]_n \in G$) and $G' = d\mathbb{Z}$, we see that d divides n, as claimed. □

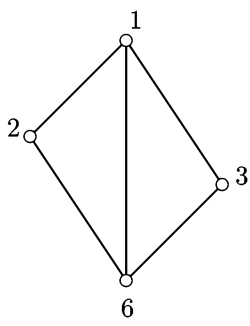
As a consequence of Proposition 6.11, there is a bijection between the set of subgroups of $\mathbb{Z}/n\mathbb{Z}$ and the set of positive divisors of n. For example, $\mathbb{Z}/12\mathbb{Z}$ has

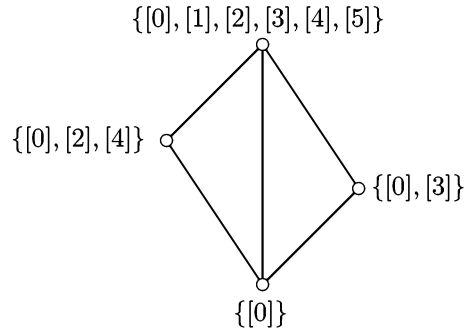
[29] We are secretly appealing to the 'well-ordering principle'. That every set of positive integers should have a smallest element is one of those fact about $\mathbb{Z}$—like the availability of division-with-remainder—that we are assuming the reader is already familiar with.

exactly 6 subgroups, because 12 has 6 positive divisors: 1, 2, 3, 4, 6, and 12. Here is the corresponding list of subgroups:

$$\begin{aligned}
\langle [1]_{12}\rangle &= \{[0]_{12}, [1]_{12}, [2]_{12}, [3]_{12}, [4]_{12}, [5]_{12}, [6]_{12}, [7]_{12}, [8]_{12}, [9]_{12}, [10]_{12}, [11]_{12}\},\\
\langle [2]_{12}\rangle &= \{[0]_{12}, [2]_{12}, [4]_{12}, [6]_{12}, [8]_{12}, [10]_{12}\},\\
\langle [3]_{12}\rangle &= \{[0]_{12}, [3]_{12}, [6]_{12}, [9]_{12}\},\\
\langle [4]_{12}\rangle &= \{[0]_{12}, [4]_{12}, [8]_{12}\},\\
\langle [6]_{12}\rangle &= \{[0]_{12}, [6]_{12}\},\\
\langle [12]_{12}\rangle &= \{[0]_{12}\}.
\end{aligned}$$

Also note that if d_1, d_2 are both divisors of n, and $d_1 \mid d_2$, then $\langle [d_1]_n\rangle \supseteq \langle [d_2]_n\rangle$. That is, the correspondence between subgroups of $\mathbb{Z}/n\mathbb{Z}$ and divisors of n preserves the natural *lattice* structure carried by these sets. We can draw these lattices for $\mathbb{Z}/6\mathbb{Z}$ as follows:





where lines connect multiples in one picture and subsets in the other. The reader will draw the lattice of subgroups of S_3, noting that it looks completely different from the one for $\mathbb{Z}/6\mathbb{Z}$.

Contemplating subgroups of cyclic groups has pretty (and useful) 'number-theoretic' consequences; cf. Exercise 6.14.

6.5. Monomorphisms. We end this section with some categorical considerations.

If H is a subgroup of G, the inclusion $H \hookrightarrow G$ is an example of a *monomorphism* in Grp in the 'categorical' sense of §I.4.2. In fact, it is easy to characterize all monomorphisms $\varphi \in \mathrm{Hom}_{\mathsf{Grp}}(G, G')$ (where G, G' are any groups):

Proposition 6.12. *The following are equivalent:*

(a) φ *is a monomorphism;*

(b) $\ker\varphi = \{e_G\}$*;*

(c) $\varphi : G \to G'$ *is injective (as a set-function).*

Proof. (a) $\Longrightarrow$ (b): Assume (a) holds, and consider the two parallel compositions

$$\ker\varphi \underset{e}{\overset{i}{\rightrightarrows}} G \xrightarrow{\varphi} G',$$

where i is the inclusion and e is the trivial map. Both $\varphi \circ i$ and $\varphi \circ e$ are the trivial map; since φ is a monomorphism, this implies $i = e$. But $i = e$ implies that $\ker \varphi$ is trivial, that is, (b) holds.

(b) $\Longrightarrow$ (c): Assume $\ker \varphi = \{e_G\}$. Then

$$\varphi(g_1) = \varphi(g_2) \implies \varphi(g_1)\varphi(g_2)^{-1} = e_{G'} \implies \varphi(g_1 g_2^{-1}) = e_{G'}$$
$$\implies g_1 g_2^{-1} \in \ker \varphi \implies g_1 g_2^{-1} = e_G \implies g_1 = g_2.$$

This shows that φ is injective, as needed.

(c) $\Longrightarrow$ (a): If φ is injective, then it satisfies the defining property for monomorphisms in Set: that is, for any set Z and any two set-functions $\alpha', \alpha'' : Z \to G$,

$$\varphi \circ \alpha' = \varphi \circ \alpha'' \iff \alpha' = \alpha''.$$

This must hold in particular if Z has a group structure and α', α'' are group homomorphisms, so φ is a monomorphism in Grp. □

The equivalence (a) $\Longleftrightarrow$ (c) may lead the reader to think that from the point of view of monomorphisms, Grp and Set are pretty much alike. This is not quite so: while it is true that homomorphisms with a left-inverse are necessarily monomorphisms, as in Set (cf. Exercise 6.15), the converse is not true in Grp (cf. Exercise 6.16).

Exercises

6.1. ¬ (If you know about matrices.) The group of invertible $n \times n$ matrices with entries in $\mathbb{R}$ is denoted $\mathrm{GL}_n(\mathbb{R})$ (Example 1.5). Similarly, $\mathrm{GL}_n(\mathbb{C})$ denotes the group of $n \times n$ invertible matrices with *complex* entries. Consider the following sets of matrices:

- $\mathrm{SL}_n(\mathbb{R}) = \{M \in \mathrm{GL}_n(\mathbb{R}) \mid \det(M) = 1\}$;
- $\mathrm{SL}_n(\mathbb{C}) = \{M \in \mathrm{GL}_n(\mathbb{C}) \mid \det(M) = 1\}$;
- $\mathrm{O}_n(\mathbb{R}) = \{M \in \mathrm{GL}_n(\mathbb{R}) \mid MM^t = M^t M = I_n\}$;
- $\mathrm{SO}_n(\mathbb{R}) = \{M \in \mathrm{O}_n(\mathbb{R}) \mid \det M = I_n\}$;
- $\mathrm{U}_n(\mathbb{C}) = \{M \in \mathrm{GL}_n(\mathbb{C}) \mid MM^\dagger = M^\dagger M = I_n\}$;
- $\mathrm{SU}_n(\mathbb{C}) = \{M \in \mathrm{U}_n(\mathbb{C}) \mid \det(M) = 1\}$.

Here I_n stands for the $n \times n$ *identity matrix*, M^t is the *transpose* of M, $M^\dagger$ is the *conjugate transpose* of M, and $\det(M)$ denotes the *determinant*[30] of M. Find all possible inclusions among these sets, and prove that in every case the smaller set is a subgroup of the larger one.

These sets of matrices have compelling geometric interpretations: for example, $\mathrm{SO}_3(\mathbb{R})$ is the group of 'rotations' in $\mathbb{R}^3$. [8.8, 9.1, III.1.4, VI.6.16]

[30] If you are not familiar with some of these notions, that's ok: leave this exercise and similar ones alone if that is the case. We will come back to linear algebra and matrices in Chapter VI and following.

6.2. ¬ Prove that the set of 2×2 matrices

$$\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$$

with a, b, d in $\mathbb{C}$ and $ad \neq 0$ is a subgroup of $\mathrm{GL}_2(\mathbb{C})$. More generally, prove that the set of $n \times n$ complex matrices $(a_{ij})_{1 \leq i,j \leq n}$ with $a_{ij} = 0$ for $i > j$ and $a_{11} \cdots a_{nn} \neq 0$ is a subgroup of $\mathrm{GL}_n(\mathbb{C})$. (These matrices are called 'upper triangular', for evident reasons.) [IV.1.20]

6.3. ¬ Prove that every matrix in $\mathrm{SU}_2(\mathbb{C})$ may be written in the form

$$\begin{pmatrix} a+bi & c+di \\ -c+di & a-bi \end{pmatrix}$$

where $a, b, c, d \in \mathbb{R}$ and $a^2 + b^2 + c^2 + d^2 = 1$. (Thus, $\mathrm{SU}_2(\mathbb{C})$ may be realized as a three-dimensional sphere embedded in $\mathbb{R}^4$; in particular, it is *simply connected.*) [8.9, III.2.5]

6.4. ▷ Let G be a group, and let $g \in G$. Verify that the image of the exponential map $\epsilon_g : \mathbb{Z} \to G$ is a cyclic group (in the sense of Definition 4.7). [§6.3, §7.5]

6.5. Let G be a *commutative* group, and let $n > 0$ be an integer. Prove that $\{g^n \mid g \in G\}$ is a subgroup of G. Prove that this is not necessarily the case if G is not commutative.

6.6. Prove that the union of a family of subgroups of a group G is not necessarily a subgroup of G. In fact:

- Let H, H' be subgroups of a group G. Prove that $H \cup H'$ is a subgroup of G only if $H \subseteq H'$ or $H' \subseteq H$.
- On the other hand, let $H_0 \subseteq H_1 \subseteq H_2 \subseteq \cdots$ be subgroups of a group G. Prove that $\bigcup_{i \geq 0} H_i$ *is* a subgroup of G.

6.7. ¬ Show that *inner* automorphisms (cf. Exercise 4.8) form a subgroup of $\mathrm{Aut}(G)$; this subgroup is denoted $\mathrm{Inn}(G)$. Prove that $\mathrm{Inn}(G)$ is cyclic if and only if $\mathrm{Inn}(G)$ is trivial if and only if G is abelian. (Hint: Assume that $\mathrm{Inn}(G)$ is cyclic; with notation as in Exercise 4.8, this means that there exists an element $a \in G$ such that $\forall g \in G\ \exists n \in \mathbb{Z}\ \gamma_g = \gamma_a^n$. In particular, $gag^{-1} = a^n a a^{-n} = a$. Thus a commutes with every g in G. Therefore....) Deduce that if $\mathrm{Aut}(G)$ is cyclic, then G is abelian. [7.10, IV.1.5]

6.8. Prove that an *abelian* group G is finitely generated if and only if there is a surjective homomorphism

$$\underbrace{\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}}_{n \text{ times}} \twoheadrightarrow G$$

for some n.

6.9. Prove that every finitely generated subgroup of $\mathbb{Q}$ is cyclic. Prove that $\mathbb{Q}$ is not finitely generated.

6.10. ¬ The set of 2×2 matrices with integer entries and determinant 1 is denoted $\mathrm{SL}_2(\mathbb{Z})$:

$$\mathrm{SL}_2(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ such that } a, b, c, d \in \mathbb{Z},\ ad - bc = 1 \right\}.$$

Prove that $\mathrm{SL}_2(\mathbb{Z})$ is generated by the matrices

$$s = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

(Hint: This is a little tricky. Let H be the subgroup generated by s and t. Given a matrix $m = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ in $\mathrm{SL}_2(\mathbb{Z})$, it suffices to show that you can obtain the identity by multiplying m by suitably chosen elements of H. Prove that $\begin{pmatrix} 1 & -q \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ -q & 1 \end{pmatrix}$ are in H, and note that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & -q \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & b - qa \\ c & d - qc \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -q & 1 \end{pmatrix} = \begin{pmatrix} a - qb & b \\ c - qd & d \end{pmatrix}.$$

Note that if c and d are both nonzero, one of these two operations may be used to decrease the absolute value of one of them. Argue that suitable applications of these operations reduce to the case in which $c = 0$ or $d = 0$. Prove directly that $m \in H$ in that case.) [7.5]

6.11. Since direct sums are coproducts in Ab, the classification theorem for abelian groups mentioned in the text says that every finitely generated *abelian group* is a coproduct of cyclic groups in Ab. The reader may be tempted to conjecture that every finitely generated *group* is a coproduct *in* Grp. Show that this is not the case, by proving that S_3 is not a coproduct of cyclic groups.

6.12. Let m, n be positive integers, and consider the subgroup $\langle m, n \rangle$ of $\mathbb{Z}$ they generate. By Proposition 6.9,

$$\langle m, n \rangle = d\mathbb{Z}$$

for some positive integer d. What is d, in relation to m, n?

6.13. ¬ Draw and compare the lattices of subgroups of $C_2 \times C_2$ and C_4. Draw the lattice of subgroups of S_3, and compare it with the one for C_6. [7.1]

6.14. ▷ If m is a positive integer, denote by $\phi(m)$ the number of positive integers $r < m$ that are *relatively prime* to m (that is, for which the gcd of r and m is 1); this is called *Euler's ϕ- (or 'totient') function.* For example, $\phi(12) = 4$. In other words, $\phi(m)$ is the order of the group $(\mathbb{Z}/m\mathbb{Z})^*$; cf. Proposition 2.6.

Put together the following observations:

- $\phi(m)$ = the number of generators of C_m,
- every element of C_n generates a subgroup of C_n,
- the discussion following Proposition 6.11 (in particular, every subgroup of C_n is isomorphic to C_m, for some $m \mid n$),

to obtain a proof of the formula

$$\sum_{m>0, m|n} \phi(m) = n.$$

(For example, $\phi(1)+\phi(2)+\phi(3)+\phi(4)+\phi(6)+\phi(12) = 1+1+2+2+2+4 = 12$.) [4.14, §6.4, 8.15, V.6.8, §VII.5.2]

6.15. ▷ Prove that if a group homomorphism $\varphi : G \to G'$ has a left-inverse, that is, a group homomorphism $\psi : G' \to G$ such that $\psi \circ \varphi = \mathrm{id}_{G'}$, then φ is a monomorphism. [§6.5, 6.16]

6.16. ▷ Counterpoint to Exercise 6.15: the homomorphism $\varphi : \mathbb{Z}/3\mathbb{Z} \to S_3$ given by

$$\varphi([0]) = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad \varphi([1]) = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \quad \varphi([2]) = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

is a monomorphism; show that it has *no* left-inverse in Grp. (Knowing about *normal* subgroups will make this problem particularly easy.) [§6.5]

---

7. Quotient groups

7.1. Normal subgroups. Before tackling 'quotient groups', we should clarify in what sense kernels are *special* subgroups, as claimed in §6.2.

Definition 7.1. A subgroup N of a group G is *normal* if $\forall g \in G$, $\forall n \in N$,

$$gng^{-1} \in N.$$ ⌟

Note that *every* subgroup of a commutative group is normal (because then $\forall g \in G$, $gng^{-1} = n \in N$). However, in general *not all* subgroups are normal: examples may be found already in S_3 (cf. Exercise 7.1). There exist noncommutative groups in which every subgroup is normal (one example is the 'quaternionic group' Q_8; cf. Exercise III.1.12 (iv)), but they are very rare.

Lemma 7.2. *If $\varphi : G \to G'$ is any group homomorphism, then $\ker\varphi$ is a normal subgroup of G.*

Proof. We already know that $\ker\varphi$ is a subgroup of G; to verify it is normal note that $\forall g \in G$, $\forall n \in \ker\varphi$

$$\varphi(gng^{-1}) = \varphi(g)\varphi(n)\varphi(g^{-1}) = \varphi(g)e_{G'}\varphi(g)^{-1} = e_{G'},$$

proving that $gng^{-1} \in \ker\varphi$. □

Loosely speaking, therefore, *kernel* $\Longrightarrow$ *normal*. In fact more is true, as we will see in a little while; for now I don't want to spoil the surprise for the reader. (Can the reader guess?)

There is a convenient shorthand to express conditions such as normality: if $g \in G$ and $A \subseteq G$ is any subset, we denote by gA, Ag, respectively, the following subsets of G:

$$gA := \{h \in G \,|\, (\exists a \in A) : h = ga\},$$

$$Ag := \{h \in G \mid (\exists a \in A) : h = ag\}.$$

Then the normality condition can be expressed by

$$(\forall g \in G): \quad gNg^{-1} \subseteq N,$$

or in a number of other ways:

$$gNg^{-1} = N \quad \text{or} \quad gN \subseteq Ng \quad \text{or} \quad gN = Ng$$

for all $g \in G$. The reader should check that these are indeed equivalent conditions (Exercise 7.3) and keep in mind that '$gN = Ng$' does *not* mean that g commutes with every element of N; it means that if $n \in N$, then there are elements $n', n'' \in N$, in general different from n, such that $gn = n'g$ (so that $gN \subseteq Ng$) and $ng = gn''$ (so that $Ng \subseteq gN$).

7.2. Quotient group. Recall that we have the notion of a quotient of a *set* by an equivalence relation (§I.1.5) and that this notion satisfies a universal property (clumsily stated in §I.5.3). It is natural to investigate this notion in Grp.

We consider then an equivalence relation $\sim$ on (the set underlying) a group G; we seek a group $G/\!\sim$ and a group homomorphism $\pi : G \to G/\!\sim$ satisfying the appropriate universal property, that is, initial with respect to group homomorphisms $\varphi : G \to G'$ such that $a \sim b \implies \varphi(a) = \varphi(b)$.

It is natural to try to construct the *group* $G/\!\sim$ by defining an operation $\bullet$ on the *set* $G/\!\sim$. The situation is tightly constrained by the requirement that the quotient map $\pi : G \to G/\!\sim$ (as in §I.2.6) be a group homomorphism: for if $[a] = \pi(a)$, $[b] = \pi(b)$ are elements of $G/\!\sim$ (that is, equivalence classes with respect to $\sim$), then the homomorphism condition *forces*

$$[a] \bullet [b] = \pi(a) \bullet \pi(b) = \pi(ab) = [ab].$$

But is this operation well-defined? This amounts to conditions on the equivalence relation, which we proceed to unearth.

For the operation to be well-defined 'in the first factor', it is necessary that if $[a] = [a']$, then $[ab] = [a'b]$ regardless of what b is; that is,

$$(\forall g \in G): \quad a \sim a' \implies ag \sim a'g.$$

Similarly, for the operation to be well-defined in the second factor we need

$$(\forall g \in G): \quad a \sim a' \implies ga \sim ga'.$$

Luckily, this is all that there is to it:

Proposition 7.3. *With notation as above, the operation*

$$[a] \bullet [b] := [ab]$$

defines a group structure on $G/\!\sim$ *if and only if* $\forall a, a', g \in G$

$$a \sim a' \implies ga \sim ga' \text{ and } ag \sim a'g.$$

In this case the quotient function $\pi : G \to G/\!\sim$ *is a homomorphism and is universal with respect to homomorphisms* $\varphi : G \to G'$ *such that* $a \sim a' \implies \varphi(a) = \varphi(a')$.

Proof. We have already noted that the condition is necessary. To prove it is sufficient, with the stated consequences, assume $\forall a, a', g \in G$

$$a \sim a' \implies ga \sim ga' \text{ and } ag \sim a'g.$$

Then the operation

$$[a] \bullet [b] := [ab]$$

is well-defined, and we have to verify that it defines a group structure on $G/\sim$. The associativity of $\bullet$ is inherited from the associativity of G: $\forall a, b, c \in G$

$$([a] \bullet [b]) \bullet [c] = [ab] \bullet [c] = [(ab)c] = [a(bc)] = [a] \bullet [bc] = [a] \bullet ([b] \bullet [c]).$$

The class $[e_G]$ is an identity with respect to this operation: $\forall g \in G$

$$[g] \bullet [e_G] = [ge_G] = [g], \quad [e_G] \bullet [g] = [e_G g] = [g].$$

The class $[g^{-1}]$ is the inverse of $[g]$:

$$[g^{-1}] \bullet [g] = [g^{-1}g] = [e_G], \quad [g] \bullet [g^{-1}] = [gg^{-1}] = [e_G].$$

This shows $G/\sim$ is indeed a group, and we have already observed that $\pi : G \to G/\sim$ is a homomorphism: this is what led us to the definition of $\bullet$.

To prove that $G/\sim$ satisfies the universal property, assume

$$\varphi : G \to G'$$

is a group homomorphism such that $a \sim a' \implies \varphi(a) = \varphi(a')$. Since (cf. §I.5.3) the *set* $G/\sim$ satisfies the corresponding universal property in Set, we know that there exists a unique set-function

$$\tilde{\varphi} : G/\sim \to G',$$

defined[31] by $\tilde{\varphi}([a]) := \varphi(a)$. So we only need to check that this function $\tilde{\varphi}$ is in fact a group homomorphism, and this is immediate:

$$\tilde{\varphi}([a] \bullet [b]) = \tilde{\varphi}([ab]) = \varphi(ab) = \varphi(a)\varphi(b) = \tilde{\varphi}([a])\tilde{\varphi}([b])$$

for all $[a], [b] \in G/\sim$, as needed. □

We will say that $\sim$ is *compatible with the group structure* of G if the condition given in Proposition 7.3 holds. Since the operation $\bullet$ on the quotient $G/\sim$ is uniquely determined by the operation on G, we yield to the usual abuse of language and omit it. If $\sim$ is compatible, we will call $G/\sim$ the *quotient group* of G by $\sim$.

7.3. Cosets. The conditions obtained in Proposition 7.3,

(†) $$(\forall g \in G): \quad a \sim b \implies ga \sim gb,$$

(††) $$(\forall g \in G): \quad a \sim b \implies ag \sim bg,$$

lead to a complete description of *all* compatible relations on a group G. In fact, each of these two conditions leads to a description of the relations satisfying it, and we will analyze them separately; the reader should keep in mind that we have a group structure on $G/\sim$ only if *both* are satisfied.

Let's begin with (†). Here is the description:

[31] The point of §I.5.3 is precisely that this function is well-defined.

Proposition 7.4. *Let $\sim$ be an equivalence relation on a group G, satisfying (†). Then*

- *the equivalence class of e_G is a subgroup H of G; and*
- $a \sim b \iff a^{-1}b \in H \iff aH = bH$.

Proof. Let $H \subseteq G$ be the equivalence class of the identity; $H \neq \emptyset$ as $e_G \in H$. For $a, b \in H$, we have $e_G \sim b$ and hence $b^{-1} \sim e_G$ (applying (†), multiplying on the left by b^{-1}); hence $ab^{-1} \sim a$ (by (†) again, multiplying on the left by a); and hence

$$ab^{-1} \sim a \sim e_G$$

by the transitivity of $\sim$ and since $a \in H$. This shows that $ab^{-1} \in H$ for all $a, b \in H$, proving that H is a subgroup (by Proposition 6.2).

Next, assume $a, b \in G$ and $a \sim b$. Multiplying on the left by a^{-1}, (†) implies $e_G \sim a^{-1}b$, that is, $a^{-1}b \in H$. Since H is closed under the operation, this implies $a^{-1}bH \subseteq H$, hence $bH \subseteq aH$; as $\sim$ is symmetric, the same reasoning gives $aH \subseteq bH$; and hence $aH = bH$. Thus, we have proved

$$a \sim b \implies a^{-1}b \in H \implies aH = bH.$$

Finally, assume $aH = bH$. Then $a = ae_G \in bH$, and hence $a^{-1}b \in H$. By definition of H, this means $e_G \sim a^{-1}b$. Multiplying on the left by a shows (by (†) again) that $a \sim b$, completing the proof. □

Proposition 7.4 shows that the equivalence classes of an equivalence relation satisfying (†) are in fact all in the form

$$aH$$

for a fixed subgroup H, as a ranges in G. These important subsets determined by a subgroup H deserve a name.

Definition 7.5. The *left-cosets* of a subgroup H in a group G are the sets aH, for $a \in G$. The *right-cosets* of H are the sets Ha, $a \in G$. ⌟

Now, a 'converse' to Proposition 7.4 holds:

Proposition 7.6. *If H is any subgroup of a group G, the relation $\sim_L$ defined by*

$$(\forall a, b \in G): \quad a \sim_L b \iff a^{-1}b \in H$$

is an equivalence relation satisfying (†).

Proof. This is straightforward and is mostly left to the reader (Exercise 7.8). To see that the relation satisfies (†), note that

$$a \sim_L b \implies a^{-1}b \in H \implies a^{-1}(g^{-1}g)b \in H \implies (ga)^{-1}(gb) \in H \implies ga \sim_L gb$$

for all $g \in G$. □

Taken together, Propositions 7.4 and 7.6 show

Proposition 7.7. *There is a one-to-one correspondence between subgroups of G and equivalence relations on G satisfying (†); for the relation $\sim_L$ corresponding to a subgroup H, $G/\sim_L$ may be described as the set of* left-cosets *aH of H.*

The reader should have no difficulty producing the mirror statements (and proofs) giving a similarly exhaustive description of all equivalence relations satisfying (††). The end result will be

Proposition 7.8. *There is a one-to-one correspondence between subgroups of G and equivalence relations on G satisfying* (††)*; for the relation $\sim_R$ corresponding to a subgroup H, $G/\sim_R$ may be described as the set of* right-cosets Ha *of* H.

The relation corresponding to H in this *second* way is defined by

$$a \sim_R b \iff ab^{-1} \in H \iff Ha = Hb.$$

What may be surprising at first is that the relations $\sim_L$ and $\sim_R$ corresponding to the *same* subgroup H may very well *not* be the same relation. That is, left-cosets and right-cosets of a subgroup need not coincide. Of course $eH = He = H$, and more generally

$$(\forall h \in H): \quad hH = Hh = H.$$

Further

$$(\forall a \in G): \quad a \in aH \cap Ha;$$

hence, if $aH = Hb$, then in fact necessarily $aH = Ha$. This is of course automatically true if G is commutative, but it is simply not the case in general.

Example 7.9. Let $G = S_3$, and let H be the subgroup consisting of the identity and the $1 \leftrightarrow 2$ switch:

$$H = \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} \right\}.$$

Then

$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} H = \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \right\},$$

while

$$H \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} = \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \right\}.$$

This state of affairs simply reflects the fact that the two conditions (†) and (††) are different: there is no reason to expect that if one holds, the other one should also hold (unless G is commutative, of course). Once more, keep in mind that *both* have to hold for the quotient $G/\sim$ to be a group, compatibly with the operation in G; cf. Proposition 7.3.

7.4. Quotient by normal subgroups. To stress the main point again, an arbitrary subgroup H of G leads to two partitions of G, which we have denoted

$$G/\sim_L = \{aH \mid a \in G\}, \quad G/\sim_R = \{Ha \mid a \in G\}.$$

The relation $\sim_L$ satisfies property (†) listed at the beginning of §7.3; $\sim_R$ satisfies (††). A priori, these relations, and hence the corresponding partitions, are different.

The condition that $\sim_L$ and $\sim_R$ *coincide* (as is necessarily the case, for example, if G is commutative) translates into a condition on H: for such 'special' subgroups, (†) and (††) get back together. The good news is that this condition is easy to identify and is not new to the reader.

Proposition 7.10. *The relations $\sim_L$, $\sim_R$ corresponding to a subgroup H coincide if and only if H is normal.*

Proof. Two relations coincide if the corresponding partitions agree. Therefore

$$\sim_L = \sim_R \iff \text{left- and right-cosets of } H \text{ coincide} \iff (\forall g \in G) : gH = Hg.$$

But this is one of the equivalent conditions defining the notion of normal subgroup (cf §7.1), proving the statement. □

The innocent-looking Proposition 7.10 is of fundamental importance. If H is normal, then the *one* equivalence relation $\sim = \sim_L = \sim_R$ corresponding to H satisfies both (†) and (††) (by Proposition 7.4 and its mirror statement), and hence (by Proposition 7.3) the quotient

$$G/\sim \; = \{aH \mid a \in G\} = \{Ha \mid a \in G\}$$

has a natural group structure.

Definition 7.11. Let H be a normal subgroup of a group G. The *quotient group of G modulo H*, denoted[32] G/H, is the group $G/\sim$ obtained from the relation $\sim$ defined above. In terms of (left-) cosets, the product in G/H is defined by

$$(aH)(bH) := (ab)H.$$

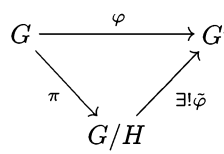
The identity element $e_{G/H}$ of the quotient group G/H is the coset of the identity, $e_G H = H$. ⌟

By Proposition 7.3, the quotient function

$$\pi : G \to G/H$$

sending $g \in G$ to $gH = Hg$ is a group homomorphism and is universal with respect to group homomorphisms $\varphi : G \to G'$ such that $aH = bH \implies \varphi(a) = \varphi(b)$. This universal property is extremely useful, so we will grace it with theorem status:

Theorem 7.12. *Let H be a normal subgroup of a group G. Then for every group homomorphism $\varphi : G \to G'$ such that $H \subseteq \ker\varphi$ there exists a unique group homomorphism $\tilde{\varphi} : G/H \to G'$ so that the diagram*



commutes.

Proof. We only need to match the stated universal property with the one we proved in Proposition 7.3, and indeed,

$$H \subseteq \ker\varphi \iff (\forall h \in H) : \varphi(h) = e_{G'}$$

is equivalent to

$$(\forall a, b \in G) : ab^{-1} \in H \implies \varphi(ab^{-1}) = e_{G'}$$

[32] In a large display we sometime use the full 'fraction' notation $\frac{G}{H}$.

that is, to

$$(\forall a, b \in G) : ab^{-1} \in H \implies \varphi(a) = \varphi(b)$$

and finally, keeping in mind how the relation $\sim$ corresponding to H is defined,

$$(\forall a, b \in G) : a \sim b \implies \varphi(a) = \varphi(b),$$

the condition giving the universal property in Proposition 7.3. □

7.5. Example. The reader is already very familiar with an important class of examples: the cyclic groups $\mathbb{Z}/n\mathbb{Z}$. Indeed, in §2.3 we defined $\mathbb{Z}/n\mathbb{Z}$ as the set of equivalence classes in $\mathbb{Z}$ with respect to the congruence equivalence relation

$$(\forall a, b \in \mathbb{Z}) : \quad a \equiv b \mod n \iff n \mid (b - a).$$

Now we recognize that $n \mid (b - a)$ is equivalent to

$$b - a \in n\mathbb{Z},$$

which is the relation $\sim_L$ corresponding (in 'abelian' notation) to the subgroup $n\mathbb{Z}$ of $\mathbb{Z}$. This subgroup is of course normal, since $\mathbb{Z}$ is abelian. The 'congruence classes mod n' are nothing but the cosets of the subgroup $n\mathbb{Z}$ in $\mathbb{Z}$; using abelian notation for cosets, we could write

$$[a]_n = a + (n\mathbb{Z}).$$

Of course the operation defined on $\mathbb{Z}/n\mathbb{Z}$ in §2.3 matches precisely the one defined above for quotient groups. This justifies the notation $\mathbb{Z}/n\mathbb{Z}$ introduced in §2.3.

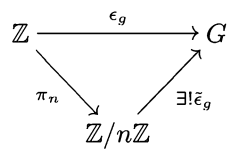
The reader can already appreciate in this simple context the usefulness of Theorem 7.12. Let $g \in G$ be an element of order n and consider the exponential map

$$\epsilon_g : \mathbb{Z} \to G, \quad N \mapsto g^N.$$

By Corollary 1.11,

$$\ker \epsilon_g = \{N \in Z \mid N \text{ is a multiple of } |g|\} = n\mathbb{Z}.$$

Theorem 7.12 then implies right away that ϵ_g factors through the quotient:



That is, there is an induced map

$$\mathbb{Z}/n\mathbb{Z} \to \langle g \rangle.$$

In fact, the 'canonical decomposition' of §I.2.8 implies that this is an isomorphism (verifying that $\langle g \rangle$ is cyclic in the sense of Definition 4.7, as the reader should have checked 'by hand' already in Exercise 6.4). We will formalize this observation in general in the next section.

Also note that $|g| = n = |\langle g \rangle|$ in this case.

7.6. kernel $\iff$ normal. If H is a *normal* subgroup, we have now constructed in gory detail a group G/H and a surjective homomorphism

$$\pi : G \to G/H.$$

What is the kernel of π? The identity of G/H is the coset $e_G H$, that is, H itself. Therefore

$$\ker \pi = \{g \in G \mid gH = H\} = H.$$

This observation completes the circle of ideas begun in §7.1: there we had noticed that every kernel (of a group homomorphism) is a *normal* subgroup; and now we have verified that every normal subgroup is in fact a *kernel* (of some group homomorphism). We encapsulate this in the slogan

$$\text{kernel} \iff \text{normal}:$$

in group theory[33], 'kernel' and 'normal subgroup' are equivalent concepts.

For example, every subgroup in an *abelian* group is the kernel of some homomorphism: yet another indication that life is simpler in Ab than in Grp.

Exercises

7.1. ▷ List all subgroups of S_3 (cf. Exercise 6.13) and determine which subgroups are normal and which are not normal. [§7.1]

7.2. Is the *image* of a group homomorphism necessarily a *normal* subgroup of the target?

7.3. ▷ Verify that the equivalent conditions for normality given in §7.1 are indeed equivalent. [§7.1]

7.4. Prove that the relation defined in Exercise 5.10 on a free abelian group $F = F^{ab}(A)$ is compatible with the group structure. Determine the quotient $F/\sim$ as a better known group.

7.5. ¬ Define an equivalence relation $\sim$ on $\mathrm{SL}_2(\mathbb{Z})$ by letting $A \sim A' \iff A' = \pm A$. Prove that $\sim$ is compatible with the group structure. The quotient $\mathrm{SL}_2(\mathbb{Z})/\sim$ is denoted $\mathrm{PSL}_2(\mathbb{Z})$ and is called the *modular group;* it would be a serious contender in a contest for 'the most important group in mathematics', due to its role in algebraic geometry and number theory. Prove that $\mathrm{PSL}_2(\mathbb{Z})$ is generated by the (cosets of the) matrices

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}.$$

(You will not need to work very hard, if you use the result of Exercise 6.10.) Note that the first has order 2 in $\mathrm{PSL}_2(\mathbb{Z})$, the second has order 3, and their product has infinite order. [9.14]

[33] We will run into analogous observations in *ring* theory, where we will verify that kernels and *ideals* coincide, and for *modules,* as kernels and submodules again coincide.

7.6. Let G be a group, and let n be a positive integer. Consider the relation

$$a \sim b \iff (\exists g \in G)\, ab^{-1} = g^n.$$

- Show that in general $\sim$ is *not* an equivalence relation.
- Prove that $\sim$ is an equivalence relation if G is commutative, and determine the corresponding subgroup of G.

7.7. Let G be a group, n a positive integer, and let $H \subseteq G$ be the subgroup generated by all elements of order n in G. Prove that H is normal.

7.8. ▷ Prove Proposition 7.6. [§7.3]

7.9. State and prove the 'mirror' statements of Propositions 7.4 and 7.6, leading to the description of relations satisfying (††).

7.10. ¬ Let G be a group, and $H \subseteq G$ a subgroup. With notation as in Exercise 6.7, show that H is normal in G if and only if $\forall \gamma \in \mathrm{Inn}(G)$, $\gamma(H) \subseteq H$.

Conclude that if H is normal in G, then there is an interesting homomorphism $\mathrm{Inn}(G) \to \mathrm{Aut}(H)$. [8.25]

7.11. ▷ Let G be a group, and let $[G,G]$ be the subgroup of G generated by all elements of the form $aba^{-1}b^{-1}$. (This is the *commutator* subgroup of G; we will return to it in §IV.3.3.) Prove that $[G,G]$ is normal in G. (Hint: With notation as in Exercise 4.8, $g \cdot aba^{-1}b^{-1} \cdot g^{-1} = \gamma_g(aba^{-1}b^{-1})$.) Prove that $G/[G,G]$ is commutative. [7.12, §IV.3.3]

7.12. ▷ Let $F = F(A)$ be a free group, and let $f : A \to G$ be a set-function from the set A to a *commutative* group G. Prove that f induces a unique homomorphism $F/[F,F] \to G$, where $[F,F]$ is the commutator subgroup of F defined in Exercise 7.11. (Use Theorem 7.12.) Conclude that $F/[F,F] \cong F^{ab}(A)$. (Use Proposition I.5.4.) [§6.4, 7.13, VI.1.20]

7.13. ¬ Let A, B be sets and $F(A)$, $F(B)$ the corresponding free groups. Assume $F(A) \cong F(B)$. If A is finite, prove that B is also and $A \cong B$. (Use Exercise 7.12 to upgrade Exercise 5.10.) [5.10, VI.1.20]

---

8. Canonical decomposition and Lagrange's theorem

We will collect in this section a number of observations on the structure of quotient groups. All these results are straightforward, given the background work done so far. Some of them are often given fancy names such as *first isomorphism theorem* in the literature; we are not too fond of such terminology: the universal property proven in Theorem 7.12 is really the only thing we need to take along, and it serves us wonderfully well. The 'isomorphism theorems' are all immediate applications of this universal property.

8.1. Canonical decomposition. The first observation comes from the *canonical decomposition* for set-functions, obtained in §I.2.8: every set-functions may be viewed as the composition of a surjective map, followed by a bijective map, followed by an injective map. We now know enough to state the corresponding (very useful) results in Grp:

Theorem 8.1. *Every group homomorphism $\varphi : G \to G'$ may be decomposed as follows:*

$$G \twoheadrightarrow G/\ker\varphi \xrightarrow[\tilde{\varphi}]{\sim} \operatorname{im}\varphi \hookrightarrow G' \qquad (\text{with } G \xrightarrow{\varphi} G')$$

where the isomorphism $\tilde{\varphi}$ in the middle is the homomorphism induced by φ (as in Theorem 7.12).

It is important that the reader agree that we have already proved anything that deserves to be proven here. We know that the projection on the left and the inclusion on the right are homomorphisms and $\tilde{\varphi}$ comes from Theorem 7.12. The decomposition is the same one obtained at the level of set-functions in §I.2.8; in particular, the function in the middle is a bijection. Since bijective homomorphisms are isomorphisms (Proposition 4.3), it is an isomorphism.

Theorem 8.1 should induce the following Pavlovian reaction: exposed to any group homomorphism $\varphi : G \to G'$, the reader should instantaneously view $G/\ker\varphi$ as (canonically identified with) a subgroup of G'. What is usually called the *'first isomorphism theorem'* is the particular case corresponding to surjective homomorphisms:

Corollary 8.2. *Suppose $\varphi : G \to G'$ is a* surjective *group homomorphism. Then*

$$G' \cong \frac{G}{\ker\varphi}.$$

Proof. $im\varphi = G'$ in Theorem 8.1. □

This result is very useful—it comes in extremely handy when proving that two groups are isomorphic, both in theoretical contexts (as we will see in the rest of this section) and in concrete instances.

Example 8.3. If $H_1 \subseteq G_1$ and $H_2 \subseteq G_2$, then the product $H_1 \times H_2$ may be viewed as a subset of $G_1 \times G_2$. It is clear that if G_1, G_2 are groups and H_1, H_2 are subgroups, then $H_1 \times H_2$ is a subgroup of $G_1 \times G_2$. The following claim is a prototype application of Corollary 8.2:

Claim 8.4. *If $H_1 \subseteq G_1$ and $H_2 \subseteq G_2$ are* normal *subgroups, then $H_1 \times H_2$ is a normal subgroup of the group $G_1 \times G_2$ and*

$$\frac{G_1 \times G_2}{H_1 \times H_2} \cong \frac{G_1}{H_1} \times \frac{G_2}{H_2}.$$

Indeed, composing the projections

$$\pi_1 : G_1 \times G_2 \to G_1, \quad \pi_2 : G_1 \times G_2 \to G_2$$

with the morphisms to the quotients gives surjective homomorphisms

$$\pi_1 : G_1 \times G_2 \to \frac{G_1}{H_1}, \quad \pi_2 : G_1 \times G_2 \to \frac{G_2}{H_2}$$

and hence a homomorphism

$$\pi : G_1 \times G_2 \to \frac{G_1}{H_1} \times \frac{G_2}{H_2}$$

by the universal property of products. Explicitly,

$$\pi(g_1, g_2) = (g_1 H_1, g_2 H_2) :$$

in particular, π is surjective and

$$\begin{aligned} \ker \pi &= \{(g_1, g_2) \in G_1 \times G_2 \,|\, (g_1 H_1, g_2 H_2) = (H_1, H_2)\} \\ &= \{(g_1, g_2) \in G_1 \times G_2 \,|\, g_1 \in H_1, g_2 \in H_2\} \\ &= H_1 \times H_2. \end{aligned}$$

The claim then follows immediately from Corollary 8.2.

The result (of course) extends to more factors in the product. Any such check should become second nature and is usually left to the reader. ⌟

Example 8.5. As a particular case of Claim 8.4, take $H_1 = \{e_{G_1}\} \subseteq G_1$ and $H_2 = G_2 \subseteq G_2$:

$$\frac{G_1 \times G_2}{G_2} \cong \frac{G_1}{\{e_{G_1}\}} \times \frac{G_2}{G_2} \cong G_1,$$

where on the left we identify G_2 with the subgroup $\{e_{G_1}\} \times G_2$. For instance[34] (cf. §4.1)

$$\frac{C_6}{C_3} \cong \frac{C_2 \times C_3}{C_3} \cong C_2.$$ ⌟

Example 8.6. The cyclic group C_3 may be viewed as a subgroup of the dihedral group D_6: the rotations of a triangle give a copy of C_3 inside D_6. Then C_3 is normal in D_6, and

$$\frac{D_6}{C_3} \cong C_2.$$

This can of course be checked 'by hand'. But note that there is an evident surjective homomorphism $D_6 \to C_2$, whose kernel is C_3: map an element σ of D_6 to the identity in C_2 if it does *not* flip the triangle (that is, precisely when $\sigma \in C_3$), and map it to the other element if it does. Corollary 8.2 implies the stated facts immediately. ⌟

Example 8.7. One can give a *circle* (denoted S^1) a group structure by identifying its points with rotations of a plane about a point and adding them accordingly. The function

$$\rho : \mathbb{R}^1 \to S^1$$

[34] Abuses of language such as the formula which follows—in which one is not explicitly specifying how to realize C_3 as a subgroup of C_6, because there is really only one way to do it—are unfortunately commonplace.

mapping a number r to the result of a rotation by $2\pi r$ radians is then a surjective group homomorphism; this is the identity precisely when we rotate by an *integer* multiple of 2π. Hence

$$\ker \rho = \mathbb{Z} \subseteq \mathbb{R}.$$

By Corollary 8.2, therefore,

$$\frac{\mathbb{R}}{\mathbb{Z}} \cong S^1.$$

(Cf. Exercise I.1.6.) Geometrically, this amounts to 'wrapping' $\mathbb{R}$ infinitely many times around the circle, realizing $\mathbb{R}$ as the 'universal cover' of S^1; here, $\mathbb{Z}$ plays the role of 'fundamental group' of S^1. ⌟

8.2. Presentations. Every group is a quotient of a free group, and every abelian group is a quotient of a free abelian group. Indeed, every group G can be surjected upon by a free group, and in many ways (at the very least, $F(G)$ will do!). Abelian groups may be likewise surjected upon by free abelian groups. Then Corollary 8.2 produces the needed isomorphism of G with a quotient of a free group.

A *presentation* of a group G is an explicit isomorphism

$$G \cong \frac{F(A)}{R}$$

where A is a set and R is a subgroup of 'relations'. In other words, a presentation is an explicit surjection

$$\rho : F(A) \twoheadrightarrow G$$

of which R is the kernel. This is especially useful if A is small, and R may be described very explicitly; usually this is done by listing 'enough' *relations,* that is, a set $\mathscr{R}$ of words $r_\alpha \in R = \ker \rho$ generating it in the sense that R is the smallest *normal* subgroup[35] of $F(A)$ containing $\mathscr{R}$.

Thus, a presentation of a group G is usually encoded as a pair $(A|\mathscr{R})$, where A is a set and $\mathscr{R} \subseteq F(A)$ is a set of words, such that $G \cong A/R$ with R as above.

A group is *finitely presented* if it admits a presentation $(A, \mathscr{R})$ in which both A and $\mathscr{R}$ are finite. Finitely presented groups are not (necessarily) 'small': for example, the free group on finitely many generators is (trivially) finitely presented.

We have already run into several examples of presentations. For instance, the free group $F(A)$ is presented by $(A|\emptyset)$. More interestingly, the description of S_3 given in §2.1 'presents' S_3 as a quotient of the free group $F(\{x, y\})$ (cf. Example 5.3) by the smallest normal subgroup containing x^2, y^3, and $yx = xy^2$: $(x, y|x^2, y^3, xyxy)$ in shorthand. From this point of view, it is clear that groups admitting the same presentation (example: S_3 and D_6) are isomorphic.

The situation is less idyllic than it may seem at first, though: even if a presentation of a group G is known, it may be very hard to establish whether two explicit

[35]Note that this is a different requirement than the one adopted in §6.3, in which normality plays no role.

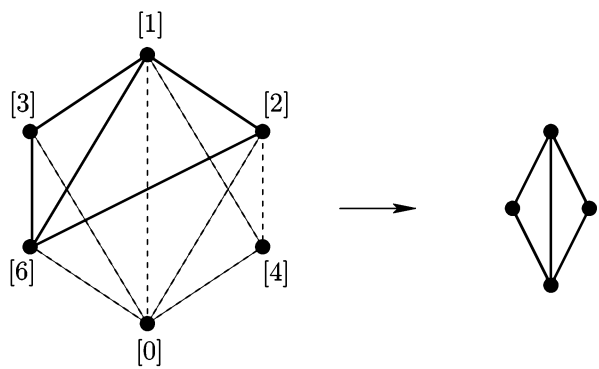
combinations of the generators coincide in G. This is known as the *word problem*, and it has been shown to be undecidable in general[36].

In any case, now that we know about presentations of groups, finding coproducts in Grp should be straightforward: see Exercise 8.7.

There is a 'mirror' statement analogous to the fact that all groups are quotients of free groups: every group may be realized as a *subgroup* of a symmetric group. This elementary observation goes under the name of *Cayley's theorem;* its natural place is within the discussion of *group actions* (cf. Theorem 9.5).

8.3. Subgroups of quotients. The lattice of subgroups (cf. §6.4) of a quotient G/H can be described very explicitly in terms of the lattice of subgroups of the group G: simply keep the part of the lattice of subgroups of G corresponding to subgroups which *contain* H.

Example 8.8. Here is the effect of this operation on the lattice of subgroups of $C_{12} \cong \mathbb{Z}/12\mathbb{Z}$ (labeled by generators; cf. §6.4), after quotienting by $H = \langle [6] \rangle \cong C_2$:



The result matches the lattice of subgroups of $C_6 \cong C_{12}/C_2$. ⌟

Here is why this works. First note that if $H \subseteq K$ are subgroups of a group G and H is normal in G, then H is normal in K.

Proposition 8.9. *Let H be a normal subgroup of a group G. Then for every subgroup K of G containing H, K/H may be identified with a subgroup of G/H. The function*

$$u : \{\textit{subgroups } K \textit{ of } G \textit{ containing } H\} \to \{\textit{subgroups of } G/H\}$$

defined by $u(K) = K/H$ is a bijection preserving inclusions.

Proof. The group K/H consists of the cosets $aH \in G/H$ with $a \in K$, and in this sense it is a subset (and clearly a subgroup) of G/H. It is also clear that if $H \subseteq K \subseteq L$, then $u(K) = K/H \subseteq L/H = u(L)$; that is, u preserves inclusions.

[36] That is, there is no general algorithm that, given a presentation of a group G and two words in the generators, will establish (in a finite time) whether those two words represent the same element of G.

Thus we simply have to verify that u is a bijection, and for this it suffices to produce an inverse function

$$v : \{\text{subgroups of } G/H\} \to \{\text{subgroups } K \text{ of } G \text{ containing } H\}.$$

Then let K' be a subgroup of G/H; define $v(K')$ to be the subset of G:

$$K := \pi^{-1}(K') = \{a \in G \mid aH \in K'\},$$

where $\pi : G \to G/H$ is the canonical projection. Then K is a subgroup of G (by Lemma 6.4) and contains H (because $H = \pi^{-1}(e)$ and $e \in K'$). The reader will check that u and v are inverses of each other. □

In fact, the correspondence is even nicer, in the sense that it preserves *normality*. The following statement is often called the *third isomorphism theorem:*

Proposition 8.10. *Let H be a normal subgroup of a group G, and let N be a subgroup of G containing H. Then N/H is normal in G/H if and only if N is normal in G, and in this case*

$$\frac{G/H}{N/H} \cong \frac{G}{N}.$$

Proof. If N is normal, then consider the projection

$$G \to \frac{G}{N} :$$

the subgroup H is contained in N, which is the kernel of this homomorphism, so we get (by the universal property of quotients, Theorem 7.12) an induced homomorphism

$$\frac{G}{H} \to \frac{G}{N}.$$

The subgroup N/H of G/H is the kernel of this homomorphism; therefore it is normal.

Conversely, if N/H is normal in G/H, consider the composition

$$G \twoheadrightarrow \frac{G}{H} \twoheadrightarrow \frac{G/H}{N/H}.$$

The kernel of this homomorphism is N; therefore N is normal. Further, this homomorphism is surjective; hence the stated isomorphism $(G/H)/(N/H) \cong G/N$ follows immediately from Corollary 8.2. □

8.4. HK/H vs. $K/(H \cap K)$. Section 8.3 deals with two 'nested' subgroups $H \subseteq K$ of a group G. What if H, K are *not* nested?

The notation introduced in §7.1 extends to subsets of G: if $A \subseteq G$, $B \subseteq G$, then AB denotes the subset

$$AB := \{ab \mid a \in A, b \in B\}.$$

It would be nice if HK were guaranteed to be a subgroup of G as soon as H and K are subgroups, but this is simply not the case in general, if G is not commutative. It *is*, however, the case if one of the subgroups is normal. The following is often called the *second isomorphism theorem.*

Proposition 8.11. *Let H, K be subgroups of a group G, and assume that H is normal in G. Then*

- *HK is a subgroup of G, and H is normal in HK;*
- *$H \cap K$ is normal in K, and*

$$\frac{HK}{H} \cong \frac{K}{H \cap K}.$$

Proof. To verify that HK is a subgroup of G when H is normal, note that HK is the union of all cosets Hk, with $k \in K$; that is,

$$HK = \pi^{-1}(\pi(K)),$$

where $\pi : G \to G/H$ is the canonical projection. Since $\pi(K)$ is a subgroup of G/H, HK is a subgroup by Lemma 6.4. It is clear that H is normal in HK.

For the second part, consider the homomorphism

$$\varphi : K \to HK/H$$

sending $k \in K$ to the coset Hk (that is, the inclusion $K \hookrightarrow HK$ followed by the canonical projection to the quotient). This is *surjective:* indeed, every element of HK/H may be written as a coset

$$Hhk, \quad h \in H, k \in K;$$

but $Hhk = Hk$, so $Hhk = \varphi(k)$ is in the image of φ. By the omnipresent Corollary 8.2,

$$\frac{HK}{H} \cong \frac{K}{\ker \varphi}.$$

What is $\ker \varphi$?

$$\ker \varphi = \{k \in K \mid \varphi(k) = e\} = \{k \in K \mid Hk = H\} = \{k \in K \mid k \in H\} = H \cap K,$$

with the stated result. □

8.5. The index and Lagrange's theorem. The notation G/H is used to denote the *set* of *left-cosets*[37] of H, even when H is not normal in G. Thus G/H is a set in general, and it is a group when H is in fact normal in G.

Definition 8.12. The *index* of H in G, denoted $[G : H]$, is the number of elements $|G/H|$ of G/H, when this is finite, and ∞ otherwise. ⌟

Thus, $[G : H]$ (if finite) denotes the number of left-cosets of H in G, regardless of whether H is normal in G.

Lemma 8.13. *Let H be a subgroup of a group G. Then $\forall g \in G$ the functions*

$$H \to gH, \quad h \mapsto gh,$$
$$H \to Hg, \quad h \mapsto hg$$

are bijections.

[37] This may seem an arbitrary choice (why not *right*-cosets?). It is. Writing from left to right gives us a bias towards *left*-actions, and G acts nicely on the *left* on the set of *left*-cosets; this will make better sense when we get to Example 9.4. In any case, there is a bijection between the set of left-cosets and the set of right-coset: Exercise 9.10.

Proof. Both functions are surjective by definition of coset. Cancellation implies that they are injective. □

Corollary 8.14 (Lagrange's theorem). *If G is a* finite *group and $H \subseteq G$ is a subgroup, then $|G| = [G : H] \cdot |H|$. In particular, $|H|$ is a divisor of $|G|$.*

Proof. Indeed, G is the disjoint union of $|G/H|$ distinct cosets gH, and $|gH| = |H|$ by Lemma 8.13. □

Lagrange's theorem is more useful than it may appear at first.

Example 8.15. The order $|g|$ of any element g of a finite group G is a divisor of $|G|$: indeed, $|g|$ equals the order of the subgroup $\langle g \rangle$ generated by g.

Note: Therefore, $g^{|G|} = e_G$ for all finite groups G, all $g \in G$. ⌟

Example 8.16. If $|G|$ is a prime integer p, then necessarily $G \cong \mathbb{Z}/p\mathbb{Z}$.

Indeed, let $g \in G$ be any element other than the identity; then $\langle g \rangle$ is a subgroup of G, of order > 1. By Lagrange's theorem, $|\langle g \rangle| = p = |G|$; that is, $G \cong \langle g \rangle$ is cyclic of order p, as claimed. ⌟

Example 8.17 (Fermat's little theorem). Let p be a prime integer, and let a be any integer. Then $a^p \equiv a \mod p$.

Indeed, this is immediate if a is a multiple of p; if a is not a multiple of p, then the class $[a]_p$ modulo p is nonzero, so it is an element of the group $(\mathbb{Z}/p\mathbb{Z})^*$, which has order $p - 1$. Thus

$$[a]_p^{p-1} = [1]_p$$

(Example 8.15); hence $[a]_p^p = [a]_p$ as claimed. ⌟

Warning: However, do not ask too much of Lagrange's theorem. For example, it does *not* say that if d is a divisor of $|G|$, then there exists a subgroup of G of order d (the smallest counterexample is A_4, a group of order 12, which does not contain subgroups of order 6; the reader will verify this in Exercise IV.4.17); it does not even say that if p is a prime divisor of $|G|$, then there is an element of order p in G. This latter statement happens to be true, but for 'deeper' reasons. The abelian case is easy (cf. Exercise 8.17). The general case is called *Cauchy's theorem,* and we will deal with it later on (cf. Theorem IV.2.1).

The index is a well-behaved invariant. It is clearly *multiplicative,* in the sense that if $H \subseteq K$ are subgroups of G, then

$$[G : H] = [G : K] \cdot [K : H],$$

provided that these numbers are finite. Also, if H and K are subgroups of G and H is normal (so that HK is a subgroup as well; cf. Proposition 8.11), then

$$|HK| = \frac{|H| \cdot |K|}{|H \cap K|}$$

(again, provided this has a chance of making sense, that is, if the orders are finite): this follows immediately from the isomorphism in Proposition 8.11 and index considerations. In fact, the formula holds even without assuming that one of the subgroups is normal in G. Do you see why? (Exercise 8.21.)

Further, if H and G are finite, then Lemma 8.13 implies immediately that the index of H in G, defined as the number of *left*-cosets of H in G, also equals the number of *right*-cosets of H. It is in fact easy to show that there always is a bijection between the set G/H of left-cosets and the set of right-cosets (cf. Exercise 9.10), regardless of finiteness hypotheses. The set of right-cosets of H in G is often (reasonably) denoted $H\backslash G$.

8.6. Epimorphisms and cokernels. The reader may expect that a mirror statement to Proposition 6.12 should hold for group *epimorphisms*. This is almost true: a homomorphism $\varphi : G \to H$ is an *epimorphism* (in the category Grp) if and only if it is *surjective*. However, while one implication is easy, the proofs we know for epimorphism $\Longrightarrow$ surjective in Grp are somewhat cumbersome.

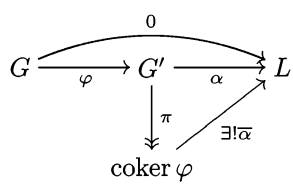
The situation is leaner (as usual) in Ab: there is in Ab a good notion of *cokernel;* this is part of what makes Ab an 'abelian category'.

As is often the case, the reader may now want to pause a moment and try to guess the right definition. Keeping in mind the universal property for kernels (Proposition 6.6), can the reader come up with the universal property defining 'cokernels'? Can the reader prove that these exist in Ab and detect epimorphisms? Don't look ahead!

Here is how the story goes. The universal property is (of course) obtained by reversing the arrows in the property for kernels: given a homomorphism $\varphi : G \to G'$ of *abelian* groups, we want an abelian group $\operatorname{coker}\varphi$ equipped with a homomorphism

$$\pi : G' \to \operatorname{coker}\varphi$$

which is initial with respect to all morphisms α such that $\alpha \circ \varphi = 0$. That is, every homomorphism $\alpha : G \to L$ such that $\alpha \circ \varphi$ is the trivial map must factor (uniquely) through $\operatorname{coker}\varphi$:



Cokernels exist in Ab: because the image of φ is a subgroup of G', hence a *normal* subgroup of G' since G' is abelian; the condition that $\alpha \circ \varphi$ is trivial says that $\operatorname{im}\varphi \subseteq \ker\alpha$, and hence

$$\frac{G'}{\operatorname{im}\varphi} \cong \operatorname{coker}\varphi$$

satisfies the universal property, by Theorem 7.12.

The 'problem' in Grp is that $\operatorname{im}\varphi$ is not guaranteed to be normal in G'; thus the situation is more complex.

Also note that, in the abelian case, $G'/\operatorname{im}\varphi$ automatically satisfies a stronger universal property: as stated, but with respect to any *set-function* $G' \to L$ which is constant on cosets of $\operatorname{im}\varphi$.

We can now state a true mirror of Proposition 6.12, in Ab:

Proposition 8.18. *Let $\varphi : G \to G'$ be a homomorphism of abelian groups. The following are equivalent:*

(a) *φ is an epimorphism;*

(b) *$\operatorname{coker}\varphi$ is trivial;*

(c) *$\varphi : G \to G'$ is surjective (as a set-function).*

Proof. (a) $\Longrightarrow$ (b): Assume (a) holds, and consider the two parallel compositions

$$G \xrightarrow{\varphi} G' \underset{e}{\overset{\pi}{\rightrightarrows}} \operatorname{coker}\varphi,$$

where π is the canonical projection and e is the trivial map. Both $\pi \circ \varphi$ and $e \circ \varphi$ are the trivial map; since φ is an epimorphism, this implies $\pi = e$. But $\pi = e$ implies that $\operatorname{coker}\varphi$ is trivial, that is, (b) holds.

(b) $\Longrightarrow$ (c): If $\operatorname{coker}\varphi = G'/\operatorname{im}\varphi$ is trivial, then $\operatorname{im}\varphi = G'$; hence φ is surjective.

(c) $\Longrightarrow$ (a): If φ is surjective, then it satisfies the universal property for epimorphisms in Set: for any set Z and any two set-functions α' and $\alpha'' : G' \to Z$,

$$\alpha' \circ \varphi = \alpha'' \circ \varphi \iff \alpha' = \alpha''.$$

This must hold in particular if Z is endowed with a group structure and α', α'' are group homomorphisms, so φ is an epimorphism in Grp. □

A cokernel may be defined in Grp: the universal property for the cokernel of $\varphi : G \to G'$ is satisfied by G'/N, where N is the smallest[38] normal subgroup of G' containing $\operatorname{im}\varphi$ (Exercise 8.22). *But* Proposition 8.18 fails, because the implication (b) $\Longrightarrow$ (c) does not hold: in Grp it is no longer true that only surjective homomorphisms have trivial cokernel (cf. Exercise 8.23).

Exercises

8.1. If a group H may be realized as a subgroup of two groups G_1 and G_2 and if

$$\frac{G_1}{H} \cong \frac{G_2}{H},$$

does it follows that $G_1 \cong G_2$? Give a proof or a counterexample.

8.2. ¬ Extend Example 8.6 as follows. Suppose G is a group and $H \subseteq G$ is a subgroup *of index* 2, that is, such that there are precisely two (say, left-) cosets of H in G. Prove that H is normal in G. [9.11, IV.1.16]

8.3. Prove that every finite group is finitely presented.

[38] The intersection of any family of normal subgroups is normal, as the reader may readily check, so this subgroup exists.

8.4. Prove that $(a, b | a^2, b^2, (ab)^n)$ is a presentation of the dihedral group D_{2n}. (Hint: With respect to the generators defined in Exercise 2.5, set $a = x$ and $b = xy$; prove you can get the relations given here from the ones you obtained in Exercise 2.5, and conversely.)

8.5. Let a, b be distinct elements of order 2 in a group G, and assume that ab has finite order $n \geq 3$. Prove that the subgroup generated by a and b in G is isomorphic to the dihedral group D_{2n}. (Use the previous exercise.)

8.6. ¬ Let G be a group, and let A be a set of generators for G; assume A is finite. The corresponding *Cayley graph*[39] is a directed graph whose set of vertices is in one-to-one correspondence with G, and two vertices g_1, g_2 are connected by an edge if $g_2 = g_1 a$ for an $a \in A$; this edge may be labeled a and oriented from g_1 to g_2. For example, the graph drawn in Example 5.3 for the free group $F(\{x, y\})$ on two generators x, y is the corresponding Cayley graph (with the convention that horizontal edges are labeled x and point to the right and vertical edges are labeled y and point up).

Prove that if a Cayley graph of a group is a tree, then the group is free. Conversely, prove that free groups admit Cayley graphs that are trees. [§5.3, 9.15]

8.7. ▷ Let $(A|\mathscr{R})$, resp., $(A'|\mathscr{R}')$, be a presentation for a group G, resp., G' (cf. §8.2); we may assume that A, A' are disjoint. Prove that the group $G * G'$ presented by

$$(A \cup A' \,|\, \mathscr{R} \cup \mathscr{R}')$$

satisfies the universal property for the *coproduct* of G and G' in Grp. (Use the universal properties of both free groups and quotients to construct natural homomorphisms $G \to G * G'$, $G' \to G * G'$.) [§3.4, §8.2, 9.14]

8.8. ¬ (If you know about matrices (cf. Exercise 6.1).) Prove that $\mathrm{SL}_n(\mathbb{R})$ is a *normal subgroup* of $\mathrm{GL}_n(\mathbb{R})$, and 'compute' $\mathrm{GL}_n(\mathbb{R})/\mathrm{SL}_n(\mathbb{R})$ as a well-known group. [VI.3.3]

8.9. ¬ (Ditto.) Prove that $\mathrm{SO}_3(\mathbb{R}) \cong \mathrm{SU}_2(\mathbb{C})/\{\pm I_2\}$, where I_2 is the identity matrix. (Hint: It so happens that every matrix in $\mathrm{SO}_3(\mathbb{R})$ can be written in the form

$$\begin{pmatrix} a^2+b^2-c^2-d^2 & 2(bc-ad) & 2(ac+bd) \\ 2(ad+bc) & a^2-b^2+c^2-d^2 & 2(cd-ab) \\ 2(bd-ac) & 2(ab+cd) & a^2-b^2-c^2+d^2 \end{pmatrix}$$

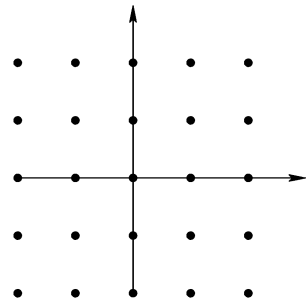
where $a, b, c, d \in \mathbb{R}$ and $a^2+b^2+c^2+d^2 = 1$. Proving this fact is not hard, but at this stage you will probably find it computationally demanding. Feel free to assume this, and use Exercise 6.3 to construct a surjective homomorphism $\mathrm{SU}_2(\mathbb{C}) \to \mathrm{SO}_3(\mathbb{R})$; compute the kernel of this homomorphism.)

If you know a little topology, you can now conclude that the fundamental group[40] of $\mathrm{SO}_3(\mathbb{R})$ is C_2. [9.1, VI.1.3]

[39] Warning: This is one of several alternative conventions.

[40] If you really want to believe this fact, remember that $\mathrm{SO}_3(\mathbb{R})$ parametrizes rotations in $\mathbb{R}^3$. Hold a tray with a glass of water on top of your extended right hand. You should be able to rotate the tray clockwise by a full 360° without spilling the water, and your muscles will tell you that the corresponding loop in $SO_3(\mathbb{R})$ is *not* trivial. But then you will be able to rotate the tray *again*

8.10. View $\mathbb{Z} \times \mathbb{Z}$ as a subgroup of $\mathbb{R} \times \mathbb{R}$:



Describe the quotient

$$\frac{\mathbb{R} \times \mathbb{R}}{\mathbb{Z} \times \mathbb{Z}}$$

in terms analogous to those used in Example 8.7. (Can you 'draw a picture' of this group? Cf. Exercise I.1.6.)

8.11. (Notation as in Proposition 8.10.) Prove 'by hand' (that is, without invoking universal properties) that N is normal in G if and only if N/H is normal in G/H.

8.12. (Notation as in Proposition 8.11.) Prove 'by hand' (that is, by using Proposition 6.2) that HK is a subgroup of G if H is normal.

8.13. ¬ Let G be a finite commutative group, and assume $|G|$ is odd. Prove that every element of G is a square. [8.14]

8.14. Generalize the result of Exercise 8.13: if G is a group of order n and k is an integer relatively prime to n, then the function $G \to G$, $g \mapsto g^k$ is surjective.

8.15. Let a, n be positive integers. Prove that n divides $\phi(a^n - 1)$, where ϕ is Euler's ϕ-function; see Exercise 6.14. (Hint: Example 8.15.)

8.16. Generalize Fermat's little theorem to congruences modulo arbitrary (that is, possibly nonprime) integers. Note that it is *not* true that $a^n \equiv a \bmod n$ for all a and n: for example, 2^4 is not congruent to 2 modulo 4. *What* is true? (This generalization is known as *Euler's theorem.*)

8.17. ▷ Assume G is a finite abelian group, and let p be a prime divisor of $|G|$. Prove that there exists an element in G of order p. (Hint: Let g be any element of G, and consider the subgroup $\langle g \rangle$; use the fact that this subgroup is cyclic to show that there is an element $h \in \langle g \rangle$ in G of *prime* order q. If $q = p$, you are done; otherwise, use the quotient $G/\langle h \rangle$ and induction.) [§8.5, 8.18, 8.20, §IV.2.1]

8.18. Let G be an abelian group of order $2n$, where n is odd. Prove that G has *exactly one* element of order 2. (It has at least one, for example by Exercise 8.17. Use Lagrange's theorem to establish that it cannot have more than one.) Does the same conclusion hold if G is not necessarily commutative?

---

a full 360° clockwise without spilling any water, taking it back to the original position. Thus, the square of the loop *is* (homotopically) trivial, as it should be if the fundamental group is cyclic of order 2.

8.19. Let G be a finite group, and let d be a proper divisor of $|G|$. Is it necessarily true that there exists an element of G of order d? Give a proof or a counterexample.

8.20. ▷ Assume G is a finite abelian group, and let d be a divisor of $|G|$. Prove that there exists a *subgroup* $H \subseteq G$ of order d. (Hint: induction; use Exercise 8.17.) [§IV.2.2]

8.21. ▷ Let H, K be subgroups of a group G. Construct a bijection between the set of cosets hK with $h \in H$ and the set of left-cosets of $H \cap K$ in H. If H and K are finite, prove that

$$|HK| = \frac{|H| \cdot |K|}{|H \cap K|}.$$

[§8.5]

8.22. ▷ Let $\varphi : G \to G'$ be a group homomorphism, and let N be the smallest normal subgroup containing $\operatorname{im}\varphi$. Prove that G'/N satisfies the universal property of $\operatorname{coker}\varphi$ in Grp. [§8.6]

8.23. ▷ Consider the subgroup

$$H = \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} \right\}$$

of S_3. Show that the cokernel of the inclusion $H \hookrightarrow S_3$ is trivial, although $H \hookrightarrow S_3$ is not surjective. [§8.6]

8.24. ▷ Show that epimorphisms in Grp do not necessarily have right-inverses. [§I.4.2]

8.25. Let H be a commutative normal subgroup of G. Construct an interesting homomorphism from G/H to $\operatorname{Aut}(H)$. (Cf. Exercise 7.10.)

---

9. Group actions

9.1. Actions. As mentioned in §4.1, an *action* of a group G on an object A of a category C is simply a homomorphism

$$\sigma : G \to \operatorname{Aut}_{\mathsf{C}}(A).$$

The way to interpret this is that every element $g \in G$ determines a 'transformation of A into itself', i.e., an isomorphism of A in C, and this happens compatibly with the operation of G and composition in C.

In a rather strong sense, we really only care about groups because they act on things: knowing that G acts on A tells us something about A; group actions are one key tool in the study of geometric and algebraic entities.

In fact, group actions are one key tool in the study of *groups* themselves: one of the best ways to 'understand' a group is to let it act on an object A, hoping that the corresponding homomorphism σ is an isomorphism, or at least an injective monomorphism. For example, we were lucky with D_6 in §2.2: we let D_6 act on a set with three elements (the vertices of an equilateral triangle) and observed that the resulting σ is an isomorphism. Thus $D_6 \cong S_3$. We would be almost as lucky

by letting D_8 act on the vertices of a square: then σ would at least realize D_8 as an explicit *subgroup* of S_4, which simplifies its analysis.

Definition 9.1. An action of a group G on an object A of a category C is *faithful* (or *effective*) if the corresponding $\sigma : G \to \mathrm{Aut}_{\mathsf{C}}(A)$ is injective. ⌟

The case $\mathsf{C} = \mathsf{Set}$ is already very rich, and we focus on it in this chapter.

9.2. Actions on sets. Spelling out our definition of action in case A is a *set*, so that $\mathrm{Aut}_{\mathsf{C}}(A)$ is the symmetric group S_A, we get the following:

Definition 9.2. An *action* of a group G on a set A is a set-function

$$\rho : G \times A \to A$$

such that $\rho(e_G, a) = a$ for all $a \in A$ and

$$(\forall g, h \in G), (\forall a \in A) : \quad \rho(gh, a) = \rho(g, \rho(h, a)).$$ ⌟

Indeed, given a function ρ satisfying these conditions, we can define $\sigma : G \to \mathrm{Hom}_{\mathsf{Set}}(A, A)$ by $\sigma(g)(a) = \rho(g, a)$. (This defines $\sigma(g)$ as a set-function $A \to A$, as needed.) This function preserves the operation, because

$$\begin{aligned}\sigma(gh)(a) &= \rho(gh, a) = \rho(g, \rho(h, a)) = \sigma(g)(\rho(h, a)) = \sigma(g)(\sigma(h)(a)) \\ &= \sigma(g) \circ \sigma(h)(a).\end{aligned}$$

In particular, this verifies that $\sigma(g^{-1})$ acts as the inverse of $\sigma(g)$: because $(\forall a \in A)$

$$\sigma(g^{-1}) \circ \sigma(g)(a) = \sigma(g^{-1}g)(a) = \sigma(e_G)(a) = \rho(e_G, a) = a.$$

Thus the image of σ consists of invertible set-functions; σ is acting as a function

$$\sigma : G \to S_A,$$

and we have verified that this is a homomorphism, as needed.

Conversely, given a homomorphism $\sigma : G \to S_A$, define $\rho : G \times A \to A$ by $\rho(g, a) = \sigma(g)(a)$; the same argument (read backwards) shows that ρ satisfies the needed properties.

It is unpleasant to carry ρ along. In practice, one just writes ga for $\rho(g, a)$; the requirements in Definition 9.2 amount then to $e_G a = a$ for all $a \in A$ and

$$(\forall g, h \in G), (\forall a \in A) : \quad (gh)a = g(ha),$$

'as if' ρ defined an *associative* operation.

If G acts on A, then $e_G a = a$ for all $a \in A$; the action of a group G on a set A is faithful if and only if the identity e_G is the *only* element g of G such that $ga = a$ for all $a \in A$, that is, 'fixing' every element of A. An action is *free* if the identity e_G is the only element fixing *any* element of A.

Example 9.3. Every *group* G acts in a natural way on the underlying *set* G. The function $\rho : G \times G \to G$ is simply the operation in the group:

$$(\forall g, a \in G) : \quad \rho(g, a) = ga.$$

In this case the defining property *really* is associativity. This is referred to as the action *by left-multiplication*[41] of G on itself. There is (at least) another very natural way to act with G on itself, by *conjugation:* define $\rho : G \times G \to G$ by

$$\rho(g,h) = ghg^{-1}.$$

This is indeed an action: $\forall g,h,k \in G$,

$$\rho(g,\rho(h,k)) = g\rho(h,k)g^{-1} = g(hkh^{-1})g^{-1} = (gh)k(gh)^{-1} = \rho(gh,k). \quad \lrcorner$$

Example 9.4. More generally, G acts by left-multiplication on the set G/H of *left*-cosets (cf. §8.5) of any subgroup H: act by $g \in G$ on $aH \in G/H$ by sending it to $(ga)H$. $\lrcorner$

These examples of actions are extremely useful in studying groups, as we will see in Chapter IV. For instance, an immediate consequence is the following counterpart to §8.2:

Theorem 9.5 (Cayley's theorem)**.** *Every group acts* faithfully *on some set. That is, every group may be realized as a subgroup of a permutation group.*

Proof. Indeed, simply observe that the left-multiplication action of G on itself is manifestly faithful. □

The notion defined in Definition 9.2 is, for the sake of precision, called a *left*-action. A *right*-action would associate to each pair (g,a) with $g \in G$ and $a \in A$ an element $ag \in A$; our make-believe associativity would now say

$$a(gh) = (ag)h$$

for all $a \in A$ and $g,h \in G$. This is a different requirement than the one given in Definition 9.2; multiplication *on the right* in a group G gives a prototypical example of a right-action of G (on itself).

Every right-action may be turned into a left-action with due care (cf. Exercise 9.3). Therefore it is not restrictive to just consider left-actions; from now on, an 'action' will be understood to be a *left*-action, unless stated otherwise.

9.3. Transitive actions and the category G-Set.

Definition 9.6. An action of a group G on a set A is *transitive* if $\forall a,b \in A\ \exists g \in G$ such that $b = ga$. $\lrcorner$

For example, the left-multiplication action of a group on itself is transitive. Transitive actions are the basic ingredients making up *every* action; this is seen by means of the following important concepts.

Definition 9.7. The *orbit* of $a \in A$ under an action of a group G is the set

$$O_G(a) := \{ga \mid g \in G\}. \quad \lrcorner$$

[41] This is *left*-multiplication in the sense that the 'acting' element g of G is placed to the left of the element a 'acted upon'.

Definition 9.8. Let G act on a set A, and let $a \in A$. The *stabilizer* subgroup of a consists of the elements of G which fix a:

$$\mathrm{Stab}_G(a) := \{g \in G \mid ga = a\}. \qquad \lrcorner$$

Orbits of an action of a group G on a set A form a partition of A; and we have an induced, *transitive* action of G on each orbit. Therefore we can, in a sense, 'understand' all actions if we understand transitive actions. This will be accomplished in a moment, by studying actions related to stabilizers.

For any group G, sets endowed with a (left) G-action form in a natural way a category G-Set: objects are pairs (ρ, A), where $\rho : G \times A \to A$ is an action (as in Definition 9.2) and morphisms between two objects are set-functions which are compatible with the actions. That is, a morphism

$$(\rho, A) \to (\rho', A')$$

in G-Set amounts to a set-function $\varphi : A \to A'$ such that the diagram

$$\begin{array}{ccc} G \times A & \xrightarrow{\mathrm{id}_G \times \varphi} & G \times A' \\ {\scriptstyle\rho}\downarrow & & \downarrow{\scriptstyle\rho'} \\ A & \xrightarrow[\varphi]{} & A' \end{array}$$

commutes. In the usual shorthand notation omitting the ρ's, this means that $\forall g \in G$, $\forall a \in A$,

$$g\varphi(a) = \varphi(ga);$$

that is, the action 'commutes' with φ. Such functions are called (G-)*equivariant.*

We therefore have a notion of *isomorphism* of G-sets (defined as in §I.4.1); the reader should expect (and should verify) that these are nothing but the *equivariant bijections.*

Among G-sets we single out the sets G/H of left-cosets of subgroups H of G; as noted in Example 9.4, G acts on G/H by left-multiplication.

Proposition 9.9. *Every transitive left-action of G on a set A is isomorphic to the left-multiplication of G on G/H, for $H =$ the stabilizer of any $a \in A$.*

Proof. Let G act transitively on a set A, let $a \in A$ be any element, and let $H = \mathrm{Stab}_G(a)$. We claim that there is an equivariant bijection

$$\varphi : G/H \to A$$

defined by

$$\varphi(gH) := ga$$

for all $g \in G$.

Indeed, first of all φ is well-defined: if $g_1 H = g_2 H$, then $g_1^{-1} g_2 \in H$, hence $(g_1^{-1} g_2)a = a$, and it follows that $g_1 a = g_2 a$ as needed. To verify that φ is bijective, define a function $\psi : A \to G/H$ by sending an element ga of A to gH; ψ is well-defined because if $g_1 a = g_2 a$, then $g_1^{-1}(g_2 a) = a$, so $g_1^{-1} g_2 \in H$ and $g_1 H = g_2 H$. It is clear that φ and ψ are inverses of each other; hence φ is a bijection.

Equivariance is immediate: $\varphi(g'(gH)) = g'ga = g'\varphi(gH)$. $\square$

Corollary 9.10. *If O is an orbit of the action of a finite group G on a set A, then O is a finite set and*

$$|O| \quad \textit{divides} \quad |G|.$$

Proof. By Proposition 9.9 there is a bijection between O and $G/\operatorname{Stab}_G(a)$ for any element $a \in O$; thus

$$|O| \cdot |\operatorname{Stab}_G(a)| = |G|$$

by Corollary 8.14. □

Corollary 9.10 upgrades Lagrange's theorem to orbits of any action; it is extremely useful, as it provides a very strong constraint on group actions.

Example 9.11. There are *no* transitive actions of S_3 on a set with 5 elements.

Indeed, 5 does not divide 6. ⌟

Ultimately, almost everything we will prove in Chapter IV on the structure of finite groups will be a consequence of 'counting arguments' stemming from applications of Corollary 9.10 to actions of a group by conjugation or left-multiplication.

There may seem to be an element of arbitrariness in the statement of Proposition 9.9: what if we change the element a of which we are taking the stabilizer? The stabilizer may change, but it does so in a controlled way:

Proposition 9.12. *Suppose a group G acts on a set A, and let $a \in A$, $g \in G$, $b = ga$. Then*

$$\operatorname{Stab}_G(b) = g \operatorname{Stab}_G(a) g^{-1}.$$

Proof. Indeed, assume $h \in \operatorname{Stab}_G(a)$; then

$$(ghg^{-1})(b) = gh(g^{-1}g)a = gha = ga = b :$$

thus $ghg^{-1} \in \operatorname{Stab}_G(b)$. This proves the $\supseteq$ inclusion; $\subseteq$ follows by the same argument, noting that $a = g^{-1}b$. □

For example, if $\operatorname{Stab}_G(a)$ happens to be normal, then it is really independent of a (in any given orbit). In any case, there is an isomorphism of G-sets beween G/H and $G/(gHg^{-1})$, as follows from these considerations (and as the reader will independently check in Exercise 9.13).

Exercises

9.1. (Once more, if you are already familiar with a little linear algebra...) The matrix groups listed in Exercise 6.1 all come with evident actions on a vector space: if M is an $n \times n$ matrix with (say) real entries, multiplication to the right by a column n-vector $\mathbf{v}$ returns a column n-vector $M\mathbf{v}$, and this defines a left-action on $\mathbb{R}^n$ viewed as the space of column n-vectors.

- Prove that, through this action, matrices $M \in \mathrm{O}_n(\mathbb{R})$ preserve lengths and angles in $\mathbb{R}^n$.
- Find an interesting action of $\mathrm{SL}_2(\mathbb{C})$ on $\mathbb{R}^3$. (Hint: Exercise 8.9.)

9.2. The effect of the matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

on the plane is to respectively flip the plane about the y-axis and to rotate it $90°$ clockwise about the origin. With this in mind, construct an action of D_8 on $\mathbb{R}^2$.

9.3. If $G = (G, \cdot)$ is a group, we can define an 'opposite' group $G^\circ = (G, \bullet)$ supported on the same set G, by prescribing

$$(\forall g, h \in G): \quad g \bullet h := h \cdot g.$$

- Verify that G° is indeed a group.
- Show that the 'identity': $G^\circ \to G$, $g \mapsto g$ is an isomorphism if and only if G is commutative.
- Show that $G^\circ \cong G$ (even if G is not commutative!).
- Show that giving a *right*-action of G on a set A is the same as giving a homomorphism $G^\circ \to S_A$, that is, a *left*-action of G° on A.
- Show that the notions of left- and right-actions coincide 'on the nose' for *commutative* groups. (That is, if $(g, a) \mapsto ag$ defines a right-action of a commutative group G on a set A, then setting $ga = ag$ defines a left-action).
- For any group G, explain how to turn a right-action of G into a left-action of G. (Note that the simple 'flip' $ga = ag$ does *not* work in general if G is not commutative.)

9.4. As mentioned in the text, *right*-multiplication defines a right-action of a group on itself. Find *another* natural right-action of a group on itself.

9.5. Prove that the action by left-multiplication of a group on itself is free.

9.6. Let O be an orbit of an action of a group G on a set. Prove that the induced action of G on O is transitive.

9.7. Prove that stabilizers are indeed subgroups.

9.8. For G a group, verify that G-Set is indeed a category, and verify that the isomorphisms in G-Set are precisely the equivariant bijections.

9.9. Prove that G-Set has products and coproducts and that every finite object of G-Set is a coproduct of objects of the type $G/H = \{\text{left-cosets of } H\}$, where H is a subgroup of G and G acts on G/H by left-multiplication.

9.10. Let H be any subgroup of a group G. Prove that there is a bijection between the set G/H of *left*-cosets of H and the set $H\backslash G$ of *right*-cosets of H in G. (Hint: G acts on the right on the set of right-cosets; use Exercise 9.3 and Proposition 9.9.)

9.11. ¬ Let G be a finite group, and let H be a subgroup of index p, where p is the *smallest prime dividing* $|G|$. Prove that H is normal in G, as follows:

- Interpret the action of G on G/H by left-multiplication as a homomorphism $\sigma : G \to S_p$.
- Then $G/\ker\sigma$ is (isomorphic to) a subgroup of S_p. What does this say about the index of $\ker\sigma$ in G?
- Show that $\ker\sigma \subseteq H$.
- Conclude that $H = \ker\sigma$, by index considerations.

Thus H is a kernel, proving that it is normal. (This exercise generalizes the result of Exercise 8.2.) [9.12]

9.12. ¬ Generalize the result of Exercise 9.11, as follows. Let G be a group, and let $H \subseteq G$ be a subgroup of index n. Prove that H contains a subgroup K that is normal in G and such that $[G : K]$ divides the gcd of $|G|$ and $n!$. (In particular, $[G : K] \leq n!$.) [IV.2.23]

9.13. ▷ Prove 'by hand' that for all subgroups H of a group G and $\forall g \in G$, G/H and $G/(gHg^{-1})$ (endowed with the action of G by left-multiplication) are isomorphic in G-Set. [§9.3]

9.14. ¬ Prove that the modular group $\mathrm{PSL}_2(\mathbb{Z})$ is isomorphic to the coproduct $C_2 * C_3$. (Recall that the modular group $\mathrm{PSL}_2(\mathbb{Z})$ is generated by $x = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $y = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, satisfying the relations $x^2 = y^3 = e$ in $\mathrm{PSL}_2(\mathbb{Z})$ (Exercise 7.5). The task is to prove that x and y satisfy *no* other relation: this will show that $\mathrm{PSL}_2(\mathbb{Z})$ is presented by $(x, y \,|\, x^2, y^3)$, and we have agreed that this is a presentation for $C_2 * C_3$ (Exercise 3.8 or 8.7). Reduce this to verifying that no products

$$(y^{\pm 1}x)(y^{\pm 1}x)\cdots(y^{\pm 1}x) \quad \text{or} \quad (y^{\pm 1}x)(y^{\pm 1}x)\cdots(y^{\pm 1}x)y^{\pm 1}$$

with one or more factors can equal the identity. This latter verification is traditionally carried out by cleverly exploiting an action[42]. Let the modular group act on the set of *irrational* real numbers by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}(r) = \frac{ar+b}{cr+d}.$$

Check that this does define an action of $\mathrm{PSL}_2(\mathbb{Z})$, and note that

$$y(r) = 1 - \frac{1}{r}, \quad y^{-1}(r) = \frac{1}{1-r}, \quad yx(r) = 1 + r, \quad y^{-1}x(r) = \frac{r}{1+r}.$$

[42] The modular group acts on $\mathbb{C} \cup \{\infty\}$ by *Möbius transformations*. The observation that it suffices to act on $\mathbb{R} \smallsetminus \mathbb{Q}$ for the purpose of this verification is due to Roger Alperin.

Now complete the verification with a case-by-case analysis. For example, a product $(y^{\pm 1}x)(y^{\pm 1}x)\cdots(y^{\pm 1}x)y$ cannot equal the identity in $\mathrm{PSL}_2(\mathbb{Z})$ because if it did, it would act as the identity on $\mathbb{R}\smallsetminus\mathbb{Q}$, while if $r<0$, then $y(r)>0$, and both yx and $y^{-1}x$ send positive irrationals to positive irrationals.) [3.8]

9.15. ¬ Prove that every (finitely generated) group G acts freely on any corresponding Cayley graph. (Cf. Exercise 8.6. Actions on a directed graph are defined as actions on the set of vertices preserving incidence: if the vertices v_1, v_2 are connected by an edge, then so must be gv_1, gv_2 for every $g\in G$.) In particular, conclude that every free group acts freely on a tree. [9.16]

9.16. ▷ The converse of the last statement in Exercise 9.15 is also true: only free groups can act freely on a tree. Assuming this, prove that every subgroup of a free group (on a finite set) is free. [§6.4]

9.17. ▷ Consider G as a G-set, by acting with left-multiplication. Prove that $\mathrm{Aut}_{G-\mathsf{Set}}(G)\cong G$. [§2.1]

9.18. Show how to construct a *groupoid* carrying the information of the action of a group G on a set A. (Hint: A will be the set of objects of the groupoid. What will be the morphisms?)

---

10. Group objects in categories

10.1. Categorical viewpoint. The definition of *group* (Definition 1.2) is firmly grounded on the category Set: *A group is a* set G *endowed with a binary operation...*. However, we have noticed along the way (for example, in §3) that what is really behind it is a pair of functions:

$$m : G\times G\to G,\quad \iota : G\to G$$

satisfying certain properties (which translate into associativity, existence of inverses, etc.). Much of what we have seen could be expressed exclusively in terms of these functions, systematically replacing considerations on 'elements' by suitable commutative diagrams and enforcing universal properties as a means to define key notions such as the quotient of a group by a subgroup. For example, homomorphisms may be defined purely in terms of the commutativity of a diagram: cf. Definition 3.1.

This point of view may be transferred easily to categories other than Set, and the corresponding notions are very important in modern mathematics.

Definition 10.1. Let C be a category with (finite) products and with a final object 1.

A *group object* in C consists of an object G of C and of morphisms

$$m : G\times G\to G,\quad e : 1\to G,\quad \iota : G\to G$$

in C such that the diagrams

$$\begin{array}{ccc} (G\times G)\times G & \xrightarrow{m\times \mathrm{id}_G} & G\times G \\ \downarrow\wr & & \downarrow m \\ G\times(G\times G) & \xrightarrow{\mathrm{id}_G\times m} & G \end{array}$$

$$\begin{array}{ccc} 1\times G & \xrightarrow{e\times \mathrm{id}_G} & G\times G \\ & \searrow^{\wr} & \downarrow m \\ & & G \end{array} \qquad \begin{array}{ccc} G\times 1 & \xrightarrow{\mathrm{id}_G\times e} & G\times G \\ & \searrow^{\wr} & \downarrow m \\ & & G \end{array}$$

$$\begin{array}{ccccc} G & \xrightarrow{\Delta} & G\times G & \xrightarrow{\mathrm{id}_G\times \iota} & G\times G \\ \downarrow & & & & \downarrow m \\ 1 & & \xrightarrow{\quad e \quad} & & G \end{array} \qquad \begin{array}{ccccc} G & \xrightarrow{\Delta} & G\times G & \xrightarrow{\iota\times \mathrm{id}_G} & G\times G \\ \downarrow & & & & \downarrow m \\ 1 & & \xrightarrow{\quad e \quad} & & G \end{array}$$

commute. ⌟

Comments. The morphism $\Delta = \mathrm{id}_G \times \mathrm{id}_G$ is the 'diagonal' morphism $G \to G \times G$ induced by the universal property for products and the identity map(s) $G \to G$. Likewise, the other unnamed morphisms in these diagrams are all uniquely determined by suitable universal properties. For example, there is a unique morphism $\epsilon : G \to 1$ because 1 is final. The composition with the projection

$$G \xrightarrow{\epsilon\times \mathrm{id}_G} 1\times G \longrightarrow G$$

is the identity; so is

$$1\times G \longrightarrow G \xrightarrow{\epsilon\times \mathrm{id}_G} 1\times G$$

(why?); therefore the projection $1 \times G \to G$ is indeed an isomorphism, as indicated.

The reader will hopefully realize immediately (Exercise 10.2) that our original definition of groups given in §1 is precisely equivalent to the definition of group object *in* Set*:* the commutativity of the given diagrams codifies associativity and the existence of two-sided identity and inverses.

Most interesting categories the reader will encounter (not necessarily in this book), such as the category of topological spaces, differentiable manifolds, algebraic varieties, schemes, etc., will carry 'their own' notion of group object. For example, a *topological group* is a group object in the category of topological spaces; a *Lie group* is a group object in the category of differentiable manifolds, etc.

Exercises

10.1. Define all the unnamed maps appearing in the diagrams in the definition of group object, and prove they are indeed isomorphisms when so indicated. (For the projection $1 \times G \to G$, what is left to prove is that the composition

$$1 \times G \to G \to 1 \times G$$

is the identity, as mentioned in the text.)

10.2. ▷ Show that *groups*, as defined in §1.2, are 'group objects in the category of sets'. [§10.1]

10.3. Let $(G, \cdot)$ be a group, and suppose $\circ : G \times G \to G$ is a group homomorphism (w.r.t. $\cdot$) such that $(G, \circ)$ is *also* a group. Prove that $\circ$ and $\cdot$ coincide. (Hint: First prove that the identity with respect to the two operations must be the same.)

10.4. Prove that every *abelian* group has exactly one structure of group object *in the category* Ab.

10.5. By the previous exercise, a group object in Ab is nothing other than an abelian group. What is a group object in Grp?

Rings and modules

1. Definition of ring

In this chapter we will do for *rings* and *modules* what we have done in Chapter II for groups: describe them in general terms, with particular attention to distinguished subobjects and quotients. More detailed information on these structures will be deferred to later chapters: in Chapter V we will look more carefully at several interesting classes of rings, and modules (over commutative rings) will take center stage in our rapid overview of linear algebra in Chapter VI and following. In this chapter we will also include a brief jaunt into homological algebra, a topic that will entertain us greatly in Chapter IX.

1.1. Definition. Rings (and modules) are defined by 'decorating' abelian groups with additional data. As motivation for the introduction of such structures, note that all number-based examples of groups that we have encountered, such as $\mathbb{Z}$ or $\mathbb{R}$, are endowed with an operation of *multiplication* as well as the 'addition' making them into (abelian) groups. The 'ring axioms' will reflect closely the properties and compatibilities of these two operations in such examples.

These examples are, however, very special. A more sophisticated motivation for the introduction of rings arises by analyzing further the structure of homomorphisms of abelian groups. Recall (§II.4.4) that if G, H are abelian groups, then $\mathrm{Hom}_{\mathsf{Ab}}(G,H)$ is also an abelian group. In particular, if G is an abelian group, then so is the set of *endomorphisms* $\mathrm{End}_{\mathsf{Ab}}(G) = \mathrm{Hom}_{\mathsf{Ab}}(G,G)$. More is true: morphisms from an object of a category to itself may be composed with each other (by definition of category!). Thus, *two* operations coexist in $\mathrm{End}_{\mathsf{Ab}}(G)$: addition (inherited from G, making $\mathrm{End}_{\mathsf{Ab}}(G)$ an abelian group), and composition. These two operations are compatible with each other in a sense captured by the *ring axioms:*

Definition 1.1. A *ring* $(R,+,\cdot)$ is an abelian group $(R,+)$ endowed with a *second* binary operation $\cdot$, satisfying on its own the requirements of being associative and having a two-sided identity, i.e.,

- $(\forall r, s, t \in R)$: $\quad (r \cdot s) \cdot t = r \cdot (s \cdot t)$,
- $(\exists 1_R \in R)\ (\forall r \in R)$: $\quad r \cdot 1_R = r = 1_R \cdot r$

(which make $(R, \cdot)$ a *monoid*), and further interacting with $+$ via the following *distributive* properties:

- $(\forall r, s, t \in R)$: $\quad (r + s) \cdot t = r \cdot s + r \cdot t$ and $t \cdot (r + s) = t \cdot r + t \cdot s$. ⌟

The notation $\cdot$ is often omitted in formulas, and we usually refer to rings by the name of the underlying set.

Warning: What we are calling a 'ring', others may call a 'ring with identity' or a 'ring with 1': it is not uncommon to exclude the axiom of existence of a 'multiplicative identity' from the list of axioms defining a ring. Rings without identity are sometimes called *rngs*, but we are not sure this should be encouraged[1]. The reader should check conventions carefully when approaching the literature. Examples of structures *without* a multiplicative identity abound: for example, the set $2\mathbb{Z}$ of *even* integers, with the usual addition and multiplication, satisfies all the ring axioms given above with the exception of the existence of 1 (and is therefore a rng). But in these notes all rings will have 1.

Of course the multiplicative identity is necessarily unique: the argument given for Proposition II.1.6 works *verbatim*.

The identity element of the abelian group underlying a ring is denoted 0_R (or simply 0, in context) and is called the 'additive' identity. This is a special element with respect to multiplication:

Lemma 1.2. *In a ring R,*

$$0 \cdot r = 0 = r \cdot 0$$

for all $r \in R$.

Proof. Indeed, $0 = 0 + 0$; hence, applying distributivity,

$$r \cdot 0 = r \cdot (0 + 0) = r \cdot 0 + r \cdot 0,$$

from which $r \cdot 0 = 0$ by cancellation (in the *group* $(R, +)$). The equality $0 \cdot r = 0$ is proven similarly. □

It is equally easy to check that multiplication behaves as expected on 'subtraction'. In fact, if -1 denotes the additive inverse of 1, then the additive inverse $-r$ of any $r \in R$ is the result of the multiplication $(-1) \cdot r$: indeed, using distributivity,

$$r + (-1) \cdot r = 1 \cdot r + (-1) \cdot r = (1 - 1) \cdot r = 0 \cdot r = 0$$

(by Lemma 1.2) from which $(-1)r = -r$ follows by (additive) cancellation.

[1]The term 'rng' was introduced with this meaning by Jacobson; but essentially at the same time Mac Lane introduced Rng as the name for the category of rings *with* identity. Hoping to steer clear of this clash of terminology we have opted to call this category 'Ring'.

1.2. First examples and special classes of rings.

Example 1.3. We can define a ring structure on a trivial group $\{*\}$ by letting $* \cdot * = *$ (as well as $* + * = *$); this is often called the *zero-ring*. Note that $0 = 1$ in this ring (cf. Exercise 1.1). ⌟

Example 1.4. More interesting examples are the number-based groups such as $\mathbb{Z}$ or $\mathbb{R}$, with the usual operations. These are very well known to our reader, who will realize immediately that they satisfy the requirements given in Definition 1.1; but they are very special. Why? ⌟

To begin with, note that multiplication is commutative in these examples; this is not among the requirements we have posed on rings in the official definition given above.

Definition 1.5. A ring R is *commutative* if

- $(\forall r, s \in R): \quad r \cdot s = s \cdot r.$ ⌟

Commutative rings (with identity) form an extremely important class of rings; *commutative algebra* is the subfield of algebra studying them. We will focus on commutative rings in later chapters; in this chapter we will develop some of the basic theory for the more general case of arbitrary rings (with 1).

Example 1.6. An example of a noncommutative ring that is (likely) familiar to the readers is the ring of 2×2 matrices with, say, real entries: matrices can be added 'componentwise', and they can be multiplied as recalled in Example II.1.5; the two operations satisfy the requirements in Definition 1.1.

Square matrices of any size, and with entries in any ring, form a ring (Exercise 1.4). ⌟

Example 1.7. The reader is already familiar with a large class of (commutative) rings: the groups $\mathbb{Z}/n\mathbb{Z}$, endowed with the multiplication defined in §II.2.3 (that is: $[a]_n \cdot [b]_n := [ab]_n$; this is well-defined (cf. Exercise II.2.14)) satisfy the ring axioms listed above. ⌟

The rings $\mathbb{Z}/n\mathbb{Z}$ prompt us to highlight an important point. Another reason why rings such as $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, ... are special is that *multiplicative* cancellation by nonzero elements holds in these rings. Of course *additive* cancellation is automatic, since rings are in particular (abelian) *groups;* and multiplicative cancellation clearly fails in general since one cannot 'cancel 0' (by Lemma 1.2). But even the fact that

$$(\forall a \in R), a \neq 0: \quad a \cdot b = a \cdot c \implies b = c,$$

which holds, for example, in $\mathbb{Z}$, does *not* follow from the ring axioms.

Indeed, this cancellation property does not hold in *all* rings; it may well fail in the rings $\mathbb{Z}/n\mathbb{Z}$. For example,

$$[2]_6 \cdot [4]_6 = [8]_6 = [2]_6 = [2]_6 \cdot [1]_6$$

even though $[4]_6 \neq [1]_6$.

The problem here is that in $\mathbb{Z}/6\mathbb{Z}$ there are elements $a \neq 0$ such that $a \cdot b = 0$ for some $b \neq 0$ (take $a = [2]_6$, $b = [3]_6$).

Definition 1.8. An element a in a ring R is a *(left-)zero-divisor* if there exist elements $b \neq 0$ in R for which $ab = 0$. ⌟

The reader will have no difficulty figuring out what a *right*-zero-divisor should be. The element 0 is a zero-divisor in all nonzero rings R; the zero ring is the only ring without zero-divisors(!).

Proposition 1.9. *In a ring R,*

- *$a \in R$ is* not *a left- (resp., right-) zero-divisor if and only if left (resp., right) multiplication by a is an* injective *function $R \to R$.*

In other words, a is not a left- (resp., right-) zero-divisor if and only if multiplicative left- (resp., right-) cancellation by the element a holds in R.

Proof. Let's verify the 'left' statement (the 'right' statement is of course entirely analogous). Assume a is not a left-zero-divisor and $ab = ac$ for $b, c \in R$. Then, by distributivity,

$$a(b - c) = ab - ac = 0,$$

and this implies $b - c = 0$ since a is not a left-zero-divisor; that is, $b = c$. This proves that left-multiplication is injective in this case.

Conversely, if a is a left-zero-divisor, then $\exists b \neq 0$ such that $ab = 0 = a \cdot 0$; this shows that left-multiplication is not injective in this case, concluding the proof. □

Rings such as $\mathbb{Z}$, $\mathbb{Q}$, etc., are commutative rings *without* (nonzero) zero-divisors. Such rings are very special, but very important, and they deserve their own terminology:

Definition 1.10. An *integral domain* is a nonzero commutative ring R (with 1) such that

$$(\forall a, b \in R): \quad ab = 0 \implies a = 0 \text{ or } b = 0.$$ ⌟

Chapter V will be entirely devoted to integral domains.

An element which is *not* a zero-divisor is called a *non-zero-divisor*. Thus, integral domains are those nonzero commutative rings in which every nonzero element is a non-zero-divisor. By Proposition 1.9, multiplicative cancellation by nonzero elements holds in integral domains. The rings $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are all integral domains. As we have seen, some $\mathbb{Z}/n\mathbb{Z}$ are *not* integral domains.

Here is one of those places where the reader can do him/herself a great favor by pausing a moment and figuring something out: answer the question, *which $\mathbb{Z}/n\mathbb{Z}$ are integral domains?* This is entirely within reach, given what the reader knows already. Don't read ahead before figuring this out—this question will be answered within a few short paragraphs, spoiling all the fun.

There are even subtler reasons why $\mathbb{Z}$ is a very special ring: we will see in due time that it is a 'UFD' *(unique factorization domain);* in fact, it is a 'PID' *(principal ideal domain);* in fact, it is more special still!, as it is a 'Euclidean domain'. All of this will be discussed in Chapter V, particularly §V.2.

However, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are more special than all of that and then some, since they are *fields*.

Definition 1.11. An element u of a ring R is a *left-unit* if $\exists v \in R$ such that $uv = 1$; it is a *right-unit* if $\exists v \in R$ such that $vu = 1$. *Units* are two-sided units. ⌟

Proposition 1.12. *In a ring R:*

- *u is a left- (resp., right-) unit if and only if left- (resp., right-) multiplication by u is a* surjective *functions $R \to R$;*
- *if u is a left- (resp., right-) unit, then right- (resp., left-) multiplication by u is injective; that is, u is not a right- (resp., left-) zero-divisor;*
- *the inverse of a two-sided unit is unique;*
- *two-sided units form a group under multiplication.*

Proof. These assertions are all straightforward. For example, denote by $\rho_u : R \to R$ *right*-multiplication by u, so that $\rho_u(r) = ru$. If u is a right-unit, let $v \in R$ be such that $vu = 1$; then $\forall r \in R$

$$\rho_u \circ \rho_v(r) = \rho_u(rv) = (rv)u = r(vu) = r1_R = r.$$

That is, ρ_v is a right-inverse to ρ_u, and therefore ρ_u is surjective (Proposition I.2.1).

Conversely, if ρ_u is surjective, then there exists a v such that $1_R = \rho_{(u)}(v) = vu$, so that u is a right-unit.

This checks the first statement, for *right*-units.

For the second statement, denote by $\lambda_u : R \to R$ *left*-multiplication by u: $\lambda_u(r) = ur$. Assume u is a *right*-unit, and let v be such that $vu = 1_R$; then $\forall r \in R$

$$\lambda_v \circ \lambda_u(r) = \lambda_v(ur) = v(ur) = (vu)r = 1_R r = r.$$

That is, λ_v is a left-inverse to λ_u, so λ_u is injective (Proposition I.2.1 again).

The rest of the proof is left to the reader (Exercise 1.9). □

Since the inverse of a two-sided unit u is unique, we can give it a name; of course we denote it by u^{-1}. The reader should keep in mind that inverses of left- or right-units are *not* unique in general, so the 'inverse notation' is not appropriate for them.

Definition 1.13. A *division ring* is a ring in which every nonzero element is a two-sided unit. ⌟

We will mostly be concerned with the commutative case, which has its own name:

Definition 1.14. A *field* is a nonzero commutative ring R (with 1) in which every nonzero element is a unit. ⌟

The whole of Chapter VII will be devoted to studying fields.

By Proposition 1.12 (second part), every field is an integral domain, but not conversely: indeed, $\mathbb{Z}$ is an integral domain, *but* it is not a field. Remember:

$$\text{field} \Rightarrow \text{integral domain},$$

$$\text{integral domain} \not\Rightarrow \text{field}.$$

There is a situation, however, in which the two notions coincide:

Proposition 1.15. *Assume R is a* finite *commutative ring; then R is an integral domain if* and only if *it is a field.*

Proof. One implication holds for all rings, as pointed out above; thus we only have to verify that if R is a finite integral domain, then it is a field. This amounts to verifying that if a is a non-zero-divisor in a *finite* (commutative) ring R, then it is a unit in R.

Now, if a is a non-zero-divisor, then multiplication by a in R is injective (Proposition 1.9); hence it is surjective, as the ring is finite, by the pigeon-hole principle; hence a is a unit, by Proposition 1.12. □

Remark 1.16. A little surprisingly, the hypothesis of commutativity in Proposition 1.15 is actually superfluous: a theorem known as *Wedderburn's little theorem* shows that *finite division rings are necessarily commutative.* The reader will prove this fact in a distant future (Exercise VII.5.14). ⌟

Example 1.17. The *group of units* in the ring $\mathbb{Z}/n\mathbb{Z}$ is precisely the group $(\mathbb{Z}/n\mathbb{Z})^*$ introduced in §II.2.3: indeed, a class $[m]_n$ is a unit if and only if (right-) multiplication by $[m]_n$ is surjective (by Proposition 1.12), if and only if the map $a \mapsto a[m]_n$ is surjective, if and only if $[m]_n$ generates $\mathbb{Z}/n\mathbb{Z}$, if and only if $\gcd(m,n) = 1$ (Corollary II.2.5), if and only if $[m]_n \in (\mathbb{Z}/n\mathbb{Z})^*$.

In particular, those n for which all nonzero elements of $\mathbb{Z}/n\mathbb{Z}$ are units (that is, for which $\mathbb{Z}/n\mathbb{Z}$ is a *field*) are precisely those $n \in \mathbb{Z}$ for which $\gcd(m,n) = 1$ for all m that are *not* multiples of n; this is the case if and only if n is prime. Putting this together with Proposition 1.15, we get the pretty classification (for integers $p \neq 0$)

$$\mathbb{Z}/p\mathbb{Z} \text{ integral domain} \iff \mathbb{Z}/p\mathbb{Z} \text{ field} \iff p \text{ prime},$$

which the reader is well advised to remember firmly. ⌟

Example 1.18. The rings $\mathbb{Z}/p\mathbb{Z}$, with p prime, are *not* the only finite fields. In fact, for every prime p and every integer $r > 0$ there is a (unique, in a suitable sense) multiplication on the product group

$$\underbrace{\mathbb{Z}/p\mathbb{Z} \times \cdots \times \mathbb{Z}/p\mathbb{Z}}_{r \text{ times}}$$

making it into a field. A discussion of these fields will have to wait until we have accumulated much more material (cf. §VII.5.1), but the reader could already try to construct small examples 'by hand' (cf. Exercise 1.11). ⌟

1.3. Polynomial rings. We will study polynomial rings in some depth, especially over fields; they are another class of examples that is to some extent already familiar to our reader. I will capitalize on this familiarity and avoid a truly formal (and truly tedious) definition.

Definition 1.19. Let R be a ring. A *polynomial* $f(x)$ in the *indeterminate* x and with *coefficients* in R is a *finite* linear combination of nonnegative 'powers' of x with coefficients in R:

$$f(x) = \sum_{i\geq 0} a_i x^i = a_0 + a_1 x + a_2 x^2 + \cdots,$$

where all a_i are elements of R (the *coefficients*) and we require $a_i = 0$ for $i \gg 0$. Two polynomials are taken to be equal if all the coefficients are equal:

$$\sum_{i\geq 0} a_i x^i = \sum_{i\geq 0} b_i x^i \iff (\forall i \geq 0): \quad a_i = b_i. \qquad \lrcorner$$

The set of polynomials in x over R is denoted by $R[x]$. Since all but finitely many a_i are assumed to be 0, one usually employs the notation

$$f(x) = a_0 + a_1 x + \cdots + a_n x^n$$

for $\sum_{i\geq 0} a_i x^i$, if $a_i = 0$ for $i > n$.

At this point the reader should just view all of this as a *notation:* a polynomial really stands for an element of an infinite direct sum of the group $(R, +)$. The 'polynomial' notation is more suggestive as it hints at what *operations* we are going to impose on $R[x]$: if

$$f(x) = \sum_{i\geq 0} a_i x^i \quad \text{and} \quad g(x) = \sum_{i\geq 0} b_i x^i,$$

then we define

$$f(x) + g(x) := \sum_{i\geq 0} (a_i + b_i) x^i$$

and

$$f(x) \cdot g(x) := \sum_{k\geq 0} \sum_{i+j=k} a_i b_j x^{i+j}.$$

To clarify this latter definition, see how it works for small k: $f(x) \cdot g(x)$ equals

$$a_0 b_0 + (a_0 b_1 + a_1 b_0)x + (a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + (a_0 b_3 + a_1 b_2 + a_2 b_1 + a_3 b_0)x^3 + \cdots,$$

that is, business as usual.

It is essentially straightforward (Exercise 1.13) to check that $R[x]$, with these operations, is a ring; the identity 1 of R is the identity of $R[x]$, when viewed as a polynomial (that is, $1_{R[x]} = 1_R + 0x + 0x^2 + \cdots$).

The *degree* of a nonzero polynomial $f(x) = \sum_{i\geq 0} a_i x^i$, denoted $\deg f(x)$, is the largest integer d for which $a_d \neq 0$. This notion is very useful, but really behaves well (Exercise 1.14) only if R is an integral domain: for example, note that over $R = \mathbb{Z}/6\mathbb{Z}$

$$\deg([1] + [2]x) = 1, \quad \deg([1] + [3]x) = 1, \quad \text{but}$$
$$\deg(([1] + [2]x) \cdot (1 + [3]x)) = \deg([1] + [5]x) = 1 \neq 1 + 1.$$

Polynomials of degree 0 (together with 0) are called *constants*; they form a 'copy' of R in $R[x]$, since the operations $+, \cdot$ on constant polynomials are nothing but the original operations in R, up to this identification. It is sometimes convenient to assign to the polynomial 0 the degree $-\infty$.

Polynomial rings in more indeterminates may be obtained by iterating this construction:

$$R[x, y, z] := R[x][y][z];$$

elements of this ring may be written as 'ordinary' polynomials in three indeterminates and are manipulated as usual. It can be checked easily that the construction does not really depend on the order in which the indeterminates are listed, in the sense that different orderings lead to *isomorphic* rings (in the sense soon to be defined officially). Different indeterminates commute with each other; constructions analogous to polynomial rings, but with noncommuting indeterminates, are also very important, but we will not develop them in this book (we will glance at one such notion in Example VIII.4.17).

We will occasionally consider a polynomial ring in *infinitely many* indeterminates: for example, we will denote by $R[x_1, x_2, \dots]$ the case of countably many indeterminates. Keep in mind, however, that polynomials are *finite* linear combinations of *finite* products of the indeterminates; in particular, every given element of $R[x_1, x_2, \dots]$ only involves *finitely many* indeterminates. An honest definition of this ring involves *direct limits,* which await us in §VIII.1.4.

Rings of *power series* may be defined and are very useful; the ring of series $\sum_{i=0}^{\infty} a_i x^i = a_0 + a_1 x + a_2 x^2 + \dots$ in x with coefficients in R, and evident operations, is denoted $R[[x]]$. Regrettably, we will only occasionally encounter these rings in this book.

The ring $R[x]$ is (clearly) commutative if R is commutative; it is an integral domain if R is an integral domain (Exercise 1.15); but it has no chances of being a field even if R is a field, since x has no inverse in $R[x]$. The question of which properties of R are 'inherited' by $R[x]$ is subtle and important, and we will give it a great deal of attention in later sections.

1.4. Monoid rings. The polynomial ring is an instance of a rather general construction, which is occasionally very useful. A *semigroup* is a set endowed with an associative operation; a *monoid* is a semigroup with an identity element. Thus a group is a monoid in which every element has an inverse; positive integers with ordinary addition form a semigroup, while the set $\mathbb{N}$ of *natural numbers* (that is, nonnegative integers[2]) is a monoid under addition.

Given a monoid $(M, \cdot)$ and a ring R, we can obtain a new ring $R[M]$ as follows. Elements of $R[M]$ are formal linear combinations

$$\sum_{m \in M} a_m \cdot m$$

where the 'coefficients' r_m are elements of R and $a_m \neq 0$ for at most finitely many summands (hence, as in §1.3, as an abelian group $R[M]$ is nothing but the direct sum $R^{\oplus M}$). Operations in $R[M]$ are defined by

$$\left(\sum_{m \in M} a_m \cdot m\right) + \left(\sum_{m \in M} b_m \cdot m\right) = \sum_{m \in M} (a_m + b_m) \cdot m,$$

[2]Some disagree, and insist that $\mathbb{N}$ should not include 0.

$$(\sum_{m\in M} a_m \cdot m)\cdot(\sum_{m\in M} b_m \cdot m) = \sum_{m\in M}\sum_{m_1m_2=m} (a_{m_1}b_{m_2})\cdot m.$$

The identity in $R[M]$ is $1_R \cdot 1_M$, viewed as a formal sum in which all other summands have 0 as coefficient.

The reader will hopefully see the similarity with the construction of the polynomial ring $R[x]$ in §1.3; in fact (Exercise 1.17) the polynomial ring $R[x]$ may be interpreted as $R[\mathbb{N}]$.

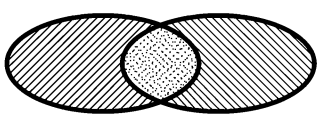
Group rings are the result of this construction when M is in fact a group. The group ring $R[\mathbb{Z}]$ is a ring of 'Laurent polynomials' $R[x, x^{-1}]$, allowing for negative as well as positive exponents.

Exercises

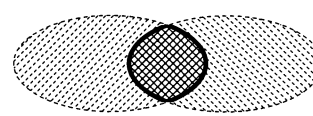
1.1. ▷ Prove that if $0 = 1$ in a ring R, then R is a zero-ring. [§1.2]

1.2. ¬ Let S be a set, and define operations on the power set $\mathscr{P}(S)$ of S by setting $\forall A, B \in \mathscr{P}(S)$

$$A + B := (A \cup B) \smallsetminus (A \cap B), \quad A \cdot B = A \cap B :$$



$A + B$



$A \cdot B$

(where the solid black contour indicates the set included in the operation). Prove that $(\mathscr{P}(S), +, \cdot)$ is a commutative ring. [2.3, 3.15]

1.3. ¬ Let R be a ring, and let S be any set. Explain how to endow the set R^S of set-functions $S \to R$ of two operations $+$, $\cdot$ so as to make R^S into a ring, such that R^S is just a copy of R if S is a sigleton. [2.3]

1.4. ▷ The set of $n \times n$ matrices with entries in a ring R is denoted $\mathcal{M}_n(R)$. Prove that componentwise addition and matrix multiplication make $\mathcal{M}_n(R)$ into a ring, for any ring R. The notation $\mathfrak{gl}_n(R)$ is also commonly used, especially for $R = \mathbb{R}$ or $\mathbb{C}$ (although this indicates one is considering them as *Lie algebras*) in parallel with the analogous notation for the corresponding groups of units; cf. Exercise II.6.1. In fact, the parallel continues with the definition of the following sets of matrices:

- $\mathfrak{sl}_n(\mathbb{R}) = \{M \in \mathfrak{gl}_n(\mathbb{R}) \mid \operatorname{tr}(M) = 0\}$;
- $\mathfrak{sl}_n(\mathbb{C}) = \{M \in \mathfrak{gl}_n(\mathbb{C}) \mid \operatorname{tr}(M) = 0\}$;
- $\mathfrak{so}_n(\mathbb{R}) = \{M \in \mathfrak{sl}_n(\mathbb{R}) \mid M + M^t = 0\}$;
- $\mathfrak{su}_n(\mathbb{C}) = \{M \in \mathfrak{sl}_n(\mathbb{C}) \mid M + M^\dagger = 0\}$.

Here $\operatorname{tr} M$ is the *trace* of M, that is, the sum of its diagonal entries. The other notation matches the notation used in Exercise II.6.1. Can we make rings of these sets

by endowing them with ordinary addition and multiplication of matrices? (These sets are all Lie algebras; cf. Exercise VI.1.4.) [§1.2, 2.4, 5.9, VI.1.2, VI.1.4]

1.5. Let R be a ring. If a, b are zero-divisors in R, is $a+b$ necessarily a zero-divisor?

1.6. ¬ An element a of a ring R is *nilpotent* if $a^n = 0$ for some n.

- Prove that if a and b are nilpotent in R and $ab = ba$, then $a+b$ is also nilpotent.
- Is the hypothesis $ab = ba$ in the previous statement necessary for its conclusion to hold?

[3.12]

1.7. Prove that $[m]$ is nilpotent in $\mathbb{Z}/n\mathbb{Z}$ if and only if m is divisible by all prime factors of n.

1.8. Prove that $x = \pm 1$ are the only solutions to the equation $x^2 = 1$ in an integral domain. Find a ring in which the equation $x^2 = 1$ has more than 2 solutions.

1.9. ▷ Prove Proposition 1.12. [§1.2]

1.10. Let R be a ring. Prove that if $a \in R$ is a right-unit and has two or more left-inverses, then a is *not* a left-zero-divisor and *is* a right-zero-divisor.

1.11. ▷ Construct a field with 4 elements: as mentioned in the text, the underlying abelian group will have to be $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$; $(0,0)$ will be the zero element, and $(1,1)$ will be the multiplicative identity. The question is what $(0,1)\cdot(0,1)$, $(0,1)\cdot(1,0)$, $(1,0)\cdot(1,0)$ must be, in order to get a *field*. [§1.2, §V.5.1]

1.12. ▷ Just as complex numbers may be viewed as combinations $a + bi$, where $a, b \in \mathbb{R}$ and i satisfies the relation $i^2 = -1$ (and commutes with $\mathbb{R}$), we may construct a ring[3] $\mathbb{H}$ by considering linear combinations $a + bi + cj + dk$ where $a, b, c, d \in \mathbb{R}$ and i, j, k commute with $\mathbb{R}$ and satisfy the following relations:

$$i^2 = j^2 = k^2 = -1, \quad ij = -ji = k, \quad jk = -kj = i, \quad ki = -ik = j.$$

Addition in $\mathbb{H}$ is defined componentwise, while multiplication is defined by imposing distributivity and applying the relations. For example,

$$(1+i+j)\cdot(2+k) = 1\cdot 2 + i\cdot 2 + j\cdot 2 + 1\cdot k + i\cdot k + j\cdot k = 2+2i+2j+k-j+i = 2+3i+j+k.$$

(i) Verify that this prescription does indeed define a ring.

(ii) Compute $(a + bi + cj + dk)(a - bi - cj - dk)$, where $a, b, c, d \in \mathbb{R}$.

(iii) Prove that $\mathbb{H}$ is a division ring.

Elements of $\mathbb{H}$ are called *quaternions*. Note that $Q_8 := \{\pm 1, \pm i, \pm j, \pm k\}$ forms a subgroup of the group of units of $\mathbb{H}$; it is a noncommutative group of order 8, called the *quaternionic group*.

(iv) List all subgroups of Q_8, and prove that they are all normal.

(v) Prove that Q_8, D_8 are not isomorphic.

(vi) Prove that Q_8 admits the presentation $(x, y \mid x^2y^{-2}, y^4, xyx^{-1}y)$.

[§II.7.1, 2.4, IV.1.12, IV.5.16, IV.5.17, V.6.19]

[3]The letter $\mathbb{H}$ is chosen in honor of William Rowan Hamilton.

1.13. ▷ Verify that the multiplication defined in $R[x]$ is associative. [§1.3]

1.14. ▷ Let R be a ring, and let $f(x), g(x) \in R[x]$ be nonzero polynomials. Prove that

$$\deg(f(x) + g(x)) \leq \max(\deg(f(x)), \deg(g(x))).$$

Assuming that R is an integral domain, prove that

$$\deg(f(x) \cdot g(x)) = \deg(f(x)) + \deg(g(x)).$$

[§1.3]

1.15. ▷ Prove that $R[x]$ is an integral domain if and only if R is an integral domain. [§1.3]

1.16. Let R be a ring, and consider the ring of power series $R[[x]]$ (cf. §1.3).

(i) Prove that a power series $a_0 + a_1x + a_2x^2 + \cdots$ is a unit in $R[[x]]$ if and only if a_0 is a unit in R. What is the inverse of $1 - x$ in $R[[x]]$?

(ii) Prove that $R[[x]]$ is an integral domain if and only if R is.

1.17. ▷ Explain in what sense $R[x]$ agrees with the monoid ring $R[\mathbb{N}]$. [§1.4]

---

2. The category Ring

2.1. Ring homomorphisms. *Ring homomorphisms* are defined in the natural way: if R, S are rings, a function $\varphi : R \to S$ is a ring homomorphism if it preserves *both* operations and the identity element. That is, φ must be a homomorphism of the underlying abelian groups,

$$(\forall a, b \in R): \quad \varphi(a + b) = \varphi(a) + \varphi(b),$$

it must preserve the operation of multiplication,

$$(\forall a, b \in R): \quad \varphi(ab) = \varphi(a)\varphi(b),$$

and finally

$$\varphi(1_R) = 1_S.$$

It is evident that rings form a category, with ring homomorphisms as morphisms. We will denote this category by Ring.

The zero-ring is clearly final in Ring. However, note that it is *not* initial: because of the requirement that ring homomorphisms send 1 to 1, the only rings to which zero-rings map homomorphically are the zero-rings (Exercise 2.1).

The category Ring *does* have initial objects: the ring of integers $\mathbb{Z}$ (with the usual operations $+$, $\cdot$) is initial in Ring. Indeed, for every ring R we can define a *group* homomorphism $\varphi : \mathbb{Z} \to R$ by

$$(\forall n \in \mathbb{Z}): \quad \varphi(n) = n \cdot 1_R,$$

that is, as the 'exponential map' ϵ_{1_R} corresponding to $1_R \in R$; cf. §II.4.1. But φ is in fact a *ring* homomorphism, since $\varphi(1) = 1_R$, and

$$\varphi(mn) = (mn)1_R = m(n1_R) \overset{!}{=} (m1_R) \cdot (n1_R) = \varphi(m) \cdot \varphi(n),$$

where the equality $\stackrel{!}{=}$ holds by the distributivity axiom[4]. This ring homomorphism is unique, since it is determined by the requirement that $\varphi(1) = 1_R$ and by the fact that φ preserves addition. Thus for every ring R there exists one and only one ring homomorphism $\mathbb{Z} \to R$, showing that $\mathbb{Z}$ is initial in Ring.

This should already convince the reader that $\mathbb{Z}$ is a *very special* ring. There is in fact nothing arbitrary about $\mathbb{Z}$: knowledge of the *group* $\mathbb{Z}$ determines the *ring* structure of $\mathbb{Z}$, as will be underscored below. This is one of the main reasons why we included the 'identity' axiom in the list in Definition 1.1: there are many nonisomorphic structures of 'ring without identity' on the group $(\mathbb{Z}, +)$ (cf. Exercise 2.15), but only one structure of ring *with* identity (Exercise 2.16).

Ring homomorphisms preserve units: that is, if u is a (left-, resp., right-) unit in R and $\varphi : R \to S$ is a ring homomorphism, then $\varphi(u)$ is a (left-, resp., right-) unit. Indeed, if v is a (right-, say) inverse of u, then

$$\varphi(u)\varphi(v) = \varphi(uv) = \varphi(1_R) = 1_S,$$

so that $\varphi(v)$ is a (right-) inverse of $\varphi(u)$.

On the other hand, the image of a non-zero-divisor by a ring homomorphism may well be a zero-divisor: for example, the canonical projection $\pi : \mathbb{Z} \to \mathbb{Z}/6\mathbb{Z}$ is a ring homomorphism, and 2 is a *non*-zero-divisor in $\mathbb{Z}$, yet $\pi(2) = [2]_6$ *is* a zero-divisor.

2.2. Universal property of polynomial rings. Polynomial rings satisfy a universal property not unlike the one for free groups explored in §II.5.2. The simplest (and possibly most useful) case is for the polynomial rings $\mathbb{Z}[x_1, \cdots, x_n]$, and with respect to *commutative* rings; we will leave the reader the pleasure of stating and proving fancier notions.

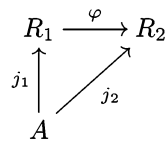
Let $A = \{a_1, \ldots, a_n\}$ be a set of order n. Consider the category $\mathscr{R}_A$ whose objects are pairs (j, R), where R is a commutative ring[5] and

$$j : A \to R$$

is a set-function (cf. §II.5.2!); morphisms

$$(j_1, R_1) \to (j_2, R_2)$$

are commutative diagrams



in which φ is a *ring* homomorphism.

For example, $(i, \mathbb{Z}[x_1, \cdots, x_n])$ is an object of $\mathscr{R}_A$, where $i : A \to \mathbb{Z}[x_1, \cdots, x_n]$ sends a_k to x_k.

[4]The reader should parse this display carefully, as there is a potentially confusing mix of two operations: multiples (such as $m1_R$) and multiplication in R (explicitly denoted here by $\cdot$).

[5]For the reader interested in generalizations: only the requirement that $j(a_1), \ldots, j(a_n)$ commute with every element of R is needed here.

Proposition 2.1. *$(i, \mathbb{Z}[x_1, \cdots, x_n])$ is initial in $\mathscr{R}_A$.*

Proof. Let (j, R) be an arbitrary object of $\mathscr{R}_A$; we have to show that there is a unique morphism $(i, \mathbb{Z}[x_1, \cdots, x_n]) \to (j, R)$, that is, there exists exactly one ring homomorphism $\varphi : \mathbb{Z}[x_1, \cdots, x_n] \to R$ such that

$$\begin{array}{ccc} \mathbb{Z}[x_1, \cdots, x_n] & \xrightarrow{\varphi} & R \\ {\scriptstyle i}\uparrow & \nearrow_{j} & \\ A & & \end{array}$$

commutes.

As usual in these verifications, the key point is that the requirements posed on φ force its definition. The postulated commutativity of the diagram means that $\varphi(x_k) = j(a_k)$ for $k = 1, \cdots, n$. Then, since φ must be a ring homomorphism, necessarily

$$\begin{aligned} \varphi\Big(\sum m_{i_1 \cdots i_n} x_1^{i_1} \cdots x_n^{i_n}\Big) &= \sum \varphi(m_{i_1 \cdots i_n}) \varphi(x_1)^{i_1} \cdots \varphi(x_m)^{i_m} \\ &= \sum \iota(m_{i_1 \cdots i_n}) j(a_1)^{i_1} \cdots j(a_m)^{i_m}, \end{aligned}$$

where $\iota : \mathbb{Z} \to R$ is the unique ring homomorphism (as $\mathbb{Z}$ is initial in Ring).

Thus, if φ exists, then it is unique. On the other hand, the formula we just obtained clearly[6] preserves the operations and sends 1 to 1, so it does define a ring homomorphism, concluding the proof. □

Example 2.2. For $n = 1$, Proposition 2.1 says that if s is any element of a ring S, then there is a unique ring homomorphism $\mathbb{Z}[x] \to S$ sending x to s and 'extending' the unique ring homomorphism $\iota : \mathbb{Z} \to S$. In this case the commutativity of S is immaterial (why?). ⌟

Example 2.3. More generally, let $\alpha : R \to S$ be a fixed ring homomorphism, and let $s \in S$ be an element commuting with $\alpha(r)$ for all $r \in R$. Then there is a unique ring homomorphism $\overline{\alpha} : R[x] \to S$ extending α and sending x to s (Exercise 2.6).

In particular, we get an 'evaluation map' for polynomials over commutative rings as follows. Given a polynomial $f(x) = \sum_{i \geq 0} a_i x^i \in R[x]$, every $r \in R$ determines an element

$$f(r) = \sum_{i \geq 0} a_i r^i :$$

this may be viewed as $\overline{\alpha}(f(x))$, where $\overline{\alpha}$ is obtained as above with $\alpha = \mathrm{id}_R : R \to R$ and $s = r$.

Thus, every polynomial $f(x)$ determines a *polynomial function* $f : R \to R$, defined by $r \mapsto f(r)$. It is a good idea to keep the two concepts of 'polynomial' and 'polynomial *function*' well distinct (cf. Exercise 2.7). ⌟

[6] Well, it is really clear that it preserves addition; multiplication requires a bit of work, which the reader would be well advised to perform. This is where the commutativity of R is used.

2.3. Monomorphisms and epimorphisms. The *kernel* of a homomorphism $\varphi : R \to S$ of *rings* is

$$\ker \varphi := \{r \in R \,|\, \varphi(r) = 0\} :$$

that is, it is nothing but the kernel of φ when the latter is viewed as a homomorphism of *groups*. As such, $\ker \varphi$ is a subgroup of $(R, +)$; it satisfies an even stronger requirement[7], which we will explore at great length in §3.

The reader may hope that an analog to Proposition II.6.12 holds in Ring, and this is indeed the case:

Proposition 2.4. *For a ring homomorphism $\varphi : R \to S$, the following are equivalent:*

(a) *φ is a monomorphism;*

(b) $\ker \varphi = \{0\}$*;*

(c) *φ is injective (as a set-function).*

Proof. We prove (a) $\implies$ (b), leaving the rest to the reader. Assume $\varphi : R \to S$ is a monomorphism and $r \in \ker \varphi$. Applying the extension property of Example 2.2, we obtain unique ring homomorpshisms $\mathrm{ev}_r : \mathbb{Z}[x] \to R$ such that $\mathrm{ev}_r(x) = r$ and $\mathrm{ev}_0 : \mathbb{Z}[x] \to R$ such that $\mathrm{ev}(x) = 0$. Consider the parallel ring homomorphisms:

$$\mathbb{Z}[x] \underset{\mathrm{ev}_0}{\overset{\mathrm{ev}_r}{\rightrightarrows}} R \xrightarrow{\varphi} S :$$

since $\varphi(r) = 0 = \varphi(0)$, the two compositions $\varphi \circ \mathrm{ev}_r$, $\varphi \circ \mathrm{ev}_0$ agree (because they agree on $\mathbb{Z}$ and they agree on x); hence $\mathrm{ev}_r = \mathrm{ev}_0$ since φ is a monomorphism. Therefore

$$r = \mathrm{ev}_r(x) = \mathrm{ev}_0(x) = 0.$$

This proves $r \in \ker \varphi \implies r = 0$, that is, (b). □

By Proposition 2.4, if $S \to R$ is a monomorphism, then S may be identified with a subset of R; the following definition formalizes this situation.

Definition 2.5. A *subring* S of a ring R is a ring whose underlying set is a subset of R and such that the inclusion function $S \hookrightarrow R$ is a ring homomorphism. ⌟

Equivalently, a subring of R is a subset $S \subseteq R$ which contains 1_R and satisfies the ring axioms with respect to the operations $+$, $\cdot$ induced from R. It is immediately checked that $S \subseteq R$ is a subring if it is a subgroup of $(R, +)$, it is closed with respect to $\cdot$, *and* it contains 1_R. As a nonexample, the zero-ring is *not* a subring of a nonzero ring R (because it does not contain 1_R).

Proposition 2.4 may induce the reader to believe that Ring and Grp are rather similar from this general categorical standpoint. But a new phenomenon occurs concerning *epimorphisms*. A *surjective* map of rings is certainly an epimorphism in Ring, since it is already an epimorphism in Set—we have run into this argument earlier (e.g., '(c) $\implies$ (a)' in Proposition II.8.18); but unlike as in Set, Grp, and Ab,

[7]Note that $\ker \varphi$ cannot be a sub*ring* in any reasonable sense, according to our definition of ring, since it does not contain a multiplicative identity in all but very pathological situations. It is a subring if the identity requirement is omitted.

epimorphisms *need not* be surjective[8] in Ring. Indeed, consider the inclusion homomorphism of rings

$$\iota : \mathbb{Z} \hookrightarrow \mathbb{Q} :$$

ι is not surjective, and hence it is not an epimorphism in Set or Ab (can the reader 'describe' $\operatorname{coker} \iota$? Exercise 2.12); but it *is* an epimorphism of rings. Indeed, if α_1, α_2 are parallel ring homomorphisms

$$\mathbb{Z} \overset{\iota}{\hookrightarrow} \mathbb{Q} \underset{\alpha_2}{\overset{\alpha_1}{\rightrightarrows}} R$$

and α_1, α_2 agree[9] on $\mathbb{Z}$, say $\alpha = \alpha_1|_{\mathbb{Z}} = \alpha_2|_{\mathbb{Z}}$, then they must agree on $\mathbb{Q}$: because for $p, q \in \mathbb{Z}$, $q \neq 0$,

$$\alpha_i\left(\frac{p}{q}\right) = \alpha_i(p)\alpha_i(q^{-1}) = \alpha(p)\alpha(q)^{-1} \qquad (i = 1, 2)$$

is the same for both[10]. Thus, ι satisfies the categorical requirement for epimorphisms (cf. §I.2.6).

Warning: Thus, in Ring, a homomorphism may well be both a monomorphism and an epimorphism *without* being an isomorphism!

2.4. Products. *Products* exist in Ring: if R_1, R_2 are rings, then $R_1 \times R_2$ may be defined by endowing the direct product of *groups* $R_1 \times R_2$ (cf. §II.3.4) with componentwise multiplication. Thus, both operations on $R_1 \times R_2$ are defined componentwise: $\forall (a_1, b_1) \in R_1$, $\forall (a_2, b_2) \in R_2$,

$$(a_1, b_1) + (a_2, b_2) := (a_1 + a_2, b_1 + b_2),$$

$$(a_1, b_1) \cdot (a_2, b_2) := (a_1 \cdot a_2, b_1 \cdot b_2).$$

The identity in $R_1 \times R_2$ is $(1_{R_1}, 1_{R_2})$. The reader will check (Exercise 2.13) that $R_1 \times R_2$ is indeed a *categorical* product in Ring.

The reader should keep in mind that in general this is *not* the only ring structure one can define on the direct product of underlying groups. For example, as mentioned in §1.2, one can define a *field* whose underlying group is $\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$, while the product ring $\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$ is very far from being a field (why?).

The situation with *coproducts* is more unfortunate—dealing with this in any generality (even for the case of *commutative* rings) requires *tensor products,* and by the time we develop tensor products (§VIII.2), we will have almost forgotten this question. But the universal property reviewed in §2.2 suffices to deal with simple examples, and the reader should work out one template case now, for fun (Exercise 2.14).

Incidentally—with the case of abelian groups in mind (§II.3.5), the reader may be tempted to consider a 'direct sum of rings', agreeing with the product in the finite

[8] Unfortunately, some references *define* ring epimorphisms as 'surjective ring homomorphisms'; this should be discouraged.

[9] As they must, since $\mathbb{Z}$ is initial.

[10] Note that $\alpha(q)$ must have a double-sided inverse in R since q has one in $\mathbb{Q}$, namely $1/q$. In particular, $\alpha_1(q^{-1}) = \alpha_2(q^{-1})$ must agree, since they must equal this unique inverse of $\alpha(q)$; cf. Proposition 1.12.

case and maybe giving something interesting in general. This is less promising than it looks: it does *not* satisfy the universal property of coproducts in Ring (why?), and, further, the 'infinite version' is not a ring because it is missing the identity.

2.5. $\mathrm{End}_{\mathsf{Ab}}(G)$**.** This is as good a place as any to expand a little on one of the main examples of rings, and one that is not quite as special as the examples reviewed in §1.2.

For every abelian group G, the group $\mathrm{End}_{\mathsf{Ab}}(G) := \mathrm{Hom}_{\mathsf{Ab}}(G, G)$ of endomorphisms of G is a *ring,* under the operations of addition and composition (as discussed in the paragraph preceding Definition 1.1). Indeed, associativity follows from the category axioms, and distributivity is immediately checked.

It is instructive to examine carefully the endomorphism ring of the *abelian group* $(\mathbb{Z}, +)$:

Proposition 2.6. $\mathrm{End}_{\mathsf{Ab}}(\mathbb{Z}) \cong \mathbb{Z}$ as rings.

Proof. Consider the function

$$\varphi : \mathrm{End}_{\mathsf{Ab}}(\mathbb{Z}) \to \mathbb{Z}$$

defined by

$$\varphi(\alpha) = \alpha(1)$$

for all group homomorphisms $\alpha : \mathbb{Z} \to \mathbb{Z}$. Then φ is a group homomorphism: the addition in $\mathrm{End}_{\mathsf{Ab}}(\mathbb{Z})$ is defined so that $\forall n \in \mathbb{Z}$

$$(\alpha + \beta)(n) = \alpha(n) + \beta(n)$$

(cf. §II.4.4); in particular

$$\varphi(\alpha + \beta) = (\alpha + \beta)(1) = \alpha(1) + \beta(1) = \varphi(\alpha) + \varphi(\beta).$$

Further, φ is a *ring* homomorphism. Indeed, for α, β in $\mathrm{End}_{\mathsf{Ab}}(\mathbb{Z})$ denote $\alpha(1)$ by a; then

$$\alpha(n) = n\alpha(1) = na = an$$

for all $n \in \mathbb{Z}$; in particular,

$$\alpha(\beta(1)) = a\beta(1) = \alpha(1)\beta(1).$$

Therefore,

$$\varphi(\alpha \circ \beta) = (\alpha \circ \beta)(1) = \alpha(\beta(1)) = \alpha(1)\beta(1) = \varphi(\alpha)\varphi(\beta)$$

as needed. Also, $\varphi(\mathrm{id}_{\mathbb{Z}}) = \mathrm{id}_{\mathbb{Z}}(1) = 1$.

Finally, φ has an inverse: for $a \in \mathbb{Z}$, let $\psi(a)$ be the homomorphism $\alpha : \mathbb{Z} \to \mathbb{Z}$ defined by

$$(\forall n \in \mathbb{Z}) : \quad \alpha(n) = an;$$

the reader will easily check that ψ is a ring homomorphism and inverse to φ.

Therefore φ is a ring isomorphism, verifying the statement. □

Proposition 2.6 gives a sense in which the ring structure on $\mathbb{Z}$ is truly 'natural': this structure arises naturally from categorical considerations, there is nothing 'arbitrary' about it.

More generally, the rings $\mathrm{End}_{\mathsf{Ab}}(G)$ and other rings arising as endomorphism rings of other structures (e.g., vector spaces) are arguably the most important class of examples of rings. The entire theory of *modules* is based on the observation that $\mathrm{End}_{\mathsf{Ab}}(G)$ is a ring for every abelian group G.

Some of the notions reviewed in §1.2 are expressed very concretely in these endomorphism rings; for example, the *group of units* in $\mathrm{End}_{\mathsf{Ab}}(G)$ is nothing but $\mathrm{Aut}_{\mathsf{Ab}}(G)$.

Every ring R interacts with the ring of endomorphisms of the underlying abelian group $(R,+)$. For $r \in R$, define left- and right-multiplication by r by λ_r, μ_r, respectively. That is, $\forall a \in R$

$$\lambda_r(a) = ra, \quad \mu_r(a) = ar.$$

The following observation is nothing more than easy notation juggling, but it is useful; it is a 'ring analog' of Cayley's theorem (Theorem II.9.5):

Proposition 2.7. *Let R be a ring. Then the function $r \mapsto \lambda_r$ is an injective ring homomorphism*

$$\lambda : R \to \mathrm{End}_{\mathsf{Ab}}(R).$$

Proof. For any $r \in R$ and for all $a, b \in R$, distributivity gives

$$\lambda_r(a+b) = r(a+b) = ra + rb = \lambda_r(a) + \lambda_r(b) :$$

this shows that λ_r is indeed an endomorphism of the *group* $(R,+)$, that is, $\lambda_r \in \mathrm{End}_{\mathsf{Ab}}(R)$.

The function $\lambda : R \to \mathrm{End}_{\mathsf{Ab}}(R)$ defined by the assignment $r \mapsto \lambda_r$ is clearly injective, since if $r \neq s$, then

$$\lambda_r(1) = r \neq s = \lambda_s(1),$$

so that $\lambda_r \neq \lambda_s$.

We have to verify that λ is in fact a homomorphism of rings. Recall that the addition in $\mathrm{End}_{\mathsf{Ab}}(R)$ is inherited from R (cf. §II.4.4): for all $r, s \in R$, $\lambda_r + \lambda_s$ is defined by

$$(\forall a \in R): \quad (\lambda_r + \lambda_s)(a) = \lambda_r(a) + \lambda_s(a).$$

Thus, the fact that λ preserves addition is an immediate consequence of distributivity: for all $r, s, a \in R$,

$$\lambda_{r+s}(a) = (r+s)a = ra + sa = \lambda_r(a) + \lambda_s(a).$$

As for multiplication, it is associativity's turn to do its job:

$$\lambda_{rs}(a) = (rs)a = r(sa) = r\lambda_s(a) = \lambda_r(\lambda_s(a)) = (\lambda_r \circ \lambda_s)(a).$$

Of course λ_1 is the identity, completing the verification. □

The function $\mu : R \to \mathrm{End}_{\mathsf{Ab}}(R)$ defined by $r \mapsto \mu_r$ is 'almost' a ring homomorphism: the reader will check (Exercise 2.18) that

$$\mu_{r+s} = \mu_r + \mu_s,$$

$$\mu_{rs} = \mu_s \circ \mu_r,$$

and $\mu_1 = \mathrm{id}_R$. That is, μ would in fact be a ring homomorphism if we 'reversed' multiplication[11] in R. Of course this issue disappears if R happens to be commutative (since then $\lambda = \mu$).

Exercises

2.1. ▷ Prove that if there is a homomorphism from a zero-ring to a ring R, then R is a zero-ring [§2.1]

2.2. Let R and S be rings, and let $\varphi : R \to S$ be a function preserving both operations $+$, $\cdot$.

- Prove that if φ is *surjective,* then necessarily $\varphi(1_R) = 1_S$.
- Prove that if $\varphi \neq 0$ and S is an integral domain, then $\varphi(1_R) = 1_S$.

(Therefore, in both cases φ is in fact a ring homomorphism).

2.3. Let S be a set, and consider the power set ring $\mathscr{P}(S)$ (Exercise 1.2) and the ring $(\mathbb{Z}/2\mathbb{Z})^S$ you constructed in Exercise 1.3. Prove that these two rings are isomorphic. (Cf. Exercise I.2.11.)

2.4. Define functions $\mathbb{H} \to \mathfrak{gl}_4(\mathbb{R})$ and $\mathbb{H} \to \mathfrak{gl}_2(\mathbb{C})$ (cf. Exercises 1.4 and 1.12) by

$$a + bi + cj + dk \mapsto \begin{pmatrix} a & b & c & d \\ -b & a & -d & c \\ -c & d & a & -b \\ -d & -c & b & a \end{pmatrix},$$

$$a + bi + cj + dk \mapsto \begin{pmatrix} a+bi & c+di \\ -c+di & a-bi \end{pmatrix}$$

for all $a, b, c, d \in \mathbb{R}$. Prove that both functions are injective ring homomorphisms. Thus, quaternions may be viewed as real or complex matrices.

2.5. ¬ The *norm* of a quaternion $w = a + bi + cj + dk$, with $a, b, c, d \in \mathbb{R}$, is the real number $N(w) = a^2 + b^2 + c^2 + d^2$.

Prove that the function from the multiplicative group $\mathbb{H}^*$ of nonzero quaternions to the multiplicative group $\mathbb{R}^+$ of positive real numbers, defined by assigning to each nonzero quaternion its norm, is a homomorphism. Prove that the kernel of this homomorphism is isomorphic to $\mathrm{SU}_2(\mathbb{C})$ (cf. Exercise II.6.3). [4.10, IV.5.17, V.6.19]

[11] This issue is entirely analogous to the business of right-actions vs. left-actions; cf. §II.9.3, especially Exercise II.9.3.

2.6. ▷ Verify the 'extension property' of polynomial rings, stated in Example 2.3. [§2.2]

2.7. ▷ Let $R = \mathbb{Z}/2\mathbb{Z}$, and let $f(x) = x^2 - x$; note $f(x) \neq 0$. What is the polynomial function $R \to R$ determined by $f(x)$? [§2.2, §V.4.2, §V.5.1]

2.8. Prove that every subring of a field is an integral domain.

2.9. ¬ The *center* of a ring R consists of the elements a such that $ar = ra$ for all $r \in R$. Prove that the center is a subring of R.

Prove that the center of a division ring is a field. [2.11, IV.2.17, VII.5.14, VII.5.16]

2.10. ¬ The *centralizer* of an element a of a ring R consists of the elements $r \in R$ such that $ar = ra$. Prove that the centralizer of a is a subring of R, for every $a \in R$.

Prove that the center of R is the intersection of all its centralizers.

Prove that every centralizer in a division ring is a division ring. [2.11, IV.2.17, VII.5.16]

2.11. ¬ Let R be a division ring consisting of p^2 elements, where p is a prime. Prove that R is commutative, as follows:

- If R is not commutative, then its center C (Exercise 2.9) is a proper subring of R. Prove that $|C|$ would then consist of p elements.
- Let $r \in R$, $r \notin C$. Prove that the centralizer of r (Exercise 2.10) contains both r and C.
- Deduce that the centralizer of r is the whole of R.
- Derive a contradiction, and conclude that R had to be commutative (hence, a field).

This is a particular case of Wedderburn's theorem: every finite division ring is a field. [IV.2.17, VII.5.16]

2.12. ▷ Consider the inclusion map $\iota : \mathbb{Z} \hookrightarrow \mathbb{Q}$. Describe the cokernel of ι in Ab and its cokernel in Ring (as defined by the appropriate universal property in the style of the one given in §II.8.6). [§2.3, §5]

2.13. ▷ Verify that the 'componentwise' product $R_1 \times R_2$ of two rings satisfies the universal property for products in a category, given in §I.5.4. [§2.4]

2.14. ▷ Verify that $\mathbb{Z}[x_1, x_2]$ (along with the evident morphisms) satisfies the universal property for the coproduct of two copies of $\mathbb{Z}[x]$ in the category of *commutative* rings. Explain why it does not satisfy it in Ring. [§2.4]

2.15. ▷ For $m > 1$, the abelian groups $(\mathbb{Z}, +)$ and $(m\mathbb{Z}, +)$ are manifestly isomorphic: the function $\varphi : \mathbb{Z} \to m\mathbb{Z}$, $n \mapsto mn$ is a group isomorphism. Use this isomorphism to transfer the structure of 'ring without identity' $(m\mathbb{Z}, +, \cdot)$ back onto $\mathbb{Z}$: give an explicit formula for the 'multiplication' $\bullet$ this defines on $\mathbb{Z}$ (that is, such that $\varphi(a \bullet b) = \varphi(a) \cdot \varphi(b)$). Explain why structures induced by different positive integers m are nonisomorphic as 'rings without 1'.

(This shows that there are many different ways to give a structure of ring without identity to the *group* $(\mathbb{Z}, +)$. Compare this observation with Exercise 2.16.) [§2.1]

2.16. ▷ Prove that there is (up to isomorphism) only one structure of ring *with identity* on the abelian group $(\mathbb{Z}, +)$. (Hint: Let R be a ring whose underlying group is $\mathbb{Z}$. By Proposition 2.7, there is an injective ring homomorphism $\lambda : R \to \mathrm{End}_{\mathsf{Ab}}(R)$, and the latter is isomorphic to $\mathbb{Z}$ by Proposition 2.6. Prove that λ is surjective.) [§2.1, 2.15]

2.17. ¬ Let R be a ring, and let $E = \mathrm{End}_{\mathsf{Ab}}(R)$ be the ring of endomorphisms of the underlying abelian group $(R, +)$. Prove that the center of E is isomorphic to the center of R. (Prove that if $\alpha \in E$ commutes with all right-multiplications by elements of R, then α is left-multiplication by an element of R; then use Proposition 2.7.) In particular, this shows that if R is commutative, then the center of $\mathrm{End}_R(R)$ is isomorphic to R. [VIII.3.15]

2.18. ▷ Verify the statements made about right-multiplication μ, following Proposition 2.7. [§2.5]

2.19. Prove that for $n \in \mathbb{Z}$ a positive integer, $\mathrm{End}_{\mathsf{Ab}}(\mathbb{Z}/n\mathbb{Z})$ is isomorphic to $\mathbb{Z}/n\mathbb{Z}$ *as a ring.*

---

3. Ideals and quotient rings

3.1. Ideals. In both Set and Grp we have been able to 'classify' surjective morphisms: in both cases, a surjective morphism is, up to natural identifications, a *quotient* by a (suitable) equivalence relation. In Grp, we have seen that such equivalence relations arise in fact from certain substructures: normal subgroups.

The situation in Ring is analogous. We will establish a canonical decomposition for rings, modeled after Theorems I.2.7 and II.8.1; the corresponding version of the 'first isomorphism theorem' (Corollary II.8.2) will identify every surjective ring homomorphism with a quotient by a suitable substructure. The role of normal subgroups will be played by *ideals.*

Definition 3.1. Let R be a ring. A subgroup I of $(R, +)$ is a *left-ideal* of R if $rI \subseteq I$ for all $r \in R$; that is,

$$(\forall r \in R)(\forall a \in I): \quad ra \in I;$$

it is a *right*-ideal if $Ir \subseteq I$ for all $r \in R$; that is,

$$(\forall r \in R)(\forall a \in I): \quad ar \in I.$$

A *two-sided ideal* is a subgroup I which is both a left- and a right-ideal. ⌟

Of course in a commutative ring there is no distinction between left- and right-ideals. Even in the general setting, we will almost exclusively be concerned with two-sided ideals; thus I will omit qualifiers, and an *ideal* of a ring will implicitly be two-sided.

Remark 3.2. As seen in §2.3, a *subring* of R is a subset $S \subseteq R$ which contains 1_R and satisfies the ring axioms with respect to the operations $+, \cdot$ induced from R.

Ideals are close to being subrings: they are subgroups, and they are closed with respect to multiplication. *But* the only ideal of a ring R containing 1_R is R itself: this is an immediate consequence of the 'absorption properties' stated in Definition 3.1. Thus ideals are in general *not* subrings; they are 'rngs'. ⌟

Ideals are considerably more important than subrings in the development of the theory of rings[12]. Of course the *image* of a ring homomorphism is necessarily a subring of the target; but a lesson learned from §II.8.1 and following is that *kernels* really capture the structure of a homomorphism, and kernels are ideals:

Example 3.3. Let $\varphi : R \to S$ be any ring homomorphism. Then $\ker \varphi$ is an ideal of R.

Indeed, we know already that $\ker \varphi$ is a subgroup; we have to verify the absorption properties. These are an immediate consequence of Lemma 1.2: for all $r \in R$, all $a \in \ker \varphi$, we have

$$\varphi(ra) = \varphi(r)\varphi(a) = 0 \cdot a = 0,$$

$$\varphi(ar) = \varphi(a)\varphi(r) = a \cdot 0 = 0.$$

More generally, it is easy to verify that the inverse image of an ideal is an ideal (Exercise 3.2) and $\{0_S\}$ is clearly an ideal of S.

Similarly to the situation with normal subgroups in the context of groups, we will soon see that 'kernels of ring homomorphisms' and 'ideals' are in fact equivalent concepts. ⌟

3.2. Quotients. Let I be a subgroup of the abelian group $(R, +)$ of a ring R. Subgroups of abelian groups are automatically normal, so we have a quotient *group* R/I, whose elements are the cosets of I:

$$r + I$$

(written, of course, in additive notation). Further, we have a surjective *group* homomorphism

$$\pi : R \to \frac{R}{I}, \quad r \mapsto r + I.$$

As we have explored in great detail for groups, this construction satisfies a suitable universal property with respect to group homomorphisms (Theorem II.7.12). *Of course* we are now going to ask under what circumstances this construction can be performed in Ring, satisfying the analogous universal property with respect to ring homomorphisms.

That is, what should we ask of I, in order to have a ring structure on R/I, so that π becomes a ring homomorphism? Go figure this out for yourself, before reading ahead!

[12] Arguably, the reason is that ideals are precisely the *submodules* of a ring R.

As so often is the case, the requirement is precisely what tells us the answer. Since π will have to be a ring homomorphism, the multiplication in R/I *must* be as follows: for all $(a+I)$, $(b+I)$ in R/I,

$$(a+I)\cdot(b+I) = \pi(a)\cdot\pi(b) \overset{!?}{=} \pi(ab) = ab+I,$$

where the equality $\overset{!?}{=}$ is forced if π is to be a ring homomorphism.

This says that there is only one sensible ring structure on R/I, given by

$$(\forall a, b \in R): \quad (a+I)(b+I) := ab+I.$$

The reader should realize right away that *if* this operation is well-defined, then it does make R/I into a ring: associativity will be inherited from the associativity in R, and the identity will simply be the coset $1+I$ of 1.

So all is well, if the proposed operation is well-defined. *But* is it going to be well-defined?

Example 3.4. It need not be, if I is an arbitrary subgroup of R. For example, take $\mathbb{Z}$ as a subgroup of $\mathbb{Q}$; then

$$0+\mathbb{Z} = 1+\mathbb{Z}$$

$(=\mathbb{Z})$ as elements of the group $\mathbb{Q}/\mathbb{Z}$, and $\frac{1}{2}+\mathbb{Z}$ is another coset, yet

$$0\cdot\frac{1}{2}+\mathbb{Z} = \mathbb{Z} \neq \frac{1}{2}+\mathbb{Z} = 1\cdot\frac{1}{2}+\mathbb{Z}.$$ ⌟

Assume then that the operation *is* well-defined, so that R/I is a ring and $\pi: R \to R/I$ is a ring homomorphism. What does this say about I?

Answer: I is the kernel of $R \to R/I$, so necessarily I must be an ideal, as seen in Example 3.3!

Conversely, let us assume I is an ideal of R, and verify that the proposed prescription for the operation in R/I is well-defined. For this, suppose

$$a'+I = a''+I \quad \text{and} \quad b'+I = b''+I;$$

recall that this means that $a''-a' \in I$, $b''-b' \in I$; then

$$a''b''-a'b' = a''b''-a''b'+a''b'-a'b' = a''(b''-b')+(a''-a')b' \in I,$$

using both the left-absorption and right-absorption properties of Definition 3.1. This says precisely that

$$a'b'+I = a''b''+I,$$

proving that the operation is well-defined.

Summarizing, we have verified that R/I is a ring, in such a way that the canonical projection $\pi: R \to R/I$ is a *ring* homomorphism, if and only if I is an *ideal* of R.

Definition 3.5. This ring R/I is called the *quotient ring* of R *modulo* I. ⌟

Example 3.6. We know that all subgroups of $(\mathbb{Z},+)$ are of the form $n\mathbb{Z}$ for a nonnegative integer n (Proposition II.6.9). It is immediately verified that all subgroups of $\mathbb{Z}$ are in fact *ideals* of the ring $(\mathbb{Z},+,\cdot)$. The quotients $\mathbb{Z}/n\mathbb{Z}$ are of course nothing but the rings so-denoted in §1.2 (and earlier).

The fact that $\mathbb{Z}$ is initial in Ring now prompts a natural definition. For a ring R, let $f : \mathbb{Z} \to R$ be the unique ring homomorphism, defined by $a \mapsto a \cdot 1_R$. Then $\ker f = n\mathbb{Z}$ for a well-defined nonnegative integer n determined by R.

Definition 3.7. The *characteristic* of R is this nonnegative integer n. ⌟

Thus, the characteristic of R is $n > 0$ if the order of 1_R *as an element of* $(R, +)$ is a positive integer n, while the characteristic is 0 if the order of 1_R is ∞. ⌟

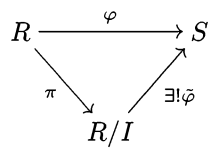
We have now fulfilled our promise to identify the notions of ideal and kernel of ring homomorphisms: every kernel is an ideal (cf. Example 3.3); and on the other hand every ideal I is the kernel of the ring homomorphisms $R \to R/I$. We have the slogan:

$$\text{kernel} \iff \text{ideal}$$

in the context of ring theory.

The key universal property holds in this context as it does for groups; cf. Theorem II.7.12. We reproduce the statement here for reference, but the reader should realize that very little needs to be proven at this point: the needed (group) homomorphism exists and is unique by Theorem II.7.12, and verifying it is a *ring* homomorphism is immediate.

Theorem 3.8. *Let I be a two-sided ideal of a ring R. Then for every ring homomorphism $\varphi : R \to S$ such that $I \subseteq \ker \varphi$ there exists a unique ring homomorphism $\tilde{\varphi} : R/I \to S$ so that the diagram*



commutes.

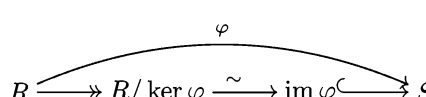
As a reminder to the lazy reader, $\tilde{\varphi}$ is defined by

$$\tilde{\varphi}(r + I) := \varphi(r);$$

(part of) the content of the theorem is that this function is well-defined (if $I \subseteq \ker \varphi$), and it is a ring homomorphism.

3.3. Canonical decomposition and consequences. As the reader should now expect, Theorem 3.8 is the key element in a standard decomposition of every ring homomorphism. This is entirely analogous to the decomposition of set-functions studied in §I.2.8 and the decomposition of group homomorphisms obtained in §II.8.1. Here is the statement:

Theorem 3.9. *Every ring homomorphism $\varphi : R \to S$ may be decomposed as follows:*



where the isomorphism $\tilde{\varphi}$ in the middle is the homomorphism induced by φ (as in Theorem 3.8).

The reader will realize that this statement requires no proof at this point: the decomposition holds at the level of groups (by Theorem II.8.1) and the maps are all ring homomorphisms as observed earlier in this section.

The 'first isomorphism theorem' for rings is an immediate corollary:

Corollary 3.10. *Suppose $\varphi : R \to S$ is a* surjective *ring homomorphism. Then*

$$S \cong \frac{R}{\ker \varphi}.$$

As in the group case, the reader should develop the healthy instinctive reaction of viewing every surjective homomorphism of rings as a quotient (by an ideal), up to a natural identification.

Equally instinctive should be the realization that *ideals of a quotient R/I* are in one-to-one correspondence with *ideals of R containing I*. Once more, not a whole lot is new here: we already know (cf. §II.8.3) that the function

$$u : \{\text{subgroups } J \text{ of } R \text{ containing } I\} \to \{\text{subgroups of } R/I\}$$

defined by $u(J) = J/I$ is a bijection preserving inclusions; it takes a moment to check that J/I is an *ideal* of R/I if and only if J is an ideal of R. The main content of this observation is best packaged in the corresponding 'third isomorphism theorem':

Proposition 3.11. *Let I be an ideal of a ring R, and let J be an ideal of R containing I. Then J/I is an ideal of R/I, and*

$$\frac{R/I}{J/I} \cong \frac{R}{J}.$$

Proof. Since $I \subseteq J = \ker(R \to R/J)$, we have an induced ring homomorphism

$$\varphi : R/I \to R/J$$

by Theorem 3.8. Explicitly, $\varphi(r + I) = r + J$; φ is manifestly surjective. Since

$$\ker \varphi = \{r + I \,|\, \varphi(r + I) = J\} = \{r + I \,|\, r + J = J\} = \{r + I \,|\, r \in J\} = J/I,$$

we see that J/I is an ideal (since it is a kernel), and the stated isomorphism follows from Corollary 3.10. □

What about the 'second' isomorphism theorem? This would be a relation between the ideals

$$\frac{I + J}{I}, \quad \frac{J}{I \cap J}$$

of the rings R/I, $R/(I \cap J)$, respectively, assuming I and J are ideals of R (and where it is immediately checked that $I + J$ and $I \cap J$ are indeed ideals[13]).

The reader may want to go back to §II.8.4 for the version of this story for groups. Our feeling is that Ring is not the best place to play this game, since

[13]The notation $I + J$ should not surprise the reader: I, J are subgroups of the abelian group R, so we know what it means to 'add' them; cf. §II.7.1.

$(I+J)/I$, $J/(I\cap J)$ are not even rings according to my conventions. *Modules* will be a more natural context for this result.

In any case, the reader will benefit the most from exploring the matter on his/her own; cf. Exercise 3.17.

Exercises

3.1. Prove that the image of a ring homomorphism $\varphi : R \to S$ is a subring of S. What can you say about φ if its image is an ideal of S? What can you say about φ if its kernel is a subring of R?

3.2. ▷ Let $\varphi : R \to S$ be a ring homomorphism, and let J be an ideal of S. Prove that $I = \varphi^{-1}(J)$ is an ideal of R. [§3.1]

3.3. ¬ Let $\varphi : R \to S$ be a ring homomorphism, and let J be an ideal of R.

- Show that $\varphi(J)$ need not be an ideal of S.
- Assume that φ is surjective; then prove that $\varphi(J)$ *is* an ideal of S.
- Assume that φ is surjective, and let $I = \ker\varphi$; thus we may identify S with R/I. Let $\overline{J} = \varphi(J)$, an ideal of R/I by the previous point. Prove that

$$\frac{R/I}{\overline{J}} \cong \frac{R}{I+J}.$$

(Of course this is just a rehash of Proposition 3.11.) [4.11]

3.4. Let R be a ring such that every subgroup of $(R,+)$ is in fact an ideal of R. Prove that $R \cong \mathbb{Z}/n\mathbb{Z}$, where n is the characteristic of R.

3.5. ¬ Let J be a *two-sided* ideal of the ring $\mathcal{M}_n(R)$ of $n\times n$ matrices over a ring R. Prove that a matrix $A \in \mathcal{M}_n(R)$ belongs to J if and only if the matrices obtained by placing any entry of A in any position, and 0 elsewhere, belong to J. (Hint: Carefully contemplate the operation $\left(\begin{smallmatrix}0&0&0\\0&0&0\\1&0&0\end{smallmatrix}\right)\left(\begin{smallmatrix}a&b&c\\d&e&f\\g&h&i\end{smallmatrix}\right)\left(\begin{smallmatrix}0&0&0\\1&0&0\\0&0&0\end{smallmatrix}\right) = \left(\begin{smallmatrix}0&0&0\\0&0&0\\b&0&0\end{smallmatrix}\right)$.) [3.6]

3.6. ¬ Let J be a two-sided ideal of the ring $\mathcal{M}_n(R)$ of $n \times n$ matrices over a ring R, and let $I \subseteq R$ be the set of $(1,1)$ entries of matrices in J. Prove that I is a two-sided ideal of R and J consists precisely of those matrices whose entries all belong to I. (Hint: Exercise 3.5.) [3.9]

3.7. Let R be a ring, and let $a \in R$. Prove that Ra is a left-ideal of R and aR is a right-ideal of R. Prove that a is a left-, resp. right-, unit if and only if $R = aR$, resp. $R = Ra$.

3.8. ▷ Prove that a ring R is a division ring if and only if its only left-ideals and right-ideals are $\{0\}$ and R.

In particular, a commutative ring R is a field if and only if the only ideals of R are $\{0\}$ and R. [3.9, §4.3]

3.9. ¬ Counterpoint to Exercise 3.8: It is *not* true that a ring R is a division ring if and only if its only two-sided ideals are $\{0\}$ and R. A nonzero ring with this property is said to be *simple;* by Exercise 3.8, fields are the only simple *commutative* rings.

Prove that $\mathcal{M}_n(\mathbb{R})$ is simple. (Use Exercise 3.6.) [4.20]

3.10. ▷ Let $\varphi : k \to R$ be a ring homomorphism, where k is a field and R is a nonzero ring. Prove that φ is *injective.* [§V.4.2, §V.5.2]

3.11. Let R be a ring containing $\mathbb{C}$ as a subring. Prove that there are no ring homomorphisms $R \to \mathbb{R}$.

3.12. ▷ Let R be a *commutative* ring. Prove that the set of nilpotent elements of R is an ideal of R. (Cf. Exercise 1.6. This ideal is called the *nilradical* of R.)

Find a noncommutative ring in which the set of nilpotent elements is *not* an ideal. [3.13, 4.18, V.3.13, §VII.2.3]

3.13. ¬ Let R be a commutative ring, and let N be its nilradical (cf. Exercise 3.12). Prove that R/N contains no nonzero nilpotent elements. (Such a ring is said to be *reduced.*) [4.6, VII.2.8]

3.14. ¬ Prove that the characteristic of an integral domain is either 0 or a prime integer. Do you know any ring of characteristic 1? [V.4.17]

3.15. ¬ A ring R is[14] *Boolean* if $a^2 = a$ for all $a \in R$. Prove that $\mathscr{P}(S)$ is Boolean, for every set S (cf. Exercise 1.2). Prove that every Boolean ring is commutative, and has characteristic 2. Prove that if an integral domain R is Boolean, then $R \cong \mathbb{Z}/2\mathbb{Z}$. [4.23, V.6.3]

3.16. ¬ Let S be a set and $T \subseteq S$ a subset. Prove that the subsets of S contained in T form an ideal of the power set ring $\mathscr{P}(S)$. Prove that if S is finite, then *every* ideal of $\mathscr{P}(S)$ is of this form. For S infinite, find an ideal of $\mathscr{P}(S)$ that is *not* of this form. [V.1.5]

3.17. ▷ Let I, J be ideals of a ring R. State and prove a precise result relating the ideals $(I + J)/I$ of R/I and $J/(I \cap J)$ of $R/(I \cap J)$. [§3.3]

---

4. Ideals and quotients: Remarks and examples. Prime and maximal ideals

4.1. Basic operations. It is often convenient to define ideals in terms of a set of *generators.*

Let $a \in R$ be any element of a ring. Then the subset $I = Ra$ of R is a left-ideal of R. Indeed, for all $r \in R$ we have

$$rI = rRa \subseteq Ra$$

as needed. Similarly, aR is right-ideal.

[14] After George Boole.

In the commutative case, these two subsets coincide and are denoted (a). This is the *principal* ideal *generated by a*. For example, the zero-ideal $\{0\} = (0)$ and the whole ring $R = (1)$ are both principal ideals.

In general (Exercise 4.1), if $\{I_\alpha\}_{\alpha \in A}$ is a family of ideals of a ring R, then the sum $\sum_\alpha I_\alpha$ is an ideal of R. If a_α is any collection of elements of a commutative ring R, then

$$(a_\alpha)_{\alpha \in A} := \sum_{\alpha \in A} (a_\alpha)$$

is the ideal *generated by the elements* a_α. In particular,

$$(a_1, \ldots, a_n) = (a_1) + \cdots + (a_n)$$

is the smallest ideal of R containing $a_1, \ldots, a_n$; this ideal consists of the elements of R that may be written as

$$r_1 a_1 + \cdots + r_n a_n$$

for $r_1, \ldots, r_n \in R$. An ideal I of a commutative ring R is *finitely generated* if $I = (a_1, \ldots, a_n)$ for some $a_1, \ldots, a_n \in R$.

Example 4.1. It is a good idea to get used to a bit of 'calculus' of ideals and quotients in terms of generators; judicious use of the isomorphism theorems yields convenient statements. For example, let R be a commutative ring, and let $a, b \in R$; denote by $\overline{b}$ the class of b in $R/(a)$. Then

$$(R/(a))/(\overline{b}) \cong R/(a, b).$$

Indeed, this is a particular case of Proposition 3.11, since

$$(\overline{b}) = \frac{(a, b)}{(a)}$$

as ideals of $R/(a)$. ⌟

Note that principal ideals are (very special) finitely generated ideals. These notions are so important that we give special names to rings in which they are satisfied by *every* ideal.

Definition 4.2. A commutative ring R is *Noetherian* if every ideal of R is finitely generated. ⌟

Definition 4.3. An integral domain R is a *PID* ('Principal Ideal Domain') if every ideal of R is principal. ⌟

Thus, PIDs are (very special) Noetherian rings. In due time we will deal at length with these classes of rings (cf. Chapter V); Noetherian rings are very important in number theory and algebraic geometry.

The reader is already familiar with an important PID:

Proposition 4.4. *$\mathbb{Z}$ is a PID.*

Proof. Let $I \subseteq \mathbb{Z}$ be an ideal. Since I is a subgroup, $I = n\mathbb{Z}$ for some $n \in \mathbb{Z}$, by Proposition II.6.9. Since $n\mathbb{Z} = (n)$, this shows that I is principal. □

The fact that $\mathbb{Z}$ is a PID captures precisely 'why' greatest common divisors behave as they do in $\mathbb{Z}$: if m, n are integers, then the ideal (m, n) must be principal, and hence

$$(m, n) = (d)$$

for some (positive) integer d. This integer is manifestly the gcd of m and n: since $m \in (d)$ and $n \in (d)$, then $d \mid m$ and $d \mid n$, etc.

If k is a field, the ring of polynomials $k[x]$ is also a PID; proving this is easy, using the 'division with remainder' that we will run into very soon (§4.2); the reader should work this out on his/her own (Exercise 4.4) now. This fact will be absorbed in the general theory when we review the general notion of 'Euclidean domain', in §V.2.4.

By contrast, the ring $\mathbb{Z}[x]$ is *not* a PID: indeed, the reader should be able to verify that the ideal $(2, x)$ cannot be generated by a single element. As we will see in due time, greatest common divisors make good sense in a ring such as $\mathbb{Z}[x]$, but the matter is a little more delicate, since this ring is not a PID[15].

There are several more basic operations involving ideals; for now, the following two will suffice.

- Again assume that $\{I_\alpha\}_{\alpha \in A}$ is a collection of ideals of a ring R. Then the intersection $\bigcap_{\alpha \in A} I_\alpha$ is (clearly) an ideal of R; it is the largest ideal contained in all of the ideals I_α.
- If I, J are ideals of R, then IJ denotes the ideal *generated* by all products ij with $i \in I$, $j \in J$. More generally, if $I_1, \ldots, I_n$ are ideals in R, then the 'product' $I_1 \cdots I_n$ denotes the ideal generated by all products $i_1 \cdots i_n$ with $i_k \in I_k$.

The reader should note the clash of notation: in the context of groups (especially §II.8.4) IJ would mean *something else.* Watch out!

It is clear that $IJ \subseteq I \cap J$: every element ij with $i \in I$ and $j \in J$ is in I (because I is a right-ideal) and in J (because J is a left-ideal); therefore $I \cap J$ contains all products ij, and hence it must contain the ideal IJ they generate. Sometime the product agrees with the intersection:

$$(4) \cap (3) = (12) = (4) \cdot (3) \quad \text{in } \mathbb{Z};$$

and sometime it does not:

$$(4) \cap (6) = (12) \neq (24) = (4) \cdot (6).$$

The matter of whether $IJ = I \cap J$ is often subtle; a prototype situation in which this equality holds is given in Exercise 4.5.

4.2. Quotients of polynomial rings. We have already observed that the quotient $\mathbb{Z}/n\mathbb{Z}$ is our familiar ring of congruence classes modulo n. Quotients of *polynomial* rings by principal ideals are a good source of 'concrete', but maybe less familiar, examples.

[15] It is, however, a 'UFD', that is, a 'unique factorization domain'. This suffices for a good notion of gcd; cf. §V.2.1.

Let R be a (nonzero) ring, and let

$$f(x) = x^d + a_{d-1}x^{d-1} + \cdots + a_1 x + a_0 \in R[x]$$

be a polynomial; for convenience, we are assuming that $f(x)$ is *monic,* that is, its leading coefficient (the coefficient of the highest power of x appearing in $f(x)$) is 1. In terms of ideals, this is not a serious requirement if the coefficient ring R is a field (Exercise 4.7), but it may be substantial otherwise: for example, $(2x) \subsetneq \mathbb{Z}[x]$ cannot be generated by a monic polynomial. Also note that a monic polynomial is necessarily a non-zero-divisor (Exercise 4.8) and that if $f(x)$ is monic, then $\deg(f(x)q(x)) = \deg f(x) + \deg q(x)$ for all polynomials $q(x)$.

It is convenient to assume that $f(x)$ is monic because we can then *divide* by $f(x)$, with remainder. That is, if $g(x) \in R[x]$ is another polynomial, then there exist polynomials $q(x)$, $r(x) \in R[x]$ such that

$$g(x) = f(x)q(x) + r(x)$$

and[16] $\deg r(x) < \deg f(x)$. This is simply the process of 'long division' of polynomials, which is surely familiar to the reader, and can be performed over any ring when dividing by *monic* polynomials[17].

The situation appears then to be similar to the situation in $\mathbb{Z}$, where we also have division with remainder. Quotients and remainders are uniquely[18] determined by $g(x)$ and $f(x)$:

Lemma 4.5. *Let $f(x)$ be a monic polynomial, and assume*

$$f(x)q_1(x) + r_1(x) = f(x)q_2(x) + r_2(x)$$

with both $r_1(x)$ and $r_2(x)$ polynomials of degree $< \deg f(x)$. Then $q_1(x) = q_2(x)$ and $r_1(x) = r_2(x)$.

Proof. Indeed, we have

$$f(x)(q_1(x) - q_2(x)) = r_2(x) - r_1(x);$$

if $r_2(x) \neq r_1(x)$, then $r_2(x) - r_1(x)$ has degree $< \deg f(x)$, while $f(x)(q_1(x) - q_2(x))$ has degree $\geq \deg f(x)$, giving a contradiction. Therefore $r_1(x) = r_2(x)$, and $q_1(x) = q_2(x)$ follows right away since monic polynomials are non-zero-divisors. □

The preceding considerations may be summarized in a rather efficient way in the language of ideals and cosets. We will now restrict ourselves to the *commutative* case, mostly for notational convenience, but also because this will guarantee that ideals are two-sided ideals, so that quotients are defined *as rings* (cf. §3.2).

---

[16]Note: With the convention that the degree of the polynomial 0 is $-\infty$, the condition $\deg r(x) < \deg f(x)$ is satisfied by $r(x) = 0$.

[17]The key point is that if $n > d$, then for all $a \in R$ we have $ax^n = ax^{n-d}f(x) + h(x)$ for some polynomial $h(x)$ of degree $< n$. Arguing inductively, this shows that we may perform division by $f(x)$ with remainder for all 'monomials' ax^n, and hence (by linearity) for all polynomials $g(x) \in R[x]$.

[18]This assertion has to be taken with a grain of salt in the noncommutative case, as different quotients and remainders may arise if we divide 'on the left' rather than 'on the right'.

Assume then that R is a commutative ring. What we have shown is that, if $f(x)$ is monic, then for every $g(x) \in R[x]$ there exists a *unique* polynomial $r(x)$ of degree $< \deg f(x)$ and such that

$$g(x) + (f(x)) = r(x) + (f(x))$$

as cosets of the principal ideal $(f(x))$ in $R[x]$.

Refining this observation leads to a useful group-theoretic statement. Note that polynomials of degree $< d$ may be seen as elements of a direct sum

$$R^{\oplus d} = \underbrace{R \oplus \cdots \oplus R}_{d \text{ times}}:$$

indeed, the function $\psi : R^{\oplus d} \to R[x]$ defined by

$$\psi((r_0, r_1, \cdots, r_{d-1})) = r_0 + r_1 x + \cdots + r_{d-1} x^{d-1}$$

is clearly an injective homomorphism of abelian groups, hence an isomorphism onto its image, and this consists precisely of the polynomials of degree $< d$. We will glibly identify $R^{\oplus d}$ with this set of polynomials for the purpose of the discussion that follows.

The next result may be seen as a way to concoct many different and interesting *ring* structures on the direct sum $R^{\oplus d}$:

Proposition 4.6. *Let R be a commutative ring, and let $f(x) \in R[x]$ be a monic polynomial of degree d. Then the function*

$$\varphi : R[x] \to R^{\oplus d}$$

defined by sending $g(x) \in R[x]$ to the remainder of the division of $g(x)$ by $f(x)$ induces an isomorphism *of abelian groups*

$$\frac{R[x]}{(f(x))} \cong R^{\oplus d}.$$

Proof. The given function φ is well-defined by Lemma 4.5, and it is surjective since it has a right inverse (that is, the function $\psi : R^{\oplus d} \to R[x]$ defined above).

We claim that φ is a homomorphism of abelian groups. Indeed, if

$$g_1(x) = f(x)q_1(x) + r_1(x) \quad \text{and} \quad g_2(x) = f(x)q_2(x) + r_2(x)$$

with $\deg r_1(x) < d$, $\deg r_2(x) < d$, then

$$g_1(x) + g_2(x) = f(x)(q_1(x) + q_2(x)) + (r_1(x) + r_2(x))$$

and $\deg(r_1(x) + r_2(x)) < d$: this implies (again by Lemma 4.5)

$$\varphi(g_1(x) + g_2(x)) = r_1(x) + r_2(x) = \varphi(g_1(x)) + \varphi(g_2(x)).$$

By the first isomorphism theorem for abelian groups, then, φ induces an isomorphism

$$\frac{R[x]}{\ker \varphi} \cong R^{\oplus d}.$$

On the other hand, $\varphi(g(x)) = 0$ if and only if $g(x) = f(x)q(x)$ for some $q(x) \in R[x]$, that is, if and only if $g(x)$ is in the principal ideal generated by $f(x)$. This shows $\ker \varphi = (f(x))$, concluding the proof. □

Example 4.7. Assume $f(x)$ is monic of degree 1: $f(x) = x - a$ for some $a \in R$. Then the remainder of $g(x)$ after division by $f(x)$ is simply the 'evaluation' $g(a)$ (cf. Example 2.3): indeed,

$$g(x) = (x - a)q(x) + r$$

for some $r \in R$ (the remainder must have degree < 1; hence it is a constant); evaluating at a gives

$$g(a) = (a - a)q(a) + r = 0 \cdot q(a) + r = r$$

as claimed. In particular, $g(a) = 0$ if and only if $g(x) \in (x - a)$.

The content of Proposition 4.6 in this case is that the evaluation map

$$R[x] \to R, \quad g(x) \mapsto g(a)$$

induces an isomorphism

$$\frac{R[x]}{(x - a)} \cong R$$

of abelian groups; the reader will verify (either by hand or by invoking Corollary 3.10) that this is in fact an isomorphism of rings. ⌟

Example 4.8. It is fun to analyze higher-degree examples. For every monic $f(x) \in R[x]$ of degree d, Proposition 4.6 gives a *different* ring (that is, $R[x]/(f(x))$) isomorphic to $R^{\oplus d}$ as a group; one can then use this isomorphism to define a new ring structure onto the group $R^{\oplus d}$.

For $d = 1$ all these structures are isomorphic (as seen in Example 4.7); but interesting structures already arise in degree 2. For a concrete example, apply this procedure with $f(x) = x^2 + 1$: Proposition 4.6 gives an isomorphism of groups

$$R \oplus R \cong \frac{R[x]}{(x^2 + 1)};$$

what multiplication does this isomorphism induce on $R \oplus R$? Take two elements (a_0, a_1), (b_0, b_1) of $R \oplus R$. With the notation used in Proposition 4.6, we have

$$(a_0, a_1) = \varphi(a_0 + a_1 x), \quad (b_0, b_1) = \varphi(b_0 + b_1 x).$$

Now a bit of high-school algebra gives

$$\begin{aligned}(a_0 + a_1 x)(b_0 + b_1 x) &= a_0 b_0 + (a_0 b_1 + a_1 b_0)x + a_1 b_1 x^2 \\ &= (x^2 + 1)a_1 b_1 + ((a_0 b_0 - a_1 b_1) + (a_0 b_1 + a_1 b_0)x)\end{aligned}$$

which shows

$$\varphi((a_0 + a_1 x)(b_0 + b_1 x)) = (a_0 b_0 - a_1 b_1, a_0 b_1 + a_1 b_0).$$

Therefore, the multiplication induced on $R \oplus R$ by this procedure is defined by

$$(a_0, a_1) \cdot (b_0, b_1) = (a_0 b_0 - a_1 b_1, a_0 b_1 + a_1 b_0).$$

This recipe may seem somewhat arbitrary, but note that upon taking $R = \mathbb{R}$, the ring of real numbers, and identifying pairs $(x, y) \in \mathbb{R} \oplus \mathbb{R}$ with *complex numbers*

$x + iy$, the multiplication we obtained on $\mathbb{R} \oplus \mathbb{R}$ matches precisely the ordinary multiplication in $\mathbb{C}$. Therefore,

$$\frac{\mathbb{R}[x]}{(x^2+1)} \cong \mathbb{C}$$

as rings. In other words, this procedure constructs the ring $\mathbb{C}$ 'from scratch', starting from $\mathbb{R}[x]$. ⌟

The point is that the polynomial equation $x^2 + 1 = 0$ has no solutions in $\mathbb{R}$; the quotient $\mathbb{R}[x]/(x^2+1)$ produces a ring containing a copy of $\mathbb{R}$ and in which the polynomial *does* have roots (that is, $\pm$ the class of x in the quotient). The fact that this turns out to be isomorphic to $\mathbb{C}$ may not be too surprising, considering that $\mathbb{C}$ is precisely a ring containing a copy of $\mathbb{R}$ and in which x^2+1 does have roots (that is, $\pm i$).

Such constructions are the "algebraist's way" to solve equations. We will come back to them in §V.5.2, and in a sense the whole of Chapter VII will be devoted to this topic.

4.3. Prime and maximal ideals. Other 'qualities' of ideals are best expressed in terms of quotients. We are still assuming that our rings are commutative—both due to unforgivable laziness and because that is the only context in which we will use these notions.

Definition 4.9. Let $I \neq (1)$ be an ideal of a commutative ring R.

- I is a *prime ideal* if R/I is an integral domain.
- I is a *maximal ideal* if R/I is a field. ⌟

Example 4.10. For all $a \in R$, the ideal $(x-a)$ is prime in $R[x]$ if and only if R is an integral domain; it is maximal if and only if R is a field. Indeed, $R[x]/(x-a) \cong R$, as we have seen in Example 4.7.

The ideal $(2, x)$ is maximal in $\mathbb{Z}[x]$, since

$$\frac{\mathbb{Z}[x]}{(2,x)} \overset{!}{\cong} \frac{\mathbb{Z}[x]/(x)}{(2)} \cong \frac{\mathbb{Z}}{(2)} = \mathbb{Z}/2\mathbb{Z}$$

is a field (for the isomorphism $\overset{!}{\cong}$, cf. Example 4.1). ⌟

Of course these notions may be translated into terms not involving quotients at all, and it is largely a matter of æsthetic preference whether prime and maximal ideals should be defined as in Definition 4.9 or by the following equivalent conditions:

Proposition 4.11. *Let $I \neq (1)$ be an ideal of a commutative ring R. Then*

- *I is prime if and only if for all $a, b \in R$*

$$ab \in I \implies (a \in I \text{ or } b \in I);$$

- *I is maximal if and only if for all ideals J of R*

$$I \subseteq J \implies (I = J \text{ or } J = R).$$

Proof. The ring R/I is an integral domain if and only if $\forall \overline{a}, \overline{b} \in R/I$

$$\overline{a} \cdot \overline{b} = 0 \implies (\overline{a} = 0 \text{ or } \overline{b} = 0).$$

This condition translates immediately to the given condition in R, with $\overline{a} = a + I$, $\overline{b} = b + I$, since the 0 in R/I is I.

As for maximality, the given condition follows from the correspondence between ideals of R/I and ideals of R containing I (§3.3) and the observation that a commutative ring is a field if and only if its only ideals are (0) and (1) (Exercise 3.8). □

From the formulation in terms of quotients, it is completely clear that

$$\text{maximal} \implies \text{prime};$$

indeed, fields are integral domains. This fact is of course easy to check in terms of the other description, but the argument is a little more cumbersome (Exercise 4.14). Prime ideals are not necessarily maximal, but note the following:

Proposition 4.12. *Let I be an ideal of a commutative ring R. If R/I is* finite, *then I is prime if and only if it is maximal.*

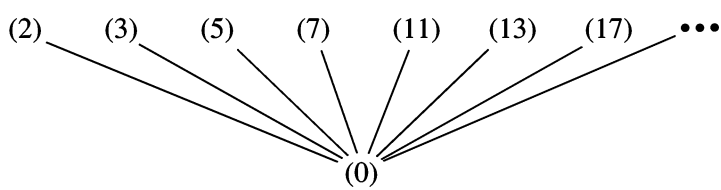
Proof. This follows immediately from Proposition 1.15. □

For example, let (n) be an ideal of $\mathbb{Z}$, with $n > 0$; then

$$(n) \text{ prime} \iff (n) \text{ maximal} \iff n \text{ is prime as an integer.}$$

Indeed, for nonzero n the ring $\mathbb{Z}/n\mathbb{Z}$ is finite, so Proposition 4.12 applies; cf. Example 1.17.

In general, the set of prime ideals of a commutative ring R is called[19] the *spectrum* of R, denoted $\operatorname{Spec} R$. We can 'draw' $\operatorname{Spec} \mathbb{Z}$ as follows:



Actually, attributing the fact that nonzero prime ideals of $\mathbb{Z}$ are maximal to the finiteness of the quotients (as we have just done) is slightly misleading; a better 'explanation' is that $\mathbb{Z}$ is a PID, and this phenomenon is common to all PIDs:

Proposition 4.13. *Let R be a PID, and let I be a nonzero ideal in R. Then I is prime if and only if it is maximal.*

Proof. Maximal ideals are prime in every ring, so we only need to verify that nonzero prime ideals are maximal in a PID; we will use the characterization of prime and maximal ideals obtained in Proposition 4.11. Let $I = (a)$ be a prime ideal in R, with $a \neq 0$, and assume $I \subseteq J$ for an ideal of R. As R is a PID, $J = (b)$

[19] Believe it or not, the term is borrowed from functional analysis.

for some $b \in R$. Since $I = (a) \subseteq (b) = J$, we have that $a = bc$ for some $c \in R$. But then $b \in (a)$ or $c \in (a)$, since $I = (a)$ is prime.

If $b \in (a)$, then $(b) \subseteq (a)$; and $I = J$ follows. If $c \in (a)$, then $c = da$ for some $d \in R$. But then

$$a = bc = bda,$$

from which $bd = 1$ since cancellation by the nonzero a holds in R (since R is an integral domain). This implies that b is a unit, and hence $J = (b) = R$.

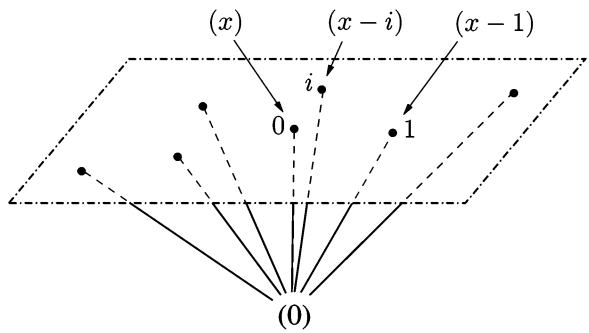
That is, we have shown that if $I \subseteq J$, then either $I = J$ or $J = R$: thus I is maximal, by Proposition 4.11. □

Example 4.14. Let k be a field. Then nonzero prime ideals in $k[x]$ are maximal, since $k[x]$ is a PID (as the reader has hopefully checked by now; cf. Exercise 4.4). Therefore, a picture of $\operatorname{Spec} k[x]$ would look pretty much like the picture of $\operatorname{Spec} \mathbb{Z}$ shown above: with the maximal ideals at one level, all containing the prime (and nonmaximal) ideal (0).

This picture is particularly appealing for fields such as $k = \mathbb{C}$, which are *algebraically closed,* that is, in which for every $f(x) \in k[x]$ there exists $r \in k$ such that $f(r) = 0$. It would take us too far afield to discuss this notion at any length now; but the reader should be aware that $\mathbb{C}$ is algebraically closed. We will come back to this[20]. Assuming this fact, it is easy to verify (Exercise 4.21) that the maximal ideals in $\mathbb{C}[x]$ are all and only the ideals

$$(x - z)$$

where z ranges over all complex numbers. That is, stretching our imagination a little, we could come up with the following picture for $\operatorname{Spec} \mathbb{C}[x]$:



There is a 'complex line[21] worth' of maximal ideals: for each $z \in \mathbb{C}$ we have the maximal $(x - z)$; the prime (0) is contained in all the maximal ideals; and there are no other prime ideals.

The picture for $\operatorname{Spec} \mathbb{C}[x]$ may serve as justification for the fact that, in algebraic geometry, $\mathbb{C}[x]$ is the ring corresponding to the 'affine line' $\mathbb{C}$; it is the ring of an algebraic *curve.* It turns out that the fact that there is exactly 'one level' of maximal

[20] A particularly pleasant proof may be given using elementary complex analysis, as a consequence of Liouville's theorem, or the maximum modulus principle; cf. §V.5.3.

[21] We know: it looks like a *plane.* But it is a *line,* as a complex entity.

ideals over (0) in $\mathbb{C}[x]$ reflects precisely the fact that the corresponding geometric object has dimension 1.

In general, the *(Krull) dimension* of a commutative ring R is the length of the longest chain of prime ideals in R. Thus, Proposition 4.13 tells us that PIDs, such as $\mathbb{Z}$, have 'dimension 1'. In the lingo of algebraic geometry, they all correspond to curves. ⌟

Example 4.15. For examples of rings of higher dimension, consider

$$k[x_1, \ldots, x_n],$$

where k is a field. Note that there are chains of prime ideals of length n

$$(0) \subsetneq (x_1) \subsetneq (x_1, x_2) \subsetneq \cdots \subsetneq (x_1, \ldots, x_n)$$

in this ring. (Why are these ideals prime? Cf. Exercise 4.13.) This says that $k[x_1, \ldots, x_n]$ has dimension $\geq n$. One can show that n is in fact the longest length of a chain of prime ideals in $k[x_1, \ldots, x_n]$; that is, the Krull dimension of $k[x_1, \ldots, x_n]$ is precisely n. In algebraic geometry, the ring $\mathbb{C}[x_1, \ldots, x_n]$ corresponds to the n-dimensional complex space $\mathbb{C}^n$.

Dealing precisely with these notions is not so easy, however. Even the seemingly simple statement that the maximal ideals of $\mathbb{C}[x_1, \ldots, x_n]$ are all and only the ideals $(x_1 - z_1, \ldots, x_n - z_n)$, for $(z_1, \ldots, z_n) \in \mathbb{C}^n$, requires a rather deep result, known as *Hilbert's Nullstellensatz.* ⌟

We will come back to all of this and get a very small taste of algebraic geometry in §VII.2, after we develop (much) more machinery.

Exercises

4.1. ▷ Let R be a ring, and let $\{I_\alpha\}_{\alpha \in A}$ be a family of ideals of R. We let

$$\sum_{\alpha \in A} I_\alpha := \left\{ \sum_{\alpha \in A} r_\alpha \text{ such that } r_\alpha \in I_\alpha \text{ and } r_\alpha = 0 \text{ for all but finitely many } \alpha \right\}.$$

Prove that $\sum_\alpha I_\alpha$ is an ideal of R and that it is the smallest ideal containing all of the ideals I_α. [§4.1]

4.2. ▷ Prove that the homomorphic image of a Noetherian ring is Noetherian. That is, prove that if $\varphi : R \to S$ is a surjective ring homomorphism and R is Noetherian, then S is Noetherian. [§6.4]

4.3. Prove that the ideal $(2, x)$ of $\mathbb{Z}[x]$ is not principal.

4.4. ▷ Prove that if k is a field, then $k[x]$ is a PID. (Hint: Let $I \subseteq k[x]$ be any ideal. If $I = (0)$, then I is principal. If $I \neq (0)$, let $f(x)$ be a monic polynomial in I of minimal degree. Use division with remainder to construct a proof that $I = (f(x))$, arguing as in the proof of Proposition II.6.9.) [§4.1, §4.3, §V.2.4, §V.4.1, §VI.7.2, §VII.1.2]

4.5. ▷ Let I, J be ideals in a ring R, such that $I+J=(1)$. Prove that $IJ=I\cap J$. [§4.1]

4.6. Let I, J be ideals in a ring R. Assume that $R/(IJ)$ is reduced (that is, it has no nonzero nilpotent elements; cf. Exercise 3.13). Prove that $IJ=I\cap J$.

4.7. ▷ Let $R=k$ be a field. Prove that every nonzero (principal) ideal in $k[x]$ is generated by a unique *monic* polynomial. [§4.2, §VI.7.2]

4.8. ▷ Let R be a ring and $f(x)\in R[x]$ a *monic* polynomial. Prove that $f(x)$ is not a (left- or right-) zero-divisor. [§4.2, 4.9]

4.9. Generalize the result of Exercise 4.8, as follows. Let R be a ring, and let $f(x)$ be a left-zero-divisor in $R[x]$. Prove that $\exists b\in R$, $b\neq 0$, such that $f(x)b=0$. (Hint: Let $f(x)=a_dx^d+\cdots+a_0$, and let $g(x)=b_ex^e+\cdots+b_0$ be a nonzero polynomial of minimal degree e such that $f(x)g(x)=0$. Deduce that $a_dg(x)=0$, and then prove $a_{d-i}g(x)=0$ for all i, by induction. What does this say about b_e?)

4.10. ¬ Let d be an integer that is not the square of an integer, and consider the subset of $\mathbb{C}$ defined by[22]

$$\mathbb{Q}(\sqrt{d}):=\{a+b\sqrt{d}\,|\,a,b\in\mathbb{Q}\}.$$

- Prove that $\mathbb{Q}(\sqrt{d})$ is a subring of $\mathbb{C}$.
- Define a function $N:\mathbb{Q}(\sqrt{d})\to\mathbb{Z}$ by $N(a+b\sqrt{d}):=a^2-b^2d$. Prove that $N(zw)=N(z)N(w)$ and that $N(z)\neq 0$ if $z\in\mathbb{Q}(\sqrt{d})$, $z\neq 0$.

The function N is a 'norm'; it is very useful in the study of $\mathbb{Q}(\sqrt{d})$ and of its subrings. (Cf. also Exercise 2.5.)

- Prove that $\mathbb{Q}(\sqrt{d})$ is a field and in fact the smallest subfield of $\mathbb{C}$ containing both $\mathbb{Q}$ and $\sqrt{d}$. (Use N.)
- Prove that $\mathbb{Q}(\sqrt{d})\cong\mathbb{Q}[t]/(t^2-d)$. (Cf. Example 4.8.)

[V.1.17, V.2.18, V.6.13, VII.1.12]

4.11. Let R be a commutative ring, $a\in R$, and $f_1(x),\dots,f_r(x)\in R[x]$.

- Prove the equality of ideals

$$(f_1(x),\dots,f_r(x),x-a)=(f_1(a),\dots,f_r(a),x-a).$$

- Prove the useful substitution trick

$$\frac{R[x]}{(f_1(x),\dots,f_r(x),x-a)}\cong\frac{R}{(f_1(a),\dots,f_r(a))}.$$

(Hint: Exercise 3.3.)

4.12. ▷ Let R be a commutative ring and $a_1,\cdots,a_n$ elements of R. Prove that

$$\frac{R[x_1,\dots,x_n]}{(x_1-a_1,\dots,x_n-a_n)}\cong R.$$

[§VII.2.2]

[22] Of course there are *two* 'square roots of d'; but the definition of $\mathbb{Q}(\sqrt{d})$ does not depend on which one is used.

4.13. ▷ Let R be an integral domain. For all $k = 1, \ldots, n$ prove that $(x_1, \ldots, x_k)$ is prime in $R[x_1, \ldots, x_n]$. [§4.3]

4.14. ▷ Prove 'by hand' that maximal ideals are prime, *without* using quotient rings. [§4.3]

4.15. Let $\varphi : R \to S$ be a homomorphism of commutative rings, and let $I \subseteq S$ be an ideal. Prove that if I is a prime ideal in S, then $\varphi^{-1}(I)$ is a prime ideal in R. Show that $\varphi^{-1}(I)$ is not necessarily maximal if I is maximal.

4.16. Let R be a commutative ring, and let P be a prime ideal of R. Suppose 0 is the only zero-divisor of R contained in P. Prove that R is an integral domain.

4.17. ¬ (If you know a little topology...) Let K be a compact topological space, and let R be the ring of continuous real-valued functions on K, with addition and multiplication defined pointwise.

(i) For $p \in K$, let $M_p = \{f \in R \mid f(p) = 0\}$. Prove that M_p is a maximal ideal in R.

(ii) Prove that if $f_1, \ldots, f_r \in R$ have no common zeros, then $(f_1, \ldots, f_r) = (1)$. (Hint: Consider $f_1^2 + \cdots + f_r^2$.)

(iii) Prove that every maximal ideal M in R is of the form M_p for some $p \in K$. (Hint: You will use the compactness of K and (ii).)

Conclude that $p \mapsto M_p$ defines a bijection from K to the set of maximal ideals of R. (The set of maximal ideals of a commutative ring R is called the *maximal spectrum* of R; it is contained in the (prime) spectrum $\operatorname{Spec} R$ defined in §4.3. Relating commutative rings and 'geometric' entities such as topological spaces is the business of *algebraic geometry*.)

The compactness hypothesis is necessary: cf. Exercise V.3.10. [V.3.10]

4.18. Let R be a commutative ring, and let N be its nilradical (Exercise 3.12). Prove that N is contained in every prime ideal of R. (Later on the reader will check that the nilradical is in fact the intersection of all prime ideals of R: Exercise V.3.13.)

4.19. Let R be a commutative ring, let P be a prime ideal in R, and let I_j be ideals of R.

(i) Assume that $I_1 \cdots I_r \subseteq P$; prove that $I_j \subseteq P$ for some j.

(ii) By (i), if $P \supseteq \bigcap_{j=1}^{r} I_j$, then P contains one of the ideals I_j. Prove or disprove: if $P \supseteq \bigcap_{j=1}^{\infty} I_j$, then P contains one of the ideals I_j.

4.20. Let M be a two-sided ideal in a (not necessarily commutative) ring R. Prove that M is maximal if and only if R/M is a simple ring (cf. Exercise 3.9).

4.21. ▷ Let k be an algebraically closed field, and let $I \subseteq k[x]$ be an ideal. Prove that I is maximal if and only if $I = (x - c)$ for some $c \in k$. [§4.3, §V.5.2, §VII.2.1, §VII.2.2]

4.22. Prove that $(x^2 + 1)$ is maximal in $\mathbb{R}[x]$.

4.23. A ring R has Krull dimension 0 if every prime ideal in R is maximal. Prove that fields and Boolean rings (Exercise 3.15) have Krull dimension 0.

4.24. Prove that the ring $\mathbb{Z}[x]$ has Krull dimension ≥ 2. (It is in fact exactly 2; thus it corresponds to a *surface* from the point of view of algebraic geometry.)

5. Modules over a ring

We have emphasized the parallels between the basic theory of groups and the basic theory of rings; there are also important differences. In Grp, one takes the quotient of a *group,* by a (normal sub)*group,* and the results is a *group:* throughout the process one never leaves Grp. The situation is in a sense even better in Ab, where the normality condition is automatic; one simply takes the quotient of an abelian group by an abelian (sub)group, obtaining an abelian group.

The situation in Ring is not nearly as neat. One of the three characters in the story is an *ideal:* which is not a ring according to the axioms listed in Definition 1.1, in all but the most pathological cases. In other words, the kernel of a homomorphism of rings is (usually) not a ring. Also, cokernels do not behave as one would hope (cf. Exercise 2.12). Even if one relaxes the ring definition, giving up the identity and making ideals subrings, there is no reasonably large class of examples in which the 'ideal' condition is automatically satisfied by substructures. In short, Ring is not a particularly pleasant category.

Modules will fix all these problems. If R is a ring and $I \subseteq R$ is a two-sided ideal, then all three structures R, I, and R/I are *modules over* R. The category of R-modules is the prime example of a well-behaved category: in this category kernels and cokernels exist and do precisely what they ought to. The category Ab is a particular case of this construction, since it is the category of modules over $\mathbb{Z}$ in disguise (Example 5.4). The category of modules over a ring R will share many of the excellent properties of Ab; and we will get a brand new, well-behaved category for each ring R. These are all examples[23] of the important notion of *abelian category.*

5.1. Definition of (left-)R-module. In short, R-modules are abelian groups endowed with an action of R. To flesh out this idea, recall that 'actions' in general denote homomorphisms into some kind of endomorphism structure: for example, we defined *group actions* in §II.9.1 as group homomorphisms from a fixed group to the groups of automorphisms of objects of a category.

We can give an analogous definition of the action of a *ring* on an *abelian group.* Indeed, recall that if M is an abelian group, then $\mathrm{End}_{\mathsf{Ab}}(M) := \mathrm{Hom}_{\mathsf{Ab}}(M, M)$ is a ring in a natural way (cf. §2.5). A *left-action* of a ring R on M is then simply a homomorphism of rings

$$\sigma : R \to \mathrm{End}_{\mathsf{Ab}}(M);$$

[23] In fact, it can be shown ('Freyd-Mitchell's embedding theorem') that every small abelian category is equivalent to a subcategory of the category of left-modules over a ring. So to some extent we can understand abelian categories by understanding categories of modules well enough. We will come back to this in Chapter IX.

we will say that σ makes M into a *left-R-module.*

Similarly to the situation with group actions, it is convenient to spell out this definition.

Claim 5.1. *The datum of a homomorphism σ as above is precisely the same as the datum of a function*

$$\rho : R \times M \to M$$

satisfying the following requirements: $(\forall r, s \in R)\ (\forall m, n \in M)$

- $\rho(r, m+n) = \rho(r, m) + \rho(r, n)$;
- $\rho(r+s, m) = \rho(r, m) + \rho(s, m)$;
- $\rho(rs, m) = \rho(r, \rho(s, m))$;
- $\rho(1, m) = m$.

The proof of this claim follows precisely the pattern of the discussion in the beginning of §II.9.2, so it is left to the reader (Exercise 5.2). Of course the relation between ρ and σ is given by

$$\rho(r, m) = \sigma(r)(m).$$

As always, carrying ρ around is inconvenient, so it is common to omit mention of it: $\rho(r, m)$ is often just denoted rm; this makes the requirements listed above a little more readable. Summarizing and adopting this shorthand,

Definition 5.2. A left-R-module structure on an abelian group M consists of a map $R \times M \to M$, $(r, m) \mapsto rm$, such that

- $r(m+n) = rm + rn$;
- $(r+s)m = rm + sm$;
- $(rs)m = r(sm)$;
- $1m = m$. ⌟

Right-R-modules are defined analogously. The reader can glance back at §II.9, especially Exercise II.9.3, for a reminder on right- vs. left-actions; the issues here are analogous. Thus, for example, a right-R-module structure may be identified with a left-R°-module structure, where R° is the 'opposite' ring obtained by reversing the order of multiplication (Exercise 5.1). However, note that R and R° have no good reasons to be isomorphic in general (while every group is isomorphic to its opposite).

These issues become immaterial if R is *commutative:* then the identity $R \to R^\circ$ is an isomorphism, and left-modules/right-modules are identical concepts. The reader will not miss much by adopting the blanket assumption that all rings mentioned in this section are commutative. It is occasionally important to make this hypothesis explicit (for example in dealing with *algebras,* cf. Example 5.6), but most of the material we are going to review works *verbatim* for, say, left-modules over an arbitrary ring as for modules over a commutative ring. We will write 'module' for 'left-module', for convenience; it will be the reader's responsibility to take care of appropriate changes, if necessary, to adapt the various concepts to *right*-modules.

5.2. The category R-Mod. The reader should spend some time getting familiar with the notion of module, by proving simple properties such as

$$(\forall m \in M): \quad 0 \cdot m = 0,$$
$$(\forall m \in M): \quad (-1) \cdot m = -m,$$

where M is a module over some ring (Exercise 5.3). The following fancier-sounding (but equally trivial) property is also useful in developing a feel for modules:

Proposition 5.3. *Every abelian group is a $\mathbb{Z}$-module, in exactly one way.*

Proof. Let G be an abelian group. A $\mathbb{Z}$-module structure on G is a ring homomorphism

$$\mathbb{Z} \to \mathrm{End}_{\mathsf{Ab}}(G).$$

Since $\mathbb{Z}$ is initial in Ring (§2.1), there exists exactly one such homomorphism, proving the statement. □

Thus, 'abelian group' and '$\mathbb{Z}$-module' are one and the same notion. Quite concretely, the action of $n \in \mathbb{Z}$ on an element a of an abelian group simply yields the ordinary 'multiple' na; this operation is trivially compatible with the operations in $\mathbb{Z}$.

A *homomorphism of R-modules* is a homomorphism of (abelian) groups which is compatible with the module structure. That is, if M, N are R-modules and $\varphi : M \to N$ is a function, then φ is a homomorphism of R-modules if and only if

- $(\forall m_1 \in M)(\forall m_2 \in M): \quad \varphi(m_1 + m_2) = \varphi(m_1) + \varphi(m_2)$;
- $(\forall r \in R)(\forall m \in M): \quad \varphi(rm) = r\varphi(m)$.

It is hopefully clear that the composition of two R-module homomorphisms is an R-module homomorphism and that the identity is an R-module homomorphism:

R-modules form a category

which we will denote[24] 'R-Mod'.

Example 5.4. The category $\mathbb{Z}$-Mod of $\mathbb{Z}$-modules is 'the same as' the category Ab: indeed, every abelian group is a $\mathbb{Z}$-module in exactly one way (Proposition 5.3), and $\mathbb{Z}$-module homomorphisms are simply homomorphisms of abelian groups. ⌟

Example 5.5. If $R = k$ is a field, R-modules are called *k-vector spaces*. We will call the category of vector spaces over a field k 'k-Vect'; this is just another name for k-Mod. Morphisms in k-Vect are often called[25] *linear maps*. 'Linear algebra' is the study of k-Vect (extended to R-Mod when possible); Chapters VI and VIII will be devoted to this subject. ⌟

[24] If R is not commutative, we should agree on whether R-Mod denotes the category of *left*-modules or of *right*-modules. We will mean 'left-modules'.

[25] This term is also used for homomorphisms of R modules for more general rings R, but not as frequently.

Example 5.6. Any homomorphism of rings $\alpha : R \to S$ may be used to define an interesting R-module: define $\rho : R \times S \to S$ by

$$\rho(r, s) := \alpha(r)s$$

for all $r \in R$ and $s \in S$. The operation on the right is simply multiplication in S, and the axioms of Definition 5.2 are immediate consequence of the ring axioms and of the fact that α is a homomorphism. For instance, taking $S = R$ and $\alpha = \mathrm{id}_R$ makes R a (left-) module over itself.

It is common to write rs rather than $\alpha(r)s$.

The ring operation in S and the R-module structure induced by α are linked even more tightly if we require R to be *commutative* and α to map R to the *center* of S: that is, if we require $\alpha(r)$, s to commute for every $r \in R$, $s \in S$. Indeed, in this case the left-module structure defined above and the right-module structure defined analogously would coincide; further, with these requirements the ring operation in S

$$(s_1, s_2) \mapsto s_1 s_2$$

is compatible with the R-module structure in the sense that[26]

$$(r_1 s_1)(r_2 s_2) = \alpha(r_1)s_1\alpha(r_2)s_2 = \alpha(r_1)\alpha(r_2)s_1 s_2 = (r_1 r_2)(s_1 s_2)$$

$\forall r_1, r_2 \in R$, $\forall s_1, s_2 \in S$: that is, we can 'move' the action of R at will through products in S.

Due to their importance, these examples deserve their own official name:

Definition 5.7. Let R be a commutative ring. An *R-algebra* is a ring homomorphism $\alpha : R \to S$ such that $\alpha(R)$ is contained in the center of S. ⌟

The usual abuse of language leads us to refer to an R-algebra by the name of the target S of the homomorphism. Thus, an R-algebra 'is' an R-module S with a compatible ring structure, or, if you prefer, a ring S with a compatible R-module structure. An R-algebra S is a *division* algebra if S is a division ring.

There is an evident notion of 'R-algebra homomorphism' (preserving both the ring and module structure), and we thus get a category R-Alg. The situation simplifies substantially if S is itself commutative, in which case the condition on the center is unnecessary. 'Commutative R-algebras' form a category, which the attentive reader will recognize as a coslice category (Example I.3.7) in the category of commutative rings.

Also, note that $\mathbb{Z}$-Alg is just another name for Ring (why?).

The polynomial rings $R[x_1, \ldots, x_n]$, as well as all their quotients, are commutative R-algebras. This is a particularly important class of examples; for $R = k$ an algebraically closed field, these are the rings used in 'classical' affine algebraic geometry (cf. §VII.2.3). ⌟

The trivial group 0 has a unique module structure over any ring R and is a zero-object in R-Mod, that is, it is both initial and final. As in the other main categories we have encountered, a bijective homomorphism of R-modules is automatically an

[26] More generally, this choice makes the multiplication in S *'R-bilinear'*.

isomorphism in R-Mod (Exercise 5.12). In these and many other respects, the category R-Mod (for any commutative ring R) and Ab are similar.

If R is commutative, the similarity goes further: just as in the category Ab, each set $\mathrm{Hom}_{R\text{-Mod}}(M,N)$ may itself be seen as an object of the category (cf. §II.4.4)[27]. Indeed, let M and N be R-modules. Since homomorphisms of R-modules are in particular homomorphisms of abelian groups,

$$\mathrm{Hom}_{R\text{-Mod}}(M,N) \subseteq \mathrm{Hom}_{\mathsf{Ab}}(M,N)$$

as sets (up to natural identifications). The operation making $\mathrm{Hom}_{\mathsf{Ab}}(M,N)$ into an abelian group, as in §II.4.4, clearly preserves $\mathrm{Hom}_{R\text{-Mod}}(M,N)$; it follows that the latter is an abelian group. For $r \in R$ and $\varphi \in \mathrm{Hom}_{R\text{-Mod}}(M,N)$, the prescription

$$(\forall m \in M): \quad (r\varphi)(m) := r\,\varphi(m)$$

defines a function[28] $r\varphi : M \to N$. This function is an R-module homomorphism if R is commutative, because $(\forall a \in R)$, $(\forall m \in M)$

$$(r\varphi)(am) = r\varphi(am) = (ra)\varphi(m) \overset{!}{=} (ar)\varphi(m) = a(r\varphi(m)).$$

Thus, we have a natural action of R on the abelian group $\mathrm{Hom}_{R\text{-Mod}}(M,N)$, and it is immediate to verify that this makes $\mathrm{Hom}_{R\text{-Mod}}(M,N)$ into an R-module.

Watch out: if R is not commutative, then in general $\mathrm{Hom}_{R\text{-Mod}}(M,N)$ is 'just' an abelian group. More structure is available if M, N are *bimodules;* cf. §VIII.3.2.

5.3. Submodules and quotients. Since R-modules are an 'enriched' version of abelian groups, we can progress through the usual constructions very quickly, by just pointing out that the analogous constructions in Ab are preserved by the R-module structure.

A *submodule* N of an R-module M is a subgroup preserved by the action of R. That is, for all $r \in R$ and $n \in N$, the element rn (defined by the R-module structure of M) is in fact in N. Put otherwise, and perhaps more transparently, N is itself an R-module, and the inclusion $N \subseteq M$ is an R-module homomorphism.

Example 5.8. We can view R itself as a (left-)R-module (cf. Example 5.6); the submodules of R are then precisely the (left-)*ideals* of R. ⌟

Example 5.9. Both the kernel and the image of a homomorphism $\varphi : M \to M'$ of R-modules are submodules (of M, M', respectively). ⌟

Example 5.10. If $r \in R$ and M is an R-module, $rM = \{rm \,|\, m \in M\}$ is a submodule of M. If I is any ideal of R, then $IM = \{rm \,|\, r \in I, m \in M\}$ is a submodule or R.

If N is a submodule of M, then it is in particular a (normal) subgroup of the abelian group $(M,+)$; thus we may define the quotient M/N as an abelian group.

[27] Notational convention: One often writes $\mathrm{Hom}_R(M,N)$ for $\mathrm{Hom}_{R\text{-Mod}}(M,N)$. We will not adopt this convention here but we will use it freely in later chapters.

[28] Parse the notation carefully: $r\varphi$ is the name of a function on the left, while $r\varphi(m)$ on the right is the action of the element $r \in R$ on the element $\varphi(m) \in N$, defined by the R-module structure on N.

Of course it would be desirable to see this as a module, and as usual there is only one reasonable way to do so: we will want the canonical projection

$$\pi : M \to M/N$$

to be an R-module homomorphism, and this forces

$$r(m+N) = r\pi(m) = \pi(rm) = rm + N$$

for all $m \in M$. That is, we are led to define the action of R on M/N by

$$r(m+N) := rm + N.$$

Claim 5.11. *For all submodules N, this prescription does define a structure of R-module on M/N.*

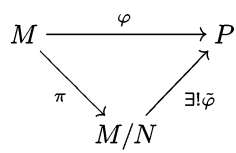
The proof of this claim is immediate and is left to the reader. The R-module M/N is (of course) called the *quotient of M by N*.

Example 5.12. If R is a ring and I is a two-sided ideal of R, then all three of I, R, and the quotient ring R/I are R-modules: I is a submodule of R, and the rings R and R/I are in fact R-algebras if R is commutative (cf. Example 5.6). ⌟

Example 5.13. If R is not commutative and I is just a (say) left-ideal, then the quotient R/I is not defined as a *ring,* but it is defined as a left-*module* (the quotient of the module R by the submodule I). The action of R on R/I is given by left-multiplication: $r(a+I) = ra + I$.

The reader should now expect a universal property for quotients, and here it is:

Theorem 5.14. *Let N be a submodule of an R-module M. Then for every homomorphism of R-modules $\varphi : M \to P$ such that $N \subseteq \ker\varphi$ there exists a unique homomorphism of R-modules $\tilde{\varphi} : M/N \to P$ so that the diagram*



commutes.

As in previous appearances of such statements, this is an immediate consequence of the set-theoretic version (§I.5.3) and of easy notation matching and compatibility checks. For an even faster proof, one can just apply Theorem II.7.12 and verify that $\tilde{\varphi}$ is an R-module homomorphism.

Since every submodule N is then the kernel of the canonical projection $M \to M/N$, our recurring slogan becomes, in the context of R-Mod

$$\text{kernel} \iff \text{submodule}:$$

unlike as in Grp or Ring, being a kernel poses no restriction on the relevant substructures. Put otherwise, 'every monomorphism in R-Mod is a kernel'; this is one of the distinguishing features of an abelian category.

5.4. Canonical decomposition and isomorphism theorems. The discussion now proceeds along the same lines as for (abelian) groups; the statements of the key facts and a few comments should suffice, as the proofs are nothing but a rehashing of the proofs of analogous statements we have encountered previously. *Of course* the reader should take the following statements as assignments and provide all the needed details.

In the context of R-modules, the canonical decomposition takes the following form:

Theorem 5.15. *Every R-module homomorphism $\varphi : M \to M'$ may be decomposed as follows:*

$$M \twoheadrightarrow M/\ker\varphi \xrightarrow[\tilde{\varphi}]{\sim} \operatorname{im}\varphi \hookrightarrow M'$$

where the isomorphism $\tilde{\varphi}$ in the middle is the homomorphism induced by φ (as in Theorem 5.14).

The 'first isomorphism theorem' is the following consequence:

Corollary 5.16. *Suppose $\varphi : M \to M'$ is a* surjective *R-module homomorphism. Then*

$$M' \cong \frac{M}{\ker\varphi}.$$

If M is an R-module and N is a submodule of M, then there is a bijection (cf. §II.8.3)

$$u : \{\text{submodules } P \text{ of } M \text{ containing } N\} \to \{\text{submodules of } M/N\}$$

preserving inclusions, and the 'third isomorphism theorem' holds:

Proposition 5.17. *Let N be a submodule of an R-module M, and let P be a submodule of R containing N. Then P/N is a submodule of M/N, and*

$$\frac{M/N}{P/N} \cong \frac{M}{P}.$$

We also have a version of the 'second isomorphism theorem' (cf. Proposition II.8.11), with simplifications due to the fact that normality is not an issue in the theory of modules:

Proposition 5.18. *Let N, P be submodules of an R-module M. Then*

- *$N + P$ is a submodule of M;*
- *$N \cap P$ is a submodule of P, and*

$$\frac{N+P}{N} \cong \frac{P}{N \cap P}.$$

More generally, it is hopefully clear that the *sum* $\sum_\alpha N_\alpha$ and *intersection* $\bigcap_\alpha N_\alpha$ of any family $\{N_\alpha\}_\alpha$ of submodules of an R-module M (which are defined as *subgroups* of the abelian group M; cf. for example Lemma II.6.3) are submodules of M.

Exercises

5.1. ▷ Let R be a ring. The *opposite* ring R° is obtained from R by reversing the multiplication: that is, the product $a \bullet b$ in R° is defined to be $ba \in R$. Prove that the identity map $R \to R^\circ$ is an isomorphism if and only if R is commutative. Prove that $\mathcal{M}_n(\mathbb{R})$ is isomorphic to its opposite (*not* via the identity map!). Explain how to turn right-R-modules into left-R-modules and conversely, if $R \cong R^\circ$. [§5.1, VIII.5.19]

5.2. ▷ Prove Claim 5.1. [§5.1]

5.3. ▷ Let M be a module over a ring R. Prove that $0 \cdot m = 0$ and that $(-1) \cdot m = -m$, for all $m \in M$. [§5.2]

5.4. ¬ Let R be a ring. A nonzero R-module M is *simple* (or *irreducible*) if its only submodules are $\{0\}$ and M. Let M, N be simple modules, and let $\varphi : M \to N$ be a homomorphism of R-modules. Prove that either $\varphi = 0$ or φ is an isomorphism. (This rather innocent statement is known as *Schur's lemma.*) [5.10, 6.16, VI.1.16]

5.5. Let R be a ring, viewed as an R-module over itself, and let M be an R-module. Prove that $\mathrm{Hom}_{R\text{-}\mathsf{Mod}}(R, M) \cong M$ as R-modules.

5.6. Let G be an abelian group. Prove that if G has a structure of $\mathbb{Q}$-vector space, then it has only one such structure. (Hint: First prove that every element of G has necessarily infinite order.)

5.7. Let K be a field, and let $k \subseteq K$ be a subfield of K. Show that K is a vector space over k (and in fact a k-algebra) in a natural way. In this situation, we say that K is an *extension* of k.

5.8. What is the initial object of the category R-Alg?

5.9. ¬ Let R be a commutative ring, and let M be an R-module. Prove that the operation of composition on the R-module $\mathrm{End}_{R\text{-}\mathsf{Mod}}(M)$ makes the latter an R-algebra in a natural way.

Prove that $\mathcal{M}_n(R)$ (cf. Exercise 1.4) is an R-algebra, in a natural way. [VI.1.12, VI.2.3]

5.10. Let R be a commutative ring, and let M be a simple R-module (cf. Exercise 5.4). Prove that $\mathrm{End}_{R\text{-}\mathsf{Mod}}(M)$ is a division R-algebra.

5.11. ▷ Let R be a commutative ring, and let M be an R-module. Prove that there is a bijection between the set of $R[x]$-module structures on M and $\mathrm{End}_{R\text{-}\mathsf{Mod}}(M)$. [§VI.7.1]

5.12. ▷ Let R be a ring. Let M, N be R-modules, and let $\varphi : M \to N$ be a homomorphism of R-modules. Assume φ is a bijection, so that it has an inverse φ^{-1} as a set-function. Prove that φ^{-1} is a homomorphism of R-modules. Conclude that a bijective R-module homomorphism is an isomorphism of R-modules. [§5.2, §VI.2.1, §IX.1.3]

5.13. Let R be an integral domain, and let I be a nonzero *principal* ideal of R. Prove that I is isomorphic to R as an R-module.

5.14. ▷ Prove Proposition 5.18. [§5.4]

5.15. Let R be a commutative ring, and let I, J be ideals of R. Prove that $I \cdot (R/J) \cong (I+J)/J$ as R-modules.

5.16. ¬ Let R be a commutative ring, M an R-module, and let $a \in R$ be a nilpotent element, determining a submodule aM of M. Prove that $M = 0 \iff aM = M$. (This is a particular case of *Nakayama's lemma,* Exercise VI.3.8.) [VI.3.8]

5.17. ▷ Let R be a commutative ring, and let I be an ideal of R. Noting that $I^j \dot{I}^k \subseteq I^{j+k}$, define a ring structure on the direct sum

$$\mathrm{Rees}_R(I) := \bigoplus_{j \geq 0} I^j = R \oplus I \oplus I^2 \oplus I^3 \oplus \cdots .$$

The homomorphism sending R identically to the first term in this direct sum makes $\mathrm{Rees}_R(I)$ into an R-algebra, called the *Rees algebra* of I. Prove that if $a \in R$ is a non-zero-divisor, then the Rees algebra of (a) is isomorphic to the polynomial ring $R[x]$ (as an R-algebra). [5.18]

5.18. With notation as in Exercise 5.17 let $a \in R$ be a non-zero-divisor, let I be any ideal of R, and let J be the ideal aI. Prove that $\mathrm{Rees}_R(J) \cong \mathrm{Rees}_R(I)$.

---

6. Products, coproducts, etc., in R-Mod

We have stated several times that categories such as R-Mod are 'well-behaved'. We will explore in this section the sense in which this can be formalized at this stage. The bottom line is that these categories enjoy the same nice properties that we have noted along the way for the category Ab.

We will also include here some general considerations on finitely generated modules and algebras.

As in the previous section, we will write 'module' for 'left-module'; the reader should make appropriate adaptations to the case of right-modules. Little will be lost by assuming that all rings appearing here are commutative (thereby removing the distinction between left- and right-modules).

6.1. Products and coproducts. As in Ab, products and coproducts exist, and *finite* products and coproducts coincide, in R-Mod. Indeed, recall the construction of the *direct sum* of two abelian groups (§II.3.5): if M and N are abelian groups, then $M \oplus N$ denotes their product, with componentwise operation. If M and N are R-modules, we can give an R-module structure to $M \oplus N$ by prescribing $\forall r \in R$

$$r(m,n) := (rm, rn).$$

This defines the *direct sum* of M, N, as an R-module. Note that $M \oplus N$ comes together with several homomorphisms of R-modules:

$$\pi_M : M \oplus N \to M, \quad \pi_N : M \oplus N \to N$$

sending (m,n) to m, n, respectively, and

$$i_M : M \to M \oplus N, \quad i_N : N \to M \oplus N$$

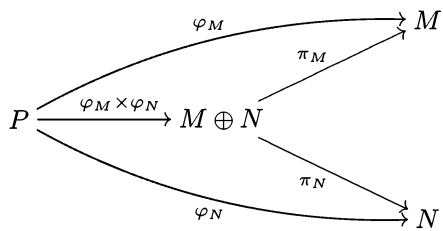
sending m to $(m,0)$ and n to $(0,n)$.

Proposition 6.1. *The direct sum $M \oplus N$ satisfies the universal properties of both the product and the coproduct of M and N.*

Proof. *Product:* Let P be an R-module, and let $\varphi_M : P \to M$, $\varphi_N : P \to N$ be two R-module homomorphisms. The definition of an R-module homomorphism

$$\varphi_M \times \varphi_N : P \to M \oplus N$$

is forced by the needed commutativity of the diagram



That is,

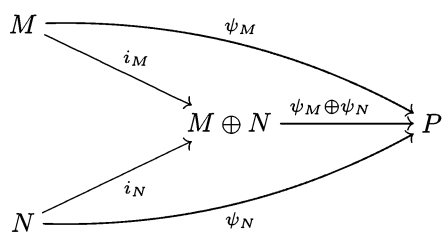
$$(\forall p \in P): \quad (\varphi_M \times \varphi_N)(p) := (\varphi_M(p), \varphi_N(p)).$$

This *is* an R-module homomorphism, and it is the unique one making the diagram commute; therefore $M \oplus N$ works as a product of M and N.

Coproduct: View the preceding argument through a mirror! Let P be an R-module, and let $\psi_M : M \to P$, $\psi_N : N \to P$ be two R-module homomorphisms. The definition of an R-module homomorphism

$$\psi_M \oplus \psi_N : M \oplus N \to P$$

is forced by the needed commutativity of the diagram



That is[29], $(\forall m \in M)(\forall n \in N)$

$$\begin{aligned}(\psi_M \oplus \psi_N)(m,n) &= (\psi_M \oplus \psi_N)(m,0) + (\psi_M \oplus \psi_N)(0,n) \\ &= (\psi_M \oplus \psi_N) \circ i_M(m) + (\psi_M \oplus \psi_N) \circ i_N(n) \\ &= \psi_M(m) + \psi_N(n).\end{aligned}$$

This *is* an R-module homomorphism: it is a homomorphism of abelian groups by virtue of the commutativity of addition in P, and it clearly preserves the action of R.

[29]This should look familiar; cf. Exercise II.3.3.

Since it is the unique R-module homomorphism making the diagram commute, this verifies that $M \oplus N$ works as a *co*product of M and N. □

It may seem like a good idea to write $M \times N$ rather than $M \oplus N$ when viewing the latter as the *product* of M and N; but in due time (§VIII.2) we will encounter $M \times N$ again, in the context of 'bilinear maps', and in that context $M \times N$ is *not* viewed as an R-module.

The reader should also work out the *fibered* versions of these constructions; cf. Exercises 6.10 and 6.11.

The fact that finite products and coproducts agree in R-Mod does not extend to the infinite case (Exercise 6.7).

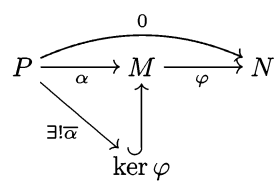
6.2. Kernels and cokernels. The facts that R-Mod has a zero-object (the 0-module), its Hom sets are abelian groups (§5.2), and it has (finite) products and coproducts make R-Mod an *additive* category. The fact that R-Mod has well-behaved kernels and cokernels, which we review next, upgrades it to the status of *abelian category*. (We will come back to these general definitions in §IX.1.)

In general, monomorphisms and epimorphisms do not automatically satisfy good properties, even when objects of a given category are realized by adding structure to sets. For example, we have seen that the precise relationship between 'surjective morphism' and 'epimorphism' may be rather subtle: epimorphisms are surjective in Grp, but for complicated reasons (§II.8.6); and there are epimorphisms that are not surjective in Ring (§2.3).

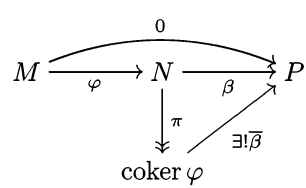
The situation in R-Mod is as simple as it can be. Recall that we have identified universal properties for kernels and cokernels (cf. §II.8.6); in the category R-Mod these would go as follows: if

$$\varphi : M \to N$$

is a homomorphism of R-modules, then $\ker \varphi$ is final with respect to the property of factoring R-module homomorphisms $\alpha : P \to M$ such that $\varphi \circ \alpha = 0$:



while $\operatorname{coker} \varphi$ is initial with respect to the property of factoring R-module homomorphisms $\beta : N \to P$ such that $\beta \circ \varphi = 0$:



Proposition 6.2. *The following hold in* R-Mod*:*

- *kernels and cokernels exist;*

- *φ is a monomorphism* $\iff$ $\ker\varphi$ *is trivial* $\iff$ *φ is injective as a set-function;*
- *φ is an epimorphism* $\iff$ $\operatorname{coker}\varphi$ *is trivial* $\iff$ *φ is surjective as a set-function.*

Further, every monomorphism identifies its source with the kernel of some morphism, and every epimorphism identifies its target with the cokernel of some morphism.

This proposition of course simply generalizes to R-Mod facts we know already from our study of Ab, and a quick review should suffice for the careful reader. Kernels exist: indeed, the 'standard' definition of kernel satisfies the universal properties spelled out above (same argument as in Proposition II.6.6). Cokernels exist: indeed, let

$$\operatorname{coker}\varphi = \frac{N}{\operatorname{im}\varphi};$$

if $\beta : N \to P$ is such that $\beta \circ \varphi = 0$, then $\operatorname{im}\varphi \subseteq \ker\beta$; so β must factor uniquely through $N/\operatorname{im}\varphi$ by the universal property of quotients, Theorem 5.14. That is, $N/\operatorname{im}\varphi$ does satisfy the universal property for cokernels.

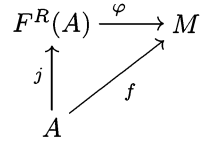
The proofs of all the implications in the second and third points in Proposition 6.2 follow familiar patterns from (for example) Proposition II.6.12 and Proposition II.8.18. The last sentence of Proposition 6.2 simply reiterates the slogan *submodule* $\iff$ *kernel* and its mirror statement (which is just as true). Further details are left to the reader.

By the way, whatever happened to the conditions characterizing monomorphisms and epimorphisms in Set (Proposition I.2.1)? In Set, a function is a monomorphism if and only if it has a left-inverse, and it is an epimorphism if and only if it has a right-inverse. We have learned not to expect any of this to happen in more general categories. Modules are not an exception: the function $\mathbb{Z} \to \mathbb{Z}$ defined by 'multiplication by 2' is a monomorphism without a left-inverse, and the projection $\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$ is an epimorphism without a right-inverse. We will come back to this point in §7.2.

6.3. Free modules and free algebras. The universal property of *free R-modules* is modeled after the properties defining the other free objects we have encountered: the goal is to define the R-module containing a given set A 'in the most efficient way'. Again, the situation will match the case of abelian groups closely, so the reader may want to refer back to §II.5.4.

The universal property formalizing the heuristic requirement goes as follows: given a set A, we seek an R-module $F^R(A)$, called a *free R-module* on the set A, together with a set-function $j : A \to F^R(A)$, such that for all R-modules M and set-functions $f : A \to M$ *there exists a unique* R-module homomorphism

$\varphi : F^R(A) \to M$ such that the diagram



commutes. Abstract nonsense guarantees that such an R-module is unique up to isomorphism if it exists at all (Proposition I.5.4) and that the function $j : A \to F^R(A)$ is necessarily injective (cf. Exercise II.5.3). The question is, does it exist?

The answer will not appear to be very exciting, since it generalizes directly the case of abelian groups (that is, $\mathbb{Z}$-modules). Given any set A, define the (possibly infinite) direct sum $N^{\oplus A}$ of an R-module N as follows:

$$N^{\oplus A} := \{\alpha : A \to N \mid \alpha(a) \neq 0 \text{ for only finitely many elements } a \in A\}.$$

Of course this agrees with the definition given for abelian groups in §II.5.4; $N^{\oplus A}$ has an evident R-module structure, obtained by defining, for all $r \in R$ and $a \in A$,

$$(r\alpha)(a) := r(\alpha(a)).$$

For $N = R$ we may define a function $j : A \to R^{\oplus A}$ by sending $a \in A$ to the function $j_a : A \to R$:

$$(\forall x \in A): \quad j_a(x) := \begin{cases} 1 & \text{if } x = a, \\ 0 & \text{if } x \neq a. \end{cases}$$

Claim 6.3. $F^R(A) \cong R^{\oplus A}$.

The proof of this claim matches precisely the proof of Proposition II.5.6 and is left to the reader (Exercise 6.1). The key is that every element of $R^{\oplus A}$ may be written uniquely as a *finite* sum

$$\sum_{a \in A} r_a\, a$$

(shorthand for $\sum_{a \in A} r_a j(a)$); incidentally, this is how elements of 'the free R-module on A' are often written—this is legal, by virtue of Claim 6.3.

In particular, for $A = \{1, \ldots, n\}$ a *finite* set, Claim 6.3 states that the R-module $R^{\oplus n}$, with $j : A \to R^{\oplus n}$ defined by

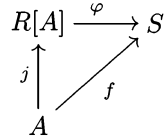
$$j(i) := (0, \cdots, 0, \underset{i\text{-th place}}{1}, 0, \cdots, 0) \in R^{\oplus n},$$

satisfies the universal property for $F^R(\{1, \ldots, n\})$.

This is all entirely analogous to the story for $\mathbb{Z}$-modules. The situation becomes a little more interesting if we switch from R-modules to *commutative R-algebras;* the category is different, so we should expect a different answer. The finite case is essentially the only one we will need in these notes, so assume $A = \{\mathbf{1}, \ldots, \mathbf{n}\}$ is a finite set. In this case, we write $R[A]$ for the polynomial ring $R[x_1, \ldots, x_n]$; we have a set-function $j : A \to R[A]$, defined by $j(\mathbf{i}) = x_i$.

Proposition 6.4. *$R[A]$ is a free commutative R-algebra on the set A.*

Proof. The statement translates into the following: for every commutative R-algebra S and every set-function $f : A \to S$, *there exists a unique* R-algebra homomorphism $\varphi : R[A] \to S$ such that the diagram



commutes. Since S is an R-algebra, we have a fixed homomorphism of rings $\alpha : R \to S$ (cf. Example 5.6). Then we may construct $\varphi : R[A] = R[x_1, \dots, x_n] \to S$ by applying n times the 'extension property' of Example 2.3: extend α to $R[x_1]$ so as to map x_1 to $f(\mathbf{1})$, then to $R[x_1, x_2] = R[x_1][x_2]$ so as to map x_2 to $f(\mathbf{2})$, etc. Note that each extension is *uniquely* determined by its requirements.

This gives φ as a ring homomorphism and shows that it is unique. The reader will verify that φ is also (automatically) an R-module homomorphism, and hence a homomorphism of R-algebras, concluding the proof. □

After the fact, the reader may want to revisit §2.2 and recognize that the 'universal property of polynomial rings $\mathbb{Z}[x_1, \dots, x_n]$' given there was really a version of their role as free objects in the category of commutative rings, a.k.a. commutative $\mathbb{Z}$-algebras.

It is not difficult to identify free objects in the larger category R-Alg: they consist of 'noncommutative polynomial rings' $R\langle A\rangle$, with variables from the set A but without any condition relating ab and ba for $a \neq b$ in A. More precisely, $R\langle A\rangle$ is isomorphic to the *monoid ring* (cf. §1.4) over the free *monoid* on A, consisting of all finite strings of elements in A, with operation defined by concatenation[30]. We will encounter this ring again, but only in the distant future (Example VIII.4.17).

6.4. Submodule generated by a subset; Noetherian modules. Let M be an R-module, and let $A \subseteq M$ be a subset of M. By the universal property of free modules, there is a unique homomorphism of R-modules

$$\varphi_A : R^{\oplus A} \to M.$$

The image of this homomorphism is a submodule of M, the *submodule generated by* A in M, usually denoted $\langle A\rangle$ (or $\langle a_1, \dots, a_n\rangle$ if $A = \{a_1, \dots, a_n\}$ is finite). Thus,

$$\langle A\rangle = \{\sum_{a\in A} r_a\, a \,|\, r_a \neq 0 \text{ for only finitely many elements } a \in A\}.$$

It is hopefully clear that $\langle A\rangle$ is the smallest submodule of M containing A.

The module M is *finitely generated* if $M = \langle A\rangle$ for a *finite* set A, that is, if and only if there is a surjective homomorphism of R-modules

$$R^{\oplus n} \twoheadrightarrow M$$

for some n. One of the highlights of Chapter VI will be the classification of *finitely generated modules over PIDs* (Theorem VI.5.6). We have already briefly mentioned

[30]This construction is similar to the free *group* on A, but without the complication of the presence of inverses and of possible cancellations.

the case for $\mathbb{Z}$ (recall that $\mathbb{Z}$ is a PID and $\mathbb{Z}$-modules are nothing but abelian groups!) back in §II.6.3.

Finitely generated R-modules are tremendously important, but they are not as well-behaved as one might hope at first. For example, it may be that a module M is finitely generated, but some *submodule* of M is *not* finitely generated!

Example 6.5. Let $R = \mathbb{Z}[x_1, x_2, \dots]$, a polynomial ring on infinitely many indeterminates. Then R is finitely generated as an R-module: indeed, 1 generates it. However, the ideal

$$(x_1, x_2, \dots)$$

of R generated by all indeterminates is *not* finitely generated as an R-module (Exercise 6.14). ⌟

Definition 6.6. An R-module M is *Noetherian* if every submodule of M is finitely generated as an R-module. ⌟

Thus, a ring R is Noetherian in the sense of Definition 4.2 if and only if it is Noetherian 'as a module over itself'. The ring in Example 6.5 is not Noetherian.

We will study the Noetherian condition more carefully later on (§V.1.1); but we can already see one reason why this is a good, 'solid' notion.

Proposition 6.7. *Let M be an R-module, and let N be a submodule of M. Then M is Noetherian if and only if both N and M/N are Noetherian.*

Proof. If M is Noetherian, then so is M/N (same proof as for Exercise 4.2), and so is N (because every submodule of N is a submodule of M, so it is finitely generated because M is Noetherian). This proves the 'only if' part of the statement.

For the converse, assume N and M/N are Noetherian, and let P be a submodule of M; we have to prove that P is finitely generated. Since $P \cap N$ is a submodule of N and N is Noetherian, $P \cap N$ is finitely generated. By the 'second isomorphism theorem', Proposition 5.18,

$$\frac{P}{P \cap N} \cong \frac{P+N}{N},$$

and hence $P/(P \cap N)$ is isomorphic to a submodule of M/N. Since M/N is Noetherian, this shows that $P/(P \cap N)$ is finitely generated.

It follows that P itself is finitely generated, by Exercise 6.18. □

Corollary 6.8. *Let R be a Noetherian ring, and let M be a finitely generated R-module. Then M is Noetherian (as an R-module).*

Proof. Indeed, by hypothesis there is an onto homomorphism $R^{\oplus n} \twoheadrightarrow M$ of R-modules; hence (by the first isomorphism theorem, Corollary 5.16) M is isomorphic to a quotient of $R^{\oplus n}$. By Proposition 6.7, it suffices to prove that $R^{\oplus n}$ is Noetherian.

This may be done by induction. The statement is true for $n = 1$ by hypothesis. For $n > 1$, assume we know that $R^{\oplus(n-1)}$ is Noetherian; since $R^{\oplus(n-1)}$ may be viewed as a submodule of $R^{\oplus n}$, in such a way that

$$\frac{R^{\oplus n}}{R^{\oplus(n-1)}} \cong R$$

(Exercise 6.4), and R is Noetherian, it follows that $R^{\oplus n}$ is Noetherian, again by applying Proposition 6.7. □

6.5. Finitely generated vs. finite type. If S is an R-algebra, it may be 'finitely generated' in two very different ways: as an R-module and as an R-algebra. It is important to keep these two concepts well distinct, although unfortunately the language used to express them is very similar.

The following definitions differ in three small details...

"S is finitely generated as a module over R if there is an onto homomorphism of R-modules from the free R-module on a finite set to S."

"S is finitely generated as an algebra over R if there is an onto homomorphism of R-algebras from the free R-algebra on a finite set to S."

The mathematical difference is more substantial than it may appear. As we have seen in §6.3, the free *R-module* over a finite set $A = \{1, \dots, n\}$ is isomorphic to $R^{\oplus n}$; the free commutative *R-algebra* over A is isomorphic to $R[x_1, \dots, x_n]$. Thus, a commutative[31] ring S is finitely generated *as an R-module* if there is an onto homomorphism of R-modules

$$R^{\oplus n} \twoheadrightarrow S$$

for some n; it is finitely generated *as an R-algebra* if there is an onto homomorphism of R-algebras

$$R[x_1, \dots, x_n] \twoheadrightarrow S$$

for some n. In other words, S is finitely generated as an R module if and only if $S \cong R^{\oplus n}/M$ for some n and a submodule M of $R^{\oplus n}$; it is a finite-type R-algebra if and only if $S \cong R[x_1, \dots, x_n]/I$ for some n and an ideal I of $R[x_1, \dots, x_n]$.

We say that S is *finite* in the first case[32] and *of finite type* in the second. It is clear that 'finite' $\implies$ 'finite type'; it should be just as clear that the converse does not hold.

Example 6.9. The polynomial ring $R[x]$ is a finite-type R-algebra, but it is not finite as an R-module. ⌟

The distinction, while macroscopic in general, may evaporate in special, important cases. For example, one can prove that if k and K are fields and $k \subseteq K$, then K is of finite type over k if *and only if* it is in fact finite as a k-module (that is, it is a *finite-dimensional* k-vector space). This is one version of *Hilbert's Nullstellensatz,* a deep result we already mentioned in Example 4.15 and that we will prove (in an important class of examples) in §VII.2.2.

David Hilbert's name is associated to another important result concerning finite-type R-algebras: if R is Noetherian (as a ring, that is, as an R-module)

[31] We are mostly interested in the commutative case, so we will make this hypothesis here; the only change in the general case is typographical: $\langle \cdots \rangle$ rather than $[\cdots]$. Also, note that a commutative ring is finitely generated as an algebra if and only if it is finitely generated as a *commutative* algebra; cf. Exercise 6.15.

[32] This is particularly unfortunate, since S may very well be an *infinite* set.

and S is a finite-type R-algebra, then S is also Noetherian (as a ring, that is, as an S-module). This is an immediate consequence of the so-called *Hilbert's basis theorem.*

The proof of Hilbert's basis theorem is completely elementary: it could be given here as an exercise, with a few key hints; we will see it in §V.1.1.

Exercises

6.1. ▷ Prove Claim 6.3. [§6.3]

6.2. Prove or disprove that if R is ring and M is a nonzero R-module, then M is not isomorphic to $M \oplus M$.

6.3. Let R be a ring, M an R-module, and $p : M \to M$ an R-module homomorphism such that $p^2 = p$. (Such a map is called a *projection.*) Prove that $M \cong \ker p \oplus \operatorname{im} p$.

6.4. ▷ Let R be a ring, and let $n > 1$. View $R^{\oplus(n-1)}$ as a submodule of $R^{\oplus n}$, via the injective homomorphism $R^{\oplus(n-1)} \hookrightarrow R^{\oplus n}$ defined by

$$(r_1, \ldots, r_{n-1}) \mapsto (r_1, \ldots, r_{n-1}, 0).$$

Give a one-line proof that

$$\frac{R^{\oplus n}}{R^{\oplus(n-1)}} \cong R.$$

[§6.4]

6.5. ▷ (Notation as in §6.3.) For any ring R and any two sets A_1, A_2, prove that $(R^{\oplus A_1})^{\oplus A_2} \cong R^{\oplus(A_1 \times A_2)}$. [§VIII.2.2]

6.6. ¬ Let R be a ring, and let $F = R^{\oplus n}$ be a finitely generated free R-module. Prove that $\mathrm{Hom}_{R\text{-}\mathsf{Mod}}(F, R) \cong F$. On the other hand, find an example of a ring R and a nonzero R-module M such that $\mathrm{Hom}_{R\text{-}\mathsf{Mod}}(M, R) = 0$. [6.8]

6.7. ▷ Let A be any set.

- For any family $\{M_a\}_{a \in A}$ of modules over a ring R, define the *product* $\prod_{a \in A} M_a$ and coproduct $\bigoplus_{a \in A} M_a$. If $M_a \cong R$ for all $a \in A$, these are denoted R^A, $R^{\oplus A}$, respectively.
- Prove that $\mathbb{Z}^{\mathbb{N}} \ncong \mathbb{Z}^{\oplus \mathbb{N}}$. (Hint: Cardinality.)

[§6.1, 6.8]

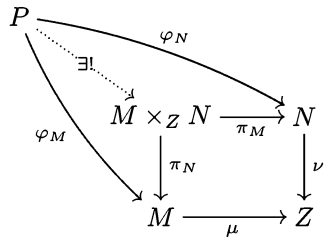
6.8. Let R be a ring. If A is any set, prove that $\mathrm{Hom}_{R\text{-}\mathsf{Mod}}(R^{\oplus A}, R)$ satisfies the universal property for the *product* of the family $\{R_a\}_{a \in A}$, where $R_a \cong R$ for all a; thus, $\mathrm{Hom}_{R\text{-}\mathsf{Mod}}(R^{\oplus A}, R) \cong R^A$. Conclude that $\mathrm{Hom}_{R\text{-}\mathsf{Mod}}(R^{\oplus A}, R)$ is *not* isomorphic to $R^{\oplus A}$ in general (cf. Exercises 6.6 and 6.7.)

6.9. ¬ Let R be a ring, F a nonzero free R-module, and let $\varphi : M \to N$ be a homomorphism of R-modules. Prove that φ is onto if and only if for all R-module homomorphisms $\alpha : F \to N$ there exists an R-module homomorphism $\beta : F \to M$

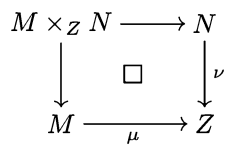
such that $\alpha = \beta \circ \varphi$. (Free modules are *projective,* as we will see in Chapter VIII.) [7.8, VI.5.5]

6.10. ▷ (Cf. Exercise I.5.12.) Let M, N, and Z be R-modules, and let $\mu : M \to Z$, $\nu : N \to Z$ be homomorphisms of R-modules.

Prove that R-Mod has 'fibered products': there exists an R-module $M \times_Z N$ with R-module homomorhisms $\pi_M : M \times_Z N \to M$, $\pi_N : M \times_Z N \to N$, such that $\mu \circ \pi_M = \nu \circ \pi_N$, and which is universal with respect to this requirement. That is, for every R-module P and R-module homomorphisms $\varphi_M : P \to M$, $\varphi_N : P \to N$ such that $\mu \circ \varphi_M = \nu \circ \varphi_N$, there exists a unique R-module homomorphism $P \to M \times_Z N$ making the diagram

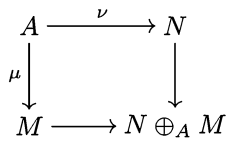


commute. The module $M \times_Z N$ may be called the *pull-back* of M along ν (or of N along μ, since the construction is symmetric). 'Fiber diagrams'



are commutative, but 'even better' than commutative; they are often decorated by a square, as shown here. [§6.1, 6.11, §IX.1.4]

6.11. ▷ Define a notion of *fibered coproduct* of two R-modules M, N, along an R-module A, in the style of Exercise 6.10 (and cf. Exercise I.5.12)



Prove that fibered coproducts exist in R-Mod. The fibered coproduct $M \oplus_A N$ is called the *push-out* of M along ν (or of N along μ). [§6.1]

6.12. Prove Proposition 6.2.

6.13. Prove that every homomorphic image of a finitely generated module is finitely generated.

6.14. ▷ Prove that the ideal $(x_1, x_2, \dots)$ of the ring $R = \mathbb{Z}[x_1, x_2, \dots]$ is not finitely generated (as an ideal, i.e., as an R-module). [§6.4]

6.15. ▷ Let R be a commutative ring. Prove that a *commutative* R-algebra S is finitely generated *as an algebra* over R if and only if it is finitely generated *as a commutative algebra* over R. (Cf. §6.5.) [§6.5]

6.16. ▷ Let R be a ring. A (left-)R-module M is *cyclic* if $M = \langle m \rangle$ for some $m \in M$. Prove that simple modules (cf. Exercise 5.4) are cyclic. Prove that an R-module M is cyclic if and only if $M \cong R/I$ for some (left-)ideal I. Prove that every quotient of a cyclic module is cyclic. [6.17, §VI.4.1]

6.17. ¬ Let M be a cyclic R-module, so that $M \cong R/I$ for a (left-)ideal I (Exercise 6.16), and let N be another R-module.

- Prove that $\mathrm{Hom}_{R\text{-}\mathsf{Mod}}(M, N) \cong \{n \in N \mid (\forall a \in I), an = 0\}$.
- For $a, b \in \mathbb{Z}$, prove that $\mathrm{Hom}_{R\text{-}\mathsf{Mod}}(\mathbb{Z}/a\mathbb{Z}, \mathbb{Z}/b\mathbb{Z}) \cong \mathbb{Z}/\gcd(a, b)\mathbb{Z}$.

[7.7]

6.18. ▷ Let M be an R-module, and let N be a submodule of M. Prove that if N and M/N are both finitely generated, then M is finitely generated. [§6.4]

---

7. Complexes and homology

In many contexts, modules arise not 'one at a time' but in whole series: for example, a real manifold of dimension d has one 'homology' group for each dimension from 0 to d. It is necessary to develop a language capable of dealing with whole sequences of modules at once. This is the language of *homological algebra,* of which we will get a tiny taste in this section, and a slightly heartier course in Chapter IX.

7.1. Complexes and exact sequences. A *chain complex of R-modules* (or, for simplicity, a *complex*) is a sequence of R-modules and R-module homomorphisms

$$\cdots \xrightarrow{d_{i+2}} M_{i+1} \xrightarrow{d_{i+1}} M_i \xrightarrow{d_i} M_{i-1} \xrightarrow{d_{i-1}} \cdots$$

such that $(\forall i) : d_i \circ d_{i+1} = 0$.

The notation $(M_\bullet, d_\bullet)$ may be used to denote a complex, or simply $M_\bullet$ for simplicity (but do not forget that the homomorphisms d_i are part of the information carried by a complex).

A complex may be infinite in both directions; 'tails' of 0's are (usually) omitted. Several possible alternative conventions may be used: for example, indices may be increasing rather than decreasing, giving a *co*chain complex (whose *homology* is called *cohomology;* this will be our choice in Chapter IX). Such choices are clearly mathematically immaterial, at least for the simple considerations which follow.

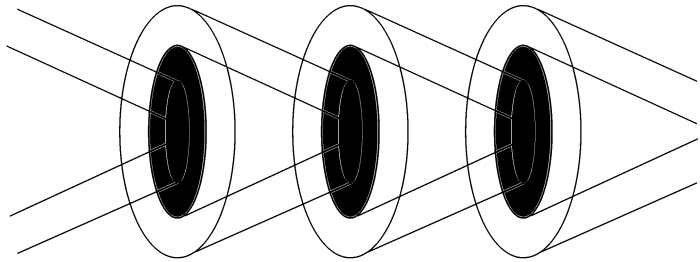
The homomorphisms d_i are called *boundary,* or *differentials,* due to important examples from geometry. Note that the defining condition

$$d_i \circ d_{i+1} = 0$$

is equivalent to the requirement

$$\operatorname{im} d_{i+1} \subseteq \ker d_i.$$

We carry in our minds an image such as

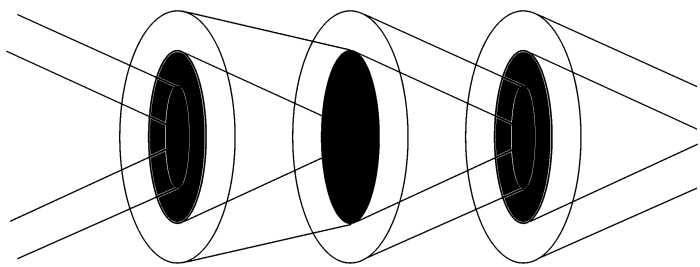


when we think of a complex. The ovals are the modules M_i; the fat black dots are the 0 elements; the gray ovals, getting squashed to zero at each step, are the kernels; and we thus visualize the fact that the image of 'the preceding homomorphism' falls inside the kernel of 'the next homomorphism'.

The picture is inaccurate in that it hints that the 'difference' between the image of d_{i+1} and the kernel of d_i (that is, the areas colored in a lighter shade of gray) should be the same for all i; this is of course not the case in general. In fact, almost the whole point about complexes is to 'measure' this difference, which is called the *homology* of the complex (cf. §7.3). We say that a complex is *exact* 'at M_i' if it has no homology there; that is,

$$\operatorname{im} d_{i+1} = \ker d_i.$$

Visually,



This complex appears to be exact at the oval in the middle.

For example, if M_i = a trivial module (usually denoted simply by 0), then the complex is necessarily exact at M_i, since then $\operatorname{im} d_{i+1} = \ker d_i = 0$.

A complex is *exact* and is often called an *exact sequence* if it is exact at all its modules.

Example 7.1. A complex

$$\cdots \longrightarrow 0 \longrightarrow L \xrightarrow{\alpha} M \longrightarrow \cdots$$

is exact at L if and only if α is a monomorphism.

Indeed, exactness at L is equivalent to $\ker \alpha$ = image of the trivial homomorphism $0 \to L$, that is, to

$$\ker \alpha = 0.$$

This is equivalent to the injectivity of α (Proposition 6.2). ⌟

Example 7.2. A complex

$$\cdots \longrightarrow M \xrightarrow{\beta} N \longrightarrow 0 \longrightarrow \cdots$$

is exact at N if and only if β is an epimorphism.

Indeed, the complex is exact at N if and only if $\operatorname{im}\beta =$ kernel of the trivial homomorphism $N \to 0$, that is, $\operatorname{im}\beta = N$. ⌟

Definition 7.3. A *short exact sequence* is an exact complex of the form

$$0 \longrightarrow L \xrightarrow{\alpha} M \xrightarrow{\beta} N \longrightarrow 0\,. \qquad ⌟$$

As seen in the previous two examples, exactness at L and N is equivalent to α being injective and β being surjective. The extra piece of data carried by a short exact sequence is the exactness at M, that is,

$$\operatorname{im}\alpha = \ker\beta;$$

by the first isomorphism theorem (Corollary 5.16), we then have

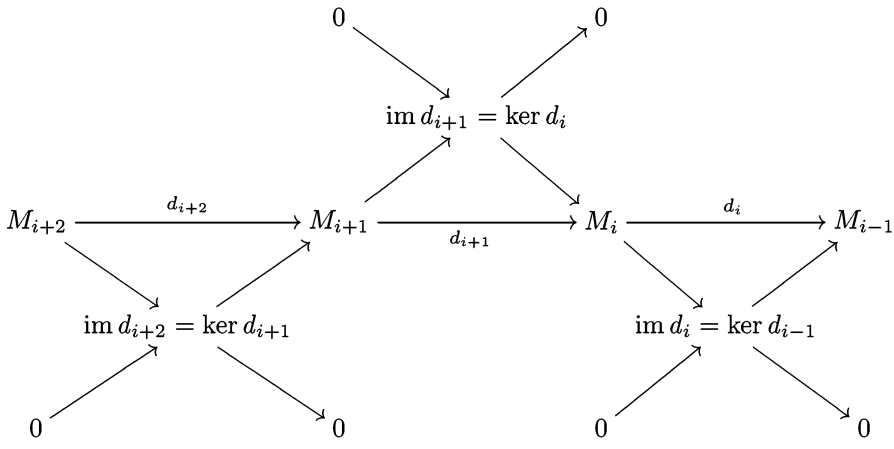
$$N \cong \frac{M}{\ker\beta} = \frac{M}{\operatorname{im}\alpha}.$$

All in all, we have good material to work on some more Pavlovian conditioning: at the sight of a short exact sequence as above, the reader should instinctively identify L with a submodule of M (via the injective map α) and N with the quotient M/L (via the isomorphism induced by the surjective map β, under the auspices of the first isomorphism theorem).

Short exact sequences abound in nature. For example, a single homomorphism $\varphi : M \to M'$ gives rise immediately to a short exact sequence

$$0 \longrightarrow \ker\varphi \longrightarrow M \longrightarrow \operatorname{im}\varphi \longrightarrow 0\,.$$

In fact, one important reason to focus on short exact sequences is that this observation allows us to break up every exact complex into a large number of short exact sequences: contemplate the impressive diagram



The diagonal sequences are short exact sequences, and they interlock nicely by the exactness of the horizontal complex.

This observation simplifies many arguments; cf. for example Exercise 7.5.

7.2. Split exact sequences. A particular case of short exact sequence arises by considering the second projection from a direct sum: $M_1 \oplus M_2 \to M_2$; there is then an exact sequence

$$0 \longrightarrow M_1 \longrightarrow M_1 \oplus M_2 \longrightarrow M_2 \longrightarrow 0\ ,$$

obtained by identifying M_1 with the kernel of the projection. These short exact sequences are said to 'split'; more generally, a short exact sequence

$$0 \longrightarrow M_1 \longrightarrow N \longrightarrow M_2 \longrightarrow 0$$

'splits' if it is isomorphic to one of these sequences in the sense that there is a commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & M_1 & \longrightarrow & N & \longrightarrow & M_2 & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle\sim} & & \downarrow{\scriptstyle\sim} & & \downarrow{\scriptstyle\sim} & & \\ 0 & \longrightarrow & M_1' & \longrightarrow & M_1' \oplus M_2' & \longrightarrow & M_2' & \longrightarrow & 0 \end{array}$$

in which the vertical maps are all isomorphisms[33].

Example 7.4. The exact sequence of $\mathbb{Z}$-modules

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\cdot 2} \mathbb{Z} \longrightarrow \tfrac{\mathbb{Z}}{2\mathbb{Z}} \longrightarrow 0$$

is *not* split. ⌟

Splitting sequences give us the opportunity to go back to a question we left dangling at the end of §6.2: what should we make of the condition of 'having a left- (resp., right-) inverse' for a homomorphism? We realized that this condition is stronger than the requirement of being a monomorphisms (resp., an epimorphism); can we give a more explicit description of such morphisms?

Proposition 7.5. *Let $\varphi : M \to N$ be an R-module homomorphism. Then*

- *φ has a left-inverse if and only if the sequence*

$$0 \longrightarrow M \xrightarrow{\varphi} N \longrightarrow \operatorname{coker}\varphi \longrightarrow 0$$

  splits.

- *φ has a right-inverse if and only if the sequence*

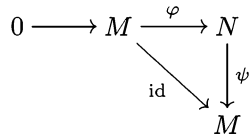
$$0 \longrightarrow \ker\varphi \longrightarrow M \xrightarrow{\varphi} N \longrightarrow 0$$

  splits.

[33] In fact, this last requirement is somewhat redundant; cf. Exercise 7.11.

Proof. We will prove the first part and leave the other as an exercise to the reader (Exercise 7.6).

If the sequence splits, then φ may be identified with the embedding of M into a direct sum $M \oplus M'$, and the projection $M \oplus M' \to M$ gives a left-inverse of φ. Conversely, assume that φ has a left-inverse ψ:



Then we claim that N is isomorphic to $M \oplus \ker\psi$ and that φ corresponds to the identification of M with the first factor: $M \to M \oplus \ker\psi \cong N$. The isomorphism $M \oplus \ker\psi \to N$ is given by

$$(m, k) \mapsto \varphi(m) + k;$$

its inverse $N \to M \oplus \ker\psi$ is

$$n \mapsto (\psi(n), n - \varphi\psi(n)).$$

The element $n - \varphi\psi(n)$ is in $\ker\psi$ as it should be, since

$$\psi(n - \varphi\psi(n)) = \psi(n) - \psi\varphi\psi(n) = \psi(n) - \psi(n) = 0.$$

All necessary verifications are immediate and are left to the reader. □

Because of Proposition 7.5, R-module homomorphisms with a left-inverse are called *split monomorphisms,* and homomorphisms with a right-inverse are called *split epimorphisms.*

We will come back to split exact sequences (in the more demanding context of Grp) in §IV.5.2 and then later again when we return to modules and, more generally, to abelian categories (Chapters VIII and IX).

7.3. Homology and the snake lemma.

Definition 7.6. The *i-th homology* of a complex

$$M_\bullet : \cdots \xrightarrow{d_{i+2}} M_{i+1} \xrightarrow{d_{i+1}} M_i \xrightarrow{d_i} M_{i-1} \xrightarrow{d_{i-1}} \cdots$$

of R-modules is the R-module

$$H_i(M_\bullet) := \frac{\ker d_i}{\operatorname{im} d_{i+1}}.$$ ⌟

That is, $H_i(M_\bullet)$ is a module capturing the 'light gray annulus' in the heuristic picture of a complex. Of course

$$H_i(M_\bullet) = 0 \iff \operatorname{im} d_{i+1} = \ker d_i \iff \text{the complex } M_\bullet \text{ is exact at } M_i :$$

that is, the homology modules are a measure of the 'failure of a complex from being exact'.

Example 7.7. In fact, homology should be thought of as a (vast) generalization of the notions of kernel and cokernel. Indeed, consider the (very) particular case in which $M_\bullet$ is the complex

$$0 \longrightarrow M_1 \xrightarrow{\varphi} M_0 \longrightarrow 0\ .$$

Then

$$H_1(M_\bullet) \cong \ker\varphi, \quad H_0(M_\bullet) \cong \operatorname{coker}\varphi. \qquad \lrcorner$$

We will end this very brief excursion into more abstract territories by indicating how a commutative diagram involving *two* short exact sequences generates a 'long exact sequence' in homology. This is actually a particular case of a more general construction—according to which a suitable commutative diagram involving *three* complexes yields a *really long* 'long exact homology sequence'. We will come back to this general construction when we deal more extensively with homological algebra in Chapter IX. The reader is also likely to learn about it in a course on algebraic topology, where this fact is put to impressive use in studying invariants of manifolds.

In the simple form we will analyze, this is affectionately known as the *snake lemma*. Consider two short exact sequences linked by homomorphisms, so as to form a commutative diagram[34]:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & L_1 & \xrightarrow{\alpha_1} & M_1 & \xrightarrow{\beta_1} & N_1 & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle\lambda} & & \downarrow{\scriptstyle\mu} & & \downarrow{\scriptstyle\nu} & & \\ 0 & \longrightarrow & L_0 & \xrightarrow{\alpha_0} & M_0 & \xrightarrow{\beta_0} & N_0 & \longrightarrow & 0 \end{array}$$

Lemma 7.8 (The snake lemma)**.** *With notation as above, there is an exact sequence*

$$0 \longrightarrow \ker\lambda \longrightarrow \ker\mu \longrightarrow \ker\nu \xrightarrow{\delta} \operatorname{coker}\lambda \longrightarrow \operatorname{coker}\mu \longrightarrow \operatorname{coker}\nu \longrightarrow 0\ .$$

Remark 7.9. Most of the homomorphisms in this sequence are induced in a completely straightforward way from the corresponding homomorphisms λ, μ, ν. The one 'surprising' homomorphism is the one denoted δ; we will discuss its definition below. $\lrcorner$

Remark 7.10. In view of Example 7.7, we could have written the sequence in this statement as

$$\begin{array}{ccccccccc} 0 & \longrightarrow & H_1(L_\bullet) & \longrightarrow & H_1(M_\bullet) & \longrightarrow & H_1(N_\bullet) & & \\ & & & & & \overset{\delta}{\searrow} & & & \\ & & H_0(L_\bullet) & \longrightarrow & H_0(M_\bullet) & \longrightarrow & H_0(N_\bullet) & \longrightarrow & 0 \end{array}$$

where $L_\bullet$ is the complex $0 \longrightarrow L_0 \xrightarrow{\lambda} L_1 \longrightarrow 0$, etc. The snake lemma generalizes to arbitrary complexes $L_\bullet$, $M_\bullet$, $N_\bullet$, producing a 'long exact homology

[34] In fact, it is better to view this diagram as *three* (very short) complexes linked by R-module homomorphisms α_i, β_i so that 'the rows are exact'. In fact, one can define a category *of complexes,* and this diagram is nothing but a 'short exact sequence of complexes'; this is the approach we will take in Chapter IX.

sequence' of which this is just the tail end. As mentioned above, we will discuss this rather straightforward generalization later (§IX.3.3). ⌟

Remark 7.11. A popular version of the snake lemma does not assume that α_1 is injective and β_0 is surjective: that is, we could consider a commutative diagram of exact sequences

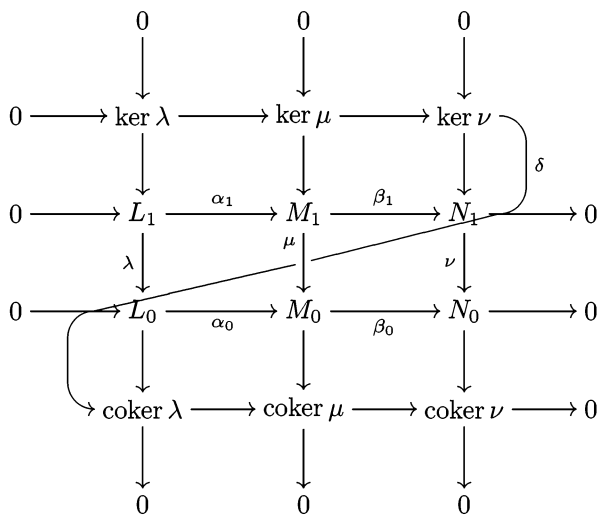
$$\begin{array}{ccccccc} & & L_1 & \longrightarrow & M_1 & \longrightarrow & N_1 \longrightarrow 0 \\ & & \downarrow{\scriptstyle\lambda} & & \downarrow{\scriptstyle\mu} & & \downarrow{\scriptstyle\nu} \\ 0 & \longrightarrow & L_0 & \longrightarrow & M_0 & \longrightarrow & N_0 \end{array}$$

The lemma will then state that there is 'only' an exact sequence

$$\ker\lambda \longrightarrow \ker\mu \longrightarrow \ker\nu \xrightarrow{\delta} \operatorname{coker}\lambda \longrightarrow \operatorname{coker}\mu \longrightarrow \operatorname{coker}\nu\,. \qquad ⌟$$

PROVING the snake lemma is something that should not be done in public, and it is notoriously useless to write down the details of the verification for others to read: the details are all essentially obvious, but they lead quickly to a notational quagmire. Such proofs are collectively known as the sport of *diagram chase,* best executed by pointing several fingers at different parts of a diagram on a blackboard, while enunciating the elements one is manipulating and stating their fate[35].

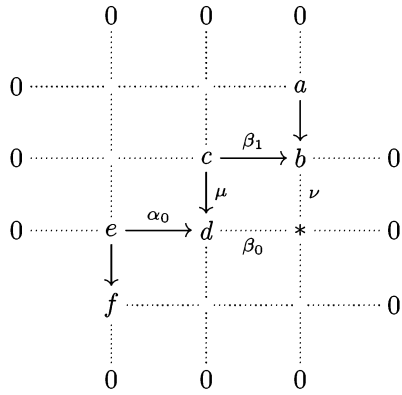
Nevertheless, we should explain where the 'connecting' homomorphism δ comes from, since this is the heart of the statement of the snake lemma and of its proof. Here is the whole diagram, including kernels and cokernels; thus, columns are exact (as well as the two original sequences, placed horizontally):



By the way, we trust that the reader now sees why this lemma is called the *snake* lemma.

[35]Real purists chase diagrams in arbitrary categories, thus without the benefit of talking about 'elements', and we will practice this skill later on (Chapter IX). For example, the snake lemma can be proven by appealing to universal property after universal property of kernels and cokernels, without ever choosing elements anywhere. But the performing technique of pointing fingers at a board while monologuing through the argument remains essentially the same.

Definition of the snaking homomorphism δ. Let $a \in \ker \nu$. We claim that a can be mapped through the diagram all the way to $\operatorname{coker} \lambda$, along the solid arrows marked here:



Indeed,

- $\ker \nu \subseteq N_1$; so view a as an element b of N_1.
- β_1 is surjective, so $\exists c \in M_1$, mapping to b.
- Let $d = \mu(c)$ be the image of c in M_0.
- What is the image of d in the spot marked $*$? By the commutativity of the diagram, it must be the same as $\nu(b)$. However, b was the image in N_1 of $a \in \ker \nu$, so $\nu(b) = 0$. Thus, $d \in \ker \beta_0$. Since rows are exact, $\ker \beta_0 = \operatorname{im} \alpha_0$; therefore, $\exists e \in L_0$, mapping to d.
- Finally, let $f \in \operatorname{coker} \mu$ be the image of e.

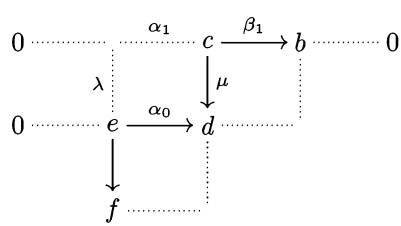
We want to set $\delta(a) := f$.

Is this legal? At two steps in the chase we have taken preimages:

- $\exists c \in M_1$ such that $\beta_1(c) = b$,
- $\exists e \in L_0$ such that $\alpha_0(e) = d$.

The second step does not involve a choice: because α_0 is injective by assumption, so the element e mapping to d is uniquely determined by d. But there *was* a choice involved in the first step: in order to verify that δ is well-defined, we have to show that choosing *some other* c would not affect the proposed value f for $\delta(a)$.

This is proved by another chase. Here is the relevant part of the diagram:



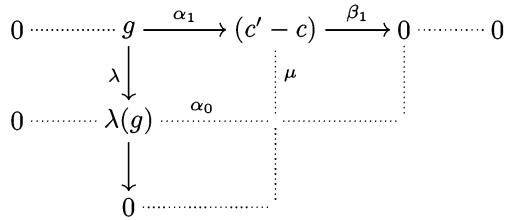
Suppose we choose a different c' mapping to the same b:

$$0 \xrightarrow{\alpha_1} c' \xrightarrow{\beta_1} b \longrightarrow 0.$$

Then $\beta_1(c'-c)=0$; by exactness, $\exists g \in L_1$ such that $(c'-c)=\alpha_1(g)$:

$$0 \longrightarrow g \xrightarrow{\alpha_1} (c'-c) \xrightarrow{\beta_1} 0 \longrightarrow 0.$$

Now the point is that, since columns form complexes, g dies in coker λ:



and it follows (by the commutativity of the diagram and the injectivity of α_0) that changing c to c' modifies e to $e+\lambda(g)$ and f to $f+0=f$. That is, f is indeed independent of the choice.

Thus δ *is* well-defined!

This is a tiny part of the proof of the snake lemma, but it probably suffices to demonstrate why reading a written-out version of a diagram chase may be supremely uninformative.

The rest of the proof (left to the reader(!) but we are not listing this as an official exercise for fear that someone might actually turn a solution in for grading) amounts to many, many similar arguments. The definition of the maps induced on kernels and cokernels is substantially less challenging than the definition of the connecting morphism δ described above. Exactness at most spots in the sequence

$$0 \longrightarrow \ker\lambda \longrightarrow \ker\mu \longrightarrow \ker\nu \xrightarrow{\delta} \operatorname{coker}\lambda \longrightarrow \operatorname{coker}\mu \longrightarrow \operatorname{coker}\nu \longrightarrow 0$$

is also reasonably straightforward; most of the work will go into proving exactness at $\ker\nu$ and $\operatorname{coker}\lambda$.

Dear reader: don't shy away from trying this, for it is excellent, indispensable practice. Miss this opportunity and you will forever feel unsure about such manipulations.

The snake lemma streamlines several facts, which would not be hard to prove individually, but become really straightforward once the lemma is settled. For example,

Corollary 7.12. *In the same situation presented in the snake lemma (notation as in §7.3), assume that μ is surjective and ν is injective. Then λ is surjective and ν is an isomorphism.*

Proof. Indeed, μ surjective $\implies \operatorname{coker}\mu=0$; ν injective $\implies \ker\nu=0$ (Proposition 6.2). Feeding this information into the sequence of the snake lemma gives an

exact sequence

$$0 \longrightarrow \ker\lambda \longrightarrow \ker\mu \longrightarrow 0 \longrightarrow \operatorname{coker}\lambda \longrightarrow 0 \longrightarrow \operatorname{coker}\nu \longrightarrow 0$$

Exactness implies $\operatorname{coker}\lambda = \operatorname{coker}\nu = 0$ (Exercise 7.1); hence λ and ν are surjective, with the stated consequences. □

Several more such statements may be proved just as easily; the reader should experiment to his or her heart's content.

Exercises

7.1. ▷ Assume that the complex

$$\cdots \longrightarrow 0 \longrightarrow M \longrightarrow 0 \longrightarrow \cdots$$

is exact. Prove that $M \cong 0$. [§7.3]

7.2. Assume that the complex

$$\cdots \longrightarrow 0 \longrightarrow M \longrightarrow M' \longrightarrow 0 \longrightarrow \cdots$$

is exact. Prove that $M \cong M'$.

7.3. Assume that the complex

$$\cdots \longrightarrow 0 \longrightarrow L \longrightarrow M \xrightarrow{\varphi} M' \longrightarrow N \longrightarrow 0 \longrightarrow \cdots$$

is exact. Show that, up to natural identifications, $L = \ker\varphi$ and $N = \operatorname{coker}\varphi$.

7.4. Construct short exact sequences of $\mathbb{Z}$-modules

$$0 \longrightarrow \mathbb{Z}^{\oplus\mathbb{N}} \longrightarrow \mathbb{Z}^{\oplus\mathbb{N}} \longrightarrow \mathbb{Z} \longrightarrow 0$$

and

$$0 \longrightarrow \mathbb{Z}^{\oplus\mathbb{N}} \longrightarrow \mathbb{Z}^{\oplus\mathbb{N}} \longrightarrow \mathbb{Z}^{\oplus\mathbb{N}} \longrightarrow 0.$$

(Hint: David Hilbert's Grand Hotel.)

7.5. ▷ Assume that the complex

$$\cdots \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow \cdots$$

is exact and that L and N are Noetherian. Prove that M is Noetherian. [§7.1]

7.6. ▷ Prove the 'split epimorphism' part of Proposition 7.5. [§7.2]

7.7. ▷ Let

$$0 \longrightarrow M \longrightarrow N \longrightarrow P \longrightarrow 0$$

be a short exact sequence of R-modules, and let L be an R-module.

(i) Prove that there is an exact sequence[36]

$$0 \longrightarrow \operatorname{Hom}_{R\text{-Mod}}(P, L) \longrightarrow \operatorname{Hom}_{R\text{-Mod}}(N, L) \longrightarrow \operatorname{Hom}_{R\text{-Mod}}(M, L).$$

[36] In general, this will be a sequence of abelian groups; if R is commutative, so that each $\operatorname{Hom}_{R\text{-Mod}}$ is an R-module (§5.2), then it will be an exact sequence of R-modules.

(ii) Redo Exercise 6.17. (Use the exact sequence $0 \to I \to R \to R/I \to 0$.)

(iii) Construct an example showing that the rightmost homomorphism in (i) need not be onto.

(iv) Show that if the original sequence splits, then the rightmost homomorphism in (i) *is* onto.

[7.9, VIII.3.14, §VIII.5.1]

7.8. ▷ Prove that every exact sequence

$$0 \longrightarrow M \longrightarrow N \longrightarrow F \longrightarrow 0$$

of R modules, with F *free,* splits. (Hint: Exercise 6.9.) [§VIII.5.4]

7.9. Let

$$0 \longrightarrow M \longrightarrow N \longrightarrow F \longrightarrow 0$$

be a short exact sequence of R-modules, with F free, and let L be an R-module. Prove that there is an exact sequence

$$0 \longrightarrow \mathrm{Hom}_{R\text{-}\mathsf{Mod}}(F, L) \longrightarrow \mathrm{Hom}_{R\text{-}\mathsf{Mod}}(N, L) \longrightarrow \mathrm{Hom}_{R\text{-}\mathsf{Mod}}(M, L) \longrightarrow 0\ .$$

(Cf. Exercise 7.7.)

7.10. ▷ In the situation of the snake lemma, assume that λ and ν are isomorphisms. Use the snake lemma and prove that μ is an isomorphism. This is called the 'short five-lemma,' as it follows immediately from the five-lemma (cf. Exercise 7.14), as well as from the snake lemma. [VIII.6.21, IX.2.4]

7.11. ▷ Let

$$(*) \qquad 0 \longrightarrow M_1 \longrightarrow N \longrightarrow M_2 \longrightarrow 0$$

be an exact sequence of R-modules. (This may be called an 'extension' of M_2 by M_2.) Suppose there is *any* R-module homomorphism $N \to M_1 \oplus M_2$ making the diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & M_1 & \longrightarrow & N & \longrightarrow & M_2 & \longrightarrow & 0 \\ & & \| & & \downarrow & & \| & & \\ 0 & \longrightarrow & M_1 & \longrightarrow & M_1 \oplus M_2 & \longrightarrow & M_2 & \longrightarrow & 0 \end{array}$$

commute, where the bottom sequence is the standard sequence of a direct sum. Prove that (*) splits. [§7.2]

7.12. ¬ Practice your diagram chasing skills by proving the 'four-lemma': if

$$\begin{array}{ccccccc} A_1 & \longrightarrow & B_1 & \longrightarrow & C_1 & \longrightarrow & D_1 \\ \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle \gamma} & & \downarrow{\scriptstyle \delta} \\ A_0 & \longrightarrow & B_0 & \longrightarrow & C_0 & \longrightarrow & D_0 \end{array}$$

is a commutative diagram of R-modules with exact rows, α is an epimorphism, and β, δ are monomorphisms, then γ is an isomorphism. [7.13, IX.2.3]

7.13. Prove another[37] version of the 'four-lemma' of Exercise 7.12: if

$$\begin{array}{ccccccc} B_1 & \longrightarrow & C_1 & \longrightarrow & D_1 & \longrightarrow & E_1 \\ \downarrow{\scriptstyle\beta} & & \downarrow{\scriptstyle\gamma} & & \downarrow{\scriptstyle\delta} & & \downarrow{\scriptstyle\epsilon} \\ B_0 & \longrightarrow & C_0 & \longrightarrow & D_0 & \longrightarrow & E_0 \end{array}$$

is a commutative diagram of R-modules with exact rows, β and δ are epimorphisms, and ϵ is a monomorphism, then γ is an isomorphism.

7.14. ¬ Prove the 'five-lemma': if

$$\begin{array}{ccccccccc} A_1 & \longrightarrow & B_1 & \longrightarrow & C_1 & \longrightarrow & D_1 & \longrightarrow & E_1 \\ \downarrow{\scriptstyle\alpha} & & \downarrow{\scriptstyle\beta} & & \downarrow{\scriptstyle\gamma} & & \downarrow{\scriptstyle\delta} & & \downarrow{\scriptstyle\epsilon} \\ A_0 & \longrightarrow & B_0 & \longrightarrow & C_0 & \longrightarrow & D_0 & \longrightarrow & E_0 \end{array}$$

is a commutative diagram of R-modules with exact rows, β and δ are isomorphisms, α is an epimorphism, and ϵ is a monomorphism, then γ is an isomorphism. (You can avoid the needed diagram chase by pasting together results from previous exercises.) [7.10]

7.15. ¬ Consider the following commutative diagram of R-modules:

$$\begin{array}{ccccccccc} & & 0 & & 0 & & 0 & & \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 0 & \longrightarrow & L_2 & \longrightarrow & M_2 & \longrightarrow & N_2 & \longrightarrow & 0 \\ & & \downarrow & & \downarrow{\scriptstyle\alpha} & & \downarrow & & \\ 0 & \longrightarrow & L_1 & \longrightarrow & M_1 & \longrightarrow & N_1 & \longrightarrow & 0 \\ & & \downarrow & & \downarrow{\scriptstyle\beta} & & \downarrow & & \\ 0 & \longrightarrow & L_0 & \longrightarrow & M_0 & \longrightarrow & N_0 & \longrightarrow & 0 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ & & 0 & & 0 & & 0 & & \end{array}$$

Assume that the three rows are exact and the two rightmost columns are exact. Prove that the left column is exact. Second version: assume that the three rows are exact and the two leftmost columns are exact; prove that the right column is exact. This is the 'nine-lemma'. (You can avoid a diagram chase by applying the snake lemma; for this, you will have to turn the diagram by 90°.) [7.16]

7.16. In the same situation as in Exercise 7.15, assume that the three rows are exact and that the leftmost and rightmost columns are exact.

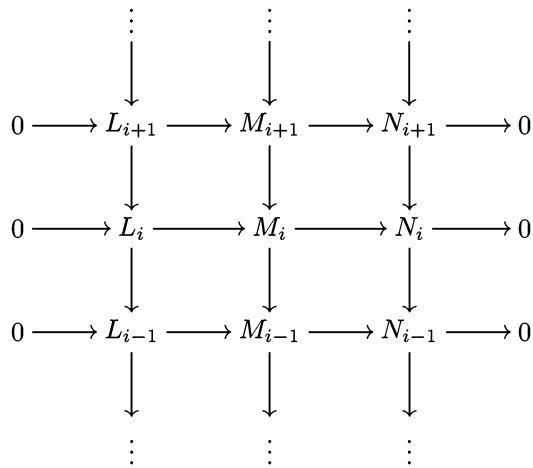
- Prove that α is a monomorphism and β is an epimorphism.
- Is the central column necessarily exact?

[37] It is in fact unnecessary to prove *both* versions, but to realize this one has to view the matter from the more general context of abelian categories; cf. Exercise IX.2.3.

(Hint: No. Place $\mathbb{Z} \oplus \mathbb{Z}$ in the middle, and surround it artfully with six copies of $\mathbb{Z}$ and two 0's.)

- Assume further that the central column is a complex (that is, $\beta \circ \alpha = 0$); prove that it is then necessarily exact.

7.17. ¬ Generalize the previous two exercises as follows. Consider a (possibly infinite) commutative diagram of R-modules:



in which the central column is a complex and every row is exact. Prove that the left and right columns are also complexes. Prove that if any two of the columns are exact, so is the third. (The first part is straightforward. The second part will take you a couple of minutes now due to the needed diagram chases, and a couple of seconds later, once you learn about the long exact (co)homology sequence in §IX.3.3.) [IX.3.12]

Groups, second encounter

In this chapter we return to Grp and study several topics of a less 'general' nature than those considered in Chapter II. Most of what we do here will apply exclusively to *finite* groups; this is an important example in its own right, as it has spectacular applications (for example, in Galois theory; cf. §VII.7), and it is a good subject from the expository point of view, since it gives us the opportunity to see several general concepts at work in a context that is complex enough to carry substance, but simple enough (in this tiny selection of elementary topics) to be appreciated easily.

1. The conjugation action

1.1. Actions of groups on sets, reminder. Groups really shine when you let them act on something. This section will make this point very effectively, since we will get surprisingly precise results on finite groups by extremely simple-minded applications of the elementary facts concerning group actions that we established back in §II.9.

Recall that we proved (Proposition II.9.9) that every *transitive* (left-) action of a group G on a set S is, up to a natural notion of isomorphism, 'left-multiplication on the set of left-cosets G/H'. Here, H may be taken to be the stabilizer $\mathrm{Stab}_G(a)$ of any element $a \in S$, that is (Definition II.9.8) the subgroup of G fixing a. This fact applies to the orbits of *every* left-action of G on a set; in particular, the number of elements in a finite orbit O equals the index of the stabilizer of any $a \in O$; in particular (Corollary II.9.10) the number of elements $|O|$ of an orbit must divide the order $|G|$ of G, if G is finite.

These considerations may be packaged into a useful 'counting' formula, which we could call the *class formula* for that action; this name is usually reserved to the particular case of the action of G onto itself by conjugation, which we will explore more carefully below.

In order to state the formula, assume G acts on a set S; for $a \in S$, let G_a denote the stabilizer $\mathrm{Stab}_G(a)$. Also, let Z be the set of *fixed points* of the action:

$$Z = \{a \in S \mid (\forall g \in G) : ga = a\}.$$

Note that $a \in Z \iff G_a = G$; we could say that $a \in Z$ if and only if the orbit of a is 'trivial', in the sense that it consists of a alone.

Proposition 1.1. *Let S be a finite set, and let G be a group acting on S. With notation as above,*

$$|S| = |Z| + \sum_{a \in A} [G : G_a],$$

where $A \subseteq S$ has exactly one element for each nontrivial orbit of the action.

Proof. The orbits form a partition of S, and Z collects the trivial orbits; hence

$$|S| = |Z| + \sum_{a \in A} |O_a|,$$

where O_a denotes the orbit of a. By Proposition II.9.9, the order $|O_a|$ equals the index of the stabilizer of a, yielding the statement. □

The main strength of Proposition 1.1 rests in the fact that, if G is finite, each summand $[G : G_a]$ *divides* the order of G (and is > 1). This can be a strong constraint, when some information is known about $|G|$. For example, let's see what this says when G is a *p-group:*

Definition 1.2. A p-group is a finite group whose order is a power of a prime integer p. ⌟

Corollary 1.3. *Let G be a p-group acting on a finite set S, and let Z be the fixed point set of the action. Then*

$$|Z| \equiv |S| \mod p.$$

Proof. Indeed, each summand $[G : G_a]$ in Proposition 1.1 is a power of p larger than 1; hence it is 0 mod p. □

For instance, in certain situations this can be used to establish[1] that $Z \neq \emptyset$: see Exercise 1.1. Such immediate consequences of Proposition 1.1 will assist us below, in the proof of Sylow's theorems.

[1] In this sense, Proposition 1.1 is an instance of a class of results known as 'fixed point theorems'. The reader will likely encounter a few such theorems in topology courses, where the role of the 'size' of a set may be played by (for example) the Euler characteristic of a topological space.

1.2. Center, centralizer, conjugacy classes. Recall (Example II.9.3) that every group G acts on itself in at least two interesting ways: by (left-) multiplication and by *conjugation*. The latter action is defined by the following $\rho : G \times G \to G$:

$$\rho(g, a) = gag^{-1}.$$

As we know (§II.9.2), this datum is equivalent to the datum of a certain group homomorphism:

$$\sigma : G \to S_G$$

from G to the permutation group on G.

This action highlights several interesting objects:

Definition 1.4. The *center* of G, denoted $Z(G)$, is the subgroup $\ker \sigma$ of G. ⌟

Concretely, the center[2] of G is

$$Z(G) = \{g \in G \mid (\forall a \in G) : ga = ag\}.$$

Indeed, $\sigma(g)$ is the identity in S_G if and only if $\sigma(g)$ acts as the identity on G; that is, if and only if $gag^{-1} = a$ for all $a \in G$; that is, if and only if g commutes with all elements of G. In other words, the center is the set of fixed points in G under the conjugation action.

Note that the center of a group G is automatically normal in G: this is nearly immediate to check 'by hand', but there is no need to do so since it is a kernel by definition and kernels are normal.

A group G is commutative if and only if $Z(G) = G$, that is, if and only if the conjugation action is trivial on G.

In general, feel happy when you discover that the center of a group is not trivial: this will often allow you to set up proofs by induction on the number of elements of the group, by mod-ing out by the center (this is, roughly, how we will prove the first Sylow theorem). Or note the following useful fact, which comes in handy when trying to prove that a group is commutative:

Lemma 1.5. *Let G be a finite group, and assume $G/Z(G)$ is cyclic. Then G is commutative (and hence $G/Z(G)$ is in fact trivial).*

Proof. (Cf. Exercise 1.5.) As $G/Z(G)$ is cyclic, there exists an element $g \in G$ such that the class $gZ(G)$ generates $G/Z(G)$. Then $\forall a \in G$

$$aZ(G) = (gZ(G))^r$$

for some $r \in \mathbb{Z}$; that is, there is an element $z \in Z(G)$ of the center such that $a = g^r z$.

If now a, b are in G, use this fact to write

$$a = g^r z, \quad b = g^s w$$

for some $s \in \mathbb{Z}$ and $w \in Z(G)$; but then

$$ab = (g^r z)(g^s w) = g^{r+s} zw = (g^s w)(g^r z) = ba,$$

[2] Why 'Z'? 'Center' ist Zentrum auf Deutsch.

where we have used the fact that z and w commute with every element of G. As a and b were arbitrary, this proves that G is commutative. □

Next, the stabilizer of $a \in G$ under conjugation has a special name:

Definition 1.6. The *centralizer* (or *normalizer*) $Z_G(a)$ of $a \in G$ is its stabilizer under conjugation. ⌟

Thus,

$$Z_G(a) = \{g \in G \mid gag^{-1} = a\} = \{g \in G \mid ga = ag\}$$

consists of those elements in G which commute *with* a. In particular, $Z(G) \subseteq Z_G(a)$ for all $a \in G$; in fact, $Z(G) = \bigcap_{a \in G} Z_G(a)$. Clearly $a \in Z(G) \iff Z_G(a) = G$.

If there is no ambiguity concerning the group G containing a, the index G may be dropped.

Definition 1.7. The *conjugacy class* of $a \in G$ is the orbit $[a]$ of a under the conjugation action. Two elements a, b of G are *conjugate* if they belong to the same conjugacy class. ⌟

The notation $[a]$ is not standard; $C(a)$ is used more frequently, but we are not fond of it. Using $[a]$ reminds us that these are nothing but the equivalence classes of elements of G under a certain interesting equivalence relation.

Note that $[a] = \{a\}$ if and only if $gag^{-1} = a$ for all $g \in G$; that is, if and only if $ga = ag$ for all $g \in G$; that is, if and only if $a \in Z(G)$.

1.3. The Class Formula. The 'official' Class Formula for a finite group G is the particular case of Proposition 1.1 for the conjugation action.

Proposition 1.8 (Class formula)**.** *Let G be a finite group. Then*

$$|G| = |Z(G)| + \sum_{a \in A} [G : Z(a)],$$

where $A \subseteq G$ is a set containing one representative for each nontrivial conjugacy class in G.

Proof. The set of fixed points is $Z(G)$, and the stabilizer of a is the centralizer $Z(a)$; apply Proposition 1.1. □

The class formula is surprisingly useful. In applying it, keep in mind that every summand on the right (that is, both $|Z(G)|$ and each $[G : Z(a)]$) is a divisor of $|G|$; this fact alone often suffices to draw striking conclusions about G.

Possibly the most famous such application is to *p-groups,* via Corollary 1.3:

Corollary 1.9. *Let G be a nontrivial p-group. Then G has a nontrivial center.*

Proof. Since $|Z(G)| \equiv |G| \bmod p$ and $|G| > 1$ is a power of p, necessarily $|Z(G)|$ is a multiple of p. As $Z(G) \neq \emptyset$ (since $e_G \in Z(G)$), this implies $|Z(G)| \geq p$. □

For example, it follows immediately (from Corollary 1.9 and Lemma 1.5; cf. Exercise 1.6) that if p is prime, then every group of order p^2 is commutative.

In general, the class formula poses a strong constraint on what can go on in a group.

Example 1.10. Consider a group G of order 6; what are the possibilities for its class formula?

If G is commutative, then the class formula will tell us very little:

$$6 = 6.$$

If G is not commutative, then its center must be trivial (as a consequence of Lagrange's theorem and Lemma 1.5); so the class formula is $6 = 1 + \cdots$, where $\cdots$ collects the sizes of the nontrivial conjugacy classes. But each of these summands must be larger than 1, smaller than 6, and must divide 6; that is, there are no choices:

$$6 = 1 + 2 + 3$$

is the only possibility. The reader should check that this is indeed the class formula for S_3; in fact, S_3 is the only noncommutative group of order 6 up to isomorphism (Exercise 1.13). ⌟

Another useful observation is that *normal* subgroups must be unions of conjugacy classes: because if H is a normal subgroup, $a \in H$, and $b = gag^{-1}$ is conjugate to a, then

$$b \in gHg^{-1} = H.$$

To stick with the $|G| = 6$ example, note that every subgroup of a group must contain the identity and its size must divide the order of the group; it follows that a normal subgroup of a noncommutative group of order 6 cannot have order 2, since 2 cannot be written as sums of orders of conjugacy classes (including the class of the identity).

1.4. Conjugation of subsets and subgroups. We may also act by conjugation on the power set of G: if $A \subseteq G$ is a subset and $g \in G$, the *conjugate* of A is the subset gAg^{-1}. By cancellation, the conjugation map $a \mapsto gag^{-1}$ is a *bijection* between A and gAg^{-1}.

This leads to terminology analogous to the one introduced in §1.2.

Definition 1.11. The *normalizer* $N_G(A)$ of A is its stabilizer under conjugation. The *centralizer* of A is the subgroup $Z_G(A) \subseteq N_G(A)$ fixing each element of A. ⌟

Thus, $g \in N_G(A)$ if and only if[3] $gAg^{-1} = A$, and $g \in Z_G(A)$ if and only if $\forall a \in A$, $gag^{-1} = a$.

For $A = \{a\}$ a singleton, we have $N_G(\{a\}) = Z_G(\{a\}) = Z_G(a)$. In general, $Z_G(A) \subsetneq N_G(A)$.

If H is a subgroup of G, every conjugate gHg^{-1} of H is also a subgroup of G; conjugate subgroups have the same order.

[3]If A is finite (but not in general), this condition is equivalent to $gAg^{-1} \subseteq A$.

Remark 1.12. The definition implies immediately that $H \subseteq N_G(H)$ and that H is normal in G if and only if $N_G(H) = G$. More generally, the normalizer $N_G(H)$ of H in G is (clearly) the *largest subgroup of G in which H is normal.* ⌟

One could apply Proposition 1.1 to the conjugation action on subsets or subgroup; however, there are too many subsets, and one has little control over the number of subgroups. Other numerical considerations involving the number of conjugates of a given subset or subgroups may be very useful.

Lemma 1.13. *Let $H \subseteq G$ be a subgroup. Then (if finite) the number of subgroups conjugate to H equals the index $[G : N_G(H)]$ of the normalizer of H in G.*

Proof. This is again an immediate consequence of Proposition II.9.9. □

Corollary 1.14. *If $[G : H]$ is finite, then the number of subgroups conjugate to H is finite and divides $[G : H]$.*

Proof.

$$[G : H] = [G : N_G(H)] \cdot [N_G(H) : H]$$

(cf. §II.8.5). □

One of the celebrated Sylow theorems will strengthen this statement substantially in the case in which H is a *maximal* p-group contained in a finite group G. For a statement concerning the size of the normalizer of an arbitrary p-subgroup of a group, see Lemma 2.9.

Another useful numerical tool is the observation that if H and K are subgroups of a group G and $H \subseteq N_G(K)$—so that $gKg^{-1} = K$ for all $g \in H$—then conjugation by $g \in H$ gives an *automorphism* of K. Indeed, we have already observed that conjugation is a bijection, and it is immediate to see that it is a homomorphism: $\forall k_1, k_2 \in K$

$$(gk_1g^{-1})(gk_2g^{-1}) = gk_1(g^{-1}g)k_2g^{-1} = g(k_1k_2)g^{-1}.$$

Thus, conjugation gives a set-function

$$\gamma : H \to \mathrm{Aut}_{\mathsf{Grp}}(K).$$

The reader will check that this is a group homomorphism and will determine $\ker \gamma$ (Exercise 1.21).

This is especially useful if H is finite and some information is available concerning $\mathrm{Aut}_{\mathsf{Grp}}(K)$ (for an example, see Exercise 4.14). A classic application is presented in Exercise 1.22.

Exercises

1.1. ▷ Let p be a prime integer, let G be a p-group, and let S be a set such that $|S| \not\equiv 0 \bmod p$. If G acts on S, prove that the action must have fixed points. [§1.1, §2.3]

1.2. Find the center of D_{2n}. (The answer depends on the parity of n. Use a presentation.)

1.3. Prove that the center of S_n is trivial for $n \geq 3$. (Suppose that $\sigma \in S_n$ sends a to $b \neq a$, and let $c \neq a, b$. Let τ be the permutation that acts solely by swapping b and c. Then compare the action of $\sigma\tau$ and $\tau\sigma$ on a.)

1.4. ▷ Let G be a group, and let N be a subgroup of $Z(G)$. Prove that N is normal in G. [§2.2]

1.5. ▷ Let G be a group. Prove that $G/Z(G)$ is isomorphic to the group $\mathrm{Inn}(G)$ of *inner* automorphisms of G. (Cf. Exercise II.4.8.) Then prove Lemma 1.5 again by using the result of Exercise II.6.7. [§1.2]

1.6. ▷ Let p, q be prime integers, and let G be a group of order pq. Prove that either G is commutative or the center of G is trivial. Conclude (using Corollary 1.9) that every group of order p^2, for a prime p, is commutative. [§1.3]

1.7. Prove or disprove that if p is prime, then every group of order p^3 is commutative.

1.8. ▷ Let p be a prime number, and let G be a p-group: $|G| = p^r$. Prove that G contains a normal subgroup of order p^k for every nonnegative $k \leq r$. [§2.2]

1.9. ¬ Let p be a prime number, G a p-group, and H a nontrivial normal subgroup of G. Prove that $H \cap Z(G) \neq \{e\}$. (Hint: Use the class formula.) [3.11]

1.10. Prove that if G is a group of odd order and $g \in G$ is conjugate to g^{-1}, then $g = e_G$.

1.11. Let G be a finite group, and suppose there exist representatives $g_1, \ldots, g_r$ of the r distinct conjugacy classes in G, such that $\forall i, j$, $g_i g_j = g_j g_i$. Prove that G is commutative. (Hint: What can you say about the sizes of the conjugacy classes?)

1.12. Verify that the class formula for both D_8 and Q_8 (cf. Exercise III.1.12) is $8 = 2 + 2 + 2 + 2$. (Also note that $D_8 \not\cong Q_8$.)

1.13. ▷ Let G be a noncommutative group of order 6. As observed in Example 1.10, G must have trivial center and exactly two conjugacy classes, of order 2 and 3.

- Prove that if every element of a group has order ≤ 2, then the group is commutative. Conclude that G has an element y of order 3.
- Prove that $\langle y \rangle$ is normal in G.
- Prove that $[y]$ is the conjugacy class of order 2 and $[y] = \{y, y^2\}$.
- Prove that there is an $x \in G$ such that $yx = xy^2$.

- Prove that x has order 2.
- Prove that x and y generate G.
- Prove that $G \cong S_3$.

[§1.3, §2.5]

1.14. Let G be a group, and assume $[G : Z(G)] = n$ is finite. Let $A \subseteq G$ be any subset. Prove that the number of conjugates of A is at most n.

1.15. Suppose that the class formula for a group G is $60 = 1 + 15 + 20 + 12 + 12$. Prove that the only *normal* subgroups of G are $\{e\}$ and G.

1.16. ▷ Let G be a finite group, and let $H \subseteq G$ be a subgroup of index 2. For $a \in H$, denote by $[a]_H$, resp., $[a]_G$, the conjugacy class of a in H, resp., G. Prove that either $[a]_H = [a]_G$ or $[a]_H$ is half the size of $[a]_G$, according to whether the centralizer $Z_G(a)$ is not or is contained in H. (Hint: Note that H is normal in G, by Exercise II.8.2; apply Proposition II.8.11.) [§4.4]

1.17. ¬ Let H be a proper subgroup of a finite group G. Prove that G is *not* the union of the conjugates of H. (Hint: You know the number of conjugates of H; keep in mind that any two subgroups overlap, at least at the identity.) [1.18, 1.20]

1.18. Let S be a set endowed with a transitive action of a finite group G, and assume $|S| \geq 2$. Prove that there exists a $g \in G$ without fixed points in S, that is, such that $gs \neq s$ for all $s \in S$. (Hint: By Proposition II.9.9, you may assume $S = G/H$, with H proper in G. Use Exercise 1.17.)

1.19. Let H be a proper subgroup of a finite group G. Prove that there exists a $g \in G$ whose conjugacy class is disjoint from H.

1.20. Let $G = \mathrm{GL}_2(\mathbb{C})$, and let H be the subgroup consisting of upper triangular matrices (Exercise II.6.2). Prove that G *is* the union of the conjugates of H. Thus, the finiteness hypothesis in Exercise 1.17 is necessary. (Hint: Equivalently, prove that every 2×2 matrix is conjugate to a matrix in H. You will use the fact that $\mathbb{C}$ is algebraically closed; see Example III.4.14.)

1.21. ▷ Let H, K be subgroups of a group G, with $H \subseteq N_G(K)$. Verify that the function $\gamma : H \to \mathrm{Aut}_{\mathsf{Grp}}(K)$ defined by conjugation is a homomorphism of groups and that $\ker \gamma = H \cap Z_G(K)$, where $Z_G(K)$ is the centralizer of K. [§1.4, 1.22]

1.22. ▷ Let G be a finite group, and let H be a cyclic subgroup of G of order p. Assume that p is the smallest prime dividing the order of G and that H is normal in G. Prove that H is contained in the center of G.

(Hint: By Exercise 1.21 there is a homomorphism $\gamma : G \to \mathrm{Aut}_{\mathsf{Grp}}(H)$; by Exercise II.4.14, $\mathrm{Aut}_{\mathsf{Grp}}(H)$ has order $p - 1$. What can you say about γ?) [§1.4]

---

2. The Sylow theorems

2.1. Cauchy's theorem. The 'Sylow theorems' consist of three statements concerning p-subgroups (cf. Definition 1.2) of a given finite group G. The form we will

give for the first of these statements will tell us that G contains p-groups of all sizes allowed by Lagrange's theorem: if p is a prime and p^k divides $|G|$, then G contains a subgroup of order p^k. The proof of this statement is an easy induction, provided the statement for $k = 1$ is known: that is, provided that one has established

Theorem 2.1 (Cauchy's theorem). *Let G be a finite group, and let p be a prime divisor of $|G|$. Then G contains an element of order p.*

As it happens, only the *abelian* version of this statement is needed for the proof of the first Sylow theorem; then the full statement of Cauchy's theorem follows from the first Sylow theorem itself. Since the (diligent) reader has already proved Cauchy's theorem for abelian groups (in Exercise II.8.17), we could directly move on to Sylow theorems.

However, there is a quick proof[4] of the full statement of Cauchy's theorem which does not rely on Sylow and is a good illustration of the power of the general 'class formula for arbitrary actions' (Proposition 1.1). We will present this proof, while also encouraging the reader to go back and (re)do Exercise II.8.17 now.

Proof of Theorem 2.1. Consider the set S of p-tuples of elements of G:

$$\{a_1, \dots, a_p\}$$

such that $a_1 \cdots a_p = e$. We claim that $|S| = |G|^{p-1}$: indeed, once $a_1, \dots, a_{p-1}$ are chosen (arbitrarily), then a_p is determined as it is the inverse of $a_1 \cdots a_{p-1}$.

Therefore, p divides the order of S as it divides the order of G.

Also note that if $a_1 \cdots a_p = e$, then

$$a_2 \cdots a_p a_1 = e$$

(even if G is not commutative): because if a_1 is a left-inverse to $a_2 \cdots a_p$, then it is also a right-inverse to it.

Therefore, we may act with the group $\mathbb{Z}/p\mathbb{Z}$ on S: given $[m]$ in $\mathbb{Z}/p\mathbb{Z}$, with $0 \leq m < p$, act by $[m]$ on

$$\{a_1, \dots, a_p\}$$

by sending it to

$$\{a_{m+1}, \dots, a_p, a_1, \dots, a_m\}:$$

as we just observed, this is still an element of S.

Now Corollary 1.3 implies

$$|Z| \equiv |S| \equiv 0 \mod p,$$

where Z is the set of fixed points of this action. Fixed points are p-tuples of the form

$$(*) \qquad \{a, \dots, a\};$$

and note that $Z \neq \emptyset$, since $\{e, \dots, e\} \in Z$. Since $p \geq 2$ and p divides $|Z|$, we conclude that $|Z| > 1$; therefore there exists some element in Z of the form (*), with $a \neq e$.

[4] This argument is apparently due to James McKay.

This says that there exists an $a \in G$, $a \neq e$, such that $a^p = e$, proving the statement. □

We should remark that the proof given here proves a more precise result than the raw statement of Theorem 2.1: every element of order p in G generates a cyclic subgroup of G of order p, and we are able to say something about the number of such subgroups.

Claim 2.2. *Let G be a finite group, let p be a prime divisor of $|G|$, and let N be the number of cyclic subgroups of G of order p. Then $N \equiv 1 \bmod p$.*

The proof of this fact is left to the reader (as an incentive to really understand the proof of Theorem 2.1).

Claim 2.2, coupled with the simple observation that if there is *only* 1 cyclic subgroup H of order p, then that subgroup must be normal (Exercise 2.2), suffices for interesting applications.

Definition 2.3. A group G is *simple* if its only normal subgroups are $\{e\}$ and G itself. ⌟

Simple groups occupy a special place in the theory of groups: one can 'break up' any finite group into basic constituents which are simple groups; we will see how this is done in §3.1. Thus, it is important to be able to tell whether a group is simple or not[5].

Example 2.4. Let p be a positive prime integer. If $|G| = mp$, with $1 < m < p$, then G is not simple.

Indeed, consider the subgroups of G with p elements. By Claim 2.2, the number of such subgroups is $\equiv 1 \bmod p$. Thus, if there is more than one such subgroup, then there must be at least $p + 1$. Any two distinct subgroups of prime order can only meet at the identity (why?); therefore this would account for at least

$$1 + (p + 1)(p - 1) = p^2$$

elements in G. Since $|G| = mp < p^2$, this is impossible. Therefore there is only one cyclic subgroup of order p in G, which must be normal as mentioned above, proving that G is not simple. ⌟

2.2. Sylow I. Let p be a prime integer. A *p-Sylow subgroup* of a finite group G is a subgroup of order p^r, where $|G| = p^r m$ and $(p, m) = 1$. That is, $P \subseteq G$ is a p-Sylow subgroup if it is a p-group and p does not divide $[G : P]$.

If p does *not* divide the order of $|G|$, then G contains a p-Sylow subgroup: namely, $\{e\}$. This is not very interesting; what is interesting is that G contains a p-Sylow subgroup even when p *does* divide the order of G:

Theorem 2.5 (First Sylow theorem)**.** *Every finite group contains a p-Sylow subgroup, for all primes p.*

[5]In fact, a complete list of all finite simple groups is known: this is the classification result mentioned at the end of §II.6.3, arguably one of the deepest and hardest results in mathematics.

The first Sylow theorem follows from the seemingly stronger statement:

Proposition 2.6. *If p^k divides the order of G, then G has a subgroup of order p^k.*

The statements are actually easily seen to be equivalent, by Exercise 1.8; in any case, the standard argument proving Theorem 2.5 proves Proposition 2.6, and we see no reason to hide this fact. Here is the argument:

Proof of Proposition 2.6. If $k = 0$, there is nothing to prove, so we may assume $k \geq 1$ and in particular that $|G|$ is a multiple of p.

Argue by induction on $|G|$: if $|G| = p$, again there is nothing to prove; if $|G| > p$ and G contains a proper subgroup H such that $[G : H]$ is relatively prime to p, then p^k divides the order of H, and hence H contains a subgroup of order p^k by the induction hypothesis, and thus so does G.

Therefore, we may assume that all proper subgroups of G have index divisible by p. By the class formula (Proposition 1.8), p divides the order of the center $Z(G)$. By Cauchy's theorem[6], $\exists a \in Z(G)$ such that a has order p. The cyclic subgroup $N = \langle a \rangle$ is contained in $Z(G)$, and hence it is normal in G (Exercise 1.4). Therefore we can consider the quotient G/N.

Since $|G/N| = |G|/p$ and p^k divides $|G|$ by hypothesis, we have that p^{k-1} divides the order of G/N. By the induction hypothesis, we may conclude that G/N contains a subgroup of order p^{k-1}. By the structure of the subgroups of a quotient (§II.8.3, especially Proposition II.8.9), this subgroup must be of the form P/N, for P a subgroup of G.

But then $|P| = |P/N| \cdot |N| = p^{k-1} \cdot p = p^k$, as needed. □

There are slicker ways to prove Theorem 2.5. We will see a pretty (and insightful) alternative in §2.3; but the proof given above is easy to remember and is a good template for similar arguments.

Remark 2.7. The diligent reader worked out in Exercise II.8.20 a *stronger* statement than Proposition 2.6, for *abelian* groups. The arguments are similar; the advantage in the abelian case is that any cyclic subgroup produced by Cauchy's theorem is automatically normal, while ensuring normality requires a few twists and turns in the general case (and, as a result, yields a weaker statement). ⌟

2.3. Sylow II. Theorem 2.5 tells us that *some* maximal p-group in G attains the largest size allowed by Lagrange's theorem, that is, the maximal power of the prime p dividing $|G|$.

One can be more precise: the second Sylow theorem tells us that *every* maximal p-group in $|G|$ is in fact a p-Sylow subgroup. It is as large as is allowed by Lagrange's theorem.

The situation is in fact even better: all p-Sylow subgroups are conjugates of each other[7]. Moreover, even better than this, *every* p-group inside G must be contained in a conjugate of any fixed p-Sylow subgroup.

[6]Note that, as mentioned in §2.1, we only need the *abelian* case of this theorem.

[7]Of course if P is a p-Sylow subgroup of G, then so are all conjugates gPg^{-1} of P.

The proof of this very precise result is very easy!

Theorem 2.8 (Second Sylow theorem). *Let G be a finite group, let P be a p-Sylow subgroup, and let $H \subseteq G$ be a p-group. Then H is contained in a conjugate of P: there exists $g \in G$ such that $H \subseteq gPg^{-1}$.*

Proof. Act with H on the set of left-cosets of P, by left-multiplication. Since there are $[G : P]$ cosets and p does not divide $[G : P]$, we know this action must have fixed points (Exercise 1.1): let gP be one of them. This means that $\forall h \in H$:

$$hgP = gP;$$

that is, $g^{-1}hgP = P$ for all h in H; that is, $g^{-1}Hg \subseteq P$; that is, $H \subseteq gPg^{-1}$, as needed. □

We can obtain an even more complete picture of the situation. Suppose we have constructed a chain

$$H_0 = \{e\} \subseteq H_1 \subseteq \cdots \subseteq H_k$$

of p-subgroups of a group G, where $|H_i| = p^i$. By Theorem 2.8 we know that H_k is contained in some p-Sylow subgroup, of order p^r = the maximum power of p dividing the order of G. But we claim that the chain can in fact be continued one step at a time all the way up to the Sylow subgroup:

$$H_0 = \{e\} \subseteq H_1 \subseteq \cdots \subseteq H_k \subseteq H_{k+1} \subseteq \cdots \subseteq H_r;$$

and, further, H_k may be assumed to be normal in H_{k+1}. The following lemma will simplify the proof of this fact considerably and will also help us prove the *third* Sylow theorem.

Lemma 2.9. *Let H be a p-group contained in a finite group G. Then*

$$[N_G(H) : H] \equiv [G : H] \mod p.$$

Proof. If H is trivial, then $N_G(H) = G$ and the two numbers are equal.

Assume then that H is nontrivial, and act with H on the set of left-cosets of H in G, by left-multiplication. The fixed points of this action are the cosets gH such that $\forall h \in H$

$$hgH = gH,$$

that is, such that $g^{-1}hg \in H$ for all $h \in H$; in other words, $H \subseteq gHg^{-1}$, and hence (by order considerations) $gHg^{-1} = H$. This means precisely that $g \in N_G(H)$. Therefore, the set of fixed points of the action consists of the set of cosets of H in $N_G(H)$.

The statement then follows immediately from Corollary 1.3. □

As a consequence, if H_k is not a *p-Sylow* subgroup 'already', in the sense that p 'still' divides $[G : H_k]$, then p must also divide $[N_G(H_k) : H_k]$. Another application of Cauchy's theorem tells us how to obtain the next subgroup H_{k+1} in the chain. More precisely, we have the following result.

Proposition 2.10. *Let H be a p-subgroup of a finite group G, and assume that H is* not *a p-Sylow subgroup. Then there exists a p-subgroup H' of G containing H, such that $[H' : H] = p$ and H is normal in H'.*

Proof. Since H is not a p-Sylow subgroup of G, p divides $[N_G(H) : H]$, by Lemma 2.9. Since H is normal in $N_G(H)$, we may consider the quotient group $N_G(H)/H$, and p divides the order of this group. By Theorem 2.1, $N_G(H)/H$ has an element of order p; this generates a subgroup of order p of $N_G(H)/H$, which must be (cf. §II.8.3) in the form H'/H for a subgroup H' of $N_G(H)$.

It is straightforward to verify that H' satisfies the stated requirements. □

The statement about 'chains of p-subgroups' follows immediately from this result.

Note that Cauchy's theorem and Proposition 2.10 provide a new proof of Proposition 2.6 and hence of the first Sylow theorem.

2.4. Sylow III. The third (and last) Sylow theorem gives a good handle on the number of p-Sylow subgroups of a given finite group G. This is especially useful in establishing the existence of normal subgroups of G: since all p-Sylow subgroups of a group are conjugates of each other (by the second Sylow theorem), if there is *only one* p-Sylow subgroup, then that subgroup must be normal[8].

Theorem 2.11. *Let p be a prime integer, and let G be a finite group of order $|G| = p^r m$. Assume that p does not divide m. Then the number of p-Sylow subgroups of G divides m and is congruent to 1 modulo p.*

Proof. Let N_p denote the number of p-Sylow subgroups of G.

By Theorem 2.8, the p-Sylow subgroups of G are the conjugates of any given p-Sylow subgroup P. By Lemma 1.13, N_p is the index of the normalizer $N_G(P)$ of P; thus (Corollary 1.14) it divides the index m of P. In fact,

$$m = [G : P] = [G : N_G(P)] \cdot [N_G(P) : P] = N_p \cdot [N_G(P) : P].$$

Now, by Lemma 2.9 we have

$$m = [G : P] \equiv [N_G(P) : P] \mod p;$$

multiplying by N_p, we get

$$mN_p \equiv m \mod p.$$

Since $m \not\equiv 0 \bmod p$ and p is prime, this implies

$$N_p \equiv 1 \mod p,$$

as needed. □

Of course there are other ways to prove Theorem 2.11: see for example Exercise 2.11.

[8]For an alternative viewpoint, see Exercise 2.2.

2.5. Applications. Consequences stemming from the group actions we have encountered, and especially the Sylow theorems, may be applied to establish exquisitely precise facts about individual groups as well as whole classes of groups; this is often based on some simple but clever numerology.

The following examples are exceedingly simple-minded but will hopefully convey the flavor of what can be done with the tools we have built in the previous two sections. More examples may be found among the exercises at the end of this section.

2.5.1. *More nonsimple groups.*

Claim 2.12. *Let G be a group of order mp^r, where p is a prime integer and $1 < m < p$. Then G is not simple.*

(Cf. Example 2.4.)

Proof. By the third Sylow theorem, the number N_p of p-Sylow subgroups divides m and is in the form $1 + kp$. Since $m < p$, this forces $k = 0$, $N_p = 1$. Therefore G has a normal subgroup of order p^r; hence it is not simple. □

Of course the same argument gives the same conclusion for every group of order mp^r, where $(m, p) = 1$ and the only divisor d of m such that $d \equiv 1 \mod p$ is $d = 1$.

Example 2.13. There are no simple groups of order 2002.

Indeed[9],

$$2002 = 2 \cdot 7 \cdot 11 \cdot 13;$$

the divisors of $2 \cdot 7 \cdot 13$ are

$$1\,,\ 2\,,\ 7\,,\ 13\,,\ 14\,,\ 26\,,\ 91\,,\ 182\,:$$

of these, only 1 is congruent to $1 \bmod 11$. Thus there is a normal subgroup of order 11 in every group of order 2002. ⌟

The reader should not expect the third Sylow theorem to always yield its fruits so readily, however.

Example 2.14. There are no simple groups of order 12.

Note that $3 \equiv 1 \bmod 2$ and $4 \equiv 1 \bmod 3$: thus the argument used above does not guarantee the existence of either a normal 2-Sylow subgroup or a normal 3-Sylow subgroup.

However, suppose that there is more than one 3-Sylow subgroup. Then there must be 4, by the third Sylow theorem. Since any two such subgroups must intersect in the identity, this accounts for exactly 8 elements of order 3. Excluding these leaves us with the identity and 3 elements of order 2 or 4; that is just enough room to fit *one* 2-Sylow subgroup. This subgroup will then have to be normal.

Thus, either there is a 3-Sylow normal subgroup or there is a 2-Sylow normal subgroup—either way, the group is not simple. ⌟

[9]It is safe to guess that this statement has been assigned on hundreds of algebra tests across the world in the year 2002.

Even this more refined counting will often fail, and one has to dig deeper.

Example 2.15. There are no simple groups of order 24.

Indeed, let G be a group of order 24, and consider its 2-Sylow subgroups; by the third Sylow theorem, there are either 1 or 3 such subgroups. If there is 1, the 2-Sylow subgroup is normal and G is not simple. Otherwise, G acts (nontrivially) by conjugation on this set of three 2-Sylow subgroups; this action gives a nontrivial homomorphism $G \to S_3$, whose kernel is a proper, nontrivial normal subgroup of G—thus again G is not simple. ⌟

The reader should practice by selecting a random number n and trying to say as much as he/she can, in general, about groups of order n. Beware: such problems are a common feature of qualifying exams.

2.5.2. *Groups of order pq, $p < q$ prime.*

Claim 2.16. *Assume $p < q$ are prime integers and $q \not\equiv 1 \bmod p$. Let G be a group of order pq. Then G is cyclic.*

Proof. By the third Sylow theorem, G has a unique (hence normal) subgroup H of order p. Indeed, the number N_p of p-Sylow subgroups must divide q, and q is prime, so $N_p = 1$ or q. Necessarily $N_p \equiv 1 \bmod p$, and $q \not\equiv 1 \bmod p$ by hypothesis; therefore $N_p = 1$.

Since H is normal, conjugation gives an action of G on H, hence (by Exercise 1.21) a homomorphism $\gamma : G \to \mathrm{Aut}(H)$. Now H is cyclic of order p, so $|\mathrm{Aut}(H)| = p - 1$ (Exercise 4.14); the order of $\gamma(G)$ must divide both pq and $p-1$, and it follows that γ is the trivial map.

Therefore, conjugation is *trivial* on H: that is, $H \subseteq Z(G)$. Lemma 1.5 implies that G is abelian.

Finally, an abelian group of order pq, with $p < q$ primes, is necessarily cyclic: indeed it must contain elements g, h of order p, q, respectively (for example by Cauchy's theorem), and then $|gh| = pq$ by Exercise II.1.14. □

For example, this statement 'classifies' all groups of order 15, 33, 35, 51, ...: such groups are necessarily cyclic.

The argument given in the proof is rather 'high-brow', as it involves the automorphism group of H; that is precisely why we gave it. For low-brow alternatives, see Exercise 2.18 or Remark 5.4.

The condition $q \not\equiv 1 \bmod p$ in Claim 2.16 is clearly necessary: indeed, $|S_3| = 2 \cdot 3$ is the product of two distinct primes, and yet S_3 is *not* cyclic. The argument given in the proof shows that if $G = |pq|$, with $p < q$ prime, and G has a normal subgroup of order p, then G is cyclic. If $q \equiv 1 \bmod p$, it can be shown that there is in fact a unique noncommutative group of order pq up to isomorphism: the reader will work this out after learning about semidirect products (Exercise 5.12). But we are in fact already in the position of obtaining rather sophisticated information about this group, even without knowing its construction in general (Exercise 2.19).

For fun, let's tackle the case in which $p = 2$.

Claim 2.17. *Let q be an odd prime, and let G be a noncommutative group of order $2q$. Then $G \cong D_{2q}$, the dihedral group.*

Proof. By Cauchy's theorem, $\exists y \in G$ such that y has order q. By the third Sylow theorem, $\langle y \rangle$ is the unique subgroup of order q in G (and is therefore normal). Since G is not commutative and in particular it is not cyclic, it has no elements of order $2q$; therefore, every element in the complement of $\langle y \rangle$ has order 2; let x be any such element.

The conjugate xyx^{-1} of y by x is an element of order q, so $xyx^{-1} \in \langle y \rangle$. Thus, $xyx^{-1} = y^r$ for some r between 0 and $q-1$.

Now observe that

$$(y^r)^r = (xyx^{-1})^r = xy^r x^{-1} = x^2 y (x^{-1})^2 = y$$

since $|x| = 2$. Therefore, $y^{r^2-1} = e$, which implies

$$q \mid (r^2 - 1) = (r-1)(r+1)$$

by Corollary II.1.11. Since q is prime, this says that $q \mid (r-1)$ or $q \mid (r+1)$; since $0 \leq r \leq q-1$, it follows that $r = 1$ or $r = q-1$.

If $r = 1$, then $xyx^{-1} = y$; that is, $xy = yx$. But then the order of xy is $2q$ (by Exercise II.1.14), and G is cyclic, a contradiction.

Therefore $r = q-1$, and we have established the relations

$$\begin{cases} x^2 = e, \\ y^q = e, \\ yx = xy^{q-1}. \end{cases}$$

These are the relations satisfied by generators x, y of D_{2q}, as the reader hopefully verified in Exercise II.2.5; the statement follows. □

Claim 2.17 yields a classification of groups of order $2q$, for q an odd prime: such a group must be either abelian (and hence cyclic, by the usual considerations) or isomorphic to a dihedral group. For $q = 3$, we recover the result of Exercise 1.13: every noncommutative group of order 6 is isomorphic to $D_6 \cong S_3$.

Exercises

2.1. ▷ Prove Claim 2.2. [§2.1]

2.2. ▷ Let G be a group. A subgroup H of G is *characteristic* if $\varphi(H) \subseteq H$ for every automorphism φ of G.

- Prove that characteristic subgroups are normal.
- Let $H \subseteq K \subseteq G$, with H characteristic in K and K normal in G. Prove that H is normal in G.
- Let G, K be groups, and assume that G contains a single subgroup H isomorphic to K. Prove that H is normal in G.

- Let K be a normal subgroup of a finite group G, and assume that $|K|$ and $|G/K|$ are relatively prime. Prove that K is characteristic in G.

[§2.1, §2.4, 2.13, §3.3]

2.3. Prove that a nonzero abelian group G is simple if and only if $G \cong \mathbb{Z}/p\mathbb{Z}$ for some positive prime integer p.

2.4. ▷ Prove that a group G is simple if and only if its only homomorphic images (i.e., groups G' such that there is an onto homomorphism $G \to G'$) are the trivial group and G itself (up to isomorphism). [§3.2]

2.5. Let G be a *simple* group, and assume $\varphi : G \to G'$ is a nontrivial group homomorphism. Prove that φ is injective.

2.6. Prove that there are no simple groups of order 4, 8, 9, 16, 25, 27, 32, or 49. In fact, prove that no p-group of order $\geq p^2$ is simple.

2.7. Prove that there are no simple groups of order 6, 10, 14, 15, 20, 21, 22, 26, 28, 33, 34, 35, 38, 39, 42, 44, 46, 51, 52, 55, 57, or 58. (Hint: Example 2.4.)

2.8. Let G be a finite group, p a prime integer, and let N be the intersection of the p-Sylow subgroups of G. Prove that N is a *normal* p-subgroup of G and that every normal p-subgroup of G is contained in N. (In other words, G/N is final with respect to the property of being a homomorphic image of G of order $|G|/p^\alpha$ for some α.)

2.9. ¬ Let P be a p-Sylow subgroup of a finite group G, and let $H \subseteq G$ be a p-subgroup. Assume $H \subseteq N_G(P)$. Prove that $H \subseteq P$. (Hint: P is normal in $N_G(P)$, so PH is a subgroup of $N_G(P)$ by Proposition II.8.11, and $|PH/P| = |H/(P \cap H)|$. Show that this implies that PH is a p-group, and hence $PH = P$ since P is a maximal p-subgroup of G. Deduce that $H \subseteq P$.) [2.10]

2.10. ¬ Let P be a p-Sylow subgroup of a finite group G, and act with P by conjugation on the set of p-Sylow subgroups of G. Show that P is the unique fixed point of this action. (Hint: Use Exercise 2.9.) [2.11]

2.11. ▷ Use the second Sylow theorem, Corollary 1.14, and Exercise 2.10 to paste together an alternative proof of the third Sylow theorem. [§2.4]

2.12. Let P be a p-Sylow subgroup of a finite group G, and let $H \subseteq G$ be a subgroup containing the normalizer $N_G(P)$. Prove that $[G : H] \equiv 1 \bmod p$.

2.13. ¬ Let P be a p-Sylow subgroup of a finite group G.

- Prove that if P is normal in G, then it is in fact characteristic in G (cf. Exercise 2.2).
- Let $H \subseteq G$ be a subgroup containing the Sylow subgroup P. Assume P is normal in H and H is normal in G. Prove that P is normal in G.
- Prove that $N_G(N_G(P)) = P$.

[3.12]

2.14. Prove that there are no simple groups of order 18, 40, 45, 50, or 54.

2.15. Classify all groups of order $n \leq 15$, $n \neq 8, 12$: that is, produce a list of nonisomorphic groups such that every group of order $n \neq 8, 12$, $n \leq 15$ is isomorphic to one group in the list.

2.16. ¬ Let G be a noncommutative group of order 8.

- Prove that G contains elements of order 4 and no elements of order 8.
- Let y be an element of order 4. Prove that G is generated by y and by an element $x \notin \langle y \rangle$, such that $x^2 = e$ or $x^2 = y^2$.
- In either case, $G = \{e, y, y^2, y^3, x, yx, y^2x, y^3x\}$. Prove that the multiplication table of G is determined by whether $x^2 = e$ or $x^2 = y^2$, and by the value of xy.
- Prove that necessarily $xy = y^3x$. (Hint: To eliminate $xy = y^2x$, multiply on the right by y.)
- Prove that $G \cong D_8$ or $G \cong Q_8$.

[6.2, VII.6.6]

2.17. ¬ Let R be a *division ring* (Definition III.1.13), and assume $|R| = 64$. Prove that R is necessarily commutative (hence, a field), as follows:

- The group of units of R has order 63. Prove it has a commutative subgroup G of order 9. (Sylow.)
- Prove that R is the only sub-division ring of R containing G.
- Prove that the set of elements of R commuting with every element of G is a sub-division ring of R containing G. (Cf. Exercise III.2.10.)
- Conclude that G is contained in the center of R. Recall that the center of R is a sub-division ring of R (cf. Exercise III.2.9), and conclude that R is commutative.

Like Exercise III.2.11, this is a particular case of a theorem of Wedderburn, according to which *every* finite division ring is a field. [VII.5.16]

2.18. ▷ Give an alternative proof of Claim 2.16 as follows: use the third Sylow theorem to count the number of elements of order p and q in G; use this to show that there are elements in G of order neither 1 nor p nor q; deduce that G is cyclic. [§2.5]

2.19. ▷ Let G be a noncommutative group of order pq, where $p < q$ are primes.

- Show that $q \equiv 1 \bmod p$.
- Show that the center of G is trivial.
- Draw the lattice of subgroups of G.
- Find the number of elements of each possible order in G.
- Find the number and size of the conjugacy classes in G.

[§2.5]

2.20. How many elements of order 7 are there in a simple group of order 168?

2.21. Let $p < q < r$ be prime integers, and let G be a group of order pqr. Prove that G is not simple.

2.22. Let G be a finite group, $n = |G|$, and p be a prime divisor of n. Assume that the only divisor of n that is congruent to 1 modulo p is 1. Prove that G is simple.

2.23. ¬ Let N_p denote the number of p-Sylow subgroups of a group G. Prove that if G is simple, then $|G|$ divides $N_p!$ for all primes p in the factorization of $|G|$. More generally, prove that if G is simple and H is a subgroup of G of index N, then $|G|$ divides $N!$. (Hint: Exercise II.9.12.) This problem capitalizes on the idea behind Example 2.15. [2.25]

2.24. ▷ Prove that there are no noncommutative simple groups of order less than 60. If you have sufficient stamina, prove that the next possible order for a noncommutative simple group is 168. (Don't feel too bad if you have to cheat and look up a few particularly troublesome orders > 60.) [§4.4]

2.25. ¬ Assume that G is a *simple* group of order 60.

- Use Sylow's theorems and simple numerology to prove that G has either five or fifteen 2-Sylow subgroups, accounting for fifteen elements of order 2 or 4. (Exercise 2.23 will likely be helpful.)
- If there are fifteen 2-Sylow subgroups, prove that there exists an element $g \in G$ of order 2 contained in at least two of them. Prove that the centralizer of g has index 5.

Conclude that every simple group[10] of order 60 contains a subgroup of index 5. [4.22]

---

3. Composition series and solvability

We have claimed that *simple* groups (in the sense of Definition 2.3) are the 'basic constituents' of all finite groups. Among other things, the material in this section will (partially) justify this claim.

3.1. The Jordan-Hölder theorem. A *series* of subgroups G_i of a group G is a decreasing sequence of subgroups starting from G:

$$G = G_0 \supsetneq G_1 \supsetneq G_2 \supsetneq \cdots .$$

The length of a series is the number of strict inclusions.

A series is *normal* if G_{i+1} is normal in G_i for all i. We will be interested in the *maximal length of a normal series* in G; if finite, we will denote this number[11] by $\ell(G)$. The number $\ell(G)$ is a measure of how far G is from being simple. Indeed, $\ell(G) = 0$ if and only if G is trivial, and $\ell(G) = 1$ if and only if G is nontrivial and *simple:* for a simple nontrivial group, the only maximal normal series is

$$G \supsetneq \{e\}.$$

[10]The reader will prove later (Exercise 4.22) that there is in fact *only one* simple group of order 60 up to isomorphism and that this group contains exactly five 2-Sylow subgroups. The result obtained here will be needed to establish this fact.

[11]There does not appear to be a standard notation for this concept.

Definition 3.1. A *composition series* for G is a normal series

$$G = G_0 \supsetneq G_1 \supsetneq G_2 \supsetneq \cdots \supsetneq G_n = \{e\}$$

such that the successive quotients G_i/G_{i+1} are simple. ⌟

It is clear (by induction on the order) that finite groups have composition series, while infinite groups do not necessarily have one (Exercise 3.3). It is also clear that if a normal series has maximal length $\ell(G)$, then it is a composition series. What is *not* clear is that the converse holds: conceivably, there could exist maximal normal series of different lengths (the longest ones having length $\ell(G)$). For example, why can't there be a finite group G with $\ell(G) = 3$ and two different composition series

$$G \supsetneq G_1 \supsetneq G_2 \supsetneq \{e\}$$

and

$$G \supsetneq G_1' \supsetneq \{e\}$$

(that is: a finite group G with $\ell(G) = 3$ and a simple normal subgroup G_1' such that G/G_1' is simple)?

Part of the content of the *Jordan-Hölder* theorem is that (luckily) this cannot happen. In fact, the theorem is much more precise: not only do all composition series have the same length, but they also have the same quotients (appearing, however, in possibly different orders).

Theorem 3.2 (Jordan-Hölder). *Let G be a group, and let*

$$G = G_0 \supsetneq G_1 \supsetneq G_2 \supsetneq \cdots \supsetneq G_n = \{e\},$$

$$G = G_0' \supsetneq G_1' \supsetneq G_2' \supsetneq \cdots \supsetneq G_m' = \{e\}$$

be two composition series for G. Then $m = n$, and the lists of quotient groups $H_i = G_i/G_{i+1}$, $H_i' = G_i'/G_{i+1}'$ agree (up to isomorphism) after a permutation of the indices.

Proof. Let

$$(*) \qquad G = G_0 \supsetneq G_1 \supsetneq G_2 \supsetneq \cdots \supsetneq G_n = \{e\}$$

be a composition series. Argue by induction on n: if $n = 0$, then G is trivial, and there is nothing to prove. Assume $n > 0$, and let

$$(**) \qquad G = G_0' \supsetneq G_1' \supsetneq G_2' \supsetneq \cdots \supsetneq G_m' = \{e\}$$

be another composition series for G. If $G_1 = G_1'$, then the result follows from the induction hypothesis, since G_1 has a composition series of length $n - 1 < n$.

We may then assume $G_1 \neq G_1'$. Note that $G_1 G_1' = G$: indeed, $G_1 G_1'$ is normal in G (Exercise 3.5), and $G_1 \subsetneq G_1 G_1'$; but there are no proper normal subgroups between G_1 and G since G/G_1 is simple.

Let $K = G_1 \cap G_1'$, and let

$$K \supsetneq K_1 \supsetneq K_2 \supsetneq \cdots \supsetneq K_r = \{e\}$$

be a composition series for K. By Proposition II.8.11 (the "second isomorphism theorem"),

$$\frac{G_1}{K} = \frac{G_1}{G_1 \cap G_1'} \cong \frac{G_1 G_1'}{G_1'} = \frac{G}{G_1'} \qquad \text{and} \qquad \frac{G_1'}{K} \cong \frac{G}{G_1}$$

are simple. Therefore, we have two new composition series for G:

$$\begin{array}{ccccccccccc} G & \supsetneq & G_1 & \supsetneq & K & \supsetneq & K_1 & \supsetneq & \cdots & \supsetneq & \{e\} \\ \| & & & & \| & & \| & & & & \| \\ G & \supsetneq & G_1' & \supsetneq & K & \supsetneq & K_1 & \supsetneq & \cdots & \supsetneq & \{e\} \end{array}$$

which only differ at the first step. These two series trivially have the same length and the same quotients (the first two quotients get switched from one series to the other).

Now we claim that the first of these two series has the same length and quotients as the series (*). Indeed,

$$G_1 \supsetneq K \supsetneq K_1 \supsetneq K_2 \supsetneq \cdots \supsetneq K_r = \{e\}$$

is a composition series for G_1: by the induction hypothesis, it must have the same length and quotients as the composition series

$$G_1 \supsetneq G_2 \supsetneq \cdots \supsetneq G_n = \{e\};$$

verifying our claim (and note that, in particular, $r = n - 2$).

By the same token, applying the induction hypothesis to the series of length $n - 1$, i.e.,

$$G_1' \supsetneq K \supsetneq K_1 \supsetneq K_2 \supsetneq \cdots \supsetneq K_{n-2} = \{e\},$$

shows that the second series has the same length and quotients as (**), and the statement follows. □

3.2. Composition factors; Schreier's theorem. Two normal series are *equivalent* if they have the same length and the same quotients (up to order). The Jordan-Hölder theorem shows that any two *maximal* finite series of a group are equivalent. That is, the (isomorphism classes of the) quotients of a composition series depend only on the group, not on the chosen series. These are the *composition factors* of the group. They form a multiset[12] of *simple* groups: the 'basic constituents' of our loose comment back in §2.

It is clear that two isomorphic groups must have the same composition factors. Unfortunately, it is not possible to reconstruct a group from its composition factors alone (Exercise 3.4). One has to take into account the way the simple groups are 'glued' together; we will come back to this point in §5.2.

The intuition that the composition factors of a group are its basic constituents is reinforced by the following fact: if G is a group with a composition series, then the composition factors of every normal subgroup N of G are composition factors of G and the remaining ones are the composition factors of the quotient G/N.

[12] See §I.2.2 for a reminder on *multisets:* they are sets of elements counted with multiplicity. For example, the composition factors of $\mathbb{Z}/4\mathbb{Z}$ form the multiset consisting of *two* copies of $\mathbb{Z}/2\mathbb{Z}$.

Example 3.3. Let $G = \mathbb{Z}/6\mathbb{Z} = \{[0], [1], [2], [3], [4], [5]\}$. Then

$$\{[0], [1], [2], [3], [4], [5]\} \supsetneq \{[0], [3]\} \supsetneq \{[0]\}$$

is a composition series for G; the quotients are $\mathbb{Z}/3\mathbb{Z}$, $\mathbb{Z}/2\mathbb{Z}$, respectively. The (normal) subgroup $N = \{[0], [2], [4]\}$ 'turns off' the second factor: indeed, intersecting the series with N gives

$$\{[0], [2], [4]\} \supsetneq \{[0]\} = \{[0]\},$$

a series with composition factor $\mathbb{Z}/3\mathbb{Z}$. On the other hand, 'mod-ing out by N' turns off the first factor: keeping in mind $[3] + N = [1] + N$, etc., we find

$$\{[0] + N, [1] + N\} = \{[0] + N, [1] + N\} \supsetneq \{[0] + N\},$$

a series with lone composition factor $\mathbb{Z}/2\mathbb{Z}$. ⌟

This phenomenon holds in complete generality:

Proposition 3.4. *Let G be a group, and let N be a normal subgroup of G. Then G has a composition series if and only if both N and G/N have composition series. Further, if this is the case, then*

$$\ell(G) = \ell(N) + \ell(G/N),$$

and the composition factors of G consist of the collection of composition factors of N and of G/N.

Proof. If G/N has a composition series, the subgroups appearing in it correspond to subgroups of G containing N, with isomorphic quotients, by Proposition II.8.10 (the "third isomorphism theorem"). Thus, if both G/N and N have composition series, juxtaposing them produces a composition series for G, with the stated consequence on composition factors.

The converse is a little trickier. Assume that G has a composition series

$$G = G_0 \supsetneq G_1 \supsetneq G_2 \supsetneq \cdots \supsetneq G_n = \{e\}$$

and that N is a normal subgroup of G. Intersecting the series with N gives a sequence of subgroups of the latter:

$$N = G \cap N \supseteq G_1 \cap N \supseteq \cdots \supseteq \{e\} \cap N = \{e\}$$

such that $G_{i+1} \cap N$ is normal in $G_i \cap N$, for all i. We claim that this becomes a composition series for N once repetitions are eliminated. Indeed, this follows once we establish that

$$\frac{G_i \cap N}{G_{i+1} \cap N}$$

is either trivial (so that $G_{i+1} \cap N = G_i \cap N$, and the corresponding inclusion may be omitted) or isomorphic to G_i/G_{i+1} (hence simple, and one of the composition factors of G). To see this, consider the homomorphism

$$G_i \cap N \hookrightarrow G_i \twoheadrightarrow \frac{G_i}{G_{i+1}} :$$

the kernel is clearly $G_{i+1} \cap N$; therefore (by the first isomorphism theorem) we have an injective homomorphisms

$$\frac{G_i \cap N}{G_{i+1} \cap N} \hookrightarrow \frac{G_i}{G_{i+1}}$$

identifying $(G_i \cap N)/(G_{i+1} \cap N)$ with a subgroup of G_i/G_{i+1}. Now, this subgroup is *normal* (because N is normal in G) and G_i/G_{i+1} is simple; our claim follows.

As for G/N, obtain a sequence of subgroups from a composition series for G:

$$\frac{G}{N} \supseteq \frac{G_1 N}{N} \supseteq \frac{G_2 N}{N} \supseteq \cdots \supseteq \frac{\{e_G\}N}{N} = \{e_{G/N}\},$$

such that $(G_{i+1}N)/N$ is normal in $(G_iN)/N$. As above, we have to check that

$$\frac{(G_iN)/N}{(G_{i+1}N)/N}$$

is either trivial or isomorphic to G_i/G_{i+1}. By the third isomorphism theorem, this quotient is isomorphic to $(G_iN)/(G_{i+1}N)$. This time, consider the homomorphism

$$G_i \hookrightarrow G_iN \twoheadrightarrow \frac{G_iN}{G_{i+1}N} :$$

this is *surjective* (check!), and the subgroup G_{i+1} of the source is sent to the identity element in the target; hence (by Theorem II.7.12) there is an onto homomorphism

$$\frac{G_i}{G_{i+1}} \twoheadrightarrow \frac{G_i + N}{G_{i+1} + N}.$$

Since G_i/G_{i+1} is simple, it follows that $(G_i + N)/(G_{i+1} + N)$ is either trivial or isomorphic to it (Exercise 2.4), as needed.

Summarizing, we have shown that if G has a composition series and N is normal in G, then both N and G/N have composition series. The first part of the argument yields the statement on lengths and composition factors, concluding the proof. □

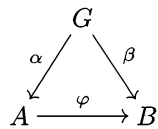
One nice consequence of the Jordan-Hölder theorem is the following observation. A series is a *refinement* of another series if all terms of the first appear in the second.

Proposition 3.5. *Any two normal series of a finite group ending with* $\{e\}$ *admit equivalent refinements.*

Proof. Refine the series to a composition series; then apply the Jordan-Hölder theorem. □

In fact, *Schreier's theorem* asserts that this holds for *all* groups (while the argument given here only works for groups admitting a composition series, e.g., finite groups). Proving this in general is reasonably straightforward, from judicious applications of the second isomorphism theorem (cf. Exercise 3.7).

3.3. The commutator subgroup, derived series, and solvability. It has been a while since we have encountered a universal object; here is one. For any group G, consider the category whose objects are group homomorphisms $\alpha : G \to A$ from G to a *commutative* group and whose morphisms $\alpha \to \beta$ are (as the reader should expect) commutative diagrams



where φ is a homomorphism.

Does this category have an initial object? That is, given a group G, does there exist a *commutative* group which is universal with respect to the property of being a homomorphic image of G?

Yes.

Such a group may well be thought of as the closest 'commutative approximation' of the given group G. To verify that this universal object exists, we introduce the following important notion. (The diligent reader has begun exploring this territory already, in Exercise II.7.11.)

Definition 3.6. Let G be a group. The *commutator* subgroup of G is the subgroup generated by all elements

$$ghg^{-1}h^{-1}$$

with $g, h \in G$. ⌟

The element $ghg^{-1}h^{-1}$ is often denoted $[g, h]$ and is called the *commutator* of g and h. Thus, g, h commute with each other if and only if $[g, h] = e$.

In the same notational style, the commutator subgroup of G should be denoted $[G, G]$; this is a bit heavy, and the common shorthand for it is G', which offers the possibility of iterating the notation. Thus, G'' may be used to denote the commutator subgroup *of the* commutator subgroup of G, and $G^{(i)}$ denotes the i-th iterate. We will adopt this notation in this subsection for convenience, but not elsewhere in this book (as we want to be able to 'prime' any letter we wish, for any reason).

First we record the following trivial, but useful, remark:

Lemma 3.7. *Let $\varphi : G_1 \to G_2$ be a group homomorphism. Then $\forall g, h \in G_1$ we have*

$$\varphi([g, h]) = [\varphi(g), \varphi(h)]$$

and $\varphi(G_1') \subseteq G_2'$.

This simple observation makes the key properties of the commutator subgroup essentially immediate (cf. Exercise II.7.11):

Proposition 3.8. *Let G' be the commutator subgroup of G. Then*

- *G' is normal in G;*
- *the quotient G/G' is commutative;*

- *if $\alpha : G \to A$ is a homomorphism of G to a commutative group, then $G' \subseteq \ker \alpha$;*
- *the natural projection $G \to G/G'$ is universal in the sense explained above.*

Proof. These are all easy consequences of Lemma 3.7:

—By Lemma 3.7, the commutator subgroup is *characteristic,* hence normal (cf. Exercise 2.2).

—By Lemma 3.7, the commutator of any two cosets gG', hG' is the coset of the commutator $[g,h]$; hence it is the identity in G/G'. As noted above, this implies that G/G' is commutative.

—Let $\alpha : G \to A$ be a homomorphism to a commutative group. By Lemma 3.7, $\alpha(G') \subseteq A' = \{e\}$: that is, $G' \subseteq \ker \alpha$.

—The universality follows from the previous point and from the universal property of quotients (Theorem II.7.12). □

Taking successive commutators of a group produces a descending sequence of subgroups,

$$G \supseteq G' \supseteq G'' \supseteq G''' \supseteq \cdots,$$

which is 'normal' in the sense indicated in §3.1.

Definition 3.9. Let G be a group. The *derived* series of G is the sequence of subgroups

$$G \supseteq G' \supseteq G'' \supseteq G''' \supseteq \cdots.$$ ⌟

The derived series may or may not end with the identity of G. For example, if G is commutative, then the sequence gets there right away:

$$G \supseteq G' = \{e\};$$

however, if G is simple and noncommutative, then it gets stuck at the first step:

$$G = G' = G'' = \cdots$$

(indeed, G' is normal and $\neq \{e\}$ as G is noncommutative; but then $G' = G$ since G is simple).

Definition 3.10. A group is *solvable* if its derived series terminates with the identity. ⌟

For example, abelian groups are solvable.

The importance of this notion will be most apparent in the relatively distant future because of a brilliant application of Galois theory (§VII.7.4). But we can already appreciate it in the way it relates to the material we just covered. A normal series is *abelian,* resp., *cyclic,* if all quotients are abelian, resp., cyclic[13].

Proposition 3.11. *For a finite group G, the following are equivalent:*

(i) All composition factors of G are cyclic.

(ii) G admits a cyclic series ending in $\{e\}$.

[13]Thus, a composition series should be called 'simple'; to our knowledge, it is not.

(iii) G admits an abelian series ending in $\{e\}$.

(iv) G is solvable.

Proof. (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii) are trivial. (iii) $\Longrightarrow$ (i) is obtained by refining an abelian series to a composition series (keeping in mind that the simple abelian groups are cyclic p-groups).

(iv) $\Longrightarrow$ (iii) is also trivial, since the derived series is abelian (by the second point in Proposition 3.8).

Thus, we only have to prove (iii) $\Longrightarrow$ (iv). For this, let

$$G = G_0 \supsetneq G_1 \supsetneq G_2 \supsetneq \cdots \supsetneq G_n = \{e\}$$

be an abelian series. Then we claim that $G^{(i)} \subseteq G_i$ for all i, where $G^{(i)}$ denotes the i-th 'iterated' commutator subgroup.

This can be verified by induction. For $i = 1$, G/G_1 is commutative; thus $G' \subseteq G_1$, by the third point in Proposition 3.8. Assuming we know $G^{(i)} \subseteq G_i$, the fact that G_i/G_{i+1} is abelian implies $G_i' \subseteq G_{i+1}$, and hence

$$G^{(i+1)} = (G^{(i)})' \subseteq G_i' \subseteq G_{i+1},$$

as claimed.

In particular we obtain that $G^{(n)} \subseteq G_n = \{e\}$: that is, the derived series terminates at $\{e\}$, as needed. □

Example 3.12. All p-groups are solvable. Indeed, the composition factors of a p-group are simple p-groups (what else could they be?), hence cyclic. ⌟

Corollary 3.13. *Let N be a normal subgroup of a group G. Then G is solvable if and only if both N and G/N are solvable.*

Proof. This follows immediately from Proposition 3.4 and the formulation of solvability in terms of composition factors given in Proposition 3.11. □

The *Feit-Thompson* theorem asserts that every finite group of *odd* order is solvable. This result is many orders of magnitude beyond the scope of this book: the original 1963 proof runs about 250 pages.

Exercises

3.1. Prove that $\mathbb{Z}$ has normal series of arbitrary lengths. (Thus, $\ell(\mathbb{Z})$ is not finite.)

3.2. Let G be a finite *cyclic* group. Compute $\ell(G)$ in terms of $|G|$. Generalize to finite solvable groups.

3.3. ▷ Prove that every finite group has a composition series. Prove that $\mathbb{Z}$ does not have a composition series. [§3.1]

3.4. ▷ Find an example of two nonisomorphic groups with the same composition factors. [§3.2]

3.5. ▷ Show that if H, K are *normal* subgroups of a group G, then HK is a normal subgroup of G. [§3.1]

3.6. Prove that $G_1 \times G_2$ has a composition series if and only if both G_1 and G_2 do, and explain how the corresponding composition factors are related.

3.7. ▷ Locate and understand a proof of (the general form of) Schreier's theorem that does not use the Jordan-Hölder theorem. Then obtain an alternative proof of the Jordan-Hölder theorem, using Schreier's. [§3.2]

3.8. ▷ Prove Lemma 3.7. [§3.3]

3.9. Let G be a nontrivial p-group. Construct explicitly an abelian series for G, using the fact that the center of a nontrivial p-group is nontrivial (Corollary 1.9). This gives an alternative proof of the fact that p-groups are solvable (Example 3.12).

3.10. ¬ Let G be a group. Define inductively an increasing sequence $Z_0 = \{e\} \subseteq Z_1 \subseteq Z_2 \subseteq \cdots$ of subgroups of G as follows: for $i \geq 1$, Z_i is the subgroup of G corresponding (as in Proposition II.8.9) to the center of G/Z_{i-1}.

- Prove that each Z_i is normal in G, so that this definition makes sense.

A group is[14] *nilpotent* if $Z_m = G$ for some m.

- Prove that G is nilpotent if and only if $G/Z(G)$ is nilpotent.
- Prove that p-groups are nilpotent.
- Prove that nilpotent groups are solvable.
- Find a solvable group that is not nilpotent.

[3.11, 3.12, 5.1]

3.11. ¬ Let H be a nontrivial normal subgroup of a nilpotent group G (cf. Exercise 3.10). Prove that H intersects $Z(G)$ nontrivially. (Hint: Let $r \geq 1$ be the smallest index such that $\exists h \neq e$, $h \in H \cap Z_r$. Contemplate a well-chosen commutator $[g, h]$.) Since p-groups are nilpotent, this strengthens the result of Exercise 1.9. [3.14]

[14] There are many alternative characterizations for this notion that are equivalent to the one given here but not too trivially so.

3.12. Let H be a proper subgroup of a finite nilpotent group G (cf. Exercise 3.10). Prove that $H \subsetneq N_G(H)$. (Hint: $Z(G)$ is nontrivial. First dispose of the case in which H does not contain $Z(G)$, and then use induction to deal with the case in which H does contain $Z(G)$.) Deduce that every Sylow subgroup of a finite nilpotent group is normal[15]. (Use Exercise 2.13.)

3.13. ¬ For a group G, let $G^{(i)}$ denote the iterated commutator, as in §3.3. Prove that each $G^{(i)}$ is characteristic (hence normal) in G. [3.14]

3.14. Let H be a nontrivial normal subgroup of a solvable group G.

- Prove that H contains a nontrivial *commutative* subgroup that is normal in G. (Hint: Let r be the largest index such that $K = H \cap G^{(r)}$ is nontrivial. Prove that K is commutative, and use Exercise 3.13 to show it is normal in G.)
- Find an example showing that H need not intersect the center of G nontrivially (cf. Exercise 3.11).

3.15. Let p, q be prime integers, and let G be a group of order p^2q. Prove that G is solvable. (This is a particular case of *Burnside's theorem:* for p, q primes, every group of order p^aq^b is solvable.)

3.16. ▷ Prove that every group of order < 120 and $\neq 60$ is solvable. [§4.4, §VII.7.4]

3.17. Prove that the Feit-Thompson theorem is equivalent to the assertion that every noncommutative finite simple group has even order.

---

4. The symmetric group

4.1. Cycle notation. It is time to give a second look at *symmetric groups.* Recall that S_n denotes the group of permutations (i.e., automorphisms in Set) of the set $\{\mathbf{1}, \ldots, \mathbf{n}\}$. In §II.2 we denoted elements of S_n in a straightforward but inconvenient way:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 8 & 1 & 2 & 7 & 5 & 3 & 4 & 6 \end{pmatrix}$$

would stand for the element in S_8 sending $\mathbf{1}$ to $\mathbf{8}$, $\mathbf{2}$ to $\mathbf{1}$, etc.

There is clearly "too much" information here (the first row should be implicit), and at the same time it seems hard to find out anything interesting about a permutation from this notation. For example, can the reader say anything about the conjugates of σ in S_8? For maximal enlightenment, try to do Exercise 4.1 now, and then try again after absorbing the material in §4.2. In short, we should be able to do better.

As often is the case, thinking in terms of actions helps. By its very definition, the group S_n acts on the set $\{\mathbf{1}, \ldots, \mathbf{n}\}$; so does every subgroup of S_n. Given a permutation $\sigma \in S_n$, consider the cyclic group $\langle\sigma\rangle$ generated by σ and its action on $\{\mathbf{1}, \ldots, \mathbf{n}\}$. The orbits of this action form a partition of $\{\mathbf{1}, \ldots, \mathbf{n}\}$; therefore, every

[15] This property characterizes finite nilpotent groups; cf. Exercise 5.1.

$\sigma \in S_n$ determines a partition of $\{\mathbf{1}, \ldots, \mathbf{n}\}$. For example, the element $\sigma \in S_8$ given above splits $\{\mathbf{1}, \ldots, \mathbf{8}\}$ into three orbits:

$$\{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{6}, \mathbf{8}\}, \quad \{\mathbf{4}, \mathbf{7}\}, \quad \{\mathbf{5}\}.$$

The action of $\langle\sigma\rangle$ is transitive on each orbit. This means that one can get from any element of the orbit to any other element and then back to the original one by applying σ enough times. In the example,

$$\mathbf{1} \mapsto \mathbf{8} \mapsto \mathbf{6} \mapsto \mathbf{3} \mapsto \mathbf{2} \mapsto \mathbf{1}, \quad \mathbf{4} \mapsto \mathbf{7} \mapsto \mathbf{4}, \quad \mathbf{5} \mapsto \mathbf{5}.$$

Definition 4.1. A (nontrivial) *cycle* is an element of S_n with exactly one nontrivial orbit. For distinct $a_1, \ldots, a_r$ in $\{\mathbf{1}, \ldots, \mathbf{n}\}$, the notation

$$(a_1 a_2 \ldots a_r)$$

denotes the cycle in S_n with nontrivial orbit $\{a_1, \ldots, a_r\}$, acting as

$$a_1 \mapsto a_2 \mapsto \cdots \mapsto a_r \mapsto a_1.$$

In this case, r is the *length* of the cycle. A cycle of length r is called an *r-cycle.* ⌟

The identity is considered a cycle of length 1 in a trivial way and is denoted by (1) (and could just as well be denoted by (i) for any i).

Note that $(a_1 a_2 \ldots a_r) = (a_2 \ldots a_r a_1)$ according to the notation introduced in Definition 4.1: the notation determines the cycle, but a nontrivial cycle only determines the notation 'up to a cyclic permutation'.

Two cycles are *disjoint* if their nontrivial orbits are. The following observation deserves to be highlighted, but it does not seem to deserve a proof:

Lemma 4.2. *Disjoint cycles commute.*

The next one gives us the alternative notation we were looking for.

Lemma 4.3. *Every $\sigma \in S_n$, $\sigma \neq e$, can be written as a product of disjoint nontrivial cycles, in a unique way up to permutations of the factors.*

Proof. As we have seen, every $\sigma \in S_n$ determines a partition of $\{\mathbf{1}, \ldots, \mathbf{n}\}$ into orbits under the action of $\langle\sigma\rangle$. If $\sigma \neq e$, then $\langle\sigma\rangle$ has nontrivial orbits. As σ acts as a cycle on each orbit, it follows that σ may be written as a product of cycles.

The proof of the uniqueness is left to the reader (Exercise 4.2). □

The *cycle notation* for $\sigma \in S_n$ is the (essentially) unique expression of σ as a product of disjoint cycles found in Lemma 4.3 (or (1) for $\sigma = e$). In our running example,

$$\sigma = (18632)(47),$$

and keep in mind that this expression is unique *cum grano salis:* we could write

$$\sigma = (63218)(47) = (74)(21863) = (32186)(74) = \cdots,$$

and all of these are 'the same' cycle notation for σ.

4.2. Type and conjugacy classes in S_n. The cycle notation has obviously annoying features—such as the not-too-unique uniqueness pointed out a moment ago, or the fact that

$$(123)$$

can be an element of S_3 just as well as of S_{100000}, and only the context can tell. However, it is invaluable as it gives easy access to quite a bit of important information on a given permutation. In fact, much of this information is carried already by something even simpler than the cycle decomposition.

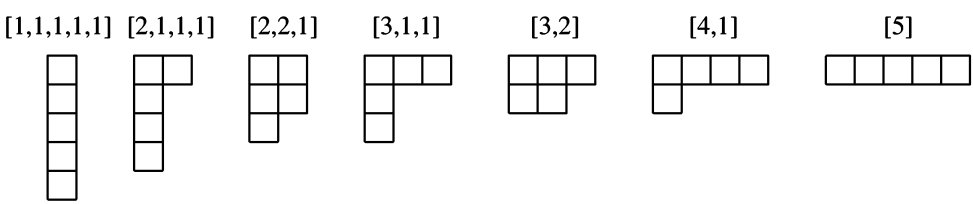
A *partition* of an integer $n > 0$ is a nonincreasing[16] sequence of positive integers whose sum is n. It is easy to enumerate partitions for small values of n. For example, 5 has 7 distinct partitions:

$$\begin{aligned} 5 &= 1+1+1+1+1 \\ &= 2+1+1+1 \\ &= 2+2+1 \\ &= 3+1+1 \\ &= 3+2 \\ &= 4+1 \\ &= 5. \end{aligned}$$

The partition $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$ may be denoted

$$[\lambda_1, \ldots, \lambda_r];$$

for example, the fourth partition listed above would be denoted $[3,1,1]$. A nicer 'visual' representation is by means of the corresponding *Young* (or *Ferrers*) *diagram*, obtained by stacking λ_1 boxes on top of λ_2 boxes on top of λ_3 boxes on top of.... For example, the diagrams corresponding to the seven partitions listed above are



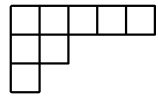
Definition 4.4. The *type* of $\sigma \in S_n$ is the partition of n given by the sizes of the orbits of the action of $\langle\sigma\rangle$ on $\{\mathbf{1}, \ldots, \mathbf{n}\}$. ⌟

It is hopefully clear (from the argument proving Lemma 4.3) that the type of $\sigma \in S_n$ is simply given by the lengths of the cycles in the decomposition of σ as the product of disjoint cycles, together with as many 1's as needed. In our running example,

$$\sigma = (18632)(47) \in S_8$$

has type $[5,2,1]$:

[16] Of course this choice is arbitrary, and nondecreasing sequences would do just as well.



The main reason why 'types' are introduced is a consequence of the following simple observation.

Lemma 4.5. *Let $\tau \in S_n$, and let $(a_1 \dots a_r)$ be a cycle. Then*

$$\tau(a_1 \dots a_r)\tau^{-1} = (a_1\tau^{-1} \dots a_r\tau^{-1}).$$

The funny notation $a_1\tau^{-1}$ stands for the action of the permutation τ^{-1} on $a_1 \in \{\mathbf{1}, \dots, \mathbf{n}\}$; recall that we agreed in §II.2.1 that we would let our permutations act *on the right*, for consistency with the usual notation for products in groups.

Proof. This is verified by checking that both sides act in the same way on $\{\mathbf{1}, \dots, \mathbf{n}\}$. For example, for $1 \leq i < r$

$$(a_i\tau^{-1})(\tau(a_1 \dots a_r)\tau^{-1}) = a_i(a_1 \dots a_r)\tau^{-1} = a_{i+1}\tau^{-1}$$

as it should; the other cases are left to the reader. □

By the usual trick of judiciously inserting identity factors $\tau^{-1}\tau$, this formula for computing conjugates extends immediately to any product of cycles:

$$\tau(a_1 \dots a_r) \cdots (b_1 \dots b_s)\tau^{-1} = (a_1\tau^{-1} \dots a_r\tau^{-1}) \cdots (b_1\tau^{-1} \dots b_s\tau^{-1}).$$

This holds whether the cycles are disjoint or not. However, since τ is a bijection, disjoint cycles remain disjoint after conjugation. This is essentially all there is to the following important observation:

Proposition 4.6. *Two elements of S_n are conjugate in S_n if and only if they have the same type.*

Proof. The 'only if' part of this statement follows immediately from the preceding considerations: conjugating a permutation yields a permutation of the same type. As for the 'if' part, suppose

$$\sigma_1 = (a_1 \dots a_r)(b_1 \dots b_s) \cdots (c_1 \dots c_t)$$

and

$$\sigma_2 = (a'_1 \dots a'_r)(b'_1 \dots b'_s) \cdots (c'_1 \dots c'_t)$$

are two permutations with the same type, written in cycle notation, with $r \geq s \geq \cdots \geq t$ (so the type is $[r, s, \dots, t]$). Let τ be any permutation such that $a_i = a'_i\tau$, $b_j = b'_j\tau$, $\dots$, $c_k = c'_k\tau$ for all i, j, $\dots$, k. Then Lemma 4.5 implies $\sigma_2 = \tau\sigma_1\tau^{-1}$, so σ_1 and σ_2 are conjugate, as needed. □

Example 4.7. In S_8,

$$(18632)(47) \quad \text{and} \quad (12345)(67)$$

must be conjugate, since they have the same type. The proof of Proposition 4.6 tells us that

$$\tau(18632)(47)\tau^{-1} = (12345)(67)$$

for

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 8 & 6 & 3 & 2 & 4 & 7 & 5 \end{pmatrix},$$

and of course this may be checked by hand in a second. Running this check, and especially staring at the second row in τ *vis-à-vis* the cycle notation of the first permutation, should clarify everything. ⌟

Summarizing, the *type* (or the corresponding Young diagram) tells us everything about conjugation in S_n.

Corollary 4.8. *The number of conjugacy classes in S_n equals the number of partitions of n.*

For example, there are 7 conjugacy classes in S_5, indexed by the Tetris look-alikes drawn above.

It is also reasonably straightforward to compute the number of elements in each conjugacy class, in terms of the type. For example, in order to count the number of permutations of type $[2,2,1]$ in S_5, note that there are $5! = 120$ ways to fill the corresponding Young diagram with the numbers $1, \ldots, 5$:

a_1	a_2
a_3	a_4
a_5	

that is, 120 ways to write a permutation as a product of two 2-cycles:

$$(a_1 a_2)(a_3 a_4);$$

but switching $a_1 \leftrightarrow a_2$ and $a_3 \leftrightarrow a_4$, as well as switching the two cycles, gives the same permutation. Therefore there are

$$\frac{120}{2 \cdot 2 \cdot 2} = 15$$

permutations of type $[2,2,1]$. Performing this computation for all Young diagrams for S_5 gives us the size of each conjugacy class, that is, the class formula (cf. §1.2) for S_5:

$$120 = 1 + 10 + 15 + 20 + 20 + 30 + 24.$$

Example 4.9. There are no normal subgroups of size 30 in S_5.

Indeed, normal subgroups are unions of conjugacy classes (§1.3); since the identity is in every subgroup and $30 - 1 = 29$ cannot be written as a sum of the numbers appearing in the class formula for S_5, there is no such subgroup. ⌟

This observation will be dwarfed by much stronger results that we will prove soon (such as Theorem 4.20, Corollary 4.21); but it is remarkable that such precise statements can be established with so little work.

4.3. Transpositions, parity, and the alternating group. For $n \geq 1$, consider the polynomials

$$\Delta_n = \prod_{1 \leq i < j \leq n} (x_i - x_j) \in \mathbb{Z}[x_1, \ldots, x_n],$$

that is,

$$\begin{aligned}
\Delta_1 &= 1, \\
\Delta_2 &= x_1 - x_2, \\
\Delta_3 &= (x_1 - x_2)(x_1 - x_3)(x_2 - x_3), \\
\Delta_4 &= (x_1 - x_2)(x_1 - x_3)(x_1 - x_4)(x_2 - x_3)(x_2 - x_4)(x_3 - x_4),
\end{aligned}$$

We can act with any $\sigma \in S_n$ on Δ_n, by permuting the indices according to σ:

$$\Delta_n \sigma := \prod_{1 \leq i < j \leq n} (x_{i\sigma} - x_{j\sigma}).$$

For example,

$$\Delta_4(1234) = (x_2 - x_3)(x_2 - x_4)(x_2 - x_1)(x_3 - x_4)(x_3 - x_1)(x_4 - x_1) = -\Delta_4.$$

In general, it is clear that $\Delta_n \sigma$ is still the product of all binomials $(x_i - x_j)$, where the factors are permuted and some factors may change sign in the process. Hence, $\Delta_n \sigma = \pm \Delta_n$, where the factor ± 1 depends on σ.

Definition 4.10. The *sign* of a permutation $\sigma \in S_n$, denoted $(-1)^\sigma$, is determined by the action of σ on Δ_n:

$$\Delta_n \sigma = (-1)^\sigma \Delta_n.$$

We say that a permutation is *even* if its sign is $+1$ and *odd* if its sign is -1. ⌟

Note that $\forall \sigma, \tau \in S_n$ we have

$$\Delta_n(\sigma\tau) = (\Delta_n \sigma)\tau :$$

it follows that $(-1)^{\sigma\tau} = (-1)^\sigma (-1)^\tau$. Viewing $\{-1, +1\}$ as a group under multiplication[17], we see that the 'sign' function

$$\epsilon : S_n \to \{-1, +1\}, \quad \epsilon(\sigma) := (-1)^\sigma$$

is a *homomorphism.*

Here is a different viewpoint on this sign function. A *transposition* is a cycle of length 2. Every permutation is a product of transpositions:

Lemma 4.11. *Transpositions generate S_n.*

Proof. Indeed, by Lemma 4.3 it suffices to show that every *cycle* is a product of transpositions, and indeed

$$(a_1 \ldots a_r) = (a_1 a_2)(a_1 a_3) \cdots (a_1 a_r),$$

as may be checked by applying[18] both sides to every element of $\{\mathbf{1}, \ldots, \mathbf{n}\}$. □

[17] This is the group of units in $\mathbb{Z}$; of course it is isomorphic to C_2.

[18] Don't forget that permutations act *on the right.*

Of course a given permutation may be written as a product of transpositions in many different ways. However, whether an *even* number of transpositions or an *odd* number is needed only depends on σ. Indeed,

Lemma 4.12. *Let $\sigma = \tau_1 \cdots \tau_r$ be a product of transpositions. Then σ is even, resp., odd, according to whether r is even, resp., odd.*

Proof. This follows immediately from the facts that ϵ is a homomorphism and the sign of a transposition is -1: indeed, (ij) acts on Δ_n by permuting its factors and changing the sign of exactly one factor (namely, the factor $\pm(x_i - x_j)$). □

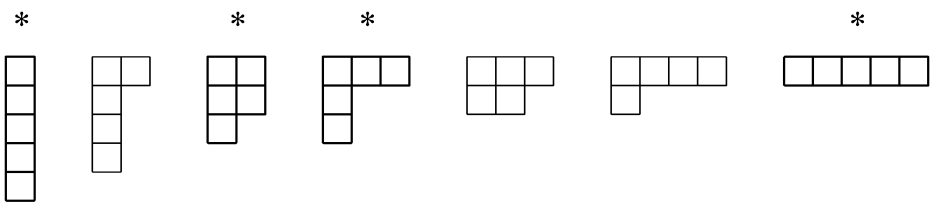
Definition 4.13. The *alternating group* on $\{\mathbf{1}, \ldots, \mathbf{n}\}$, denoted A_n, consists of all *even* permutations $\sigma \in S_n$. ⌟

The alternating group is a *normal* subgroup of S_n, and

$$[S_n : A_n] = 2$$

for $n \geq 2$: indeed, $A_n = \ker \epsilon$, and ϵ is surjective for $n \geq 2$.

It is very easy to tell whether a permutation σ belongs to A_n in terms of the type of σ. Indeed (by the argument proving Lemma 4.11) a cycle is *even*, resp., *odd*, if it has *odd*, resp., *even*, length[19]. Pictorially, a permutation $\sigma \in S_n$ is even if and only if n and the number of rows in the Young diagram of σ have the same parity. Take $n = 5$ for example:



The starred types correspond to even permutations. Adding up the sizes of the corresponding conjugacy classes,

$$1 + 15 + 20 + 24 = 60,$$

confirms that A_5 has index 2 in S_5.

However, do not read too much into this computation: we are not claiming that these are the sizes of the conjugacy classes *in* A_5, only that these are the conjugacy classes *in* S_5 making up the normal subgroup A_5. Conjugacy in A_n is rather interesting and ties in nicely with the issue of the *simplicity* of A_n.

4.4. Conjugacy in A_n; simplicity of A_n and solvability of S_n. Denote by $[\sigma]_{S_n}$, resp., $[\sigma]_{A_n}$, the conjugacy class of an even permutation σ in S_n, resp., A_n. Clearly $[\sigma]_{A_n} \subseteq [\sigma]_{S_n}$; we proceed to compare these two sets.

Lemma 4.14. *Let $n \geq 2$, and let $\sigma \in A_n$. Then $[\sigma]_{A_n} = [\sigma]_{S_n}$ or the size of $[\sigma]_{A_n}$ is half the size of $[\sigma]_{S_n}$, according to whether the centralizer $Z_{S_n}(\sigma)$ is not or is contained in A_n.*

[19] Note the unfortunate terminology clash. For example, 3-cycles are *even* permutations.

Proof. (Cf. Exercise 1.16.) Note that

$$Z_{A_n}(\sigma) = A_n \cap Z_{S_n}(\sigma) :$$

this follows immediately from the definition of centralizer (Definition 1.6). Now recall that the centralizer of σ is its stabilizer under conjugation, and therefore the size of the conjugacy class of σ equals the index of its centralizer.

If $Z_{S_n}(\sigma) \subseteq A_n$, then $Z_{A_n}(\sigma) = Z_{S_n}(\sigma)$, so that

$$[S_n : Z_{S_n}(\sigma)] = [S_n : Z_{A_n}(\sigma)] = [S_n : A_n][A_n : Z_{A_n}(\sigma)] = 2 \cdot [A_n : Z_{A_n}(\sigma)];$$

therefore, $[\sigma]_{A_n}$ is half the size of $[\sigma]_{S_n}$ in this case.

If $Z_{S_n}(\sigma) \not\subseteq A_n$, then note that $A_n Z_{S_n}(\sigma) = S_n$: indeed, $A_n Z_{S_n}(\sigma)$ is a subgroup of S_n (because A_n is normal; cf. Proposition II.8.11), and it properly contains A_n, so it must equal S_n as A_n has index 2 in S_n. Applying the second isomorphism theorem (Proposition II.8.11) gives

$$[A_n : Z_{A_n}(\sigma)] = [A_n : A_n \cap Z_{S_n}(\sigma)] = [A_n Z_{S_n}(\sigma) : Z_{S_n}(\sigma)] = [S_n : Z_{S_n}(\sigma)],$$

so the classes have the same size. Since $[\sigma]_{A_n} \subseteq [\sigma]_{S_n}$ in any case, it follows that $[\sigma]_{A_n} = [\sigma]_{S_n}$, completing the proof. □

Therefore, conjugacy classes of even permutations either are preserved from S_n to A_n or they split into two distinct, equal-sized classes. We are now in a position to give precise conditions determining which happens when.

Proposition 4.15. *Let $\sigma \in A_n$, $n \geq 2$. Then the conjugacy class of σ in S_n splits into two conjugacy classes in A_n precisely if the type of σ consists of distinct odd numbers.*

Proof. By Lemma 4.14, we have to verify that $Z_{S_n}(\sigma)$ is contained in A_n precisely when the stated condition is satisfied; that is, we have to show that

$$\sigma = \tau\sigma\tau^{-1} \implies \tau \text{ is even}$$

precisely when the type of σ consists of distinct odd numbers.

Write σ in cycle notation (including cycles of length 1):

$$\sigma = (a_1 \dots a_\lambda)(b_1 \dots b_\mu) \cdots (c_1 \dots c_\nu),$$

and recall (Lemma 4.5) that

$$\tau\sigma\tau^{-1} = (a_1\tau^{-1} \dots a_\lambda\tau^{-1})(b_1\tau^{-1} \dots b_\mu\tau^{-1}) \cdots (c_1\tau^{-1} \dots c_\nu\tau^{-1}).$$

Assume that $\lambda, \mu \dots, \nu$ are odd and distinct. If $\tau\sigma\tau^{-1} = \sigma$, then conjugation by τ must preserve each cycle in σ, as all cycle lengths are distinct:

$$\tau(a_1 \dots a_\lambda)\tau^{-1} = (a_1 \dots a_\lambda), \quad \text{etc.},$$

that is,

$$(a_1\tau^{-1} \dots a_\lambda\tau^{-1}) = (a_1 \dots a_\lambda), \quad \text{etc.}$$

This means that τ acts as a cyclic permutation on (e.g.) $a_1, \dots, a_\lambda$ and therefore in the same way as a power of $(a_1 \dots a_\lambda)$. It follows that

$$\tau = (a_1 \dots a_\lambda)^r (b_1 \dots b_\mu)^s \cdots (c_1 \dots c_\nu)^t$$

for suitable $r, s, \ldots, t$. Since all cycles have odd lengths, each cycle is an even permutation; and τ must then be even as it is a product of even permutations. This proves that $Z_{S_n}(\sigma) \subseteq A_n$ if the stated condition holds.

Conversely, assume that the stated condition does *not* hold: that is, either some of the cycles in the cycle decomposition have even length or all have odd length but two of the cycles have the same length.

In the first case, let τ be an even-length cycle in the cycle decomposition of σ. Note that $\tau\sigma\tau^{-1} = \sigma$: indeed, τ commutes with itself and with all cycles in σ other than τ. Since τ has even length, then it is odd as a permutation: this shows that $Z_{S_n}(\sigma) \not\subseteq A_n$, as needed.

In the second case, without loss of generality assume $\lambda = \mu$, and consider the odd permutation

$$\tau = (a_1 b_1)(a_2 b_2) \cdots (a_\lambda b_\lambda) :$$

conjugating by τ simply interchanges the first two cycles in σ; hence $\tau\sigma\tau^{-1} = \sigma$. As τ is odd, this again shows that $Z_{S_n}(\sigma) \not\subseteq A_n$, and we are done. □

Example 4.16. Looking again at A_5, we have noted in §4.3 that the types of the even permutations in S_5 are $[1,1,1,1,1]$, $[2,2,1]$, $[3,1,1]$, and $[5]$. By Proposition 4.15 the conjugacy classes corresponding to the first three types are preserved in A_5, while the last one splits.

Therefore there are exactly 5 conjugacy classes in A_5, and the class formula for A_5 is

$$60 = 1 + 15 + 20 + 12 + 12.$$ ⌟

Finally! We can now complete a circle of thought begun in the first section of this chapter. Along the way, the reader has hopefully checked that every simple group of order < 60 is commutative (Exercise 2.24); and we now see why 60 is special:

Corollary 4.17. *The alternating group A_5 is a simple noncommutative group of order 60.*

Proof. A normal subgroup of A_5 is necessarily the union of conjugacy classes, contains the identity, and has order equal to a divisor of 60 (by Lagrange's theorem). The divisors of 60 other than 1 and 60 are

$$2\,,\ 3\,,\ 4\,,\ 5\,,\ 6\,,\ 10\,,\ 12\,,\ 15\,,\ 20\,,\ 30;$$

counting the elements other than the identity would give one of

$$1\,,\ 2\,,\ 3\,,\ 4\,,\ 5\,,\ 9\,,\ 11\,,\ 14\,,\ 19\,,\ 29$$

as a sum of numbers $\neq 1$ from the class formula for A_5. But this simply does not happen. □

The reader will check that A_6 is simple, by the same method (Exercise 4.21). It is in fact the case that *all* groups A_n, $n \geq 5$, are simple (and noncommutative), implying that S_n is not solvable for $n \geq 5$, and these facts are rather important in view of applications to Galois theory (cf., e.g., Corollary VII.7.16). Note that A_2 is trivial, $A_3 \cong \mathbb{Z}/3\mathbb{Z}$ is simple and abelian, and A_4 is *not* simple (Exercise 2.24).

The alternating group A_5 is also called the *icosahedral (rotation) group:* indeed, it is the group of symmetries of an icosahedron obtained through rigid motions. (Can the reader verify[20] this fact?)

The simplicity of A_n for $n \geq 5$ may be established by studying 3-*cycles.* First of all, it is natural to wonder whether every even permutation may be written as a product of 3-cycles, and this is indeed so:

Lemma 4.18. *The alternating group A_n is generated by 3-cycles.*

Proof. Since every even permutation is a product of an even number of 2-cycles, it suffices to show that every product of *two* 2-cycles may be written as product of 3-cycles. Therefore, consider a product

$$(ab)(cd)$$

with $a \neq b$, $c \neq d$. If $(ab) = (cd)$, then this product is the identity, and there is nothing to prove. If $\{a, b\}$, $\{c, d\}$ have exactly one element in common, then we may assume $c = a$ and observe

$$(ab)(ad) = (abd).$$

If $\{a, b\}$, $\{c, d\}$ are disjoint, then

$$(ab)(cd) = (abc)(adc),$$

and we are done. □

Now we can capitalize on our study of conjugacy in A_n:

Claim 4.19. *Let $n \geq 5$. If a* normal *subgroup of A_n contains a 3-cycle, then it contains all 3-cycles.*

Proof. Normal subgroups are unions of conjugacy classes, so we just need to verify that 3-cycles form a conjugacy class in A_n, for $n \geq 5$. But they do in S_n, and the type of a 3-cycle is $[3, 1, 1, \ldots]$ for $n \geq 5$; hence the conjugacy class does not split in A_n, by Proposition 4.15. □

The general statement now follows easily by tying up loose ends:

Theorem 4.20. *The alternating group A_n is simple for $n \geq 5$.*

Proof. We have already checked this for $n = 5$, and the reader has checked it for $n = 6$. For $n > 6$, let N be a nontrivial normal subgroup of A_n; we will show that necessarily $N = A_n$, by proving that N contains 3-cycles.

Let $\tau \in N$, $\tau \neq (1)$, and let $\sigma \in A_n$ be a 3-cycle. Since the center of A_n is trivial (Exercise 4.14) and 3-cycles generate A_n, we may assume that τ and σ do not commute, that is, the commutator

$$[\tau, \sigma] = \tau(\sigma\tau^{-1}\sigma^{-1}) = (\tau\sigma\tau^{-1})\sigma^{-1}$$

[20] It is good practice to start with smaller examples: for instance, the tetrahedral rotation group is isomorphic to A_4.

is not the identity. This element is in N (as is evident from the first expression, as N is normal) and is a product of two 3-cycles (as is evident from the second expression, since the conjugate of a 3-cycle is a 3-cycle).

Therefore, replacing τ by $[\tau, \sigma]$ if necessary, we may assume that $\tau \in N$ is a nonidentity permutation acting on ≤ 6 elements: that is, on a subset of a set $T \subseteq \{\mathbf{1}, \ldots, \mathbf{n}\}$ with $|T| = 6$. Now we may view A_6 as a subgroup of A_n, by letting it act on T. The subgroup $N \cap A_6$ of A_6 is then normal (because N is normal) and nontrivial (because $\tau \in N \cap A_6$ and $\tau \neq (1)$). Since A_6 is simple (Exercise 4.21), this implies $N \cap A_6 = A_6$. In particular, N contains 3-cycles.

By Claim 4.19, this implies that N contains *all* 3-cycles. By Lemma 4.18, it follows that $N = A_n$, as needed. □

Corollary 4.21. *For $n \geq 5$, the group S_n is not solvable.*

Proof. Since A_n is simple, the sequence

$$S_n \supsetneq A_n \supsetneq \{(1)\}$$

is a composition series for S_n. It follows that the composition factors of S_n are $\mathbb{Z}/2\mathbb{Z}$ and A_n. By Proposition 3.11, S_n is not solvable. □

In particular, S_5 is a nonsolvable group of order 120. This is in fact the smallest order of a nonsolvable group; cf. Exercise 3.16.

Exercises

4.1. ▷ Compute the number of elements in the conjugacy class of

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 8 & 1 & 2 & 7 & 5 & 3 & 4 & 6 \end{pmatrix}$$

in S_8. [§4.1]

4.2. ▷ Suppose

$$(a_1 \ldots a_r)(b_1 \ldots b_s) \cdots (c_1 \ldots c_t) = (d_1 \ldots d_u)(e_1 \ldots e_v) \cdots (f_1 \ldots f_w)$$

are two products of disjoint cycles. Prove that the factors agree up to order. (Hint: The two corresponding partitions of $\{\mathbf{1}, \ldots, \mathbf{n}\}$ must agree.) [§4.1]

4.3. Assume σ has type $[\lambda_1, \ldots, \lambda_r]$ and that the λ_i's are pairwise relatively prime. What is $|\sigma|$? What can you say about $|\sigma|$, without the additional hypothesis on the numbers λ_i?

4.4. Make sense of the 'Taylor series' of the infinite product

$$\frac{1}{(1-x)} \cdot \frac{1}{(1-x^2)} \cdot \frac{1}{(1-x^3)} \cdot \frac{1}{(1-x^4)} \cdot \frac{1}{(1-x^5)} \cdots \cdot .$$

Prove that the coefficient of x^n in this series is the number of partitions of n.

4.5. Find the class formula for S_n, $n \leq 6$.

4.6. Let N be a *normal* subgroup of S_4. Prove that $|N| = 1$, 4, 12, or 24.

4.7. ▷ Prove that S_n is generated by (12) and $(12\ldots n)$.

(Hint: It is enough to get all transpositions. What is the conjugate of (12) by $(12\ldots n)$?) [4.9, §VII.7.5]

4.8. ¬ For $n > 1$, prove that the subgroup H of S_n consisting of permutations fixing **1** is isomorphic to S_{n-1}. Prove that there are no proper subgroups of S_n properly containing H. [VII.7.17]

4.9. By Exercise 4.7, S_4 is generated by (12) and (1234). Prove that (13) and (1234) generate a copy of D_8 in S_4. Prove that every subgroup of S_4 of order 8 is conjugate to $\langle(13),(1234)\rangle$. Prove there are exactly 3 such subgroups. For all $n \geq 3$ prove that S_n contains a copy of the dihedral group D_{2n}, and find generators for it.

4.10. ¬ • Prove that there are exactly $(n-1)!$ n-cycles in S_n.

• More generally, find a formula for the size of the conjugacy class of a permutation of given type in S_n. [4.11]

4.11. Let p be a prime integer. Compute the number of p-Sylow subgroups of S_p. (Use Exercise 4.10.) Use this result and Sylow's third theorem to prove again the 'only if' implication in Wilson's theorem (cf. Exercise II.4.16.)

4.12. ▷ A subgroup G of S_n is *transitive* if the induced action of G on $\{\mathbf{1},\ldots,\mathbf{n}\}$ is transitive.

- Prove that if $G \subseteq S_n$ is transitive, then $|G|$ is a multiple of n.
- List the transitive subgroups of S_3.
- Prove that the following subgroups of S_4 are all transitive:
  - $\langle(1234)\rangle \cong C_4$ and its conjugates,
  - $\langle(12)(34),(13)(24)\rangle \cong C_2 \times C_2$,
  - $\langle(12)(34),(1234)\rangle \cong D_8$ and its conjugates,
  - A_4, and S_4.

  With a bit of stamina, you can prove that these are the *only* transitive subgroups of S_4.

[§VII.7.5]

4.13. (If you know about determinants.) Prove that the sign of a permutation σ, as defined in Definition 4.10, equals the determinant of the matrix M_σ defined in Exercise II.2.1.

4.14. ▷ Prove that the center of A_n is trivial for $n \geq 4$. [§4.4]

4.15. Justify the 'pictorial' recipe given in §4.3 to decide whether a permutation is even.

4.16. The number of conjugacy classes in A_n, $n \geq 2$, is (allegedly)

$$1, 3, 4, 5, 7, 9, 14, 18, 24, 31, 43, \ldots.$$

Check the first several numbers in this list by finding the class formulas for the corresponding alternating groups.

4.17. ▷ • Find the class formula for A_4.

• Use it to prove that A_4 has no subgroup of order 6. [§II.8.5]

4.18. For $n \geq 5$, let H be a proper subgroup of A_n. Prove that $[A_n : H] \geq n$. Prove that A_n does have a subgroup of index n for all $n \geq 3$.

4.19. Prove that there are no nontrivial actions of A_n on any set S with $|S| < n$. Construct[21] a nontrivial action of A_4 on a set S, $|S| = 3$. Is there a nontrivial action of A_4 on a set S with $|S| = 2$?

4.20. ¬ Find all fifteen elements of order 2 in A_5, and prove that A_5 has exactly five 2-Sylow subgroups. [4.22]

4.21. ▷ Prove that A_6 is simple, by using its class formula (as is done for A_5 in the proof of Corollary 4.17). [§4.4]

4.22. ¬ Verify that A_5 is the *only* simple group of order 60, up to isomorphism. (Hint: By Exercise 2.25, a simple group G of order 60 contains a subgroup of index 5. Use this fact to construct a homomorphism $G \to S_5$, and prove that the image of this homomorphism must be A_5.) Note that A_5 has exactly five 2-Sylow subgroups; cf. Exercise 4.20. Thus, the other possibility contemplated in Exercise 2.25 does not occur. [2.25]

---

5. Products of groups

We already know that products exist in Grp (see §II.3.4); here we analyze this notion further and explore variations on the same theme, with an eye towards the question of determining the information needed to reconstruct a group from its composition factors.

5.1. The direct product. Recall from §II.3.4 that the *(direct) product* of two groups H, K is the group supported on the set $H \times K$, with operation defined componentwise. We have checked (Proposition II.3.4) that the direct product satisfies the universal property defining products in the category Grp.

There are situations in which the direct product of two subgroups N, H of a group G may be realized as a subgroup of G. Recall (Proposition II.8.11) that if one of the subgroups is *normal*, then the subset NH of G is in fact a subgroup of G. The relation between NH and $N \times H$ depends on how N and H intersect in G, so we take a look at this intersection.

The 'commutator' $[A, B]$ of two subsets A, B of G (see §3.3) is the subgroup generated by all commutators $[a, b]$ with $a \in A$, $b \in B$.

Lemma 5.1. *Let N, H be normal subgroups of a group G. Then*

$$[N, H] \subseteq N \cap H.$$

[21] You can think algebraically if you want; if you prefer geometry, visualize pairs of opposite sides on a tetrahedron.

Proof. It suffices to verify this on generators; that is, it suffices to check that

$$[n,h] = n(hn^{-1}h^{-1}) = (nhn^{-1})h^{-1} \in N \cap H$$

for all $n \in N$, $h \in H$. But the first expression and the normality of N show that $[n,h] \in N$; the second expression and the normality of H show that $[n,H] \in H$. □

Corollary 5.2. *Let N, H be normal subgroups of a group G. Assume $N \cap H = \{e\}$. Then N, H commute with each other:*

$$(\forall n \in N)\,(\forall h \in H) \quad nh = hn.$$

Proof. By Lemma 5.1, $[N,H] = \{e\}$ if $N \cap H = \{e\}$; the result follows immediately. □

In fact, under the same hypothesis more is true:

Proposition 5.3. *Let N, H be normal subgroups of a group G, such that $N \cap H = \{e\}$. Then $NH \cong N \times H$.*

Proof. Consider the function

$$\varphi : N \times H \to NH$$

defined by $\varphi(n,h) = nh$. Under the stated hypothesis, φ is a group homomorphism: indeed

$$\begin{aligned} \varphi((n_1,h_1)\cdot(n_2,h_2)) &= \varphi((n_1n_2, h_1h_2)) \\ &= n_1n_2h_1h_2 \\ &= n_1h_1n_2h_2 \end{aligned}$$

since N, H commute by Corollary 5.2

$$= \varphi((n_1,h_1))\cdot\varphi((n_2,h_2)).$$

The homomorphism φ is surjective by definition of NH. To verify it is injective, consider its kernel:

$$\ker\varphi = \{(n,h) \in N \times H \mid nh = e\}.$$

If $nh = e$, then $n \in N$ and $n = h^{-1} \in H$; thus $n = e$ since $N \cap H = \{e\}$. Using the same token for h, we conclude $h = e$; hence $(n,h) =$ the identity in $N \times H$, proving φ is injective.

Thus φ is an isomorphism, as needed. □

Remark 5.4. This result gives an alternative argument for the proof of Claim 2.16: if $|G| = pq$, with $p < q$ prime integers, and G contains normal subgroups H, K of order p, q, respectively (as is the case if $q \not\equiv 1 \bmod p$, by Sylow), then $H \cap K = \{e\}$ necessarily, and then Proposition 5.3 shows $HK \cong H \times K$. As $|HK| = |G| = pq$, this proves $G \cong H \times K \cong \mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/q\mathbb{Z}$. Finally, $(1,1)$ has order pq in this group, so G is cyclic, with the same conclusion we obtained in Claim 2.16. ⌟

5.2. Exact sequences of groups; extension problem. Of course, the hypothesis that both subgroups N, H are normal is necessary for the result of Proposition 5.3: for example, the permutations (123) and (12) generate subgroups N, H of S_3 meeting only at $\{e\}$, and N is normal in S_3, but $S_3 = NH$ is *not* isomorphic to the direct product of N and H. It is natural to examine this more general situation.

Let N, H be subgroups of a group G, with N normal (but with no *a priori* assumptions on H) and such that $N \cap H = \{e\}$; assume $G = NH$. We are after a description of the structure of G in terms of the structure of N and H.

It is notationally convenient to use the language of *exact sequences,* introduced for modules in §III.7.1. A (short) exact sequence of groups is a sequence of groups and group homomorphisms

$$1 \longrightarrow N \xrightarrow{\varphi} G \xrightarrow{\psi} H \longrightarrow 1$$

where ψ is surjective and φ identifies N with $\ker\psi$. In other words (by the first isomorphism theorem), use φ to identify N with a subgroup of G; then the sequence is exact if N is normal in G and ψ induces an isomorphism $G/N \to H$.

The reader should pause a moment and check that if G, N, H are *abelian,* then this notion matches precisely the notion of short exact sequence of abelian groups (i.e., $\mathbb{Z}$-modules) from §III.7.1; a notational difference is that here the trivial group is denoted[22] '1' rather than '0'.

Of course there always is an exact sequence

$$1 \longrightarrow N \longrightarrow N \times H \longrightarrow H \longrightarrow 1 \ :$$

map $n \in N$ to (n, e_H) and $(n, h) \in N \times H$ to h. However, keep in mind that this is a very special case: to reiterate the example mentioned above, there also is an exact sequence

$$1 \longrightarrow C_3 \longrightarrow S_3 \longrightarrow C_2 \longrightarrow 1 \ ,$$

yet $S_3 \ncong C_3 \times C_2$.

Definition 5.5. Let N, H be groups. A group G is an *extension of H by N* if there is an exact sequence of groups

$$1 \longrightarrow N \longrightarrow G \longrightarrow H \longrightarrow 1 \ . \qquad \lrcorner$$

The *extension problem* aims to describe all extensions of two given groups, up to isomorphism. For example, there are two extensions of C_2 by C_3: namely $C_6 \cong C_3 \times C_2$ and S_3; we will soon be able to verify that, up to isomorphism, there are no other extensions.

The extension problem is the 'second half' of the classification problem: the first half consists of determining all simple groups, and the second half consists of figuring out how these can be put together to construct any group[23]. For example,

[22] This is not unreasonable, since groups are more often written 'multiplicatively' rather than additively, so the identity element is more likely to be denoted 1 rather than 0.

[23] As mentioned earlier, the first half has been settled, although the complexity of the work leading to its solution justifies some skepticism concerning the absolute correctness of the proof. The status of the second half is, as far as we know, (even) murkier.

if

$$G = G_0 \supsetneq G_1 \supsetneq G_2 \supsetneq G_3 \supsetneq G_4 = \{e\}$$

is a composition series, with (simple) quotients $H_i = G_i/G_{i+1}$, then G is an *extension of H_0 by an extension of H_1 by an extension of H_2 by H_3:* knowing the composition factors of G and the extension process, it should in principle be possible to reconstruct G.

We are going to 'solve' the extension problem *in the particular case in which H is also a subgroup of G*, intersecting N at $\{e\}$.

Definition 5.6. An exact sequence of groups

$$1 \longrightarrow N \longrightarrow G \longrightarrow H \longrightarrow 1$$

(or the corresponding extension) is said to *split* if H may be identified with a subgroup of G, so that $N \cap H = \{e\}$.

We encountered this terminology in §III.7.2 for modules, thus for *abelian* groups. Note that the notion examined there appears to be more restrictive than Definition 5.6, since it requires G to be isomorphic to a direct product $N \times H$. This apparent mismatch evaporates because of Proposition 5.3: in the abelian case, every split extension (according to Definition 5.6) is in fact a direct product.

Of course split extensions are anyway very special, since quotients of a group G are usually not isomorphic to subgroups of G, even in the abelian case (cf. Exercise 5.4).

Lemma 5.7. *Let N be a normal subgroup of a group G, and let H be a subgroup of G such that $G = NH$ and $N \cap H = \{e\}$. Then G is a split extension of H by N.*

Proof. We have to construct an exact sequence

$$1 \longrightarrow N \longrightarrow G \longrightarrow H \longrightarrow 1\ ;$$

we let $N \to G$ be the inclusion map, and we prove that $G/N \cong H$. For this, consider the composition

$$\alpha : H \hookrightarrow G \twoheadrightarrow G/N.$$

Then α is surjective: indeed, since $G = NH$, $\forall g \in G$ we have $g = nh$ for some $n \in N$ and $h \in H$, and then

$$gN = nhN = h(h^{-1}nh)N = hN = \alpha(h).$$

Further, $\ker \alpha = \{h \in H \mid hN = N\} = N \cap H = \{e\}$; therefore α is also injective, as needed. □

To recap, if in the situation of Lemma 5.7 we also require that H be *normal* in G, then G is necessarily isomorphic to the 'trivial' extension $N \times H$: this is what we have proved in Proposition 5.3. We are seeking to describe the extension 'even if' H is not normal in G.

5.3. Internal/semidirect products. The attentive reader should have noticed that the key to Proposition 5.3 is really Corollary 5.2: if both N and H are normal and $N \cap H = \{e\}$, then N and H commute with each other. This is what ultimately causes the extension NH to be trivial. Now, recall that as soon as N is normal, then every subgroup H of G acts on N by conjugation: in fact (cf. Exercise 1.21) conjugation determines a homomorphism

$$\gamma : H \to \mathrm{Aut}_{\mathsf{Grp}}(N), \quad h \mapsto \gamma_h.$$

(Explicitly, for $h \in H$ the automorphism $\gamma_h : N \to N$ acts by $\gamma_h(n) := hnh^{-1}$.) The subgroups H and N commute precisely when γ is *trivial.* Corollary 5.2 shows that if N and H are both normal and $N \cap H = \{e\}$, then γ is indeed trivial.

This is the crucial remark. The next several considerations may be summarized as follows: if N is normal in G, H is a subgroup of G, $N \cap H = \{e\}$ and $G = NH$, then the extension G of H by N may be reconstructed from the conjugation action $\gamma : H \to \mathrm{Aut}_{\mathsf{Grp}}(N)$. The reader is advised to stare at the following triviality, which is the motivating observation for the general discussion:

$$(*) \qquad (\forall n_1, n_2 \in N), (\forall h_1, h_2 \in H) \qquad n_1 h_1 n_2 h_2 = (n_1 (h_1 n_2 h_1^{-1}))\,(h_1 h_2).$$

This says that if we know the conjugation action of H on N, then we can recover the operation in G from this information and from the operations in N and H.

Here is the general discussion. It is natural to abstract the situation and begin with *any two groups N, H and an arbitrary homomorphism*[24]

$$\theta : H \to \mathrm{Aut}_{\mathsf{Grp}}(N), \quad h \mapsto \theta_h.$$

Define an operation $\bullet_\theta$ on the *set* $N \times H$ as follows: for $n_1, n_2 \in N$ and $h_1, h_2 \in H$, let

$$(n_1, h_1) \bullet_\theta (n_2, h_2) := (n_1 \theta_{h_1}(n_2), h_1 h_2).$$

This will look more reasonable once it is compared with (*)!

Lemma 5.8. *The resulting structure $(N \times H, \bullet_\theta)$ is a group, with identity element (e_N, e_H).*

Proof. The reader should carefully verify this. For example, inverses exist because

$$(n_1, h_1) \bullet_\theta (\theta_{h_1^{-1}}(n_1^{-1}), h_1^{-1}) = (n_1 \theta_{h_1}(\theta_{h_1^{-1}}(n_1^{-1})), h_1 h_1^{-1}) = (n_1 n_1^{-1}, e_H) = (e_N, e_H)$$

and similarly in the reverse order. □

Definition 5.9. The group $(N \times H, \bullet_\theta)$ is a *semidirect product* of N and H and is denoted by $N \rtimes_\theta H$. ⌟

For example, the ordinary direct product is a semidirect product and corresponds to $\theta =$ the trivial map. If the reader feels a little uneasy about giving one name (*semidirect product*) for a whole host of different groups supported on the Cartesian product, welcome to the club. In fact, it gets worse still: it is not

[24]The reason why we are not denoting the image of h by θ as $\theta(h)$ is that this is an automorphism of N and we dislike the notation $\theta(h)(n)$ for the image of $n \in N$ obtained by applying the automorphism corresponding to h. The alternative $\theta_h(n)$ looks a little easier to parse.

uncommon to omit 'θ' from the notation and simply write $N \rtimes H$ for a semidirect product[25].

In any case, the notation $\bullet_\theta$ is too heavy to carry around, so we generally revert back to the usual simple juxtaposition of elements in order to denote multiplication in $N \rtimes_\theta H$. The following proposition checks that semidirect products are split extensions:

Proposition 5.10. *Let N, H be groups, and let $\theta : H \to \mathrm{Aut}_{\mathsf{Grp}}(N)$ be a homomorphism; let $G = N \rtimes_\theta H$ be the corresponding semidirect product. Then*

- *G contains isomorphic copies of N and H;*
- *the natural projection $G \to H$ is a surjective homomorphism, with kernel N; thus N is normal in G, and the sequence*

$$1 \longrightarrow N \longrightarrow N \rtimes_\theta H \longrightarrow H \longrightarrow 1$$

  is (split) exact;
- *$N \cap H = \{e_G\}$;*
- *$G = NH$;*
- *the homomorphism θ is realized by conjugation in G: that is, for $h \in H$ and $n \in N$ we have*

$$\theta_h(n) = hnh^{-1}$$

  in G.

Proof. The functions $N \to G$, $H \to G$ defined for $n \in N$, $h \in H$ by

$$n \mapsto (n, e_H), \quad h \mapsto (e_N, h)$$

are manifestly injective homomorphisms, allowing us to identify N, H with the corresponding subgroups of G. It is clear that $N \cap H = \{(e_N, e_H)\} = \{e_G\}$, and

$$(n, e_H) \bullet_\theta (e_N, h) = (n, h)$$

shows that $G = NH$.

The projection $G \to H$ defined by

$$(n, h) \mapsto h$$

is a surjective homomorphism, with kernel N; therefore N is normal in G. Finally,

$$(e_N, h) \bullet_\theta (n, e_H) \bullet_\theta (e_N, h)^{-1} = (\theta_h(n), h) \bullet_\theta (e_N, h^{-1}) = (\theta_h(n), e_H),$$

as claimed in the last point. □

Our original goal of 'reconstructing' a given split extension of a group H by a group N is a sort of converse to this proposition. More precisely,

[25] This is actually OK, if N and H are given as subgroups of a common group G, in which case the implicit action is just conjugation.

Proposition 5.11. *Let N, H be subgroups of a group G, with N normal in G. Assume that $N \cap H = \{e\}$, and $G = NH$. Let $\gamma : H \to \mathrm{Aut}_{\mathsf{Grp}}(N)$ be defined by conjugation: for $h \in H$, $n \in N$,*

$$\gamma_h(n) = hnh^{-1}.$$

Then $G \cong N \rtimes_\gamma H$.

Proof. Define a function

$$\varphi : N \rtimes_\gamma H \to G$$

by $\varphi(n,h) = nh$; this is clearly a bijection. We need to verify that φ is a homomorphism, and indeed $(\forall n \in N)$, $(\forall h \in H)$:

$$\begin{aligned}\varphi((n_1,h_1) \bullet_\gamma (n_2,h_2)) &= \varphi((n_1\gamma_{h_1}(n_2), h_1h_2)) \\ &= \varphi((n_1(h_1n_2h_1^{-1}), h_1h_2)) \\ &= n_1h_1n_2(h_1^{-1}h_1)h_2 = (n_1h_1)(n_2h_2) \\ &= \varphi((n_1,h_1))\varphi((n_2,h_2))\end{aligned}$$

as needed. □

When realized 'within' a group, as in the previous proposition, the semidirect product is sometime called an *internal* product.

Remark 5.12. If N and H commute, then the conjugation action of H on N is trivial; therefore γ is the trivial map, and the semidirect product $N \rtimes_\gamma H$ is the direct product $N \times H$. Thus, Proposition 5.11 recovers the result of Proposition 5.3 in this case. ⌟

Example 5.13. The automorphism group of C_3 is isomorphic to the cyclic group C_2: if $C_3 = \{e, y, y^2\}$, then the two automorphisms of C_3 are

$$\mathrm{id} : \begin{cases} e \mapsto e, \\ y \mapsto y, \\ y^2 \mapsto y^2, \end{cases} \qquad \sigma : \begin{cases} e \mapsto e, \\ y \mapsto y^2, \\ y^2 \mapsto y. \end{cases}$$

Therefore, there are two homomorphisms $C_2 \to \mathrm{Aut}_{\mathsf{Grp}}(C_3)$: the trivial map, and the isomorphism sending the identity to id and the nonidentity element to σ. The semidirect product corresponding to the trivial map is the direct product $C_3 \times C_2 \cong C_6$; the other semidirect product $C_3 \rtimes C_2$ is isomorphic to S_3. This can of course be checked by hand (and you should do it, for fun); but it also follows immediately from Proposition 5.11, since $N = \langle (123) \rangle$, $H = \langle (12) \rangle \subseteq S_3$ satisfy the hypotheses of this result. ⌟

The reader should contemplate carefully the slightly more general case of dihedral groups (Exercise 5.11); this enriches (and in a sense explains) the discussion presented in Claim 2.17.

In fact, semidirect products shed light on all groups of order pq, for $p < q$ primes; the reader should be able to complete the classification of these groups begun in §2.5.2 and show that if $q \equiv 1 \bmod p$, then there is exactly one such noncommutative group up to isomorphism (Exercise 5.12).

The reader would in fact be well-advised to try to use semidirect products to classify groups of small order: if a nontrivial normal subgroup N is found (typically by applying Sylow's theorems), with some luck the classification is reduced to the study of possible homomorphisms from known groups to $\mathrm{Aut}_{\mathsf{Grp}}(N)$ and can be carried out. See Exercise 5.15 for a further illustration of this technique.

Exercises

5.1. ▷ Let G be a finite group, and let $P_1, \ldots, P_r$ be its nontrivial Sylow subgroups. Assume all P_i are normal in G.

- Prove that $G \cong P_1 \times \cdots \times P_r$. (Induction on r; use Proposition 5.3.)
- Prove that G is nilpotent. (Hint: Mod out by the center, and work by induction on $|G|$. What is the center of a direct product of groups?)

Together with Exercise 3.10, this shows that a finite group is nilpotent if and only if each of its Sylow subgroups is normal. [3.12, §6.1]

5.2. Let G be an extension of H by N. Prove that the composition factors of G are the collection of the composition factors of H and those of N.

5.3. Let

$$G = G_0 \supsetneq G_1 \supsetneq \cdots \supsetneq G_r = \{e\}$$

be a normal series. Show how to 'connect' $\{e\}$ to G by means of r exact sequences of groups, involving only $\{e\}$, G, and the quotients $H_i = G_i/G_{i+1}$.

5.4. ▷ Prove that the sequence

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\cdot 2} \mathbb{Z} \longrightarrow \mathbb{Z}/2\mathbb{Z} \longrightarrow 0$$

is exact but does not split. [§5.2]

5.5. In Proposition III.7.5 we have seen that if an exact sequence

$$0 \longrightarrow M \xrightarrow{\varphi} N \longrightarrow N/(\varphi(M)) \longrightarrow 0$$

of *abelian* groups splits, then φ has a left-inverse. Is this necessarily the case for split sequences of *groups?*

5.6. Prove Lemma 5.8.

5.7. Let N be a group, and let $\alpha : N \to N$ be an automorphism of N. Prove that α may be realized as conjugation, in the sense that there exists a group G containing N as a normal subgroup and such that $\alpha(n) = gng^{-1}$ for some $g \in G$.

5.8. Prove that any semidirect product of two solvable groups is solvable. Show that semidirect products of nilpotent groups need not be nilpotent.

5.9. ▷ Prove that if $G = N \rtimes H$ is commutative, then $G \cong N \times H$. [§6.1]

5.10. Let N be a normal subgroup of a finite group G, and assume that $|N|$ and $|G/N|$ are relatively prime. Assume there is a subgroup H in G such that $|H| = |G/N|$. Prove that G is a semidirect product of N and H.

5.11. ▷ For all $n > 0$ express D_{2n} as a semidirect product $C_n \rtimes_\theta C_2$, finding θ explicitly. [§5.3]

5.12. ▷ Classify groups G of order pq, with $p < q$ prime: show that if $|G| = pq$, then either G is cyclic or $q \equiv 1 \bmod p$ and there is exactly one isomorphism class of noncommutative groups of order pq in this case. (You will likely have to use the fact that $\mathrm{Aut}_{\mathsf{Grp}}(C_q) \cong C_{q-1}$ if q is prime; cf. Exercise II.4.15.) [§2.5, §5.3]

5.13. ¬ Let $G = N \rtimes_\theta H$ be a semidirect product, and let K be the subgroup of G corresponding to $\ker\theta \subseteq H$. Prove that K is the kernel of the action of G on the set G/H of left-cosets of H. [5.14]

5.14. Recall that $S_3 \cong \mathrm{Aut}_{\mathsf{Grp}}(C_2 \times C_2)$ (Exercise II.4.13). Let ι be this isomorphism. Prove that $(C_2 \times C_2) \rtimes_\iota S_3 \cong S_4$. (Hint: Exercise 5.13.)

5.15. ▷ Let G be a group of order 28.

- Prove that G contains a normal subgroup N of order 7.
- Recall (or prove again) that, up to isomorphism, the only groups of order 4 are C_4 and $C_2 \times C_2$. Prove that there are two homomorphisms $C_4 \to \mathrm{Aut}_{\mathsf{Grp}}(N)$ and two homomorphisms $C_2 \times C_2 \to \mathrm{Aut}_{\mathsf{Grp}}(N)$ up to the choice of generators for the sources.
- Conclude that there are four groups of order 28 up to isomorphism: the two direct products $C_4 \times C_7$, $C_2 \times C_2 \times C_7$, and two noncommutative groups.
- Prove that the two noncommutative groups are D_{28} and $C_2 \times D_{14}$.

[§5.3]

5.16. Prove that the quaternionic group Q_8 (cf. Exercise III.1.12) cannot be written as a semidirect product of two nontrivial subgroups.

5.17. Prove that the multiplicative group $\mathbb{H}^*$ of nonzero quaternions (cf. Exercise III.1.12) is isomorphic to a semidirect product $\mathrm{SU}_2(\mathbb{C}) \rtimes \mathbb{R}^+$. (Hint: Exercise III.2.5.) Is this semidirect product in fact direct?

---

6. Finite abelian groups

We will end this chapter by treating in some detail the classification theorem for finite abelian groups mentioned in §II.6.3.

6.1. Classification of finite abelian groups. Now that we have acquired more familiarity with products, we are in a position to classify all finite *abelian* groups[26].

In due time (Proposition VI.2.11, Exercise VI.2.19) we will in fact be able to classify all *finitely generated* abelian groups: as mentioned in Example II.6.3, all

[26]Of course fancier *semidirect* products will not be needed here; cf. Exercise 5.9.

such groups are products of cyclic groups[27]. In particular, this is the case for finite abelian groups: this is what we prove in this section.

Since in this section we exclusively deal with abelian groups, we revert to the abelian style of notations: thus the operation will be denoted $+$; the identity will be 0; direct *products* will be called direct *sums* (and denoted $\oplus$); and so on.

First of all, we will congeal into an explicit statement a simple observation that has been with us in one form or another since at least as far back as[28] Exercise II.4.9.

Lemma 6.1. *Let G be an abelian group, and let H, K be subgroups such that $|H|$, $|K|$ are relatively prime. Then $H + K \cong H \oplus K$.*

Proof. By Lagrange's theorem (Corollary II.8.14), $H \cap K = \{0\}$. Since subgroups of abelian groups are automatically normal, the statement follows from Proposition 5.3. □

Now let G be a *finite* abelian group. For each prime p, the p-Sylow subgroup of G is unique, since it is automatically normal in G. Since the distinct nontrivial Sylow subgroups of G are p-groups for different primes p, Lemma 6.1 immediately implies the following result.

Corollary 6.2. *Every finite abelian group is the direct sum of its nontrivial Sylow subgroups.*

(The diligent reader knew already that this had to be the case, since abelian groups are nilpotent; cf. Exercise 5.1.) Thus, we already know that every finite abelian group is a direct sum of p-groups, and our main task amounts to classifying *abelian* p-groups for a fixed prime p. This is somewhat technical; we will get there by a seemingly roundabout path.

Lemma 6.3. *Let G be an abelian p-group, and let $g \in G$ be an element of maximal order. Then the exact sequence*

$$0 \longrightarrow \langle g \rangle \longrightarrow G \longrightarrow G/\langle g \rangle \longrightarrow 0$$

splits.

Put otherwise, there is a subgroup L of G such that L maps isomorphically to $G/\langle g \rangle$ via the canonical projection, that is, such that $\langle g \rangle \cap L = \{0\}$ and $\langle g \rangle + L = G$. Note that it will follow that $G \cong \langle g \rangle \oplus L$, by Proposition 5.3.

The main technicality needed in order to prove this lemma is the following particular case:

Lemma 6.4. *Let p be a prime integer and $r \geq 1$. Let G be a noncyclic abelian group of order p^{r+1}, and let $g \in G$ be an element of order p^r. Then there exists an element $h \in G$, $h \notin \langle g \rangle$, such that $|h| = p$.*

---

[27] The natural context to prove this more general result is that of modules over Euclidean rings or even principal ideal domains.

[28] In fact, this observation will really find its most appropriate resting place when we prove the *Chinese Remainder Theorem,* Theorem V.6.1.

Lemma 6.4 is a special case of Lemma 6.3 in the sense that, with notation as in the statement, necessarily $\langle h\rangle \cong G/\langle g\rangle$, and in fact $G \cong \langle g\rangle \oplus \langle h\rangle$ (and the reader is warmly encouraged to understand this before proceeding!). That is, we can split off the 'large' cyclic subgroup $\langle g\rangle$ as a direct summand of G, provided that G is not cyclic and not much larger than $\langle g\rangle$. Lemma 6.3 claims that this can be done whenever $\langle g\rangle$ is a maximal cyclic subgroup of G. We will be able to prove this more general statement easily once the particular case is settled.

Proof of Lemma 6.4. Denote $\langle g\rangle$ by K, and let h' be *any* element of G, $h' \notin K$. The subgroup K is normal in G since G is abelian; the quotient group G/K has order p. Since $h' \notin K$, the coset $h'+K$ has order p in G/K; that is, $ph' \in K$. Let $k = ph'$.

Note that $|k|$ divides p^r; hence it is a power of p. Also $|k| \neq p^r$, otherwise $|h'| = p^{r+1}$ and G would be cyclic, contrary to the hypothesis.

Therefore $|k| = p^s$ for some $s < r$; k generates a subgroup $\langle k\rangle$ of the cyclic group K, of order p^s. By Proposition II.6.11, $\langle k\rangle = \langle p^{r-s}g\rangle$. Since $s < r$, $\langle k\rangle \subseteq \langle pg\rangle$; thus, $k = mpg$ for some $m \in \mathbb{Z}$.

Then let $h = h' - mg$: $h \neq 0$ (since $h' \notin K$), and

$$ph = ph' - p(mg) = k - k = 0,$$

showing that $|h| = p$, as stated. □

Proof of Lemma 6.3. Argue by induction on the order of G; the case $|G| = p^0 = 1$ requires no proof. Thus we will assume that G is nontrivial and that the statement is true for every p-group smaller than G.

Let $g \in G$ be an element of maximal order, say p^r, and denote by K the subgroup $\langle g\rangle$ generated by g; this subgroup is normal, as G is abelian. If $G = K$, then the statement holds trivially. If not, G/K is a nontrivial p-group, and hence it contains an element of order p by Cauchy's theorem (Theorem 2.1). This element generates a subgroup of order p in G/K, corresponding to a subgroup G' of G of order p^{r+1}, containing K. This subgroup is not cyclic (otherwise the order of g is not maximal).

That is, we are in the situation of Lemma 6.4: hence we can conclude that there is an element $h \in G'$ (and hence $h \in G$) with $h \notin K$ and $|h| = p$. Let $H = \langle h\rangle \subseteq G$ be the subgroup generated by h, and note that $K \cap H = \{0\}$.

Now work modulo H. The quotient group G/H has smaller size than G, and $g+H$ generates a cyclic subgroup $K' = (K+H)/H \cong K/(K\cap H) \cong K$ of maximal order in G/H. By the induction hypothesis, there is a subgroup L' of G/H such that $K' + L' = G/H$ and $K' \cap L' = \{0_{G/H}\}$. This subgroup L' corresponds to a subgroup L of G containing H.

Now we claim that (i) $K + L = G$ and (ii) $K \cap L = \{0\}$. Indeed, we have the following:

(i) For any $a \in G$, there exist $mg + H \in K'$, $\ell + H \in L'$ such that $a + H = mg + \ell + H$ (since $K' + L' = G/H$). This implies $a - mg \in L$, and hence $a \in K + L$ as needed.

(ii) If $a \in K \cap L$, then $a + H \in K' \cap L' = \{0_{G/H}\}$, and hence $a \in H$. In particular, $a \in K \cap H = \{0\}$, forcing $a = 0$, as needed.

(i) and (ii) imply the lemma, as observed in the comments following the statement. □

Now we are ready to state the classification theorem; the proof is quite straightforward after all this preparation work. We first give the statement in a somewhat coarse form, as a corollary of the previous considerations:

Corollary 6.5. *Let G be a finite abelian group. Then G is a direct sum of cyclic groups, which may be assumed to be cyclic p-groups.*

Proof. As noted in Corollary 6.2, G is a direct sum of p-groups (as a consequence of the Sylow theorems). We claim that every abelian p-group P is a direct sum of cyclic p-groups.

To establish this, argue by induction on $|P|$. There is nothing to prove if P is trivial. If P is not trivial, let g be an element of P of maximal order. By Lemma 6.3

$$P = \langle g \rangle \oplus P'$$

for some subgroup P' of P; by the induction hypothesis P' is a direct sum of cyclic p-groups, concluding the proof. □

6.2. Invariant factors and elementary divisors. Here is a more precise version of the classification theorem. It is common to state the result in two equivalent forms.

Theorem 6.6. *Let G be a finite nontrivial abelian group. Then*

- *there exist prime integers $p_1, \ldots, p_r$ and positive integers n_{ij} such that $|G| = \prod_{i,j} p_i^{n_{i,j}}$ and*
$$G \cong \bigoplus_{i.j} \frac{\mathbb{Z}}{p_i^{n_{ij}}\mathbb{Z}};$$
- *there exist positive integers $1 < d_1 \mid \cdots \mid d_s$ such that $|G| = d_1 \cdots d_s$ and*
$$G \cong \frac{\mathbb{Z}}{d_1\mathbb{Z}} \oplus \cdots \oplus \frac{\mathbb{Z}}{d_s\mathbb{Z}}.$$

Further, these decompositions are uniquely determined by G.

The first form is nothing but a more explicit version of the statement of Corollary 6.5, so it has already been proven. We will explain how to obtain the second form from the first. The uniqueness statement[29] is left to the reader (Exercise 6.1).

The prime powers appearing in the first form of Theorem 6.6 are called the *elementary divisors* of G; the integers d_i appearing in the second form are called *invariant factors*. To go from elementary divisors to invariant factors, collect the

[29]Of course the 'uniqueness' statement only holds up to trivial manipulation such as a permutation of the factors. The claim is that the factors themselves are determined by G, in the sense that two direct sums of either form given in the statement are isomorphic only if their factors match.

elementary divisors in a table, listing (for example) prime powers according to increasing primes in the horizontal direction and decreasing exponents in the vertical direction; then the invariant factors are obtained as products of the factors in each row:

$d_r =$	$p_1^{n_{11}}$	$p_2^{n_{21}}$	$p_3^{n_{31}}$	$\cdots$
$d_{r-1} =$	$p_1^{n_{12}}$	$p_2^{n_{22}}$	$p_3^{n_{32}}$	$\cdots$
$d_{r-2} =$	$p_1^{n_{13}}$	$p_2^{n_{23}}$	$p_3^{n_{33}}$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Conversely, given the invariant factors d_i, obtain the rows of this table by factoring d_i into prime powers: the condition $d_1 \mid \cdots \mid d_r$ guarantees that these will be decreasing.

Repeated applications of Lemma 6.1 show that if $d = p_1^{n_1} \cdots p_r^{n_r}$ for *distinct* primes p_i and positive n_i (as is the case in each row of the table), then

$$\frac{\mathbb{Z}}{d\mathbb{Z}} \cong \frac{\mathbb{Z}}{p_1^{n_1}\mathbb{Z}} \oplus \cdots \oplus \frac{\mathbb{Z}}{p_r^{n_r}\mathbb{Z}},$$

proving that the two decompositions given in Theorem 6.6 are indeed equivalent.

This will likely be much clearer once the reader works through a few examples.

Example 6.7. Here are the two decompositions for a (random) group of order $29160 = 2^3 \cdot 3^6 \cdot 5$:

$$\left(\frac{\mathbb{Z}}{2\mathbb{Z}} \oplus \frac{\mathbb{Z}}{2\mathbb{Z}} \oplus \frac{\mathbb{Z}}{2\mathbb{Z}} \oplus \frac{\mathbb{Z}}{3\mathbb{Z}} \oplus \frac{\mathbb{Z}}{3\mathbb{Z}} \oplus \frac{\mathbb{Z}}{3^2\mathbb{Z}} \oplus \frac{\mathbb{Z}}{3^2\mathbb{Z}} \oplus \frac{\mathbb{Z}}{5\mathbb{Z}}\right) \cong \left(\frac{\mathbb{Z}}{3\mathbb{Z}} \oplus \frac{\mathbb{Z}}{6\mathbb{Z}} \oplus \frac{\mathbb{Z}}{18\mathbb{Z}} \oplus \frac{\mathbb{Z}}{90\mathbb{Z}}\right)$$

and here is the corresponding table of invariant factors/elementary divisors:

$90 =$	2	3^2	5
$18 =$	2	3^2	
$6 =$	2	3	
$3 =$		3	

Example 6.8. There are exactly 6 isomorphism classes of abelian groups of order 360. Indeed, $360 = 2^3 \cdot 3^2 \cdot 5$; the six possible tables of elementary divisors are shown below. In terms of invariant factors, the six distinct abelian groups of order 360 (up to isomorphism, by the uniqueness part of Theorem 6.6) are therefore

$$\frac{\mathbb{Z}}{360\mathbb{Z}}, \quad \frac{\mathbb{Z}}{2\mathbb{Z}} \oplus \frac{\mathbb{Z}}{180\mathbb{Z}}, \quad \frac{\mathbb{Z}}{2\mathbb{Z}} \oplus \frac{\mathbb{Z}}{2\mathbb{Z}} \oplus \frac{\mathbb{Z}}{90\mathbb{Z}},$$
$$\frac{\mathbb{Z}}{3\mathbb{Z}} \oplus \frac{\mathbb{Z}}{120\mathbb{Z}}, \quad \frac{\mathbb{Z}}{6\mathbb{Z}} \oplus \frac{\mathbb{Z}}{60\mathbb{Z}}, \quad \frac{\mathbb{Z}}{2\mathbb{Z}} \oplus \frac{\mathbb{Z}}{6\mathbb{Z}} \oplus \frac{\mathbb{Z}}{30\mathbb{Z}}.$$

360=	2^3	3^2	5

180=	2^2	3^2	5
2=	2		

90=	2	3^2	5
2=	2		
2=	2		

120=	2^3	3	5
3=		3	

60=	2^2	3	5
6=	2	3	

30=	2	3	5
6=	2	3	
2=	2		

6.3. Application: Finite subgroups of multiplicative groups of fields. Any classification theorem is useful in that it potentially reduces the proof of general facts to explicit verifications. Here is one example illustrating this strategy:

Lemma 6.9. *Let G be a finite abelian group, and assume that for every integer n the number of elements $g \in G$ such that $ng = 0$ is at most n. Then G is cyclic.*

The reader should try to prove this 'by hand', to appreciate the fact that it is not entirely trivial. It does become essentially immediate once we take the classification of finite abelian groups into account. Indeed, by Theorem 6.6

$$G \cong \frac{\mathbb{Z}}{(d_1)} \oplus \cdots \oplus \frac{\mathbb{Z}}{(d_s)}$$

for some positive integers $1 < d_1 \mid \cdots \mid d_s$. But if $s > 1$, then $|G| > d_s$ and $d_s g = 0$ for all $g \in G$ (so that the order of g divides d_s), contradicting the hypothesis. Therefore $s = 1$; that is, G is cyclic.

Lemma 6.9 is the key to a particularly nice proof of the following important fact, a weak form of which[30] we ran across back in Example II.4.6. Recall that the set F^* of nonzero elements of a field F is a commutative group under multiplication. Also recall (Example III.4.7) that a polynomial $f(x) \in F[x]$ is divisible by $(x - a)$ if and only if $f(a) = 0$; since a nonzero polynomial of degree n over a field can have at most n linear factors, this shows[31] that if $f(x) \in F[x]$ has degree n, then $f(a) = 0$ for at most n distinct elements $a \in F$.

Theorem 6.10. *Let F be a field, and let G be a finite subgroup of the multiplicative group $(F, \cdot)$. Then G is cyclic.*

Proof. By the considerations preceding the statement, for every n there are at most n elements $a \in F$ such that $a^n - 1 = 0$, that is, at most n elements $a \in G$ such that $a^n = 1$. Lemma 6.9 implies then that G is cyclic. □

[30] The diligent reader has proved that particular case in Exercise II.4.11. The proof hinted at in that exercise upgrades easily to the general case presented here. My point is not that the classification theorem is *necessary* in order to prove statements such as Theorem 6.10; our point is that it makes such statements nearly evident.

[31] Unique factorization in $F[x]$ is secretly needed here. We will deal with this issue more formally later; cf. Lemma V.5.1.

As a (very) particular case, the multiplicative group $((\mathbb{Z}/p\mathbb{Z})^*, \cdot)$ is cyclic: this is the fact pointed to in Example II.4.6.

Preview of coming attractions: Finitely generated (as opposed to just finite) abelian groups are also direct sums of cyclic groups. The only difference between the classification of finitely generated abelian groups and the classification of finite abelian groups explored here is the possible presence of a 'free' factor $\mathbb{Z}^{\oplus r}$ in the decomposition. The reader will first prove this fact in Exercise VI.2.19, as a consequence of 'Gaussian elimination over integral domains', and then recover it again as a particular case of the classification theorem for finitely generated modules over PIDs, Theorem VI.5.6. Neither Gaussian elimination nor the very general Theorem VI.5.6 are any harder to prove than the particular case of *finite abelian groups* laboriously worked out by hand in this section—a common benefit of finding the right general point of view is that, as a rule, proofs simplify. Technical work such as that performed in order to prove Lemma 6.3 is absorbed into the work necessary to build up the more general apparatus; the need for such technicalities evaporates in the process.

Exercises

6.1. ▷ Prove that the decomposition of a finite abelian group G as a direct sum of cyclic p-groups is unique. (Hint: The prime factorization of $|G|$ determines the primes, so it suffices to show that if

$$\frac{\mathbb{Z}}{p^{r_1}\mathbb{Z}} \oplus \cdots \oplus \frac{\mathbb{Z}}{p^{r_m}\mathbb{Z}} \cong \frac{\mathbb{Z}}{p^{s_1}\mathbb{Z}} \oplus \cdots \oplus \frac{\mathbb{Z}}{p^{s_n}\mathbb{Z}},$$

with $r_1 \geq \cdots \geq r_m$ and $s_1 \geq \cdots \geq s_n$, then $m = n$ and $r_i = s_i$ for all i. Do this by induction, by considering the group pG obtained as the image of the homomorphism $G \to G$ defined by $g \mapsto pg$.) [§6.2, §VI.5.3, VI.5.12]

6.2. Complete the classification of groups of order 8 (cf. Exercise 2.16).

6.3. Let G be a noncommutative group of order p^3, where p is a prime integer. Prove that $Z(G) \cong \mathbb{Z}/p\mathbb{Z}$ and $G/Z(G) \cong \mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$.

6.4. Classify abelian groups of order 400.

6.5. Let p be a prime integer. Prove that the number of distinct isomorphism classes of abelian groups of order p^r equals the number of partitions of the integer r.

6.6. ▷ How many *abelian* groups of order 1024 are there, up to isomorphism? [§II.6.3]

6.7. ¬ Let $p > 0$ be a prime integer, G a finite abelian group, and denote by $\rho : G \to G$ the homomorphism defined by $\rho(g) = pg$.

- Let A be a finite abelian group such that $pA = 0$. Prove that $A \cong \mathbb{Z}/p\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/p\mathbb{Z}$.
- Prove that $p \ker \rho$ and $p(\operatorname{coker} \rho)$ are both 0.

- Prove that $\ker \rho \cong \operatorname{coker} \rho$.
- Prove that every subgroup of G of order p is contained in $\ker \rho$ and that every subgroup of G of index p contains $\operatorname{im} \rho$.
- Prove that the number of subgroups of G of order p equals the number of subgroups of G of index p.

[6.8]

6.8. ¬ Let G be a finite abelian p-group, with elementary divisors $p^{n_1}, \ldots, p^{n_r}$ ($n_1 \geq n_2 \geq \cdots$). Prove that G has a subgroup H with invariant divisors $p^{m_1}, \ldots, p^{m_s}$ ($m_1 \geq m_2 \geq \cdots$) if and only if $s \leq r$ and $m_i \leq n_i$ for $i = 1, \ldots, s$. (Hint: One direction is immediate. For the other, with notation as in Exercise 6.7, compare $\ker \rho$ for H and G to establish $s \leq r$; this also proves the statement if all $n_i = 1$. For the general case use induction, noting that if $G \cong \bigoplus_i \mathbb{Z}/p^{n_i}\mathbb{Z}$, then $\rho(G) \cong \bigoplus_i \mathbb{Z}/p^{n_i-1}\mathbb{Z}$.)

Prove that the same description holds for the homomorphic images of G. [6.9]

6.9. Let H be a subgroup of a finite abelian group G. Prove that G contains a subgroup isomorphic to G/H. (Reduce to the case of p-groups; then use Exercise 6.8.) Show that both hypotheses 'finite' and 'abelian' are needed for this result. (Hint: Q_8 has a unique subgroup of order 2.)

6.10. The *dual* of a finite group G is the abelian group $G^\vee := \operatorname{Hom}_{\mathsf{Grp}}(G, \mathbb{C}^*)$, where $\mathbb{C}^*$ is the multiplicative group of $\mathbb{C}$.

- Prove that the image of every $\sigma \in G^\vee$ consists of *roots of* 1 in $\mathbb{C}$, that is, roots of polynomials $x^n - 1$ for some n.
- Prove that if G is a finite abelian group, then $G \cong G^\vee$. (Hint: First prove this for cyclic groups; then use the classification theorem to generalize to the arbitrary case.)

In Example VIII.6.5 we will encounter another notion of 'dual' of a group.

6.11. • Use the classification theorem for finite abelian groups (Theorem 6.6) to classify all finite modules over the ring $\mathbb{Z}/n\mathbb{Z}$.

• Prove that if p is prime, all finite modules over $\mathbb{Z}/p\mathbb{Z}$ are free[32].

6.12. Let G, H, K be finite abelian groups such that $G \oplus H \cong G \oplus K$. Prove that $H \cong K$.

6.13. ¬ Let G, H be finite abelian groups such that, for all positive integers n, G and H have the same number of elements of order n. Prove that $G \cong H$. (Note: The 'abelian' hypothesis is necessary! $C_4 \times C_4$ and $Q_8 \times C_2$ are nonisomorphic groups both with 1 element of order 1, 3 elements of order 2, and 12 elements of order 4.) [§II.4.3]

6.14. Let G be a finite abelian p-group, and assume G has only one subgroup of order p. Prove that G is cyclic. (This is in some sense a converse to Proposition II.6.11. You are welcome to try to prove it 'by hand', but use of the classification theorem will simplify the argument considerably.)

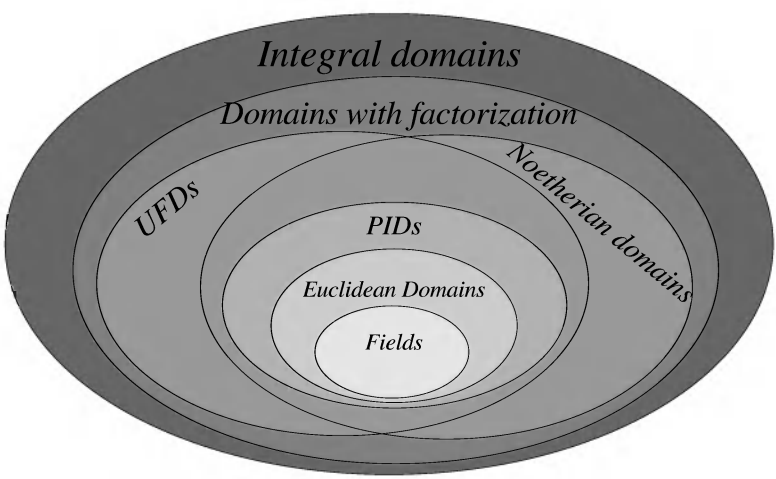
[32] As we will see in Proposition VI.4.10, this property characterizes fields.

6.15. Let G be a finite abelian group, and let $a \in G$ be an element of *maximal* order in G. Prove that the order of every $b \in G$ divides $|a|$. (This essentially reproduces the result of Exercise II.1.15.)

6.16. Let G be an abelian group of order n, and assume that G has at most one subgroup of order d for all $d \mid n$. Prove that G is cyclic.

Irreducibility and factorization in integral domains

We move our attention back to *rings* and analyze several useful classes of integral domains. One guiding theme in this chapter is the issue of *factorization:* we will address the problem of existence and uniqueness of factorizations of elements in a ring, abstracting good factorization properties of rings such as $\mathbb{Z}$ or $k[x]$ (for k a field) to whole classes of integral domains. The reader may want to associate the following picture with the first part of this chapter:



Blanket assumption: all rings considered in this chapter will be *commutative*[1]. In fact, most of the special classes of rings we will consider will be *integral domains,*

[1]Also, recall that all our rings have 1; cf. Definition III.1.1.

that is, commutative rings with 1 and with no nonzero zero-divisors (cf. Definition III.1.10).

1. Chain conditions and existence of factorizations

1.1. Noetherian rings revisited. Let R be a commutative ring. Recall that R is said to be *Noetherian* if every ideal of R is finitely generated (Definition III.4.2). In fact, this is a special case of the corresponding definition for *modules*: a module M over a ring R is Noetherian if every submodule of M is finitely generated (Definition III.6.6). In §III.6.4 we have verified that this condition is preserved through exact sequences: if M, N, P are R-modules and

$$0 \longrightarrow N \longrightarrow M \longrightarrow P \longrightarrow 0$$

is an exact sequence of R-modules, then M is Noetherian if and only if both N and P are Noetherian (Proposition III.6.7). An easy and useful consequence of this fact is that every finitely generated module over a Noetherian ring is Noetherian (Corollary III.6.8).

The Noetherian condition may be expressed in alternative ways, and it is useful to acquire some familiarity with them.

Proposition 1.1. *Let R be a commutative ring, and let M be an R-module. Then the following are equivalent:*

(1) M is Noetherian; that is, every submodule of M is finitely generated.

(2) Every ascending chain of submodules of M stabilizes; that is, if

$$N_1 \subseteq N_2 \subseteq N_3 \subseteq \cdots$$

is a chain of submodules of M, then $\exists i$ such that $N_i = N_{i+1} = N_{i+2} = \ldots$.

(3) Every nonempty family of submodules of M has a maximal element w.r.t. inclusion.

The second condition listed here is called the *ascending chain condition* (a.c.c.) for submodules. For $M = R$, Proposition 1.1 tells us (among other things) that a *ring* is Noetherian if and only if the ascending chain condition holds for its *ideals*.

Proof. (1) $\Longrightarrow$ (2): Assume that M is Noetherian, and let

$$N_1 \subseteq N_2 \subseteq N_3 \subseteq \cdots$$

be a chain of submodules of M. Consider the union

$$N = \bigcup_i N_i :$$

the reader will verify that N is a submodule of M. Since M is Noetherian, N is finitely generated, say $N = \langle n_1, \ldots, n_r \rangle$. Now $n_k \in N \implies n_k \in N_i$ for some i; by picking the largest such i, we see that $\exists i$ such that all $n_1, \cdots, n_r$ are contained in N_i. But then $N \subseteq N_i$, and since $N_i \subseteq N_{i+1} \subseteq \cdots$ are all contained in $N = N_i$, it follows that $N_i = N_{i+1} = N_{i+2} = \ldots$ as needed.

(2) $\Longrightarrow$ (3): Arguing contrapositively, assume that M admits a family $\mathscr{F}$ of submodules that does *not* have a maximal element. Construct an infinite ascending

chain as follows: let N_1 be any element of $\mathscr{F}$; since N_1 is not maximal in $\mathscr{F}$, there exists an element N_2 of $\mathscr{F}$ such that $N_1 \subsetneq N_2$; since N_3 is not maximal in $\mathscr{F}$, there exists an element N_3 of $\mathscr{F}$ such that $N_2 \subsetneq N_3$; etc. The chain

$$N_1 \subsetneq N_2 \subsetneq N_3 \subsetneq \cdots$$

does not stabilize, showing that (2) does not hold.

(3) $\Longrightarrow$ (1): Assume (3) holds, and let N be a submodule of M. Then the family $\mathscr{F}$ of *finitely generated* subsets of N is nonempty (as $(0) \in \mathscr{F}$); hence it has a maximal element N'. Say that $N' = \langle n_1, \ldots, n_r \rangle$. Now we claim that $N' = N$: indeed, let $n \in N$; the submodule $\langle n_1, \ldots, n_r, n \rangle$ is finitely generated, and therefore it is in $\mathscr{F}$; as it contains N' and N' is maximal, necessarily $\langle n_1, \ldots, n_r, n \rangle = N$; in particular $n \in N'$, as needed.

This shows that $N = N'$ is finitely generated, and since $N \subseteq M$ was arbitrary, this implies that M is Noetherian. □

Noetherian rings are a very useful and flexible class of rings. In §III.6.5 we mentioned the important fact that *every finite-type algebra over a Noetherian ring is Noetherian.* 'Finite-type (commutative) algebra' is just a fancy name for a quotient of a polynomial ring (§III.6.5), so this is what the fact states:

Theorem 1.2. *Let R be a Noetherian ring, and let J be an ideal of the polynomial ring $R[x_1, \ldots, x_n]$. Then the ring $R[x_1, \ldots, x_n]/J$ is Noetherian.*

Note that finite-type R algebras are (in general) very far from being finitely generated *as modules over* R (cf. again §III.6.5), so it would be foolish to expect them to be Noetherian *as R-modules.* The fact that they turn out to be Noetherian *as rings* (that is, as modules over themselves) provides us with a huge class of examples of Noetherian rings, among which are the rings of (classical) algebraic geometry and number theory. Thus, entire fields of mathematics are a little more manageable thanks to Theorem 1.2.

The proof of this deep fact is surprisingly easy. By Exercise 1.1, it suffices to prove that

$$R \text{ Noetherian} \implies R[x_1, \ldots, x_n] \text{ Noetherian};$$

and an immediate induction reduces the statement to the following particular case, which carries a distinguished name:

Lemma 1.3 (Hilbert's basis theorem). *R Noetherian $\Longrightarrow$ $R[x]$ Noetherian.*

Proof. Assume R is Noetherian, and let I be an ideal of $R[x]$. We have to prove that I is finitely generated.

Recall that if $f(x) = a_d x^d + a_{d-1} x^{d-1} + \cdots + a_0 \in R[x]$ and $a_d \neq 0$, then a_d is called the *leading coefficient* of $f(x)$. Consider the following subset of R:

$$A = \{0\} \cup \{a \in R \mid a \text{ is a leading coefficient of an element of } I\}.$$

It is clear that A is an *ideal* of R (Exercise 1.6); since R is Noetherian, A is finitely generated. Thus there exist elements $f_1(x), \ldots, f_r(x) \in I$ whose leading coefficients $a_1, \ldots, a_r$ generate A as an ideal of R.

Now let d_i be the degree of $f_i(x)$, and let d be the maximum among these degrees. Consider the sub-R-module

$$M = \langle 1, x, x^2, \dots, x^{d-1} \rangle \subseteq R[x],$$

that is, the R-module consisting of polynomials of degree $< d$. Since M is finitely generated as a module over R, it is Noetherian as an R-module (by Corollary III.6.8). Therefore, the submodule

$$M \cap I$$

of M is finitely generated over R, say by $g_1(x), \dots, g_s(x) \in I$.

Claim 1.4.

$$I = (f_1(x), \dots, f_r(x), g_1(x), \dots, g_s(x)).$$

This claim implies the statement of the theorem. To prove the claim, we only need to prove the $\subseteq$ inclusion; to this end, let $\alpha(x) \in I$ be an arbitrary polynomial in I. If $\deg \alpha(x) \geq d$, let a be the leading coefficient of $\alpha(x)$. Then $a \in A$, so $\exists b_1, \dots, b_r \in R$ such that

$$a = b_1 a_1 + \cdots + b_r a_r.$$

Letting $e = \deg \alpha(x)$, so that $e \geq d_i$ for all i, this says that

$$\alpha(x) - b_1 x^{e-d_1} f_1(x) - \cdots - b_r x^{e-d_r} f_r(x)$$

has degree $< e$. Iterating this procedure, we obtain a finite list of polynomials $\beta_1(x), \dots, \beta_r(x) \in R[x]$ such that

$$\alpha(x) - \beta_1(x) f_1(x) - \cdots - \beta_r(x) f_r(x)$$

has degree $< d$. But this places this element in $M \cap I$; therefore $\exists c_1, \dots, c_s \in R$ such that

$$\alpha(x) - \beta_1(x) f_1(x) - \cdots - \beta_r(x) f_r(x) = c_1 g_1(x) + \cdots + c_s g_s(x),$$

and we are done, since this verifies that

$$\begin{aligned} \alpha(x) = \beta_1(x) f_1(x) + \cdots + \beta_r(x) f_r(x) + c_1 g_1(x) + \cdots + c_s g_s(x) \\ \in (f_1(x), \dots, f_r(x), g_1(x), \dots, g_s(x)), \end{aligned}$$

completing the proof of Claim 1.4, hence of Lemma 1.3, hence of Theorem 1.2. □

1.2. Prime and irreducible elements. Let R be a (commutative) ring, and let $a, b \in R$. We say that a *divides* b, or that a is a *divisor* of b, or that b is a *multiple* of a, if $b \in (a)$, that is,

$$(\exists c \in R), \quad b = ac.$$

We use the notation $a \mid b$.

Two elements a, b are *associates* if $(a) = (b)$, that is, if $a \mid b$ and $b \mid a$.

Lemma 1.5. *Let a, b be nonzero elements of an* integral domain *R. Then a and b are associates if and only if $a = ub$, for u a unit in R.*

Proof. Assume a and b are associates. Then $\exists c, d \in R$ such that

$$b = ac, \quad a = bd;$$

therefore $a = bd = acd$, i.e.,

$$a(1 - cd) = 0.$$

Since cancellation by nonzero elements hold in integral domains, this implies $cd = 1$. Thus c is a unit, as needed.

The converse is left to the reader. □

Incidentally, here the reader sees why it is convenient to restrict our attention to integral domains: for this argument to work, a must be a non-zero-divisor. For example, in $\mathbb{Z}/6\mathbb{Z}$ the classes $[2]_6$, $[4]_6$ of 2 and 4 are associates according to our definition, yet $[4]_6$ cannot be written as $[2]_6$ times a unit. Away from the comfortable environment of integral domains, even harmless-looking statements such as Lemma 1.5 may fail.

The notions reviewed above generalize directly the corresponding notions in $\mathbb{Z}$. We are going to explore analogs of other common notions in $\mathbb{Z}$, such as 'primality' and 'irreducibility', in more general integral domains.

Definition 1.6. Let R be an integral domain.

- An element $a \in R$ is *prime* if the ideal (a) is prime; that is, a is not a unit and (cf. Proposition III.4.11)

$$a \mid bc \implies (a \mid b \quad \text{or} \quad a \mid c).$$

- An element $a \in R$ is *irreducible* if a is not a unit and

$$a = bc \implies (b \text{ is a unit or } c \text{ is a unit}).$$

There are useful alternative ways to think about the notion of 'irreducible': a nonunit a is irreducible if and only if

- $a = bc$ implies that a is an associate of b or of c;
- $a = bc$ implies that $(a) = (b)$ or $(a) = (c)$ (Lemma 1.5);
- $(a) \subseteq (b) \implies (b) = (a)$ or $(b) = (1)$ (Exercise 1.12);
- (a) is maximal *among proper principal ideals* (just rephrasing the previous point!).

It is important to realize that primality and irreducibility are *not equivalent;* this is somewhat counterintuitive since they *are* equivalent in $\mathbb{Z}$, as the reader should verify[2] (Exercise 1.13). What *is* true in general is that *prime* is stronger than *irreducible:*

Lemma 1.7. *Let R be an integral domain, and let $a \in R$ be a nonzero prime element. Then a is irreducible.*

[2]This fact will be fully explained by the general theory, so the reader should work this out right away.

Proof. Since (a) is prime, $(a) \neq (1)$; hence a is not a unit. If $a = bc$, then $bc = a \in (a)$; therefore $b \in (a)$ or $c \in (a)$ since (a) is prime. Assuming without loss of generality $b \in (a)$, we have $(b) \subseteq (a)$. On the other hand $a = bc$ implies $(a) \subseteq (b)$: hence $(a) = (b)$, that is, a and b are associates, as needed. □

We will soon see under what circumstances the converse statement holds.

1.3. Factorization into irreducibles; domains with factorizations.

Definition 1.8. Let R be an integral domain. An element $r \in R$ *has a factorization* (or *decomposition*) *into irreducibles* if there exist irreducible elements $q_1, \ldots, q_n$ such that $r = q_1 \cdots q_n$.

This factorization is *unique* if the elements q_i are determined by r up to order and associates, that is, if whenever

$$r = q_1' \cdots q_m'$$

is another factorization of r into irreducibles, then $m = n$ and q_i' is an associate of q_i after (possibly) shuffling the factors. ⌟

Definition 1.9. An integral domain R is a *domain with factorizations* (or *'factorizations exist in R'*) if every nonzero, nonunit element $r \in R$ has a factorization into irreducibles. ⌟

Definition 1.10. An integral domain R is *factorial,* or a *unique factorization domain* (abbreviated *UFD*), if every nonzero, nonunit element $r \in R$ has a unique factorization into irreducibles. ⌟

The terminology introduced in Definition 1.9 does not appear to be too standard; by contrast, UFDs are famous.

We will study the unique factorization condition in §2. For now, it seems worth spending a little time contemplating the mere existence of factorizations. Interestingly, this condition is implied by an *ascending chain condition,* for a special class of ideals.

Proposition 1.11. *Let R be an integral domain, and let r be a nonzero, nonunit element of R. Assume that every ascending chain of* principal *ideals*

$$(r) \subseteq (r_1) \subseteq (r_2) \subseteq (r_3) \subseteq \cdots$$

stabilizes. Then r has a factorization into irreducibles.

Proof. Assume that r does *not* have a factorization into irreducible elements. In particular, r is itself not irreducible; thus $\exists r_1, s_1 \in R$ such that $r = r_1 s_1$ and $(r) \subsetneq (r_1)$, $(r) \subsetneq (s_1)$. If both r_1, s_1 have factorizations into irreducibles, the product of these factorizations gives a factorization of r; thus we may assume that (e.g.) r_1 does not have a factorization into irreducibles. Therefore we have

$$(r) \subsetneq (r_1)$$

and r_1 does not have a factorization; iterating this argument constructs an infinitely increasing chain

$$(r) \subsetneq (r_1) \subsetneq (r_2) \subsetneq (r_3) \subsetneq \cdots,$$

contradicting our hypothesis. □

Thus, factorizations exist in integral domains in which the ascending chain condition holds *for principal ideals.* This has the following immediate and convenient consequence:

Corollary 1.12. *Let R be a Noetherian domain. Then factorizations exist in R.*

Proof. By Proposition 1.1, Noetherian domains satisfy the ascending chain condition for *all* ideals. □

Corollary 1.12 verifies part of the picture presented at the beginning of the chapter: the class of Noetherian domains is contained in the class of domains with factorization. This inclusion is proper: for instance, the standard example of a non-Noetherian ring,

$$\mathbb{Z}[x_1, x_2, x_3, \dots],$$

(Example III.6.5) *does* have factorizations. Indeed, every given polynomial $f \in \mathbb{Z}[x_1, x_2, \dots]$ involves only finitely many variables[3], so it belongs to a subring $\mathbb{Z}[x_1, \cdots, x_n]$ isomorphic to an ordinary polynomial ring over $\mathbb{Z}$; further, this subring contains every divisor of f. It follows easily (cf. Exercise 1.15) that the ascending chain condition for principal ideals holds in $\mathbb{Z}[x_1, x_2, x_3, \dots]$, because it holds in $\mathbb{Z}[x_1, \dots, x_n]$ (since this ring is Noetherian, by Hilbert's basis theorem).

Exercises

Remember that in this section all rings are taken to be *commutative.*

1.1. ▷ Let R be a Noetherian ring, and let I be an ideal of R. Prove that R/I is a Noetherian ring. [§1.1]

1.2. Prove that if $R[x]$ is Noetherian, so is R. (This is a 'converse' to Hilbert's basis theorem.)

1.3. Let k be a field, and let $f \in k[x]$, $f \notin k$. For every subring R of $k[x]$ containing k and f, define a homomorphism $\varphi : k[t] \to R$ by extending the identity on k and mapping t to f. This makes every such R a $k[t]$-algebra (Example III.5.6).

- Prove that $k[x]$ is finitely generated as a $k[t]$-module.
- Prove that every subring R as above is finitely generated as a $k[t]$-module.
- Prove that every subring of $k[x]$ containing k is a Noetherian ring.

1.4. Let R be the ring of real-valued continuous functions on the interval $[0, 1]$. Prove that R is not Noetherian.

1.5. Determine for which sets S the power set ring $\mathscr{P}(S)$ is Noetherian. (Cf. Exercise III.3.16.)

[3]Remember that polynomials are *finite* linear combinations of monomials; cf. §III.1.3.

1.6. ▷ Let I be an ideal of $R[x]$, and let $A \subseteq R$ be the set defined in the proof of Theorem 1.2. Prove that A is an ideal of R. [§1.1]

1.7. Prove that if R is a Noetherian ring, then the ring of power series $R[[x]]$ (cf. §III.1.3) is also Noetherian. (Hint: The order of a power series $\sum_{i=0}^{\infty} a_i x^i$ is the smallest i for which $a_i \neq 0$; the *dominant coefficient* is then a_i. Let $A_i \subseteq R$ be the set of dominant coefficients of series of order i in I, together with 0. Prove that A_i is an ideal of R and $A_0 \subseteq A_1 \subseteq A_2 \subseteq \cdots$. This sequence stabilizes since R is Noetherian, and each A_i is finitely generated for the same reason. Now adapt the proof of Lemma 1.3.)

1.8. Prove that every ideal in a Noetherian ring R contains a finite product of prime ideals. (Hint: Let $\mathscr{F}$ be the family of ideals that do not contain finite products of prime ideals. If $\mathscr{F}$ is nonempty, it has a maximal element M since R is Noetherian. Since $M \in \mathscr{F}$, M is not itself prime, so $\exists a, b \in R$ s.t. $a \notin M$, $b \notin M$, yet $ab \in M$. What's wrong with this?)

1.9. ¬ Let R be a commutative ring, and let $I \subseteq R$ be a proper ideal. The reader will prove in Exercise 3.12 that the set of prime ideals containing I has minimal elements (the *minimal primes* of I). Prove that if R is Noetherian, then the set of minimal primes of I is finite. (Hint: Let $\mathscr{F}$ be the family of ideals that do *not* have finitely many minimal primes. If $\mathscr{F} \neq \emptyset$, note that $\mathscr{F}$ must have a maximal element I, and I is not prime itself. Find ideals J_1, J_2 strictly larger than I, such that $J_1 J_2 \subseteq I$, and deduce a contradiction.) [VI.4.10]

1.10. ¬ By Proposition 1.1, a ring R is Noetherian if and only if it satisfies the a.c.c. for ideals. A ring is *Artinian* if it satisfies the d.c.c. (descending chain condition) for ideals. Prove that if R is Artinian and $I \subseteq R$ is an ideal, then R/I is Artinian. Prove that if R is an Artinian integral domain, then it is a field. (Hint: Let $r \in R$, $r \neq 0$. The ideals (r^n) form a descending sequence; hence $(r^n) = (r^{n+1})$ for some n. Therefore....) Prove that Artinian rings have Krull dimension 0 (that is, prime ideals are maximal in Artinian rings[4]). [2.11]

1.11. Prove that the 'associate' relation is an equivalence relation.

1.12. ▷ Let R be an integral domain. Prove that $a \in R$ is irreducible if and only if (a) is maximal among proper principal ideals of R. [§1.2, §2.3]

1.13. ▷ Prove that prime $\iff$ irreducible in $\mathbb{Z}$. [§1.2, §2.3]

1.14. For a, b in a commutative ring R, prove that the class of a in $R/(b)$ is prime if and only if the class of b in $R/(a)$ is prime.

1.15. ▷ Identify $S = \mathbb{Z}[x_1, \ldots, x_n]$ in the natural way with a subring of the polynomial ring in countably infinitely many variables $R = \mathbb{Z}[x_1, x_2, x_3, \ldots]$. Prove that if $f \in S$ and $(f) \subseteq (g)$ in R, then $g \in S$ as well. Conclude that the ascending chain condition for principal ideals holds in R, and hence R is a domain with factorizations. [§1.3, §4.3]

[4]One can prove that Artinian rings are necessarily Noetherian; in fact, a ring is Artinian if and only if it is Noetherian and has Krull dimension 0. Thus, the d.c.c. implies the a.c.c., while the a.c.c. implies the d.c.c. if and only if all prime ideals are maximal.

1.16. Let

$$R = \frac{\mathbb{Z}[x_1, x_2, x_3, \dots]}{(x_1 - x_2^2, x_2 - x_3^2, \dots)}.$$

Does the ascending chain condition for principal ideals hold in R?

1.17. ▷ Consider the subring of $\mathbb{C}$:

$$\mathbb{Z}[\sqrt{-5}] := \{a + bi\sqrt{5} \mid a, b \in \mathbb{Z}\}.$$

- Prove that this ring is isomorphic to $\mathbb{Z}[t]/(t^2 - 5)$.
- Prove that it is a Noetherian integral domain.
- Define a 'norm' N on $\mathbb{Z}[\sqrt{-5}]$ by setting $N(a + bi\sqrt{5}) = a^2 + 5b^2$. Prove that $N(zw) = N(z)N(w)$. (Cf. Exercise III.4.10.)
- Prove that the units in $\mathbb{Z}[\sqrt{-5}]$ are ± 1. (Use the preceding point.)
- Prove that 2, 3, $1 + i\sqrt{5}$, $1 - i\sqrt{5}$ are all irreducible nonassociate elements of $\mathbb{Z}[\sqrt{-5}]$.
- Prove that no element listed in the preceding point is prime. (Prove that the rings obtained by mod-ing out the ideals generated by these elements are not integral domains.)
- Prove that $\mathbb{Z}[\sqrt{-5}]$ is not a UFD.

[§2.2, 2.18, 6.14]

---

2. UFDs, PIDs, Euclidean domains

2.1. Irreducible factors and greatest common divisor. An integral domain R is a UFD if factorizations exist in R and are *unique* in the sense of Definition 1.8.

Thus, in a UFD all elements (other than 0 and the units) determine a multiset (a set of elements 'with multiplicity'; cf. §I.1.1) of irreducible *factors*, determined up to the associate relation. We can also agree that units have *no* factors; that is, the corresponding multiset is $\emptyset$.

The following trivial remark is at the root of most elementary facts about UFDs, such as the characterization of Theorem 2.5:

Lemma 2.1. *Let R be a UFD, and let a, b, c be nonzero elements of R. Then*

- $(a) \subseteq (b)$ $\iff$ *the multiset of irreducible factors of b is contained in the multiset of irreducible factors of a;*
- *a and b are associates (that is, $(a) = (b)$)* $\iff$ *the two multisets coincide;*
- *the irreducible factors of a product bc are the collection of all irreducible factors of b and of c.*

The proof is left to the reader (Exercise 2.1). The advantage of working in a UFD resides in the fact that ring-theoretic statements about elements of the ring often reduce to straightforward set-theoretic statements about multisets of irreducible elements, by means of Lemma 2.1.

One important instance of this mechanism is the existence of *greatest common divisors*. We have liberally used this notion (at least for integers) in previous chapters; now we can appreciate it from a more technical perspective.

Definition 2.2. Let R be an integral domain, and let $a, b \in R$. An element $d \in R$ is a *greatest common divisor* (often abbreviated 'gcd') of a and b if $(a, b) \subseteq (d)$ and (d) is the smallest principal ideal in R with this property. ⌟

In other words, d is a gcd of a and b if $d \mid a$, $d \mid b$, and

$$c \mid a\,,\ c \mid b \implies c \mid d.$$

This definition is immediately extended to any finite number of elements.

Note that greatest common divisors are not defined uniquely by this prescription: if d is a greatest common divisor of a and b, so is every associate of d. Thus, the notation '$\gcd(a, b)$' should only be used for the associate class formed by all greatest common divisors of a, b. Of course, language is often (harmlessly) abused on this point. For example, the fact that we can talk about *the* greatest common divisor of two integers is due to the fact that in $\mathbb{Z}$ there is a convenient way to choose a distinguished element in each class of associate integers (that is, the nonnegative one).

Also note that greatest common divisors need not *exist* (cf. Exercise 2.5); but they do exist in UFDs:

Lemma 2.3. *Let R be a UFD, and let a, b be nonzero elements of R. Then a, b have a greatest common divisor.*

Proof. We can write

$$a = uq_1^{\alpha_1} \cdots q_r^{\alpha_r}, \quad b = vq_1^{\beta_1} \cdots q_r^{\beta_r}$$

where u and v are units, the elements q_i are irreducible, q_i is not an associate of q_j for $i \neq j$, and $\alpha_i \geq 0$, $\beta_i \geq 0$ (so that the multisets of irreducible factors of a, resp., b, consist of those q_i for which $\alpha_i > 0$, resp., $\beta_i > 0$; the units u, v are included since the irreducible factors are only defined up to the associate relation).

We claim that

$$d = q_1^{\min(\alpha_1, \beta_1)} \cdots q_r^{\min(\alpha_r, \beta_r)}$$

is a gcd of a and b. Indeed, d is clearly a divisor of a and b; and if c also divides a and b, then the multiset of factors of c must be contained in both multisets of factors for a and b (by Lemma 2.1); that is,

$$c = wq_1^{\gamma_1} \cdots q_r^{\gamma_r}$$

with w a unit and $\gamma_i \leq \alpha_i$, $\gamma_i \leq \beta_i$. This implies $\gamma_i \leq \min(\alpha_i, \beta_i)$, and hence $c \mid d$ (again by Lemma 2.1), as needed. □

Of course the argument given in the proof generalizes one of the standard ways to compute greatest common divisors in $\mathbb{Z}$: find smallest exponents in prime factorizations. But note that this is not the only way to compute the gcd in $\mathbb{Z}$; we will come back to this point in a moment. In fact, greatest common divisors in $\mathbb{Z}$ have properties that should not be expected in a more general domain: for

example, the result of Exercise II.2.13 does *not* generalize to arbitrary domains (and not even to arbitrary UFDs), as the reader will check in Exercise 2.4.

2.2. Characterization of UFDs. It is easy to construct integral domains where unique factorization fails. The diligent reader has already analyzed one example (Exercise 1.17); for another, in the domain[5]

$$R = \frac{\mathbb{C}[x,y,z,w]}{(xw-yz)}$$

the (classes of the) elements x, y, z, w are irreducible and not associates of one another; since $xw - yz = 0$ in R, the element $r = xw$ has two distinct factorizations into irreducibles: $r = xw = yz$.

Note that this ring is Noetherian, by Theorem 1.2 (and in particular factorizations do exist in R). Thus, there are Noetherian integral domains that are *not* UFDs.

Also note that this ring provides an example in which the converse to Lemma 1.7 does *not* hold: indeed, (the class of) x is irreducible, but the quotient

$$\left(\frac{\mathbb{C}[x,y,z,w]}{(xw-yz)}\Big/(x)\right) \cong \frac{\mathbb{C}[x,y,z,w]}{(x,xw-yz)} = \frac{\mathbb{C}[x,y,z,w]}{(x,yz)}$$

is not an integral domain (because $y \neq 0$, $z \neq 0$, and yet $yz = 0$ in this ring); that is, x is not prime.

In fact, and maybe a little surprisingly, the issue of unique factorization is inextricably linked with the relation between primality and irreducibility. Indeed, the 'converse' to Lemma 1.7 does hold in UFDs:

Lemma 2.4. *Let R be a UFD, and let a be an irreducible element of R. Then a is prime.*

Proof. The element a is not a unit, by definition of irreducible. Assume $bc \in (a)$: thus $(bc) \subseteq (a)$, and by Lemma 2.1 the irreducible factors of a, that is, a itself, must be among the factors of b or of c. We have $b \in (a)$ in the first case and $c \in (a)$ in the second. This shows that (a) is a prime ideal, as needed. □

In fact, more is true. Provided that the ascending chain condition for principal ideals holds, then UFDs are *characterized* by the equivalence between irreducibility and primality.

Theorem 2.5. *An integral domain R is a UFD if and only if*

- *the a.c.c. for principal ideals holds in R and*
- *every irreducible element of R is prime.*

Proof. ($\Longrightarrow$) Assume that R is a UFD. Lemma 2.4 shows that irreducible elements of R are prime. To prove that the a.c.c. for principal ideals holds, consider an ascending chain

$$(r_1) \subsetneq (r_2) \subsetneq (r_3) \subsetneq \cdots .$$

[5] In algebraic geometry, this is the ring of a 'quadric cone in $\mathbb{A}^4$'. The vertex of this cone (at the origin) is a *singular* point, and this has to do with the fact that R is not a UFD.

By Lemma 2.1, this chain determines a corresponding *descending* chain of multisets of irreducible factors. A descending chain of finite multisets clearly stabilizes, and it follows (by Lemma 2.1 again) that $(r_i) = (r_{i+1}) = (r_{i+2}) = \dots$ for large enough i.

($\Longleftarrow$) Now assume that R satisfies the a.c.c. for principal ideals and irreducibles are prime. Proposition 1.11 implies that factorizations exist in R; we have to verify uniqueness. Let $q_1, \dots, q_m$ and $q'_1, \dots, q'_n$ be irreducible elements of R, and assume

$$q_1 \cdots q_m = q'_1 \cdots q'_n.$$

Then $q'_1 \cdots q'_n \in (q_1)$, and (q_1) is a prime ideal (by hypothesis); thus $q'_i \in (q_1)$ for some i, which we may assume to be 1 after changing the order of the factors. Therefore $q'_1 = uq_1$ for some $u \in R$. Since q'_1 is irreducible and q_1 is not a unit, necessarily u is a unit. Thus q_1 and q'_1 are associates. Canceling q_1 and replacing q'_2 by uq'_2, we find

$$q_2 \cdots q_m = q'_2 \cdots q'_n.$$

Repeating this process matches all factors one by one. It is clear that $m = n$, because otherwise we will obtain $1 =$ a product of irreducibles, contradicting the fact that irreducibles are not units. □

Theorem 2.5 is a pleasant statement, but it does not make it particularly easy to check whether a given ring is in fact a UFD. The situation is not unlike that with Noetherian rings: the a.c.c. is a sharp equivalent formulation of the Noetherian condition, but in practice one relies more often on other tools, such as Hilbert's basis theorem, in order to establish that a given ring is Noetherian. Does an analog of Hilbert's basis theorem hold for unique factorization domains?

Answering this question requires some preparatory work, and we will come back to it in §4.

2.3. PID $\Longrightarrow$ UFD. There are simple ways to produce examples of UFDs. Recall (from §III.4) that a *principal ideal domain* (abbreviated PID) is an integral domain in which every ideal is principal. In §III.4 we have observed that

- PIDs are Noetherian.
- $\mathbb{Z}$ and $k[x]$ (where k is a field) are PIDs.
- If R is a PID and $a, b \in R$, then d is a greatest common divisor for a and b if and only if $(a, b) = (d)$. In particular, if d is a greatest common divisor of a and b in a PID R, then d is a linear combination of a and b: $\exists r, s \in R$ such that $d = ra + sb$.
- If I is a nonzero ideal in a PID, then I is prime if and only if it is maximal.

We now add one important point to this list:

Proposition 2.6. *If R is a PID, then it is a UFD.*

Proof. Let R be a PID. Since PIDs are Noetherian, the a.c.c. holds in R (for all ideals, hence in particular for *principal* ideals); we verify that irreducible elements are prime in R, which implies that R is a UFD, by Theorem 2.5.

Let $a \in R$ be an irreducible element, and assume $bc \in (a)$; we have to show that either b or c is in (a). If $b \in (a)$, we are done; thus, assume $b \notin (a)$. Then $(a,b) \supsetneq (a)$; as R is a PID, $\exists d \in R$ such that $(d) = (a,b)$, and hence $(d) \supsetneq (a)$. This implies that $(a,b) = (d) = (1)$, because a is irreducible and ideals generated by irreducible elements are maximal among principal ideals (cf. Exercise 1.12).

Hence $\exists r, s \in R$ such that $ra + sb = 1$. Multiplying by c, we obtain that

$$c = rac + sbc \in (a)$$

as needed. □

In particular, $k[x]$ is a UFD if k is a field, and $\mathbb{Z}$ is a UFD—which hopefully will not come as a big surprise to our readers[6]. Note that at this point we have recovered the fact stated in Exercise 1.13 in full glory.

Proposition 2.6 justifies another feature of the picture given at the beginning of this chapter: the class of UFDs is contained in the class of PIDs. We will soon see that the inclusion is proper, that is, that there are UFDs which are not PIDs; for example, $\mathbb{Z}[x]$ is not a PID (Exercise 2.12), yet

$$\mathbb{Z}[x] \text{ is a unique factorization domain}$$

as we will soon be able to prove (Theorem 4.14). In fact, there are UFDs which are not Noetherian, as represented in the picture; however, these examples are best discussed after more material has been developed (Example 4.18).

The reader can already get a feel for the gap between UFD and PID, by contemplating the fact (recalled above and essentially tautological) that, *in a PID*, greatest common divisors of a, b are linear combinations of a and b. This is a very strong requirement: for example, it characterizes PIDs among Noetherian domains, as the reader will check (Exercise 2.7); it does not hold in general UFDs (Exercise 2.4).

2.4. Euclidean domain $\Longrightarrow$ PID. The excellent properties of $\mathbb{Z}$ and $k[x]$ (where k is a field) make these rings even more special than PIDs: they are *Euclidean domains.*

Informally, Euclidean domains are rings in which one may perform a 'division with remainder': this is the case for $\mathbb{Z}$ and for $k[x]$, as observed in §III.4. The point is that in both $\mathbb{Z}$ and $k[x]$ one can define a notion of 'size' of an element: $|n|$ for an integer n and $\deg f(x)$ for a polynomial $f(x)$. In both cases, one has control over the size of the 'remainder' in a division. The definition of Euclidean domain simply abstracts this mechanism.

For the purpose of this discussion[7], a *valuation* on an integral domain R is any function $v : R \setminus \{0\} \to \mathbb{Z}^{\geq 0}$.

[6] This fact is known as the *fundamental theorem of arithmetic.*

[7] Entire libraries have been written on the subject of *valuations,* studying a more precise notion than what is needed here.

Definition 2.7. A *Euclidean valuation* on an integral domain R is a valuation satisfying the following property[8]: for all $a \in R$ and all nonzero $b \in R$ there exist $q, r \in R$ such that

$$a = qb + r,$$

with either $r = 0$ or $v(r) < v(b)$. An integral domain R is a *Euclidean domain* if it admits a Euclidean valuation. ⌟

We say that q is the *quotient* of the division and r is the *remainder*. Division with remainder in $\mathbb{Z}$ and in $k[x]$ (where k is a field) provide examples, so that $\mathbb{Z}$ and $k[x]$ are Euclidean domains.

Proposition 2.8. *Let R be a Euclidean domain. Then R is a PID.*

The proof is modeled after the instances encountered for $\mathbb{Z}$ (Proposition III.4.4) and $k[x]$ (which the reader has hopefully worked out in Exercise III.4.4).

Proof. Let I be an ideal of R; we have to prove that I is principal. If $I = \{0\}$, there is nothing to show; therefore, assume $I \neq \{0\}$. The valuation maps the nonzero elements of I to a subset of $\mathbb{Z}^{\geq 0}$; let $b \in I$ be an element with the smallest valuation. Then we claim that $I = (b)$; therefore I is principal, as needed. Since clearly $(b) \subseteq I$, we only need to verify that $I \subseteq (b)$.

For this, let $a \in I$ and apply division with remainder: we have

$$a = qb + r$$

for some q, r in R, with $r = 0$ or $v(r) < v(b)$. But

$$r = a - qb \in I :$$

by the minimality of $v(b)$ among nonzero elements of I, we cannot have $v(r) < v(b)$. Therefore $r = 0$, showing that $a = qb \in (b)$, as needed. □

Proposition 2.8 justifies one more feature of our picture at the beginning of the chapter: the class of Euclidean domains is contained in the class of principal ideal domains.

This inclusion is proper, as suggested in the picture. Producing an explicit example of a PID which is not a Euclidean domain is not so easy, but the gap between PIDs and Euclidean domains can in fact be described very sharply: PID may be characterized as domains satisfying a weaker requirement than 'division with remainder'.

More precisely, a 'Dedekind-Hasse valuation' is a valuation v such that $\forall a, b$, either $(a, b) = (b)$ (that is, b divides a) or there exists $r \in (a, b)$ such that $v(r) < v(b)$. This latter condition amounts to requiring that there exist $q, s \in R$ such that $as = bq + r$ with $v(r) < v(b)$; hence a Euclidean valuation (for which we may in fact choose $s = 1$) is a Dedekind-Hasse valuation. It is not hard to show that an integral domain is a PID if and only if it admits a Dedekind-Hasse valuation (Exercise 2.21). For example, this can be used to show that the ring $\mathbb{Z}[(1 + \sqrt{-19})/2]$ is a PID: the

[8] It is not uncommon to also require that $v(ab) \geq v(b)$ for all nonzero $a, b \in R$; but this is not needed in the considerations that follow, and cf. Exercise 2.15.

norm considered in Exercise 2.18 in order to prove that this ring is *not* a Euclidean domain turns out to be a Dedekind-Hasse valuation[9]. Thus, this ring gives an example of a PID that is not a Euclidean domain.

One excellent feature of Euclidean domains, and the one giving them their names, is the presence of an effective algorithm computing greatest common divisors: the *Euclidean algorithm.* As Euclidean domains are PIDs, and hence UFDs, we know that they do have greatest common divisors. However, the 'algorithm' obtained by distilling the proof of Lemma 2.3 is highly impractical: if we had to factor two integers a, b in order to compute their gcd, this would make it essentially impossible (with current technologies and factorization algorithms) for integers of a few hundred digits. The Euclidean algorithm *bypasses* the necessity of factorization: greatest common divisors of thousand-digit integers may be computed in a fraction of a second.

The key lemma on which the algorithm is based is the following trivial general fact:

Lemma 2.9. *Let $a = bq + r$ in a ring R. Then $(a, b) = (b, r)$.*

Proof. Indeed, $r = a - bq \in (a, b)$, proving $(b, r) \subseteq (a, b)$; and $a = bq + r \in (b, r)$, proving $(a, b) \subseteq (b, r)$. □

In particular,

$$(\forall c \in R), \quad (a, b) \subseteq (c) \iff (b, r) \subseteq (c);$$

that is, the set of common divisors of a, b and the set of common divisors of b, r coincide. Therefore,

Corollary 2.10. *Assume $a = bq + r$. Then a, b have a gcd if and only if b, r have a gcd, and in this case* $\gcd(a, b) = \gcd(b, r)$.

Of course '$\gcd(a, b) = \gcd(b, r)$' means that the two classes of associate elements coincide.

These considerations hold over any integral domain; assume now that R is a Euclidean domain. Then we can use division with remainder to gain some control over the remainders r. Given two elements a, b in R, with $b \neq 0$, we can apply division with remainder repeatedly:

$$\begin{aligned} a &= bq_1 + r_1, \\ b &= r_1 q_2 + r_2, \\ r_1 &= r_2 q_2 + r_3, \end{aligned}$$

as long as the remainder r_i is nonzero.

Claim 2.11. *This process terminates: that is, $r_N = 0$ for some N.*

[9]This boils down to a case-by-case analysis, which we are happily leaving to our most patient readers.

Proof. Each line in the table is a division with remainder. If no r_i were zero, we would have an infinite decreasing sequence

$$v(b) > v(r_1) > v(r_2) > v(r_3) > \cdots$$

of *nonnegative* integers, which is nonsense. □

Thus the table of divisions with remainders must be as follows: letting $r_0 = b$,

$$\begin{aligned} a &= r_0 q_1 + r_1, \\ b &= r_1 q_2 + r_2, \\ r_1 &= r_2 q_2 + r_3, \end{aligned}$$

$$\begin{aligned} r_{N-3} &= r_{N-2} q_{N-2} + r_{N-1}, \\ r_{N-2} &= r_{N-1} q_{N-1} \end{aligned}$$

with $r_{N-1} \neq 0$.

Proposition 2.12. *With notation as above, r_{N-1} is a gcd of a, b.*

Proof. By Corollary 2.10,

$$\gcd(a, b) = \gcd(b, r_1) = \gcd(r_1, r_2) = \cdots = \gcd(r_{N-2}, r_{N-1}).$$

But $r_{N-2} = r_{N-1} q_{N-1}$ gives $r_{N-2} \in (r_{N-1})$; hence $(r_{N-2}, r_{N-1}) = (r_{N-1})$. Therefore r_{N-1} is a gcd for r_{N-2} and r_{N-1}, hence for a and b, as needed. □

The ring of integers and the polynomial ring over a field are both Euclidean domains. Fields are Euclidean domains (as represented in the picture at the beginning of the chapter), but not for a very interesting reason: the remainder of the division by a nonzero element in a field is *always* zero, so every function qualifies as a 'Euclidean valuation' for trivial reasons.

We will study another interesting Euclidean domain later in this chapter (§6.2).

Exercises

2.1. ▷ Prove Lemma 2.1. [§2.1]

2.2. Let R be a UFD, and let a, b, c be elements of R such that $a \mid bc$ and $\gcd(a, b) = 1$. Prove that a divides c.

2.3. Let n be a positive integer. Prove that there is a one-to-one correspondence preserving multiplicities between the irreducible factors of n (as an integer) and the composition factors of $\mathbb{Z}/n\mathbb{Z}$ (as a group). (In fact, the Jordan-Hölder theorem may be used to prove that $\mathbb{Z}$ is a UFD.)

2.4. ▷ Consider the elements x, y in $\mathbb{Z}[x, y]$. Prove that 1 is a gcd of x and y, and yet 1 is *not* a linear combination of x and y. (Cf. Exercise II.2.13.) [§2.1, §2.3]

2.5. ▷ Let R be the subring of $\mathbb{Z}[t]$ consisting of polynomials with no term of degree 1: $a_0 + a_2t^2 + \cdots + a_dt^d$.

- Prove that R is indeed a subring of $\mathbb{Z}[t]$, and conclude that R is an integral domain.
- List all common divisors of t^5 and t^6 in R.
- Prove that t^5 and t^6 have no gcd in R.

[§2.1]

2.6. Let R be a domain with the property that the intersection of any family of principal ideals in R is necessarily a principal ideal.

- Show that greatest common divisors exist in R.
- Show that UFDs satisfy this property.

2.7. ▷ Let R be a Noetherian domain, and assume that for all nonzero a, b in R, the greatest common divisors of a and b are linear combinations of a and b. Prove that R is a PID. [§2.3]

2.8. Let R be a UFD, and let $I \neq (0)$ be an ideal of R. Prove that every descending chain of principal ideals containing I must stabilize.

2.9. ¬ The *height* of a prime ideal P in a ring R is (if finite) the maximum length h of a chain of prime ideals $P_0 \subsetneq P_1 \subsetneq \cdots \subsetneq P_h = P$ in R. (Thus, the Krull dimension of R, if finite, is the maximum height of a prime ideal in R.) Prove that if R is a UFD, then every prime ideal of height 1 in R is principal. [2.10]

2.10. ¬ It is a consequence of a theorem known as *Krull's Hauptidealsatz* that every nonzero element in a Noetherian domain is contained in a prime ideal of height 1. Assuming this, prove a converse to Exercise 2.9, and conclude that a Noetherian domain R is a UFD if and only if every prime ideal of height 1 in R is principal. [4.16]

2.11. Let R be a PID, and let I be a nonzero ideal of R. Show that R/I is an artinian ring (cf. Exercise 1.10), by proving explicitly that the d.c.c. holds in R/I.

2.12. ▷ Prove that if $R[x]$ is a PID, then R is a field. [§2.3, §VI.7.1]

2.13. ▷ For a, b, c positive integers with $c > 1$, prove that $c^a - 1$ divides $c^b - 1$ if and only if $a \mid b$. Prove that $x^a - 1$ divides $x^b - 1$ in $\mathbb{Z}[x]$ if and only if $a \mid b$. (Hint: For the interesting implications, write $b = ad + r$ with $0 \leq r < a$, and take 'size' into account.) [§VII.5.1, VII.5.13]

2.14. ▷ Prove that if k is a field, then $k[[x]]$ is a Euclidean domain. [§4.3]

2.15. ▷ Prove that if R is a Euclidean domain, then R admits a Euclidean valuation $\overline{v}$ such that $\overline{v}(ab) \geq \overline{v}(b)$ for all nonzero $a, b \in R$. (Hint: Since R is a Euclidean domain, it admits a valuation v as in Definition 2.7. For $a \neq 0$, let $\overline{v}(a)$ be the minimum of all $v(ab)$ as $b \in R$, $b \neq 0$. To see that R is a Euclidean domain with respect to $\overline{v}$ as well, let a, b be nonzero in R, with $b \nmid a$; choose q, r so that $a = bq + r$, with $v(r)$ minimal; assume that $\overline{v}(r) \geq \overline{v}(b)$, and get a contradiction.) [§2.4, 2.16]

2.16. Let R be a Euclidean domain with Euclidean valuation v; assume that $v(ab) \geq v(b)$ for all nonzero $a, b \in R$ (cf. Exercise 2.15). Prove that associate elements have the same valuation and that units have minimum valuation.

2.17. ¬ Let R be a Euclidean domain that is not a field. Prove that there exists a nonzero, nonunit element c in R such that $\forall a \in R$, $\exists q, r \in R$ with $a = qc + r$ and either $r = 0$ or r a unit. [2.18]

2.18. ▷ For an integer d, denote by $\mathbb{Q}(\sqrt{d})$ the smallest subfield of $\mathbb{C}$ containing $\mathbb{Q}$ and $\sqrt{d}$, with norm N defined as in Exercise III.4.10. See Exercise 1.17 for the case $d = -5$; in this problem, you will take $d = -19$.

Let $\delta = (1 + i\sqrt{19})/2$, and consider the following subring of $\mathbb{Q}(\sqrt{-19})$:

$$\mathbb{Z}[\delta] := \left\{ a + b\,\frac{1 + i\sqrt{19}}{2} \mid a, b \in \mathbb{Z} \right\}.$$

- Prove that the smallest values of $N(z)$ for $z = a + b\delta \in \mathbb{Z}[\delta]$ are 0, 1, 4, 5. Prove that $N(a + b\delta) \geq 5$ if $b \neq 0$.
- Prove that the units in $\mathbb{Z}[\delta]$ are ± 1.
- If $c \in \mathbb{Z}[\delta]$ satisfies the condition specified in Exercise 2.17, prove that c must divide 2 or 3 in $\mathbb{Z}[\delta]$, and conclude that $c = \pm 2$ or $c = \pm 3$.
- Now show that $\nexists q \in \mathbb{Z}[\delta]$ such that $\delta = qc + r$ with $c = \pm 2, \pm 3$ and $r = 0, \pm 1$.

Conclude that $\mathbb{Z}[(1 + \sqrt{-19})/2]$ is not a Euclidean domain. [§2.4, 6.14]

2.19. ¬ A *discrete valuation* on a field k is a homomorphism of abelian groups $v : (k^*, \cdot) \to (\mathbb{Z}, +)$ such that $v(a + b) \geq \min(v(a), v(b))$ for all $a, b \in k^*$ such that $a + b \in k^*$.

- Prove that the set $R := \{a \in k^* \mid v(a) \geq 0\} \cup \{0\}$ is a subring of k.
- Prove that R is a Euclidean domain.

Rings arising in this fashion are called *discrete valuation rings,* abbreviated DVR. They arise naturally in number theory and algebraic geometry. Note that the Krull dimension of a DVR is 1 (Example III.4.14); in algebraic geometry, DVRs correspond to particularly nice points on a 'curve'.

- Prove that the ring of rational numbers a/b with b *not* divisible by a fixed prime integer p is a DVR.

[2.20, VIII.1.19]

2.20. ¬ As seen in Exercise 2.19, DVRs are Euclidean domains. In particular, they must be PIDs. Check this directly, as follows. Let R be a DVR, and let $t \in R$ be an element such that $v(t) = 1$. Prove that if $I \subseteq R$ is any nonzero ideal, then $I = (t^k)$ for some $k \geq 1$. (The element t is called a 'local parameter' of R.) [4.13, VII.2.18]

2.21. ▷ Prove that an integral domain is a PID if and only if it admits a Dedekind-Hasse valuation. (Hint: For the $\Longleftarrow$ implication, adapt the argument in Proposition 2.8; for $\Longrightarrow$, let $v(a)$ be the size of the multiset of irreducible factors of a.) [§2.4]

2.22. ¬ Suppose $R \subseteq S$ is an inclusion of integral domains, and assume that R is a PID. Let $a, b \in R$, and let $d \in R$ be a gcd for a and b in R. Prove that d is also a gcd for a and b in S. [5.2]

2.23. Compute $d = \gcd(5504227617645696, 2922476045110123)$. Further, find a, b such that $d = 5504227617645696\,a + 2922476045110123\,b$.

2.24. ▷ Prove that there are infinitely many prime integers. (Hint: Assume by contradiction that $p_1, \ldots, p_N$ is a complete list of all positive prime integers. What can you say about $p_1 \cdot \cdots \cdot p_N + 1$? This argument was already known to Euclid, more than 2,000 years ago.) [2.25, §5.2, 5.11]

2.25. ¬ Variation on the theme of Euclid from Exercise 2.24: Let $f(x) \in \mathbb{Z}[x]$ be a nonconstant polynomial such that $f(0) = 1$. Prove that infinitely many primes divide the numbers $f(n)$, as n ranges in $\mathbb{Z}$. (If $p_1, \ldots, p_N$ were a complete list of primes dividing the numbers $f(n)$, what could you say about $f(p_1 \cdots p_N x)$?)

Once you are happy with this, show that the hypothesis $f(0) = 1$ is unnecessary. (If $f(0) = a \neq 0$, consider $f(p_1 \cdots p_N a x)$. Finally, note that there is nothing special about 0.) [VII.5.18]

---

3. Intermezzo: Zorn's lemma

3.1. Set theory, reprise. We leave ring theory for a moment and take a little detour to contemplate an issue from *set theory*. As remarked at the very outset, only naive set theory is used in this book; all set-theoretic operations we have used so far are nothing more than a formalization of intuitive ideas regarding collections of objects. However, we will occasionally need to refer to a less 'intuitively obvious' set-theoretic statement: for example, this statement is needed in order to show that every ideal in a ring is contained in a maximal ideal (Proposition 3.5).

This set-theoretic fact is *Zorn's lemma*. An *order relation* on a set Z is a relation $\preceq$ which is reflexive, transitive, and *antisymmetric:* the first two terms are familiar to the reader, and the third means that

$$(\forall a, b \in Z), \quad a \preceq b \text{ and } b \preceq a \implies a = b.$$

Typical prototypes are the $\leq$ relation on $\mathbb{Z}$ or the inclusion relation $\subseteq$ among subsets of a given set. We use $a \prec b$ to denote $a \preceq b$ *and* $b \neq a$.

A pair $(Z, \preceq)$, consisting of a set Z and an order relation $\preceq$ on Z, is called a *poset*, for *partially ordered set*. The qualifier 'partially' is not necessary, but convenient, as it reminds us that if $a, b \in Z$, then it is not necessarily the case that $a \preceq b$ or $b \preceq a$: for example, $\subseteq$ does not satisfy this additional requirement in general (while $(\mathbb{Z}, \leq)$ does). An order is *total* if it does satisfy this additional requirement. A *totally ordered set* is not called a 'toset', as would seem reasonable, but rather a *chain*.

An element m of a poset Z is *maximal* if nothing comes properly 'after' it in the order:

$$(\forall a \in Z), \quad m \preceq a \implies m = a.$$

For example, maximal ideals in a ring are maximal elements in the set of *proper* ideals, that is, ideals $\neq (1)$ (by Proposition III.4.11). An *upper bound* for a subset S of a poset Z is an element $u \in Z$ coming after every element of S:

$$(\forall a \in S), \quad a \preceq u.$$

Notions of 'smallest' (or, rather, 'least') or 'largest' are defined in the evident way.

Posets may or may not have maximal elements, upper bounds, etc.: for example, the family of *finite* subsets of an infinite set does *not* have maximal elements.

All this terminology comes together in the following statement:

Lemma 3.1 (Zorn's lemma). *Let Z be a nonempty poset. Assume that every chain in Z has an upper bound in Z; then there exists a maximal element in Z.*

The status of Zorn's lemma is peculiar: on the one hand, it is complex enough that no one we know of finds it 'intuitively clear'; on the other hand, it turns out to be logically *equivalent*[10] to the axiom of choice, which *does* look reasonable to most people, and to the 'well-ordering theorem', which we find intuitively *un*reasonable. We will not prove all these equivalences (the diligent reader will prove them in the exercises and should in any case have no trouble locating more detailed proofs); but we will attempt to describe the situation in general terms and explain why the unreasonable statement implies the other two.

The *axiom of choice* states that if $\mathscr{F}$ is a family of disjoint nonempty subsets of a set Z, then we can form a new set by selecting one element x from each $X \in \mathscr{F}$. This may sound reasonable, but it raises rather subtle points: for example, it can be shown that this axiom is independent of the other axioms of (Zermelo-Fränkel) set theory. Also, it has disturbingly counterintuitive consequences, such as the Banach-Tarski paradox[11].

The subtlety of the axiom of choice boils down to the following question: *how does one* choose *the element x?* This is not controversial if Z (and hence $\mathscr{F}$) is finite, but, not surprisingly, it becomes an issue if Z is infinite.

A suitable order relation on Z would come in handy here: if $\preceq$ is an order relation on Z such that every nonempty subset of Z has a least element, then we could simply let x be the least element of X for each $X \in \mathscr{F}$. We say that Z is *well-ordered* by $\preceq$, or that $\preceq$ is a *well-ordering* on Z, if this is the case. (The abbreviation *woset* is also used in the literature, but not very often.) For example, the set $\mathbb{Z}^{>0}$ of positive numbers *is* well-ordered by $\leq$; this fact is called the *well-ordering principle.*

Thus, the statement of the axiom of choice is *really* 'clear' if the set Z is the set $\mathbb{Z}^{>0}$ of positive numbers. It may seem somewhat less so for the set $\mathbb{Z}$, since $\mathbb{Z}$ is not well-ordered by $\leq$; however, it takes a moment (Exercise 3.4) to construct a

[10] *Equivalent* in the sense that each can be derived from the other together with the other axioms of 'Zermelo-Fränkel' set theory.

[11] One can subdivide a solid ball of radius 1 into finitely many pieces, then reassemble them after rotations and translations in 3-space and obtain *two* balls of radius 1.

different relation on $\mathbb{Z}$, making $\mathbb{Z}$ a well-ordered set. Thus, the axiom of choice is also completely transparent for $Z = \mathbb{Z}$ or any countable set (like $\mathbb{Q}$) for that matter.

The well-ordering principle is at the basis of proofs by induction: in fact (cf. Exercise 3.5) it is equivalent to the so-called *principle of induction.* More is true, however. If $(Z, \preceq)$ is *any* woset, then we can consider the following 'principle of induction' on Z:

Let $S \subseteq Z$ be a subset such that $\forall a \in Z$

$$((\forall b \in Z), \quad b \prec a \implies b \in S) \implies a \in S;$$

then $S = Z$.

That is, Z is the only set S with the property that $a \in S$ if all $b \prec a$ are in S. (Note that the least element a of Z is automatically in S, since the condition on all $b \prec a$ is vacuously true in this case.)

The reader will recognize that for $Z = \mathbb{Z}^{>0}$ with the relation $\leq$, this is the ordinary principle of induction.

Claim 3.2. *Let Z be any woset. Then the principle of induction holds for Z.*

Proof. Let $S \subseteq Z$ be a subset with the property given above, and assume $S \subsetneq Z$. Then the complement T of S in Z is nonempty; hence it has a least element a. Now if $b \prec a$, then necessarily $b \in S$ (otherwise $b \in T$, contradicting the fact that a is the least element in T). But the property of S then forces $a \in S$, contradiction. Therefore T is empty; i.e., $S = Z$. □

This is remarkable, since it extends induction to uncountable sets[12], provided a well-ordering is available.

For example, if we had a well-ordering on $\mathbb{R}$, then we could prove a statement about all reals 'by induction'. Many people find this somewhat counterintuitive, and hence so seems the following amazing claim:

Theorem 3.3 (Well-ordering theorem)**.** *Every set admits a well-ordering.*

As argued above, this implies that the axiom of choice holds on every set: if you accept the well-ordering theorem, the statement of the axiom of choice becomes as evident for every set as it is for the set of positive integers. The catch is of course that the axiom of choice is *used* in the proof of the well-ordering theorem; in fact, the statements are equivalent.

In any case, the statement of the well-ordering theorem is easy to absorb (even if perhaps less 'intuitively clear' than the axiom of choice). The well-ordering theorem is equivalent to Zorn's lemma, since they are both equivalent to the axiom of choice. The good news is that the derivation of Zorn's lemma from the well-ordering theorem is reasonably straightforward:

Well-ordering theorem $\implies$ Zorn's lemma. Let $(Z, \leq)$ be a nonempty poset such that every chain in Z has an upper bound in Z. By the well-ordering theorem, there

[12]Even beyond; in the context of *ordinals* this induction principle is called *transfinite induction.*

is a well-ordering[13] $\preceq$ on Z. Define a function f from Z to the power set of Z as follows:

$$f(a) = \begin{cases} \{a\} \text{ if } (\{a\} \cup \bigcup_{b \prec a} f(b)) \textit{ is} \text{ totally ordered by } \leq; \\ \emptyset \text{ if } (\{a\} \cup \bigcup_{b \prec a} f(b)) \textit{ is not} \text{ totally ordered by } \leq. \end{cases}$$

The fact that Z is well-ordered by $\preceq$ implies that f is defined for every $a \in Z$. Indeed, let T be the set of elements of Z for which f is not defined; if $T \neq \emptyset$, T has a least element a w.r.t. $\preceq$; but then $f(b)$ is defined for all $b \prec a$, and the prescription given for $f(a)$ defines f at a, a contradiction.

Let $S = \bigcup_{a \in Z} f(a)$. It is clear that S is totally ordered by $\leq$: if $a, b \in S$, and (say) $b \prec a$, then both a and b belong to

$$\{a\} \cup \bigcup_{b \prec a} f(b),$$

which is totally ordered by $\leq$ by construction.

We claim that S is a *maximal* totally ordered subset[14] of Z:

Claim 3.4. *If $S \subseteq S' \subseteq Z$ and S' is totally ordered by $\leq$, then $S = S'$.*

Indeed, let T be the complement of S in S'. If T is nonempty, let $a \in T$ and observe that

$$\{a\} \cup \bigcup_{b \prec a} f(b)$$

is totally ordered by $\leq$, since it is a subset of $\{a\} \cup S \subseteq S'$. But then $f(a) = \{a\}$, that is, $a \in S$, a contradiction.

Since S is a chain, by the hypothesis of Zorn's lemma S has an upper bound, m. It is now clear that m must be maximal in Z w.r.t. $\leq$, verifying the statement of Zorn's lemma. Indeed, if $m' \geq m$, then $S \cup \{m'\}$ is totally ordered; hence $S = S \cup \{m'\}$ by the claim. This means $m' \in S$, and hence $m' = m$ since m is an upper bound for S. □

Zorn's lemma is the key to several basic results in algebra, the first of which we are about to encounter; the reader will sample a few more results in the exercises. The reader will also encounter equally basic applications of Zorn's lemma (or of other manifestations of the axiom of choice) in other fields: for example Tychonoff's theorem in topology and the Hahn-Banach theorem in functional analysis.

3.2. Application: Existence of maximal ideals. Recall (§III.4.3) that an ideal $\mathfrak{m}$ of a ring R is *maximal* if and only if $R/\mathfrak{m}$ is a field, if and only if no other ideal stands between I and $R = (1)$, that is, if and only if $\mathfrak{m}$ is maximal with respect to inclusion, in the family of proper ideals of R. It is not obvious from this definition that maximal ideals *exist*, but they do:

[13] Watch out: we are considering *two* orderings on Z: $\leq$ (about which we want to say something) and $\preceq$ (about which we know something already).

[14] The existence of maximal totally ordered subsets is known as the *Hausdorff maximal principle;* it is also equivalent to the axiom of choice.

Proposition 3.5. *Let $I \neq (1)$ be a proper ideal of a commutative ring R. Then there exists a maximal ideal $\mathfrak{m}$ of R containing I.*

This is immediate if R is Noetherian, by applying condition (3) from Proposition 1.1 to the family of proper ideals of R containing I. The argument for arbitrary rings is a classic application of Zorn's lemma. In fact, this is the first application given by M. Zorn in his 1935 article introducing a 'maximum principle' (now known as Zorn's lemma). The result had earlier been proven by Krull, using the well-ordering theorem.

Proof. The set $\mathscr{I}$ of proper ideals of R containing I is ordered by inclusion. Then let $\mathscr{C}$ be a chain of proper ideals, and consider

$$U := \bigcup_{J \in \mathscr{C}} J.$$

We claim that U is a proper ideal containing I; hence it is an upper bound for $\mathscr{C}$ in $\mathscr{I}$. This proves that every chain in $\mathscr{I}$ has an upper bound, and it follows that $\mathscr{I}$ has maximal elements, by Zorn's lemma.

To verify our claim, it is clear that U contains I and that it is an ideal (for example, if $a, b \in U$, then $\exists J \in \mathscr{I}$ such that $a, b \in J$; hence $a \pm b \in J$, and therefore $a \pm b \in U$). We have to check that U is *proper*. But if $U = (1)$, then $1 \in J$ for some $J \in \mathscr{I}$, contradicting the fact that $\mathscr{I}$ consists of proper ideals. □

Note that this argument relies crucially on the fact that R has a multiplicative unit 1, and indeed the stated fact is *not* true in general for rings without 1 (Exercise 3.9). Also, the use of Zorn's lemma is not just a convenient trick; the statement is known to be equivalent to the axiom of choice, by work of W. Hodges.

Exercises

3.1. Prove that every well-ordering is total.

3.2. Prove that a totally ordered set $(Z, \preceq)$ is a woset if and only if every descending chain

$$z_1 \succeq z_2 \succeq z_3 \succeq \cdots$$

in Z stabilizes.

3.3. Prove that the axiom of choice is equivalent to the statement that a set-function is surjective if and only if it has a right-inverse (cf. Exercise I.2.2).

3.4. ▷ Construct explicitly a well-ordering on $\mathbb{Z}$. Explain why you know that $\mathbb{Q}$ can be well-ordered, even without performing an explicit construction. [§3.1]

3.5. ▷ Prove that the (ordinary) principle of induction is equivalent to the statement that $\leq$ is a well-ordering on $\mathbb{Z}^{>0}$. (To prove by induction that $(\mathbb{Z}^{>0}, \leq)$ is well-ordered, assume it is known that 1 is the least element of $\mathbb{Z}^{>0}$ and that $\forall n \in \mathbb{Z}^{>0}$ there are no integers between n and $n+1$.) [§3.1]

3.6. In this exercise assume the truth of Zorn's lemma and the conventional set-theoretic constructions; you will be proving the well-ordering theorem.

Let Z be a nonempty set, and let $\mathscr{Z}$ be the set of pairs $(S, \leq)$ consisting of a subset S of Z and of a *well-ordering* $\leq$ on S. Note that $\mathscr{Z}$ is not empty (singletons can be well-ordered). Define a relation $\preceq$ on $\mathscr{Z}$ by prescribing

$$(S, \leq) \preceq (T, \leq')$$

if and only if $S \subseteq T$, $\leq$ is the restriction of $\leq'$ to S, and every element of S precedes every element of $T \smallsetminus S$ w.r.t. $\leq'$.

- Prove that $\preceq$ is an order relation in $\mathscr{Z}$.
- Prove that every chain in $\mathscr{Z}$ has an upper bound in $\mathscr{Z}$.
- Use Zorn's lemma to obtain a maximal element $(M, \leq)$ in $\mathscr{Z}$. Prove that $M = Z$.

Thus every set admits a well-ordering, as stated in Theorem 3.3.

3.7. In this exercise assume the truth of the axiom of choice and the conventional set-theoretic constructions; you will be proving the well-ordering theorem[15].

Let Z be a nonempty set. Use the axiom of choice to choose an element $\gamma(S) \notin S$ for each proper subset $S \subsetneq Z$. Call a pair $(S, \leq)$ a *γ-woset* if $S \subseteq Z$, $\leq$ is a well-ordering on S, and for every $a \in S$, $a = \gamma(\{b \in S, b < a\})$.

- Show how to begin constructing a γ-woset, and show that all γ-wosets must begin in the same way.

Define an ordering on γ-wosets by prescribing that $(U, \leq'') \preceq (T, \leq')$ if and only if $U \subseteq T$ and $\leq''$ is the restriction of $\leq'$.

- Prove that if $(U, \leq'') \prec (T, \leq')$, then $\gamma(U) \in T$.
- For two γ-wosets $(S, \leq)$ and $(T, \leq')$, prove that there is a maximal γ-woset $(U, \leq'')$ preceding both w.r.t. $\preceq$. (Note: There is no need to use Zorn's lemma!)
- Prove that the maximal γ-woset found in the previous point in fact equals $(S, \leq)$ or $(T, \leq')$. Thus, $\preceq$ is a total ordering.
- Prove that there is a maximal γ-woset $(M, \leq)$ w.r.t. $\preceq$. (Again, Zorn's lemma need not and should not be invoked.)
- Prove that $M = Z$.

Thus every set admits a well-ordering, as stated in Theorem 3.3.

3.8. Prove that every nontrivial finitely generated group has a maximal proper subgroup. Prove that $(\mathbb{Q}, +)$ has no maximal proper subgroup.

3.9. ▷ Consider the rng (= ring without 1; cf. §III.1.1) consisting of the abelian group $(\mathbb{Q}, +)$ endowed with the trivial multiplication $qr = 0$ for all $q, r \in \mathbb{Q}$. Prove that this rng has no maximal ideals. [§3.2]

3.10. ¬ As shown in Exercise III.4.17, every maximal ideal in the ring of continuous real-valued functions on a *compact* topological space K consists of the functions vanishing at a point of K.

[15] We learned this proof from notes of Dan Grayson.

Prove that there are maximal ideals in the ring of continuous real-valued functions on the *real line* that do not correspond to points of the real line in the same fashion. (Hint: Produce a proper ideal that is not contained in any maximal ideal corresponding to a point, and apply Proposition 3.5.) [III.4.17]

3.11. Prove that a UFD R is a PID if and only if every nonzero prime ideal in R is maximal. (Hint: One direction is Proposition III.4.13. For the other, assume that every nonzero prime ideal in a UFD R is maximal, and prove that every maximal ideal in R is principal; then use Proposition 3.5 to relate arbitrary ideals to maximal ideals, and prove that every ideal of R is principal.)

3.12. ¬ Let R be a commutative ring, and let $I \subseteq R$ be a proper ideal. Prove that the set of prime ideals containing I has minimal elements. (These are the *minimal primes* of I.) [1.9]

3.13. ¬ Let R be a commutative ring, and let N be its nilradical (Exercise III.3.12). Let $r \notin N$.

- Consider the family $\mathscr{F}$ of ideals of R that intersect the ideal (r) only at 0. Prove that $\mathscr{F}$ has maximal elements.
- Let I be a maximal element of $\mathscr{F}$. Prove that I is prime.
- Conclude $r \notin N \implies r$ is not in the intersection of all prime ideals of R.

Together with Exercise III.4.18, this shows that the nilradical of a commutative ring R equals the intersection of all prime ideals of R. [III.4.18, VII.2.8]

3.14. ¬ The *Jacobson radical* of a commutative ring R is the intersection of the maximal ideals in R. (Thus, the Jacobson radical contains the nilradical.) Prove that r is in the Jacobson radical if and only if $1 + rs$ is invertible for every $s \in R$. [VI.3.8]

3.15. Recall that a (commutative) ring R is Noetherian if every ideal of R is finitely generated. Assume the seemingly weaker condition that every *prime* ideal of R is finitely generated. Let $\mathscr{F}$ be the family of ideals that are not finitely generated in R. You will prove $\mathscr{F} = \emptyset$.

- If $\mathscr{F} \neq \emptyset$, prove that it has a maximal element I.
- Prove that R/I is Noetherian.
- Prove that there are ideals J_1, J_2 containing I, such that $J_1 J_2 \subseteq I$.
- Give a structure of R/I module to I/J_1J_2 and J_1/J_1J_2.
- Prove that I/J_1J_2 is a finitely generated R/I-module.
- Prove that I is finitely generated, thereby reaching a contradiction.

Thus, a ring is Noetherian if and only if its *prime* ideals are finitely generated.

4. Unique factorization in polynomial rings

We now return to regular programming and study unique factorization in polynomial rings; we will finally establish the fact (already hinted at) that $R[x]$ is a UFD if R is a UFD.

Among the necessary preparatory work we also discuss the *field of fractions* of an integral domain; this is a particular instance of the process of *localization* of a ring (or a module) at a multiplicative subset, which the diligent reader will explore a little more in the exercises.

4.1. Primitivity and content; Gauss's lemma. One issue that we have encountered already, and will encounter again, is the description of the *ideals* of the polynomial ring $R[x]$, given enough information about R. For example, we have proved that the ideals of $k[x]$ are principal if k is a field, and Hilbert's basis theorem shows that all ideals of $R[x]$ are finitely generated if all ideals of R are (Exercise III.4.4 and Lemma 1.3).

An even more naive observation is simply that every ideal I of R generates an ideal of $R[x]$:

$$IR[x] := \{a_0 + a_1x + \cdots + a_dx^d \in R[x] \mid \forall i, a_i \in I\}.$$

Lemma 4.1. *Let R be a ring, and let I be an ideal of R. Then*

$$\frac{R[x]}{IR[x]} \cong \frac{R}{I}[x].$$

The proof of this lemma is a standard application of the first isomorphism theorem and is left to the reader (Exercise 4.1).

Corollary 4.2. *If I is a prime ideal of R, then $IR[x]$ is prime in $R[x]$.*

Proof. If I is prime in R, then R/I is an integral domain; hence so is $R[x]/IR[x] \cong (R/I)[x]$, and therefore $IR[x]$ is prime in $R[x]$. □

We will use this fact in a moment.

The following definitions can be studied for every commutative ring; our main application will be to UFDs.

Definition 4.3. Let R be a commutative ring, and let

$$f = a_0 + a_1x + \cdots + a_dx^d \in R[x]$$

be a polynomial.

- f is *very primitive* if for all prime ideals $\mathfrak{p}$ of R, $f \notin \mathfrak{p}R[x]$.
- f is *primitive* if for all *principal* prime ideals $\mathfrak{p}$ of R, $f \notin \mathfrak{p}R[x]$. ⌟

The notion of 'primitive' polynomial is standard, but it is not usually presented in this way; the standard definition is the equivalent formulation given in Lemma 4.5. As for 'very primitive', this term is essentially a joke—our excuse for bringing up this notion is that it is very natural and that unfortunately some references blur the distinction between this notion and the notion of 'primitive'. We hope to ward off any possible confusion by being rather explicit on this point. *Very primitive* polynomials are primitive, but the converse does not hold in general, even for UFDs (cf. Exercise 4.3).

Perhaps the most important fact about 'primitivity' is the following easy remark.

Lemma 4.4. *Let R be a commutative ring. Then for $f, g \in R[x]$*

$$fg \text{ is primitive} \iff \text{both } f \text{ and } g \text{ are primitive.}$$

Proof. This is an easy consequence of Corollary 4.2:

$$\begin{aligned} fg \text{ primitive} &\iff \forall \mathfrak{p} \text{ prime and principal in } R,\ fg \notin \mathfrak{p}R[x] \\ &\iff \forall \mathfrak{p} \text{ prime and principal in } R,\ f \notin \mathfrak{p}R[x] \text{ and } g \notin \mathfrak{p}R[x] \\ &\iff f \text{ is primitive and } g \text{ is primitive} \end{aligned}$$

since $\mathfrak{p}R[x]$ is prime if $\mathfrak{p}$ is a prime ideal. □

The analogous equivalence holds for *very* primitive polynomials (Exercise 4.4). Lemma 4.4 will be ultimately responsible for the fact that $R[x]$ is a UFD if R is a UFD, which is our main objective in this section. If R is a UFD, the notion of 'primitive' has the following interpretation; we accompany it with the 'very primitive' case, for comparison.

Lemma 4.5. *Let R be a commutative ring and $f = a_0 + a_1 x + \cdots + a_d x^d \in R[x]$ as above.*

- *f is very primitive if and only if $(a_0, \ldots, a_d) = (1)$.*
- *If R is a UFD, then f is primitive if and only if $\gcd(a_0, \ldots, a_d) = 1$.*

Proof. If $(a_0, \ldots, a_d) = (1)$, then no prime ideal can contain all coefficients a_i, and it follows that f is very primitive. Conversely, if f is very primitive, then the coefficients of f are not all contained in any one prime ideal, and in particular they are not all contained in any maximal ideal (since maximal ideals are prime). Thus, $(a_0, \ldots, a_d) = (1)$ in this case, since every proper ideal *is* contained in a maximal ideal by Proposition 3.5. This proves the first point.

For the second point note that, in a UFD, $\gcd(a_0, \ldots, a_d) \neq 1$ if and only if there exists an irreducible element $q \in R$ such that $(a_0, \ldots, a_d) \subseteq (q)$. As (q) is then prime (by Lemma 2.4), the second point follows as well. □

With this in mind, in order to capitalize on Lemma 4.4, it is convenient to give a name to the gcd of the coefficients of a polynomial. We now assume that R is a UFD, since this is the case of greatest interest in the applications.

Definition 4.6. Let R be a UFD. The *content* of a nonzero polynomial $f \in R[x]$, denoted cont_f, is the gcd of its coefficients. ⌟

Of course the content of a polynomial is only defined up to units. We find this ambiguity distasteful. What is uniquely determined is the *principal ideal* generated by the content of f, which we will denote

$$(\text{cont}_f):$$

thus f is primitive precisely when $(\text{cont}_f) = (1)$. We will take a somewhat stubborn approach and deal with principal ideals throughout, rather with individual (but only defined up to unit) elements. The reader should keep in mind that ideals may be multiplied (cf. §III.4.1), and if (a), (b) are principal ideals, then their product $(a)(b)$ is the principal ideal (ab).

The proof of the following remarks is immediate from the definition; hence it is left to the reader (Exercise 4.5):

Lemma 4.7. *Let R be a UFD, and let $f \in R[x]$. Then*

- $(f) = (\mathrm{cont}_f)(\underline{f})$, *where $\underline{f}$ is primitive;*
- *if $(f) = (c)(g)$, with $c \in R$ and g primitive, then $(c) = (\mathrm{cont}_f)$.*

In view of its consequences, the following fact is rather profound, and it therefore deserves a name. Some references call the result of Lemma 4.4, or its main consequence (Theorem 4.14), *Gauss's lemma.* We prefer to use this name for the following statement.

Proposition 4.8 (Gauss's lemma)**.** *Let R be a UFD, and let $f, g \in R[x]$. Then*

$$(\mathrm{cont}_{fg}) = (\mathrm{cont}_f)(\mathrm{cont}_g).$$

Proof. This follows easily from our preparatory work. Write

$$(fg) = ((\mathrm{cont}_f)(\underline{f}))\,((\mathrm{cont}_g)(\underline{g})) = (\mathrm{cont}_f)(\mathrm{cont}_g)\,(\underline{fg}),$$

with $\underline{f}$, $\underline{g}$ primitive. By Lemma 4.4, $\underline{fg}$ is primitive; by Lemma 4.7 it follows that $(\mathrm{cont}_f)(\mathrm{cont}_g)$ is the content of fg, as needed. □

Note the following immediate consequence:

Corollary 4.9. *Let R be a UFD, and let $f, g \in R[x]$. Assume $(f) \subseteq (g)$. Then $(\mathrm{cont}_f) \subseteq (\mathrm{cont}_g)$.*

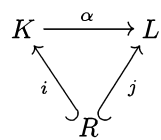
4.2. The field of fractions of an integral domain. Gauss's lemma is the key to the important observation that if R is a UFD, then so is $R[x]$. However, we need one more tool before we can prove this fact; this is one instance of the important process of *localization.*

In the case we need for our immediate goal, the process starts from any integral domain and produces a *field,* in precisely the same way the field $\mathbb{Q}$ of rational numbers may be obtained from the integral domain $\mathbb{Z}$. This construction is known as the *field of fractions* (or *field of quotients*) of the integral domain R. It satisfies the following beautiful universal property.

Given an integral domain R, consider the category $\mathscr{R}$ whose objects are pairs

$$(i, K),$$

where K is a field and $i : R \hookrightarrow K$ is an injective ring homomorphism. Morphisms are defined in the style of every analogous construction we have encountered: that is, a morphism $(i, K) \to (j, L)$ is determined by a homomorphism of fields $\alpha : K \to L$ making the following diagram



commute.

Note that α, as every homomorphism of *fields*, is necessarily injective (Exercise III.3.10); this is compatible with the requirement that $i : R \hookrightarrow K$ be injective to begin with. The injectivity of $R \hookrightarrow K$ also forces R to be an integral domain, since subrings of fields are necessarily integral domains.

Definition 4.10. The *field of fractions* $K(R)$ of R is an initial object of the category $\mathscr{R}$. ⌟

Thus, $K(R)$ is the 'smallest field containing R'.

The usual caveats apply to any such definition: the initial object of $\mathscr{R}$ carries not only the information of a field K (the field of fractions proper), but also of a specific realization of R as a subring of K; this distinction is blurred in common use of the language. Also, of course this prescription only defines the field of fractions up to isomorphism (as is true of any universal object; cf. Proposition I.5.4).

Finally, just stating the universal property does not guarantee that the sought-for initial object exists. Thus, our next task is the construction of an initial object for $\mathscr{R}$. Fortunately, the construction is essentially straightforward: we just need to formalize a notion of 'fraction' of elements of R.

Consider the set $R \times (R^*)$ of pairs (a, r) of elements of R, where $r \neq 0$. The pair (a, r) will be associated with a 'fraction' $\frac{a}{r}$; the reader would be well-advised to put this book away now and carry out the construction of $K(R)$ on his/her own, profiting from this major hint.

Here is the construction. Denote by $\dfrac{a}{r}$ the equivalence class of $(a, r) \in R \times (R^*)$ with respect to the following equivalence relation:

$$(a, r) \sim (b, s) \iff as - br = 0.$$

The reader will verify that this is indeed an equivalence relation. As a set, $K(R)$ is defined by

$$K(R) := \left\{ \frac{a}{r} \mid a \in R, r \in R, r \neq 0 \right\}.$$

It is clear that these 'fractions' behave much as ordinary fractions. For example,

$$\frac{as}{rs} = \frac{a}{r}$$

if $s \neq 0$: indeed,

$$(as)r = a(rs)$$

by associativity and commutativity in R, and this shows

$$(as, rs) \sim (a, r)$$

as needed.

We define operations on $K(R)$ as follows:

$$\frac{a}{r} + \frac{b}{s} = \frac{as + br}{rs},$$
$$\frac{a}{r} \cdot \frac{b}{s} = \frac{ab}{rs}.$$

Of course one must verify that these operations are well-defined; this is also left to the reader. The fact that R is an integral domain is used in these definitions: it guarantees that $rs \neq 0$ if $r \neq 0$ and $s \neq 0$.

Now we claim that $K(R)$ is made into a field by these operations. This is a straightforward verification; for example, distributivity amounts to the following computation:

$$\frac{a}{r}\left(\frac{b}{s}+\frac{c}{t}\right) = \frac{a}{r}\frac{(bt+cs)}{st} = \frac{a(bt+cs)}{r(st)} = \frac{a(bt)}{r(st)} + \frac{a(cs)}{r(st)} = \frac{ab}{rs} + \frac{ac}{rt}$$
$$= \frac{a}{r}\frac{b}{s} + \frac{a}{r}\frac{c}{t}.$$

The zero element is the fraction $\frac{0}{1}$ (or in fact $\frac{0}{r}$ for any nonzero $r \in R$); the multiplicative identity is the fraction $\frac{1}{1}$ (or in fact $\frac{r}{r}$ for any nonzero $r \in R$). Since

$$\frac{r}{s}\frac{s}{r} = \frac{rs}{rs} = \frac{1}{1}$$

for all $\frac{r}{s} \neq \frac{1}{0}$ (that is, for all $r \neq 0$), every nonzero element in $K(R)$ has an inverse, as promised.

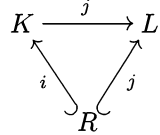
An 'inclusion map' $i : R \hookrightarrow K(R)$ is defined by

$$a \mapsto \frac{a}{1};$$

it is immediately checked that this is an injective ring homomorphism. It is common to use this map to identify R with its isomorphic copy inside $K(R)$ and simply view R as a subring of $K(R)$.

Claim 4.11. *$(i, K(R))$ is initial in $\mathscr{R}$.*

Proof. Let $j : R \hookrightarrow L$ be any injective ring homomorphism from R to a field L. We need to define an induced homomorphism $\hat{j} : K(R) \to L$ so that the diagram



commutes, and we must show that $\hat{j}$ is unique. Now, the definition of $\hat{j}$ is in fact forced upon us: if $\hat{j}$ exists as a homomorphism, then necessarily

$$\hat{j}\left(\frac{a}{r}\right) = \hat{j}\left(\frac{a}{1}\right)\hat{j}\left(\left(\frac{r}{1}\right)^{-1}\right) = \hat{j}\left(\frac{a}{1}\right)\hat{j}\left(\frac{r}{1}\right)^{-1} = (\hat{j}\circ i(a))(\hat{j}\circ i(r)^{-1})$$
$$= j(a)j(r)^{-1}.$$

Thus $\hat{j}$ is indeed unique, if it exists. On the other hand, the prescription

$$\hat{j}\left(\frac{a}{r}\right) := j(a)j(r)^{-1}$$

does define a function $K(R) \to L$: indeed, if $(a, r) \sim (b, s)$, then

$$as = br$$

in R, hence

$$j(a)j(s) = j(b)j(r)$$

in L, and (note that $j(r)$, $j(s)$ are nonzero in L since r, s are nonzero in R and j is injective)

$$j(a)j(r)^{-1} = j(b)j(s)^{-1},$$

showing that the proposed $\hat{j}$ is well-defined. The reader will verify that it is a ring homomorphism, concluding the proof of the claim. □

Example 4.12. With the notation introduced above, $K(\mathbb{Z}) = \mathbb{Q}$.

The universal property implies immediately that $F \hookrightarrow K(F)$ is an isomorphism if F is itself a field. Thus, the construction adds nothing to $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Z}/p\mathbb{Z}$, etc.

If R is any integral domain, so that $R[x]$ is also an integral domain, $K(R[x])$ is a famous field:

Definition 4.13. The field of *rational functions* with coefficients in R is the field of fractions of the ring $R[x]$. This field is denoted $R(x)$. ⌟

Elements of $R(x)$ are fractions of polynomials

$$\frac{p(x)}{q(x)}$$

with $p(x), q(x) \in R[x]$ and $q(x) \neq 0$. The term *function* is inaccurate, since the 'function' $R \to R$ given by $a \mapsto \frac{p(a)}{q(a)}$ is not defined for all $a \in R$ (not for those for which $q(a) = 0$, that is); and the function itself does not suffice to determine the element in $R(x)$ (cf. Exercise III.2.7). ⌟

4.3. R UFD $\Longrightarrow$ $R[x]$ UFD. We are now in a position to prove the analogue of Hilbert's basis theorem for unique factorization domains, that is,

Theorem 4.14. *Let R be a UFD; then $R[x]$ is a UFD.*

For example, this result (and an immediate induction) shows that the rings $\mathbb{Z}[x_1, \ldots, x_n]$ and $k[x_1, \ldots, x_n]$ (for k a field) are UFDs. Theorem 4.14 is also often called *Gauss's lemma.*

As a measure of how delicate the statement of Theorem 4.14 is, note that the power series ring $R[[x]]$ is *not* necessarily a UFD if R is a UFD; examples of this phenomenon are however not easy to construct. Of course $k[[x]]$ is a UFD if k is a field, since $k[[x]]$ is a Euclidean domain in this case (Exercise 2.14).

By Theorem 2.5, in order to prove Theorem 4.14, we have to verify that $R[x]$ satisfies the a.c.c. for principal ideals and that every irreducible element in $R[x]$ is prime, provided that R is itself a UFD. The general idea is to reduce these questions to matters in $K[x]$, where $K = K(R)$ is the field of fractions of R: as we know, $K[x]$ *is* a UFD (in fact it is a Euclidean domain, and Euclidean domain $\Longrightarrow$ PID $\Longrightarrow$ UFD, as shown in §2).

The following lemma captures the most crucial ingredient of the interaction between $R[x]$ and $K[x]$:

Lemma 4.15. *Let R be a UFD, and let $K = K(R)$ be its field of fractions. For nonzero $f, g \in R[x]$, denote by (f), (g) the principal ideals $fR[x]$, $gR[x]$ in $R[x]$, and denote by $(f)_K$, $(g)_K$ the principal ideals $fK[x]$, $gK[x]$ in $K[x]$. Assume*

- $(\text{cont}_g) \subseteq (\text{cont}_f)$ *and*
- $(g)_K \subseteq (f)_K$.

Then $(g) \subseteq (f)$.

Proof. Since $(g)_K \subseteq (f)_K$, we have $g = fh$, where $h \in K[x]$. Write $h = \frac{a}{b}\underline{h}$, where $a, b \in R$ and $\underline{h} \in R[x]$ is a *primitive* polynomial: this can be done by collecting common denominators in h, then applying the first point of Lemma 4.7. We then have

$$bg = af\underline{h}$$

in $R[x]$. By Gauss's lemma and since $\underline{h}$ is primitive,

$$(a \operatorname{cont}_f) = (b \operatorname{cont} g);$$

and since $(\text{cont}_g) \subseteq (\text{cont}_f)$ by hypothesis, we obtain

$$(a \operatorname{cont}_f) \subseteq (b \operatorname{cont}_f).$$

Since R is an integral domain and $(\text{cont}_f) \neq (0)$, this implies

$$a = bc$$

for some $c \in R$. But then $h = \frac{a}{b}\underline{h} = c\underline{h} \in R[x]$, and $g = fh \in (f)$; that is, $(g) \subseteq (f)$ in $R[x]$, as needed. □

The first application is the following description of the irreducible elements of $R[x]$; this will also be used in the proof of Theorem 4.14 and is independently interesting.

Proposition 4.16. *Let R be a UFD, and let K be its field of fractions. Let $f \in R[x]$ be a nonconstant, irreducible polynomial. Then f is irreducible as an element of $K[x]$.*

Proof. First note that f is primitive: otherwise we could factor out its content, and f would not be irreducible.

Next, assume $f = gh$, with $g, h \in K[x]$; we have to prove that either g or h is a unit in $K[x]$. Let $c, d \in K$ such that

$$g = c\underline{g}, \quad h = d\underline{h},$$

and $\underline{g}$, $\underline{h}$ are *primitive* polynomials in $R[x]$. By Lemma 4.4, $\underline{gh}$ is also primitive; thus $(\text{cont}_{\underline{gh}}) = (1) = (\text{cont}_f)$; further,

$$(f)_K = (\underline{gh})_K$$

as $cd \neq 0$ is a unit in K. By Lemma 4.15 we obtain

$$(f) = (\underline{gh})$$

as ideals of $R[x]$; that is, $f = u\underline{gh}$ with $u \in R[x]$ a unit. As f is irreducible in $R[x]$, this implies that either $\underline{g}$ or $\underline{h}$ is a unit in $R[x]$. But then g or h were units in $K[x]$, verifying that f is irreducible in $K[x]$. □

We have always found this fact almost counterintuitive: K is 'larger' than R, so one may expect that it should be 'easier' to factor polynomials over K than over R. Proposition 4.16 tells us that this is not the case: with due attention to special cases, irreducibility in $R[x]$ is 'the same as' irreducibility in $K[x]$. To be precise,

Corollary 4.17. *Let R be a UFD and K the field of fractions of R. Let $f \in R[x]$ be a nonconstant polynomial. Then f is irreducible in $R[x]$ if and only if it is irreducible in $K[x]$ and primitive.*

The proof amounts to tying up loose ends, and we leave it to the reader (Exercise 4.21).

We can now prove the main result of this section. We will make systematic use of the characterization of UFDs found in Theorem 2.5.

Proof of Theorem 4.14. We begin by verifying the a.c.c. for principal ideals in $R[x]$. Let

$$(f_1) \subseteq (f_2) \subseteq (f_3) \subseteq \cdots$$

be an ascending chain of principal ideals of $R[x]$. By Corollary 4.9, this induces an ascending chain of principal ideals

$$(\mathrm{cont}_{f_1}) \subseteq (\mathrm{cont}_{f_2}) \subseteq (\mathrm{cont}_{f_3}) \subseteq \cdots$$

in R; since R is a UFD, this chain stabilizes: that is, $(\mathrm{cont}_{f_i}) = (\mathrm{cont}_{f_{i+1}})$ for[16] $i \gg 0$. On the other hand, with notation as in Lemma 4.15 we have

$$(f_1)_K \subseteq (f_2)_K \subseteq (f_3)_K \subseteq \cdots$$

as a sequence of ideals in $K[x]$; since $K[x]$ is a UFD (because it is a PID; cf. §2.3) this sequence stabilizes. Therefore, $(f_i)_K = (f_{i+1})_K$ for $i \gg 0$.

By Lemma 4.15, $(f_i) = (f_{i+1})$ for $i \gg 0$; that is, the given chain of principal ideals stabilizes, as needed.

Next, we consider an irreducible element f of $R[x]$. We verify that (f) is a prime ideal; by Theorem 2.5 it then follows that $R[x]$ is a UFD, as stated.

If f is irreducible and constant, then f is prime in R as R is a UFD, and it follows that f is prime in $R[x]$ (by Corollary 4.2). Thus we may assume that f is nonconstant and irreducible (and in particular primitive) in $R[x]$.

By Proposition 4.16, f is irreducible as an element of $K[x]$; since $K[x]$ is a PID, $(f)_K$ is prime in $K[x]$. Consider the composition

$$\rho : R[x] \hookrightarrow K[x] \twoheadrightarrow \frac{K[x]}{(f)_K} :$$

We claim that $\ker \rho = (f)$. Indeed, the inclusion $\supseteq$ is trivial; for the other inclusion, note that $\rho(g) = 0$ implies that g is divisible by f *in* $K[x]$: that is, $(g)_K \subseteq (f)_K$; and we have $(\mathrm{cont}_g) \subseteq (\mathrm{cont}_f)$ since $(\mathrm{cont}_f) = (1)$ as f is primitive. By Lemma 4.15, we obtain that $(g) \subseteq (f)$, i.e., g is divisible by f *in* $R[x]$, as needed. Since $\ker \rho = (f)$, we find that ρ induces an *injective* homomorphism

$$\frac{R[x]}{(f)} \hookrightarrow \frac{K[x]}{(f)_K}.$$

[16] 'For $i \gg 0$' is shorthand for $(\exists N \geq 0)\,(\forall i \geq N)\ \ldots$, that is, 'for all sufficiently large i...'.

Since the ring on the right is an integral domain (as $(f)_K$ is prime in $K[x]$), so is the ring on the left. This proves that (f) is prime in $R[x]$, and we are done. □

Summarizing, if R is a UFD, then factorization in $R[x]$ is 'the same as' factorization in the polynomial ring $K[x]$ over the field of quotients of R. If $f(x) \in R[x]$, then $f(x)$ has a prime factorization in $K[x]$ for the simpler reason that $K[x]$ is a PID, hence a UFD; but if R itself is a UFD, then we know that each of the factors may be assumed to be in $R[x]$ to begin with (cf. Exercise 4.23).

Example 4.18. As mentioned already, Theorem 4.14 implies that several important rings, such as $\mathbb{Z}[x_1, \ldots, x_n]$ or $\mathbb{C}[x_1, \ldots, x_n]$, are UFDs; for example $\mathbb{Z}[x]$ is a UFD, as announced in §2.3. Further, arguing as in Exercise 1.15 to reduce to the case of finitely many indeterminates, it follows that $\mathbb{Z}[x_1, x_2, \ldots]$ is a UFD: this is an example of a non-Noetherian UFD, promised a while back (and illustrating the last missing feature of the picture presented at the beginning of the chapter). ⌟

Exercises

4.1. ▷ Prove Lemma 4.1. [§4.1]

4.2. Let R be a ring, and let I be an ideal of R. Prove or disprove that if I is maximal in R, then $IR[x]$ is maximal in $R[x]$.

4.3. ▷ Let R be a PID, and let $f \in R[x]$. Prove that f is primitive if and only if it is very primitive. Prove that this is not necessarily the case in an arbitrary UFD. [§4.1]

4.4. ▷ Let R be a commutative ring, and let $f, g \in R[x]$. Prove that

$$fg \text{ is very primitive} \iff \text{both } f \text{ and } g \text{ are very primitive.}$$

[§4.1]

4.5. ▷ Prove Lemma 4.7. [§4.1]

4.6. Let R be a PID, and let K be its field of fractions.

- Prove that every element $c \in K$ can be written as a finite sum
$$c = \sum_i \frac{a_i}{p_i^{r_i}}$$
where the p_i are nonassociate irreducible elements in R, $r_i \geq 0$, and a_i, p_i are relatively prime.
- If $\sum_i \frac{a_i}{p_i^{r_i}} = \sum_j \frac{b_j}{q_j^{s_j}}$ are two such expressions, prove that (up to reshuffling) $p_i = q_i$, $r_i = s_i$, and $a_i \equiv b_i \bmod p_i^{r_i}$.
- Relate this to the process of integration by 'partial fractions' you learned about when you took calculus.

4.7. ▷ A subset S of a commutative ring R is a *multiplicative subset* (or *multiplicatively closed*) if (i) $1 \in S$ and (ii) $s, t \in S \implies st \in S$. Define a relation on the set of pairs (a, s) with $a \in R$, $s \in S$ as follows:

$$(a, s) \sim (a', s') \iff (\exists t \in S),\ t(s'a - sa') = 0.$$

Note that if R is an integral domain and $S = R \setminus \{0\}$, then S is a multiplicative subset, and the relation agrees with the relation introduced in §4.2.

- Prove that the relation $\sim$ is an *equivalence* relation.
- Denote by $\frac{a}{s}$ the equivalence class of (a, s), and define the same operations $+$, $\cdot$ on such 'fractions' as the ones introduced in the special case of §4.2. Prove that these operations are well-defined.
- The set $S^{-1}R$ of fractions, endowed with the operations $+$, $\cdot$, is the *localization*[17] *of R at the multiplicative subset S*. Prove that $S^{-1}R$ is a commutative ring and that the function $a \mapsto \frac{a}{1}$ defines a ring homomorphism $\ell : R \to S^{-1}R$.
- Prove that $\ell(s)$ is invertible for every $s \in S$.
- Prove that $R \to S^{-1}R$ is initial among ring homomorphisms $f : R \to R'$ such that $f(s)$ is invertible in R' for every $s \in S$.
- Prove that $S^{-1}R$ is an integral domain if R is an integral domain.
- Prove that $S^{-1}R$ is the zero-ring if and only if $0 \in S$.

[4.8, 4.9, 4.11, 4.15, VII.2.16, VIII.1.4, VIII.2.5, VIII.2.6, VIII.2.12, §IX.9.1]

4.8. ¬ Let S be a multiplicative subset of a commutative ring R, as in Exercise 4.7. For every R-module M, define a relation $\sim$ on the set of pairs (m, s), where $m \in M$ and $s \in S$:

$$(m, s) \sim (m', s') \iff (\exists t \in S),\ t(s'm - sm') = 0.$$

Prove that this is an equivalence relation, and define an $S^{-1}R$-module structure on the set $S^{-1}M$ of equivalence classes, compatible with the R-module structure on M. The module $S^{-1}M$ is the *localization* of M *at* S. [4.9, 4.11, 4.14, VIII.1.4, VIII.2.5, VIII.2.6]

4.9. ¬ Let S be a multiplicative subset of a commutative ring R, and consider the localization operation introduced in Exercises 4.7 and 4.8.

- Prove that if I is an ideal of R such that $I \cap S = \emptyset$, then[18] $I^e := S^{-1}I$ is a proper ideal of $S^{-1}R$.
- If $\ell : R \to S^{-1}R$ is the natural homomorphism, prove that if J is a proper ideal of $S^{-1}R$, then $J^c := \ell^{-1}(J)$ is an ideal of R such that $J \cap S = \emptyset$.
- Prove that $(J^c)^e = J$, while $(I^e)^c = \{a \in R \,|\, (\exists s \in S)\, sa \in I\}$.
- Find an example showing that $(I^e)^c$ need not equal I, even if $I \cap S = \emptyset$. (Hint: Let $S = \{1, x, x^2, \dots\}$ in $R = \mathbb{C}[x, y]$. What is $(I^e)^c$ for $I = (xy)$?)

[4.10, 4.14]

[17] The terminology is motivated by applications to algebraic geometry; see Exercise VII.2.17.

[18] The superscript e stands for 'extension' (of the ideal from a smaller ring to a larger ring); the superscript c in the next point stands for 'contraction'.

4.10. ¬ With notation as in Exercise 4.9, prove that the assignment $\mathfrak{p} \mapsto S^{-1}\mathfrak{p}$ gives an inclusion-preserving bijection between the set of *prime* ideals of R disjoint from S and the set of ideals of $S^{-1}R$. (Prove that $(\mathfrak{p}^e)^c = \mathfrak{p}$ if $\mathfrak{p}$ is a prime ideal disjoint from S.) [4.16]

4.11. ¬ (Notation as in Exercise 4.7 and 4.8.) A ring is said to be *local* if it has a single maximal ideal.

Let R be a commutative ring, and let $\mathfrak{p}$ be a prime ideal of R. Prove that the set $S = R \setminus \mathfrak{p}$ is multiplicatively closed. The localizations $S^{-1}R$, $S^{-1}M$ are then denoted $R_\mathfrak{p}$, $M_\mathfrak{p}$.

Prove that there is an inclusion-preserving bijection between the prime ideals of $R_\mathfrak{p}$ and the prime ideals of R contained in $\mathfrak{p}$. Deduce that $R_\mathfrak{p}$ is a local ring[19]. [4.12, 4.13, VI.5.5, VII.2.17, VIII.2.21]

4.12. ¬ (Notation as in Exercise 4.11.) Let R be a commutative ring, and let M be an R-module. Prove that the following are equivalent[20]:

- $M = 0$.
- $M_\mathfrak{p} = 0$ for every prime ideal $\mathfrak{p}$.
- $M_\mathfrak{m} = 0$ for every maximal ideal $\mathfrak{m}$.

(Hint: For the interesting implication, suppose that $M \neq 0$; then the ideal $\{r \in R \mid (\forall m \in M), rm = 0\}$ is proper. By Proposition 3.5, it is contained in a maximal ideal $\mathfrak{m}$. What can you say about $M_\mathfrak{m}$?) [VIII.1.26, VIII.2.21]

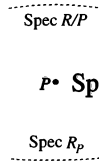
4.13. ¬ Let k be a field, and let v be a discrete valuation on k. Let R be the corresponding DVR, with local parameter t (see Exercise 2.20).

- Prove that R is local (Exercise 4.11), with maximal ideal $\mathfrak{m} = (t)$. (Hint: Note that every element of $R \setminus \mathfrak{m}$ is invertible.)
- Prove that k is the field of fractions of R.
- Now let A be a PID, and let $\mathfrak{p}$ be a prime ideal in A. Prove that the localization $A_\mathfrak{p}$ (cf. Exercise 4.11) is a DVR. (Hint: If $\mathfrak{p} = (p)$, define a valuation on the field of fractions of A in terms of 'divisibility by p'.)

[VII.2.18]

---

[19]We have the following picture in mind associated with the two operations R/P, R_P for a prime ideal P:



Can you make any sense of this?

[20]The way to think of this type of results is that a module M over R is zero if and only if it is zero 'at every point of Spec R'. Working in the localization $M_\mathfrak{p}$ amounts to looking 'near' the point $\mathfrak{p}$ of Spec R; this is what is local about localization. As in this exercise, one can often detect 'global' features by looking locally at every point.

4.14. With notation as in Exercise 4.8, define operations $N \mapsto N^e$ and $\hat{N} \mapsto \hat{N}^c$ for submodules $N \subseteq M$, $\hat{N} \subseteq S^{-1}M$, respectively, analogously to the operations defined in Exercise 4.9. Prove that $(\hat{N}^c)^e = \hat{N}$. Prove that every localization of a Noetherian module is Noetherian.

In particular, all localizations $S^{-1}R$ of a Noetherian ring are Noetherian.

4.15. ¬ Let R be a UFD, and let S be a multiplicatively closed subset of R (cf. Exercise 4.7).

- Prove that if q is irreducible in R, then $q/1$ is either irreducible or a unit in $S^{-1}R$.
- Prove that if a/s is irreducible in $S^{-1}R$, then a/s is an associate of $q/1$ for some irreducible element q of R.
- Prove that $S^{-1}R$ is also a UFD.

[4.16]

4.16. Let R be a Noetherian integral domain, and let $s \in R$, $s \neq 0$, be a prime element. Consider the multiplicatively closed subset $S = \{1, s, s^2, \dots\}$. Prove that R is a UFD if and only if $S^{-1}R$ is a UFD. (Hint: By Exercise 2.10, it suffices to show that every prime of height 1 is principal. Use Exercise 4.10 to relate prime ideals in R to prime ideals in the localization.)

On the basis of results such as this and of Exercise 4.15, one might suspect that being factorial is a local property, that is, that R is a UFD if and only if $R_{\mathfrak{p}}$ is a UFD for all primes $\mathfrak{p}$, if and only if $R_{\mathfrak{m}}$ is a UFD for all maximals $\mathfrak{m}$. This is regrettably not the case. A ring R is *locally factorial* if $R_{\mathfrak{m}}$ is a UFD for all maximal ideals $\mathfrak{m}$; factorial implies locally factorial by Exercise 4.15, but locally factorial rings that are not factorial do exist.

4.17. ▷ Let F be a field, and recall the notion of *characteristic* of a ring (Definition III.3.7); the characteristic of a field is either 0 or a prime integer (Exercise III.3.14.)

- Show that F has characteristic 0 if and only if it contains a copy of $\mathbb{Q}$ and that F has characteristic p if and only if it contains a copy of the field $\mathbb{Z}/p\mathbb{Z}$.
- Show that (in both cases) this determines the smallest subfield of F; it is called the *prime subfield* of F.

[§5.2, §VII.1.1]

4.18. ¬ Let R be an integral domain. Prove that the invertible elements in $R[x]$ are the units of R, viewed as constant polynomials. [4.20]

4.19. ▷ An element $a \in R$ in a ring is said to be *nilpotent* if $a^n = 0$ for some $n \geq 0$. Prove that if a is nilpotent, then $1 + a$ is a unit in R. [VI.7.11, §VII.2.3]

4.20. Generalize the result of Exercise 4.18 as follows: let R be a commutative ring, and let $f = a_0 + a_1 x + \cdots + a_d x^d \in R[x]$; prove that f is a unit in $R[x]$ if and only if a_0 is a unit in R and $a_1, \dots, a_d$ are nilpotent. (Hint: If $b_0 + b_1 x + \cdots + b_e x^e$ is the inverse of f, show by induction that $a_d^{i+1} b_{e-i} = 0$ for all $i \geq 0$, and deduce that a_d is nilpotent.)

4.21. ▷ Establish the characterization of irreducible polynomials over a UFD given in Corollary 4.17. [§4.3]

4.22. Let k be a field, and let f, g be two polynomials in $k[x, y] = k[x][y]$. Prove that if f and g have a nontrivial common factor in $k(x)[y]$, then they have a nontrivial common factor in $k[x, y]$.

4.23. ▷ Let R be a UFD, K its field of fractions, $f(x) \in R[x]$, and assume $f(x) = \alpha(x)\beta(x)$ with $\alpha(x)$, $\beta(x)$ in $K[x]$. Prove that there exists a $c \in K$ such that $c\alpha(x) \in R[x]$, $c^{-1}\beta(x) \in R[x]$, so that

$$f(x) = (c\alpha(x))(c^{-1}\beta(x))$$

splits $f(x)$ as a product of factors in $R[x]$.

Deduce that if $\alpha(x)\beta(x) = f(x) \in R[x]$ is monic and $\alpha(x) \in K[x]$ is monic, then $\alpha(x), \beta(x)$ are both in $R[x]$ and $\beta(x)$ is also monic. [§4.3, 4.24, §VII.5.2]

4.24. In the same situation as in Exercise 4.23, prove that the product of any coefficient of α with any coefficient of β lies in R.

4.25. Prove *Fermat's last theorem for polynomials:* the equation

$$f^n + g^n = h^n$$

has no solutions in $\mathbb{C}[t]$ for $n > 2$ and f, g, h not all constant. (Hint: First, prove that f, g, h may be assumed to be relatively prime. Next, the polynomial $1 - t^n$ factorizes in $\mathbb{C}[t]$ as $\prod_{i=1}^{n}(1 - \zeta^i t)$ for $\zeta = e^{2\pi i/n}$; deduce that $f^n = \prod_{i=1}^{n}(h - \zeta^i g)$. Use unique factorization in $\mathbb{C}[t]$ to conclude that each of the factors $h - \zeta^i g$ is an n-th power. Now let $h - g = a^n$, $h - \zeta g = b^n$, $h - \zeta^2 g = c^n$ (this is where the $n > 2$ hypothesis enters). Use this to obtain a relation $(\lambda a)^n + (\mu b)^n = (\nu c)^n$, where λ, μ, ν are suitable complex numbers. What's wrong with this?)

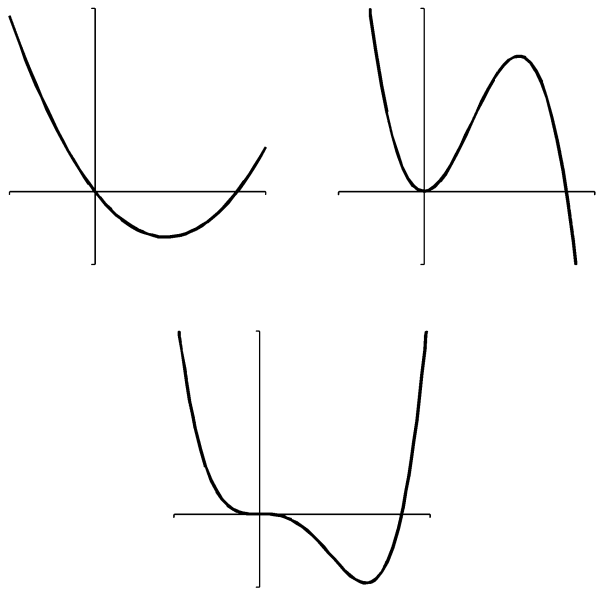
The same pattern of proof would work in any environment where unique factorization is available; if adjoining to $\mathbb{Z}$ an n-th root of 1 lead to a unique factorization domain, the full-fledged Fermat's last theorem would be as easy to prove as indicated in this exercise. This is not the case, a fact famously missed by G. Lamé as he announced a 'proof' of Fermat's last theorem to the Paris Academy on March 1, 1847.

---

5. Irreducibility of polynomials

Our work in §4.3, especially Corollary 4.17, links the irreducibility of polynomials over a UFD with their irreducibility over the corresponding field of fractions. For example, irreducibility of polynomials in $\mathbb{Z}[x]$ is 'essentially the same' as irreducibility in $\mathbb{Q}[x]$. This can be useful in both directions, provided we have ways to effectively verify irreducibility of polynomials over one kind of ring or the other. We collect here several remarks aimed at establishing whether a given polynomial is irreducible and briefly discuss the notion of algebraically closed field.

5.1. Roots and reducibility. Let R be a ring and $f \in R[x]$. An element $a \in R$ is a *root* if $f(a) = 0$. Recall (Example III.4.7) that a polynomial $f(x) \in R[x]$ is divisible by $(x - a)$ if and only if a is a root of f. More generally, we say that a is a root of f with *multiplicity* r if $(x - a)^r$ divides f and $(x - a)^{r+1}$ does *not* divide r. The reader is probably familiar with this notion from experience with calculus. For example, the graph of polynomials with roots of multiplicities 1, 2, and 3 at 0 look (near the origin), respectively, like



Lemma 5.1. *Let R be an integral domain, and let $f \in R[x]$ be a polynomial of degree n. Then the number of roots of f, counted with multiplicity, is at most n.*

Proof. The number of roots of f in R is less than or equal to the number of roots of f viewed as a polynomial over the field of fractions K of R; so we may replace R by K.

Now, $K[x]$ is a UFD, and the roots of f correspond to the irreducible factors of f of degree 1. Since the product of *all* irreducible factors of f has degree n, the number of factors of degree 1 can be at most n, as claimed. □

It is worth remembering the trick of replacing an integral domain R by its field of fractions K, used in this proof: one is often able to use convenient properties due to the fact that K is a field (the fact that $K[x]$ is a UFD, in this case), even if R is very far from satisfying such properties ($R[x]$ may not be a UFD, since R itself need not be a UFD).

Also note that the statement of Lemma 5.1 may fail if R is not an integral domain. For example, the degree 2 polynomial $x^2 + x$ has *four* roots over $\mathbb{Z}/6\mathbb{Z}$.

The innocent Lemma 5.1 has important applications: for example, we used it (without much fanfare) in the proof of Theorem IV.6.10. Also, recall that a polynomial $f \in R[x]$ determines an 'evaluation functions' (cf. Example III.2.3) $R \to R$, namely $r \mapsto f(r)$. The reader checked in Exercise III.2.7 that in general the function does *not* determine the polynomial; but polynomials *are* determined by the corresponding functions over *infinite* integral domains:

Corollary 5.2. *Let R be an* infinite *integral domain, and let $f, g \in R[x]$ be polynomials. Then $f = g$ if and only if the evaluation functions $r \mapsto f(r)$, $r \mapsto g(r)$ agree.*

Proof. Indeed, the two functions agree if and only if every $a \in R$ is a root of $f - g$; but a nonzero polynomial over R cannot have infinitely many roots, by Lemma 5.1. □

Clearly, an irreducible polynomial of degree ≥ 2 can have no roots. The converse holds for degree 2 and 3, over a field:

Proposition 5.3. *Let k be a field. A polynomial $f \in k[x]$ of degree 2 or 3 is irreducible if and only if it has no roots.*

Proof. Exercise 5.5. □

Example 5.4. Let $\mathbb{F}_2$ be the field $\mathbb{Z}/2\mathbb{Z}$.

The polynomial $f(t) = t^2 + t + 1 \in \mathbb{F}_2[t]$ is irreducible, since it has no roots: $f(0) = f(1) = 1$. Therefore the ideal $(t^2 + t + 1)$ is prime in $\mathbb{F}_2[t]$, hence maximal (because $\mathbb{F}_2[t]$ is a PID: don't forget Proposition III.4.13). This gives a one-second construction of a field with four elements:

$$\frac{\mathbb{F}_2[t]}{(t^2 + t + 1)}.$$

The reader (hopefully) constructed this field in Exercise III.1.11, by the tiresome process of searching by hand for a suitable multiplication table. The reader will now have no difficulty constructing much larger examples (cf. Exercise 5.6). ⌟

Proposition 5.3 is pleasant and can help to decide irreducibility over more general rings: for example, a primitive polynomial $f \in \mathbb{Z}[x]$ of degree 2 or 3 is irreducible if and only if it has no roots *in* $\mathbb{Q}$. Indeed, f is irreducible in $\mathbb{Z}[x]$ if and only if it is irreducible in $\mathbb{Q}[x]$ (by Corollary 4.17). However, note that e.g. $4x^2 - 1 = (2x - 1)(2x + 1)$ is primitive and reducible in $\mathbb{Z}[x]$, although it has no *integer* roots. Also, keep in mind that the statement of Proposition 5.3 may fail for polynomials of degree ≥ 4: for example, $x^4 + 2x^2 + 1 = (x^2 + 1)^2$ is *reducible* in $\mathbb{Q}[x]$, but it has *no* rational roots.

Looking for rational roots of a polynomial in $\mathbb{Z}[x]$ is in principle a finite endeavor, due to the following observation (which holds over every UFD). This is often called the 'rational root test'.

Proposition 5.5. *Let R be a UFD, and let K be its field of fractions. Let*

$$f(x) = a_0 + a_1 x + \cdots + a_n x^n \in R[x],$$

and let $c = \frac{p}{q} \in K$ *be a root of* f*, with* $p, q \in R$*,* $gcd(p, q) = 1$*. Then* $p \mid a_0$ *and* $q \mid a_n$ *in* R*.*

Proof. By hypothesis,

$$a_0 + a_1\frac{p}{q} + \cdots + a_n\frac{p^n}{q^n} = 0;$$

that is,

$$a_0q^n + a_1pq^{n-1} + \cdots + a_np^n = 0.$$

Therefore

$$a_0q^n = -p(a_1q^{n-1} + \cdots + a_np^{n-1}),$$

proving that $p \mid (a_0q^n)$. Since the gcd of p and q is one, this implies that the multiset of factors of p is contained in the multiset of irreducible factors of a_0, that is, $p \mid a_0$.

An entirely similar argument proves that $q \mid a_n$. □

Example 5.6. Looking for *rational* roots of the polynomial

$$3 - 2x + 3x^2 - 2x^3 + 3x^4 - 2x^5$$

is therefore reduced to trying fractions $\frac{p}{q}$ with $q = \pm 1, \pm 2$, $p = \pm 1, \pm 3$. As it happens, $\frac{3}{2}$ is the only root found among these possibilities, and it follows that it is the only rational root of the polynomial. ⌟

5.2. Adding roots; algebraically closed fields. A homomorphism of *fields*

$$i : k \to F$$

is necessarily injective (cf. Exercise III.3.10): indeed, its kernel is a proper ideal of k, and the only proper ideal of a field is (0). In this situation we say that F (or more properly the homomorphism i) is an *extension* of k. Abusing language a little, we then think of k as contained in F: $k \subseteq F$. Keep in mind that this is the case as soon as there is *any* ring homomorphism $k \to F$.

There is a standard procedure for constructing extensions in which a given polynomial acquires a root; we have encountered an instance of this process in Example III.4.8. In fact, the resulting field is almost universal with respect to this requirement.

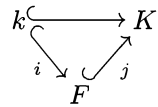
Proposition 5.7. *Let* k *be a field, and let* $f(t) \in k[t]$ *be a nonzero irreducible polynomial. Then*

$$F := \frac{k[t]}{(f(t))}$$

is a field, endowed with a natural homomorphism $i : k \hookrightarrow F$ *(obtained as the composition* $k \to k[x] \to F$*) realizing it as an extension of* k*. Further,*

- $f(x) \in k[x] \subseteq F[x]$ *has a root in* F*, namely the coset of* t*;*

- *if $k \subseteq K$ is* any *extension in which f has a root, then there exists a homomorphism $j : F \to K$ such that the diagram*



commutes.

Proof. Since k is a field, $k[t]$ is a PID; hence $(f(t))$ is a maximal ideal of $k[t]$, by Proposition III.4.13. Therefore F is indeed a field. Denoting cosets in $k[t]/(f(t)) = F$ by underlining, we have

$$f(\underline{t}) = \underline{f(t)} = 0,$$

as claimed.

To verify the second part of the statement, suppose $k \subseteq K$ is an extension and $f(u) = 0$, with $u \in K$. This means that the evaluation homomorphism

$$\epsilon : k[t] \to K$$

defined by $\epsilon(g(t)) := g(u)$ vanishes at $f(t)$; hence $(f(t)) \subseteq \ker(\epsilon)$, and the universal property of quotients gives a unique homomorphism

$$j : F = \frac{k[t]}{(f(t))} \to K$$

satisfying the stated requirement. □

We have stopped short of claiming that the extension constructed in Proposition 5.7 is universal with respect to the requirement of containing a root of f because the homomorphism j appearing in the statement is not *unique:* in fact, the proof shows that there are as many such homomorphisms as there are roots of f in the larger field K. We could say that F is *versal,* meaning 'universal without uni(queness)'. We would have full universality if we included the information of the root of f; we will come back to this construction in Chapter VII, when we analyze field extensions in greater depth.

Example 5.8. For $k = \mathbb{R}$ and $f(x) = x^2+1$, the field constructed in Proposition 5.7 is (isomorphic to) $\mathbb{C}$: this was checked carefully in Example III.4.8.

Similarly, $\mathbb{Q}[t]/(t^2 - 2)$ produces a field containing $\mathbb{Q}$ and in which there is a 'square root of 2'. There are two embeddings of this field in $\mathbb{R}$, because $\mathbb{R}$ contains two distinct square roots of 2: $\pm\sqrt{2}$. ⌟

The fact that the irreducible polynomial $f \in k[x]$ acquires a root in the extension F constructed in Proposition 5.7 implies that f has a linear (i.e., degree 1) factor over F; in particular, if $\deg(f) > 1$, then f is no longer irreducible over F.

Given any polynomial $f \in k[x]$, it is easy to construct an extension of k in which f factors completely as a product of linear factors (Exercise 5.13). The case in which this already happens in k itself for every nonzero f is very important, and hence it is given a name.

Definition 5.9. A field k is *algebraically closed* if all irreducible polynomials in $k[x]$ have degree 1. ⌟

We have encountered this notion in passing, back in Example III.4.14 (and Exercise III.4.21). The following lemma is left to the reader:

Lemma 5.10. *A field k is algebraically closed if and only if every polynomial $f \in k[x]$ factors completely as a product of linear factors, if and only if every nonconstant polynomial $f \in k[x]$ has a root in k.*

In other words, a field k is algebraically closed if the only polynomial equations with coefficients in k and *no* solutions are equations 'of degree 0', such as $1 = 0$. The result known as *Nullstellensatz* (also due to David Hilbert; we have mentioned this result in Chapter III) is a vast generalization of this observation and one of the pillars of (old-fashioned) algebraic geometry. We will come back to all of this in §VII.2.

Is any of the finite fields we have encountered (such as $\mathbb{Z}/p\mathbb{Z}$, for p prime) algebraically closed? No. Adapting Euclid's argument proving the infinitude of prime integers (Exercise 2.24) reveals that algebraically closed fields are infinite.

Proposition 5.11. *Let k be an algebraically closed field. Then k is infinite.*

Proof. By contradiction, assume that k is algebraically closed and finite; let the elements of k be $c_1, \dots, c_N$.

Then there are exactly N irreducible monic polynomials in $k[x]$, namely $(x - c_1), \dots, (x - c_N)$. Consider the polynomial

$$f(x) = (x - c_1) \cdot \dots \cdot (x - c_N) + 1 :$$

for all $c \in k$ we have

$$f(c) = (c - c_1) \cdot \dots \cdot (c - c_N) + 1 = 1 \neq 0,$$

since c equals one of the c_i's. Thus $f(x)$ is a nonconstant polynomial with no roots, contradicting Lemma 5.10. □

Since finite fields are not algebraically closed, the reader may suspect that algebraically closed fields necessarily have characteristic 0 (cf. Exercise 4.17). This is not so: there are algebraically closed fields of any characteristic. In fact, as we will see in due time, *every* field F may be embedded into an algebraically closed field; the smallest such extension is called the 'algebraic closure' of F and is denoted $\overline{F}$. Thus, for example, $\overline{\mathbb{Z}/2\mathbb{Z}}$ is an algebraically closed field of characteristic 2.

The algebraic closure $\overline{\mathbb{Q}}$ is a *countable* subfield of $\mathbb{C}$. The extension $\mathbb{Q} \subseteq \overline{\mathbb{Q}}$, and especially its 'Galois group', cf. §VII.6, is considered one of the most important objects of study in mathematics (cf. the discussion following Corollary VII.7.6).

We will construct the algebraic closure of a field rather explicitly, in §VII.2.1.

5.3. Irreducibility in $\mathbb{C}[x]$, $\mathbb{R}[x]$, $\mathbb{Q}[x]$. Every polynomial $f \in \mathbb{C}[x]$ factors completely over $\mathbb{C}$. Indeed,

Theorem 5.12. *$\mathbb{C}$ is algebraically closed.*

Gauss is credited with providing the first proof[21] of this fundamental theorem (which is indeed known as the *fundamental theorem of algebra.*)

'Algebraic' proofs of the fundamental theorem of algebra require more than we know at this point (we will encounter one in §VII.6, after we have seen a little Galois theory); curiously, a little *complex analysis* makes the statement nearly trivial. Here is a sketch of such an argument. Let $f \in \mathbb{C}[x]$ be a nonconstant polynomial; the task (cf. Lemma 5.10) consists of proving that f has a root in $\mathbb{C}$. Whatever $f(0)$ is, we can find an $r \in \mathbb{R}$ large enough that $|f(z)| > |f(0)|$ for all z on the circle $|z| = r$ (since f is nonconstant, $\lim_{z\to\infty} |f(z)| = +\infty$). The disk $|z| \leq r$ is compact, so the continuous function $|f(z)|$ has a minimum on it; by the choice of r, it must be somewhere in the interior of the disk, say at $z = a$. The *minimum modulus principle* (cf. Exercise 5.16) then implies that $f(a) = 0$, q.e.d.

By Theorem 5.12, irreducibility of polynomials in $\mathbb{C}[x]$ is as simple as it can be: a nonconstant polynomial $f \in \mathbb{C}[x]$ is irreducible if and only if it has degree 1. Every $f \in \mathbb{C}[x]$ of degree ≥ 2 is reducible.

Over $\mathbb{R}$, the situation is almost as simple:

Proposition 5.13. *Every polynomial $f \in \mathbb{R}[x]$ of degree ≥ 3 is reducible.*

The nonconstant irreducible polynomials in $\mathbb{R}[x]$ are precisely the polynomials of degree 1 and the quadratic polynomials

$$f = ax^2 + bx + c$$

with $b^2 - 4ac < 0$.

Proof. Let $f \in \mathbb{R}[x]$ be a nonconstant polynomial:

$$f = a_0 + a_1 x + \cdots + a_n x^n,$$

with all $a_i \in \mathbb{R}$. By Theorem 5.12, f has a *complex* root z:

$$a_0 + a_1 z + \cdots + a_n z^n = 0.$$

Applying complex conjugation $z \mapsto \overline{z}$ and noting that $\overline{a}_i = a_i$ since $a_i \in \mathbb{R}$,

$$a_0 + a_1 \overline{z} + \cdots + a_n \overline{z}^n = \overline{a_0} + \overline{a_1}\,\overline{z} + \cdots + \overline{a_n}\,\overline{z}^n = \overline{a_0 + a_1 z + \cdots + a_n z^n} = 0 :$$

this says that $\overline{z}$ is also a root of f. There are two possibilities:

- either $z = \overline{z}$, that is, $z = r$ was real to begin with, and hence $(x - r)$ is a factor of f in $\mathbb{R}[x]$; or
- $z \neq \overline{z}$, and then $(x - z)$ and $(x - \overline{z})$ are nonassociate irreducible factors of f in $\mathbb{C}[x]$. In this case (since $\mathbb{C}[x]$ is a UFD!)

  $$x^2 - (z + \overline{z})x + z\overline{z} = (x - z)(x - \overline{z})$$

  divides f.

[21] Actually, Gauss's first proof apparently had a gap; the first rigorous proof is due to Argand. Later, Gauss fixed the gap in his proof and provided several other proofs.

Since both $z + \overline{z}$ and $z\overline{z}$ are both real numbers, this analysis shows that every nonconstant $f \in \mathbb{R}[x]$ has an irreducible factor of degree ≤ 2, proving the first statement.

The second statement then follows immediately from Proposition 5.3 and the fact that the quadratic polynomials in $\mathbb{R}[x]$ with no real roots are precisely the polynomials $ax^2 + bx + c$ with $b^2 - 4ac < 0$. □

The following remark is an immediate consequence of Proposition 5.13 and is otherwise evident from elementary calculus (Exercise 5.17):

Corollary 5.14. *Every polynomial $f \in \mathbb{R}[x]$ of* odd *degree has a real root.*

Summarizing, the issue of irreducibility of polynomials in $\mathbb{C}[x]$ or $\mathbb{R}[x]$ is very straightforward. By contrast, irreducibility in $\mathbb{Q}[x]$ is a complicated business: as we have seen (Proposition 4.16), it is just as subtle as irreducibility in $\mathbb{Z}[x]$. There is no sharp characterization of irreducibility in $\mathbb{Z}[x]$; but the following simple-minded remarks (and Eisenstein's criterion, discussed in §5.4) suffice in many interesting cases.

One simple-minded remark is that if $\varphi : R \to S$ is a homomorphism, and a is *reducible* in R, then $\varphi(a)$ is likely to be *reducible* in S. This is just because if $a = bc$, then $\varphi(a) = \varphi(b)\varphi(c)$.

Of course this is not quite right: for example because $\varphi(a)$ and/or $\varphi(b)$ and/or $\varphi(c)$ may be units in S. But with due care for special cases this observation may be used to great effect, especially in the contrapositive form: if $\varphi(a)$ is *irreducible* in S, then a is (likely...) irreducible in R.

Here is a simple statement formalizing these remarks, for $R = \mathbb{Z}[x]$. The natural projection $\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$ induces a surjective homomorphism $\pi : \mathbb{Z}[x] \to \mathbb{Z}/p\mathbb{Z}[x]$: quite simply, $\pi(f)$ is obtained from f by reading all coefficients modulo p; we will write '$f \bmod p$' for the result of this operation.

Proposition 5.15. *Let $f \in \mathbb{Z}[x]$ be a primitive polynomial, and let p be a prime integer. Assume $f \bmod p$ has the same degree as f and is irreducible in $\mathbb{Z}/p\mathbb{Z}[x]$. Then f is irreducible in $\mathbb{Z}[x]$.*

Proof. Argue contrapositively: if f is primitive and reducible in $\mathbb{Z}[x]$ and $\deg f = n$, then $f = gh$ with $\deg g = d$, $\deg h = e$, $d + e = n$, and both d, e, positive. But then the same can be said of $f \bmod p$, so $f \bmod p$ is also reducible. □

The hypothesis on degrees is necessary, precisely to take care of the 'special cases' mentioned in the discussion preceding the statement. For example, $(2x^3 + 3x^2 + 3x + 1)$ equals $(x^2 + x + 1)$ mod 2, and the latter is irreducible in $\mathbb{Z}/2\mathbb{Z}[x]$; yet $(2x + 1)$ is a factor of the former. The point is of course that $2x + 1$ is a unit mod 2.

Warning: As we will see in a fairly distant future (Example VII.5.3), there are irreducible polynomials in $\mathbb{Z}[x]$ which are reducible modulo all primes! So Proposition 5.15 cannot be turned into a characterization of irreducibility in $\mathbb{Z}[x]$. It is however rather useful in producing examples. For instance,

Corollary 5.16. *There are irreducible polynomials in $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$ of arbitrarily large degree.*

Proof. By Proposition 4.16, the statement for $\mathbb{Z}[x]$ implies the one for $\mathbb{Q}[x]$. By Proposition 5.15, it suffices to verify that there are irreducible polynomials in $\mathbb{Z}/p\mathbb{Z}[x]$ of arbitrarily large degree, for any prime integer p. This is a particular case of Exercise 5.11. □

5.4. Eisenstein's criterion. We are giving this result in a separate subsection only on account of the fact that it is very famous; it is also very simple-minded, like the considerations leading up to Proposition 5.15, and further it is not really a 'criterion', in the sense that it also does not give a characterization of irreducibility.

The result is usually stated for $\mathbb{Z}$, but it holds for every ring R.

Proposition 5.17. *Let R be a (commutative) ring, and let $\mathfrak{p}$ be a prime ideal of R. Let*

$$f = a_0 + a_1 x + \cdots + a_n x^n \in R[x]$$

be a polynomial, and assume that

- $a_n \notin \mathfrak{p}$;
- $a_i \in \mathfrak{p}$ *for* $i = 0, \ldots, n-1$;
- $a_0 \notin \mathfrak{p}^2$.

Then f is not the product of polynomials of degree $< n$ in $R[x]$.

Proof. Argue by contradiction. Assume $f = gh$ in $R[x]$, with both $d = \deg g$ and $e = \deg h$ less than $n = \deg f$; write

$$g = b_0 + b_1 x + \cdots + b_d x^d, \quad h = c_0 + c_1 x + \cdots + c_e x^e,$$

and note that necessarily $d > 0$ and $e > 0$. Consider f modulo $\mathfrak{p}$: thus

$$\underline{f} = \underline{g}\,\underline{h} \quad \text{in } (R/\mathfrak{p})[x],$$

where $\underline{f}$ denotes f modulo $\mathfrak{p}$, etc.

By hypothesis, $\underline{f} = \underline{a}_n x^n$ modulo $\mathfrak{p}$, where $\underline{a}_n \neq 0$ in $R/\mathfrak{p}$. Since $R/\mathfrak{p}$ is an integral domain, factors of $\underline{f}$ must also be monomials: that is, necessarily

$$\underline{g} = \underline{b}_d x^d, \quad \underline{h} = \underline{c}_e x^e.$$

Since $d > 0$, $e > 0$, this implies $b_0 \in \mathfrak{p}$, $c_0 \in \mathfrak{p}$.

But then $a_0 = b_0 c_0 \in \mathfrak{p}^2$, contradicting the hypothesis. □

For example, $x^4 + 2x^2 + 2$ must be irreducible in $\mathbb{Z}[x]$ (and hence in $\mathbb{Q}[x]$): apply Eisenstein's criterion with $R = \mathbb{Z}$, $\mathfrak{p} = (2)$. More exciting applications involve whole classes of polynomials:

Example 5.18. For all n and all primes p, the polynomial $x^n - p$ is irreducible in $\mathbb{Z}[x]$. This follows immediately from Eisenstein's criterion and gives an alternative proof of Corollary 5.16. ⌟

Example 5.19. This is probably the most famous application of Eisenstein's criterion. Let p be a prime integer, and let

$$f(x) = 1 + x + x^2 + \cdots + x^{p-1} \in \mathbb{Z}[x].$$

These polynomials are called *cyclotomic;* we will encounter them again in §VII.5.2.

It may not look like it, but Eisenstein's criterion may be used to prove that $f(x)$ is irreducible. The trick (which the reader should endeavor to remember) is to apply the shift $x \to x + 1$. It is hopefully clear that $f(x)$ is irreducible if and only if $f(x+1)$ is; thus we are reduced to showing that

$$f(x+1) = 1 + (x+1) + (x+1)^2 + \cdots + (x+1)^{p-1}$$

is irreducible. This is better than it looks: since $f(x) = (x^p - 1)/(x - 1)$,

$$f(x+1) = \frac{(x+1)^p - 1}{(x+1) - 1} = x^{p-1} + \binom{p}{p-1} x^{p-2} + \cdots + \binom{p}{3} x^2 + \binom{p}{2} x + \binom{p}{1};$$

Eisenstein's criterion proves that this is irreducible, since $\binom{p}{1} = p$ and

Claim 5.20. *For p prime and $k = 1, \ldots, p-1$, p divides* $\binom{p}{k}$.

There is nothing to this, because

$$\binom{p}{k} = \frac{p!}{k!(p-k)!}$$

and p divides the numerator and does not divide the denominator; the claim follows as p is prime. It may be worthwhile noting that this fact does *not* hold if p is not prime: for example, 4 does not divide $\binom{4}{2}$. ⌟

Exercises

5.1. ¬ Let $f(x) \in \mathbb{C}[x]$ prove that a is a root of f with multiplicity r if and only if $f(a) = f'(a) = \cdots = f^{(r-1)}(a) = 0$ and $f^{(r)}(a) = 0$, where $f^{(k)}(a)$ denotes the value of the k-th derivative of f at a. Deduce that $f(x) \in \mathbb{C}[x]$ has multiple roots if and only if $\gcd(f(x), f'(x)) \neq 1$. [5.2]

5.2. Let F be a subfield of $\mathbb{C}$, and let $f(x)$ be an irreducible polynomial in $F[x]$. Prove that $f(x)$ has no multiple roots in $\mathbb{C}$. (Use Exercises 2.22 and 5.1.)

5.3. Let R be a ring, and let $f(x) = a_0 x^{2n} + a_2 x^{2n-2} + \cdots + a_2 x^2 + a_0 \in R[x]$ be a polynomial only involving *even* powers of x. Prove that if $g(x)$ is a factor of $f(x)$, so is $g(-x)$.

5.4. Show that $x^4 + x^2 + 1$ is reducible in $\mathbb{Z}[x]$. Prove that it has *no* rational roots, without finding its (complex) roots.

5.5. ▷ Prove Proposition 5.3. [§5.1]

5.6. ▷ Construct fields with 27 elements and with 121 elements. [§5.1]

5.7. Let R be an integral domain, and let $f(x) \in R[x]$ be a polynomial of degree d. Prove that $f(x)$ is determined by its value at any $d+1$ distinct elements of R.

5.8. ¬ Let K be a field, and let $a_0, \ldots, a_d$ be distinct elements of K. Given any elements $b_0, \ldots, b_d$ in K, construct explicitly a polynomial $f(x) \in K[x]$ of degree d such that $f(a_0) = b_0, \ldots, f(a_d) = b_d$, and show that this polynomial is unique. (Hint: First solve the problem assuming that only one b_i is not equal to zero.) This process is called *Lagrange interpolation.* [5.9]

5.9. ¬ Pretend you can factor integers, and then use Lagrange interpolation (cf. Exercise 5.8) to give a finite algorithm to factor *polynomials*[22] with integer coefficients over $\mathbb{Q}[x]$. Use your algorithm to factor $(x-1)(x-2)(x-3)(x-4)+1$. [5.10]

5.10. Prove that the polynomial $(x-1)(x-2)\cdots(x-n)-1$ is irreducible in $\mathbb{Q}[x]$ for all $n \geq 1$. (Hint: Think along the lines of Exercise 5.9.)

5.11. ▷ Let F be a finite field. Prove that there are irreducible polynomials in $F[x]$ of arbitrarily high degree. (Hint: Exercise 2.24. Note that Example 5.16 takes care of $\mathbb{Z}/p\mathbb{Z}$, but what about all other finite fields?) [§5.3]

5.12. Prove that applying the construction in Proposition 5.7 to an irreducible *linear* polynomial in $k[x]$ produces a field isomorphic to k.

5.13. ▷ Let k be a field, and let $f \in k[x]$ be any polynomial. Prove that there is an extension $k \subseteq F$ in which f factors completely as a product of linear terms. [§5.2, §VI.7.3]

5.14. How many different embeddings of the field $\mathbb{Q}[t]/(t^3-2)$ are there in $\mathbb{R}$? How many in $\mathbb{C}$?

5.15. Prove Lemma 5.10.

5.16. ▷ If you know about the 'maximum modulus principle' in complex analysis: formulate and prove the 'minimum modulus principle' used in the sketch of the proof of the fundamental theorem of algebra. [§5.3]

5.17. ▷ Let $f \in \mathbb{R}[x]$ be a polynomial of *odd* degree. Use the intermediate value theorem to give an 'algebra-free' proof of the fact that f has real roots. [§5.3, §VII.7.1]

5.18. Let $f \in \mathbb{Z}[x]$ be a cubic polynomial such that $f(0)$ and $f(1)$ are odd and with odd leading coefficient. Prove that f is irreducible in $\mathbb{Q}[x]$.

5.19. Give a proof of the fact that $\sqrt{2}$ is not rational by using Eisenstein's criterion.

5.20. Prove that $x^6 + 4x^3 + 1$ is irreducible by using Eisenstein's criterion.

5.21. Prove that $1 + x + x^2 + \cdots + x^{n-1}$ is *re*ducible over $\mathbb{Z}$ if n is *not* prime.

5.22. Let R be a UFD, and let $a \in R$ be an element that is not divisible by the square of some irreducible element in its factorization. Prove that $x^n - a$ is irreducible for every integer $n \geq 1$.

[22] It is in fact much harder to factor integers than integer polynomials.

5.23. Decide whether $y^5 + x^2y^3 + x^3y^2 + x$ is reducible or irreducible in $\mathbb{C}[x, y]$.

5.24. Prove that $\mathbb{C}[x, y, z, w]/(xw - yz)$ is an integral domain, by using Eisenstein's criterion. (We used this ring as an example in §2.2, but we did not have the patience to prove that it was a domain back then!)

---

6. Further remarks and examples

6.1. Chinese remainder theorem. Suppose all you know about an integer is its class modulo several numbers; can you reconstruct the integer? Also, if you are given arbitrary classes in $\mathbb{Z}/n\mathbb{Z}$ for several integers n, can you find an $N \in \mathbb{Z}$ satisfying all these congruences simultaneously?

The answer is no in both cases, for trivial reasons. If an integer N satisfies given congruences modulo $n_1, \ldots, n_k$, then adding any multiple of $n_1 \cdots n_k$ to N produces an integer satisfying the same congruences (so N cannot be entirely reconstructed from the given data); and there is no integer N such that $N \equiv 1 \bmod 2$ and $N \equiv 2 \bmod 4$, so there are plenty of congruences which cannot be simultaneously satisfied.

The *Chinese remainder theorem* (CRT) sharpens these questions so that they can in fact be answered affirmatively, and (in its modern form) it generalizes them to a much broader setting. Its statement is most pleasant for PIDs; it is very impressive[23] even in the limited context of $\mathbb{Z}$. Protoversions of the theorem have already appeared here and there in this book, e.g., Lemma IV.6.1.

We will first give the more general version, which is in some way simpler. Let R be any commutative ring.

Theorem 6.1. *Let $I_1, \ldots, I_k$ be ideals of R such that $I_i + I_j = (1)$ for all $i \neq j$. Then the natural homomorphism*

$$\varphi : R \longrightarrow \frac{R}{I_1} \times \cdots \times \frac{R}{I_k}$$

is surjective and induces an isomorphism

$$\tilde{\varphi} : \frac{R}{I_1 \cdots I_k} \xrightarrow{\sim} \frac{R}{I_1} \times \cdots \times \frac{R}{I_k}.$$

The 'natural' homomorphism φ is determined by the canonical projections $R \to R/I_j$ and the universal property of products; the homomorphism $\tilde{\varphi}$ is induced by virtue of the universal property of quotients, since $I_1 \cdots I_k \subseteq I_j$ for all j, hence $I_1 \cdots I_k \subseteq \ker \varphi$. In fact, clearly

$$\ker \varphi = I_1 \cap \cdots \cap I_k;$$

so the second part of the statement of Theorem 6.1 follows immediately from the first part, the 'first isomorphism theorem', and the following (independently interesting) observation:

[23] Allegedly, a large number of 'Putnam' problems can be solved by clever applications of the CRT.

Lemma 6.2. *Let $I_1, \dots, I_k$ be ideals of R such that $I_i + I_j = (1)$ for all $i \neq j$. Then $I_1 \cdots I_k = I_1 \cap \cdots \cap I_k$.*

Proof. A simple induction reduces the general statement to the case $k = 2$. Therefore, assume I and J are ideals of R, such that $I+J = (1)$. The inclusion $IJ \subseteq I \cap J$ holds for all ideals I, J, so the task amounts to proving $I \cap J \subseteq IJ$ when $I+J = (1)$.

If $I + J = (1)$, then there exist elements $a \in I$, $b \in J$ such that $a + b = 1$. But if $r \in I \cap J$, then

$$r = r \cdot 1 = r(a + b) = ra + rb \in IJ :$$

because $ra \in IJ$ as $r \in J$ and $a \in I$, while $rb \in IJ$ as $r \in I$ and $b \in J$. The statement follows. □

Thus, we only need to prove the first part of Theorem 6.1: the surjectivity of φ under the given hypotheses. The proof of this is a simple exercise but may be tricky to set up. The following remark streamlines it considerably:

Lemma 6.3. *Let $I_1, \dots, I_k$ be ideals of R such that $I_i + I_k = (1)$ for all $i = 1, \dots, k-1$. Then $(I_1 \cdots I_{k-1}) + I_k = (1)$.*

Proof. By hypothesis, for $i = 1, \dots, k-1$ there exists $a_i \in I_k$ such that $1 - a_k \in I_i$. Then

$$(1 - a_1) \cdots (1 - a_{k-1}) \in I_1 \cdots I_{k-1},$$

and

$$1 - (1 - a_1) \cdots (1 - a_{k-1}) \in I_k,$$

because it is a combination of $a_1, \dots, a_{k-1} \in I_k$. □

Proof of Theorem 6.1. Argue by induction on k. For $k = 1$, there is nothing to show. For $k > 1$, assume the statement is known for fewer ideals. Thus, we may assume that the natural projection induces an isomorphism

$$\frac{R}{I_1 \cdots I_{k-1}} \cong \frac{R}{I_1} \times \cdots \times \frac{R}{I_{k-1}};$$

and all that we have left to prove is that the natural homomorphism

$$R \to \frac{R}{I_1 \cdots I_{k-1}} \times \frac{R}{I_k}$$

is surjective. By Lemma 6.3, $(I_1 \cdots I_{k-1}) + I_k = (1)$; thus we are reduced to the case of *two* ideals.

Let then I, J be ideals of a commutative ring R, such that $I + J = (1)$, and let $r_I, r_J \in R$; we have to verify that $\exists r \in R$ such that $r \equiv r_I \bmod I$ and $r \equiv r_J \bmod J$. Since $I + J = (1)$, there are $a \in I$, $b \in J$ such that $a + b = 1$. Let $r = ar_J + br_I$: then

$$r = ar_J + (1 - a)r_I = r_I + a(r_J - r_I) \equiv r_I \mod I$$

as $a \in I$, and

$$r = (1 - b)r_J + br_I = r_J + b(r_I - r_J) \equiv r_J \mod J$$

as $b \in J$, as needed, and completing the proof. □

In a PID, the CRT takes the following form:

Corollary 6.4. *Let R be a PID, and let $a_1, \dots, a_k \in R$ be elements such that $\gcd(a_i, a_j) = 1$ for all $i \neq j$. Let $a = a_1 \cdots a_k$. Then the function*

$$\varphi : \frac{R}{(a)} \to \frac{R}{(a_1)} \times \cdots \times \frac{R}{(a_k)}$$

defined by $r + (a) \mapsto (r + (a_1), \dots, r + (a_k))$ is an isomorphism.

This is an immediate consequence of Theorem 6.1, since (in a PID!) $\gcd(a, b) = 1$ if and only if $(a, b) = (1)$ as ideals. This is not the case for arbitrary UFDs, and indeed the natural map

$$\mathbb{Z}[x] \to \frac{\mathbb{Z}[x]}{(2)} \times \frac{\mathbb{Z}[x]}{(x)}$$

is not surjective (check this!), even though $\gcd(2, x) = 1$. But the kernel of this map *is* $(2x)$; and this is what can be expected in general from the CRT over UFDs (Exercise 6.6).

As $\mathbb{Z}$ is a PID, Corollary 6.4 holds for $\mathbb{Z}$ and gives the promised answer to a revised version of the questions posed at the beginning of this subsection. In fact, tracing the proof in this case gives an effective procedure to solve simultaneous congruences over $\mathbb{Z}$ (or in fact over any Euclidean domain, as will be apparent from the argument).

To see this more explicitly, let $n_1, \dots, n_k$ be pairwise relatively prime integers, and let $n = n_1 \cdots n_k$; for each i, let $m_i = n/n_i$. Then n_i and m_i are relatively prime[24], and we can use the Euclidean algorithm to explicitly find integers a_i, b_i such that

$$a_i n_i + b_i m_i = 1.$$

The numbers $q_i = b_i m_i$ have the property that

$$q_i \equiv 1 \mod n_i, \quad q_i \equiv 0 \mod n_j \quad \forall j \neq i$$

(why?). These integers may be used to solve any given system of congruences modulo $n_1, \dots, n_k$: indeed, if $r_1, \dots, r_k \in \mathbb{Z}$ are given, then

$$N := r_1 q_1 + \cdots + r_k q_k$$

satisfies

$$N \equiv r_1 \cdot 0 + \cdots + r_i \cdot 1 + \cdots + r_k \cdot 0 \equiv r_i \mod n_i$$

for all i.

The same procedure can be applied over any Euclidean domain R; see Exercise 6.7 for an example.

[24] This is a particular case of Lemma 6.3 and is clear anyway from gcd considerations.

6.2. Gaussian integers. The rings $\mathbb{Z}$ and $k[x]$ (where k is a field) may be the only examples of Euclidean domains known to our reader. The next most famous example is the ring of *Gaussian integers;* this is a very pretty ring, and it has elementary but striking applications in number theory (one instance of which we will see in §6.3).

Abstractly, we may define this ring as

$$\mathbb{Z}[i] := \frac{\mathbb{Z}[x]}{(x^2+1)};$$

and we are going to verify that $\mathbb{Z}[i]$ is a Euclidean domain.

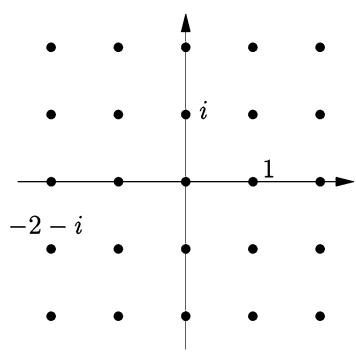
The notation $\mathbb{Z}[i]$ is justified by the discussion in §5.2 (cf. especially Proposition 5.7): $\mathbb{Z}[i]$ may be viewed as the 'smallest' ring containing $\mathbb{Z}$ and a root of x^2+1, that is, a square root i of -1. The natural embedding

$$\frac{\mathbb{Z}[x]}{(x^2+1)} \subseteq \frac{\mathbb{R}[x]}{(x^2+1)}$$

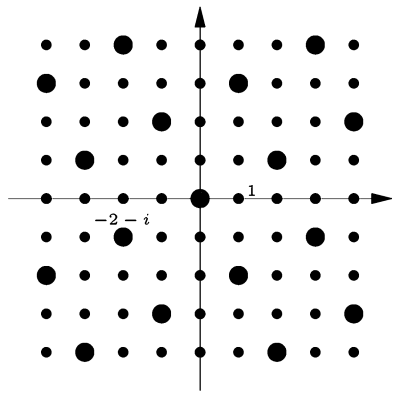
and the identification of the rightmost ring with $\mathbb{C}$ (Example III.4.8) realize $\mathbb{Z}[i]$ as a subring of $\mathbb{C}$; tracing this embedding shows that we could equivalently define

$$\mathbb{Z}[i] = \{a+bi \in \mathbb{C} \mid a, b \in \mathbb{Z}\}.$$

Thus, the reader should think of $\mathbb{Z}[i]$ as consisting of the complex numbers whose real and imaginary parts are *integers;* this may be referred to as the 'integer lattice' in $\mathbb{C}$:



This picture is particularly compelling, because it allows us to 'visualize' principal ideals in $\mathbb{Z}[i]$: simple properties of complex multiplication show that the multiples of a fixed $w \in \mathbb{Z}[i]$ form a regular lattice superimposed on $\mathbb{Z}[i]$. For example, the fattened dots in the picture



represent the ideal $(-2-i)$ in $\mathbb{Z}[i]$: an enlarged, tilted lattice superimposed on the integer lattice in $\mathbb{C}$. The reader should stop reading now and make sure to understand why this works out so neatly.

A Gaussian integer is a complex number, and as such it has a *norm*

$$N(a+bi) := (a+bi)(a-bi) = a^2 + b^2.$$

Geometrically, this is the square of the distance from the origin to $a+bi$.

The norm of a Gaussian integer is a nonnegative integer; thus N is a function

$$\mathbb{Z}[i] \to \mathbb{Z}^{\geq 0}.$$

Lemma 6.5. *The function N is a Euclidean valuation on $\mathbb{Z}[i]$; further, N is multiplicative in the sense that $\forall z, w \in \mathbb{Z}[i]$*

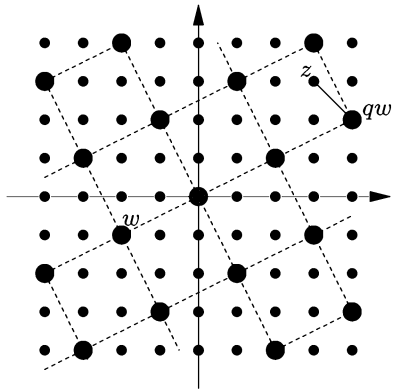
$$N(zw) = N(z)N(w).$$

Proof. The multiplicativity is an immediate consequence of the elementary properties of complex conjugation:

$$N(zw) = (zw)(\overline{zw}) = (z\overline{z})(w\overline{w}) = N(z)N(w).$$

To see that N is a Euclidean valuation, we have to show how to perform a 'division with remainder' in $\mathbb{Z}[i]$. It is not hard, but it is a bit messy, to do this algebraically; we will attempt to convince the reader that this is in fact visually evident (and then the reader can have fun producing the needed algebraic computations).

Let $z, w \in \mathbb{Z}[i]$, and assume $w \neq 0$. The ideal (w) is a lattice superimposed on $\mathbb{Z}[i]$. The given z is either one of the vertices of this lattice (in which case z is a multiple of w, so that the division z/w can be performed in $\mathbb{Z}[i]$, with remainder 0) or it sits inside one of the 'boxes' of the lattice. In the latter case, pick any of the vertices of that box, that is, a multiple qw of w, and let $r = z - qw$. The situation may look as follows:



Then we have obtained

$$z = qw + r,$$

and the norm of r is the square of the length of the segment between qw and z. Since this segment is contained in a box, and the square of the size of the box is $N(w)$, we have achieved

$$N(r) < N(w),$$

completing the proof. □

As we know, all sorts of amazing properties hold for the ring of Gaussian integers as a consequence of Lemma 6.5: $\mathbb{Z}[i]$ is a PID and a UFD; irreducible elements are prime in $\mathbb{Z}[i]$; greatest common divisors exist and may be found by applying the Euclidean algorithm, etc. Any one of these facts would seem fairly challenging in itself, but they are all immediate consequences of the existence of a Euclidean valuation and of the general considerations in §2.

The fact that the norm is multiplicative simplifies further the analysis of the ring $\mathbb{Z}[i]$.

Lemma 6.6. *The units of* $\mathbb{Z}[i]$ *are* ± 1, $\pm i$.

Proof. If u is a unit in $\mathbb{Z}[i]$, then there exists $v \in \mathbb{Z}[i]$ such that $uv = 1$. But then $N(u)N(v) = N(uv) = N(1) = 1$ by multiplicativity, so $N(u)$ is a unit in $\mathbb{Z}$. This implies $N(u) = 1$, and the only elements in $\mathbb{Z}[i]$ with norm 1 are ± 1, $\pm i$. The statement follows. □

Lemma 6.7. *Let* $q \in \mathbb{Z}[i]$ *be a prime element. Then there is a prime integer* $p \in \mathbb{Z}$ *such that* $N(q) = p$ *or* $N(q) = p^2$.

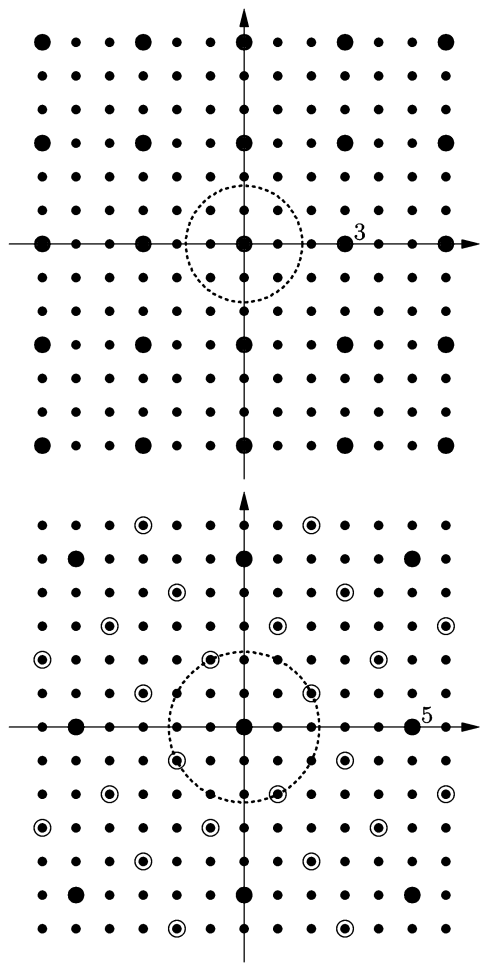
Proof. Since q is not a unit, $N(q) \neq 1$ (by Lemma 6.6). Thus $N(q)$ is a nontrivial product of (integer) primes, and since q is prime in $\mathbb{Z}[i] \supseteq \mathbb{Z}$, q must divide one of the prime integer factors of $N(q)$; let p be this integer prime. But then $q \mid p$ in $\mathbb{Z}[i]$, and it follows (by multiplicativity of the norm) that $N(q) \mid N(p) = p^2$. Since $N(q) \neq 1$, the only possibilities are $N(q) = p$ and $N(q) = p^2$, as claimed. □

Both possibilities presented in Lemma 6.7 occur, and studying this dichotomy further will be key to the result in §6.3. But we can already work out several cases 'by hand':

Example 6.8. The prime integer 3 is a prime element of $\mathbb{Z}[i]$; this can be verified by proving that 3 is irreducible in $\mathbb{Z}[i]$ (since $\mathbb{Z}[i]$ is a UFD). For this purpose, note that since $N(3) = 9$, the norm of a factor of 3 would have to be a divisor of 9, that is, 1, 3, or 9. Gaussian integers with norm 1 are units, and those with norm 9 are associates of 3 (Exercise 6.10); thus a nontrivial factor of 3 would necessarily have norm equal to 3. But there are no such elements in $\mathbb{Z}[i]$; hence 3 is indeed irreducible.

The prime integer 5 is *not* a prime element of $\mathbb{Z}[i]$: running through the same argument as we just did for 3, we find that a factor of 5 should have norm 5, and $2 + i$ does. In fact, $5 = (2 + i)(2 - i)$ is a prime factorization of 5 in $\mathbb{Z}[i]$.

Visually, what happens is that the lattice generated by 3 in $\mathbb{Z}[i]$ cannot be refined further into a tighter lattice, while the one generated by 5 does admit a refinement:



(the circles represent complex numbers with norm 3, 5, respectively). ⌟

The reader is encouraged to work out several other examples and to try to figure out what property of an integer prime makes it split in $\mathbb{Z}[i]$ (like 5 and unlike 3). But this should be done before reading on!

6.3. Fermat's theorem on sums of squares. Once armed with our knowledge about rings, playing with $\mathbb{Z}[i]$ a little further teaches us something new about $\mathbb{Z}$—this is a beautiful and famous example of the success achieved by judicious use of generalization over brute force.

First, let us complete the circle of thought begun with Lemma 6.7. Say that a prime integer p *splits in* $\mathbb{Z}[i]$ if it is *not* a prime element of $\mathbb{Z}[i]$; we have seen in Example 6.8 that 5 splits in $\mathbb{Z}[i]$, while 3 does not.

Lemma 6.9. *A positive integer prime $p \in \mathbb{Z}$ splits in $\mathbb{Z}[i]$ if and only if it is the sum of two squares in $\mathbb{Z}$.*

Proof. First assume that $p = a^2 + b^2$, with $a, b \in \mathbb{Z}$. Then

$$p = (a + bi)(a - bi)$$

in $\mathbb{Z}[i]$, and $N(a \pm bi) = a^2 + b^2 = p \neq 1$, so neither of the two factors is a unit in $\mathbb{Z}[i]$ (Lemma 6.6). Thus p is not irreducible, hence not prime, in $\mathbb{Z}[i]$.

Conversely, assume that p is not irreducible in $\mathbb{Z}[i]$: then it has an irreducible factor $q \in \mathbb{Z}[i]$, which is not an associate of p. Since $q \mid p$, by multiplicativity of the norm we have $N(q) \mid N(p) = p^2$, and hence $N(q) = p$ since q and p are not associates (thus $N(q) \neq p^2$) and q is not a unit (thus $N(q) \neq 1$). If $q = a + bi$, we find

$$p = N(q) = a^2 + b^2,$$

verifying that p is the sum of two squares and completing the proof. □

Thus, 2 splits (as $2 = 1^2 + 1^2$), and so do

$$5 = 1^2 + 2^2, \quad 13 = 2^2 + 3^2, \quad 17 = 1^2 + 4^2,$$

while

$$3, \quad 7, \quad 11, \quad 19, \quad 23,$$

remain prime if viewed as elements of $\mathbb{Z}[i]$.

The next puzzle is as follows: what *else* distinguishes the primes in the first list from the primes in the second list? Again, the reader who does not know the answer already should pause and try to come up with a conjecture and then prove that conjecture! Do not read the next statement before trying this on your own[25]!

[25]The point we are trying to make is that these are not difficult facts, and a conscientious reader really *is* in the position of discovering and proving them on his or her own. This is extremely remarkable: the theorem we are heading towards was stated by Fermat, without proof, in 1640, and had to wait about one hundred years before being proven rigorously (by Euler, using 'infinite descent'). Proofs using Gaussian integers are due to Dedekind and had to wait another hundred years. The moral is, of course, that the modern algebraic apparatus acts as a tremendous amplifier of our skills.

Lemma 6.10. *A positive* odd *prime integer p splits in $\mathbb{Z}[i]$ if and only if it is congruent to* 1 *modulo* 4.

Proof. The question is whether p is prime as an element of $\mathbb{Z}[i]$, that is, whether $\mathbb{Z}[i]/(p)$ is an integral domain. But we have isomorphisms

$$\frac{\mathbb{Z}[i]}{(p)} \cong \frac{\mathbb{Z}[x]/(x^2+1)}{(p)} \cong \frac{\mathbb{Z}[x]}{(p, x^2+1)} \cong \frac{\mathbb{Z}[x]/(p)}{(x^2+1)} \cong \frac{\mathbb{Z}/p\mathbb{Z}[x]}{(x^2+1)}$$

by virtue of the usual isomorphism theorems (including an appearance of Lemma 4.1) and taking some liberties with the language (for example, (p) means three different things, hopefully self-clarifying through the context). Therefore,

$$\begin{aligned} p \text{ splits in } \mathbb{Z}[i] &\iff \mathbb{Z}[i]/(p) \text{ is } \textit{not} \text{ an integral domain} \\ &\iff \mathbb{Z}/p\mathbb{Z}[x]/(x^2+1) \text{ is } \textit{not} \text{ an integral domain} \\ &\iff x^2+1 \text{ is not irreducible in } \mathbb{Z}/p\mathbb{Z}[x] \\ &\iff x^2+1 \text{ has a root in } \mathbb{Z}/p\mathbb{Z} \\ &\iff \text{there is an integer } n \text{ such that } n^2 \equiv -1 \bmod p, \end{aligned}$$

and we are reduced to verifying that this last condition is equivalent to $p \equiv 1 \bmod 4$, provided that p is an odd prime.

For this purpose, recall (Theorem IV.6.10) that the multiplicative group G of $\mathbb{Z}/p\mathbb{Z}$ is *cyclic;* let $g \in G$ be a generator of G. Since p is assumed to be odd, $p-1$ is an even number, say 2ℓ: thus g has order $|G| = 2\ell$. Also, denote the class of an integer n mod (p) by $\underline{n}$. Since g is a generator of G, for every integer n there is an integer m such that $\underline{n} = g^m$.

The class $\underline{-1}$ generates the unique subgroup of order 2 of G (unique by the classification of subgroups of a cyclic group, Proposition II.6.11); since g^ℓ does the same, we have

$$g^\ell = \underline{-1}.$$

Therefore, with $\underline{n} = g^m$, we see that $n^2 \equiv -1 \bmod p$ if and only if $g^{2m} = g^\ell$, that is, if and only if

$$2m \equiv \ell \bmod 2\ell.$$

Summarizing, $\exists n \in \mathbb{Z}$ such that $n^2 \equiv -1 \bmod p$ if and only if $\exists m \in \mathbb{Z}$ such that $2m \equiv \ell \bmod 2\ell$. Now this is clearly the case if and only if ℓ is even, that is, if and only if $(p-1) = 2\ell$ is a multiple of 4, that is, if and only if $p \equiv 1 \bmod 4$; and we are done. □

Putting together Lemma 6.9 and Lemma 6.10 gives the following beautiful number-theoretic statement:

Theorem 6.11 (Fermat). *A positive odd prime $p \in \mathbb{Z}$ is a sum of two squares if and only if $p \equiv 1 \bmod 4$.*

Remark 6.12. Lagrange proved (in 1770) that every positive integer is the sum of *four* squares. One way to prove this is not unlike the proof given for Theorem 6.11: it boils down to analyzing the splitting of (integer) primes in a ring; the role of $\mathbb{Z}[i]$ is taken here by a ring of 'integral quaternions'. ⌟

The reader should pause and note that there is no mention of UFDs, Euclidean domains, complex numbers, cyclic groups, etc., in the *statement* of Theorem 6.11, although a large selection of these tools was used in its *proof*. This is of course the dream of the generalizer: to set up an abstract machinery making interesting facts (nearly) evident—facts that would be extremely mysterious or that would require Fermat-grade cleverness to be understood without that machinery.

Exercises

6.1. Generalize the CRT for two ideals, as follows. Let I, J be ideals in a commutative ring R; prove that there is an exact sequence of R-modules

$$0 \longrightarrow I \cap J \longrightarrow R \xrightarrow{\varphi} \frac{R}{I} \times \frac{R}{J} \longrightarrow \frac{R}{I+J} \longrightarrow 0$$

where φ is the natural map. (Also, explain why this implies the first part of Theorem 6.1, for $k = 2$.)

6.2. Let R be a commutative ring, and let $a \in R$ be an element such that $a^2 = a$. Prove that $R \cong R/(a) \times R/(1-a)$.

Show that the multiplication in R endows the ideal (a) with a *ring* structure, with a as the identity[26]. Prove that $(a) \cong R/(1-a)$ as rings. Prove that $R \cong (a) \times (1-a)$ as rings.

6.3. Recall (Exercise 3.15) that a ring R is called *Boolean* if $a^2 = a$ for all $a \in R$. Let R be a finite Boolean ring; prove that $R \cong \mathbb{Z}/2\mathbb{Z} \times \cdots \times \mathbb{Z}/2\mathbb{Z}$.

6.4. Let R be a finite commutative ring, and let p be the smallest prime dividing $|R|$. Let $I_1, \ldots, I_k$ be proper ideals such that $I_i + I_j = (1)$ for $i \neq j$. Prove that $k \leq \log_p |R|$. (Hint: Prove $|R|^{k-1} \leq |I_1| \cdots |I_k| \leq (|R|/p)^k$.)

6.5. Show that the map $\mathbb{Z}[x] \to \mathbb{Z}[x]/(2) \times \mathbb{Z}[x]/(x)$ is not surjective.

6.6. ▷ Let R be a UFD.

- Let $a, b \in R$ such that $\gcd(a,b) = 1$. Prove that $(a) \cap (b) = (ab)$.
- Under the hypotheses of Corollary 6.4 (but only assuming that R is a UFD) prove that the function φ is injective.

[§6.1]

6.7. ▷ Find a polynomial $f \in \mathbb{Q}[x]$ such that $f \equiv 1 \bmod (x^2+1)$ and $f \equiv x \bmod x^{100}$. [§6.1]

6.8. ¬ Let $n \in \mathbb{Z}$ be a positive integer and $n = p_1^{a_1} \cdots p_r^{a_r}$ its prime factorization. By the classification theorem for finite abelian groups (or, in fact, simpler

[26] This is an extremely unusual situation. Note that this ring (a) is *not* a subring of R if $a \neq 1$ according to Definition III.2.5, since the identities in (a) and R differ.

considerations; cf. Exercise II.4.9)

$$\frac{\mathbb{Z}}{(n)} \cong \frac{\mathbb{Z}}{(p_1^{a_1})} \times \cdots \times \frac{\mathbb{Z}}{(p_r^{a_r})}$$

as abelian groups.

- Use the CRT to prove that this is in fact a *ring* isomorphism.
- Prove that
$$\left(\frac{\mathbb{Z}}{(n)}\right)^* \cong \left(\frac{\mathbb{Z}}{(p_1^{a_1})}\right)^* \times \cdots \times \left(\frac{\mathbb{Z}}{(p_r^{a_r})}\right)^*$$
(recall that $(\mathbb{Z}/n\mathbb{Z})^*$ denotes the group of units of $\mathbb{Z}/n\mathbb{Z}$).
- Recall (Exercise II.6.14) that *Euler's ϕ-function* $\phi(n)$ denotes the number of positive integers $< n$ that are relatively prime to n. Prove that
$$\phi(n) = p_1^{a_1-1}(p_1-1)\cdots p_r^{a_r-1}(p_r-1).$$

[II.2.15, VII.5.19]

6.9. Let I be an ideal of $\mathbb{Z}[i]$. Prove that $\mathbb{Z}[i]/I$ is finite.

6.10. ▷ Let $z, w \in \mathbb{Z}[i]$. Show that if z and w are associates, then $N(z) = N(w)$. Show that if $w \in (z)$ and $N(z) = N(w)$, then z and w are associates. [§6.2]

6.11. Prove that the irreducible elements in $\mathbb{Z}[i]$ are, up to associates: $1+i$; the integer primes congruent to 3 mod 4; and the elements $a \pm bi$ with a^2+b^2 an integer prime congruent to 1 mod 4.

6.12. ¬ Prove Lemma 6.5 without any 'visual' aid. (Hint: Let $z = a+bi$, $w = c+di$ be Gaussian integers, with $w \neq 0$. Then $z/w = \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}$. Find integers e, f such that $|e - \frac{ac+bd}{c^2+d^2}| \leq \frac{1}{2}$ and $|f - \frac{bc-ad}{c^2+d^2}| \leq \frac{1}{2}$, and set $q = e+if$. Prove that $|\frac{z}{w} - q| \leq \frac{1}{2}$. Why does this do the job?) [6.13]

6.13. ¬ Consider the set $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Z}\} \subseteq \mathbb{C}$.

- Prove that $\mathbb{Z}[\sqrt{2}]$ is a ring, isomorphic to $\mathbb{Z}[t]/(t^2-2)$.
- Prove that the function $N : \mathbb{Z}[\sqrt{2}] \to \mathbb{Z}$ defined by $N(a+b\sqrt{2}) = a^2 - 2b^2$ is multiplicative: $N(zw) = N(z)N(w)$. (Cf. Exercise III.4.10.)
- Prove that $\mathbb{Z}[\sqrt{2}]$ has infinitely many units.
- Prove that $\mathbb{Z}[\sqrt{2}]$ is a Euclidean domain, by using the absolute value of N as valuation. (Hint: Follow the same steps as in Exercise 6.12.)

[6.14]

6.14. Working as in Exercise 6.13, prove that $\mathbb{Z}[\sqrt{-2}]$ is a Euclidean domain. (Use the norm $N(a+b\sqrt{-2}) = a^2+2b^2$.)

If you are particularly adventurous, prove that $\mathbb{Z}[(1+\sqrt{d})/2]$ is also a Euclidean domain[27] for $d = -3, -7, -11$. (You can still use the norm defined by $N(a+b\sqrt{d}) = a^2 - db^2$; note that this is still an integer on $\mathbb{Z}[(1+\sqrt{d})/2]$, if $d \equiv 1 \bmod 4$.)

[27] You are probably wondering why we switched from $\mathbb{Z}[\sqrt{d}]$ to $\mathbb{Z}[(1+\sqrt{d})/2]$. These rings are the 'ring of integers' in $\mathbb{Q}(\sqrt{d})$; the form they take depends on the class of d modulo 4. Their study is a cornerstone of algebraic number theory.

The five values $d = -1, -2$, resp., $-3, -7, -11$, are the only ones for which $\mathbb{Z}[\sqrt{d}]$, resp., $\mathbb{Z}[(1+\sqrt{d})/2]$, is Euclidean. For the values $d = -19, -43, -67, -163$, the ring $\mathbb{Z}[(1+\sqrt{d})/2]$ is still a PID (cf. §2.4 and Exercise 2.18 for $d = -19$); the fact that there are no other negative values for which the ring of integers in $\mathbb{Q}(\sqrt{d})$ is a PID was conjectured by Gauss and only proven by Alan Baker and Harold Stark around 1966. Also, keep in mind that $\mathbb{Z}[\sqrt{-5}]$ is not even a UFD, as you have proved all by yourself in Exercise 1.17.

6.15. Give an elementary proof (using modular arithmetic) of the fact that if an integer n is congruent to 3 modulo 4, then it is not the sum of two squares.

6.16. Prove that if m and n are two integers both of which can be written as sums of two squares, then mn can also be written as the sum of two squares.

6.17. Let n be a positive integer.

- Prove that n is a sum of two squares if and only if it is the norm of a Gaussian integer $a + bi$.
- By factoring $a^2 + b^2$ in $\mathbb{Z}$ and $a + bi$ in $\mathbb{Z}[i]$, prove that n is a sum of two squares if and only if each integer prime factor p of n such that $p \equiv 3 \bmod 4$ appears with an even power in n.

6.18. ¬ One ingredient in the proof of Lagrange's theorem on four squares is the following result, which can be proven by completely elementary means. Let $p > 0$ be an odd prime integer. Then there exists an integer n, $0 < n < p$, such that np may be written as $1 + a^2 + b^2$ for two integers a, b. Prove this result, as follows:

- Prove that the numbers a^2, $0 \le a \le (p-1)/2$, represent $(p+1)/2$ distinct congruence classes mod p.
- Prove the same for numbers of the form $-1 - b^2$, $0 \le b \le (p-1)/2$.
- Now conclude, using the pigeon-hole principle.

[6.21]

6.19. ¬ Let $\mathbb{I} \subseteq \mathbb{H}$ be the set of quaternions (cf. Exercise III.1.12) of the form $\frac{a}{2}(1+i+j+k) + bi + cj + dk$ with $a, b, c, d \in \mathbb{Z}$.

- Prove that $\mathbb{I}$ is a (noncommutative) subring of the ring of quaternions.
- Prove that the norm $N(w)$ (Exercise III.2.5) of an integral quaternion $w \in \mathbb{I}$ is an integer and $N(w_1 w_2) = N(w_1)N(w_2)$.
- Prove $\mathbb{I}$ has exactly 24 units in $\mathbb{I}$: ± 1, $\pm i$, $\pm j$, $\pm k$, and $\frac{1}{2}(\pm 1 \pm i \pm j \pm k)$.
- Prove that every $w \in \mathbb{I}$ is an associate of an element $a + bi + cj + dk \in \mathbb{I}$ with $a, b, c, d \in \mathbb{Z}$.

The ring $\mathbb{I}$ is called the ring of *integral quaternions*. [6.20, 6.21]

6.20. ¬ Let $\mathbb{I}$ be as in Exercise 6.19. Prove that $\mathbb{I}$ shares most good properties of a Euclidean domain, notwithstanding the fact that it is noncommutative.

- Let $z, w \in \mathbb{I}$, with $w \neq 0$. Prove that $\exists q, r \in \mathbb{I}$ such that $z = qw + r$, with $N(r) < N(w)$. (This is a little tricky; don't feel too bad if you have to cheat and look it up somewhere.)

- Prove that every left-ideal in $\mathbb{I}$ is of the form $\mathbb{I}w$ for some $w \in \mathbb{I}$.
- Prove that every $z, w \in \mathbb{I}$, not both zero, have a 'greatest common right-divisor' d in $\mathbb{I}$, of the form $\alpha z + \beta w$ for $\alpha, \beta \in \mathbb{I}$.

[6.21]

6.21. Prove Lagrange's theorem on four squares. Use notation as in Exercises 6.19 and 6.20.

- Let $z \in \mathbb{I}$ and $n \in \mathbb{Z}$. Prove that the greatest common right-divisor of z and n in $\mathbb{I}$ is 1 if and only if $(N(z), m) = 1$ in $\mathbb{Z}$. (If $\alpha z + \beta n = 1$, then $N(\alpha)N(z) = N(1 - \beta n) = (1 - \beta n)(1 - \overline{\beta} n)$, where $\overline{\beta}$ is obtained by changing the signs of the coefficients of i, j, k. Expand, and deduce that $(N(z), n) \mid 1$.)
- For an odd prime integer p, use Exercise 6.18 to obtain an integral quaternion $z = 1 + ai + bj$ such that $p \mid N(z)$. Prove that z and p have a common right-divisor that is not a unit and not an associate of p.
- Say that $w \in \mathbb{I}$ is *irreducible* if $w = \alpha\beta$ implies that either α or β is a unit. Prove that integer primes are *not* irreducible in $\mathbb{I}$. Deduce that every positive prime integer is the norm of some integral quaternion.
- Prove that every positive integer is the norm of some integral quaternion.
- Finally, use the last point of Exercise 6.19 to deduce that every positive integer may be written as the sum of four perfect squares.

Linear algebra

In several branches of science, 'algebra' means 'linear algebra': the study of vector spaces and linear maps, that is, of the category of modules over a ring R in the very special case in which R is a *field* (and often one restricts attention to very special fields, such as $\mathbb{R}$ or $\mathbb{C}$).

This will be one of the main themes in this chapter. However, we will stress that much of what can be done over a field can in fact be done over less special rings. In fact, we will argue that working out the general theory over *integral domains* produces invaluable tools when we are working over *fields:* the paramount example being canonical forms for matrices with entries in a field, which will be obtained as a corollary of the classification of finitely generated modules over a PID.

Throughout the main body of this chapter (but not necessarily in the exercises) R will denote an integral domain; most of the theory can be extended to arbitrary commutative[1] rings without major difficulty.

1. Free modules revisited

1.1. R-Mod. For generalities on modules, see §III.5. A *module* over R is an abelian group M, endowed with an action of R. The action of $r \in R$ on $m \in M$ is denoted rm: there is a notational bias towards *left*-modules (but the distinction between left- and right-modules will be immaterial here as R is commutative). The defining axioms of a module tell us that for all $r_1, r_2, r \in R$ and $m, m_1, m_2 \in M$,

- $(r_1 + r_2)m = r_1 m + r_2 m$,
- $1m = m$ and $(r_1 r_2)m = r_1(r_2 m)$,
- $r(m_1 + m_2) = rm_1 + rm_2$.

[1]The hypothesis of commutativity is very convenient, as it allows us to identify the notion of *left-* and *right-*modules, and the fact that integral domains have fields of fractions simplifies many arguments. The reader should keep in mind that many results in this chapter can be extended to more general rings.

Modules over a ring R form the category R-Mod, which we encountered in Chapter III. This category reflects subtle and important properties of R, and we are going to attempt to uncover some of these features. As a first approximation, we look at the 'full subcategory' (cf. Exercise I.3.8) of R-Mod whose objects are *free* modules and in which morphisms are ordinary morphisms in R-Mod, that is, R-linear group homomorphisms.

The goal of the first few sections of this chapter is to give a very explicit description of this subcategory: in the case of *finitely generated* free modules, this can be done by means of *matrices* with entries in R. Later in the chapter, we will see that matrices may also be used to describe important classes of nonfree modules.

1.2. Linear independence and bases. The reader is invited to review the definition of *free* R-modules given in §III.6.3: $F^R(S)$ denotes an R-module containing a given set S and universal with respect to the existence of a set-map from S. We proved (Claim III.6.3) that the module $R^{\oplus S}$ with 'one component for each element of S' gives an explicit realization of $F^R(S)$.

The main point of this subsection will be that, for reasonable rings R, the set S can be recovered[2] 'abstractly' from the free module $F^R(S)$. For example, it will follow easily that $R^m \cong R^n$ if and only if $m = n$; this is the first indication that the category of (finitely generated) free modules does indeed admit a simple description.

Our main tool will be the famous concepts of *linearly independent subsets* and *bases*. It is easy to be imprecise in defining these notions. In order to avoid obvious traps, we will give the definitions for *indexed sets* (cf. §I.2.2), that is, for functions $i : I \to M$ from a (nonempty) indexing set I to a given module M. The reader should think of i as a selection of elements of M, allowing for the possibility that the elements $m_\alpha \in M$ corresponding to $\alpha \in I$ may not all be distinct.

Recall that for all sets I there is a canonical injection $j : I \to F^R(I)$ and any function $i : I \to M$ determines a unique R-module homomorphism $\varphi : F^R(I) \to M$ making the diagram

$$\begin{array}{ccc} F^R(I) & \xrightarrow{\varphi} & M \\ {\scriptstyle j}\uparrow & \nearrow_{i} & \\ I & & \end{array}$$

commute: this is precisely the universal property satisfied by $F^R(I)$.

Definition 1.1. We say that the indexed set $i : I \to M$ is *linearly independent* if φ is injective; i is *linearly dependent* otherwise. We say that i *generates* M if φ is surjective. ⌟

Put in a slightly messier, but perhaps more common, way, an indexed set $S = \{m_\alpha\}_{\alpha \in I}$ of elements of M is linearly independent if the only vanishing linear

[2]This is not necessarily the case if R is not commutative.

combination

$$\sum_{\alpha \in I} r_\alpha m_\alpha = 0$$

is obtained by choosing $r_\alpha = 0\ \forall \alpha \in I$; S is linearly dependent otherwise. The indexed set generates M if every element of M may be written as $\sum_{\alpha \in I} r_\alpha m_\alpha$ for some choice of r_α. (As a notational warning/reminder, keep in mind that only *finite* sums are defined in a module; therefore, in this context the notation $\sum$ stands for a finite sum. When writing $\sum_{\alpha \in I} m_\alpha$ in an ordinary module, one is implicitly assuming that $m_\alpha = 0$ for all but finitely many $\alpha \in I$.)

Using indexed sets in Definition 1.1 takes care of obvious particular cases such as m and m being linearly *de*pendent since $1 \cdot m + (-1) \cdot m = 0$: if $i : I \to M$ is not itself injective, then φ is surely not injective (by the commutativity of the diagram and the injectivity of j), so that i is linearly dependent in this case. Because of this fact, the datum of a linearly independent $i : I \to M$ amounts to a (special) choice of *distinct* elements of M; the temptation to identify the elements of I with the corresponding elements of M is historically irresistible and essentially harmless. Thus, it is common to speak about linearly dependent/independent *subsets* of M. We will conform to this common practice; the reader should parse the statements carefully and correct obvious imprecisions that may arise.

A simple application of Zorn's lemma shows that every module has *maximal* linearly independent subsets. In fact, it gives the following conveniently stronger statement:

Lemma 1.2. *Let M be an R-module, and let $S \subseteq M$ be a linearly independent subset. Then there exists a maximal linearly independent subset of M containing S.*

Proof. Consider the family $\mathscr{S}$ of linearly independent subsets of M containing S, ordered by inclusion. Since S is linearly independent, $\mathscr{S} \neq \emptyset$. By Zorn's lemma, it suffices to verify that every chain in $\mathscr{S}$ has an upper bound. Indeed, the union of a chain of linearly independent subsets containing S is also linearly independent: because any relation of linear dependence only involves finitely many elements and these elements would all belong to one subset in the chain. □

Remark 1.3. This statement is in fact known to be equivalent to the axiom of choice; therefore, the use of Zorn's lemma in one form or another cannot be bypassed. ⌟

Note that the singleton $\{2\} \subseteq \mathbb{Z}$ is a 'maximal linearly independent subset' of $\mathbb{Z}$, but it does *not* generate $\mathbb{Z}$. In general this is an additional requirement, leading to the definition of a *basis*.

Definition 1.4. An indexed set $B \to M$ is a *basis* if it generates M and is linearly independent. ⌟

Again, one often talks of bases as 'subsets' of the module M; since the images of all $b \in B$ are necessarily *distinct* elements, this is rather harmless. When B is finite (or at any rate countable), the extra information carried by the indexed set

can be encoded by *ordering* the elements of B; to emphasize this, one talks about *ordered bases.*

Bases are necessarily maximal linear independent subsets and minimal generating subsets; this holds over every ring. What will make modules over a *field,* i.e., vector spaces, so special is that the *converse* will also hold.

In any case, only very special modules admit bases:

Lemma 1.5. *An R-module M is free if and only if it admits a basis. In fact, $B \subseteq M$ is a basis if and only if the natural homomorphism $R^{\oplus B} \to M$ is an isomorphism.*

Proof. This is immediate from Definition 1.1: if $B \subseteq M$ is linearly independent and generates M, then the corresponding homomorphism $R^{\oplus B} \to M$ is injective and surjective. Conversely, if $\varphi : R^{\oplus B} \to M$ is an isomorphism, then B is identified with a subset of M which generates it (because φ is surjective) and is linearly independent (because φ is injective). □

By Lemma 1.5, the choice of a basis B of a free module M amounts to the choice of an isomorphism $R^{\oplus B} \cong M$; this will be an important observation in due time (e.g., in §2.2).

Once a basis B has been chosen for a free module M, then every element $m \in M$ can be written uniquely as a linear combination

$$m = \sum_{b \in B} r_b b$$

with $r_b \in R$. As always, remember that all but finitely many of the coefficients r_b are 0 in any such expression.

1.3. Vector spaces. Lemma 1.5 is all that is needed to prove the fundamental observation that *modules over a field are necessarily free.* Recall that modules over a field k are called *k-vector spaces* (Example III.5.5). Elements of a vector space are called (surprise, surprise) *vectors,* while elements of the field are called[3] *scalars.*

By Lemma 1.5, proving that vector spaces are free modules amounts to proving that they admit bases; Lemma 1.2 reduces the matter to the following:

Lemma 1.6. *Let $R = k$ be a field, and let V be a k-vector space. Let B be a* maximal *linearly independent subset of V; then B is a basis of V.*

Again, this should be contrasted with the situation over rings: $\{2\}$ is a maximal linearly independent subset of $\mathbb{Z}$, but it is not a basis.

Proof. Let $v \in V$, $v \notin B$. Then $B \cup \{v\}$ is not linearly independent, by the maximality of B; therefore, there exist $c_0, \dots, c_t \in k$ and (distinct) $b_1, \dots, b_t \in B$ such that

$$c_0 v + c_1 b_1 + \cdots + c_t b_t = 0,$$

[3]This terminology is also often used for free modules over any ring.

with not all $c_0, \ldots, c_t$ equal to 0. Now, $c_0 \neq 0$: otherwise we would get a linear dependence relation among elements of B. *Since k is a field,* c_0 is a unit; but then

$$v = (-c_0^{-1}c_1)b_1 + \cdots + (-c_0^{-1}c_t)b_t,$$

proving that v is in the span of B. It follows that B generates V, as needed. □

Summarizing,

Proposition 1.7. *Let $R = k$ be a field, and let V be a k-vector space. Let S be a linearly independent set of vectors of V. Then there exists a basis B of V containing S.*

In particular, V is free as a k-module.

Proof. Put Lemma 1.2, Lemma 1.5, and Lemma 1.6 together. □

We could also contemplate this situation from the 'mirror' point of view of generating sets:

Lemma 1.8. *Let $R = k$ be a field, and let V be a k-vector space. Let B be a* minimal *generating set for V; then B is a basis of V.*

Every set generating V contains a basis of V.

Proof. Exercise 1.6. □

Lemma 1.8 also fails on more general rings (Exercise 1.5). To reiterate, over fields (but not over general rings) a subset B of a vector space is a basis $\iff$ it is a maximal linearly independent subset $\iff$ it is a minimal generating set.

1.4. Recovering B from $F^R(B)$. We are ready for the 'reconstruction' of a set B (up to a bijection!) from the corresponding free module $F^R(B)$. This is the result justifying the notion of *dimension* of a vector space, or, more generally, the *rank* of a free module. Again, we prove a somewhat stronger statement.

Proposition 1.9. *Let R be an integral domain, and let M be a free R-module. Let B be a maximal linearly independent subset of M, and let S be a linearly independent subset. Then*[4] $|S| \leq |B|$.

In particular, any two maximal linearly independent subsets of a free module over an integral domain have the same cardinality.

Proof. By taking fields of fractions, the general case over an integral domain is easily reduced to the case of vector spaces over a field; see Exercise 1.7. We may then assume that $R = k$ is a field and $M = V$ is a k-vector space.

We have to prove that there is an injective map $j : S \hookrightarrow B$, and this can be done by an inductive process, replacing elements of B by elements of S 'one-by-one'. For this, let $\leq$ be a well-ordering on S, let $v \in S$, and assume we have defined j for all $w \in S$ with $w < v$. Let B' be the set obtained from B by replacing all

[4]Here, $|A|$ denotes the *cardinality* of the set A, a notion with which the reader is hopefully familiar. The reader will not lose much by only considering the case in which B, S are finite sets; but the fact is true for 'infinite-dimensional spaces' as well, as the argument shows.

$j(w)$ by w, for $w < v$, and assume (inductively) that B' is still a maximal linearly independent subset of V. Then we claim that $j(v) \in B$ may be defined so that

- $j(v) \neq j(w)$ for all $w < v$;
- the set B'' obtained from B' by replacing $j(v)$ by v is still a maximal linearly independent subset.

(Transfinite) induction (Claim V.3.2) then shows that j is defined and injective on S, as needed.

To verify our claim, since B' is a maximal linearly independent set, $B' \cup \{v\}$ is linearly dependent (as an indexed set[5]), so that there exists a linear dependence relation

$$(*) \qquad c_0 v + c_1 b_1 + \cdots + c_t b_t = 0$$

with not all c_i equal to zero and the b_i distinct in B'. Necessarily $c_0 \neq 0$ (because B' is linearly independent); also, necessarily not all the b_i with $c_i \neq 0$ are elements of S (because S is linearly independent). Without loss of generality we may then assume that $c_1 \neq 0$ and $b_1 \in B' \smallsetminus S$. This guarantees that $b_1 \neq j(w)$ for all $w < v$; we set $j(v) = b_1$.

All that is left now is the verification that the set B'' obtained by replacing b_1 by v in B' is a maximal linearly independent subset. But by using (*) to write

$$v = -c_0^{-1} c_1 b_1 - \cdots - c_0^{-1} c_t b_t,$$

this is an easy consequence of the fact that B' is a maximal linearly independent subset. Further details are left to the reader. □

Example 1.10. An *uncountable* subset of $\mathbb{C}[x]$ is necessarily linearly *dependent.* Indeed, $\mathbb{C}[x]$ has a countable basis over $\mathbb{C}$: for example, $\{1, x, x^2, x^3, \dots\}$. ⌟

Corollary 1.11. *Let R be an integral domain, and let A, B be sets. Then*

$$F^R(A) \cong F^R(B) \iff \textit{there is a bijection } A \cong B.$$

Proof. Exercise 1.8. □

Remark 1.12. We have learned in Lemma 1.2 that we can 'complete' every linearly independent subset S to a maximal one. The argument used in the proof of Proposition 1.9 shows that we can in fact do this by borrowing elements of a *given* maximal linearly independent subset. ⌟

Remark 1.13. As a particular case of Corollary 1.11, we see that if R is an integral domain, then $R^m \cong R^n$ if and only if $m = n$. This says that integral domains satisfy the 'IBN (Invariant Basis Number) property'.

Strange as it may seem, this obvious-looking fact does *not* hold over arbitrary rings: for example, the ring of endomorphisms of an infinite-dimensional vector space does not satisfy the IBN property. On the other hand, integral domains are a bit of an overshoot: all commutative rings satisfy the IBN property (Exercise 1.11).

[5]This allows for the possibility that $v \in B'$ 'already'. In this case, the reader can check that the process we are about to describe gives $j(v) = v$.

One way to think about this is that the category of finitely generated free modules over (say) an integral domain is 'classified' by $\mathbb{Z}^{\geq 0}$: up to isomorphisms, there is exactly one finitely generated free module for any nonnegative integer. The task of describing this category then amounts to describing the homomorphisms between objects corresponding to two given nonnegative integers; this will be done in §2.1. ⌟

As a byproduct of the result of Proposition 1.9, we can now give the following important definition.

Definition 1.14. Let R be an integral domain. The *rank* of a free R-module M, denoted $\mathrm{rk}_R M$, is the cardinality of a maximal linearly independent subset of M. The rank of a vector space is called the *dimension,* denoted $\dim_k V$. ⌟

This definition will in fact be adopted for more general finitely generated modules when the time comes, in §5.3.

Finite-dimensional vector spaces over a fixed field form a category. Since vector spaces are free modules (Proposition 1.7), Corollary 1.11 implies that two finite-dimensional vector spaces are isomorphic if and only if they have the same dimension.

The subscripts R, k are often omitted, if the context permits. But note that, for example, viewing the complex numbers as a real vector space, we have $\dim_{\mathbb{R}} \mathbb{C} = 2$, while $\dim_{\mathbb{C}} \mathbb{C} = 1$. So some care is warranted.

Proposition 1.9 tells us that every linearly independent subset S of a free R-module M must have cardinality lower than $\mathrm{rk}_R M$. Similarly, every *generating set* must have cardinality *higher* than the rank. Indeed,

Proposition 1.15. *Let R be an integral domain, and let M be a free R-module; assume that M is generated by S: $M = \langle S \rangle$. Then S contains a maximal linearly independent subset of M.*

Proof. By Exercise 1.7 we may assume that R is a field and $M = V$ is a vector space. Use Zorn's lemma to obtain a linearly independent subset $B \subseteq S$ which is maximal among subsets of S. Arguing as in the proof of Lemma 1.6 shows that S is in the span of B, and it follows that B generates V. Thus B is a basis, and hence a maximal linearly independent subset of V, as needed. □

Remark 1.16. We have used again the trick of switching from an integral domain to its field of fractions. The second part of the argument would not work over an arbitrary integral domain, since maximal linearly independent subsets over an integral domain are not generating sets in general.

Another standard method to reduce questions about modules over arbitrary commutative rings to vector spaces is to mod out by a maximal ideal; cf. Exercise 1.9 and following. ⌟

Exercises

1.1. ¬ Prove that $\mathbb{R}$ and $\mathbb{C}$ are isomorphic *as $\mathbb{Q}$-vector spaces.* (In particular, $(\mathbb{R},+)$ and $(\mathbb{C},+)$ are isomorphic as groups.) [II.4.4]

1.2. ¬ Prove that the sets listed in Exercise III.1.4 are all $\mathbb{R}$-vector spaces, and compute their dimensions. [1.3]

1.3. Prove that $\mathfrak{su}_2(\mathbb{C}) \cong \mathfrak{so}_3(\mathbb{R})$ as $\mathbb{R}$-vector spaces. (This is immediate, and not particularly interesting, from the dimension computation of Exercise 1.2. However, these two spaces may be viewed as the tangent spaces to $\mathrm{SU}_2(\mathbb{C})$, resp., $\mathrm{SO}_3(\mathbb{R})$, at I; the surjective homomorphism $\mathrm{SU}_2(\mathbb{C}) \to \mathrm{SO}_3(\mathbb{R})$ you constructed in Exercise II.8.9 induces a more 'meaningful' isomorphism $\mathfrak{su}_2(\mathbb{C}) \to \mathfrak{so}_3(\mathbb{R})$. Can you find this isomorphism?)

1.4. Let V be a vector space over a field k. A *Lie bracket* on V is an operation $[\cdot,\cdot] : V \times V \to V$ such that

- $(\forall u,v,w \in V), (\forall a,b \in k)$,
$$[au+bv,w] = a[u,w] + b[v,w], \quad [w,au+bv] = a[w,u] + b[w,v],$$
- $(\forall v \in V)$, $[v,v] = 0$,
- and $(\forall u,v,w \in V)$, $[[u,v],w] + [[v,w],u] + [[w,u],v] = 0$.

(This axiom is called the *Jacobi identity.*) A vector space endowed with a Lie bracket is called a *Lie algebra.* Define a category of Lie algebras over a given field. Prove the following:

- In a Lie algebra V, $[u,v] = -[v,u]$ for all $u,v \in V$.
- If V is a k-algebra (Definition III.5.7), then $[v,w] := vw - wv$ defines a Lie bracket on V, so that V is a Lie algebra in a natural way.
- This makes $\mathfrak{gl}_n(\mathbb{R})$, $\mathfrak{gl}_n(\mathbb{C})$ into Lie algebras. The sets listed in Exercise III.1.4 are all Lie algebras, with respect to a Lie bracket induced from $\mathfrak{gl}$.
- $\mathfrak{su}_2(\mathbb{C})$ and $\mathfrak{so}_3(\mathbb{R})$ are isomorphic as Lie algebras over $\mathbb{R}$.

1.5. ▷ Let R be an integral domain. Prove or disprove the following:

- Every linearly independent subset of a free R-module may be completed to a basis.
- Every generating subset of a free R-module contains a basis.

[§1.3]

1.6. ▷ Prove Lemma 1.8. [§1.3]

1.7. ▷ Let R be an integral domain, and let $M = R^{\oplus A}$ be a free R-module. Let K be the field of fractions of R, and view M as a subset of $V = K^{\oplus A}$ in the evident way. Prove that a subset $S \subseteq M$ is linearly independent in M (over R) if and only if it is linearly independent in V (over K). Conclude that the rank of M (as

an R-module) equals the dimension of V (as a K-vector space). Prove that if S generates M over R, then it generates V over K. Is the converse true? [§1.4]

1.8. ▷ Deduce Corollary 1.11 from Proposition 1.9. [§1.4]

1.9. ▷ Let R be a commutative ring, and let M be an R-module. Let $\mathfrak{m}$ be a maximal ideal in R, such that $\mathfrak{m}M = 0$ (that is, $rm = 0$ for all $r \in \mathfrak{m}$, $m \in M$). Define in a natural way a vector space structure over $R/\mathfrak{m}$ on M. [§1.4]

1.10. ¬ Let R be a commutative ring, and let $F = R^{\oplus B}$ be a free module over R. Let $\mathfrak{m}$ be a maximal ideal of R, and let $k = R/\mathfrak{m}$ be the quotient field. Prove that $F/\mathfrak{m}F \cong k^{\oplus B}$ as k-vector spaces. [1.11]

1.11. ▷ Prove that commutative rings satisfy the IBN property. (Use Proposition V.3.5 and Exercise 1.10.) [§1.4]

1.12. Let V be a vector space over a field k, and let $R = \mathrm{End}_{k\text{-}\mathsf{Vect}}(V)$ be its ring of endomorphisms (cf. Exercise III.5.9). (Note that R is *not* commutative in general.)

- Prove that $\mathrm{End}_{k\text{-}\mathsf{Vect}}(V \oplus V) \cong R^4$ as an R-module.
- Prove that R does not satisfy the IBN property if $V = k^{\oplus \mathbb{N}}$.

(Note that $V \cong V \oplus V$ if $V = k^{\oplus \mathbb{N}}$.)

1.13. ¬ Let A be an abelian group such that $\mathrm{End}_{\mathsf{Ab}}(A)$ is a field of characteristic 0. Prove that $A \cong \mathbb{Q}$. (Hint: Prove that A carries a $\mathbb{Q}$-vector space structure; what must its dimension be?) [IX.2.13]

1.14. ¬ Let V be a finite-dimensional vector space, and let $\varphi : V \to V$ be a homomorphism of vector spaces. Prove that there is an integer n such that $\ker \varphi^{n+1} = \ker \varphi^n$ and $\operatorname{im} \varphi^{n+1} = \operatorname{im} \varphi^n$.

Show that both claims may fail if V has infinite dimension. [1.15]

1.15. Consider the question of Exercise 1.14 for free R-modules F of finite rank, where R is an integral domain that is not a field. Let $\varphi : F \to F$ be an R-module homomorphism.

What property of R immediately guarantees that $\ker \varphi^{n+1} = \ker \varphi^n$ for $n \gg 0$?

Show that there is an R-module homomorphism $\varphi : F \to F$ such that $\operatorname{im} \varphi^{n+1} \subsetneq \varphi^n$ for all $n \geq 0$.

1.16. ¬ Let M be a module over a ring R. A *finite composition series* for M (if it exists) is a decreasing sequence of submodules

$$M = M_0 \supsetneq M_1 \supsetneq \cdots \supsetneq M_m = \langle 0 \rangle$$

in which all quotients M_i/M_{i+1} are *simple* R-modules (cf. Exercise III.5.4). The *length* of a series is the number of strict inclusions. The *composition factors* are the quotients M_i/M_{i+1}.

Prove a Jordan-Hölder theorem for modules: any two finite composition series of a module have the same length and the same (multiset of) composition factors. (Adapt the proof of Theorem IV.3.2.)

We say that M has *length* m if M admits a finite composition series of length m. This notion is well-defined as a consequence of the result you just proved. [1.17, 1.18, 3.20, 7.15]

1.17. Prove that a k-vector space V has finite length as a module over k (cf. Exercise 1.16) if and only if it is finite-dimensional and that in this case its length equals its dimension.

1.18. Let M be an R-module of finite length m (cf. Exercise 1.16).

- Prove that every submodule N of M has finite length $n \leq m$. (Adapt the proof of Proposition IV.3.4.)
- Prove that the 'descending chain condition' (d.c.c.) for submodules holds in M. (Use induction on the length.)
- Prove that if R is an integral domain that is not a field and F is a free R-module, then F has finite length if and only if it is the 0-module.

1.19. Let k be a field, and let $f(x) \in k[x]$ be any polynomial. Prove that there exists a multiple of $f(x)$ in which all exponents of nonzero monomials are *prime* integers. (Example: for $f(x) = 1 + x^5 + x^6$,

$$(1 + x^5 + x^6)(2x^2 - x^3 + x^5 - x^8 + x^9 - x^{10} + x^{11}) \\ = 2x^2 - x^3 + x^5 + 2x^7 + 2x^{11} - x^{13} + x^{17}.)$$

(Hint: $k[x]/(f(x))$ is a finite-dimensional k-vector space.)

1.20. ¬ Let A, B be sets. Prove that the free groups $F(A)$, $F(B)$ (§II.5) are isomorphic if and only if there is a bijection $A \cong B$. (For the interesting direction: remember that $F(A) \cong F(B) \implies F^{ab}(A) \cong F^{ab}(B)$, by Exercise II.7.12). This extends the result of Exercise II.7.13 to possibly infinite sets A, B. [II.5.10]

---

2. Homomorphisms of free modules, I

2.1. Matrices. As pointed out in §1, Corollary 1.11 amounts to a classification of free modules over an integral domain: if F is a free module, then there is a set A (determined up to a bijection) such that $F \cong R^{\oplus A}$. The choice of such an isomorphism is precisely the same thing as the choice of a basis of F (Lemma 1.5).

This is simultaneously good and bad news. The good news is that if F_1, F_2 are free, then it must be possible to 'understand'[6]

$$\mathrm{Hom}_R(F_1, F_2)$$

entirely in terms of the corresponding sets A_1, A_2 such that $F_1 \cong R^{\oplus A_1}$, $F_2 \cong R^{\oplus A_2}$. That is, this set of morphisms in R-Mod may be identified with

$$\mathrm{Hom}_R(R^{\oplus A_1}, R^{\oplus A_2}).$$

The bad news is that this identification $\mathrm{Hom}_R(F_1, F_2) \cong \mathrm{Hom}_R(R^{\oplus A_1}, R^{\oplus A_2})$ is *not* 'canonical', because it depends on the chosen isomorphisms $F_1 \cong R^{\oplus A_1}$,

[6]For notational simplicity we will denote $\mathrm{Hom}_{R\text{-Mod}}(M, N)$ by $\mathrm{Hom}_R(M, N)$; this is very common, and no confusion is likely.

$F_2 \cong R^{\oplus A_2}$, that is, on the choice of bases. We will therefore have to do some work to deal with this ambiguity (in §2.2).

The point of this subsection is to deal with the good news, that is, describe $\mathrm{Hom}_R(R^{\oplus A_1}, R^{\oplus A_2})$. This can be done in a particularly convenient way when the free modules are *finitely generated,* the case we are going to analyze more carefully. Also recall (from §III.5.2) that one of the good features of the category R-Mod is that the set of morphisms $\mathrm{Hom}_R(M,N)$ between two R-modules is *itself* an R-module, in a natural way. The task is then to describe as explicitly as possible[7]

$$\mathrm{Hom}_R(R^n, R^m)$$

as an R-module, for every choice of $m, n \in \mathbb{Z}^{\geq 0}$.

This will be done by means of *matrices* with entries in R. We trust that the reader is familiar with the general notion of an $m \times n$ matrix; we have occasionally used matrices in examples given in previous chapters, and they have showed up in several exercises. An $m \times n$ *matrix* with entries in R is simply a choice of mn elements of R. It is common to arrange these elements as an array consisting of m rows and n columns:

$$(r_{ij})_{\substack{i=1,\cdots,m \\ j=1,\cdots,n}} = \begin{pmatrix} r_{11} & r_{12} & & r_{1n} \\ r_{21} & r_{22} & & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & & r_{mn} \end{pmatrix}$$

with $r_{ij} \in R$. For any given m, n the set $\mathcal{M}_{m,n}(R)$ of $m \times n$ matrices with entries in R is an abelian group under entrywise *addition*[8]

$$(a_{ij}) + (b_{ij}) := (a_{ij} + b_{ij})$$

(cf. Example II.1.5); this is in fact an R-module, under the action

$$r(a_{ij}) := (ra_{ij})$$

for $r \in R$. From this point of view, the set of $m \times n$ matrices is simply a copy of the R-module R^{mn}.

However, there are other interesting operations on these sets. If $A = (a_{ik})$ is an $m \times p$ matrix and $B = (b_{kj})$ is a $p \times n$ matrix, then one may define the *product* of A and B as

$$A \cdot B = (a_{ik}) \cdot (b_{kj}) := (\sum_{k=1}^{p} a_{ik} b_{kj});$$

this operation is (clearly) distributive with respect to addition and compatible with the R-module structure. It is just a tiny bit messier to check that it is associative, in the sense that if A, B, C are matrices of size, respectively, $m \times p$, $p \times q$, and $q \times n$, then

$$(A \cdot B) \cdot C = A \cdot (B \cdot C).$$

[7]We are now writing R^n for what was denoted $R^{\oplus n}$ in previous chapters. This is a slight abuse of language, but it makes for easier typesetting.

[8]Note that we will be dropping the extra subscripts giving the range of the indices and will write (r_{ij}) rather than $(r_{ij})_{\substack{i=1,\cdots,m \\ j=1,\cdots,n}}$: these subscripts are an eyesore, and the context usually makes the information redundant.

The reader should have no trouble reconstructing the proof of this fact (Exercise 2.2).

In particular, we have a binary operation on the set $\mathcal{M}_n(R)$ of *square* $n \times n$-matrices, and this operation is associative, distributive w.r.t. $+$, and admits the identity element

$$\begin{pmatrix} 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

(the *identity matrix,* denoted I_n). That is, $\mathcal{M}_n(R)$ is a *ring* (Example III.1.6) and in fact an R-algebra (Exercise 2.3). Except in very special cases (such as $n = 1$) this ring is not commutative.

Matrices of type $n \times 1$ are called *column n-vectors;* matrices of type $1 \times m$ are called *row m-vectors,* for evident reasons. An element of a free R-module R^n is nothing but the choice of n elements of R, and we can arrange these elements into row or column vectors if we please; the standard choice is to represent them as *column* vectors:

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in R^n.$$

We will denote by $\mathbf{e}_i$ the elements of the 'standard basis' of R^n:

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots, \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

so that

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \sum_{j=1}^{n} v_j \mathbf{e}_j.$$

The elements $v_j \in R$ are the 'components' of $\mathbf{v}$.

Interpreting elements of R^n as column vectors, we can *act on R^n with an $m \times n$ matrix, by left-multiplication:* if $A = (a_{ij})$ is an $m \times n$ matrix and $\mathbf{v} \in R^n$ is a column vector, the product is a column vector in R^m:

$$A \cdot \mathbf{v} = \begin{pmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} a_{11}v_1 + a_{12}v_2 + \cdots + a_{1n}v_n \\ a_{21}v_1 + a_{22}v_2 + \cdots + a_{2n}v_n \\ \vdots \\ a_{m1}v_1 + a_{m2}v_2 + \cdots + a_{mn}v_n \end{pmatrix} \in R^m.$$

Lemma 2.1. *For all $m \times n$ matrices A with entries in R:*

- *The function $\varphi : R^n \to R^m$ defined by $\varphi(\mathbf{v}) = A \cdot \mathbf{v}$ is a homomorphism of R-modules.*

- *Every R-module homomorphism $R^n \to R^m$ is determined in this way by a unique $m \times n$ matrix.*

Proof. The first point follows immediately from the elementary properties of matrix multiplication recalled above: $\forall r, s \in R$, $\forall \mathbf{v}, \mathbf{w} \in R^n$

$$\varphi(r\mathbf{v} + s\mathbf{w}) = A \cdot (r\mathbf{v} + s\mathbf{w}) = rA \cdot \mathbf{v} + sA \cdot \mathbf{w} = r\varphi(\mathbf{v}) + s\varphi(\mathbf{w})$$

as needed.

For the second point, let $\varphi : R^n \to R^m$ be a homomorphism of R-modules; let a_{ij} be the i-th component of $\varphi(\mathbf{e}_j)$, so that

$$\varphi(\mathbf{e}_j) = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix}.$$

Then $A = (a_{ij})$ is an $m \times n$ matrix, and $\forall \mathbf{v} \in R^n$ with components v_j,

$$A \cdot \mathbf{v} = \begin{pmatrix} a_{11}v_1 + a_{12}v_2 + \cdots + a_{1n}v_n \\ a_{21}v_1 + a_{22}v_2 + \cdots + a_{2n}v_n \\ \vdots \\ a_{m1}v_1 + a_{m2}v_2 + \cdots + a_{mn}v_n \end{pmatrix} = \sum_{j=1}^{n} \begin{pmatrix} a_{1j}v_j \\ a_{2j}v_j \\ \vdots \\ a_{mj}v_j \end{pmatrix} = \sum_{j=1}^{n} v_j \varphi(\mathbf{e}_j)$$

$$= \varphi(\sum_{j=1}^{n} v_j \mathbf{e}_j) = \varphi(\mathbf{v}),$$

as needed. The homomorphism induced by a nonzero matrix is manifestly nontrivial; this implies that the matrix A associated to a homomorphism φ is uniquely determined by φ. □

The reader should attempt to remember the recipe associating a matrix A to a homomorphism φ as in the proof of Lemma 2.1: the j-th column of A is simply the column vector $\varphi(\mathbf{e}_j)$. The collection of these vectors determines φ—since a homomorphism is determined by its action on a set of generators.

Lemma 2.1 yields the promised explicit description of $\mathrm{Hom}_R(R^n, R^m)$:

Corollary 2.2. *The correspondence introduced in Lemma 2.1 gives an isomorphism of R-modules*

$$\mathcal{M}_{m,n}(R) \cong \mathrm{Hom}_R(R^n, R^m).$$

Proof. The reader will check that the correspondence is a bijective homomorphism of R-modules; this is enough, by Exercise III.5.12. □

This is very good news, and it gets even better. Don't forget that R-Mod is a category; that is, we can compose morphisms. Therefore, there is a function[9]

$$\mathrm{Hom}_R(R^p, R^m) \times \mathrm{Hom}_R(R^n, R^p) \to \mathrm{Hom}_R(R^n, R^m),$$

[9]Here we am reversing the order of the Hom sets on the left w.r.t. the convention used in Definition I.3.1.

mapping (φ, ψ) to $\varphi \circ \psi$; on the other hand, the matrix product gives a function

$$\mathcal{M}_{m,p}(R) \times \mathcal{M}_{p,n}(R) \to \mathcal{M}_{m,n}(R),$$

mapping (A, B) to $A \cdot B$. That is, we have a diagram

$$\begin{array}{ccc} \mathcal{M}_{m,p}(R) \times \mathcal{M}_{p,n}(R) & \longrightarrow & \mathcal{M}_{m,n}(R) \\ \sim\downarrow & & \sim\downarrow \\ \mathrm{Hom}_R(R^p, R^m) \times \mathrm{Hom}_R(R^n, R^p) & \longrightarrow & \mathrm{Hom}_R(R^n, R^m) \end{array}$$

where the vertical maps are (induced by) the isomorphisms obtained in Corollary 2.2.

Lemma 2.3. *This diagram commutes. That is, the matrix corresponding to a composition $\varphi \circ \psi$ is the product of the matrices corresponding to φ and ψ.*

Proof. This follows immediately from the associativity of matrix multiplication: for $\mathbf{v} \in R^n$ and $A \in \mathcal{M}_{m,p}(R)$, $B \in \mathcal{M}_{p,n}(R)$,

$$A \cdot (B \cdot \mathbf{v}) = (A \cdot B) \cdot \mathbf{v};$$

that is, the successive action of B and A is equivalent to the action of $A \cdot B$. □

Here is a summary of the situation. Given an integral domain R, we can construct a 'toy category' as follows: we let $\mathbb{Z}^{\geq 0}$ be the set of objects, and we define the set of morphisms $\mathrm{Hom}(n, m)$ to be the set of matrices $\mathcal{M}_{m,n}(R)$, with composition given by the matrix product (cf. Exercise I.3.6). Then we have verified that this toy category is 'essentially the same as' the category of finitely generated free R-modules and R-module homomorphisms[10].

For example, by virtue of Proposition 1.7, if $R = k$ is a field, then R-Mod $= k$-Vect consists exclusively of free k-modules. The toy category we just described is a faithful snapshot of the category of finite-dimensional k-vector spaces.

2.2. Change of basis. It is time to deal with the bad news.

Free modules are only defined *up to isomorphism,* as is any structure satisfying a universal property. Writing a free module as a direct sum $R^{\oplus A}$ amounts to making the choice of one specific realization. As we pointed out at the end of §1.2, by Lemma 1.5 this choice is equivalent to the choice of a basis of the module.

The price to pay for representing homomorphisms of free modules by matrices is that this correspondence relies on the choice of a basis: we say that it is *not canonical.* It is essential to be able to keep track of this choice. For finitely generated free modules, this also boils down to the action of a matrix, as we proceed to see.

Let F be a free module, and choose two bases A, B for F; assume that F is finitely generated, so that A, B are finite sets, and further $|A| = |B|$ $(= \mathrm{rk}\, F)$ by

[10] The reader should not take this statement too seriously: we have identified together all free modules isomorphic to a given one, which is a rather drastic operation. We will take a more formal look at this situation in Example VIII.1.8.

Proposition 1.9. The two bases correspond to two isomorphisms

$$R^{\oplus A} \xrightarrow{\varphi} F, \qquad R^{\oplus B} \xrightarrow{\psi} F\,.$$

Then

$$R^{\oplus A} \xrightarrow{\psi^{-1}\circ\varphi} R^{\oplus B}$$

is an isomorphism, which corresponds to a matrix as seen in Lemma 2.1; we may call this matrix N_A^B. Explicitly, the j-th column of N_A^B is the image of the j-th element of A, viewed as a column vector in $R^{\oplus B}$. That is, letting[11]

$$A = (\mathbf{a}_1, \dots, \mathbf{a}_r), \quad B = (\mathbf{b}_1, \dots, \mathbf{b}_r),$$

we have $N_A^B = (n_{ij})_{\substack{i=1,\cdots,r\\ j=1,\cdots,r}}$, with

$$\psi^{-1} \circ \varphi(\mathbf{a}_j) = \sum_{i=1}^{r} n_{ij}\mathbf{b}_i.$$

This equality is written *in* $R^{\oplus B}$; in practice one views the basis vectors $\mathbf{a}_j$, $\mathbf{b}_i$ as vectors *of* F and would therefore simply write

$$\mathbf{a}_j = \sum_{i=1}^{r} n_{ij}\mathbf{b}_i,$$

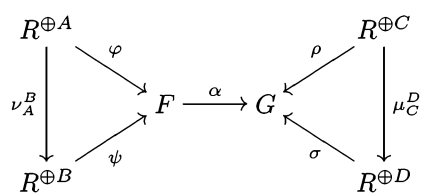
that is, the corresponding equality *in* F.

Definition 2.4. The matrix N_A^B is called the *matrix of the change of basis.* ⌟

Since it represents an isomorphism, the matrix of the change of basis is necessarily an *invertible* matrix.

To our knowledge there is no established convention on the notation for this matrix, and in any case we would find it futile to try to remember any such convention. Any choice of notation will lead to a pleasant 'calculus': for example, the definition given above implies immediately $N_B^A = (N_A^B)^{-1}$ and $N_B^C N_A^B = N_A^C$. In our experience, any such manipulation is immediately clarified by writing isomorphisms out explicitly, so no convention is necessary.

For example, let's work out the action of a change of basis on the matrix representation of a homomophism $\alpha : F \to G$ of two free modules. The diagram taking care of the needed manipulations is



Let N_A^B be the matrix of $\nu_A^B = \psi^{-1} \circ \varphi$, as above, and let M_C^D be the matrix for $\mu_C^D = \sigma^{-1} \circ \rho$.

[11] Ordering the elements of a basis is convenient, for example because it allows us to talk about the 'j-th element' of the basis. Hence we have the 'tuple' notation.

Choose the basis A for F and C for G. Then the matrix representing α will be the matrix P corresponding to

$$\rho^{-1} \circ \alpha \circ \varphi : R^{\oplus A} \to R^{\oplus C}.$$

If on the other hand we choose the basis B for F and D for G, then we represent α by the matrix Q corresponding to

$$\sigma^{-1} \circ \alpha \circ \psi : R^{\oplus B} \to R^{\oplus D}.$$

Now,

$$\sigma^{-1} \circ \alpha \circ \psi = (\mu_C^D \circ \rho^{-1}) \circ \alpha \circ (\varphi \circ (\nu_A^B)^{-1}) = \mu_C^D \circ (\rho^{-1} \circ \alpha \circ \varphi) \circ (\nu_A^B)^{-1},$$

and by Lemma 2.3 this shows

$$Q = M_C^D \cdot P \cdot (N_A^B)^{-1} = M_C^D \cdot P \cdot N_B^A.$$

This is hardly surprising, of course: starting from a vector expressed as a combination of $\mathbf{b}$'s, N_B^A converts it into a combination of $\mathbf{a}$'s; P acts on it by giving its image under α as a combination of $\mathbf{c}$'s; and M_C^D converts it into $\mathbf{d}$'s. That accomplishes Q's job, as it should.

Summarizing this discussion,

Proposition 2.5. *Let $\alpha : F \to G$ be a homomorphism of finitely generated free modules, and let P be a matrix representing it with respect to any choice of bases for F and G. Then the matrices representing α with respect to any other choice of bases are all and only the matrices of the form*

$$M \cdot P \cdot N,$$

where M and N are invertible matrices.

Definition 2.6. Two matrices $P, Q \in \mathcal{M}_{m,n}(R)$ are *equivalent* if they represent the same homomorphism of free modules $R^n \to R^m$ up to a choice of basis. ⌟

This is manifestly an equivalence relation. Properly speaking, the 'abstract' homomorphism $\alpha : F \to G$ is not represented by one matrix as much as by the whole equivalence class with respect to this relation. Proposition 2.5 gives a computational interpretation of equivalence of matrices: P and Q are equivalent if and only if there are *invertible* M and N such that $Q = MPN$.

2.3. Elementary operations and Gaussian elimination. The idea is now to capitalize on Proposition 2.5, as follows: given a homomorphism $\alpha : F \to G$ between two free modules, find 'special' bases in F and G so that the matrix of α takes a particularly convenient form. That is, look for a particularly convenient matrix in each equivalence class with respect to the relation introduced in Definition 2.6.

For this, we can start with random bases in F and G, representing α by a matrix P and then (by Proposition 2.5) multiply P on the right and left by invertible matrices, in order to bring P into whatever form is best suited for our needs.

There is an even more concrete way to deal with equivalence computationally. Consider the following three 'elementary (row/column) operations' that can be performed on a matrix P:

- switch two rows (or two columns) of P;
- add to one row (resp., column) a multiple of another row (resp., column);
- multiply all entries in one row (or column) of P by a unit of R.

Proposition 2.7. *Two matrices $P, Q \in \mathcal{M}_{m,n}(R)$ are equivalent if Q may be obtained from P by a sequence of elementary operations.*

Proof. To see that elementary operations produce equivalent matrices, it suffices (by Proposition 2.5) to express them as multiplications on the left or right[12] by invertible matrices. Indeed, these operations may be performed by suitably multiplying by the matrices obtained from the identity matrix by performing the same operation. For example, multiplying on the left by

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

interchanges the second and fourth row of a $4 \times n$ matrix; multiplying on the right by

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

adds to the third column of a $m \times 3$ matrix the c-multiple of the first column. The diligent reader will formalize this discussion and prove that all these matrices are indeed invertible (Exercise 2.5). □

The matrices corresponding to the elementary operations are called *elementary matrices.* Linear algebra over arbitrary rings would be a great deal simpler if Proposition 2.7 were an 'if and only if' statement. This boils down to the question of whether every invertible matrix may be written as a product of elementary matrices, that is, whether elementary matrices generate the 'general linear group'.

Definition 2.8. The n-th *general linear group* over the ring R, denoted $\mathrm{GL}_n(R)$, is the group of units in $\mathcal{M}_n(R)$, that is, the group of invertible $n \times n$ matrices with entries in R. ⌟

Brief mentions of this notion have occurred already; cf. Example II.1.5 and several exercises in previous chapters.

The elementary matrices are elements of $\mathrm{GL}_n(R)$; in fact, the inverse of an elementary matrix is (of course) again an elementary matrix. The following observation is surely known to the reader, in one form or another; in view of the foregoing considerations, it says that the relation introduced in Definition 2.6 is under good control over *fields.*

[12] Multiplying on the *left* acts on the rows of the matrix; multiplying on the *right* acts on its columns.

Proposition 2.9. *Let $R = k$ be a field, and let $n \geq 0$ be an integer. Then $\mathrm{GL}_n(k)$ is generated by elementary matrices.*

Thus, two matrices are equivalent over a field if and only if they are linked by a sequence of elementary operations.

Proof. Let $A = (a_{ij})$ be an $n \times n$ invertible matrix. In particular, some entry in the first column of A is nonzero; by performing a row switch if necessary, we may assume that a_{11} is nonzero. Multiplying the first row by a_{11}^{-1}, we may assume that $a_{11} = 1$:

$$\begin{pmatrix} 1 & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}.$$

Adding to the second row the $(-a_{21})$-multiple of the first row clears the $(2,1)$ entry. After performing the analogous operation on all rows, we may assume that the only nonzero entry in the first column is the $(1,1)$ entry:

$$\begin{pmatrix} 1 & a_{12} & & a_{1n} \\ 0 & a_{22} & & a_{1n} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2} & \dots & a_{nn} \end{pmatrix}.$$

Similarly, adding to the second column the $(-a_{12})$-multiple of the first column clears the $(1,2)$ entry. Performing this operation on all columns reduces the matrix to the form

$$\begin{pmatrix} 1 & 0 & & 0 \\ 0 & a_{22} & & a_{1n} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2} & \dots & a_{nn} \end{pmatrix} = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & A' \end{array}\right),$$

where A' denotes a (clearly invertible) $(n-1) \times (n-1)$ matrix[13].

Repeating the process on A' and on subsequent smaller matrices reduces A to the identity matrix I_n. In other words, I_n may be obtained from A by a sequence of elementary operations:

$$I_n = M \cdot A \cdot N,$$

where M and N are products of elementary matrices. But then

$$A = M^{-1} \cdot N^{-1}$$

is itself a product of elementary matrices, yielding the statement. □

Note that, with notation as in the preceding proof, $A^{-1} = N \cdot M$; thus, the process explained in the proof may be used to compute the inverse of a matrix (also cf. Exercise 3.5).

When applied only to the *rows* of a matrix, the simplification of a matrix by means of elementary operations is called *Gaussian elimination*. This corresponds

[13]The vertical and horizontal lines alert the reader to the fact that the sectors of the matrix are themselves matrices; this notation is called a *block matrix*.

to multiplying the given matrix *on the left* by a product of elementary matrices, and it suffices in order to reduce any square invertible matrix to the identity (Exercise 2.15).

We will sloppily call 'Gaussian elimination' the more drastic process including column operations as well as row operations; this is in line with our focus on equivalence rather than 'row-equivalence'. Applied to any rectangular matrix, this process yields the following:

Proposition 2.10. *Over a field,* every $m \times n$ matrix is equivalent to a matrix of the form

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

(where $r \leq \min(m, n)$ and '0' stands for null matrices of appropriate sizes).

Different matrices of the type displayed in Proposition 2.10 are *in*equivalent (for example by rank considerations; cf. §3.3). Thus, Proposition 2.10 describes all equivalence classes of matrices over a field and shows that for any given m, n there are in fact only *finitely many* such classes (over a field!).

2.4. Gaussian elimination over Euclidean domains. With due care, Gaussian elimination may be performed over every *Euclidean domain.* A 2×2 example should suffice to illustrate the general case: let

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathcal{M}_2(R),$$

for a Euclidean domain R, with Euclidean valuation N. After switching rows and/or columns if necessary, we may assume that $N(a)$ is the minimum of the valuations of all entries in the matrices. Division with remainder gives

$$b = aq + r$$

with $r = 0$ or $N(r) < N(a)$. Adding to the second column the $(-q)$-multiple of the first produces the matrix

$$\begin{pmatrix} a & r \\ c & d - qc \end{pmatrix}.$$

If $r \neq 0$, so that $N(r) < N(a)$, we begin again and shuffle rows and columns so that the $(1, 1)$ entry has minimum valuation. This process may be repeated, but after a finite number of steps the $(1, 2)$ entry will have to vanish: because valuations are nonnegative integers and at each iteration the valuation of the $(1, 1)$ entry decreases.

Trivial variations of the same procedure will clear the $(2, 1)$ entry as well, producing a matrix

$$\begin{pmatrix} e & 0 \\ 0 & f \end{pmatrix}.$$

Now (this is the cleverest part) we claim that we may assume that e divides f in R, with no remainder. Indeed, otherwise we can add the second row to the first,

$$\begin{pmatrix} e & f \\ 0 & f \end{pmatrix},$$

and *start all over* with this new matrix. Again, the effect of all the operations will be to decrease the valuation of the $(1,1)$ entry, so after a final number of steps we must reach the condition $e \mid f$.

The reader will have some fun describing this process in the general case. The end result is the following useful remark:

Proposition 2.11. *Let R be a Euclidean domain, and let $P \in \mathcal{M}_{m,n}(R)$. Then P is equivalent to a matrix of the form*[14]

$$\left(\begin{array}{ccc|c} d_1 & & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & d_r & 0 \\ \hline 0 & \cdots & 0 & 0 \end{array}\right)$$

with $d_1 \mid \cdots \mid d_r$.

This is called the *Smith normal form* of the matrix.

Proposition 2.11 suffices to prove a weak form of one of our main goals (a classification theorem for finitely generated modules over PIDs) *over Euclidean domains;* this will hopefully be clear in a little while, but the reader can gain the needed insight right away by contemplating Proposition 2.11 vis-à-vis the classification of finite abelian groups (Exercise 2.19).

Remark 2.12. As a consequence of the preceding considerations and arguing as in the proof of Proposition 2.9, we see that $\mathrm{GL}_n(R)$ is generated by elementary matrices if R is a Euclidean domain. The reader may think that some cleverness in handling Gaussian elimination may extend this to more general rings, but this will not go too far: there are examples of PIDs R for which $\mathrm{GL}_n(R)$ is *not* generated by elementary matrices.

On the other hand, some cleverness *does* manage to produce a Smith normal form for any matrix with entries in a PID: the presence of a good gcd suffices to adapt the procedure sketched above (but one may need more than elementary row and column operations). We will take a direct approach to this question in §5. ⌟

Exercises

2.1. Prove that the subset of $\mathcal{M}_2(R)$ consisting of matrices of the form

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

is a group under matrix multiplication and is isomorphic to $(R,+)$.

2.2. ▷ Prove that matrix multiplication is associative. [§2.1]

[14] Here $r \leq \min(m,n)$, the bottom-right 0 stands for a null $(m-r) \times (n-r)$ matrix, etc.

2.3. ▷ Prove that both $\mathcal{M}_n(R)$ and $\mathrm{Hom}_R(R^n, R^n)$ are R-algebras in a natural way and the bijection $\mathrm{Hom}_R(R^n, R^n) \cong \mathcal{M}_n(R)$ of Corollary 2.2 is an isomorphism of R-algebras. (Cf. Exercise III.5.9.) In particular, if the matrix M corresponds to the homomorphism $\varphi : R^n \to R^n$, then M is invertible in $\mathcal{M}_n(R)$ if and only if φ is an isomorphism. [§2.1, §3.2, §6.1]

2.4. Prove Corollary 2.2.

2.5. ▷ Give a formal argument proving Proposition 2.7. [§2.3]

2.6. ¬ A matrix with entries in a field is in *row echelon form* if

- its nonzero rows are all above the zero rows and
- the leftmost nonzero entry of each row is 1, and it is strictly to the right of the leftmost nonzero entry of the row above it.

The matrix is further in *reduced* row echelon form if

- the leftmost nonzero entry of each row is the only nonzero entry in its column.

The leftmost nonzero entries in a matrix in row echelon form are called *pivots.*

Prove that any matrix with entries in a field can be brought into reduced echelon form by a sequence of elementary operations on *rows.* (This is what is more properly called *Gaussian elimination.*) [2.7, 2.9]

2.7. ¬ Let M be a matrix with entries in a field and in reduced row echelon form (Exercise 2.6). Prove that if a row vector $\mathbf{r}$ is a linear combination $\sum a_i \mathbf{r}_i$ of the nonzero rows of M, then a_i equals the component of $\mathbf{r}$ at the position corresponding to the pivot on the i-th row of M. Deduce that the nonzero rows of M are linearly independent. [2.9]

2.8. ¬ Two matrices M, N are *row-equivalent* if $M = PN$ for an invertible matrix P. Prove that this is indeed an equivalence relation, and that two matrices with entries in a field are row-equivalent if and only if one may be obtained from the other by a sequence of elementary operations on rows. [2.9, 2.12]

2.9. ¬ Let k be a field, and consider row-equivalence (Exercise 2.8) on the set of $m \times n$ matrices $\mathcal{M}_{m,n}(k)$. Prove that each equivalence class contains exactly one matrix in reduced row echelon form (Exercise 2.6). (Hint: To prove uniqueness, argue by contradiction. Let M, N be different row-equivalent reduced row echelon matrices; assume that they have the minimum number of columns with this property. If the leftmost column at which M and N differ is the k-th column, use the minimality to prove that M, N may be assumed to be of the form

$$\left(\begin{array}{c|c} I_{k-1} & * \\ \hline 0 & * \end{array}\right) \quad \text{or} \quad \left(\begin{array}{c|c} I_{k-1} & * \end{array}\right).$$

Use Exercise 2.7 to obtain a contradiction.)

The unique matrix in reduced row echelon form that is row-equivalent to a given matrix M is called the *reduced echelon form* of M. [2.11]

2.10. ▷ The *row space* of a matrix M is the span of its rows; the *column space* of M is the span of its column. Prove that row-equivalent matrices have the same row space and isomorphic column spaces. [2.12, §3.3]

2.11. Let k be a field and $M \in \mathcal{M}_{m,n}(k)$. Prove that the dimension of the space spanned by the rows of M equals the number of nonzero rows in the reduced echelon form of M (cf. Exercise 2.9).

2.12. ¬ Let k be a field, and consider row-equivalence on $\mathcal{M}_{m,n}(k)$ (Exercise 2.8). By Exercise 2.10, row-equivalent matrices have the same row space. Prove that, conversely, there is exactly one row-equivalence class in $\mathcal{M}_{m,n}(k)$ for each subspace of k^n of dimension $\leq m$. [2.13, 2.14]

2.13. ¬ The set of subspaces of given dimension in a fixed vector space is called a *Grassmannian.* In Exercise 2.12 you have constructed a bijection between the Grassmannian of r-dimensional subspaces of k^n and the set of reduced row echelon matrices with n columns and r nonzero rows.

For $r = 1$, the Grassmannian is called the *projective space.* For a vector space V, the corresponding projective space $\mathbb{P}V$ is the set of 'lines' (1-dimensional subspaces) in V. For $V = k^n$, $\mathbb{P}V$ may be denoted $\mathbb{P}_k^{n-1}$, and the field k may be omitted if it is clear from the context. Show that $\mathbb{P}_k^{n-1}$ may be written as a union $k^{n-1} \cup k^{n-2} \cup \cdots \cup k^1 \cup k^0$, and describe each of these subsets 'geometrically'.

Thus, $\mathbb{P}^{n-1}$ is the union of n 'cells'[15], the largest one having dimension $n-1$ (accounting for the choice of notation). Similarly, all Grassmannians may be written as unions of cells. These are called *Schubert cells.*

Prove that the Grassmannian of $(n-1)$-dimensional subspaces of k^n admits a cell decomposition entirely analogous to that of $\mathbb{P}_k^{n-1}$. (This phenomenon will be explained in Exercise VIII.5.17.) [VII.2.20, VIII.4.7, VIII.5.17]

2.14. ▷ Show that the Grassmannian $\mathrm{Gr}_k(2,4)$ of 2-dimensional subspaces of k^4 is the union of 6 Schubert cells: $k^4 \cup k^3 \cup k^2 \cup k^2 \cup k^1 \cup k^0$. (Use Exercise 2.12; list all the possible reduced echelon forms.) [VIII.4.8]

2.15. ▷ Prove that a square matrix with entries in a field is invertible if and only if it is equivalent to the identity, if and only if it is row-equivalent to the identity, if and only if its reduced echelon form is the identity. [§2.3, 3.5]

2.16. Prove Proposition 2.10.

2.17. Prove Proposition 2.11.

2.18. Suppose $\alpha : \mathbb{Z}^3 \to \mathbb{Z}^2$ is represented by the matrix

$$\begin{pmatrix} -6 & 12 & 18 \\ -15 & 36 & 54 \end{pmatrix}$$

with respect to the standard bases. Find bases of $\mathbb{Z}^3$, $\mathbb{Z}^2$ with respect to which α is given by a matrix of the form obtained in Proposition 2.11.

[15] Here, a 'cell' is simply a subset endowed with a natural bijection with k^ℓ for some ℓ.

2.19. ▷ Prove Corollary IV.6.5 again as a corollary of Proposition 2.11. In fact, prove the more general fact that every *finitely generated* abelian group is a direct sum of cyclic groups. [§II.6.3, §IV.6.1, §IV.6.3, §2.4, §4.1, §4.3]

---

3. Homomorphisms of free modules, II

The work in §2 accomplishes the goal of describing $\mathrm{Hom}_R(F, G)$, where F and G are free R-modules of finite rank; as we have seen, this can be done most explicitly if, for example, R is a field or a Euclidean domain. We are not quite done, though: even over fields, it is important to 'understand the answer'—that is, examine the classification of homomorphisms into finitely many classes, highlighted at the end of §2.3. Also, the frequent appearance of invertible matrices makes it necessary to develop criteria to tell whether a given matrix is or is not in the general linear group. We begin by pointing out a straightforward application of the classification of homomorphisms.

3.1. Solving systems of linear equations. A moment's thought reveals that Gaussian elimination, through the auspices of statements such as Proposition 2.9, gives us a tool to solve systems of linear equations over a field or a Euclidean domain[16]. This is another point which does not seem to warrant a careful treatment or memory gymnastic: writing things out carefully should take care of producing tools as need be. To describe a typical situation, suppose

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \quad \cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \end{cases}$$

is a system of m equations in n unknowns, with a_{ij}, b_i in a Euclidean domain R. We want to find all solutions $x_1, \ldots, x_n$ in R.

With $A = \begin{pmatrix} a_{11} & & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$, and $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, this amounts to 'solving for $\mathbf{x}$' in the matrix equation

$$A \cdot \mathbf{x} = \mathbf{b}.$$

Of course, if $m = n$ and the matrix A is square and invertible, then the solutions are obtained simply as

$$\mathbf{x} = A^{-1} \cdot \mathbf{b}.$$

Here A^{-1} may be obtained through Gaussian elimination (cf. Proposition 2.9) or through determinants (§3.2).

---

[16] Of course one can deal with systems over arbitrary integral domains R by reducing to the field case, by embedding R in its field of fractions.

Even without requiring anything special of the 'matrix of coefficients' A, row and column operations will take it to the standard form presented in Proposition 2.11; that is, it will yield invertible matrices M, N such that

$$M \cdot A \cdot N = \left(\begin{array}{ccc|c} d_1 & & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & d_r & 0 \\ \hline 0 & \cdots & 0 & 0 \end{array}\right)$$

with notation as in Proposition 2.11. Gaussian elimination is a constructive procedure: watching yourself as you switch/combine rows or columns will produce M and N explicitly. Now, letting $\mathbf{y} = (y_j)$ and $\mathbf{c} = M\mathbf{b}$, the system

$$(M \cdot A \cdot N) \cdot \mathbf{y} = \mathbf{c}$$

solves itself:

$$\begin{cases} d_1 y_1 = c_1 \\ \\ d_r y_r = c_r \\ 0 = c_{r+1} \\ \\ \end{cases}$$

has solutions if and only if $d_j \mid c_j$ for all j and $c_j = 0$ for $j > r$; moreover, in this case $y_j = d_j^{-1} c_j$ for $j = 1, \ldots, r$, and y_j is arbitrary for $j > r$. This yields $\mathbf{y}$, and the reader will check that $\mathbf{x} = N\mathbf{y}$ gives all solutions to the original system.

Such arguments can be packaged into convenient explicit procedures to solve systems of linear equations. Again, it seems futile to list (or try to remember) any such procedure; it seems more important to know what is behind such techniques, so as to be able to come up with one when needed.

In any case, the reader should be able to justify such recipes rigorously. The most famous one is possibly *Cramer's rule* (Proposition 3.6) which relies on determinants and is, incidentally, essentially useless in practice for any decent-size problem.

3.2. The determinant. Let $\alpha : F \to G$ be a homomorphism of free R-modules of the same rank, and let A be the matrix representing α with respect to a choice of bases for F and G. As a consequence of Lemma 2.3 (cf. Exercise 2.3), α is an isomorphism if and only if A is a unit in $\mathcal{M}_n(R)$, that is, if and only if it is invertible as a matrix with entries in R. This may be detected by computing the *determinant* of A.

Definition 3.1. Let $A = (a_{ij}) \in \mathcal{M}_n(R)$ be a square matrix. Then the *determinant* of A is the element

$$\det(A) = \sum_{\sigma \in S_n} (-1)^\sigma \prod_{i=1}^{n} a_{i\sigma(i)} \in R. \qquad \lrcorner$$

Here S_n denotes the symmetric group on $\{1,\dots,n\}$, and we write[17] $\sigma(i)$ for the action of $\sigma \in S_n$ on $i \in \{1,\dots,n\}$; $(-1)^\sigma$ is the *sign* of a permutation (Definition IV.4.10, Lemma IV.4.12).

The reader is surely familiar with determinants, at least over fields. They satisfy a number of remarkable properties; here is a selection:

—The determinant of a matrix A equals the determinant of its *transpose* $A^t = (a^t_{ij})$, defined by setting

$$a^t_{ij} = a_{ji}$$

for all i and j (that is, the rows of A^t are the columns of A). Indeed,

$$\det A^t = \sum_{\sigma\in S_n} (-1)^\sigma \prod_{i=1}^n a^t_{i\sigma(i)} = \sum_{\sigma\in S_n} (-1)^\sigma \prod_{i=1}^n a_{\sigma(i)i} = \sum_{\sigma\in S_n} (-1)^\sigma \prod_{i=1}^n a_{i\sigma^{-1}(i)}$$

by the commutativity of the product in R; and σ^{-1} ranges over all permutations of $\{1,\dots,n\}$ as σ does the same, and with the same sign, so the right-most term equals $\det A$.

—If two rows or columns of a square matrix A agree, then $\det A = 0$. Indeed, it is enough to check this for matching columns; the case for rows follows by applying the previous observation. If columns j and j' of A are equal, the contribution to $\det A$ due to a $\sigma \in S_n$ is equal and opposite in sign to the contribution due to the product of σ and the transposition (jj'), so $\det A = 0$.

—Suppose $A = (a_{ij})$ and $B = (b_{ij})$ agree on all but at most one row: $a_{ij} = b_{ij}$ if $i \neq k$, for all j and some fixed k. Let $c_{ij} := a_{ij} = b_{ij}$ for $i \neq k$, $c_{kj} := a_{kj} + b_{kj}$, and let $C := (c_{ij})$. Then

$$\det(C) = \det(A) + \det(B).$$

This follows immediately from Definition 3.1 and distributivity. Applying this observation to the transpose matrices gives an analogous statement for matrices differing at most along a column.

We will record more officially the effect of elementary operations on determinants:

Lemma 3.2. *Let A be a square matrix with entries in an integral domain R.*

- *Let A' be obtained from A by switching two rows or two columns. Then* $\det(A') = -\det(A)$.
- *Let A' be obtained from A by adding to a row (column) a multiple of another row (column). Then* $\det(A') = \det(A)$.
- *Let A' be obtained from A by multiplying a row (column) by an element[18] $c \in R$. Then* $\det(A') = c\det(A)$.

In other words, the effect of an elementary operation on $\det A$ *is the same as multiplying* $\det A$ *by the determinant of the corresponding elementary matrix.*

[17] Consistency with previous encounters with the symmetric group (e.g., §IV.4) would demand that we write $i\sigma$; but then we would end up with things like '$a_{ii\sigma}$', which are very hard to parse.

[18] For an elementary operation, c should be a unit; this restriction is not necessary here.

Proof. These are all essentially immediate from Definition 3.1. For example, switching two columns amounts to correcting each σ in the definition by a fixed transposition, changing the sign of all contributions to the $\sum$ in the definition. The third point is immediate from distributivity. Combining the third operation and the two remarks preceding the statement yields the second point. Details are left to the reader (Exercise 3.2). □

This observation simplifies the theory of determinants drastically. If $R = k$ is a field and $P \in \mathcal{M}_n(k)$, Gaussian elimination (Proposition 2.10) shows

$$A = E_1 \cdots E_a \cdot \left(\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right) \cdot E'_1 \cdots E'_b,$$

where $r \leq n$ and E_i, E'_j are elementary matrices. Then Lemma 3.2 gives that

$$\det(A) = \prod_i \det(E_i) \prod_j \det(E'_j) \det \left(\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right).$$

(In particular, $\det A \neq 0$ only if $r = n$.) Useful facts about the determinant follow from this remark. For example,

Proposition 3.3. *Let R be a commutative ring.*

- *A square matrix $A \in \mathcal{M}_n(R)$ is invertible if and only if $\det A$ is a unit in R.*
- *The determinant is a homomorphism*[19] $\mathrm{GL}_n(R) \to (R^*, \cdot)$. *More generally, for $A, B \in \mathcal{M}_n(R)$,*

$$\det(A \cdot B) = \det(A) \det(B).$$

Proof for $R =$ a field. If $R = k$ is a field, we can use the considerations immediately preceding the statement. The first point is reduced to the case of a block matrix

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

for which it is immediate. In fact, this shows that $\det A = 0$ if and only if the linear map $k^n \to k^n$ corresponding to A is not an isomorphism. In particular, for all A, B we have $\det(AB) = 0$ if and only if AB is not an isomorphism, if and only if A or B is not an isomorphism, if and only if $\det(A) = 0$ or $\det(B) = 0$. So we only need to check the homomorphism property for invertible matrices. These are products of elementary matrices 'on the nose', and the homomorphism property then follows from Lemma 3.2. □

Before giving the (easy) extension to the case of arbitrary commutative rings, it is helpful to note the following explicit formulas. A 'submatrix' obtained from a given matrix A by removing a number of rows and columns is called a *minor* of A; more properly, this term refers to the *determinants of square submatrices* obtained

[19] Recall that $(R^*, \cdot)$ denotes the groups of units of R.

in these ways. If $A \in \mathcal{M}_n(R)$, the *cofactors* of A are the $(n-1) \times (n-1)$ minors of A, corrected by a sign. More precisely, for $A = (a_{ij})$ we will let

$$A^{(ij)} := (-1)^{i+j} \det \begin{pmatrix} a_{11} & a_{1j-1} & a_{1j+1} & a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i-11} & a_{i-1j-1} & a_{i-1j+1} & a_{i-1n} \\ a_{i+11} & a_{i+1j-1} & a_{i+1j+1} & a_{i+1n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{nj-1} & a_{nj+1} & a_{nn} \end{pmatrix}.$$

Lemma 3.4. *With notation as above,*

- *for all* $i = 1, \ldots, n$, $\det(A) = \sum_{j=1}^{n} a_{ij} A^{(ij)}$,
- *for all* $j = 1, \ldots, n$, $\det(A) = \sum_{i=1}^{n} a_{ij} A^{(ij)}$.

Proof. This is a simple (if slightly messy) induction on n, which we leave to the diligent reader. □

Of course Lemma 3.4 is simply stating the famous strategy computing a determinant by expanding it with respect to your favorite row or column. This works wonderfully for very small matrices and is totally useless for large ones, since the number of computations needed to apply it grows as the factorial of the size of the matrix. From a computational point of view, it makes much better sense to apply Gaussian elimination and use the considerations preceding Proposition 3.3.

But Lemma 3.4 has the following important implication:

Corollary 3.5. *Let* R *be a commutative ring and* $A \in \mathcal{M}_n(R)$. *Then*

$$A \cdot \begin{pmatrix} A^{(11)} & A^{(n1)} \\ \vdots & \vdots \\ A^{(1n)} & A^{(nn)} \end{pmatrix} = \begin{pmatrix} A^{(11)} & A^{(n1)} \\ \vdots & \vdots \\ A^{(1n)} & A^{(nn)} \end{pmatrix} \cdot A = \det(A)\, I_n.$$

Note the switch in the role of i and j in the matrix of cofactors. This matrix is called the *adjoint* matrix of A.

Proof. Along the diagonal of the right-hand side, this is a restatement of Lemma 3.4. Off the diagonal, one is evaluating (for example)

$$\sum_{j=1}^{n} a_{i'j} A^{(ij)}$$

for $i' \neq i$. By Lemma 3.4 this is the same as the determinant of the matrix obtained by replacing the i-th row with the i'-th row; the resulting matrix has two equal rows, so its determinant is 0, as needed. □

In particular, Corollary 3.5 proves that we can invert a matrix if we can invert its determinant:

$$A^{-1} = \det(A)^{-1} \begin{pmatrix} A^{(11)} & & A^{(n1)} \\ \vdots & & \vdots \\ A^{(1n)} & & A^{(nn)} \end{pmatrix};$$

this holds over any commutative ring, as soon as $\det A$ is a unit. In practice the computation of cofactors is 'expensive', so in any given concrete case Gaussian elimination is likely a better alternative (at least over fields); cf. Exercise 3.5.

But this formula for the inverse has good theoretical significance. For example, we are now in a position to complete the proof of Proposition 3.3.

Proof of Proposition 3.3 for commutative rings. The first point follows from the second and from what we have just seen. Indeed, we have checked that $A \in \mathcal{M}_n(R)$ admits an inverse $A^{-1} \in \mathcal{M}_n(R)$ if $\det(A)$ is a unit in R; conversely, if A admits an inverse $A^{-1} \in \mathcal{M}_n(R)$, then

$$\det(A)\det(A^{-1}) = \det(AA^{-1}) = \det(I_n) = 1$$

by the second statement, so that $\det(A^{-1})$ is the inverse of $\det(A)$ in R.

Thus, we just need to verify the second point, that is, the homomorphism property of determinants. In order to verify this over *every* (commutative) ring, it suffices to verify the 'universal' identity obtained by writing out the claimed equality, for matrices with indeterminate entries. For example, for $n = 2$ the statement is

$$\det \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix} \det \begin{pmatrix} y_1 & y_2 \\ y_3 & y_4 \end{pmatrix} = \det \begin{pmatrix} x_1y_1 + x_2y_3 & x_1y_2 + x_2y_4 \\ x_3y_1 + x_4y_3 & x_3y_2 + x_4y_4 \end{pmatrix},$$

which translates into the identity

$$(x_1x_4 - x_2x_3)(y_1y_4 - y_2y_3) = (x_1y_1 + x_2y_3)(x_3y_2 + x_4y_4) - (x_1y_2 + x_2y_4)(x_3y_1 + x_4y_3);$$

since this identity holds in $\mathbb{Z}[x_1, \ldots, y_4]$, it must hold in any commutative ring, for any choice of $x_1, \ldots, y_4$: indeed, $\mathbb{Z}$ is initial in Ring.

Now, we have verified that the homomorphism property holds over fields; in particular it holds over the field of fractions of $\mathbb{Z}[x_{11}, \ldots, x_{nn}, y_{11}, \ldots, y_{nn}]$. It follows that it does hold in $\mathbb{Z}[x_{11}, \ldots, x_{nn}, y_{11}, \ldots, y_{nn}]$, and we are done. □

The 'universal identity' argument extending the result from fields to arbitrary commutative rings is a useful device, and the reader is invited to contemplate it carefully.

As an application of determinants (and especially cofactors) we can now go back to the special case of a system of n equations in n unknowns

$$A \cdot \mathbf{x} = \mathbf{b},$$

cf. §3.1, in the case in which $\det A$ is a unit.

Proposition 3.6 (Cramer's rule). *Assume* $\det(A)$ *is a unit, and let* $A^{(j)}$ *be the matrix obtained by replacing the* j*-th column of* A *by the column vector* b*. Then*

$$x_j = \det(A)^{-1} \det(A^{(j)}).$$

Proof. Using Lemma 3.4, expand $\det(A^{(j)})$ with respect to the j-th column:

$$\det(A^{(j)}) = \sum_{i=1}^{n} A^{(ij)} b_i.$$

Therefore

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = A^{-1}\mathbf{b} = \det(A)^{-1} \begin{pmatrix} A^{(11)} & & A^{(n1)} \\ \vdots & & \vdots \\ A^{(1n)} & & A^{(nn)} \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

$$= \det(A)^{-1} \begin{pmatrix} \sum_{i=1}^{n} A^{(i1)} b_i \\ \vdots \\ \sum_{i=1}^{n} A^{(in)} b_i \end{pmatrix} = \begin{pmatrix} \det(A)^{-1} \det(A^{(1)}) \\ \vdots \\ \det(A)^{-1} \det(A^{(n)}) \end{pmatrix},$$

which gives the statement. □

3.3. Rank and nullity. According to Proposition 2.10, each equivalence class of matrices *over a field* has a representative of the type

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

The reason why two different $m \times n$ matrices of this type are surely *in*equivalent is that if $\alpha : V \to W$ (with V and W free modules, vector spaces in this case) is represented by a matrix of this form, then r is the *dimension of the image* of α. Therefore, matrices with different r cannot represent the same α.

This integer r is called the *rank* of the matrix and deserves some attention. We will discuss it for matrices with entries in a field, leaving generalizations to more general rings to the reader (for now at least).

The *column (row) space* of a matrix P over a field k is the span of the columns (rows) of P. The *column (row) rank* of P is the dimension of the column (row) space of P.

Proposition 3.7. *The row rank of a matrix over a field k equals its column rank.*

Proof. Equivalent matrices have the same ranks. Indeed, let $P \in \mathcal{M}_{m,n}(k)$; the row space of P consists of all row vectors

$$\begin{pmatrix} a_1 & & a_m \end{pmatrix} = \begin{pmatrix} v_1 & & v_m \end{pmatrix} \cdot P$$

obtained as each v_i ranges in k. If $Q = MPN$, with M and N invertible, let $\begin{pmatrix} w_1 & \cdots & w_m \end{pmatrix} = \begin{pmatrix} v_1 & \cdots & v_m \end{pmatrix} \cdot M^{-1}$; then

$$\begin{pmatrix} w_1 & & w_m \end{pmatrix} \cdot Q = \begin{pmatrix} v_1 & & v_m \end{pmatrix} M^{-1}(MPN) = \begin{pmatrix} a_1 & & a_m \end{pmatrix} \cdot N.$$

This shows that multiplication on the right by N maps the row space of P (isomorphically, since N is invertible) to the row space of Q; thus the two spaces have the same dimension, as claimed. Minimal variations on the same argument show that the column ranks of P and Q agree. (Cf. Exercise 2.10.)

Applying this observation and Proposition 2.10 reduces the question to matrices

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

for which row rank $= r =$ column rank, proving the statement. □

The result upgrades easily to matrices over an arbitrary integral domain R (applying the usual trick of embedding R in its field of fractions).

In view of Proposition 3.7, we can simply talk about the *rank* of a matrix:

Definition 3.8. Let $M \in \mathcal{M}_{m,n}(k)$ be a matrix over a field k. The *rank* of M is the dimension of its column (or, equivalently, row) space. ⌟

One way to rephrase Proposition 2.10 is that matrices over a field are classified up to equivalence by their rank.

The foregoing considerations translate nicely in more abstract terms for a linear map $\alpha : V \to W$ between finite-dimensional vector spaces over a field k. Using the convenient language introduced in §III.7.1, note that each α determines an exact sequence of vector spaces

$$0 \longrightarrow \ker \alpha \longrightarrow V \longrightarrow \operatorname{im} \alpha \longrightarrow 0 \,.$$

Definition 3.9. The *rank* of α, denoted $\operatorname{rk} \alpha$, is the dimension of $\operatorname{im} \alpha$. The *nullity* of α is $\dim(\ker \alpha)$. ⌟

Claim 3.10. *Let $\alpha : V \to W$ be a linear map of finite-dimensional vector spaces. Then*

$$(\text{rank of } \alpha) + (\text{nullity of } \alpha) = \dim V.$$

Proof. Let $n = \dim V$ and $m = \dim W$. By Proposition 2.10 we can represent α by an $n \times m$ matrix of the form

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

From this representation it is immediate that $\operatorname{rk} \alpha = r$ and the nullity of α is $n - r$, with the stated consequence. □

Summarizing, $\operatorname{rk} \alpha$ equals the (column) rank of any matrix P representing α; similarly, the nullity of α equals 'dim V minus the (row) rank' of P. Claim 3.10 is the abstract version of the equality of row rank and column rank.

3.4. Euler characteristic and the Grothendieck group. Against our best efforts, we cannot resist extending these simple observations to more general complexes. Claim 3.10 may be reformulated as follows:

Proposition 3.11. *Let*

$$0 \longrightarrow U \longrightarrow V \longrightarrow W \longrightarrow 0$$

be a short exact *sequence of finite-dimensional vector spaces. Then*

$$\dim(V) = \dim(U) + \dim(W).$$

Equivalently, this amounts to the relation $\dim(V/U) = \dim(V) - \dim(U)$.

Consider then a *complex of finite-dimensional vector spaces and linear maps:*

$$V_\bullet: \quad 0 \longrightarrow V_N \xrightarrow{\alpha_N} V_{N-1} \xrightarrow{\alpha_{N-1}} \cdots \xrightarrow{\alpha_2} V_1 \xrightarrow{\alpha_1} V_0 \longrightarrow 0$$

(cf. §III.7.1). Thus, $\alpha_{i-1} \circ \alpha_i = 0$ for all i. This condition is equivalent to the requirement that $\operatorname{im}(\alpha_{i+1}) \subseteq \ker(\alpha_i)$; recall that the *homology* of this complex is defined as the collection of spaces

$$H_i(V_\bullet) = \frac{\ker(\alpha_i)}{\operatorname{im}(\alpha_{i+1})}.$$

The complex is *exact* if $\operatorname{im}(\alpha_{i+1}) = \ker(\alpha_i)$ for all i, that is, if $H_i(V_\bullet) = 0$ for all i.

Definition 3.12. The *Euler characteristic* of $V_\bullet$ is the integer

$$\chi(V_\bullet) := \sum_i (-1)^i \dim(V_i). \qquad \lrcorner$$

The original motivation for the introduction of this number is topological: with suitable positions, this Euler characteristic equals the Euler characteristic obtained by triangulating a manifold and then computing the number of vertices of the triangulation, minus the number of edges, plus the number of faces, etc.

The following simple result is then a straightforward (and very useful) generalization of Proposition 3.11:

Proposition 3.13. *With notation as above,*

$$\chi(V_\bullet) = \sum_{i=0}^{N} (-1)^i \dim(H_i(V_\bullet)).$$

In particular, if $V_\bullet$ is exact, then $\chi(V_\bullet) = 0$.

Proof. There is nothing to show for $N = 0$, and the result follows directly from Proposition 3.11 if $N = 1$ (Exercise 3.15). Arguing by induction, given a complex

$$V_\bullet: \quad 0 \longrightarrow V_N \xrightarrow{\alpha_N} V_{N-1} \xrightarrow{\alpha_{N-1}} \cdots \xrightarrow{\alpha_2} V_1 \xrightarrow{\alpha_1} V_0 \longrightarrow 0,$$

we may assume that the result is known for 'shorter' complexes. Consider then the truncation

$$V'_\bullet: \quad 0 \longrightarrow V_{N-1} \xrightarrow{\alpha_{N-1}} \cdots \xrightarrow{\alpha_2} V_1 \xrightarrow{\alpha_1} V_0 \longrightarrow 0\,.$$

Then

$$\chi(V_\bullet) = \chi(V'_\bullet) + (-1)^N \dim(V_N),$$

and

$$H_i(V_\bullet) = H_i(V'_\bullet) \quad \text{for } 0 \le i \le N-2,$$

while

$$H_{N-1}(V'_\bullet) = \ker(\alpha_{N-1}), \quad H_{N-1}(V_\bullet) = \frac{\ker(\alpha_{N-1})}{\operatorname{im}(\alpha_N)}, \quad H_N(V_\bullet) = \ker(\alpha_N).$$

By Proposition 3.11 (cf. Claim 3.10),

$$\dim(V_N) = \dim(\operatorname{im}(\alpha_N)) + \dim(\ker(\alpha_N))$$

and

$$\dim(H_{N-1}(V_\bullet)) = \dim(\ker(\alpha_{N-1})) - \dim(\mathrm{im}(\alpha_N));$$

therefore

$$\dim(H_{N-1}(V'_\bullet)) - \dim(V_N) = \dim(H_{N-1}(V_\bullet)) - \dim(H_N(V_\bullet)).$$

Putting all of this together with the induction hypothesis,

$$\chi(V'_\bullet) = \sum_{i=0}^{N-1} (-1)^i \dim(H_i(V'_\bullet))$$

gives

$$\begin{aligned}
\chi(V_\bullet) &= \chi(V'_\bullet) + (-1)^N \dim(V_N) \\
&= \sum_{i=0}^{N-1} (-1)^i \dim(H_i(V'_\bullet)) + (-1)^N \dim(V_N) \\
&= \sum_{i=0}^{N-2} (-1)^i \dim(H_i(V'_\bullet)) + (-1)^{N-1}(\dim(H_{N-1}(V'_\bullet)) - \dim(V_N)) \\
&= \sum_{i=0}^{N-2} (-1)^i \dim(H_i(V_\bullet)) + (-1)^{N-1}(\dim(H_{N-1}(V_\bullet)) - \dim(H_N(V_\bullet))) \\
&= \sum_{i=0}^{N} (-1)^i \dim(H_i(V_\bullet))
\end{aligned}$$

as needed. □

In terms of the topological motivation recalled above, Proposition 3.13 tells us that the Euler characteristic of a manifold may be computed as the alternating sum of the ranks of its homology, that is, of its *Betti numbers.*

Having come this far, we cannot refrain from mentioning the next, equally simple-minded, generalization. The reader has surely noticed that the *only* tool used in the proof of Proposition 3.13 was the 'additivity' property of dimension, established in Proposition 3.11: if

$$0 \longrightarrow U \longrightarrow V \longrightarrow W \longrightarrow 0$$

is exact, then

$$\dim(V) = \dim(U) + \dim(W).$$

Proposition 3.13 is a formal consequence of this one property of dim.

With this in mind, we can reinterpret what we have just done in the following curious way. Consider the category $k\text{-}\mathsf{Vect}^f$ of finite-dimensional k-vector spaces. Each object V of $k\text{-}\mathsf{Vect}^f$ determines an isomorphism class $[V]$. Let $F(k\text{-}\mathsf{Vect}^f)$ be the free abelian group on the set of these isomorphism classes; further, let E be the subgroup generated by the elements

$$[V] - [U] - [W]$$

for all short exact sequences

$$0 \longrightarrow U \longrightarrow V \longrightarrow W \longrightarrow 0$$

in k-Vect^f. The quotient group

$$K(k\text{-Vect}^f) := \frac{F(k\text{-Vect}^f)}{E}$$

is called the *Grothendieck group* of the category k-Vect^f. The element determined by V in the Grothendieck group is still denoted $[V]$.

More generally, a *Grothendieck group* may be defined for any category admitting a notion of exact sequence.

Every complex $V_\bullet$ determines an element in $K(k\text{-Vect}^f)$, namely

$$\chi_K(V_\bullet) := \sum_i (-1)^i [V_i] \in K(k\text{-Vect}^f).$$

Claim 3.14. *With notation as above, we have the following:*

- χ_K *'is an Euler characteristic', in the sense that it satisfies the formula given in Proposition 3.13:*

$$\chi_K(V_\bullet) = \sum_i (-1)^i [H_i(V_\bullet)].$$

- χ_K *is a 'universal Euler characteristic', in the following sense. Let G be an abelian group, and let δ be a function associating an element of G to each finite-dimensional vector space, such that $\delta(V) = \delta(V')$ if $V \cong V'$ and $\delta(V/U) = \delta(V) - \delta(U)$. For $V_\bullet$ a complex, define*

$$\chi_G(V_\bullet) = \sum_i (-1)^i \delta(V_i).$$

  Then δ induces a (unique) group homomorphism

$$K(k\text{-Vect}^f) \to G$$

  mapping $\chi_K(V_\bullet)$ to $\chi_G(V_\bullet)$.

- *In particular, $\delta = \dim$ induces a group homomorphism*

$$K(k\text{-Vect}^f) \to \mathbb{Z}$$

  such that $\chi_K(V_\bullet) \mapsto \chi(V_\bullet)$.

- *This is in fact an isomorphism.*

The best way to convince the reader that this impressive claim is completely trivial is to leave its proof to the reader (Exercise 3.16). The first point is proved by adapting the proof of Proposition 3.13; the second point is a harmless mixture of universal properties; the third point follows from the second; and the last point follows from the fact that $\dim(k) = 1$ and the IBN property.

The last point is, in fact, rather anticlimactic: if the impressively abstract Grothendieck group turns out to just be a copy of the integers, why bother defining it? The answer is, of course, that this only stresses how special the category k-Vect^f is. The definition of the Grothendieck group can be given in any context in

which complexes and a notion of exactness are available (for example, in the category of finitely generated[20] modules over any ring). The formal arguments proving Claim 3.14 will go through in any such context and provide us with a useful notion of 'universal Euler characteristic'.

We will come back to complexes and homology in Chapter IX.

Exercises

3.1. Use Gaussian elimination to find all integer solutions of the system of equations

$$\begin{cases} 7x - 36y + 12z = 1, \\ -8x + 42y - 14z = 2. \end{cases}$$

3.2. ▷ Provide details for the proof of Lemma 3.2. [§3.2]

3.3. Redo Exercise II.8.8.

3.4. Formalize the discussion of 'universal identities': by what cocktail of universal properties is it true that if an identity holds in $\mathbb{Z}[x_1, \dots, x_r]$, then it holds over every commutative ring R, for every choice of $x_i \in R$? (Is the commutativity of R necessary?)

3.5. ▷ Let A be an $n \times n$ square invertible matrix with entries in a field, and consider the $n \times (2n)$ matrix $B = (A|I_n)$ obtained by placing the identity matrix to the side of A. Perform elementary row operations on B so as to reduce A to I_n (cf. Exercise 2.15). Prove that this transforms B into $(I_n|A^{-1})$.

(This is a much more efficient way to compute the inverse of a matrix than by using determinants as in §3.2.) [§2.3, §3.2]

3.6. ¬ Let R be a commutative ring and $M = \langle m_1, \dots, m_r \rangle$ a finitely generated R-module. Let $A \in \mathcal{M}_r(R)$ be a matrix such that $A \cdot \begin{pmatrix} m_1 \\ \vdots \\ m_r \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$. Prove that $\det(A)m = 0$ for all $m \in M$. (Hint: Multiply by the adjoint.) [3.7]

3.7. ¬ Let R be a commutative ring, M a finitely generated R-module, and let J be an ideal of R. Assume $JM = M$. Prove that there exists an element $b \in J$ such that $(1 + b)M = 0$. (Let $m_1, \dots, m_r$ be generators for M. Find an $r \times r$ matrix B with entries in J such that $\begin{pmatrix} m_1 \\ \vdots \\ m_r \end{pmatrix} = B \cdot \begin{pmatrix} m_1 \\ \vdots \\ m_r \end{pmatrix}$. Then use Exercise 3.6.) [3.8, VIII.1.18]

[20] Note that some finiteness condition is likely to be necessary. We cannot define a 'Grothendieck group of k-Vect' because the isomorphism classes of objects in k-Vect do not form a set: there is one for each cardinal number, and cardinal numbers do not form a set.

3.8. ¬ Let R be a commutative ring, M be a finitely generated R-module, and let J be an ideal of R *contained in the Jacobson radical* of R (Exercise V.3.14). Prove that $M = 0 \iff JM = M$. (Use Exercise 3.7. This is *Nakayama's lemma,* a result with important applications in commutative algebra and algebraic geometry. A particular case was given as Exercise III.5.16.) [III.5.16, 3.9, 5.5]

3.9. ¬ Let R be a commutative local ring, that is, a ring with a single maximal ideal $\mathfrak{m}$, and let M, N be finitely generated R-modules. Prove that if $M = \mathfrak{m}M + N$, then $M = N$. (Apply Nakayama's lemma, that is, Exercise 3.8, to M/N. Note that the Jacobson radical of R is $\mathfrak{m}$.) [3.10]

3.10. ¬ Let R be a commutative local ring, and let M be a finitely generated R-module. Note that $M/\mathfrak{m}M$ is a finite-dimensional vector space over the field $R/\mathfrak{m}$; let $m_1, \ldots, m_r \in M$ be elements whose cosets mod $\mathfrak{m}M$ form a basis of $M/\mathfrak{m}M$. Prove that $m_1, \ldots, m_r$ generate M.

(Show that $\langle m_1, \ldots, m_r \rangle + \mathfrak{m}M = M$; then apply Nakayama's lemma in the form of Exercise 3.9.) [5.5, VIII.2.24]

3.11. Explain how to use Gaussian elimination to find bases for the row space and the column space of a matrix over a field.

3.12. ¬ Let R be an integral domain, and let $M \in \mathcal{M}_{m,n}(R)$, with $m < n$. Prove that the columns of M are linearly dependent over R. [5.6]

3.13. Let k be a field. Prove that a matrix $M \in \mathcal{M}_{m,n}(k)$ has rank $\leq r$ if and only if there exist matrices $P \in \mathcal{M}_{m,r}(k)$, $Q \in \mathcal{M}_{r,n}(k)$ such that $M = PQ$. (Thus the rank of M is the smallest such integer.)

3.14. Generalize Proposition 3.11 to the case of finitely generated free modules over any integral domain. (Embed the integral domain in its field of fractions.)

3.15. ▷ Prove Proposition 3.13 for the case $N = 1$. [§3.4]

3.16. ▷ Prove Claim 3.14. [§3.4]

3.17. Extend the definition of Grothendieck group of vector spaces given in §3.4 to the category of vector spaces of *countable* (possibly infinite) dimension, and prove that it is the trivial group.

3.18. Let Ab^{fg} be the category of finitely generated abelian groups. Define a Grothendieck group of this category in the style of the construction of $K(k\text{-}\mathsf{Vect}^f)$, and prove that $K(\mathsf{Ab}^{fg}) \cong \mathbb{Z}$.

3.19. ¬ Let Ab^f be the category of finite abelian groups. Prove that assigning to every finite abelian group its order extends to a homomorphism from the Grothendieck group $K(\mathsf{Ab}^f)$ to the multiplicative group $(\mathbb{Q}^*, \cdot)$. [3.20]

3.20. Let $R\text{-}\mathsf{Mod}^f$ be the category of modules of finite *length* (cf. Exercise 1.16) over a ring R. Let G be an abelian group, and let δ be a function assigning an element of G to every *simple* R-module. Prove that δ extends to a homomorphism from the Grothendieck group of $R\text{-}\mathsf{Mod}^f$ to G.

Explain why Exercise 3.19 is a particular case of this observation.

(For another example, letting $\delta(M) = 1 \in \mathbb{Z}$ for every simple module M shows that length itself extends to a homomorphism from the Grothendieck group of $R\text{-}\mathsf{Mod}^f$ to $\mathbb{Z}$.)

---

4. Presentations and resolutions

After this excursion into the idyllic world of free modules, we can come back to earth and see if we have learned something that may be useful for more general situations. Modules over a field are necessarily free (Proposition 1.7), not so for modules over more general rings. In fact, this property will turn out to be a characterization of fields (Proposition 4.10).

It is important that we develop some understanding of nonfree modules. In this section we will see that homomorphisms of free modules carry enough information to allow us to deal with many nonfree modules.

4.1. Torsion. There are several ways in which a module M may fail to be free: the most spectacular one is that M may have *torsion.*

Definition 4.1. Let M be an R-module. An element $m \in M$ is a *torsion* element if $\{m\}$ is linearly dependent, that is, if $\exists r \in R$, $r \neq 0$, such that $rm = 0$. The subset of torsion elements of M is denoted $\mathrm{Tor}_R(M)$. A module M is *torsion-free* if $\mathrm{Tor}_R(M) = \{0\}$. A *torsion* module is a module M in which every element is a torsion element. ⌟

The subscript R is usually omitted if there is no uncertainty about the base ring.

A commutative ring is torsion-free as a module over itself if and only if it is an integral domain; this is a good reason to limit the discussion to integral domains in this chapter. Also, the reader will check (Exercise 4.1) that *if R is an integral domain,* then $\mathrm{Tor}(M)$ is a *submodule* of M. Equally easy is the following observation:

Lemma 4.2. *Submodules and direct sums of torsion-free modules are torsion-free. Free modules over an integral domain are torsion-free.*

Proof. The first statement is immediate; the second follows from the first, since an integral domain is torsion-free as a module over itself. □

Lemma 4.2 gives a good source of torsion-free modules: for example, ideals in an integral domain R are torsion-free (because they are submodules of the free module R^1). In fact, ideals provide us with examples of another mechanism in which a module may fail to be free.

Example 4.3. Let $R = \mathbb{Z}[x]$, and let $I = (2, x)$. Then I is not a free R-module. More generally, let I be any nonprincipal ideal of an integral domain R; then I is a torsion-free module which is not free.

Indeed, if I were free, then its rank would have to be 1 at most, by Proposition 1.9 (a basis for I would be a linearly independent subset of R, and R has rank 1 over itself); thus one element would suffice to generate I, and I would be principal. ⌟

What is behind this example is a characterization of PIDs in terms of 'torsion-free submodules of a rank-1 free module' (Exercise 4.3). This is a facet of the main result towards which we are heading, that is, the classification of finitely generated modules over PIDs (Theorem 5.6). The gist of this classification is that finitely generated modules over a PID can be decomposed into *cyclic* modules. We have essentially proved this fact already for modules over Euclidean domains (it follows from Proposition 2.11; see Exercise 2.19), and we have looked in great detail at the particular case of $\mathbb{Z}$-modules, a.k.a. abelian groups (§IV.6); we are almost ready to deal with the general case of PIDs.

Definition 4.4. An R-module M is *cyclic* if it is generated by a singleton, that is, if $M \cong R/I$ for some ideal I of R. ⌟

The equivalence in the definition is hopefully clear to our reader, as an immediate consequence of the first isomorphism theorem for modules (Corollary III.5.16). If not, go back and (re)do Exercise III.6.16.

Cyclic modules are witness to the difference between fields and more general rings: over a field k, a cyclic module is just a 1-dimensional vector space, that is a 'copy of k'; over more general rings, cyclic modules may be very interesting (think of the many hours spent contemplating cyclic *groups*). In fact, we can tell that a ring is a field by just looking at its cyclic modules:

Lemma 4.5. *Let R be an integral domain. Assume that every cyclic R-module is torsion-free. Then R is a field.*

Proof. Let $c \in R$, $c \neq 0$; then $M = R/(c)$ is a cyclic module. Note that $\mathrm{Tor}(M) = M$: indeed, the class of 1 generates $R/(c)$ and belongs to $\mathrm{Tor}(M)$ since $c \cdot 1$ is 0 mod (c) and $c \neq 0$. However, by hypothesis M is torsion-free; that is, $\mathrm{Tor}(M) = \{0\}$. Therefore $M = \mathrm{Tor}(M)$ is the zero module.

This shows $R/(c)$ is the zero R-module; that is, $(c) = (1)$. Therefore, c is a unit. Thus every nonzero element of R is a unit, proving that R is a field. □

Lemma 4.5 is a simple-minded illustration of the fact that we can study a ring R by studying the module structure over R, that is, the category R-Mod, and that we may not even need to look at the whole of R-Mod to be able to draw strong conclusions about R.

4.2. Finitely presented modules and free resolutions. The 'right' way to think of a cyclic R-module M is as a module which admits an epimorphism from R,

viewing the latter as the free rank-1 R-module[21]:

$$R^1 \longrightarrow M \longrightarrow 0.$$

The fact that M is surjected upon by a free R-module is nothing special. In fact, *every* module M admits such an epimorphism:

$$R^{\oplus A} \longrightarrow M \longrightarrow 0$$

provided that we are willing to take A large enough; if we are desperate, $A = M$ will surely do. This is immediate from the universal property of free modules; if the reader does not agree, it is time to go back and review §III.6.3. What makes cyclic modules special is that A can be chosen to be a singleton.

We are now going to focus on a case which is also special, but not quite as special as cyclic modules: *finitely generated* modules are modules for which we can choose A to be a *finite* set (cf. §III.6.4). Thus, we will assume that M admits an epimorphism from a finite-rank free module:

$$R^m \xrightarrow{\pi} M \longrightarrow 0$$

for some integer m. The image by π of the m vectors in a basis of R^m is a set of generators for M.

Finitely generated modules are much easier to handle than arbitrary modules. For example, an ideal of R can tell us whether a finitely generated module is torsion.

Definition 4.6. The *annihilator* of an R-module M is

$$\mathrm{Ann}_R(M) := \{r \in R \,|\, \forall m \in M, rm = 0\}. \qquad \lrcorner$$

The subscript is usually omitted. The reader will check (Exercise 4.4) that $\mathrm{Ann}(M)$ is an *ideal* of R and that if M is a finitely generated module and R is an integral domain, then M is torsion if and only if $\mathrm{Ann}(M) \neq 0$.

We would like to develop tools to deal with finitely generated modules. It turns out that matrices allow us to describe a comfortably large collection of such modules.

Definition 4.7. An R-module M is *finitely presented* if for some positive integers m, n there is an exact sequence

$$R^n \xrightarrow{\varphi} R^m \longrightarrow M \longrightarrow 0.$$

Such a sequence is called a *presentation* of M. $\lrcorner$

In other words, finitely presented modules are cokernels (cf. III.6.2) of homomorphisms between finitely generated *free* modules. Everything about M must be encoded in the homomorphism φ; therefore, we should be able to describe the module M by studying the matrix corresponding to φ.

There is a gap between finitely presented modules and finitely generated modules, but on reasonable rings the two notions coincide:

[21] In context the exactness of a sequence of R-modules will be understood, so the displayed sequence is a way to denote the fact that there exists a surjective homomorphism of R-modules from R to M; cf. Example III.7.2. Also note the convention of denoting R by R^1 when it is viewed as a *module* over itself.

Lemma 4.8. *If R is a Noetherian ring, then* every *finitely generated R-module is finitely presented.*

Proof. If M is a finitely generated module, there is an exact sequence

$$R^m \xrightarrow{\pi} M \longrightarrow 0$$

for some m. Since R is Noetherian, R^m is Noetherian as an R-module (Corollary III.6.8). Thus $\ker \pi$ is finitely generated; that is, there is an exact sequence

$$R^n \longrightarrow \ker \pi \longrightarrow 0$$

for some n. Putting together the two sequences gives a presentation of M. □

Once we have gone one step to obtain generators and two steps to get a presentation, we should hit upon the idea to keep going:

Definition 4.9. A *resolution* of an R-module M by finitely generated free modules is an exact complex

$$\cdots \longrightarrow R^{m_3} \longrightarrow R^{m_2} \longrightarrow R^{m_1} \longrightarrow R^{m_0} \longrightarrow M \longrightarrow 0. \qquad \lrcorner$$

Iterating the argument proving Lemma 4.8 shows that if R is Noetherian, then every finitely generated module has a resolution as in Definition 4.9.

It is an important conceptual step to realize that M may be studied by studying an exact complex of free modules

$$\cdots \longrightarrow R^{m_3} \longrightarrow R^{m_2} \longrightarrow R^{m_1} \longrightarrow R^{m_0}$$

resolving M, that is, such that M is the cokernel of the last map. The R^{m_0} piece keeps track of the generators of M; R^{m_1} accounts for the relations among these generators; R^{m_2} records relations among the relations; and so on.

Developing this idea in full generality would take us too far for now: for example, we would have to deal with the fact that every module admits many different resolutions (for example, we can bump up every m_i by one by direct-summing each term in the complex with a copy of R^1, sent to itself by the maps in the complex). We will do this very carefully later on, in Chapter IX.

However, we can already learn something by considering coarse questions, such as 'how long' a resolution can be. *A priori,* there is no reason to expect a free resolutions to be 'finite', that is, such that $m_i = 0$ for $i \gg 0$. Such finiteness conditions tell us something special about the base ring R.

The first natural question of this type is, for which rings R is it the case that every finitely generated R-module M has a free resolution 'of length 0', that is, stopping at m_0? That would mean that there is an exact sequence

$$0 \longrightarrow R^{m_0} \longrightarrow M \longrightarrow 0.$$

Therefore, M itself must be free. What does this say about R?

Proposition 4.10. *Let R be an integral domain. Then R is a field if and only if every finitely generated R-module is free.*

Proof. If R is a field, then every R-module is free, by Proposition 1.7. For the converse, assume that every finitely generated R-module is free; in particular, every *cyclic* module is free; in particular, every cyclic module is *torsion-free.* But then R is a field, by Lemma 4.5. □

The next natural question concerns rings for which finitely generated modules admit free resolutions of length 1. It is convenient to phrase the question in stronger terms, that is, to require that for *every* finitely generated R-module M and *every* beginning of a free resolution

$$R^{m_0} \xrightarrow{\pi} M \longrightarrow 0,$$

the resolution can be completed to a length 1 free resolution. This would amount to demanding that there exist an integer m_1 and an R-module homomorphism $R^{m_1} \to R^{m_0}$ such that the sequence

$$0 \longrightarrow R^{m_1} \longrightarrow R^{m_0} \xrightarrow{\pi} M \longrightarrow 0$$

is exact. Equivalently, this condition requires that the module $\ker \pi$ of relations among the m_0 generators necessarily be free.

Claim 4.11. *Let R be an integral domain satisfying this property. Then R is a PID.*

Proof. Let I be an ideal of R, and apply the condition to $M = R/I$. Since we have an epimorphism

$$R^1 \xrightarrow{\pi} R/I \longrightarrow 0,$$

the condition says that $\ker \pi$ is free; that is, I is free. Since I is a free submodule of R, which is free of rank 1, I must be free of rank ≤ 1 by Proposition 1.9. Therefore I is generated by one element, as needed. □

The classification result for finitely generated modules over PIDs (Theorem 5.6), which we keep bringing up, will essentially be a converse to Claim 4.11: the mysterious condition requiring free resolutions of finitely generated modules to have length at most 1 turns out to be a *characterization* of PIDs, just as the length 0 condition is a characterization of fields (as proved in Proposition 4.10). We will work this out in §5.2.

4.3. Reading a presentation. Let us return to the brilliant idea of studying a finitely presented module M by studying a homomorphism of free modules

$$(*) \qquad \varphi : R^n \longrightarrow R^m$$

such that $M = \operatorname{coker} \varphi$. As we know, we can describe φ completely by considering a matrix A representing it, and therefore we can describe any finitely presented module by giving a matrix corresponding to (a homomorphism corresponding to) it.

In many cases, judicious use of the material developed in §2 allows us to determine the module M explicitly. For example, take

$$\begin{pmatrix} 1 & 3 \\ 2 & 3 \\ 5 & 9 \end{pmatrix};$$

this matrix corresponds to a homomorphism $\mathbb{Z}^2 \to \mathbb{Z}^3$, hence to a $\mathbb{Z}$-module, that is, a finitely generated abelian group G. The reader should figure out what G is more explicitly (in terms of the classification of §IV.6; cf. Exercise 2.19) before reading on. In the rest of this section we will simply tie up loose ends into a more concrete recipe to perform these operations.

Incidentally, a number of software packages can perform sophisticated operations on modules (say, over polynomial rings); a personal favorite is `Macaulay2`. These packages rely precisely on the correspondence between modules and matrices: with due care, every operation on modules (such as direct sums, tensors, quotients, etc.) can be executed on the corresponding matrices. For example,

Lemma 4.12. *Let A, B be matrices with entries in an integral domain R, and let M, N denote the corresponding R-modules. Then $M \oplus N$ corresponds to the block matrix*

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

Proof. This follows immediately from Exercise 4.16. □

Coming back to (*), note that the module M cannot know which bases we have chosen for R^n or R^m; that is, $M = \operatorname{coker} \varphi$ really depends on the *homomorphism* φ, not on the specific matrix representation we have chosen for φ. This is an issue that we have already encountered, and treated rather thoroughly, in §2.2 and following: 'equivalent' matrices represent the same homomorphism and hence the same module. In the context we are exploring now, Proposition 2.5 tells us that *two matrices A, B represent the same module M if there exist invertible matrices P, Q such that $B = PAQ$.*

But this is not the whole story. Two different homomorphisms φ_1, φ_2 may have isomorphic cokernels, even if they act between different modules: the extreme case being any isomorphism

$$R^m \longrightarrow R^m,$$

whose cokernel is 0 (regardless of the isomorphism and no matter what m is). Therefore, if a matrix A' corresponds to a module M, then (by Lemma 4.12) so does the block matrix

$$A = \left(\begin{array}{c|c} I_r & 0 \\ \hline 0 & A' \end{array}\right),$$

where I_r is the $r \times r$ identity matrix (and r is any nonnegative integer); in fact, I_r could be replaced here by any invertible matrix.

The following proposition attempts to formalize these observations.

Proposition 4.13. *Let A be a matrix with entries in an integral domain R, and let B be obtained from A by any sequence of the following operations:*

- *switch two rows or two columns;*
- *add to one row (resp., column) a multiple of another row (resp., column);*
- *multiply all entries in one row (or column) by a unit of R;*
- *if a unit is the only nonzero entry in a row (or column), remove the row and column containing that entry.*

Then B represents the same R-module as A, up to isomorphism.

Proof. The first three operations are the 'elementary operations' of §2.3, and they transform a matrix into an equivalent one (by Proposition 2.7); as observed above, this does not affect the corresponding module, up to isomorphism.

As for the fourth operation, if u is a unit and the only nonzero entry in (say) a row, then by applications of the second elementary operation we may assume that u is also the only nonzero entry in its column; without loss of generality we may assume that u is in fact the (1,1) entry of the matrix; that is, the matrix is in block form:

$$A = \left(\begin{array}{c|c} u & 0 \\ \hline 0 & A' \end{array}\right).$$

But then A and A' represent the same module, as needed. □

Example 4.14. The matrix with integer entries

$$\begin{pmatrix} 1 & 3 \\ 2 & 3 \\ 5 & 9 \end{pmatrix}$$

determines an abelian group G. Subtract three times the first column from the second column, obtaining

$$\begin{pmatrix} 1 & 0 \\ 2 & -3 \\ 5 & -6 \end{pmatrix};$$

the $(1,1)$ entry is a unit and the only nonzero entry in the first row, so we can remove the first row and column:

$$\begin{pmatrix} -3 \\ -6 \end{pmatrix};$$

now change the sign, and subtract twice the first row from the second, leaving

$$\begin{pmatrix} 3 \\ 0 \end{pmatrix}.$$

Therefore G is isomorphic to the cokernel of the homomorphism

$$\varphi : \mathbb{Z} \longrightarrow \mathbb{Z} \oplus \mathbb{Z}$$

mapping 1 to $(3,0)$. This homomorphism is injective and identifies $\mathbb{Z}$ with the subgroup $3\mathbb{Z} \oplus 0$ of the target. Therefore

$$G \cong \operatorname{coker} \varphi \cong \frac{\mathbb{Z} \oplus \mathbb{Z}}{3\mathbb{Z} \oplus 0} \cong \frac{\mathbb{Z}}{3\mathbb{Z}} \oplus \mathbb{Z}.$$

⌟

By virtue of Gaussian elimination, the 'algorithm' implicitly described in Proposition 4.13 will work without fail over Euclidean domains (e.g., over the polynomial ring in one variable over a field), in the sense that it will identify the finitely generated module corresponding to a matrix with an explicit direct sum of cyclic modules, as in Example 4.14. This is too much to expect over more general rings, since in general elementary transformations do not generate GL; cf. Remark 2.12.

Exercises

4.1. ▷ Prove that if R is an integral domain and M is an R-module, then $\mathrm{Tor}(M)$ is a submodule of M. Give an example showing that the hypothesis that R is an integral domain is necessary. [§4.1]

4.2. ▷ Let M be a module over an integral domain R, and let N be a torsion-free module. Prove that $\mathrm{Hom}_R(M,N)$ is torsion-free. In particular, $\mathrm{Hom}_R(M,R)$ is torsion-free. (We will run into this fact again; see Proposition VIII.5.16.) [§VIII.5.5]

4.3. ▷ Prove that an integral domain R is a PID if and only if every submodule of R itself is free. [§4.1, 5.13]

4.4. ▷ Let R be a commutative ring and M an R-module.

- Prove that $\mathrm{Ann}(M)$ is an ideal of R.
- If R is an integral domain and M is finitely generated, prove that M is torsion if and only if $\mathrm{Ann}(M) \neq 0$.
- Give an example of a torsion module M over an integral domain, such that $\mathrm{Ann}(M) = 0$. (Of course this example cannot be finitely generated!)

[§4.2, §5.3]

4.5. ¬ Let M be a module over a commutative ring R. Prove that an ideal I of R is the annihilator of an element of M if and only if M contains an isomorphic copy of R/I (viewed as an R-module).

The *associated primes* of M are the prime ideals among the ideals $\mathrm{Ann}(m)$, for $m \in M$. The set of the associated primes of a module M is denoted $\mathrm{Ass}_R(M)$. Note that every prime in $\mathrm{Ass}_R(M)$ contains $\mathrm{Ann}_R(M)$. [4.6, 4.7, 5.16]

4.6. ¬ Let M be a module over a commutative ring R, and consider the family of ideals $\mathrm{Ann}(m)$, as m ranges over the nonzero elements of M. Prove that the maximal elements in this family are prime ideals of R. Conclude that if R is Noetherian, then $\mathrm{Ass}_R(M) \neq \emptyset$ (cf. Exercise 4.5). [4.7, 4.9]

4.7. ¬ Let R be a commutative Noetherian ring, and let M be a finitely generated module over R. Prove that M admits a finite series

$$M = M_0 \supsetneq M_1 \supsetneq \cdots \supsetneq M_m = \langle 0 \rangle$$

in which all quotients M_i/M_{i+1} are of the form $R/\mathfrak{p}$ for some prime ideal $\mathfrak{p}$ of R. (Hint: Use Exercises 4.5 and 4.6 to show that M contains an isomorphic copy M'

of $R/\mathfrak{p}_1$ for some prime $\mathfrak{p}_1$. Then do the same with M/M', producing an $M'' \supseteq M'$ such that $M''/M' \cong R/\mathfrak{p}_2$ for some prime $\mathfrak{p}_2$. Why must this process stop after finitely many steps?) [4.8]

4.8. Let R be a commutative Noetherian ring, and let M be a finitely generated module over R. Prove that every prime in $\mathrm{Ass}_R(M)$ appears in the list of primes produced by the procedure presented in Exercise 4.7. (If $\mathfrak{p}$ is an associated prime, then M contains an isomorphic copy N of $R/\mathfrak{p}$. With notation as in the hint in Exercise 4.7, prove that either $\mathfrak{p}_1 = \mathfrak{p}$ or $N \cap M' = 0$. In the latter case, N maps isomorphically to a copy of $R/\mathfrak{p}$ in M/M'; iterate the reasoning.)

In particular, if M is a finitely generated module over a Noetherian ring, then $\mathrm{Ass}(M)$ is *finite.*

4.9. Let M be a module over a commutative Noetherian ring R. Prove that the union of all annihilators of nonzero elements of M equals the union of all associated primes of M. (Use Exercise 4.6.)

Deduce that the *union* of the associated primes of a Noetherian ring R (viewed as a module over itself) equals the set of zero-divisors of R.

4.10. Let R be a commutative Noetherian ring. One can prove that the minimal primes of $\mathrm{Ann}(M)$ (cf. Exercise V.1.9) are in $\mathrm{Ass}(M)$. Assuming this, prove that the *intersection* of the associated primes of a Noetherian ring R (viewed as a module over itself) equals the nilradical of R.

4.11. Review the notion of presentation *of a group,* (§II.8.2), and relate it to the notion of presentation introduced in §4.2.

4.12. Let $\mathfrak{p}$ be a prime ideal of a polynomial ring $k[x_1, \ldots, x_n]$ over a field k, and let $R = k[x_1, \ldots, x_n]/\mathfrak{p}$. Prove that every finitely generated module over R has a finite presentation.

4.13. ¬ Let R be a commutative ring. A tuple $(a_1, a_2, \ldots, a_n)$ of elements of R is a *regular sequence* if a_1 is a non-zero-divisor in R, a_2 is a non-zero-divisor modulo[22] (a_1), a_3 is a non-zero-divisor modulo (a_1, a_2), and so on.

For a, b in R, consider the following complex of R-modules:

$$(*) \qquad 0 \longrightarrow R \xrightarrow{d_2} R \oplus R \xrightarrow{d_1} R \xrightarrow{\pi} \frac{R}{(a,b)} \longrightarrow 0$$

where π is the canonical projection, $d_1(r, s) = ra + sb$, and $d_2(t) = (bt, -at)$. Put otherwise, d_1 and d_2 correspond, respectively, to the matrices

$$\begin{pmatrix} a & b \end{pmatrix}, \quad \begin{pmatrix} b \\ -a \end{pmatrix}.$$

- Prove that this is indeed a complex, for every a and b.
- Prove that if (a, b) is a regular sequence, this complex is *exact.*

The complex (*) is called the *Koszul complex* of (a, b). Thus, when (a, b) is a regular sequence, the Koszul complex provides us with a free resolution of the module $R/(a, b)$. [4.14, 5.4, VIII.4.22]

[22] That is, the class of a_2 in $R/(a_1)$ is a non-zero-divisor in $R/(a_1)$.

4.14. ¬ A Koszul complex may be defined for any sequence $a_1, \dots, a_n$ of elements of a commutative ring R. The case $n = 2$ seen in Exercise 4.13 and the case $n = 3$ reviewed here will hopefully suffice to get a gist of the general construction; the general case will be given in Exercise VIII.4.22.

Let $a, b, c \in R$. Consider the following complex:

$$0 \longrightarrow R \xrightarrow{d_3} R \oplus R \oplus R \xrightarrow{d_2} R \oplus R \oplus R \xrightarrow{d_1} R \xrightarrow{\pi} \frac{R}{(a,b,c)} \longrightarrow 0$$

where π is the canonical projection and the matrices for d_1, d_2, d_3 are, respectively,

$$\begin{pmatrix} a & b & c \end{pmatrix}, \quad \begin{pmatrix} 0 & -c & -b \\ -c & 0 & a \\ b & a & 0 \end{pmatrix}, \quad \begin{pmatrix} a \\ -b \\ c \end{pmatrix}.$$

- Prove that this is indeed a complex, for every a, b, c.
- Prove that if (a, b, c) is a regular sequence, this complex is *exact*.

Koszul complexes are very important in commutative algebra and algebraic geometry. [VIII.4.22]

4.15. ▷ View $\mathbb{Z}$ as a module over the ring $R = \mathbb{Z}[x, y]$, where x and y act by 0. Find a free resolution of $\mathbb{Z}$ over R.

4.16. ▷ Let $\varphi : R^n \to R^m$ and $\psi : R^p \to R^q$ be two R-module homomorphisms, and let

$$\varphi \oplus \psi : R^n \oplus R^p \to R^m \oplus R^q$$

be the morphism induced on direct sums. Prove that

$$\operatorname{coker}(\varphi \oplus \psi) = \operatorname{coker} \varphi \oplus \operatorname{coker} \psi.$$

[VIII.4.21]

4.17. Determine (as a better known entity) the module represented by the matrix

$$\begin{pmatrix} 1+3x & 2x & 3x \\ 1+2x & 1+2x-x^2 & 2x \\ x & x^2 & x \end{pmatrix}$$

over the polynomial ring $k[x]$ over a field.

---

5. Classification of finitely generated modules over PIDs

It is finally time to prove the classification theorem for finitely generated modules over arbitrary PIDs. We have already proved this statement in the special case of finite $\mathbb{Z}$-modules (§IV.6), and the diligent reader has worked out a proof in the less special case of finitely generated modules over Euclidean domains, in Exercise 2.19. Now we go for the real thing.

5.1. Submodules of free modules. Recall (Lemma 4.2, Example 4.3) that a submodule of a free module over an arbitrary integral domain R is necessarily torsion-free but need not be free. For example, the ideal $I = (x, y)$ of $R = k[x, y]$ (with k a field, for example) is torsion-free as an R-module, but not free: for this to become really, really evident, it is helpful to rename $x = a$, $y = b$ when these are viewed as elements of I and observe that a and b are not linearly independent over $k[x, y]$, since $ya - xb = 0$.

On the other hand, submodules of a free module over a field are automatically free: simply because *every* module over a field is free (Proposition 1.7). It is reasonable to expect that 'some property in between' being a field and being a friendly UFD such as $k[x, y]$ will guarantee that a submodule of a free module is free. We will now prove that this property is precisely that of being a *principal ideal domain.*

Proposition 5.1. *Let R be a PID, let F be a finitely generated free module over R, and let $M \subseteq F$ be a submodule. Then M is free.*

We will actually prove a more precise result, in view of the full statement of the classification theorem: we will show that *there is a basis $(x_1, \dots, x_n)$ of F and elements $a_1, \dots, a_m$ of R (with $m \leq n$) such that*

$$y_1 = a_1 x_1, \dots, \quad y_m = a_m x_m$$

form a basis of M. That is, not only do we prove that M is free, but we also show that there are 'compatible' bases of F and M.

In order to do this, we may of course assume that $M \neq 0$: otherwise there is nothing to prove. Most of the work will then go into showing that if $M \neq 0$, we can split one direct summand off M; iterating this process will prove the proposition[23]. This is where the PID condition is used, so we will single out the main technical point into the following statement.

Lemma 5.2. *Let R be a PID, let F be a finitely generated free module over R, and let $M \subseteq F$ be a nonzero submodule. Then there exist $a \in R$, $x \in F$, $y \in M$, and submodules $F' \subseteq F$ and $M' \subseteq M$, such that $y = ax$, $M' = F' \cap M$, and*

$$F = \langle x \rangle \oplus F', \quad M = \langle y \rangle \oplus M'.$$

The reader would find it instructive to pause a moment and try to imagine how the PID hypothesis may enter into the proof of this lemma. The PID hypothesis is a hypothesis on R, so the question is, where is the special copy of R to which we can apply it? Don't read on until you have spent a moment thinking about this.

It is tempting to look for this copy of R among the 'factors' of $F = R^n$: for example, we could map F to R by projecting onto the first component:

$$\pi(r_1, \dots, r_n) = r_1.$$

But there is nothing special about the first component of R^n; in fact there is nothing special about the particular basis chosen for F, that is, the particular

[23] In a loose sense, this is the same strategy we employed in the proof of the classification result for finite abelian groups in §IV.6.

representation of F as R^n. Therefore, this does not look too promising. The way out of this bind is to democratically map F to R in *every* possible way: we consider *all* homomorphisms $\varphi : F \to R$. This is a set that does not depend on any extra choice, so it is more likely to carry the information we need.

For each φ, $\varphi(M)$ is a submodule of R, that is, an ideal of R; so we have a chance to use the PID hypothesis with profit. In fact, the much weaker fact that R is *Noetherian* guarantees already that there must be some homomorphism α for which $\alpha(M)$ is maximal among all ideals $\varphi(M)$. *This* copy of R, that is, the target of such a homomorphism α, is special for a good reason, independent of inessential choices. The fact that R is a PID tells us that $\alpha(M)$ is principal, and then we are clearly in business.

Here is the formal argument:

Proof. For all $\varphi \in \mathrm{Hom}_R(F, R)$, $\varphi(M)$ is a submodule of R, that is, an ideal. The family of all these ideals is nonempty, and PIDs are Noetherian; therefore (by Proposition V.1.1) there exists a maximal element in the family, say $\alpha(M)$, for a homomorphism $\alpha : M \to R$. The fact that $M \neq 0$ implies immediately that some $\varphi(M) \neq 0$ (for example, take for φ the projection to a suitable factor of R^n); hence $\alpha(M) \neq 0$.

Since R is a PID, $\alpha(M)$ is principal: $\alpha(M) = (a)$ for some $a \in R, a \neq 0$. Since $a \in \alpha(M)$, there exists an element $y \in M$ such that $\alpha(y) = a$. These are the elements a, y mentioned in the statement.

We claim that a divides $\varphi(y)$ for all $\varphi \in \mathrm{Hom}_R(F, R)$. Indeed, let b be a generator of $(a, \varphi(y))$ (which exists since R is a PID; of course b is simply a gcd of a and $\varphi(y)$), and let $r, s \in R$ such that $b = ra + s\varphi(y)$; consider the homomorphism $\psi := r\alpha + s\varphi$. Since $a \in (b)$, we have $\alpha(M) \subseteq (b)$. On the other hand

$$b = ra + s\varphi(y) = (r\alpha + s\varphi)(y) = \psi(y) \in \psi(M);$$

therefore $(b) \subseteq \psi(M)$. It follows that $\alpha(M) \subseteq \psi(M)$, and by maximality $\alpha(M) = \psi(M)$; hence $(a) = (b)$, and in particular a divides $\varphi(y)$, as claimed.

Let $y = (s_1, \ldots, s_n)$ as an element of $F = R^n$. Each s_i is the image of y by a homomorphism $F \to R$ (that is, the i-th projection), so a divides all of them by what we just proved. Therefore $\exists r_1, \ldots, r_n \in R$ such that $s_i = ar_i$; let

$$x = (r_1, \ldots, r_n) \in F.$$

This is the element x mentioned in the statement. By construction, $y = ax$. Further, $a = \alpha(y) = \alpha(ax) = a\alpha(x)$; since R is an integral domain and $a \neq 0$, this implies $\alpha(x) = 1$.

Finally, we let $F' = \ker \alpha$ and $M' = F' \cap M$, and we can proceed to verify the stated direct sums.

First, every $z \in F$ may be written as

$$z = \alpha(z)x + (z - \alpha(z)x);$$

by linearity

$$\alpha(z - \alpha(z)x) = \alpha(z) - \alpha(z)\alpha(x) = \alpha(z) - \alpha(z) = 0,$$

that is, $z - \alpha(z)x \in \ker\alpha$. This implies that $F = \langle x\rangle + F'$. On the other hand, $rx \in F' \implies \alpha(rx) = 0 \implies r\alpha(x) = 0 \implies r = 0$: that is, $\langle x\rangle \cap F' = 0$. Therefore

$$F = \langle x\rangle \oplus F',$$

as claimed (cf. Exercise 5.1).

Second, if $z \in M$, then a divides $\alpha(z)$: indeed, $\alpha(z) \in \alpha(M) = (a)$. Writing $\alpha(z) = ca$, we have $\alpha(z)x = cax = cy$; splitting z as above, we note

$$z - \alpha(z)x = z - cy \in M \cap F' = M',$$

and this leads as before to

$$M = \langle y\rangle \oplus M',$$

concluding the proof. □

Once Lemma 5.2 is established, the proof of Proposition 5.1 is mere busywork:

Proof of Proposition 5.1. If $M = 0$, we are done. If not, applying Lemma 5.2 to $M \subseteq F$ produces an element $y_1 \in M$ and a submodule $M^{(1)} \subseteq M$ such that

$$M = \langle y_1\rangle \oplus M^{(1)}.$$

If $M^{(1)} = 0$, we are done; otherwise, apply Lemma 5.2 again (to $M^{(1)} \subseteq F$) to obtain $y_2 \in M^{(1)}$ and $M^{(2)} \subseteq M^{(1)}$ so that

$$M = \langle y_1\rangle \oplus (\langle y_2\rangle \oplus M^{(2)}).$$

This process may be continued, producing elements $y_1, \dots, y_m \in M$ such that

$$M = \langle y_1\rangle \oplus \cdots \oplus \langle y_m\rangle \oplus M^{(m)},$$

so long as the module $M^{(m)}$ is nonzero. However, by Proposition 1.9 we know $m \leq n$, since $y_1, \dots, y_m$ are linearly independent in F. It follows that the process must stop; that is, $M^{(m)} = 0$ for some $m \leq n$. That is,

$$M = \langle y_1\rangle \oplus \cdots \oplus \langle y_m\rangle$$

is free, as needed. □

Note that the proof has not used part of the result of Lemma 5.2, that is, the fact that the 'factor' $\langle y\rangle$ of M is a submodule of a corresponding factor $\langle x\rangle$ of F. This is needed in order to upgrade Proposition 5.1 along the lines mentioned after its statement. Here is that stronger statement:

Corollary 5.3. *Let R be a PID, let F be a finitely generated free module over R, and let $M \subseteq F$ be a submodule. Then there exist a basis $(x_1, \dots, x_n)$ of F and nonzero elements $a_1, \dots, a_m$ of R ($m \leq n$) such that $(a_1x_1, \dots, a_mx_m)$ is a basis of M. Further, we may assume $a_1 \mid a_2 \mid \cdots \mid a_m$.*

This statement should be compared with Proposition 2.11: it amounts to a 'Smith normal form over PIDs'; cf. Remark 2.12.

Proof. Now that we know that submodules of a free module are free, we see that the submodule $F' \subseteq F$ produced in Lemma 5.2 is free. The first part of the statement then follows from Lemma 5.2, by an inductive argument analogous to the proof of Proposition 5.1, and is left to the reader.

The most delicate part of the statement is the divisibility condition. By induction it suffices to prove $a_1 \mid a_2$, and for this we refer back to the proof of Lemma 5.2: (a_1) is maximal among the ideals $\varphi(M)$ of R, as φ ranges over all homomorphisms $F \to R$. Consider then any[24] homomorphism φ such that $\varphi(x_1) = \varphi(x_2) = 1$. Since $\varphi(y_1) = \varphi(a_1 x_1) = a_1$, we have $(a_1) \subseteq \varphi(M)$; by maximality $(a_1) = \varphi(M)$. Therefore $a_2 = \varphi(a_2 x_2) = \varphi(y_2) \in \varphi(M) = (a_1)$, proving $a_1 \mid a_2$ as needed. □

The logic of the argument given in this section is somewhat tricky: it takes Lemma 5.2 to prove Proposition 5.1; but then it takes Proposition 5.1 (proving that F' is free) to revisit Lemma 5.2 and squeeze from it the stronger Corollary 5.3.

5.2. PIDs and resolutions. Proposition 5.1 allows us to complete the circle of ideas begun in §4.2.

Proposition 5.4. *Let R be an integral domain. Then R is a PID if and only if for every finitely generated R-module M and every epimorphism*

$$R^{m_0} \xrightarrow{\pi_0} M \longrightarrow 0,$$

there exist a free R-module R^{m_1} and a homomorphism $\pi_1 : R^{m_1} \to R^{m_0}$ such that the sequence

$$0 \longrightarrow R^{m_1} \xrightarrow{\pi_1} R^{m_0} \xrightarrow{\pi_0} M \longrightarrow 0$$

is exact.

Proof. The fact that the stated condition implies that R is a PID was proved in Claim 4.11. For the converse, let $\pi_0 : R^{m_0} \to M$ be an epimorphism; then $\ker \pi_0$ is free by Proposition 5.1; the result follows by choosing any isomorphism $\pi_1 : R^{m_1} \to \ker(\pi_0)$. □

The content of Proposition 5.4 is possibly simpler than its awkward formulation: Proposition 5.4 is simply a characterization for PIDs analogous to the characterization of fields worked out in Proposition 4.10. This is another example of the fact that we can study a ring R by looking at how R-Mod is put together.

The reader can well imagine the 'length-n' version of the condition appearing in these results: we may ask for the class of integral domains R with the property that for all finitely generated modules M and all partial free resolutions of M,

$$R^{m_{n-1}} \xrightarrow{\pi_{n-1}} \cdots \xrightarrow{\pi_1} R^{m_0} \xrightarrow{\pi_0} M \longrightarrow 0\ ,$$

[24] In order to define a homomorphism on a free module F, one may define it on a basis of F and then extend it by linearity; there is no restriction on the possible choices for the images of basis elements. The reader who is not sure about this should review the universal property of free modules!

there exists a homomorphism $\pi_n : R^{m_n} \to R^{m_{n-1}}$ such that

$$0 \longrightarrow R^{m_n} \xrightarrow{\pi_n} R^{m_{n-1}} \xrightarrow{\pi_{n-1}} \cdots \xrightarrow{\pi_1} R^{m_0} \xrightarrow{\pi_0} M \longrightarrow 0$$

is exact. We have proved that this condition characterizes fields for $n = 0$ and PIDs for $n = 1$.

The alert reader should now have a *déjà vu,* as we have already encountered a notion that is 0 precisely for fields and 1 for PIDs: the *Krull dimension* of a ring; cf. Example III.4.14. Of course this is not a coincidence, but the full relation between the Krull dimension and the bound on the length of free resolutions is beyond the scope of this book[25]. 'Most' rings (even rings with finite Krull dimension) do not satisfy any such bound for all modules. The local rings satisfying the finiteness condition described above for some n correspond in the language of algebraic geometry to 'smooth' points on varieties of dimension $\leq n$.

5.3. The classification theorem. The notion of the *rank* of a free module (Definition 1.14) extends naturally to every finitely generated module M over an integral domain R.

Definition 5.5. Let R be an integral domain. The *rank* $\operatorname{rk} M$ of a finitely generated R-module M is the maximum number of linearly independent elements in M. ⌟

It should be clear that this number is finite (Exercise 5.6).

Theorem 5.6. *Let R be a PID, and let M be a finitely generated R-module. Then the following hold:*

- *There exist distinct prime ideals $(q_1), \ldots, (q_n) \in R$, positive integers r_{ij}, and an isomorphism*
$$M \cong R^{\operatorname{rk} M} \oplus \left(\bigoplus_{i,j} \frac{R}{(q_i^{r_{ij}})} \right).$$
- *There exist nonzero, nonunit ideals $(a_1), \ldots, (a_m)$ of R, such that $(a_1) \supseteq (a_2) \supseteq \cdots \supseteq (a_m)$, and an isomorphism*
$$M \cong R^{\operatorname{rk} M} \oplus \frac{R}{(a_1)} \oplus \cdots \oplus \frac{R}{(a_m)}.$$

These decompositions are unique (in the evident sense).

After all our preparatory work, this statement practically proves itself. We will describe the arguments in general terms, leaving the details to the enterprising reader.

The two forms taken by the theorem go under the name of *invariant factors* and *elementary divisors,* as in the case of abelian groups. As in the case of abelian groups, the equivalence between the two formulations amounts to careful bookkeeping, and we will not prove it formally; see §IV.6.2 for a reminder of how to go from

[25] The most persistent of readers will take a more sophisticated look at these bounds in the exercises to §IX.8.

one to the other. The role played by Lemma IV.6.1 in that discussion is taken over by the Chinese remainder theorem, Theorem V.6.1, in this more general setting.

As the two formulations are equivalent, it suffices to prove the existence and uniqueness of the decompositions given in Theorem 5.6 for any one of the two.

The existence is a direct consequence of the results in §5.1. Indeed, let M be a finitely generated module; thus there is an epimorphism

$$R^n \xrightarrow{\pi} M \longrightarrow 0$$

where n is the number of generators of M. Apply Corollary 5.3 to the submodule $\ker \pi \subseteq R^n$: there exist a basis $(x_1, \dots, x_n)$ of R^n and nonzero elements $a_1, \dots, a_m$ of R such that $(a_1 x_1, \dots, a_m x_m)$ is a basis of $\ker \pi$ and further $a_1 \mid \cdots \mid a_m$.

That is, M is presented by the $n \times m$ matrix

$$\begin{pmatrix} a_1 & 0 \\ \vdots & \vdots \\ 0 & a_m \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix}.$$

By Proposition 4.13, we may assume that $a_1, \dots, a_m$ are not units: if any one of them is, the corresponding row and column may be omitted from the matrix (and n corrected accordingly). It follows that

$$M \cong \frac{R}{(a_1)} \oplus \cdots \oplus \frac{R}{(a_m)} \oplus R^{(n-m)}$$

with $a_1 \mid \cdots \mid a_m$ nonzero nonunits, as prescribed by Theorem 5.6, and the existence is proved.

As for the uniqueness of the representations, if

$$M \cong R^r \oplus T,$$

with T a torsion module[26], then $r = \operatorname{rk} M$ and $T \cong \operatorname{Tor}_R(M)$ (Exercise 5.10). Thus, the 'free part' and the 'torsion part' in the decompositions given in Theorem 5.6 are determined uniquely.

This reduces the uniqueness question to the structure of torsion modules: assuming that

$$(*) \qquad \bigoplus_{i,j} \frac{R}{(p_i^{r_{ij}})} \cong \bigoplus_{k,\ell} \frac{R}{(q_k{}^{s_{k\ell}})},$$

with p_i, q_k irreducible elements of R, the task is to show that the range of the indices is the same and, up to reordering, $(p_i) = (q_i)$ and $r_{ij} = s_{ij}$ for all i, j.

Since a finitely generated module is torsion if and only if its annihilator is nonzero (Exercise 4.4), it is reasonable that $\operatorname{Ann}(M)$ will be of some use here.

[26] Recall that this means that every element of T is a torsion element; cf. §4.1.

Lemma 5.7. *Let M be a torsion module, expressed as in Theorem 5.6 (with* $\operatorname{rk} M = 0$*). Then* $\operatorname{Ann}(M) = (a_m)$*. Further, the prime ideals (q_i) are precisely the prime ideals of R containing* $\operatorname{Ann}(M)$*.*

Proof. By hypothesis

$$M \cong \frac{R}{(a_1)} \oplus \cdots \oplus \frac{R}{(a_m)},$$

with $a_1 \mid \cdots \mid a_m$. If $r \in \operatorname{Ann}(M)$, then

$$0 = r(1, \ldots, 1) = (r, \ldots, r).$$

In particular $r \equiv 0$ modulo (a_m); that is, $r \in (a_m)$. Thus $\operatorname{Ann}(M) \subseteq (a_m)$. For the reverse inclusion, assume $r \in (a_m)$ and $y \in M$. Identifying M with its decomposition, write $y = (y_1, \ldots, y_m)$, with $y_i \in R/(a_i)$. Since $r \in (a_m) \subseteq (a_i)$, we have $ry_i = 0$ for all i; therefore $ry = 0$, and $r \in \operatorname{Ann}(M)$ as needed.

For the second part of the statement, tracing the equivalence between the two decompositions in Theorem 5.6 shows that the q_i's are precisely the irreducible factors of a_m, so the assertion follows from the first part. □

By Lemma 5.7, the sets $\{p_i\}$, $\{q_k\}$ of irreducibles appearing in (*) must coincide (up to inessential units): the ideals they generate are precisely the prime ideals containing $\operatorname{Ann}(M)$. The fact that the sets coincide also follows from the observation that the isomorphism in (*) must match like primes, in the sense that an element of a factor

$$\frac{R}{(p_i^{r_{ij}})}$$

in the source must land in a combination of quotients in the target by powers of the same prime p_i; this follows by comparing annihilators (cf. Exercise 5.11). Therefore, uniqueness is reduced to the case of a single irreducible $q \in R$: it suffices to show that if

$$\frac{R}{(q^{r_1})} \oplus \cdots \oplus \frac{R}{(q^{r_m})} \cong \frac{\mathbb{Z}}{(q^{s_1})} \oplus \cdots \oplus \frac{\mathbb{Z}}{(q^{s_n})},$$

with $r_1 \geq \cdots \geq r_m$ and $s_1 \geq \cdots \geq s_n$, then $m = n$ and $r_i = s_i$ for all i. This situation reproduces precisely the key step the reader used in the proof of uniqueness for the particular case of abelian groups, in Exercise IV.6.1. Of course we will not spoil the reader's fun by giving any more details this time (Exercise 5.12).

Remark 5.8. To summarize, the information carried by a torsion module over a PID is equivalent to the choice of a selection of nonzero, nonunit ideals $(a_1), \ldots, (a_m)$ such that

$$(a_1) \supseteq (a_2) \supseteq \cdots \supseteq (a_m).$$

We have seen that (a_m) is in fact the annihilator ideal of the module; it would be nice if we had a similarly explicit description of all invariants $(a_1), \ldots, (a_m)$. As a preview of coming attractions, we could call the ideal

$$(a_1 \cdots a_m)$$

the *characteristic ideal* of the module[27]. In situations in which we can compute both the annihilator and the characteristic ideal, comparing them may lead to

[27] Warning: This does not seem to be standard terminology.

strong conclusions about the module. For example, a torsion module is cyclic if and only if its annihilator and characteristic ideals coincide.

Further, the prime ideals appearing in Theorem 5.6 have been characterized as the prime ideals containing the annihilator ideal. They may just as well be characterized as those prime ideals containing the characteristic ideal, as the reader will check (Exercise 5.15).

Finally, note that from this point of view it is immediate that *the characteristic ideal is contained in the annihilator ideal.* We will run into this fact again, in an important application (Theorem 6.11). ⌟

Exercises

5.1. ▷ Let N, P be submodules of a module M, such that $N \cap P = \{0\}$ and $M = N + P$. Prove that $M \cong N \oplus P$. (This is a word-for-word repetition of Proposition IV.5.3 for modules.) [§5.1]

5.2. Let R be an integral domain, and let M be a finitely generated R-module. Prove that M is torsion if and only if $\operatorname{rk} M = 0$.

5.3. Complete the proof of Corollary 5.3.

5.4. Let R be an integral domain, and assume that $a, b \in R$ are such that $a \neq 0$, $b \notin (a)$, and $R/(a)$, $R/(a,b)$ are both integral domains.

- Prove that the Krull dimension of R is at least 2.
- Prove that if R satisfies the finiteness condition discussed in §5.2 for some n, then $n \geq 2$.

You can prove this second point by appealing to Proposition 5.4. For a more concrete argument, you should look for an R-module admitting a free resolution of length 2 which cannot be shortened.

- Prove that (a, b) is a regular sequence in R (Exercise 4.13).
- Prove that the R-module $R/(a,b)$ has a free resolution of length exactly 2.

Can you see how to construct analogous situations with $n \geq 3$ elements $a_1, \ldots, a_n$?

5.5. ¬ Recall (Exercise V.4.11) that a commutative ring is *local* if it has a single maximal ideal $\mathfrak{m}$. Let R be a local ring, and let M be a *direct summand* of a finitely generated free R-module: that is, there exists an R-module N such that $M \oplus N$ is a free R-module.

- Choose elements $m_1, \ldots, m_r \in M$ whose cosets mod $\mathfrak{m}M$ are a basis of $M/\mathfrak{m}M$ as a vector space over the field $R/\mathfrak{m}$. By Nakayama's lemma, $M = \langle m_1, \ldots, m_r \rangle$ (Exercise 3.10).
- Obtain a surjective homomorphism $\pi : F = R^{\oplus r} \to M$.

- Show that π splits, giving an isomorphism $F \cong M \oplus \ker \pi$. (Apply Exercise III.6.9 to the surjective homomorphism π and the free module $M \oplus N$ to obtain a splitting $M \to F$; then use Proposition III.7.5.)
- Show $\ker \pi / \mathfrak{m} \ker \pi = 0$. Use Nakayama's lemma (Exercise 3.8) to deduce that $\ker \pi = 0$.
- Conclude that $M \cong F$ is in fact free. [VIII.2.24, VIII.6.8, VIII.6.11]

Summarizing, over a *local ring,* every *direct summand* of a finitely generated[28] free R-module is free. Using the terminology we will introduce in Chapter VIII, we would say that 'projective modules over local rings are free'. This result has strong implications in algebraic geometry, since it underlies the notion of vector bundle.

Contrast this fact with Proposition 5.1, which shows that, over a *PID, every* submodule of a finitely generated free module is free.

5.6. ▷ Let R be an integral domain, and let $M = \langle m_1, \ldots, m_r \rangle$ be a finitely generated module. Prove that $\operatorname{rk} M \leq r$. (Use Exercise 3.12.) [§5.3]

5.7. Let R be an integral domain, and let M be a finitely generated module over R. Prove that $\operatorname{rk} M = \operatorname{rk}(M/\operatorname{Tor}(M))$.

5.8. Let R be an integral domain, and let M be a finitely generated module over R. Prove that $\operatorname{rk} M = r$ if and only if M has a *free* submodule $N \cong R^r$, such that M/N is torsion.

If R is a PID, then N may be chosen so that $0 \to N \to M \to N/M \to 0$ splits.

5.9. Let R be an integral domain, and let

$$0 \longrightarrow M_1 \longrightarrow M_2 \longrightarrow M_3 \longrightarrow 0$$

be an exact sequence of finitely generated R-modules. Prove that $\operatorname{rk} M_2 = \operatorname{rk} M_1 + \operatorname{rk} M_3$.

Deduce that 'rank' defines a homomorphism from the Grothendieck group of the category of finitely generated R-modules to $\mathbb{Z}$ (cf. §3.4).

5.10. ▷ Let R be an integral domain, M an R-module, and assume $M \cong R^r \oplus T$, with T a torsion module. Prove directly (that is, without using Theorem 5.6) that $r = \operatorname{rk} M$ and $T \cong \operatorname{Tor}_R(M)$. [§5.3]

5.11. ▷ Let R be an integral domain, let M, N be R-modules, and let $\varphi : M \to N$ be a homomorphism. For $m \in M$, show that $\operatorname{Ann}(\langle m \rangle) \subseteq \operatorname{Ann}(\langle \varphi(m) \rangle)$. [§5.3]

5.12. ▷ Complete the proof of uniqueness in Theorem 5.6. (The hint in Exercise IV.6.1 may be helpful.) [§5.3]

5.13. Let M be a finitely generated module over an integral domain R.

Prove that if R is a PID, then M is torsion-free if and only if it is free. Prove that this property characterizes PIDs. (Cf. Exercise 4.3.)

5.14. Give an example of a finitely generated module over an integral domain which is *not* isomorphic to a direct sum of cyclic modules.

[28] The finite rank hypothesis is actually unnecessary, but the proof is harder without this condition.

5.15. ▷ Prove that the prime ideals appearing in the elementary divisor version of the classification theorem for a torsion module M over a PID are the prime ideals containing the characteristic ideal of M, as defined in Remark 5.8. [§5.3]

5.16. Prove that the prime ideals appearing in the elementary divisor version of the classification theorem for a module M over a PID are the associated primes of M, as defined in Exercise 4.5.

5.17. Let R be a PID. Prove that the Grothendieck group (cf. §3.4) of the category of finitely generated R-modules is isomorphic to $\mathbb{Z}$.

---

6. Linear transformations of a free module

One beautiful application of the classification theorem for finitely generated modules over PIDs is the determination of 'special' forms for matrices of linear maps of a vector space to itself.

Several fundamental concepts are associated with the general notion of a linear map from a vector space to itself, and we will review these concepts in this section. Not surprisingly, much of the discussion may be carried out for free modules over any integral domain R, and we will stay at this level of generality in (most of) the section. Theorem 5.6 will be used with great profit when R is a field, in the next section.

6.1. Endomorphisms and similarity. Let R be an integral domain. We have considered in some detail the module $\mathrm{Hom}_R(F, G)$ of R-module homomorphisms

$$F \to G$$

between two free modules. For example, we have argued that describing such homomorphisms for finitely generated free modules F, G amounts to describing matrices with entries in R, up to 'equivalence': the equivalence relation introduced in §2.2 accounts for the (arbitrary) choice of bases for F and G.

We now shift the focus a little and consider the special case in which $F = G$, that is, the R-module $\mathrm{End}_R(F)$ of *endomorphisms* α of a fixed free R-module F:

$$F \xrightarrow{\alpha} F\ .$$

Note that $\mathrm{End}_R(F)$ is in fact an *R-algebra:* the operation of composition makes it a ring, compatibly with the R-module structure (cf. Exercise 2.3).

From the point of view championed in §2, the two copies of F appearing in $\mathrm{End}_R(F) = \mathrm{Hom}_R(F, F)$ are unrelated; they are just (any) isomorphic representatives of a free module of a given rank. There are circumstances, however, in which we need to really choose one representative F and stick with it: that is, view α as acting from a selected free module F to *itself,* not just to an isomorphic copy of itself. In this situation we also say that α is a *linear transformation* of F, or an *operator* on F.

From this different point of view it makes sense to compare elements of F 'before and after' we apply α; for example, we could ask whether for some $\mathbf{v} \in F$

we may have $\alpha(\mathbf{v}) = \lambda\mathbf{v}$ for some $\lambda \in R$, or more generally whether a submodule M of F may be sent to itself by α. In other words, we can compare the action of α with the identity[29] $I : F \to F$.

In terms of matrix representations (in the finite rank case) the description of α can be carried out as we did in §2, but with one interesting twist. In §2.2 we dealt with how the matrix representation of a homomorphism changes when we change bases in the source *and* in the target. As we are now *identifying* source and target, we must choose *the same* basis for both source and target. This leads to a different notion of equivalence of matrices:

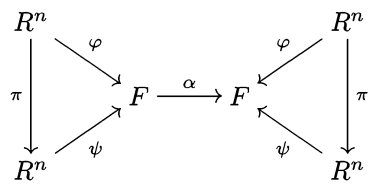
Definition 6.1. Two square matrices $A, B \in \mathcal{M}_n(R)$ are *similar* if they represent the same homomorphism $F \to F$ of a free rank-n module F to itself, up to the choice of a basis for F. ⌟

This is clearly an equivalence relation. The analog of Proposition 2.5 for this new notion is readily obtained:

Proposition 6.2. *Two matrices $A, B \in \mathcal{M}_n(R)$ are similar if and only if there exists an invertible matrix P such that*

$$B = PAP^{-1}.$$

The reader who has really understood Proposition 2.5 will not need any detailed proof of this statement: it should be apparent from staring at the butterfly diagram



which we are essentially copying from §2.2. The difference here is that we are choosing the same basis for source and target; hence the two triangles keeping track of the change of basis are the same. Suppose A (resp., B) represents α with respect to the choice of basis dictated by φ (resp., ψ); that is, A (resp., B) is the matrix of the 'top' (resp., bottom) composition

$$\varphi^{-1} \circ \alpha \circ \varphi : R^n \to R^n$$

(resp., $\psi^{-1} \circ \alpha \circ \psi$). If P is the matrix representing the change of basis π, then $B = PAP^{-1}$ simply because the diagram commutes:

$$\psi^{-1} \circ \alpha \circ \psi = (\pi \circ \varphi^{-1}) \circ \alpha \circ (\varphi \circ \pi^{-1}) = \pi \circ (\varphi^{-1} \circ \alpha \circ \varphi) \circ \pi^{-1}.$$

Proposition 6.2 suggests a useful equivalence relation among endomorphisms:

Definition 6.3. Two R-module homomorphisms of a free module F to itself,

$$\alpha, \beta : F \to F,$$

[29]The existence of the identity is one of the axioms of a category; every now and then, it is handy to be able to invoke it.

are *similar* if there exists an automorphism $\pi : F \to F$ such that

$$\beta = \pi \circ \alpha \circ \pi^{-1}. \qquad \lrcorner$$

In the finite rank case, similar endomorphisms are represented by similar matrices, and two endomorphisms α, β are similar if and only if they may be represented by the *same* matrix by choosing appropriate (possibly different) bases on F. The interesting invariants determined by similar endomorphisms will turn out (not surprisingly) to be the same.

From a group-theoretic viewpoint, similarity is an eminently natural notion to study: the group $\mathrm{GL}(F) = \mathrm{Aut}_R(F)$ of automorphisms of a free module acts on $\mathrm{End}_R(F)$ by *conjugation,* and two endomorphisms α, β are similar if and only if they are in the same orbit under this action. If the reader prefers matrices, the group $\mathrm{GL}_n(R)$ of invertible $n \times n$ matrices with entries in R acts by conjugation on the module $\mathcal{M}_n(R)$ of square $n \times n$ matrices, and two matrices are similar if and only if they are in the same orbit.

Natural questions arise in this context. For example, we should look for ways to distinguish different orbits: given two endomorphisms α, β, can we effectively decide whether α and β are similar? One way to approach this question is to determine invariants which can distinguish different orbits. Better still, we can look for 'special' representatives within each orbit, that is, of a given similarity class. In other words, given an endomorphism $\alpha : F \to F$, find a basis of F with respect to which the matrix description of α has a particular, predictable shape. Then α, β are similar if these distinguished representatives coincide.

This is what we are eventually going to squeeze out of Theorem 5.6, in the friendly case of vector spaces over a fixed field.

6.2. The characteristic and minimal polynomials of an endomorphism. Let $\alpha \in \mathrm{End}_R(F)$ be an endomorphism of a free R-module; henceforth we are often tacitly going to assume that F is finitely generated (this is necessary anyway as we are aiming to translate what we do into the language of matrices). We want to identify *invariants* of the similarity class of α, that is, quantities that will not change if we replace α with a similar linear map $\beta \in \mathrm{End}_R(F)$.

For example, the *determinant* is such a quantity:

Definition 6.4. Let $\alpha \in \mathrm{End}_R(F)$. The *determinant* of α is $\det \alpha := \det A$, where A is the matrix representing α with respect to any choice of basis of F. $\lrcorner$

Of course there is something to check here, that is, that $\det A$ does not depend on the choice of basis. But we know (Proposition 6.2) that A, B represent the same endomorphism α if and only if there exists an invertible matrix P such that $B = PAP^{-1}$; then

$$\det(B) = \det(PAP^{-1}) = \det(P)\det(A)\det(P^{-1}) = \det(A)$$

by Proposition 3.3. Thus the determinant is indeed independent of the choice of basis. Essentially the same argument shows that if α and β are similar linear transformations, then $\det(\alpha) = \det(\beta)$ (Exercise 6.3).

By Proposition 3.3, a linear transformation α is invertible if and only if $\det\alpha$ is a unit in R. If R is a field, this of course means simply $\det\alpha \neq 0$. Even if R is not a field, $\det\alpha \neq 0$ says something interesting:

Proposition 6.5. *Let α be a linear transformation of a free R-module $F \cong R^n$. Then $\det\alpha \neq 0$ if and only if α is injective.*

Proof. Embed R in its field of fractions K, and view α as a linear transformation of K^n; note that the determinant of α is the same whether it is computed over R or over K. Then α is injective as a linear transformation $R^n \to R^n$ if and only if it is injective as a linear transformation $K^n \to K^n$, if and only if it is invertible as a linear transformation $K^n \to K^n$, if and only if $\det\alpha \neq 0$. □

Of course over integral domains other than fields α may fail to be *surjective* even if $\det\alpha \neq 0$; and care is required even over fields, if we abandon the hypothesis that the free modules are finitely generated (cf. Exercises 6.4 and 6.5).

Another quantity that is invariant under similarity is the *trace*. The trace of a square matrix $A = (a_{ij}) \in M_n(R)$ is

$$\operatorname{tr}(A) := \sum_{i=1}^{n} a_{ii},$$

that is, the sum of its diagonal entries.

Definition 6.6. Let $\alpha \in \operatorname{End}_R(F)$. The *trace* of α is defined to be $\operatorname{tr}\alpha := \operatorname{tr} A$, where A is the matrix representing α with respect to any choice of basis of F. ⌟

Again, we have to check that this is independent of the choice of basis; the key is the following computation.

Lemma 6.7. *Let $A, B \in \mathcal{M}_n(R)$. Then $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.*

Proof. Let $A = (a_{ij})$, $B = (b_{ij})$. Then $AB = (\sum_{k=1}^{n} a_{ik}b_{kj})$; hence

$$\operatorname{tr}(AB) = \sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}b_{ki}.$$

This expression is symmetric in A, B, so it must equal $\operatorname{tr}(BA)$. □

With this understood, if $B = PAP^{-1}$, then

$$\operatorname{tr}(B) = \operatorname{tr}((PA)P^{-1}) = \operatorname{tr}(P^{-1}(PA)) = \operatorname{tr}((P^{-1}P)A) = \operatorname{tr}(A),$$

showing (by Proposition 6.2) that similar matrices have the same trace, as needed.

Again, the reader will check that the same token shows that similar linear transformations have the same trace.

The trace and determinant of an endomorphism α of a rank-n free module are in fact just two of a sequence of n invariants, which can be nicely collected together into the *characteristic polynomial* of α.

Definition 6.8. Let F be a free R-module, and let $\alpha \in \mathrm{End}_R(F)$. Denote by I the identity map $F \to F$. The *characteristic polynomial* of α is the polynomial

$$P_\alpha(t) := \det(tI - \alpha) \in R[t].$$

Proposition 6.9. *Let F be a free R-module of rank n, and let $\alpha \in \mathrm{End}_R(F)$.*

- *The characteristic polynomial $P_\alpha(t)$ is a monic polynomial of degree n.*
- *The coefficient of t^{n-1} in $P_\alpha(t)$ equals $-\mathrm{tr}(\alpha)$.*
- *The constant term of $P_\alpha(t)$ equals $(-1)^n \det(\alpha)$.*
- *If α and β are similar, then $P_\alpha(t) = P_\beta(t)$.*

Proof. The first point is immediate, and the third is checked by setting $t = 0$. To verify the second assertion, let $A = (a_{ij})$ be a matrix representing α with respect to any basis for F, so that

$$P_\alpha(t) = \det \begin{pmatrix} t - a_{11} & -a_{12} & & -a_{1n} \\ -a_{21} & t - a_{22} & & -a_{2n} \\ \vdots & \vdots & & \vdots \\ -a_{n1} & -a_{n2} & \dots & t - a_{nn} \end{pmatrix}.$$

Expanding the determinant according to Definition 3.1 (or in any other way), we see that the only contributions to the coefficient of t^{n-1} come from the diagonal entries, in the form

$$\sum_{i=1}^{n} t \cdots t \cdot (-a_{ii}) \cdot t \cdots t,$$

and the statement follows.

Finally, assume that α and β are similar. Then there exists an invertible π such that $\beta = \pi \circ \alpha \circ \pi^{-1}$, and hence

$$tI - \beta = \pi \circ (tI - \alpha) \circ \pi^{-1}$$

are similar (as endomorphisms of $R[t]^n$). By Exercise 6.3 these two transformations must have the same determinant, and this proves the fourth point. □

By Proposition 6.9, all coefficients in the characteristic polynomial

$$t^n - (\mathrm{tr}\,\alpha)t^{n-1} + \cdots + (-1)^n(\det \alpha)$$

are invariant under similarity; as far as we know, trace and determinant are the only ones that have special names.

Determinants, traces, and more generally the characteristic polynomial can show very quickly that two linear transformations are *not* similar; but do they tell us unfailingly when two transformations *are* similar? In general, the answer is no, even over fields. For example, the two matrices

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

both have characteristic polynomial $(t-1)^2$, but they are not similar. Indeed, I is the identity and clearly only the identity is similar to the identity: $PIP^{-1} = I$ for all invertible matrices P.

Therefore, the equivalence relation defined by prescribing that two transformations have the same characteristic polynomial is coarser than similarity. This hints that there must be other interesting quantities associated with a linear transformation and invariant under similarity.

We are going to understand this much more thoroughly in a short while (at least over fields), but we can already gain some insight by contemplating *another* kind of 'polynomial' information, which also turns out to be invariant under similarity.

As $\mathrm{End}_R(F)$ is an R-algebra, we can evaluate every polynomial

$$f(t) = r_m t^m + r_{m-1} t^{m-1} + \cdots + r_0 \in R[t]$$

at any $\alpha \in \mathrm{End}_R(F)$:

$$f(\alpha) = r_m \alpha^m + r_{m-1} \alpha^{m-1} + \cdots + r_0 \in \mathrm{End}_R(F).$$

In other words, we can perform these operations in the *ring* $\mathrm{End}_R(F)$; multiplication by $r \in R$ amounts to composition with $rI \in \mathrm{End}_R(F)$, and α^k stands for the k-fold composition $\alpha \circ \cdots \circ \alpha$ of α with itself. The set of polynomials such that $f(\alpha) = 0$ is an ideal of $R[t]$ (Exercise 6.7), which we will denote $\mathscr{I}_\alpha$ and call the *annihilator ideal* of α.

Lemma 6.10. *If α and β are similar, then $\mathscr{I}_\alpha = \mathscr{I}_\beta$.*

Proof. By hypothesis there exists an invertible π such that $\beta = \pi \circ \alpha \circ \pi^{-1}$. As

$$\beta^k = (\pi \circ \alpha \circ \pi^{-1})^k = (\pi \circ \alpha \circ \pi^{-1}) \circ (\pi \circ \alpha \circ \pi^{-1}) \circ \cdots \circ (\pi \circ \alpha \circ \pi^{-1}) = \pi \circ \alpha^k \circ \pi^{-1},$$

we see that for all $f(t) \in R[t]$ we have

$$f(\beta) = \pi \circ f(\alpha) \circ \pi^{-1}.$$

It follows immediately that $f(\alpha) = 0 \iff f(\beta) = 0$, which is the statement. □

Going back to the simple example shown above, the polynomial $t - 1$ is in the annihilator ideal of the identity, while it is *not* in the annihilator ideal of the matrix

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

An optimistic reader might now guess that two linear transformations are similar if and only if *both* their characteristic polynomials and annihilator ideals coincide. This is unfortunatley not the case in general, but we promise that the situation will be considerably clarified in short order (cf. Exercise 7.3).

In any case, even the simple example given above allows us to point out a remarkable fact. Note that the (common) characteristic polynomial $(t-1)^2$ of I and A annihilates *both:*

$$(A-1)^2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

This is not a coincidence: $P_\alpha(t) \in \mathscr{I}_\alpha$ for all linear transformations α. That is,

Theorem 6.11 (Cayley-Hamilton). *Let $P_\alpha(t)$ be the characteristic polynomial of the linear transformation $\alpha \in \mathrm{End}_R(F)$. Then*

$$P_\alpha(\alpha) = 0.$$

This beautiful observation can be proved directly by judicious use of Cramer's rule[30], in the form of Corollary 3.5; cf. Exercise 6.9. In any case, the Cayley-Hamilton theorem will become essentially evident once we connect these linear algebra considerations with the classification theorem for finitely generated modules over a PID; the adventurous reader can already look back at Remark 5.8 and figure out why the Cayley-Hamilton theorem is obvious.

If R is an arbitrary integral domain, we cannot expect too much of $R[t]$, and it seems hard to say something *a priori* concerning $\mathscr{I}_\alpha$. However, consider the field of fractions K of R (§V.4.2); viewing α as an element of $\mathrm{End}_K(K^n)$ (that is, viewing the entries of a matrix representation of α as elements of K, rather than R), α will have an annihilator ideal $\mathscr{I}_\alpha^{(K)}$ 'over K', and it is clear that

$$\mathscr{I}_\alpha = \mathscr{I}_\alpha^{(K)} \cap R[t].$$

The advantage of considering $\mathscr{I}_\alpha^{(K)} \subseteq K[t]$ is that $K[t]$ is a PID, and it follows that $\mathscr{I}_\alpha^{(K)}$ has a (unique) monic generator.

Definition 6.12. Let F be a free R-module, and let $\alpha \in \mathrm{End}_R(F)$. Let K be the field of fractions of R. The *minimal polynomial* of α is the monic generator $m_\alpha(t) \in K[t]$ of $\mathscr{I}_\alpha^{(K)}$. ⌟

With this terminology, the Cayley-Hamilton theorem amounts to the assertion that *the minimal polynomial divides the characteristic polynomial:* $m_\alpha(t) \mid P_\alpha(t)$.

Of course the situation is simplified, at least from an expository point of view, if R is itself a field: then $K = R$, $m_\alpha(t) \in R[t]$, and $\mathscr{I}_\alpha = (m_\alpha(t))$. This is one reason why we will eventually assume that R is a field.

6.3. Eigenvalues, eigenvectors, eigenspaces.

Definition 6.13. Let F be a free R-module, and let $\alpha \in \mathrm{End}_R(F)$ be a linear transformation of F. A scalar $\lambda \in R$ is an *eigenvalue* for α if there exists $\mathbf{v} \in F$, $\mathbf{v} \neq 0$, such that

$$\alpha(\mathbf{v}) = \lambda \mathbf{v}.$$ ⌟

For example, 0 is an eigenvalue for α precisely when α has a nontrivial kernel. The notion of eigenvalue is one of the most important in linear algebra, if not in algebra, if not in mathematics, if not in the whole of science. The set of eigenvalues of a linear transformation is called its *spectrum.* Spectra of operators show up everywhere, from number theory to differential equations to quantum mechanics. The spectrum *of a ring,* encountered briefly in §III.4.3, was so named because it may be interpreted (in important motivating contexts) as a spectrum in the sense

[30]Here is an even more direct 'proof': $P_\alpha(\alpha) = \det(\alpha I - \alpha) = \det(0) = 0$. Unfortunately this does not work: in the definition $\det(tI - \alpha)$ of the characteristic polynomial, t is assumed to act as a scalar and cannot be replaced by α. Applying Cramer's rule circumvents this obstacle.

of Definition 6.13. The 'spectrum of the hydrogen atom' is also a spectrum in this sense.

It is hopefully immediate that similar transformations have the same spectrum: for if $\beta = \pi \circ \alpha \circ \pi^{-1}$ and $\alpha(\mathbf{v}) = \lambda\mathbf{v}$, then

$$\beta(\pi(\mathbf{v})) = \pi \circ \alpha \circ \pi^{-1}(\pi(\mathbf{v})) = \pi \circ (\alpha(\mathbf{v})) = \pi(\lambda\mathbf{v}) = \lambda\pi(\mathbf{v});$$

as $\pi(\mathbf{v}) \neq 0$ if $\mathbf{v} \neq 0$, this shows that every eigenvalue of α is an eigenvalue of β.

If F is finitely generated, then we have the following useful translation of the notion of eigenvalue:

Lemma 6.14. *Let F be a finitely generated R-module, and let $\alpha \in \mathrm{End}_R(F)$. Then the set of eigenvalues of α is precisely the set of roots in R of the characteristic polynomial $P_\alpha(t)$.*

Proof. This is a straightforward consequence of Proposition 6.5:

$$\begin{aligned}
\lambda \text{ is an eigenvalue for } \alpha &\iff \exists \mathbf{v} \neq 0 \text{ such that } \alpha(\mathbf{v}) = \lambda I(\mathbf{v}) \\
&\iff \exists \mathbf{v} \neq 0 \text{ such that } (\lambda I - \alpha)(\mathbf{v}) = 0 \\
&\iff \lambda I - \alpha \text{ is not injective} \\
&\iff \det(\lambda I - \alpha) = 0 \\
&\iff P_\alpha(\lambda) = 0
\end{aligned}$$

as claimed. □

It is an evident consequence of Lemma 6.14 that eigenvalue considerations depend very much on the base ring R.

Example 6.15. The matrix

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

has no eigenvalues over $\mathbb{R}$, while it has eigenvalues over $\mathbb{C}$: indeed, the characteristic polynomial $t^2 + 1$ has no real roots and two complex roots. The reader should observe that, as a linear transformation of the real plane $\mathbb{R}^2$, this matrix corresponds to a 90° counterclockwise rotation; the reason why this transformation has no (real) eigenvalues is that no direction in the plane is preserved through a 90° rotation. ⌟

Example 6.16. An example with a different flavor is given by the matrix

$$\begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}.$$

This has characteristic polynomial $t^2 - 2$; hence it has eigenvalues over $\mathbb{R}$, but not over $\mathbb{Q}$. Geometrically, the corresponding transformation flips the plane about a line; but that line has irrational slope, so it contains no nonzero vectors with rational components. ⌟

As another benefit of Lemma 6.14, we may now introduce the following notion:

Definition 6.17. The *algebraic multiplicity* of an eigenvalue of a linear transformation α of a finitely generated free module is its multiplicity as a root of the characteristic polynomial of α. ⌟

For example, the identity on F has the single eigenvalue 1, with (algebraic) multiplicity equal to the *rank* of F.

The sum of the algebraic multiplicities is bounded by the degree of the characteristic polynomial, that is, the dimension of the space. In particular,

Corollary 6.18. *The number of eigenvalues of a linear transformation of R^n is at most n. If the base ring R is an algebraically closed field, then every linear transformation has exactly n eigenvalues (counted with algebraic multiplicity).*

Proof. Immediate from Lemmas 6.14, V.5.1, and V.5.10. □

There is a different notion of multiplicity of an eigenvalue, related to how big the corresponding[31] *eigenspace* may be.

Definition 6.19. Let λ be an eigenvalue of a linear transformation α of a free R-module F. Then a nonzero $\mathbf{v} \in F$ is an *eigenvector* for α, corresponding to the eigenvalue λ, if $\alpha(\mathbf{v}) = \lambda\mathbf{v}$, that is, if $\mathbf{v} \in \ker(\lambda I - \alpha)$. The submodule $\ker(\lambda I - \alpha)$ is the *eigenspace* corresponding to λ. ⌟

Definition 6.20. The *geometric multiplicity* of an eigenvalue λ is the rank of its eigenspace. ⌟

This is clearly invariant under similarity (Exercise 6.14). Geometric and algebraic multiplicities do not necessarily coincide: for example, the matrix

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

has characteristic polynomial $(t-1)^2$, and hence the single eigenvalue 1, with algebraic multiplicity 2. However, for a vector $\mathbf{v} = \binom{v_1}{v_2}$,

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 + v_2 \\ v_2 \end{pmatrix}$$

equals $1\mathbf{v}$ if and only if $v_2 = 0$: that is, the eigenspace corresponding to the single eigenvalue 1 consists of the span of the vector $\binom{1}{0}$ and has dimension 1. Thus, the geometric multiplicity of the eigenvalue is 1 in this example.

Contemplating this example will convince the reader that the geometric multiplicity of an eigenvalue is always *less than* its algebraic multiplicity. In the neatest possible situation, an operator α on a free module F of rank n may have all its n eigenvalues in the base ring R, and each algebraic multiplicity may agree with the corresponding geometric multiplicity. If this is the case, F may then be expressed as a direct sum of the eigenspaces, producing the so-called *spectral decomposition* of F determined by α (cf. Exercise 6.15 for a concrete instance of this situation). The action of α on F is then completely transparent, as it amounts to simply applying a (possibly) different scaling factor on each piece of the spectral decomposition.

[31] If we were serious about pursuing the theory over arbitrary integral domains, we would feel compelled to call this object the eigen*module* of α. However, we are eventually going to restrict our attention to the case in which the base ring is a field, so we will not bother.

If A is a matrix representing $\alpha \in \mathrm{End}_R(V)$ with respect to any basis, then α admits a spectral decomposition if and only if A is similar to a diagonal matrix:

$$A = P \begin{pmatrix} \lambda_1 & 0 & & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} P^{-1},$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of α. In this case we say that A (or α) is *diagonalizable.* A moment's thought reveals that the columns of P are then a basis of V consisting of eigenvectors of α.

Once more, we promise that the situation will become (even) clearer once we bring Theorem 5.6 into the picture.

Exercises

In these exercises, R denotes an integral domain.

6.1. Let k be an infinite field, and let n be any positive integer.

- Prove that there are finitely many *equivalence* classes of matrices in $\mathcal{M}_n(k)$.
- Prove that there are infinitely many *similarity* classes of matrices in $\mathcal{M}_n(k)$.

6.2. Let F be a free R-module of rank n, and let $\alpha, \beta \in \mathrm{End}_R(F)$. Prove that $\det(\alpha \circ \beta) = \det(\alpha)\det(\beta)$. Prove that α is invertible if and only if $\det(\alpha)$ is invertible.

6.3. ▷ Prove that if α and β are similar in the sense of Definition 6.3, then $\det(\alpha) = \det(\beta)$ and $\mathrm{tr}(\alpha) = \mathrm{tr}(\beta)$. (Do this without using Proposition 6.9!) [§6.2]

6.4. ▷ Let F be a finitely generated free R-module, and let α be a linear transformation of F. Give an example of an injective α which is not surjective; in fact, prove that α is not surjective precisely when $\det R$ is not a unit. [§6.2]

6.5. ▷ Let k be a field, and view $k[t]$ as a vector space over k in the evident way. Give an example of a k-linear transformation $k[t] \to k[t]$ which is injective but not surjective; give an example of a linear transformation which is surjective but not injective. [§6.2, §VII.4.1]

6.6. Prove that two 2×2 matrices have the same characteristic polynomial if and only if they have the same trace and determinant. Find two 3×3 matrices with the same trace and determinant but different characteristic polynomials.

6.7. ▷ Let $\alpha \in \mathrm{End}_R(F)$ be a linear transformation on a free R-module F. Prove that the set of polynomials $f(t) \in R[t]$ such that $f(\alpha) = 0$ is an ideal of $R[t]$. [§6.2]

6.8. ▷ Let $A \in \mathcal{M}_n(R)$ be a square matrix, and let A^t be its transpose. Prove that A and A^t have the same characteristic polynomial and the same annihilator ideals. [§7.2]

6.9. ▷ Prove the Cayley-Hamilton theorem, as follows. Recall that every square matrix M has an *adjoint* matrix, which we will denote $\operatorname{adj}(M)$, and that we proved (Corollary 3.5) that $\operatorname{adj}(M) \cdot M = \det(M) \cdot I$. Applying this to $M = tI - A$ (with A a matrix realization of $\alpha \in \operatorname{End}_R(F)$) gives

$$(*) \qquad \operatorname{adj}(tI - A) \cdot (tI - A) = P_\alpha(t) \cdot I.$$

Prove that there exist matrices $B_k \in \mathcal{M}_n(R)$ such that $\operatorname{adj}(tI - A) = \sum_{k=0}^{n-1} B_k t^k$; then use (*) to obtain $P_\alpha(A) = 0$, proving the Cayley-Hamilton theorem. [§6.2]

6.10. ▷ Let F_1, F_2 be free R-modules of finite rank, and let α_1, resp., α_2, be linear transformations of F_1, resp., F_2. Let $F = F_1 \oplus F_2$, and let $\alpha = \alpha_1 \oplus \alpha_2$ be the linear transformation of F restricting to α_1 on F_1 and α_2 on F_2.

- Prove that $P_\alpha(t) = P_{\alpha_1}(t) P_{\alpha_2}(t)$. That is, the characteristic polynomial is multiplicative under direct sums.
- Find an example showing that the minimal polynomial is *not* multiplicative under direct sums.

[6.11, §7.2]

6.11. ¬ Let α be a linear transformation of a finite-dimensional vector space V, and let V_1 be an invariant subspace, that is, such that $\alpha(V_1) \subseteq V_1$. Let α_1 be the restriction of α to V_1, and let $V_2 = V/V_1$. Prove that α induces a linear transformation α_2 on V_2, and (in the same vein as Exercise 6.10) show that $P_\alpha(t) = P_{\alpha_1}(t) P_{\alpha_2}(t)$. Also, prove that $\operatorname{tr} \alpha^r = \operatorname{tr} \alpha_1^r + \operatorname{tr} \alpha_2^r$, for all $r \geq 0$. [6.12]

6.12. Let α be a linear transformation of a finite-dimensional $\mathbb{C}$-vector space V. Prove the identity of formal power series with coefficients in $\mathbb{C}$:

$$\frac{1}{\det(1 - \alpha t)} = \exp\left(\sum_{r=1}^{\infty} \operatorname{tr} \alpha^r \frac{t^r}{r} \right).$$

(Hint: The left-hand side is essentially the inverse of the characteristic polynomial of α. Use Exercise 6.11 to show that both sides are multiplicative with respect to exact sequences $0 \to V_1 \to V \to V_2 \to 0$, where V_1 is an invariant subspace. Use this and the fact that α admits nontrivial invariant subspaces since $\mathbb{C}$ is algebraically closed (why?) to reduce to the case $\dim V = 1$, for which the identity is an elementary calculus exercise.)

With due care, the identity can be stated and proved (in the same way) over arbitrary fields of characteristic 0. It is an ingredient in the cohomological interpretation of the 'Weil conjectures'.

6.13. Let A be a square matrix with integer entries. Prove that if λ is a *rational* eigenvalue of A, then in fact $\lambda \in \mathbb{Z}$. (Hint: Proposition V.5.5.)

6.14. ▷ Let λ be an eigenvalue of two similar transformations α, β. Prove that the geometric multiplicities of λ with respect to α and β coincide. [§6.3]

6.15. ▷ Let α be a linear transformation on a free R-module F, and let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be eigenvectors corresponding to *pairwise distinct* eigenvalues $\lambda_1, \ldots, \lambda_n$. Prove that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent. (Hint: If not, there is a *shortest* linear

combination $r_1\mathbf{v}_{i_1} + \cdots + r_m\mathbf{v}_{i_m} = 0$ with all $r_j \in R$, $r_j \neq 0$. Compare the action of α on this linear combination with the product by λ_{i_1}.)

It follows that if F has rank n and α has n distinct eigenvalues, then α induces a spectral decomposition of F. [§6.3, VII.6.14]

6.16. ¬ The *standard inner product* on $V = \mathbb{R}^n$ is the map $V \times V \to \mathbb{R}$ defined by

$$(\mathbf{v}, \mathbf{w}) := \mathbf{v}^t \cdot \mathbf{w}$$

(viewing elements $\mathbf{v} \in V$ as column vectors). The *standard hermitian product* on $W = \mathbb{C}^n$ is the map $W \times W \to \mathbb{C}$ defined by

$$(\mathbf{v}, \mathbf{w}) := \mathbf{v}^\dagger \cdot \mathbf{w},$$

where for any matrix M, $M^\dagger$ stands for the matrix obtained by taking the complex conjugates of the entries of the transpose M^t.

These products satisfy evident linearity properties: for example, for $\lambda \in \mathbb{C}$ and $\mathbf{v}, \mathbf{w} \in W$

$$(\lambda\mathbf{v}, \mathbf{w}) = \overline{\lambda}(\mathbf{v}, \mathbf{w}).$$

Prove[32] that a matrix $M \in \mathcal{M}_n(\mathbb{R})$ belongs to $\mathrm{O}_n(\mathbb{R})$ if and only if it preserves the standard inner product on $\mathbb{R}^n$:

$$(\forall \mathbf{v}, \mathbf{w} \in \mathbb{R}^n) \qquad (M\mathbf{v}, M\mathbf{w}) = (\mathbf{v}, \mathbf{w}).$$

Likewise, prove that a matrix $M \in \mathcal{M}_n(\mathbb{C})$ belongs to $\mathrm{U}_n(\mathbb{C})$ if and only if it preserves the standard hermitian product on $\mathbb{C}^n$. [6.18]

6.17. ¬ We say that two vectors $\mathbf{v}$, $\mathbf{w}$ of $\mathbb{R}^n$ or $\mathbb{C}^n$ are *orthogonal* if $(\mathbf{v}, \mathbf{w}) = 0$. The *orthogonal complement* $\mathbf{v}^\perp$ of $\mathbf{v}$ is the set of vectors $\mathbf{w}$ that are orthogonal to $\mathbf{v}$. Prove that if $\mathbf{v} \neq 0$ in $V = \mathbb{R}^n$ or $\mathbb{C}^n$, then $\mathbf{v}^\perp$ is a subspace of V of dimension $n-1$. [7.16, VIII.5.15]

6.18. ¬ Let $V = \mathbb{R}^n$, endowed with the standard inner product defined in Exercise 6.16. A set of distinct vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ is *orthonormal* if $(\mathbf{v}_i, \mathbf{v}_j) = 0$ for $i \neq j$ and 1 for $i = j$. Geometrically, this means that each vector has length 1, and different vectors are orthogonal. The same terminology may be used in $\mathbb{C}^n$, w.r.t. the standard hermitian product.

Prove that $M \in \mathrm{O}_n(\mathbb{R})$ if and only if the columns of M are orthonormal, if and only if the rows of M are orthonormal[33].

Formulate and prove an analogous statement for $\mathrm{U}_n(\mathbb{C})$. (The group $\mathrm{U}_n(\mathbb{C})$ is called the *unitary* group. Note that, for $n = 1$, it consists of the complex numbers of norm 1.) [6.19, 7.16, VIII.5.9]

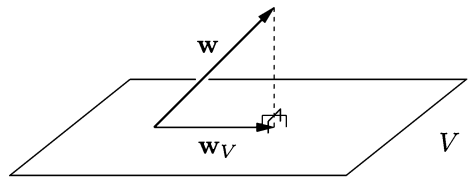
6.19. ¬ Let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ form an orthonormal set of vectors in $\mathbb{R}^n$.

---

[32] See Exercise II.6.1 for the definitions of $\mathrm{O}_n(\mathbb{R})$ and $\mathrm{U}_n(\mathbb{C})$. We trust that basic facts on inner products are not new to the reader. Among these, recall that for $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, $(\mathbf{v}, \mathbf{v})$ equals the square of the length of $\mathbf{v}$, and $(\mathbf{v}, \mathbf{w})$ equals the product of the lengths of $\mathbf{v}$ and $\mathbf{w}$ and the cosine of the angle formed by $\mathbf{v}$ and $\mathbf{w}$. Thus, the reader is essentially asked to verify that $M \in \mathrm{O}_n(\mathbb{R})$ if and only if M preserves lengths and angles, and to check an analogous statement in the complex environment.

[33] The group $\mathrm{O}_n(\mathbb{R})$ is called the *orthogonal group;* ortho*normal* group would probably be a more appropriate terminology.

- Prove that $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are linearly independent; so they form an *orthonormal basis* of the space V they span.
- Let $\mathbf{w} = a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r$ be a vector of V. Prove that $a_i = (\mathbf{v}_i, \mathbf{w})$.
- More generally, prove that if $\mathbf{w} \in \mathbb{R}^n$, then $(\mathbf{v}_i, \mathbf{w})$ is the component of $\mathbf{v}_i$ in the *orthogonal projection* $\mathbf{w}_V$ of $\mathbf{w}$ onto V. (That is, prove that $\mathbf{w} - \mathbf{w}_V$ is orthogonal to all vectors of V.)



For reasons such as these, it is convenient to work with orthonormal bases. Note that, by Exercise 6.18, a matrix is in $\mathrm{O}_n(\mathbb{R})$ if and only if its columns form an orthonormal basis of $\mathbb{R}^n$. Again, the reader should formulate parallel statements for hermitian products in $\mathbb{C}^n$. [6.20]

6.20. The *Gram-Schmidt process* takes as input a choice of linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ of $\mathbb{R}^n$ and returns vectors $\mathbf{w}_1, \ldots, \mathbf{w}_r$ spanning the same space V spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_r$ and such that $(\mathbf{w}_i, \mathbf{w}_j) = 0$ for $i \neq j$. Scaling each $\mathbf{w}_i$ by its length, this yields an orthonormal basis for V.

Here is how the Gram-Schmidt process works:

— $\mathbf{w}_1 := \mathbf{v}_1$.

— For $k \geq 1$, $\mathbf{w}_k :=$ $\mathbf{v}_k$ – orthogonal projection of $\mathbf{v}_k$ onto $\langle \mathbf{w}_1, \ldots, \mathbf{w}_{k-1} \rangle$.

(Cf. Exercise 6.19.) Prove that this process accomplishes the stated goal.

6.21. ¬ A matrix $M \in \mathcal{M}_n(\mathbb{R}^n)$ is *symmetric* if $M^t = M$. Prove that M is symmetric if and only if $(\forall \mathbf{v}, \mathbf{w} \in \mathbb{C}^n)$, $(M\mathbf{v}, \mathbf{w}) = (\mathbf{v}, M\mathbf{w})$.

A matrix $M \in \mathcal{M}_n(\mathbb{C}^n)$ is *hermitian* if $M^\dagger = M$. Prove that M is hermitian if and only if $(\forall \mathbf{v}, \mathbf{w} \in \mathbb{C}^n)$, $(M\mathbf{v}, \mathbf{w}) = (\mathbf{v}, M\mathbf{w})$.

In both cases, one may say that M is *self-adjoint;* this means that shuttling it from one side of the product to the other does not change the result of the operation.

A hermitian matrix with real entries is symmetric. It is in fact useful to think of real symmetric matrices as particular cases of hermitian matrices. [6.22]

6.22. ¬ Prove that the eigenvalues of a hermitian matrix (Exercise 6.21) are real. Also, prove that if $\mathbf{v}$, $\mathbf{w}$ are eigenvectors of a hermitian matrix, corresponding to different eigenvalues, then $(v, w) = 0$. (Thus, eigenvectors with distinct eigenvalues for a real symmetric matrix are orthogonal.) [7.20]

---

7. Canonical forms

7.1. Linear transformations of free modules; actions of polynomial rings. We have promised to use Theorem 5.6 to better understand the similarity relation

and to obtain special forms for square matrices with entries in a field. We are now ready to make good on our promise.

The questions the reader should be anxiously asking are, where is the PID? Where is the finitely generated module? Why would a classification theorem for these things have anything to tell us about matrices? Once these questions are answered, everything else will follow easily. In a sense, this is the *one* issue that we keep in our mind concerning all these considerations: once we remember how to get an interesting module out of a linear transformation of vector spaces, the rest is essentially an immediate consequence of standard notions.

Once more, while the main application will be to vector spaces and matrices over a *field,* we do not need to preoccupy ourselves with specializing the situation before we get to the key point. Therefore, let's keep working for a while longer on our given *integral domain*[34] R.

Claim 7.1. *Giving a linear transformation on a free R-module F is the same as giving an $R[t]$-module structure on F, compatible with its R-module structure.*

Here $R[t]$ is the polynomial ring in one indeterminate t over the ring R. The claim is much less impressive than it sounds; in fact, it is completely tautological. Giving an $R[t]$-module structure on F compatible with the R-module structure is the same as giving a homomorphism of R-algebras

$$\varphi : R[t] \to \mathrm{End}_R(F)$$

(this is an insignificant upgrade of the very definition of module from §III.5.1). By the universal property satisfied by polynomial rings (§III.2.2, especially Example III.2.3), giving φ as an extension of the basic R-algebra structure of $\mathrm{End}_R(F)$ is just the same as specifying the image $\varphi(t)$, that is, choosing an element of $\mathrm{End}_R(F)$. This is precisely what the claim says.

Our propensity for the use of a certain language may obfuscate the matter, which is very simple. Given a linear transformation α of a free module F, we can define the action of a polynomial

$$f(t) = r_m t^m + r_{m-1} t^{m-1} + \cdots + r_0 \in R[t]$$

on F as follows: for every $\mathbf{v} \in F$, set

$$f(t)(\mathbf{v}) := r_m \alpha^m(\mathbf{v}) + r_{m-1} \alpha^{m-1}(\mathbf{v}) + \cdots + r_0 \mathbf{v},$$

where α^k denotes the k-fold composition of α with itself; we have already run into this in §6.2. Conversely, an action of $R[t]$ on F determines in particular a map $\alpha : F \to F$, that is, 'multiplication by t':

$$\mathbf{v} \mapsto t\mathbf{v};$$

the compatibility with the R-module structure on F tells us precisely that α is a linear transformation of F. Again, this is the content of Claim 7.1.

Tautological as it is, Claim 7.1 packs a good amount of information. Indeed, the $R[t]$ module knows everything about *similarity:*

[34] In fact, the requirements that R be an integral domain and that F be free are immaterial here; cf. Exercise III.5.11.

Lemma 7.2. *Let α, β be linear transformations of a free R-module F. Then the corresponding $R[t]$-module structures on F are isomorphic if and only if α and β are similar.*

Proof. Denote by F_α, F_β the two $R[t]$-modules defined on F by α, β as per Claim 7.1.

Assume first that α and β are similar. Then there exists an invertible R-linear transformation $\pi : F \to F$ such that

$$\beta = \pi \circ \alpha \circ \pi^{-1};$$

that is, $\pi \circ \alpha = \beta \circ \pi$. We can view π as an R-linear map

$$F_\alpha \to F_\beta.$$

We claim that it is $R[t]$-linear: indeed, multiplication by t is α in F_α and β in F_β, so

$$\pi(t\mathbf{v}) = \pi \circ \alpha(\mathbf{v}) = \beta \circ \pi(\mathbf{v}) = t\pi(\mathbf{v}).$$

Thus π is an invertible $R[t]$-linear map $F_\alpha \to F_\beta$, proving that F_α and F_β are isomorphic as $R[t]$-modules.

The converse implication is obtained essentially by running this argument in reverse and is left to the reader (Exercise 7.1). □

Corollary 7.3. *There is a one-to-one correspondence between the similarity classes of R-linear transformations of a free R-module F and the isomorphism classes of $R[t]$-module structures on F.*

Of course the same statement holds for square matrices with entries in R, in the finite rank case.

Corollary 7.3 is the tool we were looking for, bridging between classifying similarity classes of linear transformations or matrices and classifying modules. The task now becomes that of translating the notions we have encountered in §6 into the module language and seeing if this teaches us anything new.

In any case, since we have classified finitely generated modules over PIDs, it is clear that we will be able to classify similarity classes of transformations of finite-rank free modules—provided that $R[t]$ is a PID. This is where the additional hypothesis that R be a field enters the discussion (cf. Exercise V.2.12).

7.2. $k[t]$-modules and the rational canonical form. It is finally time to specialize to the case in which $R = k$ is a *field* and F has finite rank; so $F = V$ is simply a finite-dimensional vector space. Let $n = \dim V$.

By the preceding discussion, choosing a linear transformation of V is the same as giving V a $k[t]$-module structure (compatible with its vector space structure); similar linear transformations correspond to isomorphic $k[t]$-modules. Then V is a finitely generated module over $k[t]$, and $k[t]$ is a PID since k is a field (Exercise III.4.4 and §V.2); therefore, we are precisely in the situation covered by the classification theorem.

Even before spelling out the result of applying Theorem 5.6, we can make use of its slogan version: every finitely generated module over a PID is a direct sum

of cyclic modules. This tells us that we can understand all linear transformations of finite-dimensional vector spaces, up to similarity, if we understand the linear transformation given by multiplication by t on a cyclic $k[t]$-module

$$V = \frac{k[t]}{(f(t))},$$

where $f(t)$ is a nonconstant monic polynomial:

$$f(t) = t^n + r_{n-1}t^{n-1} + \cdots + r_0.$$

It is worthwhile obtaining a good matrix representation of this poster case. We choose the basis

$$1, \quad t, \quad \cdots, \quad t^{n-1}$$

of V (cf. Proposition III.4.6). Recall that the columns of the matrix corresponding to a transformation consist of the images of the chosen basis (cf. the comments preceding Corollary 2.2). Since multiplication by t on V acts as

$$\begin{cases} 1 \mapsto t, \\ t \mapsto t^2, \\ \\ t^{n-1} \mapsto t^n = -r_{n-1}t^{n-1} - \cdots - r_0, \end{cases}$$

the matrix corresponding to this linear transformation is

$$\begin{pmatrix} 0 & 0 & 0 & 0 & -r_0 \\ 1 & 0 & 0 & 0 & -r_1 \\ 0 & 1 & 0 & 0 & -r_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & -r_{n-2} \\ 0 & 0 & 0 & 1 & -r_{n-1} \end{pmatrix}.$$

Definition 7.4. This is called the *companion matrix* of the polynomial $f(t)$, denoted $C_{f(t)}$. ⌟

Theorem 5.6 tells us (among other things) that every linear transformation admits a matrix representation into blocks, each of which is the companion matrix of a polynomial. Here is the statement of Theorem 5.6 in this context:

Theorem 7.5. *Let k be a field, and let V be a finite-dimensional vector space. Let α be a linear transformation on V, and endow V with the corresponding $k[t]$-module structure, as in Claim 7.1. Then the following hold:*

- *There exist distinct monic irreducible polynomials $p_1(t), \ldots, p_s(t) \in k[t]$ and positive integers r_{ij} such that*

$$V \cong \bigoplus_{i,j} \frac{k[t]}{(p_i(t)^{r_{ij}})}$$

as $k[t]$-modules.

- *There exist monic nonconstant polynomials* $f_1(t), \dots, f_m(t) \in k[t]$ *such that* $f_1(t) \mid \cdots \mid f_m(t)$ *and*

$$V \cong \frac{k[t]}{(f_1(t))} \oplus \cdots \oplus \frac{k[t]}{(f_m(t))}$$

as $k[t]$*-modules.*

Via these isomorphisms, the action of α *on* V *corresponds to multiplication by* t*.*

Further, two linear transformations α*,* β *are similar if and only if they have the same collections of invariants* $p_i(t)^{r_{ij}}$ *('elementary divisors'),* $f_i(t)$ *('invariant factors').*

Proof. Since $\dim V$ is finite, V is finitely generated as a k-module and *a fortiori* as a $k[t]$-module. The two isomorphisms are then obtained by applying Theorem 5.6. All the relevant polynomials may be chosen to be monic since every polynomial over a field is the associate of a (unique) monic polynomial (Exercise III.4.7). The fact that the action of α corresponds to multiplication by t is precisely what defines the corresponding $k[t]$-module structure on V. The statement about similar transformations follows from Corollary 7.3. □

Theorem 7.5 answers (over fields) the question raised in §6.1: we now have a list of invariants which describe completely the similarity class of a linear transformation. Further, these invariants provide us with a special matrix representation of a given linear transformation.

Definition 7.6. The *rational canonical form* of a linear transformation α of a vector space V is the block matrix

$$\left(\begin{array}{c|c|c} C_{f_1(t)} & & \\ \hline & \ddots & \\ \hline & & C_{f_m(t)} \end{array}\right),$$

where $f_1(t), \dots, f_m(t)$ are the invariant factors of α. ⌟

The rational canonical form of a square matrix is (of course) the rational canonical form of the corresponding linear transformation. The following statement is an immediate consequence of Theorem 7.5.

Corollary 7.7. *Every linear transformation admits a rational canonical form. Two linear transformations have the same rational canonical form if and only if they are similar.*

Remark 7.8. The 'rational' in rational canonical form has nothing to do with $\mathbb{Q}$; it is meant to remind the reader that this form can be found without leaving the base field. The other canonical form we will encounter will have entries in a possibly larger field, where the characteristic polynomial of the transformation factors completely. ⌟

Corollary 7.7 fulfills explicitly another wish expressed at the end of §6.1, that is, to find one distinguished representative in each similarity class of matrices/linear transformations. The rational canonical form is such a representative.

The next obvious question is how the invariants we just found relate to our more naive attempts to produce invariants of similar transformations in §6.

Proposition 7.9. *Let $f_1(t) \mid \cdots \mid f_m(t)$ be the invariant factors of a linear transformation α on a vector space V. Then the minimal polynomial $m_\alpha(t)$ equals $f_m(t)$, and the characteristic polynomial $P_\alpha(t)$ equals the product $f_1(t) \cdots f_m(t)$.*

Proof. Tracing definitions, $(m_\alpha(t))$ is the annihilator ideal of V when this is viewed as a $k[t]$-module via α (as in Claim 7.1). Therefore the equality of $m_\alpha(t)$ and $f_m(t)$ is a restatement of Lemma 5.7.

Concerning the characteristic polynomial, by Exercise 6.10 it suffices to prove the statement in the cyclic case: that is, it suffices to prove that if $f(t)$ is a monic polynomial, then $f(t)$ equals the characteristic polynomial of the companion matrix $C_{f(t)}$. Explicitly, let

$$f(t) = t^n + r_{n-1}t^{n-1} + \cdots + r_0$$

be a monic polynomial; then the task amounts to showing that

$$\det \begin{pmatrix} t & 0 & 0 & 0 & r_0 \\ -1 & t & 0 & 0 & r_1 \\ 0 & -1 & t & 0 & r_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & t & r_{n-2} \\ 0 & 0 & 0 & -1 & t + r_{n-1} \end{pmatrix} = f(t).$$

This can be done by an easy induction and is left to the reader (Exercise 7.2). □

Corollary 7.10 (Cayley-Hamilton). *The minimal polynomial of a linear transformation divides its characteristic polynomial.*

Proof. This has now become evident, as promised in §6.2. □

Finding the rational canonical form of a given matrix A amounts to working out the classification theorem for the specific $k[t]$-module corresponding to A as in Claim 7.1. As $k[t]$ is in fact a Euclidean domain, we know that this can be done by Gaussian elimination (cf. §2.4). In practice, the process consists of compiling the information obtained while diagonalizing $tI - A$ over $k[t]$ by elementary operations. The reader is invited to either produce an original algorithm or at least look it up in a standard reference to see what is involved.

Major shortcuts may come to our aid. For example, if the minimal and characteristic polynomials coincide, then one knows *a priori* that the corresponding module is cyclic (cf. Remark 5.8), and it follows that the rational canonical form is simply the companion matrix of the characteristic polynomial.

In any case, the strength of a canonical form rests in the fact that it allows us to reduce general facts to specific, standard cases. For example, the following statement is somewhat mysterious if all one knows are the definitions, but it becomes essentially trivial with a pinch of rational canonical forms:

Proposition 7.11. *Let $A \in \mathcal{M}_n(k)$ be a square matrix. Then A is similar to its transpose.*

Proof. If B is similar to A and we can prove that B is similar to its transpose B^t, then A is similar to its transpose A^t: because $B = PAP^{-1}$, $B^t = QBQ^{-1}$ give

$$A^t = (P^tQP)A(P^tQP)^{-1}.$$

Therefore, it suffices to prove the statement for matrices in rational canonical form.

Further, to prove the statement for a block matrix, it clearly suffices to prove it for each block; so we may assume that A is the companion matrix C of a polynomial $f(t)$. Since the characteristic and minimal polynomials of the transpose C^t coincide with those of C (Exercise 6.8), they are both equal to $f(t)$. It follows that the rational canonical form of C^t is again the companion matrix to $f(t)$; therefore C^t and C are similar, as needed. □

7.3. Jordan canonical form. We have derived the rational canonical form from the invariant factors of the module corresponding to a linear transformation. A useful alternative can be obtained from the elementary divisors, at least in the case in which *the characteristic polynomial factors completely* over the field k. If this is not the case, one can enlarge k so as to include all the roots of the characteristic polynomial: the reader proved this in Exercise V.5.13. The price to pay will be that the *Jordan canonical form* of a linear transformation $\alpha \in \mathrm{End}_k(V)$ may be a matrix with entries in a field larger than k. In any case, whether two transformations are similar or not is independent of the base field (Exercise 7.4), so this does not affect the issue at hand.

Given $\alpha \in \mathrm{End}_k(V)$, obtain the elementary divisor decomposition of the corresponding $k[t]$-module, as in Theorem 7.5:

$$V \cong \bigoplus_{i,j} \frac{k[t]}{(p_i(t)^{r_{ij}})}.$$

It is an immediate consequence of Proposition 7.9 and the equivalence between the elementary divisor and invariant factor formulations that the characteristic polynomial $P_\alpha(t)$ equals the product

$$\prod_{i,j} p_i(t)^{r_{ij}}.$$

Lemma 7.12. *Assume that the characteristic polynomial $P_\alpha(t)$ factors completely; that is,*

$$P_\alpha(t) = \prod_{i=1}^{s} (t - \lambda_i)^{m_i}$$

where λ_i, $i = 1, \ldots, s$, are the distinct eigenvalues of α (and m_i are their algebraic multiplicities; cf. §6.3). Then $p_i(t) = (t - \lambda_i)$, and $m_i = \sum_j r_{ij}$.

In this situation, the minimal polynomial of α *equals*

$$m_\alpha(t) = \prod_{i=1}^{s} (t - \lambda_i)^{\max_j\{r_{ij}\}}.$$

Proof. The first statement follows from uniqueness of factorizations. The statement about the minimal polynomial is immediate from Proposition 7.9 and the bookkeeping giving the equivalence of the two formulations in Theorem 7.5. □

The elementary divisor decomposition splits V into a different collection of cyclic modules than the decomposition in invariant factors: the basic cyclic bricks are now of the form $k[t]/(p(t)^r)$ for a monic prime $p(t)$; by Lemma 7.12, assuming that the characteristic polynomial factors completely over k, they are in fact of the form

$$\frac{k[t]}{((x-\lambda)^r)}$$

for some $\lambda \in k$ (which equals an eigenvalue of α) and $r > 0$. Just as we did for 'companion matrices', we now look for a basis with respect to which α (= multiplication by t; keep in mind the fine print in Theorem 7.5) has a particularly simple matrix representation. This time we choose the basis

$$(t-\lambda)^{r-1}, \quad (t-\lambda)^{r-2}, \quad \cdots, \quad (t-\lambda)^0 = 1.$$

Multiplication by t on V acts (in the first line, use the fact that $(t-\lambda)^r = 0$ in V) as

$$\begin{cases} (t-\lambda)^{r-1} \mapsto t(t-\lambda)^{r-1} = \lambda(t-\lambda)^{r-1} + (t-\lambda)^r = \lambda(t-\lambda)^{r-1}, \\ (t-\lambda)^{r-2} \mapsto t(t-\lambda)^{r-2} = (t-\lambda)^{r-1} + \lambda(t-\lambda)^{r-2}, \\ \\ \qquad 1 \mapsto t = (t-\lambda) + \lambda. \end{cases}$$

Therefore, with respect to this basis the linear transformation has matrix

$$\begin{pmatrix} \lambda & 1 & 0 & & 0 & 0 \\ 0 & \lambda & 1 & & 0 & 0 \\ 0 & 0 & \lambda & & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & & \lambda & 1 \\ 0 & 0 & 0 & & 0 & \lambda \end{pmatrix}.$$

Definition 7.13. This matrix is the *Jordan block* of size r corresponding to λ, denoted $J_{\lambda,r}$. ⌟

We can put several blocks together for a given λ: for $(r_j) = (r_1, \dots, r_\ell)$, let

$$J_{\lambda,(r_j)} := \left(\begin{array}{c|c|c} J_{\lambda,r_1} & & \\ \hline & \ddots & \\ \hline & & J_{\lambda,r_\ell} \end{array}\right).$$

With this notation in hand, we get new distinguished representatives of the similarity class of a given linear transformation:

Definition 7.14. The *Jordan canonical form* of a linear transformation α of a vector space V is the block matrix

$$\left(\begin{array}{c|c|c} J_{\lambda_1,(r_{1j})} & & \\ \hline & \ddots & \\ \hline & & J_{\lambda_s,(r_{sj})} \end{array}\right),$$

where $(t-\lambda_i)^{r_{ij}}$ are the elementary divisors of α (cf. Lemma 7.12). ⌟

This canonical form is 'a little less canonical' than the rational canonical form, in the sense that while the rational canonical form is *really* unique, some ambiguity is left in the Jordan canonical form: there is no special way to choose the order of the different blocks any more than there is a way to order[35] the eigenvalues $\lambda_1, \ldots, \lambda_s$.

Therefore, the analog of Corollary 7.7 should state that two linear transformations are similar if and only if they have the same Jordan canonical form up to a reordering of the blocks. As we have seen, every linear transformation $\alpha \in \mathrm{End}_k(V)$ admits a Jordan canonical form if $P_\alpha(t)$ factors completely over k; for example, we can always find a Jordan canonical form with entries in the algebraic closure of k.

Example 7.15. One use of the Jordan canonical form is the enumeration of all possible similarity classes of transformations with given eigenvalues. For example, there are 5 similarity classes of linear transformations with a single eigenvalue λ with algebraic multiplicity 4, over a 4-dimensional vector space: indeed, there are 5 different ways to stack together Jordan blocks corresponding to the same eigenvalue, within a 4×4 square matrix:

$$\left(\begin{array}{c|c|c|c} \lambda & 0 & 0 & 0 \\ \hline 0 & \lambda & 0 & 0 \\ \hline 0 & 0 & \lambda & 0 \\ \hline 0 & 0 & 0 & \lambda \end{array}\right), \quad \left(\begin{array}{cc|c|c} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ \hline 0 & 0 & \lambda & 0 \\ \hline 0 & 0 & 0 & \lambda \end{array}\right), \quad \left(\begin{array}{cc|cc} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ \hline 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{array}\right),$$

$$\left(\begin{array}{ccc|c} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 0 \\ \hline 0 & 0 & 0 & \lambda \end{array}\right), \quad \left(\begin{array}{cccc} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{array}\right).$$

⌟

The Jordan canonical form clarifies the difference between *algebraic* and *geometric* multiplicities of an eigenvalue. The algebraic multiplicity of λ as an eigenvalue of a linear transformation α is (of course) the number of times λ appears in the Jordan canonical form of α. Recall that the *geometric* multiplicity of λ equals the dimension of the eigenspace of λ.

Proposition 7.16. *The geometric multiplicity of λ as an eigenvalue of α equals the number of Jordan blocks corresponding to λ in the Jordan canonical form of α.*

[35] Of course if the ground field is $\mathbb{Q}$ or $\mathbb{R}$, then we can order the λ_i's in, for example, increasing order. However, this is not an option on fields like $\mathbb{C}$.

Proof. As the geometric multiplicity is clearly additive in direct sums, it suffices to show that the geometric multiplicity of λ for the transformation corresponding to a single Jordan block

$$J = \begin{pmatrix} \lambda & 1 & & 0 & 0 \\ 0 & \lambda & & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & \lambda & 1 \\ 0 & 0 & & 0 & \lambda \end{pmatrix}$$

is 1.

Let $\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_r \end{pmatrix}$ be an eigenvector corresponding to λ. Then

$$\lambda \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_r \end{pmatrix} = J \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_r \end{pmatrix} = \begin{pmatrix} \lambda v_1 + v_2 \\ \lambda v_2 + v_3 \\ \vdots \\ \lambda v_r \end{pmatrix},$$

yielding

$$v_2 = \cdots = v_r = 0.$$

That is, the eigenspace of λ is generated by

$$\mathbf{v} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and has dimension 1, as needed. □

7.4. Diagonalizability. A linear transformation of a vector space V is *diagonalizable* if it can be represented by a diagonal matrix.

The question of whether a given linear transformation is or is not diagonalizable is interesting and important: as we pointed out already in §6.3, when $\alpha \in \mathrm{End}_k(V)$ *is* diagonalizable, then the vector space V admits a corresponding *spectral decomposition:* α is diagonalizable if and only if V admits a basis of eigenvectors of α. Much more could be said on this important theme, but we are already in the position of giving useful criteria for diagonalizability.

For example, the diagonal matrix representing α will necessarily be a Jordan canonical form for α, consisting of blocks of size 1. Proposition 7.16 tells us that this can be detected by the difference between algebraic and geometric multiplicity, so we get the following characterization of diagonalizable transformations:

Corollary 7.17. *Assume the characteristic polynomial of $\alpha \in \mathrm{End}_k(V)$ factors completely over k. Then α is diagonalizable if and only if the geometric and algebraic multiplicities of all its eigenvalues coincide.*

Another view of the same result is the following:

Proposition 7.18. *Assume the characteristic polynomial of $\alpha \in \mathrm{End}_k(V)$ factors completely over k. Then α is diagonalizable if and only if the minimal polynomial of α has no multiple roots.*

Proof. Again, diagonalizability is equivalent to having all Jordan blocks of size 1 in the Jordan canonical form of α. Therefore, if the characteristic polynomial of α factors completely, then α is diagonalizable if and only if all exponents r_{ij} appearing in Theorem 7.5 equal 1. By the second part of Lemma 7.12, this is equivalent to the requirement that the minimal polynomial of α have no multiple roots. □

The condition of factorizability in these statements is necessary in order to guarantee that a Jordan canonical form exists. Not surprisingly, whether a matrix is diagonalizable or not depends on the ground field. For example,

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

is not diagonalizable over $\mathbb{R}$ (cf. Example 6.15), while it is diagonalizable over $\mathbb{C}$.

Endowing the vector space V with additional structure (such as an inner product) singles out important classes of operators, for which one can obtain precise and delicate results on the existence of spectral decompositions. For example, this can be used to prove that *real symmetric matrices are diagonalizable* (Exercise 7.20.) The reader will get a taste of these 'spectral theorems' in the exercises.

Exercises

As above, k denotes a *field.*

7.1. ▷ Complete the proof of Lemma 7.2. [§7.1]

7.2. ▷ Prove that the characteristic polynomial of the companion matrix of a monic polynomial $f(t)$ equals $f(t)$. [§7.2]

7.3. ▷ Prove that two linear transformations of a vector space of dimension ≤ 3 are similar if and only if they have the same characteristic and minimal polynomials. Is this true in dimension 4? [§6.2]

7.4. ▷ Let k be a field, and let K be a field containing k. Two square matrices $A, B \in \mathcal{M}_n(k)$ may be viewed as matrices with entries in the larger field K. Prove that A and B are similar over k if and only if they are similar over K. [§7.3]

7.5. Find the rational canonical form of a diagonal matrix

$$\begin{pmatrix} \lambda_1 & 0 & & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & \lambda_r \end{pmatrix},$$

assuming that the λ_i are all distinct.

7.6. Let

$$A = \begin{pmatrix} 6 & -10 & -10 \\ 3 & -5 & -6 \\ -1 & 2 & 3 \end{pmatrix}.$$

- Compute the characteristic polynomial of A.
- Find the minimal polynomial of A (use the Cayley-Hamilton theorem!).
- Find the invariant factors of A.
- Find the rational canonical form of A.

7.7. Let V be a k-vector space of dimension n, and let $\alpha \in \mathrm{End}_k(V)$. Prove that the minimal and characteristic polynomials of α coincide if and only if there is a vector $\mathbf{v} \in V$ such that

$$\mathbf{v}, \quad \alpha(\mathbf{v}), \quad \cdots, \quad \alpha^{n-1}(\mathbf{v})$$

is a basis of V.

7.8. Let V be a k-vector space of dimension n, and let $\alpha \in \mathrm{End}_k(V)$. Prove that the characteristic polynomial $P_\alpha(t)$ divides a power of the minimal polynomial $m_\alpha(t)$.

7.9. What is the number of distinct similarity classes of linear transformations on an n-dimensional vector space with one fixed eigenvalue λ with algebraic multiplicity n?

7.10. Classify all square matrices $A \in \mathcal{M}_n(k)$ such that $A^2 = A$, up to similarity. Describe the action of such matrices 'geometrically'.

7.11. A square matrix $A \in \mathcal{M}_n(k)$ is *nilpotent* (cf. Exercise V.4.19) if $A^k = 0$ for some integer k.

- Characterize nilpotent matrices in terms of their Jordan canonical form.
- Prove that if $A^k = 0$ for some integer k, then $A^k = 0$ for some integer k no larger than n (= the size of the matrix).
- Prove that the trace of a nilpotent matrix is 0.

7.12. ¬ Let V be a finite-dimensional k-vector space, and let $\alpha \in \mathrm{End}_k(V)$ be a diagonalizable linear transformation. Assume that $W \subseteq V$ is an invariant subspace, so that α induces a linear transformation $\alpha|_W \in \mathrm{End}_k(W)$. Prove that $\alpha|_W$ is also diagonalizable. (Use Proposition 7.18.) [7.14]

7.13. ¬ Let R be an integral domain. Assume that $A \in \mathcal{M}_n(R)$ is diagonalizable, with distinct eigenvalues. Let $B \in \mathcal{M}_n(R)$ be such that $AB = BA$. Prove that B is also diagonalizable, and in fact it is diagonal w.r.t. a basis of eigenvectors of A. (If P is such that PAP^{-1} is diagonal, note that PAP^{-1} and PBP^{-1} also commute.) [7.14]

7.14. Prove that 'commuting transformations may be simultaneously diagonalized', in the following sense. Let V be a finite-dimensional vector space, and let $\alpha, \beta \in \mathrm{End}_k(V)$ be diagonalizable transformations. Assume that $\alpha\beta = \beta\alpha$. Prove that V has a basis consisting of eigenvectors of both α and β. (Argue as in Exercise 7.13 to reduce to the case in which V is an eigenspace for α; then use Exercise 7.12.)

7.15. A *complete flag* of subspaces of a vector space V of dimension n is a sequence of nested subspaces

$$0 = V_0 \subsetneq V_1 \subsetneq \cdots \subsetneq V_{n-1} \subsetneq V_n = V$$

with $\dim V_i = i$. In other words, a complete flag is a composition series in the sense of Exercise 1.16.

Let V be a finite-dimensional vector space over an algebraically closed field. Prove that every linear transformation α of V preserves a complete flag: that is, there is a complete flag as above and such that $\alpha(V_i) \subseteq V_i$.

Find a linear transformation of $\mathbb{R}^2$ that does not preserve a complete flag.

7.16. *(Schur form)* Let $A \in \mathcal{M}_n(\mathbb{C})$. Prove that there exists a *unitary* matrix $P \in \mathrm{U}_n(\mathbb{C})$ such that

$$A = P \begin{pmatrix} \lambda_1 & * & * & & * \\ 0 & \lambda_2 & * & & * \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & & \lambda_{n-1} & * \\ 0 & 0 & \dots & 0 & \lambda_n \end{pmatrix} P^{-1}.$$

So not only is A similar to an upper triangular matrix (this is already guaranteed by the Jordan canonical form), but a matrix of change of basis P 'triangularizing' the matrix A can in fact be chosen to be in $\mathrm{U}_n(\mathbb{C})$. (Argue inductively on n. Since $\mathbb{C}$ is algebraically closed, A has at least one eigenvector $\mathbf{v}_1$, and we may assume $(\mathbf{v}_1, \mathbf{v}_1) = 1$. For the induction step, consider the complement $\mathbf{v}_1^{\perp}$; cf. Exercise 6.17, and keep in mind Exercise 6.18.)

The upper triangular matrix is a *Schur form* for A. It is (of course) not unique.

7.17. ¬ A matrix $M \in \mathcal{M}_n(\mathbb{C})$ is *normal* if $MM^\dagger = M^\dagger M$. Note that unitary matrices and hermitian matrices are both normal.

Prove that triangular normal matrices are diagonal. [7.18]

7.18. ¬ Prove the *spectral theorem for normal operators:* if M is a normal matrix, then there exists an orthonormal basis of eigenvectors of M. Therefore, normal operators are diagonalizable; they determine a spectral decomposition of the vector space on which they act. (Consider the Schur form of M; cf. Exercise 7.16. Use Exercise 7.17.) [7.19, 7.20]

7.19. Prove that a matrix $M \in \mathcal{M}_n(\mathbb{C})$ is normal *if and only if* it admits an orthonormal basis of eigenvectors. (Exercise 7.18 gives one direction; prove the converse.)

7.20. ▷ Prove that real symmetric matrices are diagonalizable and admit an orthonormal basis of eigenvectors. (This is an immediate consequence of Exercise 7.18 and Exercise 6.22.) [§7.4]

Fields

The minuscule amount of algebra we have developed so far allows us to scratch the surface of the unfathomable subject of *field theory*. Fields are of basic importance in many subjects—number theory and algebraic geometry come immediately to mind—and it is not surprising that their study has been developed to a particularly high level of sophistication. While our profoundly limited knowledge will prevent us from even hinting at this sophistication, even a cursory overview of the basic notions will allow us to deal with remarkable applications, such as the famous problems of constructibility of geometric figures. We will also get a glimpse of the beautiful subject of *Galois theory*, an amazing interplay of the theory of field extensions, the solvability of polynomials, and group theory.

1. Field extensions, I

We first deal with several basic notions in field theory, mostly inspired by easy linear algebra considerations. The keywords here are *finite, simple, finitely generated, algebraic.*

1.1. Basic definitions. The study of fields is (of course) the study of the category Fld of fields (Definition III.1.14), with ring homomorphisms as morphisms. The task is to understand what fields are and above all what they are *in relation to one another.*

The place to start is a reminder from elementary ring theory[1]: *every ring homomorphisms from a field to a nonzero ring is injective.* Indeed, the kernel of a ring homomorphism to a nonzero ring is a proper ideal (because a homomorphism maps 1 to 1 by definition and $1 \neq 0$ in nonzero rings), and the only proper ideal in a field is (0). In particular, every ring homomorphism of fields is injective (fields are nonzero rings by definition!); every morphism in Fld is a monomorphism (cf. Proposition III.2.4).

[1]The last time we made this point was back in §V.5.2.

Thus, every morphism $k \to K$ between two fields identifies the first with a subfield of the second. In other words, one can view K as a particular way to enlarge k, that is, as an *extension* of k. Field theory is first and foremost the study of *field extensions.* We will denote a field extension by $k \subseteq K$ (which is less than optimal, because there may be many ways to embed a field into another); other popular choices are K/k (which we do not like, as it suggests a quotienting operation) and

$$\begin{array}{c} K \\ | \\ k \end{array}$$

(which we will avoid most of the time, as it is hard to typeset).

The coarsest invariant of a field k is its *characteristic;* cf. Exercise V.4.17. We have a unique ring homomorphisms $i : \mathbb{Z} \to k$ ($\mathbb{Z}$ is initial in Ring: 1 must go to 1 by definition of homomorphism, and this fixes the value of $i(n)$ for all $n \in \mathbb{Z}$); the *characteristic* of k, $\operatorname{char} k$, is defined to be the nonnegative generator of the ideal $\ker i$; that is, $\operatorname{char} k = 0$ if i is injective, and $\operatorname{char} k = p > 0$ if $\ker i = (p) \neq (0)$.

Put otherwise, either $n \cdot 1$ is only 0 in k for $n = 0$, in which case $\operatorname{char} k = 0$, or $n \cdot 1 = 0$ for some $n \neq 0$; $\operatorname{char} k = p > 0$ is then the smallest positive integer for which $p \cdot 1 = 0$ in k. Since fields are integral domains, the image $i(\mathbb{Z})$ must be an integral domain; hence $\ker i$ must be a prime ideal. Therefore, the characteristic of a field is either 0 or a prime number.

If $k \subseteq K$ is an extension, then $\operatorname{char} k = \operatorname{char} K$ (Exercise 1.1). Thus, we could define and study categories Fld_0, Fld_p of fields of a given characteristic, without losing any information: these categories live within Fld, without interacting with each other[2]. Further, each of these categories has an initial object: $\mathbb{Q}$ is initial in Fld_0, and[3] $\mathbb{F}_p := \mathbb{Z}/p\mathbb{Z}$ in Fld_p; the reader has checked this easy fact in Exercise V.4.17, and now is an excellent time to contemplate it again. In each case, the initial object is called the *prime subfield* of the given field. Thus, every field is in a canonical way an extension of its prime subfield: studying fields really means studying field extensions. To a large extent, the 'small' field k will be fixed in what follows, and our main object of study will be the category Fld_k of extensions of k, with the evident (and by now familiar to the reader) notions of morphisms.

The first general remark concerning field extensions is that the larger field is an algebra, and hence a *vector space* over the smaller one (by the very definition of algebra; cf. Example III.5.6).

Definition 1.1. A field extension $k \subseteq F$ is *finite,* of *degree* n, if F has (finite) dimension $\dim F = n$ as a vector space over k. The extension is *infinite* otherwise.

The degree of a finite extension $k \subseteq F$ is denoted by $[F : k]$ (and we write $[F : k] = \infty$ if the extension is infinite). ⌟

[2]They do interact in other ways, however—for example, because $\mathbb{Z}$ is initial in Ring, so $\mathbb{Z}$ maps to all fields.

[3]We are going to denote the field $\mathbb{Z}/p\mathbb{Z}$ by $\mathbb{F}_p$ in this chapter, to stress its role as a field (rather than 'just' as a group).

In §V.5.2 we have encountered a prototypical example of finite field extension: a procedure starting from an irreducible polynomial $f(x) \in k[x]$ with coefficients in a field and producing an extension K of k in which $f(t)$ has a *root.* Explicitly,

$$K = \frac{k[t]}{(f(t))}$$

is such an extension (Proposition V.5.7). The quotient is a field because $k[t]$ is a PID; hence irreducible elements generate maximal ideals—*prime* because of the UFD property (cf. Theorem V.2.5) and then *maximal* because nonzero prime ideals are maximal in a PID (cf. Proposition III.4.13). The coset of t is a root of $f(x)$ when this is viewed as an element of $K[x]$. The degree of the extension $k \subseteq K$ equals the degree of the polynomial $f(x)$ (this is hopefully clear at this point and was checked carefully in Proposition III.4.6).

The reader may wonder whether perhaps *all* finite extensions are of this type. This is unfortunately not the case, but as it happens, it *is* the case in a large class of examples. As we often do, we promise that the situation will clarify itself considerably once we have accumulated more general knowledge; in this case, the relevant fact will be Proposition 5.19.

Also recall that we proved that these extensions are almost universal with respect to the problem of extending k so that the polynomial $f(t)$ acquires a root. We pointed out, however, that the *uni* in *uni*versal is missing; that is, if $k \subseteq F$ is an extension of k in which $f(t)$ has a root, there may be many different ways to put K in between:

$$k \subseteq K \subseteq F.$$

Such questions are central to the study of extensions, so we begin by giving a second look at this situation.

1.2. Simple extensions. Let $k \subseteq F$ be a field extension, and let $\alpha \in F$. The smallest subfield of F containing both k and α is denoted $k(\alpha)$; that is, $k(\alpha)$ is the intersection of all subfields of F containing k and α.

Definition 1.2. A field extension $k \subseteq F$ is *simple* if there exists an element $\alpha \in F$ such that $F = k(\alpha)$. ⌟

The extensions recalled above are of this kind: if $K = k[t]/(f(t))$ and α denotes the coset of t, then $K = k(\alpha)$: indeed, if a subfield of K contains the coset of t, then it must contain (the coset of) every polynomial expression in t, and hence it must be the whole of K.

We have always found the notation $k(\alpha)$ somewhat unfortunate, since it suggests that all such extensions are in some way isomorphic and possibly all isomorphic to the field $k(t)$ of rational functions in one indeterminate t (cf. Example V.4.12). This is not true, although it is clear that every element of $k(\alpha)$ may be written as a rational function in α with coefficients in k (Exercise 1.3). In any case, it is easy to classify simple extensions: they are either isomorphic to $k(t)$ or they are of the prototypical kind recalled above. Here is the precise statement.

Proposition 1.3. *Let $k \subseteq k(\alpha)$ be a simple extension. Consider the evaluation map $\epsilon : k[t] \to k(\alpha)$, defined by $f(t) \mapsto f(\alpha)$. Then we have the following:*

- *ϵ is injective if and only if $k \subseteq k(\alpha)$ is an infinite extension. In this case, $k(\alpha)$ is isomorphic to the field of rational functions $k(t)$.*
- *ϵ is not injective if and only if $k \subseteq k(\alpha)$ is finite. In this case there exists a unique monic irreducible nonconstant polynomial $p(t) \in k[t]$ of degree $n = [k(\alpha) : k]$ such that*

$$k(\alpha) \cong \frac{k[t]}{(p(t))}.$$

Via this isomorphism, α corresponds to the coset of t. The polynomial $p(t)$ is the monic polynomial of smallest degree in $k[t]$ such that $p(\alpha) = 0$ in $k(\alpha)$.

The polynomial $p(t)$ appearing in the first part of the statement is called the *minimal polynomial* of α over k.

Of course the minimal polynomial of an element α of a ('large') field depends on the base ('small') field k. For example, $\sqrt{2} \in \mathbb{C}$ has minimal polynomial $t^2 - 2$ over $\mathbb{Q}$, but $t - \sqrt{2}$ over $\mathbb{R}$.

Proof. Let $F = k(\alpha)$. By the 'first isomorphism theorem', the image of $\epsilon : k[t] \to F$ is isomorphic to $k[t]/\ker(\epsilon)$. Since F is an integral domain, so is $k[t]/\ker(\epsilon)$; hence $\ker(\epsilon)$ is a prime ideal in $k[t]$.

—Assume $\ker(\epsilon) = 0$; that is, ϵ is an injective map from the integral domain $k[t]$ to the field F. By the universal property of fields of fractions (cf. §V.4.2), ϵ extends to a unique homomorphism

$$k(t) \to F.$$

The (isomorphic) image of $k(t)$ in F is a field containing k and α; hence it equals F by definition of simple extension.

Since ϵ is injective, the powers $\alpha^0 = 1, \alpha, \alpha^2, \alpha^3, \ldots$ (that is, the images $\epsilon(t^i)$) are all distinct and linearly independent over k (because the powers $1, t, t^2, \ldots$ are linearly independent over k); therefore the extension $k \subseteq F$ is infinite in this case.

—If $\ker(\epsilon) \neq 0$, then $\ker(\epsilon) = (p(t))$ for a unique monic irreducible nonconstant polynomial $p(t)$, which has smallest degree (cf. Exercise III.4.4!) among all nonzero polynomials in $\ker(\epsilon)$. As $(p(t))$ is then maximal in $k[t]$, the image of ϵ is a subfield of F containing $\alpha = \epsilon(t)$. By definition of simple extension, $F =$ the image of ϵ; that is, the induced homomorphism

$$\frac{k[t]}{(p(t))} \to F$$

is an isomorphism. In this case $[F : k] = \deg p(t)$, as recalled in §1.1, and in particular the extension is finite, as claimed. □

The alert reader will have noticed that the proof of Proposition 1.3 is essentially a rehash of the argument proving the 'versality' part of Proposition V.5.7.

Example 1.4. Consider the extension $\mathbb{Q} \subseteq \mathbb{R}$.

The polynomial $x^2 - 2 \in \mathbb{Q}[x]$ has roots in $\mathbb{R}$: therefore, by Proposition V.5.7 there exists a homomorphism (hence a field extension)

$$\overline{\epsilon}: \quad \frac{\mathbb{Q}[t]}{(t^2-2)} \hookrightarrow \mathbb{R},$$

such that the image of (the coset of) t is a root α of $x^2 - 2$. Proposition 1.3 simply identifies the image of this homomorphism with $\mathbb{Q}(\alpha) \subseteq \mathbb{R}$.

This is hopefully crystal clear; however, note that even this simple example shows that the induced morphism $\overline{\epsilon}$ is *not unique* (hence the 'lack of uni'): because *there are more than one root of* $x^2 - 2$ *in* $\mathbb{R}$. Concretely, there are *two* possible choices for α: $\alpha = +\sqrt{2}$ and $\alpha = -\sqrt{2}$. The choice of α determines the evaluation map ϵ and therefore the specific realization of $\mathbb{Q}[t]/(t^2 - 2)$ as a subfield of $\mathbb{R}$.

The reader may be misled by one feature of this example: clearly $\mathbb{Q}(\sqrt{2}) = \mathbb{Q}(-\sqrt{2})$, and this may seem to compensate for the lack of uniqueness lamented above. The *morphism* may not be unique, but the *image* of the morphism surely is? No. The reader will check that there are three *distinct* subfields of $\mathbb{C}$ isomorphic to $\mathbb{Q}[t]/(t^3 - 2)$ (Exercise 1.5).

Ay, there's the rub. One of our main goals in this chapter will be to single out a condition on an extension $k \subseteq F$ that guarantees that no matter how we embed F in a larger extension, the images of these (possibly many different) embeddings will all coincide. Up to further technicalities, this is what makes an extension *Galois*. Thus, $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2})$ will be a Galois extension, while $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2})$ will *not* be a Galois extension. But Galois extensions will have to wait until §6, and the reader can put this issue aside for now. ⌟

One way to recover uniqueness is to incorporate the choice of the root in the data. This leads to the following refinement of the (uni)versality statement, adapted for future applications.

Proposition 1.5. *Let $i : k_1 \subseteq F_1 = k_1(\alpha_1)$, $k_2 \subseteq F_2 = k_2(\alpha_2)$ be two finite simple extensions. Let $p_1(t) \in k_1[t]$, resp., $p_2(t) \in k_2[t]$, be the minimal polynomials of α_1, resp., α_2. Let $i : k_1 \to k_2$ be an isomorphism, such that*[4]

$$i(p_1(t)) = p_2(t).$$

Then there exists a unique isomorphism $j : F_1 \to F_2$ agreeing with i on k_1 and such that $j(\alpha_1) = \alpha_2$.

Proof. Since every element of $k_1(\alpha_1)$ is a linear combination of powers of α_1 with coefficients in k_1, j is determined by its action on k_1 (which agrees with i) and by $j(\alpha_1)$, which is prescribed to be α_2. Thus an isomorphism j as in the statement is uniquely determined.

To see that the isomorphism j exists, note that as i maps $p_1(t)$ to $p_2(t)$, it induces an isomorphism

$$\frac{k_1[t]}{(p_1(t))} \xrightarrow{\sim} \frac{k_2[t]}{(p_2(t))}.$$

[4]Of course any homomorphism of rings $f : R \to S$ induces a unique ring homomorphism $f : R[t] \to S[t]$ sending t to t, named in the same way by a harmless abuse of language.

Composing with the isomorphisms found in Proposition 1.3 gives j:

$$j : k_1(\alpha_1) \xrightarrow{\sim} \frac{k_1[t]}{(p_1(t))} \xrightarrow{\sim} \frac{k_2[t]}{(p_2(t))} \xrightarrow{\sim} k_2(\alpha_2),$$

as needed. □

Thus, isomorphisms lift uniquely to simple extensions, so long as they preserve minimal polynomials. We may say that j *extends* i, and draw the following diagram of extensions:

$$\begin{array}{ccc} k_1(\alpha_1) & \xrightarrow[\sim]{j} & k_2(\alpha_2) \\ | & & | \\ k_1 & \xrightarrow[\sim]{i} & k_2 \end{array}$$

We will be particularly interested in considering isomorphisms of a field *to itself,* subject to the condition of fixing a specified subfield k (that is, extending the identity on k), that is, *automorphisms* in Fld_k. Explicitly, for $k \subseteq F$ an extension, we will analyze isomorphisms $j : F \xrightarrow{\sim} F$ such that $\forall c \in k, j(c) = c$.

We will denote the group of such automorphisms by $\mathrm{Aut}_k(F)$. *This is probably the most important object of study in this chapter,* so we should make its definition official:

Definition 1.6. Let $k \subseteq F$ be a field extension. The *group of automorphisms* of the extension, denoted $\mathrm{Aut}_k(F)$, is the group of field automorphisms $j : F \to F$ such that $j|_k = \mathrm{id}_k$. ⌟

Corollary 1.7. *Let $k \subseteq F = k(\alpha)$ be a simple finite extension, and let $p(x)$ be the minimal polynomial of α over k. Then $|\mathrm{Aut}_k(F)|$ equals the number of distinct roots of $p(x)$ in F; in particular,*

$$|\mathrm{Aut}_k(F)| \leq [F : k],$$

with equality if and only if $p(x)$ factors over F as a product of distinct linear polynomials.

Proof. Let $j \in \mathrm{Aut}_k(F)$. Since every element of F is a polynomial expression in α with coefficients in k, and j extends the identity on k, j is determined by $j(\alpha)$. Now

$$p(j(\alpha)) = j(p(\alpha)) = j(0) = 0 :$$

therefore, $j(\alpha)$ is necessarily a root of $p(x)$. This shows that $|\mathrm{Aut}_k(F)|$ is no larger than the number of roots of $p(x)$.

On the other hand, by Proposition 1.5 every choice of a root of $p(t)$ determines a lift of the identity to an element $j \in \mathrm{Aut}_k(F)$, establishing the other inequality. □

Corollary 1.7 is our first hint of the powerful interaction between group theory and the theory of field extensions; this theme will haunt us throughout the chapter.

The proof of Corollary 1.7 really sets up a bijection between the roots of an irreducible polynomial $p(t) \in k[t]$ in a larger field F and the *group* $\mathrm{Aut}_k(F)$. A particularly optimistic reader may then hope that the roots of $p(t)$ come endowed

with a *group structure,* but this is hoping for too much: the bijection depends on the choice of one root α, and no one root of $p(t)$ is 'more beautiful' than the others. However, we have established that *if $F = k(\alpha)$ is a simple extension of k, then the group $\mathrm{Aut}_k(F)$ acts faithfully and transitively on the set of roots of $p(t)$ in F.* In other words, $\mathrm{Aut}_k(F)$ can be identified with a certain subgroup of the symmetric group acting on the set of roots of $p(t)$.

More generally (Exercise 1.6) the automorphisms of any extension act on roots of polynomials. One conclusion we can draw from these preliminary considerations is that the analysis of $\mathrm{Aut}_k(F)$ is substantially simplified if $k \subseteq F$ is a *simple* extension $k(\alpha)$ and further if the minimal polynomial $p(x)$ of α factors into $\deg p(x)$ distinct linear terms in F. It will not come as a surprise that we will focus our attention on such extensions in later sections: an extension is *Galois* precisely if it is of this type. (The reader is invited to remember this fact, but its import will not be fully appreciated until we develop substantially more material.)

1.3. Finite and algebraic extensions. The dichotomy encountered in Proposition 1.3 provides us with one of the most important definitions in field theory:

Definition 1.8. Let $k \subseteq F$ be a field extension, and let $\alpha \in F$. Then α is *algebraic over k, of degree n* if $n = [k(\alpha) : k]$ is finite; α is *transcendantal over k* otherwise.

The extension $k \subseteq F$ is *algebraic* if every $\alpha \in F$ is algebraic over k. ⌟

By Proposition 1.3, $\alpha \in F$ is algebraic over k if and only if there exists a nonzero polynomial $f(x) \in k[x]$ such that $f(\alpha) = 0$. The minimal polynomial of α is the monic polynomial of smallest degree satisfying this condition; as we have seen, it is necessarily irreducible.

Also note that if α is algebraic over k, then every element of $k(\alpha)$ may in fact be written as a *polynomial* with coefficients in k.

Finite extensions are necessarily *algebraic:*

Lemma 1.9. *Let $k \subseteq F$ be a finite extension. Then every $\alpha \in F$ is algebraic over k, of degree $\leq [F : k]$.*

Proof. Since $k \subseteq k(\alpha) \subseteq F$, the dimension of $k(\alpha)$ as a k-vector space is bounded by $\dim_k F = [F : k]$. □

Concretely, if $k \subseteq F$ is finite and $\alpha \in F$, then the powers $1, \alpha, \alpha^2 \ldots$ are necessarily linearly *de*pendent; and any nontrivial linear dependence relation among them provides us with a nonzero polynomial $f(x) \in k[x]$ such that $f(\alpha) = 0$.

The literal converse of the claim that 'finite extensions are algebraic' is not true; we will see an example in a moment. However, something along these lines holds; understanding the situation requires a more careful look at finiteness conditions.

First of all, compositions of finite extensions are finite extensions, and the degree behaves nicely with respect to this operation:

Proposition 1.10. *Let $k \subseteq E \subseteq F$ be field extensions. Then $k \subseteq F$ is finite if and only if both $k \subseteq E$ and $E \subseteq F$ are finite. In this case,*

$$[F : k] = [F : E][E : k].$$

Proof. If F is finite-dimensional as a vector space over k, then so is its subspace E; and any linear dependence relation of elements of F over the field k gives one over the larger field E. It follows that if $k \subseteq F$ is finite, then so are $k \subseteq E$ and $E \subseteq F$.

Conversely, assume $k \subseteq E$ and $E \subseteq F$ are both finite. Let $(e_1, \ldots, e_m)$ be a basis for E over k, and let $(f_1, \ldots, f_n)$ be a basis for F over E. It will suffice to show that the mn products

$$(e_1 f_1, e_1 f_2, \ldots, e_m f_n)$$

form a basis for F over k.

Let $g \in F$. Then $\exists d_1, \ldots, d_n \in E$ such that

$$g = \sum_{j=1}^{n} d_j f_j,$$

since the elements f_j span F over E. Since E is spanned by the elements e_i over k, $\exists c_{1j}, \ldots, c_{mj} \in k$ such that $\forall j$

$$d_j = \sum_{i=1}^{m} c_{ij} e_i.$$

It follows that

$$g = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} e_i f_j :$$

therefore the products $e_i f_j$ span F over k. (This suffices to prove that $k \subseteq F$ is finite.)

To verify that these elements are linearly independent, assume

$$\sum_{i,j} \lambda_{ij} e_i f_j = 0$$

for $\lambda_{ij} \in k$. Then we have

$$\sum_j (\sum_i \lambda_{ij} e_i) f_j = 0,$$

implying $\sum_i \lambda_{ij} e_i = 0$ for all j, since the elements f_j are linearly independent over E. As the elements e_i are linearly independent over k, this shows that all λ_{ij} equal 0, as needed. □

The formula given in the statement should look familiar to the reader: glance at the end of §II.8.5. Of course this is not accidental; the reader should take it as another hint that the world of groups and the world of fields have deep interaction. As in the case of groups, we draw the immediate (but powerful) consequence, reminiscent of Lagrange's theorem:

Corollary 1.11. *Let $k \subseteq F$ be a finite extension, and let E be an intermediate field (that is, $k \subseteq E \subseteq F$). Then both $[E : k]$ and $[F : E]$ divide $[F : k]$.*

As with Lagrange's theorem, judicious use of this result will make otherwise mysterious statements melt into trivialities.

Example 1.12. Let $k \subseteq F$ be a field extension, and let $\alpha \in F$ be an algebraic element over k, of *odd* order. Then we claim that α may be written as a polynomial in α^2, with coefficients in k.

Indeed, $k(\alpha^2)$ is intermediate between k and $k(\alpha)$:

$$k \subseteq k(\alpha^2) \subseteq k(\alpha);$$

what can we say about the degree d of $k(\alpha)$ over $k(\alpha^2)$? Since α satisfies the polynomial $t^2 - \alpha^2 \in k(\alpha^2)[t]$, $d \leq 2$. On the other hand, d divides $[k(\alpha) : k]$ by Corollary 1.11, and $[k(\alpha) : k]$ is odd, so $d \neq 2$. Therefore $d = 1$, proving $k(\alpha) = k(\alpha^2)$, and in particular $\alpha \in k(\alpha^2)$, which is the claim. ⌟

Here is something else that should evoke fond memories for the reader: back in §III.6.5 we had encountered an important distinction between *finite* algebras and algebras *of finite type.* An algebra is *finite* over the base ring if it is finitely generated as a module, that is, if it admits an onto homomorphism (of modules) from a finitely generated free module; a commutative algebra is *of finite type* if it admits an onto homomorphism (of algebras) from a polynomial ring in finitely many variables.

Something along the same lines is going to occur here. An extension $k \subseteq F$ is finite if and only if $\dim_k F$ is finite, that is, if and only if F is a finite k-algebra. The other finiteness condition takes the following form.

Definition 1.13. A field extension $k \subseteq F$ is *finitely generated* if there exist $\alpha_1, \ldots, \alpha_n \in F$ such that

$$F = k(\alpha_1)(\alpha_2)\ldots(\alpha_n).$$ ⌟

Remark 1.14. This is not literally the same as saying that F is a finite-type k-algebra: even in the case of simple extensions, if α is *transcendental,* then the natural evaluation map $\epsilon : k[t] \to k(\alpha)$ of Proposition 1.3 is not onto. However, this is an *epimorphism* of rings in the categorical sense, as the reader will check (Exercise 1.17); thus, the 'categorical spirit' behind the finite-type condition is preserved in this context.

This subtlety raises an interesting question: what if a field extension $k \subseteq F$ is *really* a finite-type k-algebra? This turns out to be an important issue, and we will come back to it in §2.2. ⌟

We will use the notation

$$k(\alpha_1, \alpha_2, \ldots, \alpha_n)$$

for the field $k(\alpha_1)(\alpha_2)\ldots(\alpha_n)$. For $k \subseteq F$ and $\alpha_1, \ldots, \alpha_n \in F$, the elements of the field $F = k(\alpha_1, \ldots, \alpha_n)$ are all the elements of F which may be written as rational functions in the α_i's, with coefficients in k (cf. Exercise 1.3). Put otherwise, $k(\alpha_1, \ldots, \alpha_n)$ is the smallest subfield of F containing $k(\alpha_1), \ldots, k(\alpha_n)$ (it is the *composite* of these subfields). It is clear that the order of the elements α_i is irrelevant.

Coming back to the issue of finite vs. algebraic, these two notions do coincide for *finitely generated* extensions.

Proposition 1.15. *Let $k \subseteq F = k(\alpha_1, \ldots, \alpha_n)$ be a finitely generated field extension. Then the following are equivalent:*

(i) $k \subseteq F$ is a finite extension.

(ii) $k \subseteq F$ is an algebraic extension.

(iii) Each α_i is algebraic over k.

If these conditions are satisfied, then $[F : k] \leq$ the product of the degrees of α_i over k.

Proof. Lemma 1.9 shows that (i) $\implies$ (ii); (ii) $\implies$ (iii) trivially. Thus, we only need to prove that (iii) $\implies$ (i), and to bound the degree of F over k in the process.

Assume that each α_i is algebraic over k, and let d_i be the degree of α_i over k. By definition, $k \subseteq k(\alpha_i)$ is finite, of degree d_i. By Exercise 1.16, each extension

$$k(\alpha_1, \ldots, \alpha_{i-1}) \subseteq k(\alpha_1, \ldots, \alpha_i)$$

is finite, of degree $\leq d_i$. Applying Proposition 1.10 to the composition of extensions

$$k \subseteq k(\alpha_1) \subseteq k(\alpha_1, \alpha_2) \subseteq \cdots \subseteq k(\alpha_1, \ldots, \alpha_n) = F$$

proves that $k \subseteq F$ is finite and $[F : k] \leq d_1 \cdots d_n$, as needed. □

While rather straightforward, Proposition 1.15 has always seemed remarkable to us: it says (in particular) that if α and β are algebraic over a field k, then so are $\alpha \pm \beta$, $\alpha\beta$, $\alpha\beta^{-1}$, and any other rational function of α and β. If all we knew were the simple-minded definition of 'algebraic' (that is, 'root of a polynomial'), this would seem rather mysterious: given a polynomial $f(x)$ of which α is a root and a polynomial $g(x)$ of which β is a root, how do we construct a polynomial $h(x)$ such that (for example) $h(\alpha + \beta) = 0$? The answer is that we do not need to perform any such construction, by virtue of Proposition 1.15.

One immediate consequence of this observation is that the set of algebraic elements of *any* extension forms a field:

Corollary 1.16. *Let $k \subseteq F$ be a field extension. Let*

$$E = \{\alpha \in F \mid \alpha \text{ is algebraic over } k\}.$$

Then E is a field.

Example 1.17. Let $\overline{\mathbb{Q}} \subseteq \mathbb{C}$ be the set of complex numbers that are algebraic over $\mathbb{Q}$; then $\overline{\mathbb{Q}}$ is a field, by Corollary 1.16, and the extension $\mathbb{Q} \subseteq \overline{\mathbb{Q}}$ is (tautologically) algebraic. Note that $\mathbb{Q} \subseteq \overline{\mathbb{Q}}$ is *not* a finite extension, because in it there are elements of arbitrarily high degree over $\mathbb{Q}$: indeed, there exist irreducible polynomials in $\mathbb{Q}[x]$ of arbitrarily high degree, as we observed in Corollary V.5.16.

Elements of $\overline{\mathbb{Q}}$ are called *algebraic* numbers; the complex numbers in the complement of $\overline{\mathbb{Q}}$ are transcendental over $\mathbb{Q}$ (by definition); they are simply called *transcendental* numbers. Elementary cardinality considerations show that $\overline{\mathbb{Q}}$ is countable; thus, the set of transcendental numbers is uncountable. Surprisingly, it may be substantially difficult to prove that a given number is transcendental: for example, the fact that e is transcendental was only proved around 1870, by Charles Hermite. The number π is transcendental (Lindemann, ca. 1880); e^π is

transcendental (Gelfond-Schneider, 1934). No one knows whether π^e, $\pi + e$, or πe are transcendental. ⌟

Another consequence of Proposition 1.15 is the important fact that *compositions of algebraic extensions are algebraic* (whether finitely generated or not):

Corollary 1.18. *Let $k \subseteq E \subseteq F$ be field extensions. Then $k \subseteq F$ is algebraic if and only if both $k \subseteq E$ and $E \subseteq F$ are algebraic.*

Proof. If $k \subseteq F$ is algebraic, then every element of F is algebraic over k, hence over E, and every element of E is algebraic over k; thus $E \subseteq F$ and $k \subseteq E$ are algebraic.

Conversely, assume $k \subseteq E$ and $E \subseteq F$ are both algebraic, and let $\alpha \in F$. Since α is algebraic over E, there exists a polynomial

$$f(x) = x^n + e_{n-1}x^{n-1} + \cdots + e_0 \in E[x]$$

such that $f(\alpha) = 0$. This implies that in fact α is 'already' algebraic over the subfield $k(e_0, \ldots, e_{n-1}) \subseteq E$; therefore,

$$k(e_0, \ldots, e_{n-1}) \subseteq k(e_0, \ldots, e_{n-1}, \alpha)$$

is a finite extension. On the other hand,

$$k \subseteq k(e_0, \ldots, e_{n-1})$$

is a finite extension by Proposition 1.15, since each e_i is in E and therefore algebraic over k. By Proposition 1.10

$$k \subseteq k(e_0, \ldots, e_{n-1}, \alpha)$$

is finite. This implies that α is algebraic over k, as needed, by Lemma 1.9. □

We will essentially deal only with finitely generated extensions, and Proposition 1.15 will simplify our work considerably. Also, since finitely generated extensions are compositions of simple extensions, the reader should expect that we will give a careful look at automorphisms of such extensions.

It may in fact come as a surprise that, in many cases, finitely generated extensions turn out to be simple to begin with; we will prove a precise statement to this effect in due time (Proposition 5.19). The reader already has enough tools to contemplate easy (but interesting) examples, such as the following. This should serve as an encouragement to look at many more.

Example 1.19. Consider the extension $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2}, \sqrt{3})$.

—By Proposition 1.15 we know that this is a finite (hence algebraic) extension, of degree at most 4.

—Thus any five elements in $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ must be linearly dependent over $\mathbb{Q}$. We consider powers of $\sqrt{2} + \sqrt{3}$:

$$1, \quad (\sqrt{2} + \sqrt{3}), \quad (\sqrt{2} + \sqrt{3})^2, \quad (\sqrt{2} + \sqrt{3})^3, \quad (\sqrt{2} + \sqrt{3})^4$$

must be linearly dependent. Thus, there must be rational numbers $q_0, \ldots, q_3$ such that

$$(\sqrt{2} + \sqrt{3})^4 + q_3(\sqrt{2} + \sqrt{3})^3 + q_2(\sqrt{2} + \sqrt{3})^2 + q_1(\sqrt{2} + \sqrt{3}) + q_0 = 0.$$

—Elementary arithmetic shows that this relation is satisfied for $q_0 = 1$, $q_1 = 0$, $q_2 = -10$, $q_3 = 0$: that is, $\sqrt{2} + \sqrt{3}$ is a root of the polynomial

$$f(t) = t^4 - 10t^2 + 1.$$

In fact, it is easily checked that $f(t)$ vanishes at all four combinations

$$\pm\sqrt{2} \pm \sqrt{3}.$$

—It follows that

$$f(t) = (t - (-\sqrt{2} - \sqrt{3}))(t - (-\sqrt{2} + \sqrt{3}))(t - (\sqrt{2} - \sqrt{3}))(t - (\sqrt{2} + \sqrt{3}))$$

and that $f(t)$ is irreducible over $\mathbb{Q}$ (no proper factor of $f(t)$ has rational coefficients). Thus $\sqrt{2} + \sqrt{3}$ has degree 4 over $\mathbb{Q}$.

—Consider the composition of extensions

$$\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2} + \sqrt{3}) \subseteq \mathbb{Q}(\sqrt{2}, \sqrt{3}).$$

By Corollary 1.11,

$$4 = [\mathbb{Q}(\sqrt{2} + \sqrt{3}) : \mathbb{Q}] \leq [\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}].$$

On the other hand we know that $[\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}] \leq 4$; thus we can conclude that this degree is exactly 4, and it follows that

$$\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3}) :$$

the extension was simple to begin with. (In due time we will prove that this is a typical example, not a contrived one.)

—Staring now at the composition

$$\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2}) \subseteq \mathbb{Q}(\sqrt{2}, \sqrt{3}),$$

Proposition 1.10 tells us that $[\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}(\sqrt{2})] = 2$. That is, a side-effect of the computation carried out above is that the polynomial $t^2 - 3$ must be irreducible *over* $\mathbb{Q}(\sqrt{2})$; this is not surprising, but note that this method is very different from anything we encountered back in §V.5.

—The fact that the other roots of the minimal polynomial of $\sqrt{2} + \sqrt{3}$ 'looked so much like' $\sqrt{2} + \sqrt{3}$ is not surprising. Indeed, since $\mathbb{Q}(\sqrt{2}) \subseteq \mathbb{Q}(\sqrt{2}, \sqrt{3})$ is a simple extension, we know (cf. Proposition 1.5 and following) that there is an automorphism of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ fixing $\mathbb{Q}(\sqrt{2})$ (and hence $\mathbb{Q}$) and swapping $\sqrt{3}$ and $-\sqrt{3}$. Similarly, there is an automorphism fixing $\mathbb{Q}(\sqrt{3})$ and swapping $\sqrt{2}$ and $-\sqrt{2}$. These automorphisms must act on the set of roots of $f(t)$ (Exercise 1.6): applying them and their composition to $\sqrt{2} + \sqrt{3}$ produces the other three roots of $f(t)$.

—In fact, at this point we know that $G = \mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt{2}, \sqrt{3}))$ must consist of 4 elements (by Corollary 1.7) and has at least two elements of order 2 (both automorphisms found above have order 2). This is enough to conclude that G is *not* a cyclic group, and hence it must be isomorphic to the group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. ⌟

Exercises

1.1. ▷ Prove that if $k \subseteq K$ is a field extension, then $\operatorname{char} k = \operatorname{char} K$. Prove that the category Fld has no initial object. [§1.1]

1.2. Define carefully the category Fld_k of extensions of k.

1.3. ▷ Let $k \subseteq F$ be a field extension, and let $\alpha \in F$. Prove that the field $k(\alpha)$ consists of all the elements of F which may be written as a rational function in α, with coefficients in k. Why does this *not* give (in general) an onto homomorphism $k(t) \to k(\alpha)$? [§1.2, §1.3]

1.4. Let $k \subseteq k(\alpha)$ be a simple extension, with α transcendental over k. Let E be a subfield of $k(\alpha)$ properly containing k. Prove that $k(\alpha)$ is a finite extension of E.

1.5. ▷ (Cf. Example 1.4.)

- Prove that there is exactly one subfield of $\mathbb{R}$ isomorphic to $\mathbb{Q}[t]/(t^2 - 2)$.
- Prove that there are exactly three subfields of $\mathbb{C}$ isomorphic to $\mathbb{Q}[t]/(t^3 - 2)$.

From a 'topological' point of view, one of these copies of $\mathbb{Q}[t]/(t^3 - 2)$ looks very different from the other two: it is not dense in $\mathbb{C}$, but the others are. [§1.2]

1.6. ▷ Let $k \subseteq F$ be a field extension, and let $f(x) \in k[x]$ be a polynomial. Prove that $\operatorname{Aut}_k(F)$ acts on the set of roots of $f(x)$. Provide examples showing that this action need not be transitive or faithful. [§1.2, §1.3]

1.7. Let $k \subseteq F$ be a field extension, and let $\alpha \in F$ be algebraic over k.

- Suppose $p(x) \in k[x]$ is an irreducible monic polynomial such that $p(\alpha) = 0$; prove that $p(x)$ is the minimal polynomial of α over k, in the sense of Proposition 1.3.
- Let $f(x) \in k[x]$. Prove that $f(\alpha) = 0$ if and only if $p(x) \mid f(x)$.
- Show that the minimal polynomial of α is the minimal polynomial of a certain k-linear transformation of F, in the sense of Definition VI.6.12.

1.8. ¬ Let $f(x) \in k[x]$ be a polynomial over a field k of degree d, and let $\alpha_1, \ldots, \alpha_d$ be the roots of $f(x)$. For a subset $I \subseteq \{1, \ldots, d\}$, denote by α_I the sum $\sum_{i \in I} \alpha_i$. Assume that $\alpha_I \in k$ only for $I = \emptyset$ and $I = \{1, \ldots, d\}$. Prove that $f(x)$ is irreducible over k. [7.14]

1.9. Let k be a finite field. Prove that the order $|k|$ is a power of a prime integer.

1.10. ¬ Let k be a field. Prove that the ring of square $n \times n$ matrices $\mathcal{M}_n(k)$ contains an isomorphic copy of every extension of k of degree $\leq n$. (Hint: If $k \subseteq F$ is an extension of degree n and $\alpha \in F$, then 'multiplication by α' is a k-linear transformation of F.) [5.20]

1.11. ¬ Let $k \subseteq F$ be a finite field extension, and let $p(x)$ be the characteristic polynomial of the k-linear transformation of F given by multiplication by α. Prove that $p(\alpha) = 0$.

This gives an effective way to find a polynomial satisfied by an element of an extension. Use it to find a polynomial satisfied by $\sqrt{2}+\sqrt{3}$ over $\mathbb{Q}$, and compare this method with the one used in Example 1.19. [1.12]

1.12. ¬ Let $k \subseteq F$ be a finite field extension, and let $\alpha \in F$. The *norm* of α, $N_{k\subseteq F}(\alpha)$, is the determinant of the linear transformation of F given by multiplication by α (cf. Exercise 1.11, Definition VI.6.4).

Prove that the norm is multiplicative: for $\alpha, \beta \in F$,

$$N_{k\subseteq F}(\alpha\beta) = N_{k\subseteq F}(\alpha)N_{k\subseteq F}(\beta).$$

Compute the norm of a complex number viewed as an element of the extension $\mathbb{R} \subseteq \mathbb{C}$ (and marvel at the excellent choice of terminology). Do the same for elements of an extension $\mathbb{Q}(\sqrt{d})$ of $\mathbb{Q}$, where d is an integer that is not a square, and compare the result with Exercise III.4.10. [1.13, 1.14, 1.15, 4.19, 6.18, VIII.1.5]

1.13. ¬ Define the *trace* $\mathrm{tr}_{k\subseteq F}(\alpha)$ of an element α of a finite extension F of a field k by following the lead of Exercise 1.12. Prove that the trace is *additive:*

$$\mathrm{tr}_{k\subseteq F}(\alpha+\beta) = \mathrm{tr}_{k\subseteq F}(\alpha) + \mathrm{tr}_{k\subseteq F}(\beta)$$

for $\alpha, \beta \in F$. Compute the trace of an element of an extension $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{d})$, for d an integer that is not a square. [1.14, 1.15, 4.19, VIII.1.5]

1.14. ¬ Let $k \subseteq k(\alpha)$ be a simple algebraic extension, and let $x^d + a_{d-1}x^{d-1} + \cdots + a_0$ be the minimal polynomial of α over k. Prove that

$$\mathrm{tr}_{k\subseteq k(\alpha)}(\alpha) = -a_{d-1} \quad \text{and} \quad N_{k\subseteq k(\alpha)}(\alpha) = (-1)^d a_0.$$

(Cf. Exercises 1.12 and 1.13.) [4.19]

1.15. ¬ Let $k \subseteq F$ be a finite extension, and let $\alpha \in F$. Assume $[F : k(\alpha)] = r$. Prove that

$$\mathrm{tr}_{k\subseteq F}(\alpha) = r\,\mathrm{tr}_{k\subseteq k(\alpha)}(\alpha) \quad \text{and} \quad N_{k\subseteq F}(\alpha) = N_{k\subseteq k(\alpha)}(\alpha)^r.$$

(Cf. Exercises 1.12 and 1.13.) (Hint: If $f_1, \ldots, f_r$ is a basis of F over $k(\alpha)$ and α has degree d over k, then $(f_i\alpha^j)_{\substack{i=1,\cdots,r \\ j=1,\cdots,d-1}}$ is a basis of F over k. The matrix corresponding to multiplication by α with respect to this basis consists of r identical square blocks.) [4.19, 4.21]

1.16. ▷ Let $k \subseteq L \subseteq F$ be fields, and let $\alpha \in F$. If $k \subseteq k(\alpha)$ is a finite extension, then $L \subseteq L(\alpha)$ is finite and $[L(\alpha) : L] \leq [k(\alpha) : k]$. [§1.3]

1.17. ▷ Let $k \subseteq F = k(\alpha_1, \ldots, \alpha_n)$ be a finitely generated extension. Prove that the evaluation map

$$k[t_1, \ldots, t_n] \to F, \quad t_i \mapsto \alpha_i$$

is an epimorphism of rings (although it need not be onto). [§1.3]

1.18. ¬ Let R be a ring sandwiched between a field k and an algebraic extension F of k. Prove that R is a field.

Is it necessary to assume that the extension is algebraic? [1.19]

1.19. Let $k \subseteq F$ be a field extension of degree p, a prime integer. Prove that there are no subrings of F properly containing k and properly contained in F. (Use Exercise 1.18.)

1.20. Let p be a prime integer, and let $\alpha = \sqrt[p]{2} \in \mathbb{R}$. Let $g(x) \in \mathbb{Q}[x]$ be any non-constant polynomial of degree $< p$. Prove that α may be expressed as a polynomial in $g(\alpha)$ with rational coefficients.

Prove that an analogous statement for $\sqrt[4]{2}$ is false.

1.21. Let $k \subseteq F$ be a field extension, and let E be the intermediate field consisting of the elements of F which are algebraic over k. For $\alpha \in F$, prove that α is algebraic over E if and only if $\alpha \in E$. Deduce that $\overline{\mathbb{Q}}$ is algebraically closed.

1.22. Let $k \subseteq F$ be a field extension, and let $\alpha \in F$, $\beta \in F$ be algebraic, of degree d, e, resp. Assume d, e are relatively prime, and let $p(x)$ be the minimal polynomial of β over k. Prove $p(x)$ is irreducible over $k(\alpha)$.

1.23. Express $\sqrt{2}$ explicitly as a polynomial function in $\sqrt{2} + \sqrt{3}$ with rational coefficients.

1.24. Generalize the situation examined in Example 1.19: let k be a field of characteristic $\neq 2$, and let $a, b \in k$ be elements that are not squares in k; prove that $k(\sqrt{a}, \sqrt{b}) = k(\sqrt{a} + \sqrt{b})$.

Prove that $k(\sqrt{a}, \sqrt{b})$ has degree 2, resp., 4, over k according to whether ab is, resp., is not, a square in k.

1.25. ¬ Let $\xi := \sqrt{2 + \sqrt{2}}$.

- Find the minimal polynomial of ξ over $\mathbb{Q}$, and show that $\mathbb{Q}(\xi)$ has degree 4 over $\mathbb{Q}$.
- Prove that $\sqrt{2 - \sqrt{2}}$ is another root of the minimal polynomial of ξ.
- Prove that $\sqrt{2 - \sqrt{2}} \in \mathbb{Q}(\xi)$. (Hint: $(a + b)(a - b) = a^2 - b^2$.)
- By Proposition 1.5, sending ξ to $\sqrt{2 - \sqrt{2}}$ defines an automorphism of $\mathbb{Q}(\xi)$ over $\mathbb{Q}$. Find the matrix of this automorphism w.r.t. the basis $1, \xi, \xi^2, \xi^3$.
- Prove that $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\xi))$ is cyclic of order 4.

[6.6]

1.26. ¬ Let $k \subseteq F$ be a field extension, let I be an indexing set, and let $\{\alpha_i\}_{i \in I}$ be a choice of elements of F. This choice determines a homomorphism φ of k-algebras from the polynomial ring $k[I]$ on the set I to F (the polynomial ring is a free commutative k-algebra; cf. Proposition III.6.4). We say that $\{\alpha_i\}_{i \in I}$ is *algebraically independent* over k if φ is injective. For example, distinct elements $\alpha_1, \ldots, \alpha_n$ of F are algebraically independent over k if there is no nonzero polynomial $f(x_1, \ldots, x_n) \in k[x_1, \ldots, x_n]$ such that $f(\alpha_1, \ldots, \alpha_n) = 0$.

Prove that $\alpha_1, \ldots, \alpha_n$ are algebraically independent if and only if the assignment $t_1 \mapsto \alpha_1, \ldots, t_n \mapsto \alpha_n$ defines a homomorphism of k-algebras (and hence an isomorphism) from the field of rational functions $k(t_1, \ldots, t_n)$ to $k(\alpha_1, \ldots, \alpha_n)$. [1.27]

1.27. ¬ With notation and terminology as in Exercise 1.26, the indexed set $\{\alpha_i\}_{i\in I}$ is a *transcendence basis* for F over k if it is a maximal algebraically independent set in F.

- Prove that $\{\alpha_i\}_{i\in I}$ is a transcendence basis for F over k if and only if it is algebraically independent and F is algebraic over $k(\{\alpha_i\}_{i\in I})$.
- Prove that transcendence bases exist. (Zorn)
- Prove that any two transcendence bases for F over k have the same cardinality. (Mimic the proof of Proposition VI.1.9. Don't feel too bad if you prefer to deal only with the case of finite transcendence bases.)

The cardinality of a transcendence basis is called the *transcendence degree* of F over k, denoted $\mathrm{tr.deg}_{k\subseteq F}$. [1.28, 1.29, 2.19]

1.28. ¬ Let $k \subseteq E \subseteq F$ be field extensions. Prove that $\mathrm{tr.deg}_{k\subseteq F}$ is finite if and only if both $\mathrm{tr.deg}_{k\subseteq E}$ and $\mathrm{tr.deg}_{E\subseteq F}$ (see Exercise 1.27) are finite and in this case

$$\mathrm{tr.deg}_{k\subseteq F} = \mathrm{tr.deg}_{k\subseteq E} + \mathrm{tr.deg}_{E\subseteq F}\,.$$

[7.3]

1.29. ¬ An extension $k \subseteq F$ is *purely transcendental* if it admits a transcendence basis $\{\alpha_i\}_{i\in I}$ (see Exercise 1.27) such that $F = k(\{\alpha_i\}_{i\in I})$.

Prove that any field extension $f \subseteq F$ may be decomposed as a purely transcendental extension followed by an algebraic extension. (Not all field extensions may be decomposed as an algebraic extension followed by a purely transcendental extension.) [1.30]

1.30. Let $k \subseteq k(\alpha)$ be a simple extension, with α transcendental over k. Let E be a subfield of $k(\alpha)$ properly containing k. Prove that $\mathrm{tr.deg}_{k\subseteq E} = 1$.

Lüroth's theorem asserts that in this situation $k \subseteq E$ is itself a simple transcendental extension of k; that is, it is purely transcendental (Exercise 1.29).

---

2. Algebraic closure, Nullstellensatz, and a little algebraic geometry

One of the most important extensions of a field k is its *algebraic closure* $k \subseteq \overline{k}$; this was mentioned already in §V.5.2, and we can now prove that $\overline{k}$ exists and is unique (up to isomorphism). Once this circle of ideas is approached, the temptation to say a few words about the important result known as *Hilbert's Nullstellensatz,* at the cost of a small digression into algebraic geometry, will simply be irresistible.

2.1. Algebraic closure. Recall (Definition V.5.9) that a field K is *algebraically closed* if all irreducible polynomials in $K[x]$ have degree 1, that is, if every polynomial in $K[x]$ factors completely as a product of linear terms. Equivalently (Exercise III.4.21) every maximal ideal in $K[x]$ is of the form $(x - c)$, for $c \in K$.

Now that we have a little more vocabulary, we can recast this definition yet again, as follows.

Lemma 2.1. *For a field K, the following are equivalent:*

- *K is algebraically closed.*
- *K has no nontrivial algebraic extensions.*
- *If $K \subseteq L$ is any extension and $\alpha \in L$ is algebraic over K, then $\alpha \in K$.*

The proof is a straightforward application of the definitions and a good exercise (Exercise 2.1).

Definition 2.2. An *algebraic closure* of a field k is an *algebraic* extension $k \subseteq \overline{k}$ such that $\overline{k}$ is algebraically closed. ⌟

Of course, every polynomial $f(x) \in k[x]$ will split into linear factors over $\overline{k}$: so does every polynomial in the much larger $\overline{k}[x]$, as $\overline{k}$ is algebraically closed. The requirement that $k \subseteq \overline{k}$ be algebraic ensures that no other intermediate field L,

$$k \subseteq L \subsetneq \overline{k},$$

may be algebraically closed: indeed, $L \subseteq \overline{k}$ would be a nontrivial algebraic extension, contradicting Lemma 2.1. In fact, the only subfield of $\overline{k}$ containing all roots of all nonconstant polynomials in $k[x]$ is $\overline{k}$ itself (Exercise 2.2); thus, the algebraic closure of a field is 'as small as possible' subject to this requirement. Not surprisingly, $\overline{k}$ turns out to be unique up to isomorphism, so that we can speak of *the* algebraic closure of k.

Theorem 2.3. *Every field k admits an algebraic closure $k \subseteq \overline{k}$; this extension is unique up to isomorphism.*

Concerning existence, the idea is to construct 'by hand' a huge extension K of k where every polynomial $f(x) \in k[x]$ factors completely. The elements of K which are algebraic over k will form an algebraic closure of k.

The construction is done in steps, each step including 'one more root' of all nonconstant polynomials in $k[x]$. Here is a formalization of this step.

Lemma 2.4. *Let k be a field. Then there exists an extension $k \subseteq K$ such that every nonconstant polynomial $f(x) \in k[x]$ has at least one root in K.*

Proof. (This construction is apparently due to Emil Artin.) Consider a set $\mathscr{T} = \{t_f\}$ in bijection with the set of nonconstant monic polynomials $f(x) \in k[x]$, and let $k[\mathscr{T}]$ be the corresponding polynomial ring[5] in all the indeterminates t_f. Let $I \subseteq k[\mathscr{T}]$ be the ideal generated by all polynomials $f(t_f)$.

Then I is a proper ideal. Indeed, otherwise we could write

$$(*) \qquad 1 = \sum_{i=1}^{n} a_i \cdot f_i(t_{f_i}),$$

where $a_i \in k[\mathscr{T}]$. We claim that this cannot be done: indeed, we can construct an extension $k \subseteq F$ where the polynomials $f_1(x), \ldots, f_n(x)$ have roots $\alpha_1, \ldots, \alpha_n$,

[5]Here is another rare occasion where we need to consider polynomial rings with possibly infinitely many indeterminates. An element of $k[\mathscr{T}]$ is simply an ordinary polynomial with coefficients in k, involving a finite number of indeterminates from $\mathscr{T}$.

respectively (apply Proposition V.5.7 n times); view (*) as an identity in $F[\mathscr{T}]$, and plug in $t_{f_i} = \alpha_i$, obtaining

$$1 = \sum_{i=1}^{n} a_i \cdot f_i(\alpha_i) = \sum_{i=1}^{n} a_i \cdot 0 = 0,$$

which is nonsense.

Since I is proper, it is contained in a maximal ideal $\mathfrak{m}$ (Proposition V.3.5). Thus, we obtain a field extension

$$k \subseteq K := \frac{k[\mathscr{T}]}{\mathfrak{m}};$$

by construction every nonconstant monic (and hence every nonconstant) polynomial $f(x)$ has a root in K, namely the coset of t_f. □

The field constructed in Lemma 2.4 contains at least one root of each nonconstant polynomial $f(x) \in k[x]$, but this is not good enough: we are seeking a field which contains *all* roots of such polynomials. Equivalently, not only should $f(x)$ have (at least) one linear factor $x - \alpha$ in $K[x]$, but we need the quotient polynomial $f(x)/(x-\alpha) \in K[x]$ to have a linear factor and the quotient by that new factor to have one, etc. This prompts us to consider a whole chain of extensions

$$k \subseteq K_1 \subseteq K_2 \subseteq K_3 \subseteq \cdots$$

where K_1 is obtained from k by applying Lemma 2.4, K_2 is similarly obtained from K_1, etc.

Now consider the *union* L of this chain[6]. For every two $a, b \in L$, there is an i such that $a, b \in K_i$; we can define $a + b$, $a \cdot b$ in L by adopting their definition in K_i, and the result does not depend on the choice of i. It follows easily that L is a field.

Claim 2.5. *The field L is algebraically closed.*

Proof. If $f(x) \in L[x]$ is a nonconstant polynomial, then $f(x) \in K_i$ for some i; hence $f(x)$ has a root in $K_{i+1} \subseteq L$. That is, every nonconstant polynomial in $L[x]$ has a root in L, as needed. □

Proof of the existence of algebraic closures. The existence of algebraic closures is now completely transparent, because of the following simple observation:

Lemma 2.6. *Let $k \subseteq L$ be a field extension, with L algebraically closed. Let*

$$\overline{k} := \{\alpha \in L \mid \alpha \text{ is algebraic over } k\}.$$

Then $\overline{k}$ is an algebraic closure of k.

The construction reviewed above provides us with an algebraically closed field L containing any given field k, so the lemma is all we need to prove.

[6] This is an example of *direct limit,* a notion which we will introduce more formally in §VIII.1.4.

By Corollary 1.16, $\overline{k}$ is a field, and the extension $k \subseteq \overline{k}$ is tautologically algebraic. To verify that $\overline{k}$ is algebraically closed, let α be algebraic over $\overline{k}$; then

$$k \subseteq \overline{k} \subseteq \overline{k}(\alpha)$$

is a composition of algebraic extensions, so $k \subseteq \overline{k}(\alpha)$ is algebraic (Corollary 1.18), and in particular α is algebraic over k. But then $\alpha \in \overline{k}$, by definition of the latter. It follows that $\overline{k}$ is algebraically closed, by Lemma 2.1. □

Remark 2.7. Zorn's lemma was sneakily used in this proof (we used the existence of maximal ideals, which relies upon Zorn's lemma), and it is natural to wonder whether the existence of algebraic closures may be another statement equivalent to the axiom of choice. Apparently, this is not the case[7]. ⌟

Next, we deal with uniqueness. It may be tempting to try to set things up in such a way that algebraic closures would end up being solutions to a universal problem; this would guarantee uniqueness by abstract nonsense. However, we run into the same obstacle encountered in Proposition V.5.7: morphisms between extensions depend on choices of roots of polynomials, so that algebraic closures are not 'universal' in the most straightforward sense of the term.

Therefore, a bit of independent work is needed.

Lemma 2.8. *Let $k \subseteq L$ be a field extension, with L algebraically closed. Let $k \subseteq F$ be any algebraic extension. Then there exists a morphism of extensions $i : F \to L$.*

As pointed out above, i is by no means unique!

Proof. This argument also relies on Zorn's lemma. Consider the set Z of homomorphisms

$$i_K : K \to L$$

where K is an intermediate field, $k \subseteq K \subseteq F$, and i_K restricts to the identity on k; Z is nonempty, since the extension $i_k : k \subseteq L$ defines an element of Z. We give a poset structure to Z by defining

$$i_K \preceq i_{K'}$$

if $K \subseteq K' \subseteq F$ and $i_{K'}$ restricts to i_K on K. To verify that every chain C in Z has an upper bound in Z, let K_C be the union of the sources of all $i_K \in C$ (K_C is clearly a field); if $\alpha \in K_C$, define $i_{K_C}(\alpha)$ to be $i_K(\alpha)$, where i_K is any element of C such that $\alpha \in K$. This prescription is clearly independent of the chosen K and defines a homomorphism $K_C \to L$ restricting to the identity on k. This homomorphism is an upper bound for C.

By Zorn's lemma, Z admits a maximal element i_G, corresponding to an intermediate field $k \subseteq G \subseteq F$. Let $H = i_G(G)$ be the image of G in $\overline{k}$.

We claim that $G = F$: this will prove the statement, because it will imply that there is a homomorphism $i_F : F \to L$ extending the identity on k.

Arguing by contradiction, assume that there exists an $\alpha \in F \setminus G$, and consider the extension $G \subseteq G(\alpha)$. Since $\alpha \in F$ is algebraic over k, it is algebraic over G;

[7] Allegedly, the existence of algebraic closures is a consequence of the *compactness theorem for first-order logic,* which is known to be weaker than the axiom of choice.

thus, it is a root of an irreducible polynomial $p(x) \in G[x]$. Consider the induced homomorphism

$$i_G : G[x] \to H[x],$$

and let $h(x) = i_G(g(x))$. Then $h(x)$ is an irreducible polynomial over H, and it has a root β in L (this is where we use the hypothesis that L is algebraically closed!). We are in the situation considered in Proposition 1.5: we have an isomorphism of fields $i_G : G \to H$; we have simple extensions $G(\alpha)$, $H(\beta)$; and α, β are roots of irreducible polynomials $g(x)$, $h(x) = i_G(g(x))$. By Proposition 1.5, i_G lifts to an isomorphism

$$i_{G(\alpha)} : G(\alpha) \to H(\beta) \subseteq L$$

sending α to β. This contradicts the maximality of i_G; hence $G = F$, concluding the argument. □

Proof of uniqueness of algebraic closures. Let $k \subseteq \overline{k}$, $k \subseteq \overline{k}_1$ be two algebraic closures of k; we have to prove that there exists an isomorphism $\overline{k}_1 \to \overline{k}$ extending the identity on k.

Since $k \subseteq \overline{k}_1$ is algebraic and $\overline{k}$ is algebraically closed, by Lemma 2.8 there exists a homomorphism $i : \overline{k}_1 \to \overline{k}$ extending the identity on k; this homomorphism is trivially injective, since $\overline{k}_1$ is a field. It is also surjective, by Lemma 2.1: otherwise $\overline{k}_1 \subsetneq \overline{k}$ is a nontrivial algebraic extension, contradicting the fact that $\overline{k}_1$ is algebraically closed. Therefore i is an isomorphism, as needed. □

2.2. The Nullstellensatz. If K is an algebraically closed field, then every maximal ideal in $K[x]$ is of the form $(x-c)$, for $c \in K$ (Exercise III.4.21). This statement has a straightforward-looking generalization to polynomial rings in more indeterminates: *if K is algebraically closed, then every maximal ideal in $K[x_1, \ldots, x_n]$ is of the form $(x_1 - c_1, \ldots, x_n - c_n)$.*

Proving this statement is more challenging than it may seem from the looks of it; it is one facet of the famous theorem known as *Hilbert's Nullstellensatz* ("theorem on the position of zeros"). The Nullstellensatz has made a few cameo appearances earlier (see Example III.4.15 and §III.6.5 in particular), and we will spend a little time on it now. We will not prove the Nullstellensatz in its natural generality, that is, for all fields; this would take us too far. Actually, there are reasonably short proofs of the theorem in this generality, but the short arguments we have run into end up replacing a wider (and useful) context with ingenious cleverness; we do not find these arguments particularly insightful or memorable.

By contrast, the theorem has a very simple and memorable proof if we make the further assumption that the field is *uncountable*. Therefore, the reader will have a complete proof of the theorem (in the form given below, which does not assume the field to be algebraically closed) for fields such as $\mathbb{R}$ and $\mathbb{C}$ but will have to trust on faith regarding other extremely important fields such as $\overline{\mathbb{Q}}$.

The most vivid applications of the Nullstellensatz involve basic definitions in algebraic geometry, and we will get a small taste of this in the next section; but the theorem itself is best understood in the context of the considerations on *'finiteness'* mentioned in §1.3; see especially Remark 1.14. We pointed out that if $k \subseteq F$

is finitely generated *as a field extension,* then F need not be finitely generated (that is, 'finite-type') *as a k-algebra,* and we raised the question of understanding extensions F that *are* finite-type k-algebras. The following result answers this question. It is our favorite version of the Nullstellensatz; therefore we will name it so:

Theorem 2.9 (Nullstellensatz). *Let $k \subseteq F$ be a field extension, and assume that F is a finite-type k-algebra. Then $k \subseteq F$ is a finite (hence algebraic) extension.*

The reader should pause and savor this statement for a moment. As pointed out in §III.6.5, an algebra $k \to S$ may very well be of finite type (as an algebra), without being finite (i.e., finitely generated as a module): $S = k[t]$ is an obvious example. The content of Theorem 2.9 is that the distinction between finite-type and finite disappears *if S is a field.*

As mentioned above, we will only prove this statement under an additional (and somewhat unnatural) assumption, which reduces the argument to elementary linear algebra.

Proof for uncountable fields. *Assume that k is uncountable.*

Let $k \subseteq F$ be a field extension, and assume that F is finitely generated as an algebra over k; in particular, it is finitely generated as a field extension. We have to prove that $k \subseteq F$ is a finite extension, or equivalently (by Proposition 1.15) that it is algebraic, that is, that every $\alpha \in F \smallsetminus k$ is the root of a nonzero $f(x) \in k[x]$.

Now, by hypothesis F is the quotient of a polynomial ring over k:

$$k[x_1, \ldots, x_n] \twoheadrightarrow F \ ;$$

this implies that there is a *countable* basis of F as a vector space over k, because this is the case for $k[x_1, \ldots, x_n]$. Consider then the set

$$\left\{ \frac{1}{\alpha - c} \right\}_{c \in k} :$$

this is an uncountable subset of F; therefore it is linearly *de*pendent over k (Proposition VI.1.9). That is, there exist distinct $c_1, \ldots, c_m \in k$ and nonzero coefficients $\lambda_1, \ldots, \lambda_m \in k$, such that

$$\frac{\lambda_1}{\alpha - c_1} + \cdots + \frac{\lambda_m}{\alpha - c_m} = 0.$$

Expressing the left-hand side with a common denominator gives

$$\frac{f(\alpha)}{g(\alpha)} = 0,$$

for $f(x) \neq 0$ in $k[x]$ and $g(\alpha) \neq 0$ (Exercise 2.4). This implies $f(\alpha) = 0$ for a nonzero $f(x) \in k[x]$, and we are done. □

Regardless of its proof, the form of the Nullstellensatz given in Theorem 2.9 does not look much like the other results we have mentioned as associated with it. For one thing, it is not too clear why Theorem 2.9 should be called a theorem on the 'position of zeros'. The result deserves a little more attention.

The form stated at the beginning of this section is obtained as follows:

Corollary 2.10. *Let K be an algebraically closed field, and let I be an ideal of $K[x_1, \ldots, x_n]$. Then I is maximal if and only if*

$$I = (x_1 - c_1, \ldots, x_n - c_n),$$

for $c_1, \ldots, c_n \in K$.

Proof. For $c_1, \ldots, c_n \in K$

$$\frac{K[x_1, \ldots, x_n]}{(x_1 - c_1, \ldots, x_n - c_n)} \cong K$$

(Exercise III.4.12) is a field; therefore $(x_1 - c_1, \ldots, x_n - c_n)$ is maximal. Conversely, let $\mathfrak{m}$ be a maximal ideal in $K[x_1, \ldots, x_n]$, so that the quotient is a field and the natural map

$$K \to L := \frac{K[x_1, \ldots, x_n]}{\mathfrak{m}}$$

is a field extension. The field L is finitely generated as a K-algebra; therefore $K \subseteq L$ is an algebraic extension by Theorem 2.9. Since K is algebraically closed, this implies that $L = K$ (Lemma 2.1); therefore, there is a surjective homomorphism

$$\varphi : K[x_1, \ldots, x_n] \to K$$

such that $\mathfrak{m} = \ker \varphi$. Let $c_i := \varphi(x_i)$. Then

$$(x_1 - c_1, \ldots, x_n - c_n) \subseteq \ker \varphi = \mathfrak{m};$$

but $(x_1 - c_1, \ldots, x_n - c_n)$ is maximal, so this implies $\mathfrak{m} = (x_1 - c_1, \ldots, x_n - c_n)$, as needed. □

Note how very (trivially) false the statement of Corollary 2.10 is over fields which are not algebraically closed: $(x^2 + 1)$ is maximal in $\mathbb{R}[x]$. No such silliness may occur over $\mathbb{C}$, for any number of variables.

2.3. A little affine algebraic geometry. Corollary 2.10 is one of the main reasons why 'classical' algebraic geometry developed over $\mathbb{C}$ rather than fields such as $\mathbb{R}$ or $\mathbb{Q}$: it may be interpreted as saying that *if K is algebraically closed, then there is a natural bijection between the points of the product space K^n and the maximal ideals in the ring $K[x_1, \ldots, x_n]$*. This is the beginning of a fruitful dictionary translating geometry into algebra and conversely. It is worthwhile formalizing this statement and exploring other 'geometric' consequences of the Nullstellensatz.

For K a field, $\mathbb{A}^n_K$ denotes the *affine space* of dimension n over K, that is, the set of n-tuples of elements of K:

$$\mathbb{A}^n_K = \{(c_1, \ldots, c_n) \mid c_i \in K\}.$$

Elements of $\mathbb{A}^n_K$ are called *points.*

Objection: Why don't we just use 'K^n' for this object? Because the latter already stands for the standard n-dimensional *vector space* over K; so it carries with it a certain baggage: the tuple $(0, \ldots, 0)$ is special, so are the vector subspaces of K^n, etc. By constrast, no point of $\mathbb{A}^n_K$ is 'special'; we can carry any point to any other point by a simple translation. Similarly, although it is useful to consider

affine subspaces, these (unlike *vector* subspaces) are not required to go through the origin. Affine geometry and linear algebra are different in many respects.

We have already established that, by the Nullstellensatz, *if K is algebraically closed,* then there is a natural bijection between the points of $\mathbb{A}^n_K$ and the *maximal* ideals of the polynomial ring $K[x_1, \ldots, x_n]$. Concretely, the bijection works like this: the point $p = (c_1, \ldots, c_n)$ corresponds to the set of polynomials

$$\mathscr{I}(p) := \{f(\underline{x}) \in K[x_1, \ldots, x_n] \mid f(p) = 0\},$$

where $f(p)$ is the result of applying the evaluation map:

$$f(x_1, \ldots, x_n) \mapsto f(c_1, \ldots, c_n) \in K.$$

This of course is nothing but the homomorphism

$$\varphi : K[x_1, \ldots, x_n] \to K$$

defined by prescribing $x_i \mapsto c_i$; the set $\mathscr{I}(p)$ is simply $\ker \varphi$, which is the maximal ideal

$$(x_1 - c_1, \ldots, x_n - c_n).$$

This correspondence $p \mapsto \mathscr{I}(p)$ may be defined over every field; the content of Corollary 2.10 is that *every* maximal ideal of $K[x_1, \ldots, x_n]$ corresponds to a point of $\mathbb{A}^n_K$ in this way, if K is algebraically closed.

It is natural to upgrade the correspondence as follows (again, over arbitrary fields): for every subset $S \subseteq \mathbb{A}^n_K$, consider the ideal

$$\mathscr{I}(S) := \{f(\underline{x}) \in K[x_1, \ldots, x_n] \mid \forall p \in S, f(p) = 0\},$$

that is, the set of polynomials 'vanishing along S'; it is immediately checked that this is indeed an ideal of $K[x_1, \ldots, x_n]$. We can also consider a correspondence in the reverse direction, from ideals of $K[x_1, \ldots, x_n]$ to subsets of $\mathbb{A}^n_K$, defined by setting for every ideal I,

$$\mathscr{V}(I) := \{p = (c_1, \ldots, c_n) \in \mathbb{A}^n_K \mid \forall f \in I, f(c_1, \ldots, c_n) = 0\}.$$

Thus, $\mathscr{V}(I)$ is the set of common solutions of all the polynomial equations $f = 0$ as $f \in I$: the set of points 'cut out in $\mathbb{A}^n_K$' by the polynomials in I.

In its most elementary manifestation, the dictionary mentioned in the beginning of the section consists precisely of the pair of correspondences

$$\{\text{subsets of } \mathbb{A}^n_K\} \underset{\mathscr{V}}{\overset{\mathscr{I}}{\rightleftarrows}} \{\text{ideals in } K[x_1, \ldots, x_n]\} .$$

The set $\mathscr{V}(I)$ is often (somewhat improperly) called the *variety of I,* while $\mathscr{I}(S)$ is (also improperly) the *ideal of S.*

The function $\mathscr{V}$ can be defined (using the same prescription) for any *set* $A \subseteq K[x_1, \ldots, x_n]$, whether A is an ideal or not; it is clear that $\mathscr{V}(A) = \mathscr{V}(I)$ where I is the ideal generated by A (Exercise 2.5). Hilbert's basis theorem (Theorem V.1.2) tells us something rather interesting: $K[x_1, \ldots, x_n]$ is Noetherian when K is a field; hence every ideal I is generated by a finite number of elements. Thus, for every ideal I there exist polynomials $f_1, \ldots, f_r$ such that (abusing notation a little)

$$\mathscr{V}(I) = \mathscr{V}(f_1, \ldots, f_r).$$

In other words, if a set may be defined by any set of polynomial equations, it may in fact be defined by a *finite* set of polynomial equations[8].

We will pass in silence many simple-minded properties of the functions $\mathscr{V}$, $\mathscr{I}$ (they reverse inclusions, they behave reasonably well with respect to unions and intersections, etc.); the reader should feel free to research the topic at will. But we want to focus the reader's attention on the fact that (of course) $\mathscr{V}$, $\mathscr{I}$ are *not bijections;* this is a problem, if we really want to construct a dictionary between 'geometry' and 'algebra'.

For example, there are (in general) lots of subsets of $\mathbb{A}^n_K$ which are not cut out by a set of polynomial equations, and there are lots of ideals in $K[x_1, \ldots, x_n]$ which are not obtained as $\mathscr{I}(S)$ for any subset $S \subseteq \mathbb{A}^n_K$.

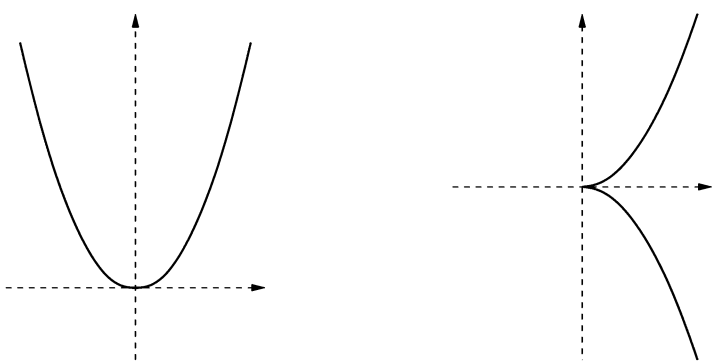
We will leave to the reader the task of finding examples of the first phenomenon (Exercise 2.6). The way around it is to restrict our attention to subsets of $\mathbb{A}^n_K$ which *are* of the form $\mathscr{V}(I)$ for some ideal I.

Definition 2.11. An *(affine) algebraic set* is a subset $S \subseteq \mathbb{A}^n_K$ such that there exists an ideal $I \subseteq K[x_1, \ldots, x_n]$ for which $S = \mathscr{V}(I)$. ⌟

This definition may seem like a cheap patch: we do not really know what the image of $\mathscr{V}$ is, so we label it in some way and proceed. This is not entirely true—the family of algebraic subsets of a given $\mathbb{A}^n_K$ has a solid, recognizable structure: it is the family of *closed subsets in a topology* on $\mathbb{A}^n_K$ (Exercise 2.7); this topology is called the *Zariski topology.*

Studying affine algebraic sets is the business of affine algebraic geometry. The aim is to understand 'geometric' properties of these sets in terms of 'algebraic' properties of corresponding ideals or other algebraic entities associated with them.

Example 2.12. Here are pictures of two algebraic subsets of $\mathbb{A}^2_{\mathbb{R}}$:



The first is $\mathscr{V}((y - x^2))$; the second is $\mathscr{V}((y^2 - x^3))$. What feature of the ideal $(y^2 - x^3)$ is responsible for the 'cusp' at $(0, 0)$ in the second picture? The reader will find out in any course in elementary algebraic geometry. ⌟

The fact that $\mathscr{I}$ is not surjective leads to interesting considerations. The standard example of an ideal that is *not* the ideal of any set is (x^2) in $K[x]$: because

[8]We distinctly remember being surprised by this fact the first time we ran into it.

wherever x^2 vanishes, so does x; hence if $x^2 \in \mathscr{I}(S)$ for a set S, then $x \in \mathscr{I}(S)$ as well. This motivates the following definitions.

Definition 2.13. Let I be an ideal in a commutative ring R. The *radical of* I is the ideal

$$\sqrt{I} := \{r \in R \mid \exists k \geq 0, r^k \in I\}.$$

An ideal I is a *radical ideal* if $I = \sqrt{I}$. ⌟

The reader should check that $\sqrt{I}$ is indeed an ideal; it may be characterized as the intersection of all prime ideals containing I (Exercise 2.8).

Example 2.14. An element of a commutative ring R is nilpotent if and only if it is in the radical of the ideal (0) (cf. Exercise V.4.19). The reader has encountered this ideal, called the *nilradical* of R, in the exercises, beginning with Exercise III.3.12.

Prime ideals $\mathfrak{p}$ are clearly radical: because $f^k \in \mathfrak{p} \implies f \in \mathfrak{p}$, and therefore $\sqrt{\mathfrak{p}} \subseteq \mathfrak{p}$ (the other inclusion holds trivially for all ideals). ⌟

Lemma 2.15. *Let K be a field, and let S be a subset of $\mathbb{A}^n_K$. Then the ideal $\mathscr{I}(S)$ is a radical ideal of $K[x_1, \ldots, x_n]$.*

Proof. The inclusion $\mathscr{I}(S) \subseteq \sqrt{\mathscr{I}(S)}$ holds for every ideal, so it is trivially satisfied. To verify the inclusion $\sqrt{\mathscr{I}(S)} \subseteq \mathscr{I}(S)$, let $f \in \sqrt{\mathscr{I}(S)}$. Then there is an integer $k \geq 0$ such that $f^k \in \mathscr{I}(S)$; that is,

$$(\forall p \in S), \quad f(p)^k = 0.$$

But then

$$(\forall p \in S), \quad f(p) = 0,$$

proving $f \in \mathscr{I}(S)$, as needed. □

In view of these considerations, for any field we can refine the correspondence between subsets of $\mathbb{A}^n_K$ and ideals of $K[x_1, \ldots, x_n]$ as follows:

$$\{\text{algebraic subsets of } \mathbb{A}^n_K\} \underset{\mathscr{V}}{\overset{\mathscr{I}}{\rightleftarrows}} \{\text{radical ideals in } K[x_1, \ldots, x_n]\};$$

as we have argued, it is necessary to do so if we want to have a good dictionary. In the rest of the section, $\mathscr{I}$ and $\mathscr{V}$ will be taken as acting between these two sets.

While we have tightened the correspondence substantially, the situation is still less than idyllic. Over arbitrary fields, the function $\mathscr{V}$ is now surjective by the very definition of an affine algebraic subset (cf. Exercise 2.9); but it is not necessarily injective. An immediate counterexample is offered over $K = \mathbb{R}$ by the ideal (x^2+1): this is maximal, hence prime, hence radical, and

$$\mathscr{V}(x^2 + 1) = \emptyset = \mathscr{V}(1).$$

Here is where the Nullstellensatz comes to our aid, and here is where we see what the Nullstellensatz has to do with the 'zeros' of an ideal (= $\mathscr{V}$ of that ideal).

Proposition 2.16 (Weak Nullstellensatz)**.** *Let K be an algebraically closed field, and let $I \subseteq K[x_1, \ldots, x_n]$ be an ideal. Then $\mathscr{V}(I) = \emptyset$ if and only if $I = (1)$.*

Proof. If $I = (1)$, then $\mathscr{V}(I) = \emptyset$ by definition.

Conversely, assume that $I \neq (1)$. By Proposition V.3.5, I is then contained in a maximal ideal $\mathfrak{m}$. Since K is algebraically closed, by Corollary 2.10 we have

$$\mathfrak{m} = (x_1 - c_1, \ldots, x_n - c_n)$$

for some $c_1, \ldots, c_n \in K$; so $\forall f(x_1, \ldots, x_n) \in I$ there exist $g_1, \ldots, g_n \in K[x_1, \ldots, x_n]$ such that

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n} g(x_1, \ldots, x_n)(x_i - c_i).$$

In particular,

$$f(c_1, \ldots, c_n) = \sum_{i=1}^{n} g(c_1, \ldots, c_n)(c_i - c_i) = 0 :$$

that is, $(c_1, \ldots, c_n) \in \mathscr{V}(I)$. This proves $\mathscr{V}(I) \neq \emptyset$ if $I \neq (1)$, and we are done. □

This is encouraging, but it does not seem to really prove that $\mathscr{V}$ is injective. As it happens, however, it does: we can derive from the 'weak' Nullstellensatz the following stronger result, which does imply injectivity:

Proposition 2.17 (Strong Nullstellensatz). *Let K be an algebraically closed field, and let $I \subseteq K[x_1, \ldots, x_n]$ be an ideal. Then*

$$\mathscr{I}(\mathscr{V}(I)) = \sqrt{I}.$$

This implies that the composition $\mathscr{I} \circ \mathscr{V}$ is the identity on the set of *radical* ideals; that is, $\mathscr{V}$ has a left-inverse, and therefore it is injective (cf. Proposition I.2.1!).

Proof. Note that $\mathscr{I}(\mathscr{V}(I))$ is a radical ideal (by Lemma 2.15). It follows immediately from the definitions that $I \subseteq \mathscr{I}(\mathscr{V}(I))$; therefore

$$\sqrt{I} \subseteq \sqrt{\mathscr{I}(\mathscr{V}(I))} = \mathscr{I}(\mathscr{V}(I)).$$

We have to verify the reverse inclusion: assume that $f \in \mathscr{I}(\mathscr{V}(I))$; that is, assume that

$$f(p) = 0$$

for all $p \in \mathbb{A}^n_K$ such that $\forall g \in I, g(p) = 0$; we have to show that there exists an m for which $f^m \in I$.

The argument is not difficult after the fact, but it is fiendishly clever[9]. Let y be an extra variable, and consider the ideal J generated by I and $1 - fy$ in $K[x_1, \ldots, x_n, y]$. More explicitly, assume

$$I = (g_1, \ldots, g_r)$$

(since $K[x_1, \ldots, x_n]$ is Noetherian, finitely many generators suffice); we can view $g_i(x_1, \ldots, x_n)$, resp., $f(x_1, \ldots, x_n) \in K[x_1, \ldots, x_n]$, as polynomials $G_i(x_1, \ldots, x_n, y)$, resp., $F(x_1, \ldots, x_n, y) \in K[x_1, \ldots, x_n, y]$, and let

$$J = (G_1, \ldots, G_r, 1 - Fy).$$

[9]This is called the *Rabinowitsch trick* and dates back to 1929. It was published in a one-page, eighteen-line article in the Mathematische Annalen.

As $K[x_1, \ldots, x_n, y]$ has $n+1$ variables, the ideal J defines a subset $\mathscr{V}(J)$ in $\mathbb{A}_K^{n+1}$.

We claim $\mathscr{V}(J) = \emptyset$.

Indeed, assume on the contrary that $\mathscr{V}(J) \neq \emptyset$, and let $p = (a_1, \ldots, a_n, b)$ in $\mathbb{A}_K^{n+1}$ be a point of $\in \mathscr{V}(J)$. Then for $i = 1, \ldots, r$

$$G_i(a_1, \ldots, a_n, y) = 0;$$

but this means

$$g_i(a_1, \ldots, a_n) = 0$$

for $i = 1, \ldots, r$; that is, $(a_1, \ldots, a_n) \in \mathscr{V}(I)$. Since f vanishes at all points of $\mathscr{V}(I)$, this implies

$$f(a_1, \ldots, a_n) = 0,$$

and then

$$(1 - Fy)(a_1, \ldots, a_n, b) = 1 - f(a_1, \ldots, a_n)b = 1 - 0 = 1.$$

But this contradicts the assumption that $p \in \mathscr{V}(J)$. Therefore, $\mathscr{V}(J) = \emptyset$.

Now use the weak Nullstellensatz: since K is algebraically closed, we can conclude $J = (1)$. Therefore, there exist polynomials $H_i(x_1, \ldots, x_n, y)$, $i = 1, \ldots, r$, and $L(x_1, \ldots, x_n, y)$, such that

$$\sum_{i=1}^{r} H_i(x_1, \ldots, x_n, y) G_i(x_1, \ldots, x_n, y) + L(x_1, \ldots, x_n, y)(1 - F(x_1, \ldots, x_n, y)y) = 1.$$

We are not done being clever: the next step is to view this equality in the field of rational functions over K rather than in the polynomial ring, which we can do, since the latter is a subring of the former. We can then plug in $y = 1/F$ and still get an equality, with the advantage of killing the last summand on the left-hand side:

$$\sum_{i=1}^{r} H_i\left(x_1, \ldots, x_n, \frac{1}{F}\right) G_i\left(x_1, \ldots, x_n, \frac{1}{F}\right) = 1.$$

Further,

$$G_i\left(x_1, \ldots, x_n, \frac{1}{F}\right) = g_i(x_1, \ldots, x_n)$$

(G_i is just the name of g_i in the larger polynomial ring and only depends on the first n variables), while

$$H_i\left(x_1, \ldots, x_n, \frac{1}{F}\right) = \frac{h_i(x_1, \ldots, x_n)}{f(x_1, \ldots, x_n)^m},$$

where $h_i \in K[x_1, \ldots, x_n]$ and m is an integer large enough to work for all $i = 1, \ldots, r$. Expressing in terms of a common denominator, we can rewrite the identity as

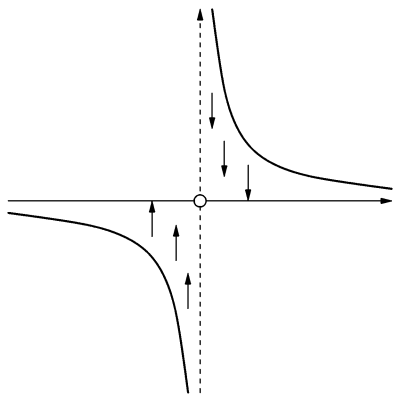
$$\frac{h_1 g_1 + \cdots + h_r g_r}{f^m} = 1,$$

or

$$f^m = h_1 g_1 + \cdots + h_r g_r.$$

We have proved this identity in the field of rational functions; as it only involves polynomials, it holds in $K[x_1, \ldots, x_n]$. The right-hand side is an element of I, so this proves $f \in \sqrt{I}$, and we are done. □

The addition of the variable y in the proof looks more reasonable (and less like a trick) if one views it geometrically. The variety $\mathscr{V}(1-xy)$ in $\mathbb{A}^2_K$ is an ordinary hyperbola:



The projection $(x,y) \mapsto x$ may be used to identify $\mathscr{V}(1-xy)$ with the complement of the origin (that is, $\mathscr{V}(x)$) in the "x-axis" (viewed as $\mathbb{A}^1_K$). Similarly, $\mathscr{V}(1-fy)$ is a way to realize the Zariski *open* set $\mathbb{A}^n_K \smallsetminus \mathscr{V}(f)$ as a Zariski *closed* subset in a space $\mathbb{A}^{n+1}_K$ of dimension one higher.

The proof of Proposition 2.17 hinges on this fact, and this fact is also an important building block in the basic set-up of algebraic geometry, since it can be used to show that the Zariski topology has a basis consisting of (sets which are isomorphic in a suitable sense to) affine algebraic sets.

We will officially record the conclusion we were able to establish concerning the functions $\mathscr{V}$, $\mathscr{I}$:

Corollary 2.18. *Let K be an algebraically closed field. Then for any $n \geq 0$ the functions*

$$\{\textit{algebraic subsets of } \mathbb{A}^n_K\} \underset{\mathscr{V}}{\overset{\mathscr{I}}{\rightleftarrows}} \{\textit{radical ideals in } K[x_1,\ldots,x_n]\}$$

are inverses of each other.

Proof. Proposition 2.17 shows that $\mathscr{I} \circ \mathscr{V}$ is the identity on radical ideals, so $\mathscr{V}$ is injective, and $\mathscr{V}$ is surjective by definition of affine algebraic set. It follows that $\mathscr{V}$ is a bijection and $\mathscr{I}$ is its inverse. □

SUMMARIZING: *If K is algebraically closed,* then studying sets defined by polynomial equations in a space K^n is 'the same thing as' studying radical ideals in a polynomial ring $K[x_1,\ldots,x_n]$.

This correspondence is actually realized even more effectively at the level of K-algebras. We say that a ring is *reduced* if it has no nonzero nilpotents; then R/I is reduced if and only if I is a radical ideal (Exercise 2.8).

Definition 2.19. Let $S \subseteq \mathbb{A}^n_K$ be an algebraic set. The *coordinate ring* of S is the quotient[10]

$$K[S] := \frac{K[x_1, \ldots, x_n]}{\mathscr{I}(S)}.$$ ⌟

Thus, the coordinate ring of an algebraic set is a *reduced, commutative K-algebra of finite type.* The reader will establish a 'concrete' interpretation of this ring, as the ring of 'polynomial functions' on S, in Exercise 2.12. One way to think about the Nullstellensatz (in its manifestation as Corollary 2.18) is that, *if K is algebraically closed,* then *every* reduced commutative K-algebra of finite type R may be realized as the coordinate ring of an affine algebraic set S; the points of S correspond in a natural way to the maximal ideals of R (Exercise 2.14).

In fact, in basic algebraic geometry one defines the *category* of affine algebraic sets over a field K in terms of this correspondence: morphisms of algebraic sets are defined so that they match homomorphisms of the corresponding (reduced, finite-type) K-algebras. (The reader will encounter a more precise definition in Example VIII.1.9.)

This dictionary allows us to translate every 'geometric' feature of an algebraic set (such as dimension, smoothness, etc.) into corresponding 'algebraic' features of the corresponding coordinate rings. Once these key algebraic features have been identified, then one can throw away the restriction of working on reduced finite-type algebras over an algebraically closed field and try to 'do geometry' on any (say commutative, Noetherian) ring. For example, it turns out that PIDs correspond to (certain) smooth curves in the affine geometric world; and then one can try to think of $\mathbb{Z}$ as a 'smooth curve' (although $\mathbb{Z}$ is *not* a reduced finite-type K-algebra over any field K!) and try to use theorems inspired by the geometry of curves to understand features of $\mathbb{Z}$—that is, use geometry to do number theory[11].

Another direction in which these simple considerations may be generalized is by 'gluing' affine algebraic sets into manifold-like objects: sets which may not be affine algebraic sets 'globally' but may be covered by affine algebraic sets. A concrete way to do this is by working in *projective space* rather than affine space; the globalization process can then be carried out in a straightforward way at the algebraic level, where it translates into the study of 'graded' rings and modules—more notions for which, regretfully, we will find no room in this book (except for a glance, in §VIII.4.3). In the second half of the twentieth century a more abstract viewpoint, championed by Alexander Grothendieck and others, has proven to be extremely effective; it has led to the intense study of *schemes,* the current language of choice in algebraic geometry. We have encountered the simplest kind of schemes when we introduced the *spectrum* of a ring R, $\operatorname{Spec} R$, back in §III.4.3.

[10]The notation '$K[S]$' is fairly standard, although it may lead to confusion with the usual polynomial rings; hopefully the context takes care of this. It makes sense to incorporate the name of the field in the notation, since one could in principle change the base field while keeping the 'same' equations encoded in the ideal $\mathscr{I}(S)$.

[11]This is of course a drastic oversimplification.

Exercises

2.1. ▷ Prove Lemma 2.1. [§2.1]

2.2. ▷ Let $k \subseteq \overline{k}$ be an algebraic closure, and let L be an intermediate field. Assume that every polynomial $f(x) \in k[x] \subseteq L[x]$ factors as a product of linear terms in $L[x]$. Prove that $L = \overline{k}$. [§2.1]

2.3. Prove that if k is a countable field, then so is $\overline{k}$.

2.4. ▷ Let k be a field, let $c_1, \ldots, c_m \in k$ be distinct elements, and let $\lambda_1, \ldots, \lambda_m$ be nonzero elements of k. Prove that

$$\frac{\lambda_1}{x - c_1} + \cdots + \frac{\lambda_m}{x - c_m} \neq 0.$$

(This fact is used in the proof of Theorem 2.9.) [§2.2]

2.5. ▷ Let K be a field, let A be a subset of $K[x_1, \ldots, x_n]$, and let I be the ideal generated by A. Prove that $\mathscr{V}(A) = \mathscr{V}(I)$ in $\mathbb{A}^n_K$. [§2.3]

2.6. ▷ Let K be your favorite infinite field. Find examples of subsets $S \subseteq \mathbb{A}^n_K$ which cannot be realized as $V(I)$ for any ideal $I \subseteq K[x_1, \ldots, x_n]$. Prove that if K is a finite field, then every subset $S \subseteq \mathbb{A}^n_K$ equals $V(I)$ for some ideal $I \subseteq K[x_1, \ldots, x_n]$. [§2.3]

2.7. ▷ Let K be a field and n a nonnegative integer. Prove that the set of algebraic subsets of $\mathbb{A}^n_K$ is the family of closed sets of a topology on $\mathbb{A}^n_K$. [§2.3]

2.8. ▷ With notation as in Definition 2.13:

- Prove that the set $\sqrt{I}$ is an ideal of R.
- Prove that $\sqrt{I}$ corresponds to the nilradical of R/I via the correspondence between ideals of R/I and ideals of R containing I.
- Prove that $\sqrt{I}$ is in fact the intersection of all prime ideals of R containing I. (Cf. Exercise V.3.13.)
- Prove that I is radical if and only if R/I is reduced. (Cf. Exercise III.3.13.)

[§2.3]

2.9. ▷ Prove that every affine algebraic set equals $\mathscr{V}(I)$ for a *radical* ideal I. [§2.3]

2.10. Prove that every ideal in a Noetherian ring contains a power of its radical.

2.11. Assume a field is *not* algebraically closed. Find a reduced finite-type K-algebra which is not the coordinate ring of any affine algebraic set.

2.12. ▷ Let K be an infinite field. A *polynomial function* on an affine algebraic set $S \subseteq \mathbb{A}^n_K$ is the restriction to S of (the evaluation function of) a polynomial $f(x_1, \ldots, x_n) \in K[x_1, \ldots, x_n]$. Polynomial functions on an algebraic S manifestly form a ring and in fact a K-algebra. Prove that this K-algebra is isomorphic to the coordinate ring of S. [§2.3, §VIII.1.3, §VIII.2.3]

2.13. Let K be an algebraically closed field. Prove that every reduced commutative K-algebra of finite type is the coordinate ring of an algebraic set S in some affine space $\mathbb{A}^n_K$.

2.14. ▷ Prove that, over an algebraically closed field K, the points of an algebraic set S correspond to the maximal ideals of the coordinate ring $K[S]$ of S, in such a way that if p corresponds to the maximal ideal $\mathfrak{m}_p$, then the value of the function $f \in K[S]$ at p equals the coset of f in $K[S]/\mathfrak{m}_p \cong K$. [§2.3, VIII.1.8]

2.15. ¬ Let K be an algebraically closed field. An algebraic subset S of $\mathbb{A}^n_K$ is *irreducible* if it can*not* be written as the union of two algebraic subsets properly contained in it. Prove that S is irreducible if and only if its ideal $\mathscr{I}(S)$ is prime, if and only if its coordinate ring $K[S]$ is an integral domain.

An irreducible algebraic set is 'all in one piece', like $\mathbb{A}^n_K$ itself, and unlike (for example) $\mathscr{V}(xy)$ in the affine plane $\mathbb{A}^2_K$ with coordinates x, y. Irreducible affine algebraic sets are called (affine algebraic) *varieties*. [2.18]

2.16. ▷ Let K be an algebraically closed field. The field of rational functions $K(x_1, \ldots, x_n)$ is the field of fractions of $K[\mathbb{A}^n_K] = K[x_1, \ldots, x_n]$; every rational function $\alpha = \frac{F}{G}$ (with $G \neq 0$ and F, G relatively prime) may be viewed as defining a function on the open set $\mathbb{A}^n_K \smallsetminus \mathscr{V}(G)$; we say that α is 'defined' for all points in the complement of $\mathscr{V}(G)$.

Let $G \in K[x_1, \ldots, x_n]$ be irreducible. The set of rational functions that are defined in the complement of $\mathscr{V}(G)$ is a subring of $K(x_1, \ldots, x_n)$. Prove that this subring may be identified with the *localization* (Exercise V.4.7) of $K[\mathbb{A}^n_K]$ at the multiplicative set $\{1, G, G^2, G^3, \ldots\}$. (Use the Nullstellensatz.)

The same considerations may be carried out for any irreducible algebraic set S, adopting as field of 'rational functions' $K(S)$ the field of fractions of the integral domain $K[S]$. [2.17, 2.19, §6.3]

2.17. ¬ Let K be an algebraically closed field, and let $\mathfrak{m}$ be a maximal ideal of $K[x_1, \ldots, x_n]$, corresponding to a point p of $\mathbb{A}^n_K$. A *germ* of a function at p is determined by an open set containing p and a function defined on that open set; in our context (dealing with rational functions and where the open set may be taken to be the complement of a function that does not vanish at p) this is the same information as a rational function defined at p, in the sense of Exercise 2.16.

Show how to identify the ring of germs with the localization $K[\mathbb{A}^n_K]_{\mathfrak{m}}$ (defined in Exercise V.4.11).

As in Exercise 2.16, the same discussion can be carried out for any algebraic set. This is the origin of the name 'localization': localizing the coordinate ring of a variety V at the maximal ideal corresponding to a point p amounts to considering only functions defined in a neighborhood of p, thus studying V 'locally', 'near p'. [V.4.7]

2.18. ¬ Let K be an algebraically closed field. Consider the two 'curves' $C_1 : y = x^2$, $C_2 : y^2 = x^3$ in $\mathbb{A}^2_K$ (pictures of the real points of these algebraic sets are shown in Example 2.12).

- Prove that $K[C_1] \cong K[t] = K[\mathbb{A}^1_K]$, while $K[C_2]$ may be identified with the subring $K[t^2, t^3]$ of $K[t]$ consisting of polynomials $a_0 + a_2t^2 + \cdots + a_dt^d$ with zero t-coefficient. (Note that every polynomial in $K[x, y]$ may be written as $f(x) + g(x)y + h(x, y)(y^2 - x^3)$ for uniquely determined polynomials $f(x)$, $g(x)$, $h(x, y)$.)
- Show that C_1, C_2 are both irreducible (cf. Exercise 2.15).
- Prove that $K[C_1]$ is a UFD, while $K[C_2]$ is not.
- Show that the Krull dimension of both $K[C_1]$ and $K[C_2]$ is 1. (This is why these sets would be called 'curves'. You may use the fact that maximal chains of prime ideals in $K[x, y]$ have length 2.)
- The origin $(0, 0)$ is in both C_1, C_2 and corresponds to the maximal ideals $\mathfrak{m}_1$, resp., $\mathfrak{m}_2$, in $K[C_1]$, resp., $K[C_2]$, generated by the classes of x and y.
- Prove that the localization $K[C_1]_{\mathfrak{m}_1}$ is a DVR (Exercise V.4.13). Prove that the localization $K[C_2]_{\mathfrak{m}_2}$ is *not* a DVR. (Note that the relation $y^2 = x^3$ still holds in this ring; prove that $K[C_2]_{\mathfrak{m}_2}$ is not a UFD.)

The fact that a DVR admits a local parameter, that is, a single generator for its maximal ideal (cf. Exercise V.2.20), is a good algebraic translation of the fact that a curve such as C_1 has a single, smooth branch through $(0, 0)$. The maximal ideal of $K[C_2]_{\mathfrak{m}_2}$ cannot be generated by just one element, as the reader may verify. [2.19]

2.19. Prove that the fields of rational functions (Exercise 2.16) of the curves C_1 and C_2 of Exercise 2.18 are isomorphic and both have transcendence degree 1 over k (cf. Exercise 1.27).

This is another reason why we should think of C_1 and C_2 as 'curves'. In fact, it can be proven that the Krull dimension of the coordinate ring of a variety equals the transcendence degree of its field of rational functions. This is a consequence of *Noether's normalization theorem,* a cornerstone of commutative algebra.

2.20. ¬ Recall from Exercise VI.2.13 that $\mathbb{P}^n_K$ denotes the 'projective space' parametrizing lines in the vector space K^{n+1}. Every such line consists of multiples of a nonzero vector $(c_0, \ldots, c_n) \in K^{n+1}$, so that $\mathbb{P}^n_K$ may be identified with the quotient (in the set-theoretic sense of §I.1.5) of $K^{n+1} \smallsetminus \{(0, \ldots, 0)\}$ by the equivalence relation $\sim$ defined by

$$(c_0, \ldots, c_n) \sim (c'_0, \ldots, c'_n) \iff (\exists \lambda \in K^*)\,,\ (c'_0, \ldots, c'_n) = (\lambda c_0, \ldots, \lambda c_n).$$

The 'point' in $\mathbb{P}^n_K$ determined by the vector $(c_0, \ldots, c_n)$ is denoted $(c_0 : \ldots : c_n)$; these are the 'projective coordinates'[12] of the point. Note that there is no 'point' $(0 : \ldots : 0)$.

Prove that the function $\mathbb{A}^n_K \to \mathbb{P}^n_K$ defined by

$$(c_1, \ldots, c_n) \mapsto (1 : c_1 : \ldots : c_n)$$

is a bijection. This function is used to realize $\mathbb{A}^n_K$ as a subset of $\mathbb{P}^n_K$. By using similar functions, prove that $\mathbb{P}^n_K$ can be covered with $n+1$ copies of $\mathbb{A}^n_K$, and relate

[12] This is a convenient abuse of language. Keep in mind that the c_i's are not determined by the point, so they are not 'coordinates' in any strict sense. Their *ratios* are, however, determined by the point.

this fact to the cell decomposition obtained in Exercise VI.2.13. (Suggestion: Work out carefully the case $n = 2$.) [2.21, VIII.4.8]

2.21. ¬ Let $F(x_0, \ldots, x_n) \in K[x_0, \ldots, x_n]$ be a *homogeneous* polynomial. With notation as in Exercise 2.20, prove that the condition '$F(c_0, \ldots, c_n) = 0$' for a point $(c_0 : \ldots : c_n) \in \mathbb{P}^n_K$ is well-defined: it does not depend on the representative $(c_0, \ldots, c_n)$ chosen for the points $(c_0 : \ldots : c_n)$. We can then define the following subset of $\mathbb{P}^n_K$:

$$\mathscr{V}(F) := \{(c_0 : \ldots : c_n) \in \mathbb{P}^n_K \mid F(c_0, \ldots, c_n) = 0\}.$$

Prove that this 'projective algebraic set' can be covered with $n + 1$ affine algebraic sets.

The basic definitions in 'projective algebraic geometry' can be developed along essentially the same path taken in this section for affine algebraic geometry, using 'homogenous ideals' (that is, ideals generated by homogeneous polynomials; see §VIII.4.3) rather than ordinary ideals. This problem shows one way to relate projective and affine algebraic sets, in one template example. [VIII.4.8, VIII.4.11]

---

3. Geometric impossibilities

Very simple considerations in field theory dispose easily of a whole class of geometric problems which had seemed unapproachable for a very long time—problems which preoccupied the Greeks and were only solved in the relatively recent past.

These problems have to do with the construction of certain geometric figures with 'straightedge and compass', that is, subject to certain very strict rules. The practical utility of these constructions would appear to be absolute zero, but the intellectual satisfaction of being able to thoroughly understand matters which stumped extremely intelligent people for a very long time is well worth the little side trip.

3.1. Constructions by straightedge and compass. We begin with two points O, P in the ordinary, real plane. You are allowed to mark ('construct') more points and other geometric figures in the plane, but only according to the following rules:

- If you have constructed two points A, B, then you can draw the line joining them (using your straightedge).
- If you have constructed two points A, B, then you can draw the circle with center at A and containing B (using your compass).
- You can mark any number of points of intersection of any two distinct lines, line and circle, or circles that you have drawn already.

Performing these actions leads to complex (and beautiful) collections of lines and circles; we say that we have 'constructed' a geometric figure if that figure appears as a subset of the whole picture.

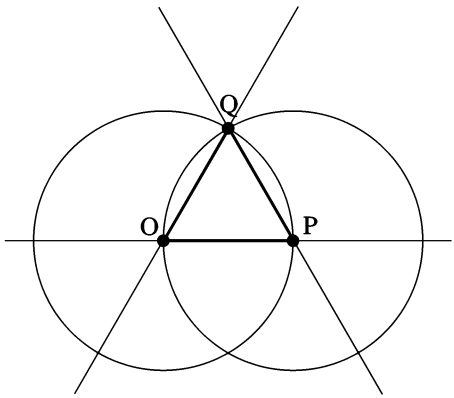
For example, here is a recipe to construct an equilateral triangle with O, P as two of its vertices:

(1) draw the circle with center O, through P;

(2) draw the circle with center P, through O;

(3) let Q be any of the two points of intersection of the two circles;

(4) draw the lines through O, P; O, Q; and P, Q.



Of course O, P, Q are vertices of an equilateral triangle.

It is a pleasant exercise to try to come up with a specific sequence of basic moves constructing a given figure or geometric configuration. The enterprising reader could try to produce the vertices of a *pentagon* by a straightedge-and-compass construction, without looking it up and before we learn enough to thwart pure geometric intuition.

The general question is to decide whether a figure can or cannot be constructed this way. Three problems of this kind became famous in antiquity:

- trisecting angles;
- squaring circles;
- doubling cubes.

For example, one assumes one has already constructed two lines forming an angle θ and asks whether one can construct two lines forming $\theta/3$. This is trivially possible for *some* constructible θ: we just constructed an angle of $\pi/3$, so evidently we were able to trisect the angle π; but can this be done for *all* constructible angles?

The answer is no. Similarly, it is not possible to construct a square whose area equals the area of a (constructible) circle, and it is not possible to construct the side of a cube whose volume is twice as large as a cube with given (constructible) side.

Field theory will allow us to establish all this[13]. Later we will return to these constructions and study the question of the constructibility of *regular polygons* with a given side: it is very easy to construct equilateral triangles (we just did), squares, hexagons; it is not too hard to construct pentagons; but the question

[13]However, for the impossibility of squaring circles we will use the transcendence of π, which we are taking on faith.

of which polygons are constructible and which are not connects beautifully with deeper questions in field theory.

The three problems listed above have an illustrious history; they apparently date back to at least 414 B.C., if one is to take this fragment from Aristophanes' *"The birds"* as an indication:

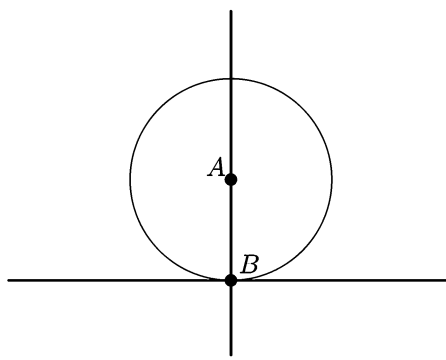
> *METON: By measuring with a straightedge there*
> *the circle becomes a square with a court within.*[14]

The problems depend on the precise restrictions imposed on the constructions. For example, one *can* trisect angles if one is allowed to mark points on the straightedge (thus making it a *ruler*); in fact, this was already known to Archimedes (Exercise 3.13).

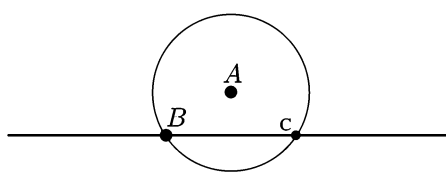
Translating these geometric problems into algebra requires establishing the feasibility of a few basic constructions.

First of all, one can construct a line containing a given point A and perpendicular to a given line ℓ (thus, ℓ contains at least another constructible point B). For this, draw the circle with center A and containing B;

—if this circle only intersects ℓ at B, then the line through A and B is perpendicular to ℓ:

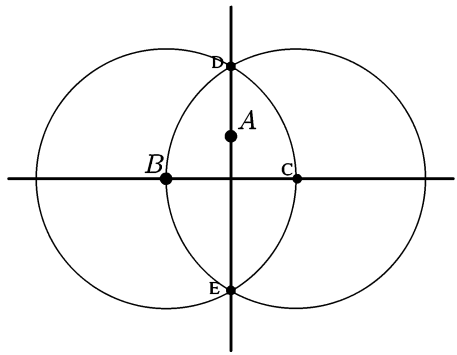


—otherwise, the circle intersects ℓ at a second point C (whether A is on ℓ or not):

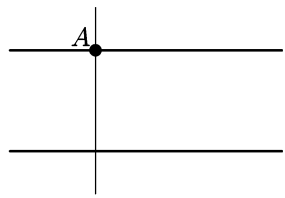


and once C is determined, the line through A and perpendicular to ℓ may be found by joining the two points of intersection D, E of the two circles with centers at B, resp., C, and containing C, resp., B:

[14]We thank Matilde Marcolli for the translation. For those who are less ignorant than we are:
ὀρθῶι μετρήσω κανόνι προστιθείς ἵνα
ὁ κύκλος γενηταί σοι τετράγωνος καν μέσωι ἀγορά

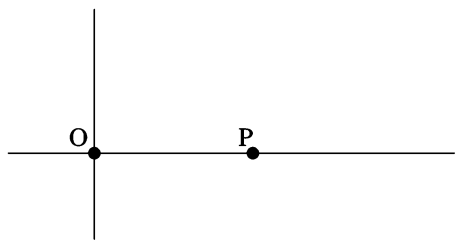


Applying this construction twice gives the line through a point A and *parallel* to a given line:

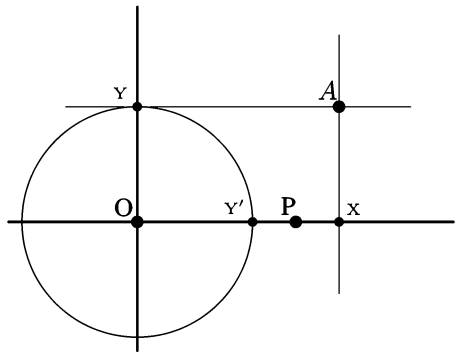


which is often handy.

Judicious application of these operations allows us to bring a Cartesian reference system into the picture. Starting with the initial two points O, P, we can construct perpendicular Cartesian axes centered at O, with P marking (say) the point $(1,0)$ on the 'x-axis':



and constructing a point $A = (x, y)$ is equivalent to constructing its projections $X = (x, 0)$, $Y = (0, y)$ onto the axes, and in fact it is equivalent to constructing the two points $X = (x, 0)$ and $Y' = (y, 0)$:



It follows that determining which figures are constructible by straightedge and compass is equivalent to determining which numbers may be realized as coordinates of a constructible point.

Definition 3.1. A real number r is *constructible* if the point $(r, 0)$ is constructible with straightedge and compass (assuming $O = (0, 0)$ and $P = (1, 0)$, as above). We will denote by $\mathscr{C}_{\mathbb{R}} \subseteq \mathbb{R}$ the set of constructible real numbers. ⌟

Also, we can identify the real plane with $\mathbb{C}$, placing O at 0 and P at 1, and we say that $z = x+iy$ is *constructible* if the point (x, y) is constructible by straightedge and compass. We will denote by $\mathscr{C}_{\mathbb{C}} \subseteq \mathbb{C}$ the set of constructible complex numbers. Summarizing the foregoing discussion, we have proved:

Lemma 3.2. *A point (x, y) is constructible by straightedge and compass if and only if $x + iy \in \mathscr{C}_{\mathbb{C}}$, if and only if $x, y \in \mathscr{C}_{\mathbb{R}}$.*

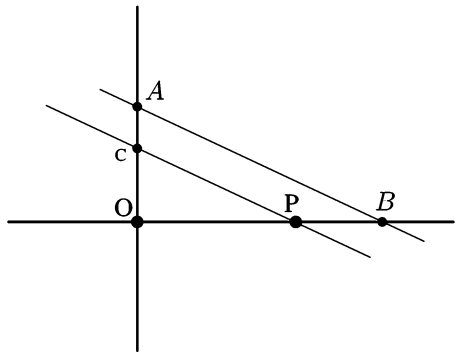
The next obvious question is what kind of structure these sets of constructible numbers carry, and this prompts us to look at a few more basic straightedge-and-compass constructions.

Lemma 3.3. *The subset $\mathscr{C}_{\mathbb{R}} \subseteq \mathbb{R}$ of constructible numbers is a subfield of $\mathbb{R}$. Likewise, $\mathscr{C}_{\mathbb{C}}$ is a subfield of $\mathbb{C}$, and in fact $\mathscr{C}_{\mathbb{C}} = \mathscr{C}_{\mathbb{R}}(i)$.*

Proof. The set $\mathscr{C}_{\mathbb{R}} \subseteq \mathbb{R}$ is nonempty, so in order to show it is a field, we only need to show that it is closed with respect to subtraction and division by a nonzero constructible number (cf. Exercise II.6.2).

The reader will check that $\mathscr{C}_{\mathbb{R}}$ is closed under subtraction (Exercise 3.2). To see that $\mathscr{C}_{\mathbb{R}}$ is closed under division, let $a, b \in \mathscr{C}_{\mathbb{R}}$, with $b \neq 0$. Since $A = (0, a)$ and $B = (b, 0)$ are constructible by hypothesis, we can construct C as the y-intersect of the line through $P = (1, 0)$ and parallel to the line through A and B, as in the following picture[15]:

[15] In the picture we are assuming $a > 0$ and $b > 1$; the reader will check that other possibilities lead to the same conclusion.



Since the triangles AOB and COP are similar, we see that $C = (0, a/b)$; it follows that a/b is constructible.

The fact that $\mathscr{C}_{\mathbb{C}}$ is a subfield of $\mathbb{C}$ is an immediate consequence of the fact that $\mathscr{C}_{\mathbb{R}}$ is a field, and the proof of this is left to the enjoyment of the reader (Exercise 3.7). The fact that $\mathscr{C}_{\mathbb{C}} = \mathscr{C}_{\mathbb{R}}(i)$ is a restatement of the fact that $x + iy \in \mathscr{C}_{\mathbb{C}}$ if and only if x and y are in $\mathscr{C}_{\mathbb{R}}$. □

We may view $\mathscr{C}_{\mathbb{R}}$ and $\mathscr{C}_{\mathbb{C}}$ as extensions of $\mathbb{Q}$, drawing a bridge between constructibility by straightedge and compass and field theory: we will be able to understand constructibility of geometric figures if we can understand the field extensions

$$\mathbb{Q} \subseteq \mathscr{C}_{\mathbb{R}} \subseteq \mathscr{C}_{\mathbb{C}}.$$

Happily, we *can* understand these extensions!

3.2. Constructible numbers and quadratic extensions. Our goal is to prove the following amazingly explicit description of $\mathscr{C}_{\mathbb{R}}$ (immediately implying one for $\mathscr{C}_{\mathbb{C}}$).

Theorem 3.4. *Let $\gamma \in \mathbb{R}$. Then $\gamma \in \mathscr{C}_{\mathbb{R}}$ if and only if there exist real numbers $\delta_1, \ldots, \delta_k$ such that $\forall j = 1, \ldots, k$*

$$[\mathbb{Q}(\delta_1, \ldots, \delta_j) : \mathbb{Q}(\delta_1, \ldots, \delta_{j-1})] = 2$$

and $\gamma \in \mathbb{Q}(\delta_1, \ldots, \delta_k)$.

In other words, $\gamma \in \mathbb{R}$ is constructible if and only if it can be placed in the top field of a sequence of (real) *quadratic extensions* over $\mathbb{Q}$. Since $\mathscr{C}_{\mathbb{C}} = \mathscr{C}_{\mathbb{R}}(i)$, the same statement holds for constructible *complex* numbers, including i in the list of δ_j (or simply allowing the δ_j to be complex numbers; cf. Exercise 3.9).

The proof of this theorem is as explicit as one can possibly hope. From a given straightedge-and-compass construction of a point (x, y) one can obtain an explicit sequence $\delta_1, \ldots, \delta_k$ as in the statement, such that x and y belong to the extension $\mathbb{Q}(\delta_1, \ldots, \delta_k)$; conversely, from any element γ of any such extension one can obtain an explicit construction of a point $(\gamma, 0)$ by straightedge and compass.

Proof. Let's first argue in the 'geometry to algebra' direction. A configuration of points, lines, and circles obtained by a straightedge-and-compass construction

may be described by the coordinates of the points and the equations of the lines and circles. Suppose that at one stage in a given construction all coordinates of all points and all coefficients in the equations of lines and circles belong to a field F; we will say that the configuration is *defined over* F.

Then we claim that for every object constructed at the next stage, there exists a number $\delta \in \mathbb{R}$, of degree *at most* 2 over F, such that the new configuration is defined over $F(\delta)$. The 'only if' part of the theorem follows by induction on the number of steps in the construction, since at the beginning the configuration (that is, the pair of points $O = (0,0)$, $P = (1,0)$) is defined over $\mathbb{Q}$.

Verifying our claim amounts to verifying it for the basic operations defining straightedge-and-compass constructions. The reader will check (Exercise 3.5) that the point of intersection of two lines defined over F has coordinates in F and that lines and circles determined by points with coordinates in F are defined over F. So $\delta = 1$ works in all these cases.

For the intersection of a line ℓ and a circle C, assume that ℓ is not parallel to the y-axis (the argument is entirely analogous otherwise) and that it does meet C; let

$$y = mx + r$$

be the equation of ℓ and let

$$x^2 + y^2 + ax + by + c = 0$$

be the equation of C. We are assuming that $a, b, c, m, r \in F$. Then the x-coordinates of the points of intersection of ℓ and C are the solutions of the equation

$$x^2 + (mx + r)^2 + ax + b(mx + r) + c = 0.$$

The 'quadratic formula' shows that these coordinates belong to the field $F(\sqrt{D})$, where D is the discriminant of this polynomial: explicitly,

$$D = (2mr + bm + a)^2 - 4(m^2 + 1)(r^2 + br + c),$$

but this is unimportant. What is important is that $D \in F$; hence $\delta = \sqrt{D}$ satisfies our requirement.

For the intersection of two (distinct) circles defined over F, nothing new is needed: if

$$\begin{cases} x^2 + y^2 + a_1 x + b_1 y + c_1 = 0, \\ x^2 + y^2 + a_2 x + b_2 y + c_2 = 0 \end{cases}$$

are two circles, subtracting the two equations shows that their points of intersection coincide with the points of intersection of a circle and a line:

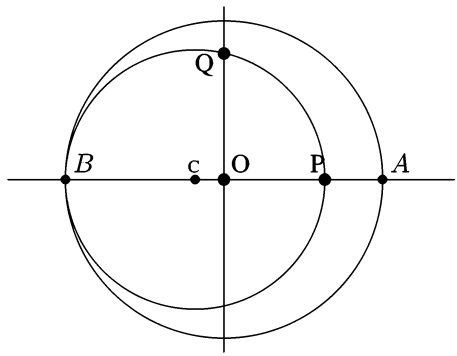
$$\begin{cases} x^2 + y^2 + a_1 x + b_1 y + c_1 = 0, \\ (a_1 - a_2)x + (b_1 - b_2 y) + (c_1 - c_2) = 0, \end{cases}$$

with the same conclusion as in the previous case.

This completes the verification of the 'only if' part of the theorem.

To prove that every element of an extension as stated is constructible, again argue by induction: it suffices to show that if (i) $\delta \in \mathbb{R}$, (ii) all elements of F are constructible, and (iii) $r = \delta^2 \in F$, then δ is constructible (note that in order to

construct an element of degree 2 over F, it suffices to construct the square root of the discriminant of its minimal polynomial). Therefore, all we have to show is that we can 'take square roots' by a straightedge-and-compass construction. Here is the picture (if $r > 1$):



If $A = (r, 0)$ is constructible, so is $B = (-r, 0)$; C is the midpoint of the segment BP (midpoints are constructible, Exercise 3.1); the circle with center C and containing P intersects the positive y-axis at a point $Q = (0, \delta)$, and elementary geometry shows that $\delta^2 = r$. Therefore δ is constructible, concluding the proof of the theorem. □

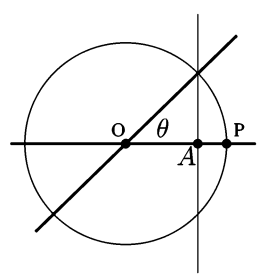
For example, one can construct an angle of 3°: allegedly

$$\cos 3^\circ = \frac{1}{8}(\sqrt{3}+1)\sqrt{5+\sqrt{5}} + \frac{1}{16}(\sqrt{6}-\sqrt{2})(\sqrt{5}-1),$$

and this expression shows that

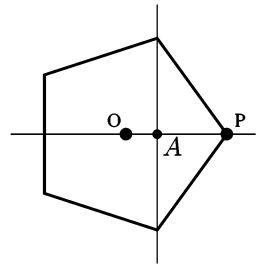
$$\cos 3^\circ \in \mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})(\sqrt{5+\sqrt{5}});$$

that is (by Theorem 3.4), $\cos 3^\circ$ is constructible. Of course once $A = (\cos\theta, 0)$ is constructed, so is the angle θ:



Here is another example of a 'constructive' application of Theorem 3.4:

Example 3.5. Regular pentagons are constructible.



Indeed, it suffices to construct the point $A = (\cos(2\pi/5), 0)$, and it so happens that $\gamma = \cos(2\pi/5)$ satisfies

$$(*) \qquad 4\gamma^2 + 2\gamma - 1 = 0$$

(in fact, γ is half of the inverse of the *golden ratio:* $\gamma = \frac{\sqrt{5}-1}{4}$). It follows that $\gamma \in \mathbb{Q}(\sqrt{5})$, and hence it is constructible by Theorem 3.4. In fact, the proof shows how to construct $\sqrt{5}$, so the reader should now have no difficulty producing a straightedge-and-compass construction of a regular pentagon (Exercise 3.6). ⌟

We will come back to constructions of regular polygons once we have studied field theory a little more thoroughly (cf. §7.2). By the way, how does one figure out that $\gamma = \cos(2\pi/5)$ should satisfy the identity (*)? This is easy modulo some complex number arithmetic; see Exercise 3.11.

3.3. Famous impossibilities. Theorem 3.4 easily settles the three problems mentioned in §3.1, via the following immediate consequence:

Corollary 3.6. *Let $\gamma \in \mathscr{C}_{\mathbb{C}}$ be a constructible number. Then $[\mathbb{Q}(\gamma) : \mathbb{Q}]$ is a power of 2.*

Proof. By Lemma 3.3 and Theorem 3.4, there exist $\delta_1 \ldots, \delta_k \in \mathbb{R}$ such that

$$\gamma \in \mathbb{Q}(\delta_1, \ldots, \delta_k, i),$$

and each δ_j has degree ≤ 2 over $\mathbb{Q}(\delta_1, \ldots, \delta_{j-1})$. Repeated application of Proposition 1.10 shows that

$$[\mathbb{Q}(\delta_1, \ldots, \delta_k, i) : \mathbb{Q}]$$

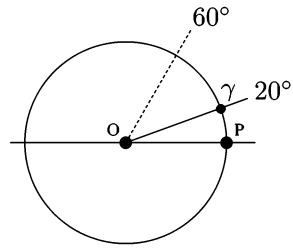
is a power of 2, and since

$$\mathbb{Q} \subseteq \mathbb{Q}(\gamma) \subseteq \mathbb{Q}(\delta_1, \ldots, \delta_k, i),$$

the statement follows from Corollary 1.11. □

In particular, constructible *real* numbers must satisfy the same condition.

Consider the problem of trisecting an angle. We know we can construct an angle of 60° (this is a subproduct of the construction of an equilateral triangle). Constructing an angle of 20° is equivalent (Exercise 3.10) to constructing the complex number γ on the unit circle and with argument $\pi/9$:



Now $\gamma^9 = -1$; that is, γ satisfies the polynomial

$$t^9 + 1 = (t^3 + 1)(t^6 - t^3 + 1).$$

It does not satisfy $t^3 + 1$; therefore its minimal polynomial over $\mathbb{Q}$ is a factor of

$$t^6 - t^3 + 1.$$

However, this polynomial is irreducible over $\mathbb{Q}$ (substitute $t \mapsto t - 1$, and apply Eisenstein's criterion), so

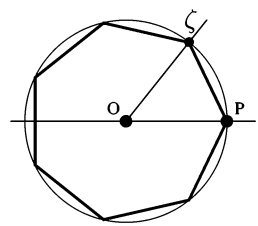
$$[\mathbb{Q}(\gamma) : \mathbb{Q}] = 6.$$

By Corollary 3.6, γ is not constructible. Therefore there exist constructible angles which cannot be trisected.

Similarly, cubes cannot be doubled because that would imply the constructibility of $\sqrt[3]{2}$, which has degree 3 over $\mathbb{Q}$, contradicting Corollary 3.6.

Squaring circles amounts to constructing π, and π is not even algebraic (although, as we have mentioned, the proof of this fact is not elementary), so this is also not possible.

As another example, constructing a regular 7-gon would amount to constructing a 7-th (complex) root of 1, ζ:



By definition, ζ satisfies

$$t^7 - 1 = (t - 1)(t^6 + t^5 + \cdots + t + 1);$$

as $\zeta \neq 1$, ζ must satisfy the *cyclotomic* polynomial $t^6 + \cdots + 1$. This is irreducible (Example V.5.19); hence again we find that ζ has degree 6 over $\mathbb{Q}$, and Corollary 3.6 implies that the regular 7-gon cannot be constructed with straightedge and compass.

Of course '7' is not too special: if p is a positive prime integer, the cyclotomic polynomial of degree $p - 1$ is irreducible (Example V.5.19 again); hence

$$[\mathbb{Q}(\zeta_p) : \mathbb{Q}] = p - 1,$$

where ζ_p is the complex p-th root of 1 with argument $2\pi/p$. Therefore, Corollary 3.6 reveals that if p is prime, then the regular p-gon can be constructed *only if* $p - 1$ is a power of 2. This is even more restrictive than it looks at first, since if $p = 2^k + 1$

is prime, then necessarily k is itself a power of 2 (Exercise 3.15). Primes of the form $2^{2^\ell}+1$ are called *Fermat primes:*

$$3\,,\ 5\,,\ 17\,,\ 257\,,\ 65537$$

(the 'next one', $2^{32}+1 = 4294967297 = 641 \cdot 6700417$, is not prime; in fact, no one knows if there are any *other* Fermat primes, and yet some people conjecture that there are infinitely many).

These considerations alone do not tell us that *if* p is a Fermat prime, *then* the regular p-gon can be constructed with straightedge and compass; but they do tell us that the next case to consider would be $p = 2^4+1 = 17$, and it just so happens that

$$\cos\frac{2\pi}{17} = \frac{\sqrt{17}-1+\sqrt{2}\sqrt{34+6\sqrt{17}+\sqrt{2}(\sqrt{17}-1)\sqrt{17-\sqrt{17}}-8\sqrt{2}\sqrt{17+\sqrt{17}}}+\sqrt{2}\sqrt{17-\sqrt{17}}}{16}$$

as noted by Gauss at age 19. Therefore (by Theorem 3.4) the 17-gon *is* constructible by straightedge and compass. As we have announced already, the situation will be further clarified soon, allowing us to bypass heavy-duty trigonometry.

Exercises

3.1. ▷ Prove that if A, B are constructible, then the midpoint of the segment AB is also constructible. Prove that if two lines ℓ_1, ℓ_2 are constructible and not parallel, then the two lines bisecting the angles formed by ℓ_1 and ℓ_2 are also constructible. [§3.2]

3.2. ▷ Prove that if a, b are constructible numbers, then so is $a-b$. [§3.1]

3.3. Find an explicit straightedge-and-compass construction for the product of two real numbers.

3.4. Show how to square a *triangle* by straightedge and compass.

3.5. ▷ Let F be a subfield of $\mathbb{R}$.

- Let $A = (x_A, y_A)$, $B = (x_B, y_B)$ be two points in $\mathbb{R}^2$, with x_A, y_A, $x_B, y_B \in F$. Prove that the line through A, B is defined over F (that is, it admits an equation with coefficients in F).
- Prove that the circle with center at A and containing B is defined over F.
- Let ℓ_1, ℓ_2 be two distinct, nonparallel lines in $\mathbb{R}^2$, defined over F, and let $(x, y) = \ell_1 \cap \ell_2$. Prove that $x, y \in F$.

[§3.2]

3.6. ▷ Devise a way to construct a regular pentagon with straightedge and compass. [§3.2]

3.7. ▷ Identify the (real) plane with $\mathbb{C}$, and place O, P at $0, 1 \in \mathbb{C}$. Let $\mathscr{C}_{\mathbb{C}}$ be the set of all constructible points, viewed as a subset of $\mathbb{C}$. Prove that $\mathscr{C}_{\mathbb{C}}$ is a sub*field* of $\mathbb{C}$. [§3.1]

3.8. For $\delta \in \mathbb{C}$, $\delta \neq 0$, let θ_δ be the argument of δ (that is, the angle formed by the line through 0 and δ with the real axis). Prove that $\delta \in \mathscr{C}_{\mathbb{C}}$ if and only if $|\delta|$, $\cos\theta_\delta$, $\sin\theta_\delta$ are all constructible *real* numbers.

3.9. ▷ Let $\gamma_1, \ldots, \gamma_k \in \mathscr{C}_{\mathbb{C}}$ be constructible complex numbers, and let K be the field $\mathbb{Q}(\gamma_1, \ldots, \gamma_k) \subseteq \mathscr{C}_{\mathbb{C}}$. Let δ be a complex number such that $[K(\delta) : K] = 2$. Prove that $\delta \in \mathscr{C}_{\mathbb{C}}$.

Deduce that there are no irreducible polynomials of degree 2 over $\mathscr{C}_{\mathbb{C}}$ and that $\mathscr{C}_{\mathbb{C}}$ is the smallest subfield of $\mathbb{C}$ with this property. [§3.2]

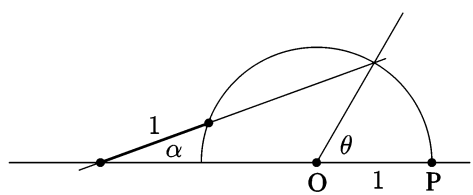
3.10. ▷ Prove that if two lines in any constructible configuration form an angle of θ, then the line through O forming an angle of θ clockwise with respect to the line OP is constructible.

Deduce that an angle θ is constructible anywhere in the plane if and only if $\cos\theta$ is constructible. [§3.3]

3.11. ▷ Verify that $\gamma = \cos(2\pi/5)$ satisfies the relation (*) given in §3.2. (If $\sigma = \sin(2\pi/5)$, note that $z = \gamma + i\sigma$ satisfies $z^5 - 1$.) [§3.2]

3.12. Prove that the angles of 1° and 2° are not constructible. (Hint: Given what we know at this point, you only need to recall that there exist trigonometric formulas for the sum of two angles; the exact shape of these formulas is not important.) For what integers n is the angle $n°$ constructible?

3.13. ▷ Prove that $\alpha = \frac{\theta}{3}$ in the following picture:



This says that angles *can* be trisected if we allow the use of a 'ruler', that is, a straightedge with markings (here, the construction works if we can mark a fixed distance of 1 on the ruler). Apparently, this construction was known to Archimedes. [§3.1]

3.14. Prove that the regular 9-gon is not constructible.

3.15. ▷ Prove that if $2^k + 1$ is prime, then k is a power of 2. [§3.3]

---

4. Field extensions, II

It is time to continue our survey of different flavors of field extensions. The keywords here are *splitting fields, normal, separable.*

4.1. Splitting fields and normal extensions. In §2 we have constructed the algebraic closure $\overline{k}$ of any given field k: *every* polynomial in $k[x]$ factors as a product of linear terms (that is, *'splits'*) in $\overline{k}[x]$, and $\overline{k}$ is the 'smallest' extension of k satisfying this property.

Here is an analogous, but more modest, requirement: given a subset $\mathscr{F} \subseteq k[x]$ of polynomials, construct an extension $k \subseteq F$ such that every polynomial *in* $\mathscr{F}$ splits as a product of linear terms over F, and require F to be as small as possible with this property. We then call F the *splitting field* for $\mathscr{F}$. In practice we will only be interested in the case in which $\mathscr{F}$ is a *finite* collection of polynomials $f_1(x), \dots, f_r(x)$; then requiring that each $f_i(x)$ splits over F is equivalent to requiring that the product $f_1(x) \cdots f_r(x)$ splits over F.

That is, for our purposes it is not restrictive to assume that $\mathscr{F}$ consists of a single polynomial $f(x) \in k[x]$.

Definition 4.1. Let k be a field, and let $f(x) \in k[x]$ be a polynomial of degree d. The *splitting field* for $f(x)$ over k is an extension F of k such that

$$f(x) = c \prod_{i=1}^{d} (x - \alpha_i)$$

splits in $F[x]$, and further $F = k(\alpha_1, \dots, \alpha_d)$ is generated over k by the roots of $f(x)$ in F. ⌟

Note that it is clear that *a* splitting field exists: given $f(x) \in k[x]$, the subfield $F \subseteq \overline{k}$ generated over k by the roots of $f(x)$ in $\overline{k}$ satisfies the requirements in Definition 4.1. But we have written *the* splitting field, and this is justified by the uniqueness part of the following basic observation, which also evaluates 'how big' this extension can be.

Lemma 4.2. *Let k be a field, and let $f(x) \in k[x]$. Then the splitting field F for $f(x)$ over k is unique up to isomorphism, and $[F : k] \leq (\deg f)!$.*

In fact, if $\iota : k' \to k$ is any isomorphism of fields and $g(x) \in k'[x]$ is such that $f(x) = \iota(g(x))$, then ι extends to an isomorphism of any splitting field of $g(x)$ over k' to any splitting field of $f(x)$ over k.

Proof. We first construct explicitly a splitting field and obtain the bound on the degree mentioned in the statement. Then we prove the second part, which implies uniqueness up to isomorphism.

The construction of a splitting field and the given bound on the degree are an easy application of our basic simple extension found in Proposition V.5.7. Arguing inductively, assume the splitting field has been constructed and the bound has been proved for all fields and all polynomials of degree $(\deg f - 1)$. Let $q(t)$ be any irreducible factor of $f(t)$ over k; then

$$k \subseteq F' := \frac{k[t]}{(q(t))}$$

is an extension of degree $\deg q \leq \deg f$, in which $q(x)$ (and hence $f(x)$) has a root (the coset α of t) and therefore a linear factor $x - \alpha$. The polynomial $h(x) :=$

$f(x)/(x-\alpha) \in F'[x]$ has degree $(\deg f - 1)$; therefore a splitting field F exists for $h(x)$, and

$$[F : F'] \leq (\deg f - 1)!.$$

The factors of $f(x)$ over F are $(x-\alpha)$ and the factors of $h(x)$, which are linear; it follows that F is a splitting field for $f(x)$ and

$$[F : k] = [F : F'][F' : k] \leq (\deg f)(\deg f - 1)! = (\deg f)!$$

as stated.

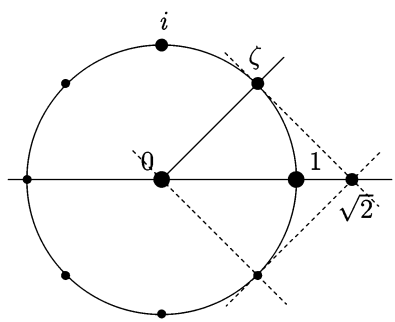
To prove that isomorphisms $\iota : k' \to k$ extend to isomorphisms of splitting fields as stated, let G be a splitting field for $g(x)$ and consider the composition $k' \to k \subseteq \overline{k}$. Since the extension $k' \subseteq G$ is algebraic, by Lemma 2.8 there exists a morphism of extensions $\iota : G \to \overline{k}$ over k'. Since G is generated over k' by the roots of $g(x)$ in G, we have

$$\iota(G) = k(\alpha_1, \ldots, \alpha_d) \subseteq \overline{k},$$

where the α_i's are the roots of $f(x) = \iota(g(x))$ in $\overline{k}$. Therefore $L := \iota(G)$ is independent of the chosen isomorphism ι and splitting field G. Applying this observation to the given ι and to the identity $k \to k$ proves the statement (as both G and F are isomorphic to L). □

Example 4.3. By definition, $\mathbb{Q}(i)$ is the splitting field of x^2+1 over $\mathbb{Q}$, and $\mathbb{C}$ is the splitting field for the same polynomial, over $\mathbb{R}$. ⌟

Example 4.4. The splitting field F of $x^8 - 1$ over $\mathbb{Q}$ is generated by $\zeta := e^{2\pi i/8}$: indeed, the roots of $x^8 - 1$ are all the 8-th roots of 1, and all of them are powers of ζ:



In fact, ζ is a root of the polynomial x^4+1, which is irreducible over $\mathbb{Q}$; therefore $F = \mathbb{Q}(\zeta)$ is 'already' the splitting field of x^4+1. The degree of F over $\mathbb{Q}$ is

$$[\mathbb{Q}(\zeta) : \mathbb{Q}] = 4,$$

way less than the bounds 8!, 4! obtained in Lemma 4.2.

To understand this splitting field (even) better, note that $i = \zeta^2$ is in F, and so is $\sqrt{2} = \zeta + \zeta^7$; thus F contains $\mathbb{Q}(i, \sqrt{2})$. Conversely, $\zeta = \frac{\sqrt{2}}{2}(1+i) \in \mathbb{Q}(i, \sqrt{2})$. Therefore, the splitting field of x^4+1 (a.k.a. the splitting field of x^8-1) is $\mathbb{Q}(i, \sqrt{2})$. Analyzing $\mathbb{Q} \subseteq \mathbb{Q}(i, \sqrt{2})$ (as we did for the extension in Example 1.19) shows that its group of automorphisms over $\mathbb{Q}$ is $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. ⌟

Example 4.5. *Variation on the theme:* $x^4 - 1$. The situation changes if instead of $x^4 + 1$ we consider $x^4 - 1$: this polynomial factors over $\mathbb{Q}$,

$$x^4 - 1 = (x - 1)(x + 1)(x^2 + 1),$$

and it follows that the splitting field is the same as for $x^2 + 1$, that is, just $\mathbb{Q}(i)$. ⌟

Example 4.6. *Variation on the theme:* $x^4 + 2$. The overoptimistic reader may now hope that the difference between the splitting fields of $x^4 + 1$ vs. $x^4 - 1$ over $\mathbb{Q}$ is just due to the fact that the first polynomial is irreducible over $\mathbb{Q}$ and the second is not. This example will nip any such guess in the bud. With notation as in Example 4.4, the roots of $x^4 + 2$ are

$$\sqrt[4]{2}\zeta\,,\ \sqrt[4]{2}\zeta^3\,,\ \sqrt[4]{2}\zeta^5\,,\ \sqrt[4]{2}\zeta^7.$$

Therefore, with $K = \mathbb{Q}(\sqrt[4]{2}\zeta, \sqrt[4]{2}\zeta^3, \sqrt[4]{2}\zeta^5, \sqrt[4]{2}\zeta^7)$ the splitting field of $x^4 + 2$,

$$K \subseteq \mathbb{Q}(\zeta, \sqrt[4]{2}) = \mathbb{Q}(i, \sqrt{2}, \sqrt[4]{2}) = \mathbb{Q}(i, \sqrt[4]{2}).$$

On the other hand, $\sqrt{2} = (\sqrt[4]{2}\zeta)^3/(\sqrt[4]{2}\zeta^3) \in K$; hence $i = (\sqrt[4]{2}\zeta)^2/\sqrt{2} \in K$; hence $\zeta = \frac{\sqrt{2}}{2}(1 + i) \in K$; hence $\sqrt[4]{2} = (\sqrt[4]{2}\zeta)/\zeta \in K$. Therefore, $\mathbb{Q}(i, \sqrt[4]{2}) \subseteq K$, and the conclusion is that the splitting field of $x^4 + 2$ equals $K = \mathbb{Q}(i, \sqrt[4]{2})$. A simple degree computation (Exercise 4.3) shows $[K : \mathbb{Q}] = 8$ and in particular the splitting field of $x^4 + 2$ is certainly not isomorphic to the splitting field of $x^4 + 1$. ⌟

Examples such as these have always left me with the impression that field theory must be rather mysterious: innocent-looking variations on the parameters of a problem may cause dramatic changes in the corresponding field extensions. On the plus side, this hints that field theory can indeed be a very precise tool in (for example) the study of polynomials.

Splitting fields will play an important role in the rest of the story. They are even more special than they may appear to be at first: it turns out that not only do they split the given polynomial, but they also automatically split any irreducible polynomial which dares touch them with a root. This makes splitting fields *normal extensions:*

Definition 4.7. A field extension $k \subseteq F$ is *normal* if for every irreducible polynomial $f(x) \in k[x]$, $f(x)$ has a root in F if and only if $f(x)$ splits as a product of linear factors over F. ⌟

Theorem 4.8. *A field extension $k \subseteq F$ is finite and normal if and only if F is the splitting field of some polynomial $f(x) \in k[x]$.*

Proof. Assume $k \subseteq F$ is finite and normal. Then F is finitely generated: $F = k(\alpha_1, \ldots, \alpha_r)$ with α_i algebraic over k. Let $p_i(t)$ be the minimal polynomial of α_i over k. As F is normal over k, each $p_i(t)$ splits completely over F, and hence so does $f(t) = p_1(t) \cdots p_r(t)$. It follows that F is the splitting field of $f(x)$.

Conversely, assume that F is a splitting field for a polynomial $f(x) \in k[x]$, and let $p(x) \in k[x]$ be an irreducible polynomial, such that F contains a root α of $p(x)$. Viewing F as a subfield of the algebraic closure $\overline{k}$, let $\beta \in \overline{k}$ be any other root of $p(x)$; we are going to show that $\beta \in F$. This will prove that F contains all roots of $p(x)$, implying that $k \subseteq F$ is normal, and $k \subseteq F$ is finite by Lemma 4.2.

By Proposition 1.5, there exists an isomorphism $\iota : k(\alpha) \to k(\beta)$ extending the identity on k and sending $\alpha \mapsto \beta$. We also consider the subfield $F(\beta) \subseteq \overline{k}$, viewed as an extension of $k(\beta)$. Putting everything in one diagram:

$$\begin{array}{ccccc} & & k(\alpha) & \hookrightarrow & F & \hookrightarrow & \overline{k} \\ & \nearrow & \downarrow{\scriptstyle \iota} & & & & \\ k & & & & & & \\ & \searrow & & & & & \\ & & k(\beta) & \hookrightarrow & F(\beta) & \hookrightarrow & \overline{k} \end{array}$$

(Note: If we join the two copies of $\overline{k}$ on the right by the identity map, the corresponding diagram does not commute: ι sends $\alpha \in \overline{k}$ in the top row to $\beta \in \overline{k}$ in the bottom row.) Now observe that F may be viewed as the splitting field of $f(x)$ *over* $k(\alpha)$*:* indeed, it contains all the roots of $f(x)$ and is generated over k (and hence over $k(\alpha)$) by these roots. By the same token, $F(\beta)$ is the splitting field for $f(x)$ over $k(\beta)$. By Lemma 4.2, ι extends to an isomorphism $\iota' : F \to F(\beta)$. Since ι restricts to the identity on k, ι' is an isomorphism of k-vector spaces; in particular, $\dim_k F = \dim_k F(\beta)$ (keep in mind that splitting fields are finite extensions).

Now consider the *different* k-linear map of k-vector spaces $i : F \to F(\beta)$ simply given by the inclusion within $\overline{k}$:

$$k \subseteq F \subseteq F(\beta) \subseteq \overline{k}.$$

Since F and $F(\beta)$ are k-vector spaces of the same *finite* dimension, i must *also* be an isomorphism. In other words,

$$[F(\beta) : F] = 1;$$

that is, $\beta \in F$ as needed. □

Remark 4.9. This argument is somewhat delicate. With notation as in the proof, the diagram

$$\begin{array}{ccc} k(\alpha) & \hookrightarrow & F \\ \downarrow{\scriptstyle \iota} & & \downarrow{\scriptstyle \iota'} \\ k(\beta) & \hookrightarrow & F(\beta) \end{array}$$

is commutative, while the diagram

$$\begin{array}{ccc} k(\alpha) & \hookrightarrow & F \\ \downarrow{\scriptstyle \iota} & & \downarrow{\scriptstyle i} \\ k(\beta) & \hookrightarrow & F(\beta) \end{array}$$

is not commutative if $\beta \neq \alpha$. Indeed, ι sends α to β, while i sends α to α. The key point in the argument is the observation that if there exists one isomorphism between two finite-dimensional vector spaces V, W, then *every* injective linear map $V \to W$ must be an isomorphism. Finite dimensionality is necessary in order to draw this conclusion; cf. Exercise VI.6.5. ⌟

Example 4.10. If a complex root of an irreducible polynomial $p(x) \in \mathbb{Q}[x]$ may be expressed as a polynomial in i and $\sqrt[4]{2}$ with rational coefficients, then *all* roots

of $p(x)$ may be expressed likewise in terms of i and $\sqrt[4]{2}$. Indeed, we have checked (Example 4.6) that $\mathbb{Q}(i, \sqrt[4]{2})$ is a splitting field over $\mathbb{Q}$; hence it is a normal extension of $\mathbb{Q}$. ⌟

4.2. Separable polynomials. Our intuition may lead us to think that if a polynomial factors as a product of linear factors and we are not 'purposely' repeating one of the factors (as in $(x-1)^2(x-2)$), then these will be *distinct*. For example, surely irreducible polynomials necessarily split as products of distinct factors in an algebraic closure, right? Wrong.

Example 4.11. Let p be a prime, and consider the field $\mathbb{F}_p(t)$ of rational functions over $\mathbb{F}_p$. Then the polynomial

$$x^p - t \in \mathbb{F}_p(t)[x]$$

is irreducible: by Eisenstein's criterion it is irreducible in $\mathbb{F}_p[t][x]$ (since (t) is prime in $\mathbb{F}_p[t]$), hence in $\mathbb{F}_p(t)[x]$ by Proposition V.4.16. Let u be a root of this polynomial in an extension L of $\mathbb{F}_p(t)$ (for example L could be the algebraic closure of $\mathbb{F}_p(t)$, or more modestly a splitting field for the polynomial). Then (Exercise 4.8)

$$x^p - t = (x-u)^p$$

in $L[x]$; that is, u has multiplicity p as a root of $f(x)$.

In other words, the minimal polynomial over $\mathbb{F}_p(t)$ of $u \in L$ *vanishes p times at u*, and there is nothing to do about this: no smaller power of $(x-u)$ than $(x-u)^p = x^p - t$ has coefficients in $\mathbb{F}_p(t)$ (a smaller power would give a nontrivial factor of $x^p - t$, and $x^p - t$ is irreducible). ⌟

We have always found this example difficult to visualize, because of intuition developed in characteristic 0, and as we will see, no such pathology can occur in characteristic 0.

Definition 4.12. Let k be a field. A polynomial $f(x) \in k[x]$ is *separable* if it has *no* multiple factors over its splitting field; $f(x)$ is *inseparable* if it *has* multiple factors over its splitting field. ⌟

Thus, the polynomial $x^p - t$ in Example 4.11 is inseparable. We suppose the terminology reflects the fact that we cannot 'separate' its roots: they come clumped together like quarks in a proton, and we cannot take them apart.

Of course Definition 4.12 does not depend on the chosen splitting field, since these are unique up to isomorphism (Lemma 4.2). In fact, to detect separability, we can use *any* field in which the polynomial splits as a product of linear factors (by Exercise 4.1). The first, somewhat surprising, observation about separability is that we can in fact detect it without leaving the field of coefficients of $f(x)$. This fact uses a notion borrowed from *calculus:* for a polynomial

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,$$

we denote by $f'(x)$ the 'derivative'

$$f'(x) = a_1 + 2a_2 x + \cdots + n a_n x^{n-1}.$$

Of course this is a purely formal operation; no limiting process is at work over arbitrary fields. Still, all expected properties of derivatives hold, as the reader should check: for example, $(fg)' = f'g + fg'$ as usual.

Lemma 4.13. *Let k be a field, and let $f(x) \in k[x]$. Then $f(x)$ is separable if and only if $f(x)$ and $f'(x)$ are relatively prime.*

By definition, $f(x)$ and $f'(x)$ are relatively prime precisely when the greatest common divisor of $f(x)$ and $f'(x)$ is 1. Note that if $k \subseteq F$, the gcd of $f(x)$ and $f'(x)$ is the same whether it is considered in $k[x]$ or in $F[x]$: for example because it can be computed by applying the Euclidean algorithm (§V.2.4), and this proceeds in exactly the same way whether it is performed in $k[x]$ or in $F[x]$.

Proof. First assume that $f(x)$ is *not* separable. Then $f(x)$ has a multiple root in a splitting field F; that is,

$$f(x) = (x - \alpha)^m g(x)$$

for some $\alpha \in F$, $g(x) \in F[x]$, and $m \geq 2$. Therefore

$$f'(x) = m(x - \alpha)^{m-1} g(x) + (x - \alpha)^m g'(x),$$

and it follows that $(x - \alpha)$ is a common factor of $f(x)$ and $f'(x)$ in $F[x]$. Therefore $\gcd(f(x), f'(x)) \neq 1$ in $F[x]$; hence $\gcd(f(x), f'(x)) \neq 1$ in $k[x]$; that is, $f(x)$, $f'(x)$ are not relatively prime.

Conversely, assume $\gcd(f(x), f'(x)) \neq 1$, so $f(x)$ and $f'(x)$ have a common irreducible factor $(x - \alpha)$ in the algebraic closure $\overline{k}$ of k. Write $f(x) = (x - \alpha)h(x)$; we have

$$f'(x) = h(x) + (x - \alpha)h'(x),$$

and it follows that $x - \alpha$ divides $h(x)$ since it divides $f'(x)$. But then

$$(x - \alpha)^2 \mid f(x);$$

hence $f(x)$ is inseparable. □

For example, the polynomial $x^p - t$ of Example 4.11 may be seen to be inseparable without invoking splitting fields: the derivative of $x^p - t$ equals $px^{p-1} = 0$ in characteristic p, and $\gcd(x^p - t, 0) = x^p - t \neq 1$.

This example captures one of the key features of inseparability:

Lemma 4.14. *Let k be a field, and let $f(x) \in k[x]$ be an inseparable* irreducible *polynomial. Then $f'(x) = 0$.*

Proof. Since $f(x)$ is inseparable, $f(x)$ and $f'(x)$ have a common irreducible factor $q(x)$ by Lemma 4.13; but as $f(x)$ is itself irreducible, $q(x)$ must be an associate of $f(x)$, and in particular its degree is larger than the degree of $f'(x)$ if $f'(x) \neq 0$. As $q(x) \mid f'(x)$, the only option is that $f'(x) = 0$. □

This already tells us that irreducible polynomials are necessarily *separable* in characteristic 0: in characteristic 0, the derivative of a nonconstant polynomial

clearly cannot vanish. In fact, Lemma 4.14 gives us a precise picture of what inseparable, irreducible polynomials must look like. If

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

is irreducible and inseparable, then the characteristic of the field must be a positive prime p, and by Lemma 4.14 we must have

$$f'(x) = \sum_{i=0}^{n} i a_i x^{i-1} = 0;$$

that is, $ia_i = 0$ for all i. Now $ia_i = 0$ is automatic if i is a multiple of p, and it implies $a_i = 0$ for all the indices i which are *not* multiples of p. Therefore, the only nonvanishing coefficients in $f(x)$ must be those corresponding to indices which *are* multiples of p:

$$f(x) = a_0 + a_p x^p + a_{2p} x^{2p} + \ldots;$$

therefore, $f(x)$ must in fact be a polynomial *in* x^p. (The reader will refine this picture further by working out Exercise 4.13 and following.)

This leads us to a precise result linking separability and a flavor of fields that we have not yet run into.

Definition 4.15. Let k be a field of characteristic $p > 0$. The *Frobenius* homomorphism is the map $k \to k$ defined by $x \mapsto x^p$. ⌟

The Frobenius homomorphism may not look like a homomorphism of rings, but it is (Exercise 4.8). It must be injective, as is every nontrivial ring homomorphism from a field; but it is not necessarily surjective.

Definition 4.16. A field k is *perfect* if $\operatorname{char} k = 0$ or if $\operatorname{char} k > 0$ and the Frobenius homomorphism is surjective. ⌟

Proposition 4.17. *Let k be a field. Then k is perfect if and only if all irreducible polynomials in $k[x]$ are separable.*

Proof. We will prove that irreducible polynomials over a perfect field are separable, leaving the other implication to the reader (Exercise 4.12).

We have already noted that irreducible polynomials are separable over fields of characteristic zero. In positive characteristic p, we have observed that an inseparable irreducible polynomial must be of the form

$$f(x) = \sum_{i=0}^{m} a_i \cdot (x^p)^i.$$

Since Frobenius is surjective, there exist b_i such that $b_i^p = a_i$. Thus

$$f(x) = \sum_{i=0}^{m} b_i^p (x^p)^i = \sum_{i=0}^{m} (b_i x^i)^p = \left(\sum_{i=0}^{m} b_i x^i \right)^p = g(x)^p,$$

where $g(x) = \sum b_i x^i$ and keeping in mind that the Frobenius map is a homomorphism. But this contradicts the irreducibility of $f(x)$, so there is no such polynomial. □

Corollary 4.18. *Finite fields are perfect. Therefore, over* finite *fields, irreducible polynomials are separable.*

Proof. The Frobenius map is injective (because it is a homomorphism of fields), so it is surjective over finite fields, by the pigeon-hole principle. Therefore finite fields are perfect, and the second part of the statement follows from Proposition 4.17. □

The reader now sees why we have chosen the somewhat unusual field $\mathbb{F}_p(t)$ in Example 4.11: we need a field of positive characteristic, but no example of irreducible inseparable polynomial can be found over $\mathbb{F}_p$, by Corollary 4.18; to concoct an example, we need to enlarge the field so that it is infinite and to throw in an element (that is, t) for which there is no p-th root, that is, which is not in the image of Frobenius.

4.3. Separable extensions and embeddings in algebraic closures. The terminology examined in the previous section extends to the language of field extensions. If $k \subseteq F$ is an extension and $\alpha \in F$, we say that α is *separable* over k if the minimal polynomial of α over k is separable; α is *inseparable* otherwise.

Definition 4.19. An algebraic field extension $k \subseteq F$ is *separable* if every $\alpha \in F$ is separable over k. ⌟

With this terminology, Proposition 4.17 may be restated in the following more impressive form:

Proposition 4.20. *A field k is perfect if and only if every algebraic extension of k is separable.*

In particular, algebraic extensions of $\mathbb{Q}$ (or any field of characteristic zero) and of every finite field are necessarily separable.

The separability condition is exceedingly convenient, and we will essentially adopt it henceforth: all extensions we will seriously consider will be separable. One convenient feature of separable extensions is the following alternative description of the separability condition.

We have seen (Lemma 2.8) that every algebraic extension $k \subseteq F$ may be embedded in an algebraic closure $k \subseteq \overline{k}$, and we pointed out that this can in general be done in many different ways. If finite, the number of different homomorphisms $F \to \overline{k}$ extending the identity on k is denoted by

$$[F : k]_s.$$

Definition 4.21. This is the *separable degree* of F over k. ⌟

It is clear that this number is independent of the chosen algebraic closure $k \subseteq \overline{k}$. What does it have to do with separability?

Lemma 4.22. *Let $k \subseteq k(\alpha)$ be a simple algebraic extension. Then $[k(\alpha) : k]_s$ equals the number of distinct roots in $\overline{k}$ of the minimal polynomial of α. In particular, $[k(\alpha) : k]_s \leq [k(\alpha) : k]$, with equality if and only if α is separable over k.*

Proof. The proof is essentially (and not by coincidence) a rehash of the proof of Corollary 1.7. Associate with each $\iota : k(\alpha) \to \overline{k}$ extending id_k the image $\iota(\alpha)$, which must be a root of the minimal polynomial of α. This correspondence is injective, since $\iota(\alpha)$ determines ι (as ι extends the identity on k). To see it is surjective, let $\beta \in \overline{k}$ be any other root, and consider the extension $k(\beta) \subseteq \overline{k}$; by Proposition 1.5 there is an isomorphism $k(\alpha) \to k(\beta)$ sending α to β, and composing with the embedding $k(\beta) \subseteq \overline{k}$ defines the corresponding ι, as needed. □

Thus, the 'separable degree' does detect separability, for simple extensions. Further, it is *multiplicative* over successive extensions:

Lemma 4.23. *Let $k \subseteq E \subseteq F$ be algebraic extensions. Then $[F : k]_s$ is finite if and only if both $[F : E]_s$, $[E : k]_s$ are finite, and in this case*

$$[F : k]_s = [F : E]_s [E : k]_s.$$

Proof. Different embeddings of E into $\overline{k}$ extend to different embeddings of F into $\overline{k}$, by Lemma 2.8; and embeddings of F into $\overline{E} = \overline{k}$ extending the identity on E extend *a fortiori* the identity on k. Therefore, if any of $[F : E]_s$, $[E : k]_s$ is infinite, then so is $[F : k]_s$.

For the converse implication, it suffices to prove the stated degree formula. But every embedding $F \subseteq \overline{k}$ extending the identity on k may be obtained in two steps: first extend the identity to an embedding $E \subseteq \overline{k}$, which can be done in $[E : k]_s$ ways; and then extend the chosen embedding $E \subseteq \overline{k} = \overline{E}$ to an embedding $F \subseteq \overline{E} = \overline{k}$, which can be done in $[F : E]_s$ ways. There are precisely $[F : E]_s [E : k]_s$ ways to do this, as stated. □

Lemmas 4.22 and 4.23 allow us to recast separability of finite extensions entirely in the light of counting embeddings in an algebraic closure:

Proposition 4.24. *Let $k \subseteq F$ be a finite extension. Then $[F : k]_s \leq [F : k]$, and the following are equivalent:*

(i) *$F = k(\alpha_1, \ldots, \alpha_r)$, where each α_i is separable over k;*

(ii) *$k \subseteq F$ is separable;*

(iii) *$[F : k]_s = [F : k]$.*

Proof. Since F is finite over k, it is finitely generated. Let $F = k(\alpha_1, \ldots, \alpha_r)$. Then using Lemma 4.23, Lemma 4.22, and Proposition 1.10,

$$\begin{aligned}[F : k]_s &= [k(\alpha_1, \ldots, \alpha_{r-1})(\alpha_r) : k(\alpha_1, \ldots, \alpha_{r-1})]_s \cdots [k(\alpha_1) : k]_s \\ &\leq [k(\alpha_1, \ldots, \alpha_{r-1})(\alpha_r) : k(\alpha_1, \ldots, \alpha_{r-1})] \cdots [k(\alpha_1) : k] \\ &= [F : k].\end{aligned}$$

This proves the stated inequality.

(i) $\Longrightarrow$ (iii): If each α_i is separable over k, then it is separable over the field $k(\alpha_1, \ldots, \alpha_{i-1})$ (Exercise 4.15), so the inequality is an equality by Lemma 4.22.

(iii) $\Longrightarrow$ (ii): Assume $[F:k]_s = [F:k]$, and let $\alpha \in F$. We have $k \subseteq k(\alpha) \subseteq F$; hence by Lemma 4.23

$$[F:k(\alpha)]_s[k(\alpha):k]_s = [F:k]_s = [F:k] = [F:k(\alpha)][k(\alpha):k].$$

Since both separable degrees are less than or equal to their plain counterparts, the equality implies

$$[k(\alpha):k]_s = [k(\alpha):k],$$

proving that α is separable, by Lemma 4.22. Therefore the extension $k \subseteq F$ is separable according to Definition 4.19.

(ii) $\Longrightarrow$ (i) is immediate from Definition 4.19, since finite extensions are finitely generated. □

For example, if α is separable over k, then every $\beta \in k(\alpha)$ is separable. This would seem rather mysterious from the definition alone: why should the fact that the minimal polynomial of α has distinct roots in $\overline{k}$ imply the same for the minimal polynomial of β? It does, by virtue of Proposition 4.24.

Remark 4.25. While we have only proved $[F:k]_s \leq [F:k]$, one can in fact prove that $[F:k]_s$ *divides* $[F:k]$: in fact, $[F:k]_s$ is the degree of a certain intermediate field F_{sep} over k. The diligent reader will prove this in Exercise 4.18.

Exercises

4.1. ▷ Let k be a field, $f(x) \in k[x]$, and let F be the splitting field for $f(x)$ over k. Let $k \subseteq K$ be an extension such that $f(x)$ splits as a product of linear factors over K. Prove that there is a homomorphism $F \to K$ extending the identity on k. [§4.2]

4.2. Describe the splitting field of $x^6 + x^3 + 1$ over $\mathbb{Q}$. Do the same for $x^4 + 4$.

4.3. ▷ Find the order of the automorphism group of the splitting field of $x^4 + 2$ over $\mathbb{Q}$ (cf. Example 4.6). [§4.1]

4.4. Prove that the field $\mathbb{Q}(\sqrt[4]{2})$ is not the splitting field of any polynomial over $\mathbb{Q}$.

4.5. ▷ Let F be a splitting field for a polynomial $f(x) \in k[x]$, and let $g(x) \in k[x]$ be a factor of $f(x)$. Prove that F contains a unique copy of the splitting field of $g(x)$. [§5.1]

4.6. Let $k \subseteq F_1$, $k \subseteq F_2$ be two finite extensions, viewed as embedded in the algebraic closure $\overline{k}$ of k. Assume that F_1 and F_2 are splitting fields of polynomials in $k[x]$. Prove that the intersection $F_1 \cap F_2$ and the composite F_1F_2 (the smallest subfield of $\overline{k}$ containing both F_1 and F_2) are both also splitting fields over k. (Theorem 4.8 is likely going to be helpful.)

4.7. ▷ Let $k \subseteq F = k(\alpha)$ be a simple algebraic extension. Prove that F is normal over k if and only if for every algebraic extension $F \subseteq K$ and every $\sigma \in \text{Aut}_k(K)$, $\sigma(F) = F$. [§6.1]

4.8. ▷ Let p be a prime, and let k be a field of characteristic p. For $a, b \in K$, prove that[16] $(a+b)^p = a^p + b^p$. [§4.2, §5.1, §5.2]

4.9. Using the notion of 'derivative' given in §4.2, prove that $(fg)' = f'g + fg'$ for all polynomials f, g.

4.10. Let $k \subseteq F$ be a finite extension in characteristic $p > 0$. Assume that p does not divide $[F : k]$. Prove that $k \subseteq F$ is separable.

4.11. ▷ Let p be a prime integer. Prove that the Frobenius homomorphism on $\mathbb{F}_p$ is the identity. (Hint: Fermat.) [§5.1]

4.12. ▷ Let k be a field, and assume that k is not perfect. Prove that there are inseparable irreducible polynomials in $k[x]$. (If $\operatorname{char} k = p$ and $u \in k$, how many roots does $x^p - u$ have in $\overline{k}$?) [§4.2]

4.13. ▷ Let k be a field of positive characteristic p, and let $f(x)$ be an irreducible polynomial. Prove that there exist an integer d and a *separable* irreducible polynomial $f_{\text{sep}}(x)$ such that

$$f(x) = f_{\text{sep}}(x^{p^d}).$$

The number p^d is called the *inseparable degree* of $f(x)$. If $f(x)$ is the minimal polynomial of an algebraic element α, the inseparable degree of α is defined to be the inseparable degree of $f(x)$. Prove that α is inseparable if and only if its inseparable degree is $\geq p$.

The picture to keep in mind is as follows: the roots of the minimal polynomial $f(x)$ of α are distributed into $\deg f_{\text{sep}}$ 'clumps', each collecting a number of coincident roots equal to the inseparable degree of α. We say that α is 'purely inseparable' if there is only one clump, that is, if all roots of $f(x)$ coincide (see Exercise 4.14). [§4.2, 4.14, 4.18]

4.14. ¬ Let $k \subseteq F$ be an algebraic extension, in positive characteristic p. An element $\alpha \in F$ is *purely inseparable* over k if $\alpha^{p^d} \in k$ for some[17] $d \geq 0$. The extension is defined to be purely inseparable if every $\alpha \in F$ is purely inseparable over k.

Prove that α is purely inseparable if and only if $[k(\alpha) : k]_s = 1$, if and only if its degree equals its *inseparability* degree (Exercise 4.13). [4.13, 4.17]

4.15. ▷ Let $k \subseteq F$ be an algebraic extension, and let $\alpha \in F$ be separable over k. For every intermediate field $k \subseteq E \subseteq F$, prove that α is separable over E. [§4.3]

4.16. ¬ Let $k \subseteq E \subseteq F$ be algebraic field extensions, and assume that $k \subseteq E$ is separable. Prove that if $\alpha \in F$ is separable over E, then $k \subseteq E(\alpha)$ is a separable extension. (Reduce to the case of finite extensions.)

Deduce that the set of elements of F which are separable over k form an intermediate field F_{sep}, such that every element $\alpha \in F$, $\alpha \notin F_{\text{sep}}$ is *inseparable* over k.

For $F = \overline{k}$, $\overline{k}_{\text{sep}}$ is called the *separable closure* of k. [4.17, 4.18]

---

[16] This is sometimes referred to as the *freshman's dream,* for painful reasons that are likely all too familiar to the reader.

[17] Note the slightly annoying clash of notation: elements of k are not inseparable, yet they are *purely* inseparable according to this definition.

4.17. ¬ Let $k \subseteq F$ be an algebraic extension, in positive characteristic. With notation as in Exercises 4.14 and 4.16, prove that the extension $F_{\text{sep}} \subseteq F$ is purely inseparable. Prove that an extension $k \subseteq F$ is purely inseparable if and only if $F_{\text{sep}} = k$. [4.18]

4.18. ▷ Let $k \subseteq F$ be a finite extension, in positive characteristic. Define the *inseparable degree* $[F : k]_i$ to be the quotient $[F : k]/[F : k]_s$.

- Prove that $[k(\alpha) : k]_i$ equals the inseparable degree of α, as defined in Exercise 4.13.
- Prove that the inseparable degree is multiplicative: if $k \subseteq E \subseteq F$ are finite extensions, then $[F : k]_i = [F : E]_i[E : k]_i$.
- Prove that a finite extension is purely inseparable if and only if its inseparable degree equals its degree.
- With notation as in Exercise 4.16, prove that $[F : k]_s = [F_{\text{sep}} : k]$ and $[F : k]_i = [F : F_{\text{sep}}]$. (Use Exercise 4.17.)

In particular, $[F : k]_s$ divides $[F : k]$; the inseparable degree $[F : k]_i$ is an integer. [§4.3]

4.19. ¬ Let $k \subseteq F$ be a finite separable extension, and let $\iota_1, \ldots, \iota_d$ be the distinct embeddings of F in $\overline{k}$ extending id_k. For $\alpha \in F$, prove that the *norm* $N_{k\subseteq F}(\alpha)$ (cf. Exercise 1.12) equals $\prod_{i=1}^{d} \iota_i(\alpha)$ and its *trace* $\mathrm{tr}_{k\subseteq F}(\alpha)$ (Exercise 1.13) equals $\sum_{i=1}^{d} \iota_i(\alpha)$. (Hint: Exercises 1.14 and 1.15.) [4.21, 4.22, 6.15]

4.20. ¬ Let $k \subseteq F$ be a finite separable extension, and let $\alpha \in F$. Prove that for all $\sigma \in \mathrm{Aut}_k(F)$, $N_{k\subseteq F}(\alpha/\sigma(\alpha)) = 1$ and $\mathrm{tr}_{k\subseteq F}(\alpha - \sigma(\alpha)) = 0$. [6.16, 6.19]

4.21. ¬ Let $k \subseteq E \subseteq F$ be finite separable extensions, and let $\alpha \in F$. Prove that

$$N_{k\subseteq F}(\alpha) = N_{k\subseteq E}(N_{E\subseteq F}(\alpha)) \quad \text{and} \quad \mathrm{tr}_{k\subseteq F}(\alpha) = \mathrm{tr}_{k\subseteq E}(\mathrm{tr}_{E\subseteq F}(\alpha)).$$

(Hint: Use Exercise 4.19: if $d = [E : k]$ and $e = [F : E]$, the de embeddings of F into $\overline{k}$ lifting id_k must divide into d groups of e each, according to their restriction to E.)

This 'transitivity' of norm and trace extends the result of Exercise 1.15 to separable extensions. The separability restriction is actually unnecessary; cf. Exercise 4.22. [4.22]

4.22. Generalize Exercises 4.19—4.21 to all finite extensions $k \subseteq F$. (For the norm, raise to power $[F : k]_i$; for the trace, multiply by $[F : k]_i$.)

---

5. Field extensions, III

The material in §4 provides us with the main tools needed to tackle several key examples. In this section we study finite and cyclotomic fields, and we return to the question of when a *finite* extension is in fact *simple* (cf. Example 1.19).

5.1. Finite fields. Let F be a finite field, and let p be its characteristic. We know (§1.1) that F may be viewed as an extension

$$\mathbb{F}_p \subseteq F$$

of $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$; let $d = [F : \mathbb{F}_p]$. Since F has dimension d as a vector space over $\mathbb{F}_p$, it is isomorphic to $\mathbb{F}_p^d$ as a vector space, and in particular $|F| = p^d$ is a power of p.

The general question we pose is whether there exist fields of cardinality equal to every power of a prime, and we aim to classifying all fields of a given cardinality. The conclusion we will reach is as neat as it can be: *for every prime power q there exists exactly one field F with q elements, up to isomorphism.*

Also recall (Remark III.1.16) that, by a theorem of Wedderburn, every finite *division ring* is in fact a finite *field.* Thus, relaxing the hypothesis of commutativity in this subsection would not lead to a different classification.

Theorem 5.1. *Let $q = p^d$ be a power of a prime integer p. Then the splitting field of the polynomial $x^q - x$ over $\mathbb{F}_p$ is a field with precisely q elements.*

Conversely, let F be a field with exactly q elements; then F is a splitting field for $x^q - x$ over $\mathbb{F}_p$. The polynomial $x^q - x$ is separable over $\mathbb{F}_p$.

Proof. Let F be the splitting field of $x^q - x$ over $\mathbb{F}_p$. Let E be the set of roots of $f(x) = x^q - x$ in F. Since $f'(x) = qx^{q-1} - 1 = -1$ (as $q = 0$ in characteristic p), we have $(f(x), f'(x)) = 1$; hence (Lemma 4.13) $f(x)$ is separable, and E consists of precisely q elements. We claim that E is a field, and it follows that $E = F$. Indeed, F is generated by the roots of $f(x)$; hence the smallest subfield of F containing E is F itself.

To see that E is a field, let $a, b \in E$. Then $a^q = a$ and $b^q = b$; it follows that

$$(a - b)^q = a^q + (-1)^q b^q = a - b$$

(using Exercise 4.8; note that $(-1)^q = -1$ if p is odd and $(-1)^q = +1 = -1$ if $p = 2$). If $b \neq 0$,

$$(ab^{-1})^q = a^q (b^q)^{-1} = ab^{-1}.$$

Thus E is closed under subtraction and division by a nonzero element, proving that E is a field and concluding the proof of the first statement.

To prove the second statement, let F be a field with exactly q elements. The nonzero elements of F form a group under multiplication, consisting of $q - 1$ elements; therefore, the (multiplicative) order of every nonzero $a \in F$ divides $q - 1$ (Example II.8.15). Therefore,

$$a \neq 0 \implies a^{q-1} = 1 \implies a^q - a = 0;$$

of course $0^q - 0 = 0$. In other words, the polynomial $x^q - x$ has q roots in F (that is, all elements of F!); it follows that F is a splitting field for $x^q - x$, as stated. □

Corollary 5.2. *For every prime power q there exists one and only one finite field of order q, up to isomorphism.*

Proof. This follows immediately from Theorem 5.1 and the uniqueness of splitting fields (Lemma 4.2). □

Since there is exactly one isomorphism class of fields of order q for any given prime power q, we can devise a notation for a field of order q; it makes sense to adopt[18] $\mathbb{F}_q$. This is called the *Galois field* of order q.

Example 5.3. Let p be a prime integer. Then we claim that the polynomial x^4+1 is *reducible*[19] over $\mathbb{F}_p$ (and therefore over every finite field).

Since $x^4+1=(x+1)^4$ in $\mathbb{F}_2[x]$, the statement holds for $p=2$. Thus, we may assume that p is an odd prime. Then we claim that x^4+1 divides $x^{p^2}-x$. Indeed, the square of every odd number is congruent to 1 mod 8 (Exercise II.2.11); hence $8 \mid (p^2-1)$; hence x^8-1 divides $x^{p^2-1}-1$ (Exercise V.2.13); hence

$$(x^4+1) \mid (x^8-1) \mid (x^{p^2-1}-1) \mid (x^{p^2}-x).$$

It follows that x^4+1 factors completely in the splitting field of $x^{p^2}-x$, that is, in $\mathbb{F}_{p^2}$. If α is a root of x^4+1 in $\mathbb{F}_{p^2}$, we have the extensions

$$\mathbb{F}_p \subseteq \mathbb{F}_p(\alpha) \subseteq \mathbb{F}_{p^2};$$

therefore (Corollary 1.11) $[\mathbb{F}_p(\alpha):\mathbb{F}_p]$ divides $[\mathbb{F}_{p^2}:\mathbb{F}_p]=2$. That is, α has degree 1 or 2 over $\mathbb{F}_p$. But then its minimal polynomial is a factor of degree 1 or 2 of (x^4+1), showing that the latter is reducible. ⌟

Theorem 5.1 has many interesting consequences, and we sample a few in the rest of this subsection.

Corollary 5.4. *Let p be a prime, and let $d \le e$ be positive integers. Then there is an extension $\mathbb{F}_{p^d} \subseteq \mathbb{F}_{p^e}$ if and only if $d \mid e$. Further, if $d \mid e$, then there is exactly one such extension, in the sense that $\mathbb{F}_{p^e}$ contains a unique copy of $\mathbb{F}_{p^d}$.*

All extensions $\mathbb{F}_{p^d} \subseteq \mathbb{F}_{p^e}$ are simple.

Proof. If there is an extension as stated, then $\mathbb{F}_p \subseteq \mathbb{F}_{p^d} \subseteq \mathbb{F}_{p^e}$; hence $[\mathbb{F}_{p^d}:\mathbb{F}_p]$ divides $[\mathbb{F}_{p^e}:\mathbb{F}_p]$ by Corollary 1.11. This says precisely that $d \mid e$.

Conversely, assume that $d \mid e$. As

$$p^e-1=(p^d-1)((p^d)^{\frac{e}{d}-1}+\cdots+1),$$

we see that p^d-1 divides p^e-1, and consequently $x^{p^d-1}-1$ divides $x^{p^e-1}-1$ (Exercise V.2.13). Therefore

$$(x^{p^d}-x) \mid (x^{p^e}-x).$$

By Theorem 5.1, $\mathbb{F}_{p^e}$ is a splitting field for the second polynomial. It follows that it contains a unique copy of the splitting field for the first polynomial (Exercise 4.5), that is, of $\mathbb{F}_{p^d}$.

For the last statement, recall that the multiplicative group of nonzero elements of a finite field is necessarily *cyclic* (Theorem IV.6.10). If $\alpha \in \mathbb{F}_{p^e}$ is a generator of this group, then α will generate $\mathbb{F}_{p^e}$ over any subfield; if $d \mid e$, this says $\mathbb{F}_{p^e}=\mathbb{F}_{p^d}(\alpha)$, so $\mathbb{F}_{p^d} \subseteq \mathbb{F}_{p^e}$ is simple. □

[18] Keep in mind that $\mathbb{F}_q$ is *not* the ring $\mathbb{Z}/q\mathbb{Z}$ unless q is prime: if q is composite, then $\mathbb{Z}/q\mathbb{Z}$ is not an integral domain.

[19] Of course x^4+1 is *irreducible* over $\mathbb{Z}$. This example shows that Proposition V.5.15 cannot be turned into an 'if and only if' statement.

These results can be translated into rather precise information on the structure of the polynomial ring over a finite field. For example,

Corollary 5.5. *Let F be a finite field. Then for all integers $n \geq 1$ there exist irreducible polynomials of degree n in $F[x]$.*

Proof. We know $F = \mathbb{F}_{p^d}$ for some prime p and some $d \geq 1$. By Corollary 5.4 there is an extension $\mathbb{F}_{p^d} \subseteq \mathbb{F}_{p^{dn}}$, generated by an element α. Then $[\mathbb{F}_{p^{dn}} : \mathbb{F}_{p^d}] = n$, and it follows that the minimal polynomial of α over $F = \mathbb{F}_{p^d}$ is an irreducible polynomial of degree n in $F[x]$. □

In fact, our analysis of extensions of finite fields tells us about explicit factorizations, leading to an inductive algorithm for the computation of irreducible polynomials in $\mathbb{F}_q[x]$:

Corollary 5.6. *Let $F = \mathbb{F}_q$ be a finite field, and let n be a positive integer. Then the factorization of $x^{q^n} - x$ in $F[x]$ consists of all irreducible monic polynomials of degree d, as d ranges over the positive divisors of n. In particular, all these polynomials factor completely in $\mathbb{F}_{q^n}$.*

Proof. By Theorem 5.1, $\mathbb{F}_{q^n}$ is the splitting field of $x^{q^n} - x$ over $\mathbb{F}_p$, and hence over $\mathbb{F}_q = F$.

If $f(x)$ is a monic irreducible polynomial of degree d, then $F[x]/(f(x)) = F(\alpha)$ is an extension of degree d of F, that is, an isomorphic copy of $\mathbb{F}_{q^d}$. By Corollary 5.4, if $d \mid n$, then there is an embedding of $\mathbb{F}_{q^d}$ in $\mathbb{F}_{q^n}$. But then α must be a root of $x^{q^n} - x$, and hence $x^{q^n} - x$ is a multiple of $f(x)$, as this is the minimal polynomial of α. This proves that every irreducible polynomial of degree $d \mid n$ is a factor of $x^{q^n} - x$.

Conversely, if $f(x)$ is an irreducible factor of $x^{q^n} - x$, then $\mathbb{F}_{q^n}$ contains a root α of $f(x)$; we have the extensions $F = \mathbb{F}_q \subseteq \mathbb{F}_q(\alpha) \subseteq \mathbb{F}_{q^n}$, and $\mathbb{F}_q(\alpha) \cong \mathbb{F}_{q^d}$ for $d = \deg \alpha$. It follows that $d \mid n$, again by Corollary 5.4. □

The picture we are trying to convey is the following: the q^n roots of $x^{q^n} - x$ clump into disjoint subsets, with each subset collecting the roots of each and every irreducible polynomial of degree $d \mid n$ in $F[x]$.

Example 5.7. Let's contemplate the case $q = 2$: $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$.

• $n = 1$: the polynomial $x^2 - x$ factors as the product of x and $(x - 1)$ (which we could write as $(x + 1)$ just as well, since we are working over $\mathbb{F}_2$). These are all the irreducible polynomials of degree 1 over $\mathbb{F}_2$.

• $n = 2$: the polynomial $x^4 - x$ must factor as the product of all irreducible polynomials of degree 1 and 2; in fact

$$x^4 - x = x(x - 1)(x^2 + x + 1),$$

and the conclusion is that there is exactly one irreducible polynomial of degree 2 over $\mathbb{F}_2$, namely $x^2 + x + 1$.

• $n = 3$: the quotient of $x^8 - x$ by $x(x-1)$ is a polynomial of degree 6, which must therefore be the product of the two irreducible polynomials of degree 3 over $\mathbb{F}_2$. It takes a moment to find them:

$$x^3 + x^2 + 1, \quad x^3 + x + 1.$$

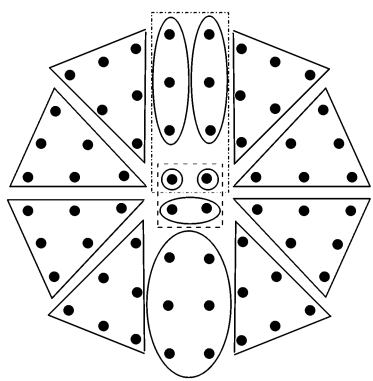
It also follows that $\mathbb{F}_8$ may be realized in two ways as a quotient of $\mathbb{F}_2[x]$ modulo an irreducible polynomial:

$$\frac{\mathbb{F}_2[x]}{(x^3 + x^2 + 1)} \cong \frac{\mathbb{F}_2[x]}{(x^3 + x + 1)}.$$

We know that these two fields must be isomorphic because of Theorem 5.1; it is instructive to find an explicit isomorphism between them (Exercise 5.3).

• $n = 4$ and 5: these cases are left to the enjoyment of the reader (Exercise 5.4).

• $n = 6$: the factorization of $x^{64} - x$ must include x, $x - 1$, the one irreducible polynomial of degree 2, and the two irreducible polynomials of degree 3 found above, and that leaves room for 9 polynomials of degree 6, which must be all and only the irreducible polynomials of degree 6 over $\mathbb{F}_2$. So the 64 elements of $\mathbb{F}_{64}$ cluster as follows:



The two dotted rectangles delimit the (unique) copies of $\mathbb{F}_4$ and $\mathbb{F}_8$ contained in $\mathbb{F}_{64}$; these intersect in the (unique) copy of $\mathbb{F}_2$.

Again, $\mathbb{F}_{64}$ may be realized by quotienting $\mathbb{F}_2[x]$ by the ideal generated by any of the irreducible polynomials of degree 6; this gives 9 'different' realizations of this field. ⌟

The picture drawn above for $\mathbb{F}_{64}$ means next to nothing, but it may help us focus on one last element of information we are going to extract regarding finite fields. Note that the effect of any automorphism of $\mathbb{F}_{64}$ must be to scramble elements in each sector represented in the picture, without mixing elements in different sectors and without interchanging sectors. Indeed, every automorphism of an extension sends roots of an irreducible polynomial to roots of the *same* polynomial.

Since extensions of finite fields are simple extensions, our previous work allows us to be much more precise. Restricting our attention to the extensions $\mathbb{F}_p \subseteq \mathbb{F}_{p^d}$, for a prime p, we know these can be realized as simple extensions by an element with minimal polynomial of degree d. This polynomial is necessarily separable

($\mathbb{F}_p$ is perfect), so Corollary 1.7 immediately gives us the size of the automorphism group:

$$|\operatorname{Aut}_{\mathbb{F}_p}(\mathbb{F}_{p^d})| = d.$$

But what is this group?

Proposition 5.8. $\operatorname{Aut}_{\mathbb{F}_p}(\mathbb{F}_{p^d})$ *is cyclic, generated by the Frobenius isomorphism.*

Proof. Let φ be the Frobenius homomorphism $\mathbb{F}_{p^d} \to \mathbb{F}_{p^d}$: $\varphi(x) = x^p$. The Frobenius homomorphism is an isomorphism on a finite field (Corollary 4.18) and restricts to the identity on $\mathbb{F}_p$ (Exercise 4.11), so $\varphi \in \operatorname{Aut}_{\mathbb{F}_p}(\mathbb{F}_{p^d})$. Since we know *a priori* the size of this group, all we have to show is that the order of φ is d.

Let $e = |\varphi|$. Then $\varphi^e = \mathrm{id}$; hence $x^{p^e} = x$ for all $x \in \mathbb{F}_{p^d}$. In other words, the (nonzero) polynomial $x^{p^e} - x$ has p^d roots in $\mathbb{F}_{p^d}$; this implies $p^d \leq p^e$ (Lemma V.5.1). Equivalently, $d \leq e$, yielding

$$|\operatorname{Aut}_{\mathbb{F}_p}(\mathbb{F}_{p^d})| \leq |\varphi|.$$

But then these two numbers must be equal, since the order of an element always divides the order of the group (Example II.8.15). □

Two last comments on finite fields are in order:

—For n large enough, the stupendously large (but finite) field $\mathbb{F}_{p^{n!}}$ works as an approximation of the algebraic closure $\overline{\mathbb{F}_p}$: indeed, this field contains roots of all polynomials of degree $\leq n$ in $\mathbb{F}_p[x]$, by Corollary 5.6.

This observation can be turned into the explicit construction of an algebraic closure of $\mathbb{F}_p$: by Corollary 5.4 there are extensions

$$\mathbb{F}_p \subseteq \mathbb{F}_{p^2} \subseteq \mathbb{F}_{p^6} \subseteq \mathbb{F}_{p^{4!}} \subseteq \cdots$$

and the union of this chain of fields[20] gives a copy of $\overline{\mathbb{F}_p}$.

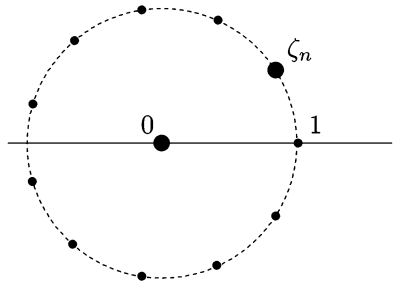
—Finite fields form a category, which (by Corollary 5.4) should strongly remind the reader of the 'toy' category encountered in Exercise I.5.6. There is an important difference, however: in that category every object has exactly one automorphism, while we have just checked that finite fields may have many automorphisms. Challenge: The reader has encountered a very similar situation in a different (but related) context. Where?

5.2. Cyclotomic polynomials and fields. In the last several sections we have often encountered roots of 1 in $\mathbb{C}$; the extensions they generate are very important.

Let n be a positive integer. *We will denote by* ζ_n *the complex number* $e^{2\pi i/n}$; thus, the n roots of the polynomial $x^n - 1$ in $\mathbb{C}$ are the n distinct powers of ζ_n; they form a cyclic subgroup of order n of the multiplicative group of $\mathbb{C}$, which we will denote μ_n.

Pictorially, the roots of 1 are placed at the vertices of a regular n-gon centered at 0, with one vertex at 1.

[20] A more formal construction requires the notion of *direct limit;* cf. §VIII.1.4.



A *primitive* n-th root of 1 is a generator of μ_n. Thus, ζ_n is primitive; from our work on cyclic groups, we know (Corollary II.2.5) that ζ_n^m is a primitive n-th root of 1 if and only if m is relatively prime to n. In particular, there are $\phi(n)$ primitive roots of 1, where ϕ is Euler's totient function (cf. Exercise II.6.14). Thus,

$$\Phi_n(x) := \prod_{\zeta \text{ primitive root of } 1} (x - \zeta) = \prod_{1 \le m \le n, (m,n)=1} (x - \zeta_n^m)$$

is a polynomial of degree $\phi(n)$.

Definition 5.9. The polynomial $\Phi_n(x)$ is called the *n-th cyclotomic polynomial.* ⌟

It is clear that $\Phi_n(x)$ is a monic polynomial of degree $\phi(n)$. It is perhaps a little less immediate that $\Phi_n(x)$ has *rational* (in fact, *integer*) coefficients and that it is *irreducible* over $\mathbb{Q}$. We will prove these facts in short order.

Example 5.10. If $n = p$ is prime, then every nonidentity element of $\mu_p \cong C_p$ is a generator: every p-th root of 1 is primitive except 1 itself. Therefore

$$\Phi_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + \cdots + 1$$

is the particular case encountered in Example V.5.19, where we proved that $\Phi_p(x)$ is indeed irreducible. ⌟

What if n is not prime?

Lemma 5.11. *For all positive integers n,*

$$x^n - 1 = \prod_{1 \le d | n} \Phi_d(x).$$

Proof. If $n = de$, then every d-th root ζ of 1 is an n-th root of 1, because $\zeta^n = \zeta^{de} = (\zeta^d)^e = 1$. In particular, every *primitive* d-th root ζ of 1 is an n-th root of 1.

On the other hand, every $\zeta \in \mu_n$ generates a subgroup H of μ_n, and $H = \mu_d$ for d equal to the order of ζ, a divisor of n (Proposition II.6.11). Thus, every $\zeta \in \mu_n$ is a primitive d-th root of 1 for some $d \mid n$.

Thus the set of n-th roots of 1 equals the union of the sets of primitive d-th roots of 1, as d ranges over all positive divisors of n. The statement follows immediately:

$$x^n - 1 = \prod_{\zeta \in \mu_n} (x - \zeta) = \prod_{1 \le d | n} \left(\prod_{\zeta \text{ primitive } d\text{-th root of } 1} (x - \zeta) \right) = \prod_{1 \le d | n} \Phi_d(x),$$

as claimed. □

This argument brings us back to simple group-theoretic considerations. The reader can check that Lemma 5.11 implies immediately the result of Exercise II.6.14, by comparing degrees of both sides of the stated identity.

Lemma 5.11 yields an inductive computation of cyclotomic polynomials; the fact that $\Phi_n(x) \in \mathbb{Z}[x]$ follows from this fact. Explicitly,

Corollary 5.12. *The cyclotomic polynomials $\Phi_n(x)$ have integer coefficients.*

Proof. Use induction on n. Note that $\Phi_1(x) = x - 1$, and assume we have shown that all $\Phi_m(x)$ have integer coefficients for $m < n$. In particular, $f(x) := \prod_{1 \le d|n, d<n} \Phi_d(x)$ is a monic polynomial with integer coefficients. Since $f(x)$ is monic, we can divide it into $x^n - 1$ with remainder, within $\mathbb{Z}[x]$: $\exists q(x), r(x) \in \mathbb{Z}[x]$ such that

$$x^n - 1 = f(x)q(x) + r(x),$$

with $r(x) = 0$ or $\deg r(x) < \deg f(x)$. On the other hand, by Lemma 5.11,

$$x^n - 1 = f(x)\Phi_n(x)$$

in $\mathbb{C}[x]$. Therefore

$$f(x)(\Phi(x) - q(x)) = r(x)$$

in $\mathbb{C}[x]$. But this forces $r(x) = 0$ (otherwise we would have $\deg r(x) > \deg f(x)$). Therefore $\Phi(x) = q(x) \in \mathbb{Z}[x]$. □

Example 5.13. The reader can spend some quality time computing explicitly the cyclotomic polynomials $\Phi_n(x)$ for several nonprime numbers n, working inductively and capitalizing on the fact that we know explicitly $\Phi_p(x)$ for prime p.

For example, $x^4 - 1 = \Phi_1(x)\Phi_2(x)\Phi_4(x)$; therefore

$$\Phi_4(x) = \frac{x^4 - 1}{x^2 - 1} = x^2 + 1.$$

Since $x^6 - 1 = \Phi_1(x)\Phi_2(x)\Phi_3(x)\Phi_6(x)$,

$$\Phi_6(x) = \frac{x^6 - 1}{(x^2 - 1)(x^2 + x + 1)} = x^2 - x + 1.$$

Since $x^{12} - 1 = \Phi_1(x)\Phi_2(x)\Phi_3(x)\Phi_4(x)\Phi_6(x)\Phi_{12}(x)$,

$$\Phi_{12}(x) = \frac{x^{12} - 1}{(x^6 - 1)(x^2 + 1)} = x^4 - x^2 + 1,$$

and so on *ad libitum.* ⌟

The optimistic reader may now start guessing that the coefficients of $\Phi_n(x)$ are always 0 or ± 1 (so would we, if we didn't know better). Apparently the first counterexample is found for $n = 105$ (Exercise 5.9), and the coefficients are known to get as large as one pleases for $n \gg 0$.

The irreducibility of $\Phi_n(x)$ is a bit trickier, in particular because *separability* sneaks into the standard argument (which is why we had to wait until now to present it).

Proposition 5.14. *For all positive n, $\Phi_n(x) \in \mathbb{Z}[x]$ is irreducible over $\mathbb{Q}$.*

(Since $\Phi_n(x)$ is monic, irreducibility in $\mathbb{Q}[x]$ is equivalent to irreducibility in $\mathbb{Z}[x]$; cf. Corollary V.4.17.)

Proof. Arguing by contradiction, assume $\Phi_n(x)$ is *reducible.* Then its roots ζ_n^m, with $(m,n) = 1$, are divided among the factors; we can choose a root ζ_n^m of one irreducible monic factor $f(x)$, such that another root ζ_n^{mp} (for some prime p not dividing n) is *not* a root of $f(x)$. Write

$$\Phi_n(x) = f(x)g(x);$$

since $\Phi_n(x) \in \mathbb{Z}[x]$ and $\Phi_n(x)$, $f(x)$ are monic, then $f(x)$ and $g(x)$ have integer coefficients (cf. Exercise V.4.23). By our choice, $f(x)$ is the minimal polynomial of ζ_n^m over $\mathbb{Q}$, and $g(\zeta_n^{mp}) = 0$.

It follows that ζ_n^m is a root of $g(x^p)$, and hence $f(x) \mid g(x^p)$. Therefore we can write

$$g(x^p) = f(x)h(x)$$

with $h(x) \in \mathbb{Z}[x]$. Reading the last equation modulo p, we get (again using Exercise 4.8, and denoting cosets by underlining)

$$\underline{g}(x)^p = \underline{f}(x)\underline{h}(x) \quad \text{in } \mathbb{F}_p[x];$$

in particular, $\underline{f}(x)$ and $\underline{g}(x)$ must have a nontrivial common factor $\underline{\ell}(x)$ in $\mathbb{F}_p[x]$. But then

$$\underline{\ell}(x)^2 \mid \underline{f}(x)\underline{g}(x) :$$

the reduction of $\Phi_n(x)$ modulo p must have a multiple factor.

This implies that $x^n - 1 \in \mathbb{F}_p[x]$ has a multiple factor; that is, it is *in*separable. However, its derivative $nx^{n-1} \in \mathbb{F}_p[x]$ is nonzero (because p does not divide n by assumption), and Lemma 4.13 implies that $x^n - 1$ *is* separable in $\mathbb{F}_p[x]$.

This contradiction shows that our assumption that $\Phi_n(x)$ is reducible must be nonsense, proving the statement. □

Definition 5.15. The splitting field $\mathbb{Q}(\zeta_n)$ for the polynomial $x^n - 1$ over $\mathbb{Q}$ is the n-th *cyclotomic field.* ⌟

By Proposition 5.14, $\mathbb{Q}(\zeta_n)$ is an extension of $\mathbb{Q}$ of degree $\phi(n)$; $\Phi_n(x)$ is the minimal polynomial of ζ_n.

As usual, we now preoccupy ourselves with the *automorphism group* of this extension, and as in previous examples the situation is very neat.

Proposition 5.16. $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\zeta_n))$ *is isomorphic to the group of units in* $\mathbb{Z}/n\mathbb{Z}$.

Proof. We know that $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\zeta_n))$ has cardinality $\phi(n)$ (Corollary 1.7; the roots are distinct since $\Phi(n)$ is separable), so all we need to do is exhibit an injective homomorphism

$$j : (\mathbb{Z}/n\mathbb{Z})^* \to \mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\zeta_n)).$$

For $[m]_n \in (\mathbb{Z}/n\mathbb{Z})^*$, that is, for m an integer relatively prime to n, let $j([m]_n)$ be the unique automorphism $\mathbb{Q}(\zeta_n) \to \mathbb{Q}(\zeta_n)$ sending ζ_n to ζ_n^m (cf. Proposition 1.5);

this is evidently independent of the representative m. Then j is clearly injective, and

$$j([m_1]_n) \circ j([m_2]_n)(\zeta_n) = j([m_1]_n)(\zeta_n^{m_2}) = \zeta_n^{m_2 m_1}$$

agrees with the action of $j([m_1]_n[m_2]_n)$, showing that j is a homomorphism and concluding the proof. □

Example 5.17. The reader should consider Examples 4.4 again: by what we have just proved, the automorphism group of the splitting field of $x^8 - 1$ is isomorphic to the group of units in $\mathbb{Z}/8\mathbb{Z}$; this group is immediately seen to be isomorphic to $(\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/2\mathbb{Z})$, confirming the claim made in Example 4.4. ⌟

The constructibility of the regular n-gon by straightedge and compass is of course equivalent to the constructibility of the complex number ζ_n. Since we now know that $[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \phi(n)$, Corollary 3.6 tells us that *the regular n-gon is constructible by straightedge and compass only if $\phi(n)$ is a power of 2*. This condition can be chased a little further (Exercise 5.19). In any case, we shall return to this theme one more time, after we absorb a little Galois theory.

5.3. Separability and simple extensions. All the examples of field extensions we have encountered so far were *simple* extensions, although sometimes this may not be completely apparent at first (cf. Example 1.19). There is a good reason for this: most of the examples we have seen were *separable* extensions, and we can now prove that a finite separable extension is necessarily simple.

Here is a nice criterion for simplicity (which does not involve any separability condition):

Proposition 5.18. *An algebraic extension $k \subseteq F$ is simple if and only if the number of distinct intermediate fields $k \subseteq E \subseteq F$ is finite.*

Proof. Assume $F = k(\alpha)$ is simple and algebraic, and let $q_k(x)$ be the minimal polynomial of α over k. Embed F in an algebraic closure $\overline{k}$ of k. If E is an intermediate field, then $F = E(\alpha)$ is also a simple, algebraic extension; denote by $q_E(x)$ the minimal polynomial of α over E. Since $q_k(x) \in E[x]$ for all E and $q_k(\alpha) = 0$, we know that each $q_E(x)$ is a factor of $q_k(x)$.

We claim that E is in fact determined by $q_E(x)$. Since $q_k(x)$ has finitely many factors in $\overline{k}$, this proves that there are only finitely many intermediate fields, that is, the 'only if' part of the statement.

To verify our claim, we will show that E is generated (over k) by the coefficients of $q_E(x)$. To see this, let E' be the subfield of E generated by k and the coefficients of $q_E(x)$. Then $q_E(x) \in E'[x]$, and since $q_E(x)$ is irreducible over E, it must be irreducible over E'. Note that $E'(\alpha) = F = E(\alpha)$. We have $E' \subseteq E \subseteq F$; therefore

$$\deg(q_E(x)) = [F : E'] = [F : E][E : E'] = \deg(q_E(x))[E : E'],$$

from which $[E : E'] = 1$; that is, $E = E'$ as needed. This concludes the proof of the 'only if' part of the statement.

To prove the 'if' part, assume that there are only finitely many intermediate fields $k \subseteq E \subseteq F$. The extension $k \subseteq F$ must be finitely generated (otherwise

we could construct an infinite sequence of subextensions $k \subseteq k(\alpha_1) \subseteq k(\alpha_1, \alpha_2) \subseteq \cdots \subseteq F$, and we would have infinitely many intermediate fields), hence finite since it is algebraic. If k is a finite field, then so is F, and the extension is automatically simple by Corollary 5.4; thus we may assume that k is infinite, and we have to show that every finitely generated algebraic extension $F = k(\alpha_1, \ldots, \alpha_r)$ is simple.

Arguing inductively, we may assume without loss of generality that $F = k(\alpha, \beta)$. For every $c \in k$, we have the intermediate field

$$F \subseteq k(c\alpha + \beta) \subseteq k(\alpha, \beta).$$

But there are only finitely many intermediate fields, while k is infinite, so for some $c \neq c'$ in k we must have

$$k(c'\alpha + \beta) = k(c\alpha + \beta).$$

It follows that

$$\alpha = \frac{(c'\alpha + \beta) - (c\alpha + \beta)}{c' - c} \in k(c\alpha + \beta) \quad \text{and} \quad \beta = (c\alpha + \beta) - c\alpha \in k(c\alpha + \beta)$$

and therefore $k(\alpha, \beta) \subseteq k(c\alpha + \beta)$. The other inclusion holds trivially, so $k(\alpha, \beta) = k(c\alpha + \beta)$ is simple. □

Here is the connection with separability:

Proposition 5.19. *Every finite separable extension is simple.*

Proof. Arguing inductively as in the proof of Proposition 5.18, we may assume $F = k(\alpha, \beta)$, with α and β separable (and in particular algebraic) over k, and we may assume k is an infinite field.

Consider the set I of embeddings $\iota : F \hookrightarrow \overline{k}$ of F in an algebraic closure of k, extending the identity of k. If $\iota \neq \iota'$ in I and x is an indeterminate, then the polynomials

$$\iota(\alpha)x + \iota(\beta), \quad \iota'(\alpha)x + \iota'(\beta)$$

are different: otherwise $\iota'(\alpha) = \iota(\alpha)$ and $\iota'(\beta) = \iota(\beta)$, so ι, ι' would act in the same way on the whole of $k(\alpha, \beta)$, forcing $\iota = \iota'$.

Therefore, the polynomial

$$f(x) = \prod_{\iota \neq \iota'} ((\iota(\alpha)x + \iota(\beta)) - (\iota'(\alpha)x + \iota'(\beta))) \in \overline{k}[x]$$

is not identically 0. Since k is infinite, it follows that $\exists c \in k$ such that $f(c) \neq 0$; that is, distinct $\iota \in I$ map

$$\gamma = c\alpha + \beta$$

to distinct elements

$$\iota(\gamma) = \iota(\alpha)c + \iota(\beta)$$

(each ι extends id_k, so $\iota(c) = \iota$). Since the cardinality of I is $[F : k]_s$ (Definition 4.21) and each $\iota(\gamma)$ is a root of the minimal polynomial of γ over k, we have

$$[F : k]_s \leq [k(\gamma) : k] \leq [F : k].$$

By assumption the extension is separable, so $[F : k]_s = [F : k]$ by Proposition 4.24, implying $[k(\gamma) : k] = [F : k]$ and finally

$$F = k(\gamma),$$

concluding the proof. □

Proposition 5.19 is a good illustration of why separability is a technically desirable condition. To stress the point even further, note that we are now in a position to extend the convenient inequality found in Corollary 1.7 to all finite *separable* extensions:

Corollary 5.20. *Let $k \subseteq F$ be a finite, separable extension. Then*

$$|\operatorname{Aut}_k(F)| \leq [F : k],$$

with equality if and only if $k \subseteq F$ is a normal extension.

Proof. Since $k \subseteq F$ is finite and separable, Proposition 5.19 implies it is simple: $F = k(\alpha)$ for some $\alpha \in F$. The inequality follows immediately from Corollary 1.7, and equality holds if and only if the minimal polynomial $f(x)$ of α factors into distinct linear factors in F. In this case F is the splitting field of $f(x)$, so it is normal over k by Theorem 4.8. Conversely, if F is normal over k, then $f(x)$ splits completely in F and has distinct roots since α is separable over k; therefore $|\operatorname{Aut}_k(F)| = [F : k]$, again by Corollary 1.7. □

Example 5.21. By Proposition 5.19, if we want to construct a *nonsimple* finite extension, we have to use inseparable elements; to produce an example, we jazz up Example 4.11 a little. Consider the field of rational functions $F = \mathbb{F}_p(u, v)$ in *two* variables over $\mathbb{F}_p$ and the subfield $k = \mathbb{F}_p(s, t)$, where $s = u^p$ and $t = v^p$. Then $k \subseteq F$ is an algebraic extension; the minimal polynomial of u over k is $x^p - s$; the minimal polynomial of v over $k(u)$ is $y^p - t$; it follows that

$$[F : k] = p^2.$$

As c ranges in k, we obtain intermediate fields

$$k \subseteq k(cu + v) \subseteq F.$$

If any two choices c, c' led to the same intermediate field, then we would deduce that $k(u, v) = k(cu+v)$ by arguing as in the proof of Proposition 5.18. In particular, we would have

$$[k(cu + v) : k] = p^2.$$

But this is not the case, since $(cu + v)^p = c^p s + t \in k(s, t) = k$; hence the minimal polynomial of $cu + v$ over k has degree at most p. Since $k = \mathbb{F}_p(s, t)$ is infinite, there are infinitely many intermediate fields, and it follows (by Proposition 5.18) that F is *not* a simple extension of k. ⌟

Exercises

5.1. Let $p > 0$ be a prime integer. Prove that the additive group of a finite field with p^d elements is isomorphic to $(\mathbb{Z}/p\mathbb{Z})^d$.

5.2. Prove that every element of a finite field F is a sum of two squares in F. (Hint: Keep in mind that the multiplicative group of F is cyclic.)

5.3. ▷ Find an explicit isomorphism

$$\frac{\mathbb{F}_2[x]}{(x^3+x^2+1)} \xrightarrow{\sim} \frac{\mathbb{F}_2[x]}{(x^3+x+1)}.$$

[§5.1]

5.4. ▷ Find all irreducible polynomials of degree 4 over $\mathbb{F}_2$, and count those of degree 5. [§5.1]

5.5. Find the number of irreducible polynomials of degree 12 over $\mathbb{F}_9$.

5.6. Write out explicitly the action of the cyclic group C_4 on $\mathbb{F}_{16}$, in terms of any realization of this field as a quotient of $\mathbb{F}_2[x]$.

5.7. Let p be a prime integer. View the Frobenius automorphism $\varphi : \mathbb{F}_{p^d} \to \mathbb{F}_{p^d}$ as a linear transformation of the $\mathbb{F}$-vector space $\mathbb{F}_{p^d}$. Find the rational canonical form of φ. (Adapt the proof of Proposition 5.8 to show that the minimal polynomial of φ is $x^d - 1$.)

5.8. For a prime p, find the factorization of $\Phi_p(x)$ over $\mathbb{F}_p$.

5.9. ▷ Find all cyclotomic polynomials $\Phi_n(x)$ for $1 \leq n \leq 15$. Compute $\Phi_{105}(x)$. [§5.2]

5.10. Find the cyclotomic polynomials $\Phi_{2^m}(x)$ for all $m \geq 0$.

5.11. Prove that if n is odd, then $\Phi_{2n}(x) = \Phi_n(-x)$. (Hint: Draw the primitive 14-th roots of 1 side-by-side to the primitive 7-th roots of 1; then go back to Exercise II.2.15 to justify the fact you observe.)

5.12. ¬ Let a, n be positive integers, with $a > 1$. Prove that if $\Phi_n(a)$ divides $a - 1$, then $n = 1$. (Remember that $\Phi_n(a)$ is a product of complex numbers $a - \zeta$, where ζ is a primitive n-th root of 1. What does this tell you about the size of $\Phi_n(a)$?) [5.14]

5.13. ¬ Let a, d, n be positive integers, with $d < n$ and $a > 1$. Assume that $a^d - 1$ divides $a^n - 1$. Prove that $\Phi_n(a)$ divides the quotient $(a^n - 1)/(a^d - 1)$. (Hint: Exercise V.2.13.) [5.14]

5.14. ▷ Let R be a finite division ring.

- Prove that the center of R (Exercise III.2.9) is isomorphic to $\mathbb{F}_q$, for q a prime power. Prove that $|R| = q^n$ for some n.

- For every $r \in R$, prove that the centralizer of x in the multiplicative group $(R^*, \cdot)$ has order $q^d - 1$ for some $d \leq n$.
- Prove that there are integers $d_1, \ldots, d_r < n$ such that

$$(*) \qquad q^n - 1 = q - 1 + \sum_{i=1}^{r} \frac{q^n - 1}{q^{d_i} - 1}.$$

  (Hint: Class equation.)
- Deduce that $\Phi_n(q)$ divides $q - 1$ and hence $n = 1$. (Use Exercises 5.12 and 5.13.)
- Conclude that R equals its center, showing that R is commutative.

Thus, every finite division ring is a field: this is Wedderburn's little theorem. The argument given here is due to Ernst Witt. [§III.1.2]

5.15. ▷ Let a, p, n be integers, with p, n positive and p prime, $p \nmid n$.

- Show that $x^n - 1$ has no multiple roots modulo p.
- Show that if p divides $\Phi_n(a)$, then $a^n \equiv 1$ modulo p. (In particular, $p \nmid a$, so $[a]_p \in (\mathbb{Z}/p\mathbb{Z})^*$.)
- Show that if p divides $\Phi_n(a)$, then $a^d \not\equiv 1$ modulo p for every $d < n$.
- Deduce that $p \mid \Phi_n(a)$ if and only if the order of $[a]_p$ in $(\mathbb{Z}/p\mathbb{Z})^*$ is n.
- Compute $\Phi_{15}(9)$, and show it is divisible by 31. Then look back at the first part of Exercise II.4.12.

[§II.4.3, 5.16, 5.17]

5.16. If q is a prime that divides $a^n - 1$ and does not divide $a^d - 1$ for any $d < n$, then q is said to be a *primitive* prime divisor of $a^n - 1$. By Exercise 5.15, if $q \nmid n$ and $q \mid \Phi_n(a)$, then q is a primitive prime divisor of $a^n - 1$. The *Birkhoff-Vandiver* theorem asserts that $a^n - 1$ has primitive prime divisors for all but a very short list of exceptions: $n = 1, a = 2$; $n = 2$, $a + 1$ a power of 2; and $n = 6$, $a = 2$.

Assuming this statement, we can give another proof of Wedderburn's theorem (published in 2003 by Nicolas Lichiardopol). Let R be a finite division ring.

- Prove that there is a prime p such that $|R| = p^n$ for some integer n.
- If $n = 1$, then $R \cong \mathbb{F}_p$. If $n = 2$ or if $p = 2$ and $n = 6$, then R is commutative: the reader has proved this in Exercises III.2.11 and IV.2.17.
- Therefore, by Birkhoff-Vandiver we may assume that $p^n - 1$ has a primitive prime divisor p'.
- Prove that there is an element a of order p' in the multiplicative group $(R^*, \cdot)$ of R.
- Prove that R is the only sub-division ring of R containing a.
- Prove that the centralizer of a in R (Exercise III.2.10) is R itself; deduce that a is in the center of R (Exercise III.2.9).
- Prove that the center of R is R: this shows that R is commutative.

Thus, R is a finite *field.*

Note that the Birkhoff-Vandiver theorem gives another rapid way to conclude Witt's proof: a primitive prime divisor of $q^n - 1$ would divide $q - 1$ by (*), showing that $n = 1$.

5.17. ¬ Let a, p, n be integers, with p, n positive and p prime, $p \nmid n$. Assume that p divides $\Phi_n(a)$. Prove that $p \equiv 1 \bmod n$. (Use Exercise 5.15.) [5.18]

5.18. ▷ Let $n > 0$ be any integer. Prove that there are infinitely many prime integers $\equiv 1 \bmod n$. (Use Exercise 5.17 together with Exercise V.2.25.)

The result of this exercise is a particular case of Dirichlet's theorem on primes in arithmetic progressions; see §7.6. [§7.6]

5.19. ▷ Prove that the regular n-gon can be constructed by straightedge and compass only if $n = 2^m p_1 \dots p_r$, where $m \geq 0$ and the factors p_i are distinct Fermat primes. (Hint: Use Exercise V.6.8.) [§5.2, §7.2]

5.20. Recall from Exercise 1.10 that every field extension of degree $\leq n$ over a field k is a sub-k-algebra of the ring of matrices $\mathcal{M}_n(k)$. Prove that if k is finite or has characteristic 0, then every extension of k contained in $\mathcal{M}_n(k)$ has degree $\leq n$.

5.21. Prove that if $k \subseteq F$ is the splitting field of a separable polynomial, then it is the splitting field of an *irreducible* separable polynomial.

5.22. Let k be an infinite field. If $F = k(\alpha_1, \dots, \alpha_r)$, with each α_i separable over k, prove that there exist $c_1, \dots, c_r \in k$ such that $F = k(c_1\alpha_1 + \dots + c_r\alpha_r)$.

5.23. ▷ Let k be a field, and let $n > 0$ be an integer. Assume that there are no irreducible polynomials of degree n in $k[x]$. Prove that there are no separable extensions of k of degree n. [§7.1]

---

6. A little Galois theory

Galois theory is a beautiful interplay of field theory and group theory, originally motivated by the problem of determining 'symmetries' among roots of a polynomial. Galois was interested in concrete relations which must necessarily hold among the roots of a polynomial, the most trivial example of which being the fact that if α, β are roots of $x^2 + px + q$, then $\alpha + \beta = -p$ and $\alpha\beta = q$. Interchanging the roots α, β has no effect on the quantities $\alpha + \beta$, $\alpha\beta$. More generally, for a higher degree polynomial there may be several quantities which are invariant under certain permutations of the roots. These quantities and the corresponding groups of permutations may be viewed as invariants determined by the polynomial, yielding a sophisticated tool to study the polynomial.

6.1. The Galois correspondence and Galois extensions. In the language of field theory, subsets of the roots of a polynomial $f(x)$ determine intermediate fields of the splitting field of $f(x)$, and groups of permutations of the roots give automorphisms of these intermediate fields. With this in mind, it is not surprising that splitting fields should come to the fore; the fact that one can characterize such

fields in terms of groups of automorphisms (Theorem 6.9 below) will be key to the whole discussion.

Before wading into these waters, we should formalize the precise relation between groups of automorphisms of an extension and intermediate fields.

Definition 6.1. Let $k \subseteq F$ be a field extension, and let $G \subseteq \mathrm{Aut}_k(F)$ be a group of automorphisms of the extension. The *fixed field* of G is the intermediate field

$$F^G := \{\alpha \in F \,|\, \forall g \in G,\, g\alpha = \alpha\}.$$ ⌟

The fact that F^G is indeed a subfield of F containing k is immediate. The notion of fixed field allows us to set up a correspondence

$$\{\text{intermediate fields } E\colon\ k \subseteq E \subseteq F\} \rightleftarrows \{\text{subgroups of } \mathrm{Aut}_k(F)\}\ ,$$

sending the intermediate field E to the subgroup $\mathrm{Aut}_E(F)$ of $\mathrm{Aut}_k(F)$ and the subgroup $G \subseteq \mathrm{Aut}_k(F)$ to the fixed field F^G.

Definition 6.2. This is known as the *Galois correspondence.* ⌟

Lemma 6.3. *The Galois correspondence is inclusion-reversing. Further, for all subgroups G of $\mathrm{Aut}_k(F)$ and all intermediate fields $k \subseteq E \subseteq F$:*

- $E \subseteq F^{\mathrm{Aut}_E(F)}$;
- $G \subseteq \mathrm{Aut}_{F^G}(F)$.

Further still, denote by E_1E_2 the smallest subfield of F containing two intermediate fields E_1, E_2, and denote by $\langle G_1, G_2\rangle$ the smallest subgroup of $\mathrm{Aut}_k(F)$ containing two subgroups G_1, G_2. Then

- $\mathrm{Aut}_{E_1E_2}(F) = \mathrm{Aut}_{E_1}(F) \cap \mathrm{Aut}_{E_2}(F)$;
- $F^{\langle G_1,G_2\rangle} = F^{G_1} \cap F^{G_2}$.

Proof. Exercise 6.1. □

Of course the reader should start wondering whether there are situations guaranteeing that the inclusions appearing in Lemma 6.3 are equalities, making the Galois correspondence a bijection. This is precisely where we are heading. The first observation is that this is not always the case:

Example 6.4. Consider the extension $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2})$. Since

$$[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$$

is prime, the only intermediate fields are $\mathbb{Q}$ and $\mathbb{Q}(\sqrt[3]{2})$ (by Corollary 1.11). Concerning $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt[3]{2}))$, since $\sqrt[3]{2} \in \mathbb{R}$, we have an extension $\mathbb{Q}(\sqrt[3]{2}) \subseteq \mathbb{R}$; since $\sqrt[3]{2}$ is the only cube root of 2 in $\mathbb{R}$, we see that the minimal polynomial $t^3 - 2$ of $\sqrt[3]{2}$ has a single root in $\mathbb{Q}(\sqrt[3]{2})$. By Corollary 1.7, $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt[3]{2}))$ consists of a single element: it is trivial.

Thus, in this example the Galois correspondence acts between a set with two elements and a singleton:

$$\{\mathbb{Q},\, \mathbb{Q}(\sqrt[3]{2})\} \rightleftarrows \{\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt[3]{2}))\} = \{e\}\ .$$

In particular, the function associating with each intermediate field the corresponding automorphism group is not injective in general. ⌟

This example shows that the inclusion $E \subseteq F^{\mathrm{Aut}_E(F)}$ in Lemma 6.3 may be proper: $\mathbb{Q}$ is properly contained in $\mathbb{Q}(\sqrt[3]{2})$, which is the fixed field of the (only) subgroup of $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt[3]{2}))$. Fortunately, the situation with the other inclusion is more constrained:

Proposition 6.5. *Let $k \subseteq F$ be a* finite *extension, and let G be a subgroup of $\mathrm{Aut}_k(F)$. Then $|G| = [F : F^G]$, and*

$$G = \mathrm{Aut}_{F^G}(F).$$

In particular, the Galois correspondence (from intermediate fields to automorphism groups) is surjective for a finite extension.

Proving this fact leads us to examining very carefully finite extensions of the type $F^G \subseteq F$. Surprisingly, these turn out to satisfy just about every property we have encountered so far:

Lemma 6.6. *Let $k \subseteq F$ be a finite extension, and let G be a subgroup of $\mathrm{Aut}_k(F)$. Then $F^G \subseteq F$ is a finite, simple, normal, separable extension.*

Remark 6.7. The key to this lemma is the following observation, which is worth highlighting.

If $\alpha \in F$ and $g \in G$, note that $g\alpha$ must be a root of the minimal polynomial of α over k; there are only finitely many roots, so the G-orbit of α consists of finitely many elements $\alpha = \alpha_1, \ldots, \alpha_n$. The group G acts on the orbit by permuting its elements; therefore, every element of G leaves the polynomial

$$q_\alpha(t) = (t - \alpha_1) \cdots (t - \alpha_n)$$

fixed. In other words, *the coefficients of this polynomial must be in the fixed field F^G*. Further, $q_\alpha(t)$ is separable since it has distinct roots. Finally, note that $\deg q_\alpha(t) \leq |G|$. ⌟

With this in mind, we are ready to prove the lemma.

Proof of Lemma 6.6. The extension $F^G \subseteq F$ is finite because $k \subseteq F$ is finite.

Let $\alpha \in F$; by the remark following the statement of the lemma, α is a root of a *separable* polynomial $q_\alpha(t)$ with coefficients in F^G. It follows that α is separable over F^G; hence the extension is separable according to Definition 4.12.

Since $F^G \subseteq F$ is finite and separable, it is simple by Proposition 5.19; let α be a generator. The polynomial $q_\alpha(t)$ splits in F, and F is generated over F^G by the roots of $q_\alpha(t)$ (indeed, $\alpha_1 = \alpha$ suffices to generate F over F^G); therefore, F is a splitting field for $q_\alpha(t)$ over F^G according to Definition 4.1. Therefore $F^G \subseteq F$ is normal by Theorem 4.8, and we are done. □

Proposition 6.5 is a consequence of Lemma 6.6:

Proof of Proposition 6.5. By Lemma 6.3, G is a subgroup of $\mathrm{Aut}_{F^G}(F)$, and in particular

$$|G| \leq |\mathrm{Aut}_{F^G}(F)|.$$

In order to verify that $G = \mathrm{Aut}_{F^G}$, it suffices to prove the converse inequality. By Lemma 6.6, $F = F^G(\alpha)$ for some $\alpha \in F$; therefore $|\mathrm{Aut}_{F^G}(F)|$ equals the number of distinct roots in F of the minimal polynomial of α over F^G (by Corollary 1.7). With notation as in Remark 6.7, α is a root of $q_\alpha(x) \in F^G[x]$; hence the minimal polynomial of α is a factor of $q_\alpha(x)$. Since by construction the number of roots of $q_\alpha(x)$ is $\leq |G|$, we obtain

$$|\mathrm{Aut}_{F^G}(F)| \leq |G|,$$

as needed.

To verify that $[F : F^G] = |G|$, just note that $[F : F^G] = |\mathrm{Aut}_{F^G}(F)|$ by Corollary 5.20, since $F^G \subseteq F$ is normal and separable by Lemma 6.6. □

Remark 6.8. The hypothesis that the extension is finite in Proposition 6.5 is necessary: the Galois correspondence is not necessarily surjective in the infinite case. Not all is lost, though: one can give a suitable topology to the group of automorphisms and limit the Galois correspondence to *closed* subgroups of the automorphism group, recovering results such as Proposition 6.5. The reader is welcome to explore this further—we will not be able to take this more general point of view here. ⌟

Theorem 6.9. *Let $k \subseteq F$ be a finite field extension. Then the following are equivalent:*

(1) F is the splitting field of a separable polynomial $f(t) \in k[t]$ over k;

(2) $k \subseteq F$ is normal and separable;

(3) $|\mathrm{Aut}_k(F)| = [F : k]$;

(4) $k = F^{\mathrm{Aut}_k(F)}$ is the fixed field of $\mathrm{Aut}_k(F)$;

(5) the Galois correspondence for $k \subseteq F$ is a bijection;

(6) $k \subseteq F$ is separable, and if $F \subseteq K$ is an algebraic extension and $\sigma \in \mathrm{Aut}_k(K)$, then $\sigma(F) = F$.

Proof. Most of the needed implications have been proven along the way.

(1) $\Longleftrightarrow$ (2) by Theorem 4.8; (2) $\Longleftrightarrow$ (3) by Corollary 5.20. (3) $\Longleftrightarrow$ (4) follows from Proposition 6.5, applied to the extension $F^{\mathrm{Aut}_k(F)} \subseteq F$: by Proposition 6.5, we have

$$[F : F^{\mathrm{Aut}_k(F)}] = |\mathrm{Aut}_k(F)|;$$

since $k \subseteq F^{\mathrm{Aut}_k(F)} \subseteq F$, it follows that $k = F^{\mathrm{Aut}_k(F)}$ if and only if $|\mathrm{Aut}_k(F)| = [F : k]$.

(2) $\Longleftrightarrow$ (6) holds by Exercise 4.7, since finite separable extensions are simple (Proposition 5.19).

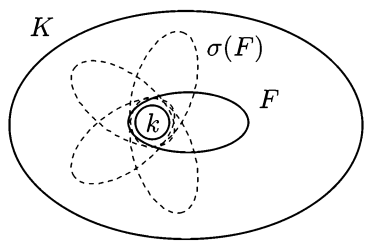
To prove (5) $\Longrightarrow$ (4), let $E = F^{\mathrm{Aut}_k(F)}$; by Proposition 6.5, $\mathrm{Aut}_E(F) = \mathrm{Aut}_k(F)$. Assuming that the Galois correspondence is bijective, it follows that $k = E = F^{\mathrm{Aut}_k(F)}$, giving (4).

Finally, we prove that (1) $\Longrightarrow$ (5). Since $k \subseteq F$ is a finite extension, we already know that the Galois correspondence from intermediate fields to subgroups of $\mathrm{Aut}_k(F)$ has a right inverse (Proposition 6.5), so it suffices to show it has a left-inverse. Therefore, it suffices to verify that every intermediate field E equals the fixed field of the corresponding subgroup $\mathrm{Aut}_E(F)$. Then let E be an intermediate field. By (1), F is the splitting field of a separable polynomial $f(t) \in k[t] \subseteq E[t]$; therefore, condition (1) holds *for the extension* $E \subseteq F$. As we have already proved that (1) $\Longleftrightarrow$ (4), this implies that $E = F^{\mathrm{Aut}_E(F)}$, and we are done. □

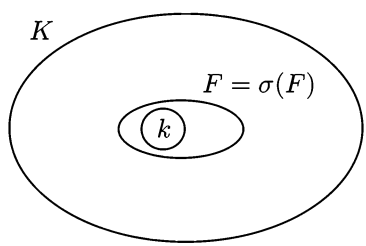
Definition 6.10. A finite extension $k \subseteq F$ is *Galois* if it satisfies any of the conditions listed in Theorem 6.9. ⌟

Warning: We will not study the Galois condition for *infinite* extensions (cf. Remark 6.8). Thus Galois extensions will implicitly be *finite* in what follows.

Condition (6) is technically useful and helps with visualizing the Galois condition. If a finite separable extension $k \subseteq F$ is *not* Galois, then F can be embedded in some larger extension $k \subseteq K$ (for example, in the algebraic closure $k \subseteq \overline{k}$) in many possible ways:



If $k \subseteq F$ *is* Galois, all these images must coincide:



Of course there are possibly still many ways[21] to embed F in K, but they all have the same image F.

The mysterious comments at the end of Example 1.4 are now hopefully clear. Galois extensions of a field k are well-defined subfields of $\overline{k}$, preserved by automorphisms of $\overline{k}$ over k. Contrast this situation with the non-Galois extension $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2})$, which admits three different embeddings in $\overline{\mathbb{Q}}$. For example, it does not make sense to talk about 'the' composite of $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt[3]{2})$, as a subfield of (for instance) $\mathbb{C}$: according to the chosen embedding of $\mathbb{Q}(\sqrt[3]{2})$, this may or may

[21] In fact, precisely $[F : k]_s = [F : k] = |\mathrm{Aut}_k(F)|$ if $K = \overline{k}$ and hence for all $K \supseteq F$.

not be a subfield of $\mathbb{R}$. But[22] it *does* make sense to talk about the composite of two Galois extensions $k \subseteq F_1$, $k \subseteq F_2$, since both F_1 and F_2 are well-defined as subfield of $\overline{k}$, so their composite F_1F_2 is independent of the choice of embeddings $F_1 \subseteq \overline{k}$, $F_2 \subseteq \overline{k}$.

Definition 6.11. If $k \subseteq F$ is a Galois extension, the corresponding automorphism group $\mathrm{Aut}_k(F)$ is called the *Galois group* of the extension. ⌟

To reiterate, the extension $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2})$ is *not* a Galois extension: as we have observed in Example 6.4, $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt[3]{2}))$ is trivial (hence, for example, condition (3) of Theorem 6.9 does not hold).

On the other hand, we have already run into several examples of Galois extensions: the examples of splitting fields we have seen in §4.1 are all Galois extensions. The 'Galois fields' of §5.1 are (surprise, surprise) Galois extensions of their prime subfields, by Theorem 5.1. The cyclotomic fields (§5.2) are Galois extensions of $\mathbb{Q}$.

A Galois extension $k \subseteq F$ is *cyclic, abelian,* etc., if the corresponding group of automorphisms $\mathrm{Aut}_k(F)$ is cyclic, abelian, etc. For example, $\mathbb{Q} \subseteq \mathbb{Q}(\zeta_8)$ is an abelian, but not cyclic, Galois extension (cf. Proposition 5.16 and Example 5.17).

6.2. The fundamental theorem of Galois theory, I. The fundamental theorem of Galois theory amounts to a more complete description of the Galois correspondence for Galois extensions.

At this stage, this is what we know:

- If $k \subseteq F$ is a (finite) Galois extension, then there is an inclusion-reversing bijection

  $$\{\text{intermediate fields } E\colon\ k \subseteq E \subseteq F\} \rightleftarrows \{\text{subgroups of } \mathrm{Aut}_k(F)\}\ :$$

  through this bijection, the intermediate field E corresponds to the subgroup $\mathrm{Aut}_E(F)$ of $\mathrm{Aut}_k(F)$, and the subgroup $G \subseteq \mathrm{Aut}_k(F)$ corresponds to the fixed field F^G.
- For every intermediate field E,

  $$[F : E] = |\mathrm{Aut}_E(F)|.$$

  The extension $E \subseteq F$ is then also a Galois extension (Exercise 6.3), and

  $$[E : k] = |\mathrm{Aut}_k(F) : \mathrm{Aut}_E(F)|\ :$$

  this follows from Lagrange's theorem (Corollary II.8.14) and its field theory counterpart (Proposition 1.10).
- The extension $k \subseteq E$ is not necessarily Galois. (Once more $\mathbb{Q}(\sqrt[3]{2})$ gives a counterexample, since it can be viewed as intermediate in the splitting field of $t^3 - 2$ over $\mathbb{Q}$.)

The fact that the Galois correspondence is a bijection may be upgraded as follows:

[22] 'Separability' does not play a role in these considerations; thus they best convey 'Galoisness' in, say, characteristic 0.

Theorem 6.12. *Let $k \subseteq F$ be a Galois extension. The Galois correspondence is an inclusion-reversing isomorphism of the lattice of intermediate subfields of $k \subseteq F$ with the lattice of subgroups of* $\mathrm{Aut}_k(F)$.

That is (with notation as in Lemma 6.3), if E_1, E_2 are intermediate fields and G_1, G_2 are the corresponding subgroups of $\mathrm{Aut}_k(F)$, *then $E_1 \cap E_2$ corresponds to $\langle G_1, G_2 \rangle$ and $E_1 E_2$ corresponds to $G_1 \cap G_2$.*

Proof. This follows immediately from Theorem 6.9 and Lemma 6.3, which gives

$$\mathrm{Aut}_{E_1 E_2}(F) = G_1 \cap G_2, \quad F^{\langle G_1, G_2 \rangle} = E_1 \cap E_2$$

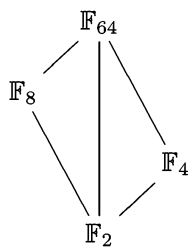
as needed. □

In this context, it is common to use the 'vertical' depiction of field extensions, with a parallel notation for subgroups:

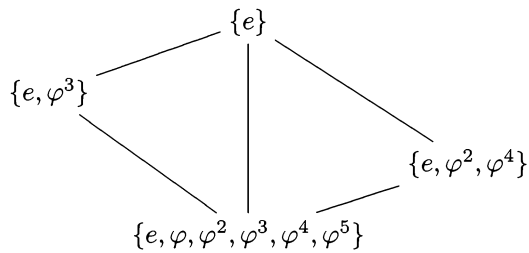
$$\begin{array}{c} F \\ | \\ E \\ | \\ k \end{array} \qquad \begin{array}{c} \{e\} \\ | \\ \mathrm{Aut}_E(F) \\ | \\ \mathrm{Aut}_k(F) \end{array}$$

The content of Theorem 6.12 is that the lattice of intermediate fields of a Galois extension $k \subseteq F$ and the lattice of subgroups of the corresponding Galois group are *identical.*

Example 6.13. For finite fields, this coincidence of lattices was essentially proven 'by hand' in Corollary 5.4 and Proposition 5.8. For example, the extension $\mathbb{F}_2 \subseteq \mathbb{F}_{64}$ is Galois, with cyclic Galois group C_6, generated by the Frobenius automorphism φ. The lattices are



,

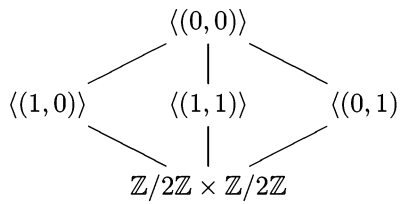


⌟

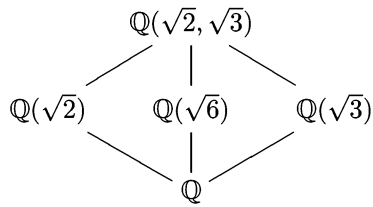
Theorem 6.12 is often used in the 'groups to fields' direction: finding the lattice of subgroups of a finite group is essentially a combinatorial problem, and through the Galois correspondence it determines the lattice of intermediate fields of a corresponding Galois extension.

Example 6.14. The extension $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3})$ studied in Example 1.19 is the splitting field of the polynomial $t^4 - 10t^2 + 1$, so it is Galois. We found that

its Galois group is $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$; the lattice of this group has no mysteries for us:

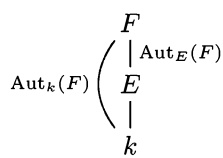


and therefore the lattice of intermediate fields is just as transparent:



(The intermediate fields are determined by recalling the generators of the corresponding subgroups, as found in Example 1.19: the flip $\sqrt{3} \mapsto -\sqrt{3}$ leaves $\mathbb{Q}(\sqrt{2})$ fixed, etc.) Galois theory tells us that there is *no other* intermediate field. ⌟

6.3. The fundamental theorem of Galois theory, II. Another popular notational device is to mark an extension by the corresponding Galois group, if the extension is Galois:



Example 6.4 shows that the extension $k \subseteq E$ need not be a Galois extension: $\mathbb{Q}(\sqrt[3]{2})$ is an intermediate field for the extension of $\mathbb{Q}$ into the splitting field for the polynomial $t^3 - 2$ (which is Galois, by part (1) of Theorem 6.9), but as we have seen, it is not Galois. The splitting field has degree 6 over $\mathbb{Q}$ and is a quadratic extension of $\mathbb{Q}(\sqrt[3]{2})$. This is in fact the typical situation: *every* finite separable extension may be enlarged to a Galois extension (Exercise 6.4).

In any case, the question of whether the extension of the base field k in an intermediate field E of a Galois extension $k \subseteq F$ is Galois is central to the theory. The outcome of the discussion will be very neat, and this is possibly the most striking part of the fundamental theorem of Galois theory:

Theorem 6.15. *Let $k \subseteq F$ be a Galois extension, and let E be an intermediate field. Then $k \subseteq E$ is Galois if and only if* $\mathrm{Aut}_E(F)$ *is* normal *in* $\mathrm{Aut}_k(F)$*; in this case, there is an isomorphism*

$$\mathrm{Aut}_k(E) \cong \frac{\mathrm{Aut}_k(F)}{\mathrm{Aut}_E(F)}.$$

The 'problem' with $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2})$ is that the automorphism group of the splitting field of $t^3 - 2$ is S_3 (as we will soon be able to verify: cf. Example 7.19). The subgroup corresponding to $\mathbb{Q}(\sqrt[3]{2})$ has index 3; hence (as we know, since we are well acquainted with S_3) it is not normal.

The fact that the Galois group of $k \subseteq E$ will turn out to be isomorphic to the quotient $\mathrm{Aut}_k(F)/\mathrm{Aut}_E(F)$ in the Galois case should not be too surprising, since we know already that in any case $[E : k]$ equals the index of $\mathrm{Aut}_E(F)$ in $\mathrm{Aut}_k(F)$, and the index equals the order of the quotient if the subgroup is normal. In general, $[E : k]$ equals the number of cosets of $\mathrm{Aut}_E(F)$ in $\mathrm{Aut}_k(F)$, and it is worth shooting for a concrete interpretation of these cosets, even when $k \subseteq E$ does not turn out to be a Galois extension.

We already know that the number of left-cosets of $\mathrm{Aut}_E(F)$ equals the index of $\mathrm{Aut}_E(F)$ in $\mathrm{Aut}_k(F)$, hence $[E : k]$, hence $[E : k]_s$ (because $k \subseteq E$ is separable, as a subextension of a separable extension), and hence there must exist a bijection between the left-cosets $g\,\mathrm{Aut}_E(F)$ as g ranges in $\mathrm{Aut}_k(F)$ and the *different embeddings $\iota : E \to \overline{k}$ in an algebraic closure of k, extending id_k*. We are now going to find a 'meaningful' bijection between these two sets.

Fix an embedding $F \subseteq \overline{k}$ of F in an algebraic closure of k, and let I be the set of embeddings $\iota : E \hookrightarrow \overline{k}$ extending the identity on k. It is a good idea to view $k \subseteq E$ as an abstract extension, rather than identifying E with a specific intermediate field in F; the fact that we can view E as an intermediate field of the extension $k \subseteq F$ simply means that $\iota(E) \subseteq F$ for *some* $\iota \in I$.

The following general observations clarify the situation considerably:

- For *all* $\iota \in I$, $\iota(E) \subseteq F$.

Indeed, $k \subseteq E$ is simple (finite and separable $\Longrightarrow$ simple, Proposition 5.19). If $E = k(\alpha)$, then ι is determined by $\iota(\alpha)$, which is necessarily a root of the minimal polynomial of α over k. But $k \subseteq F$ is normal and contains *some* root $\iota(\alpha)$ of this polynomial; hence F contains all of them (Definition 4.7). That is, $\iota(\alpha) \in F$ for all $\iota \in I$, implying $\iota(k(\alpha)) \subseteq F$ for all $\iota \in I$.

- Let $\iota \in I$, and let $g \in \mathrm{Aut}_k(F)$; then $g \circ \iota \in I$. Therefore, $\mathrm{Aut}_k(F)$ acts on I.

Here $g \circ \iota$ is interpreted in the following way: $\iota(E) \subseteq F$, as we have seen; hence g restricts to a homomorphism $\iota(E) \to F$, and $g \circ \iota$ is the composition of this homomorphism with ι.

This is an excellent moment to review the material on group actions we covered in §II.9, especially §II.9.3. Recall in particular that we were able to classify all *transitive* actions of a group on a set (Proposition II.9.9); this is precisely our situation:

- The action of $\mathrm{Aut}_k(F)$ on I is transitive.

This follows from the 'uniqueness of splitting fields', Lemma 4.2. Indeed, let $\iota_1, \iota_2 \in I$. Both $\iota_1(E)$, $\iota_2(E)$ contain k, and F is a splitting field over k (part (1) of Theorem 6.9); hence F is a splitting field (for the same polynomial) over both $\iota_1(E)$ and $\iota_2(E)$. By Lemma 4.2, there exists an automorphism $g : F \to F$ extending the

isomorphism

$$\iota_2 \circ \iota_1^{-1} : \iota_1(E) \to \iota_2(E).$$

Then $g \in \mathrm{Aut}_k(F)$ (since $\iota_2 \circ \iota_1^{-1}$ restricts to the identity on k), and the fact that g restricts to $\iota_2 \circ \iota_1^{-1}$ means precisely that $g \circ \iota_1 = \iota_2$.

Now choose one $\iota \in I$, and use it to identify E with an intermediate field of $k \subseteq F$. Then $\mathrm{Aut}_E(F) = \mathrm{Aut}_{\iota(E)}(F)$ is the subgroup of $\mathrm{Aut}_k(F)$ consisting of those elements g which restrict to the identity on $E = \iota(E)$, that is, the *stabilizer* of ι.

- *As an* $\mathrm{Aut}_k(F)$*-set,* I is isomorphic to the set of left-cosets of $\mathrm{Aut}_E(F)$ in $\mathrm{Aut}_k(F)$.

At this point, this follows immediately from Proposition II.9.9.

As promised, we have found a meaningful bijection between the two sets, 'explaining' the equality

$$[E : k] = [E : k]_s = |I| = [\mathrm{Aut}_k(F) : \mathrm{Aut}_E(F)]$$

that we had already discovered by numerical considerations (= Lagrange's theorem; cf. the beginning of §6.2). Further, we are now in a position to prove Theorem 6.15.

Proof of Theorem 6.15. As observed above, the $\mathrm{Aut}_k(F)$-stabilizer of $\iota \in I$ equals $\mathrm{Aut}_{\iota(E)}(F)$. By Proposition II.9.12, stabilizers of different ι are conjugates of each other; if $\mathrm{Aut}_E(F)$ is normal, it follows that for all $\iota \in I$

$$\mathrm{Aut}_{\iota(E)}(F) = \mathrm{Aut}_E(F).$$

This implies that $\iota(E) = E$ for all $\iota \in I$, since the Galois correspondence is bijective for the Galois extension $k \subseteq F$. It follows that $k \subseteq E$ satisfies condition (6) of Theorem 6.9; hence it is Galois.

Conversely, assume that $k \subseteq E$ is Galois; by the same condition (6), $\iota(E) = E$ for all $\iota \in I$. Restriction to $E = \iota(E)$ then defines a homomorphism

$$\rho : \mathrm{Aut}_k(F) \longrightarrow \mathrm{Aut}_k(E).$$

This homomorphism is surjective (because the $\mathrm{Aut}_k(F)$-action on I is transitive), and its kernel consists of those $g \in \mathrm{Aut}_k(F)$ which restrict to the identity on E, that is, precisely of $\mathrm{Aut}_E(F)$. This shows that $\mathrm{Aut}_E(F)$ is normal and establishes the stated isomorphism by virtue of the 'first isomorphism theorem' for groups, Corollary II.8.2. □

Remark 6.16. There is a parallel between Galois theory and the theory of covering spaces in topology. In this analogy, Galois extensions correspond to *regular* covers; the Galois group of an extension corresponds to the group of deck transformations; and Theorem 6.15 corresponds to the fact that the quotient of a regular cover by a normal subgroup of the group of deck transformations is again a regular cover.

More general (connected) covers correspond to more general algebraic extensions. A space is *simply connected* if and only if it admits no nontrivial connected covers, so this notion corresponds in field theory to the condition that a field K admits no nontrivial algebraic extensions, that is, that K is *algebraically closed.* Viewing the fundamental group of a space as the group of deck transformations

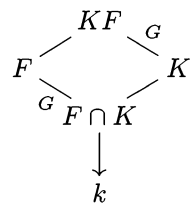
of its fundamental cover suggests that we should think of the Galois group of the algebraic closure[23] $k \subseteq \overline{k}$ as 'the fundamental group' of a field k.

In algebraic geometry this analogy is carried out to its natural consequences. A covering map of algebraic varieties $X \to Y$ determines a field extension $K(Y) \subseteq K(X)$, where $K(X)$, $K(Y)$ are the 'fields of rational functions' (in the affine case, these are just the fields of fractions of the corresponding coordinate rings; cf. Exercise 2.16). One can then use Galois theory to transfer to the algebro-geometric environment notions such as the fundamental group, without appealing to topological notions (such as 'continuous maps from S^1'), which would be problematic in, e.g., positive characteristic. ⌟

6.4. Further remarks and examples. Suppose $k \subseteq F$ and $k \subseteq K$ are two finite extensions contained in a larger extension of k (for example, we could choose specific embeddings $F \subseteq \overline{k}$, $K \subseteq \overline{k}$ in an algebraic closure of k). Then we can consider the *composite* FK (cf. Lemma 6.3) of F, K in the larger extension. It is natural to question how the Galois condition behaves under this operation.

Proposition 6.17. *Suppose $k \subseteq F$ is a Galois extension and $k \subseteq K$ is any finite extension. Then $K \subseteq KF$ is a Galois extension, and* $\mathrm{Aut}_K(KF) \cong \mathrm{Aut}_{F\cap K}(F)$.

Pictorially,



Proof. As $k \subseteq F$ is Galois, it is the splitting field of a separable polynomial $f(x) \in k[x] \subseteq K[x]$. The roots of $f(x)$ generate F over k, so they generate KF over K; in other words, KF is the splitting field of $f(x)$ over K, so $K \subseteq KF$ is Galois.

If $\sigma : KF \to KF$ is an automorphism extending the identity on K (and hence on k), then σ restricts to a homomorphism $F \to \sigma(F)$ extending the identity on k. Since $k \subseteq F$ is Galois, $\sigma(F) = F$ ((6) in Theorem 6.9). Thus, 'restriction' defines a natural group homomorphism

$$\rho : \mathrm{Aut}_K(KF) \longrightarrow \mathrm{Aut}_k(F).$$

We claim that ρ is injective. Indeed, by definition every $\sigma \in \mathrm{Aut}_K(KF)$ is the identity on K; if σ restricts to the identity on F, then σ is the identity on the composition KF; that is, $\sigma = \mathrm{id}$ in $\mathrm{Aut}_K(KF)$.

Therefore, $\mathrm{Aut}_K(KF)$ may be identified with a subgroup G of $\mathrm{Aut}_k(F)$. Since every $\sigma \in \mathrm{Aut}_K(KF)$ fixes K, the restriction $\rho(\sigma)$ fixes $F \cap K$; that is, $F^G \supseteq F \cap K$. Conversely, suppose $\alpha \in F^G$; then $K(\alpha)$ is in the fixed field of $\mathrm{Aut}_K(KF)$, which is K itself since $K \subseteq KF$ is Galois (condition (4) of Theorem 6.9). That is, $\alpha \in K$,

[23] Of course the extension $k \subseteq \overline{k}$ is not finite in general, so one needs the infinite version of Galois theory; cf. Remark 6.8.

and it follows that $F^G = F \cap K$. By Proposition 6.5, $G = \mathrm{Aut}_{F\cap K}(F)$, concluding the proof. □

We will leave to the reader the pleasure of verifying that if *both* $k \subseteq F$ and $k \subseteq K$ are Galois, then FK is also Galois *over* k, and so is $F \cap K$ (Exercise 6.13).

Example 6.18. We have studied cyclotomic fields $\mathbb{Q}(\zeta_n)$ as extensions of $\mathbb{Q}$; $\mathbb{Q}(\zeta_n)$ is the splitting field of $x^n - 1$, so these extensions are Galois; we have proved (Proposition 5.16) that $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\zeta_n))$ is isomorphic to the group of units in $\mathbb{Z}/n\mathbb{Z}$.

Now let k be any field of characteristic zero. The splitting field of $x^n - 1$ over k is the composite $k(\zeta)$ of k and $\mathbb{Q}(\zeta)$. By Proposition 6.17 the extension $k \subseteq k(\zeta)$ is Galois, and $\mathrm{Aut}_k(k(\zeta))$ is isomorphic to a *subgroup* of the group of units of $\mathbb{Z}/n\mathbb{Z}$. In particular, it is abelian. ⌟

The previous example gives an important class of *abelian* Galois extensions. Remarkably, we can classify all *cyclic* Galois extensions, if we impose some restriction on the base field k.

Proposition 6.19. *Let $k \subseteq F$ be an extension of degree m. Assume that k contains a primitive m-th root of* 1 *and* $\mathrm{char}\, k$ *does not divide*[24] m. *Then $k \subseteq F$ is Galois and cyclic if and only if $F = k(\delta)$, with $\delta^m \in k$.*

That is, under the given hypotheses on k, the *cyclic* Galois extensions of k are precisely those of the form $k \subseteq k(\sqrt[m]{c})$, with $c \in k$. This is part of *Kummer theory,* a basic tool in the study of abelian extensions. Proposition 6.19 will be a key ingredient in our application of Galois theory to the solvability of polynomial equations.

Proof. Let $\zeta \in k$ be a primitive m-th root of 1.

First assume that $F = k(\delta)$, with $\delta^m = c \in k$. Then all m roots of the polynomial $x^m - c$,

$$\delta\,,\ \zeta\delta\,,\ \zeta^2\delta\,,\ \cdots\,,\ \zeta^{m-1}\delta,$$

are in F, and F is generated by them; therefore F is the splitting field of the separable (since $\mathrm{char}\, k$ does not divide m) polynomial $x^m - c$ over k; hence (Theorem 6.9, part (1)) $k \subseteq F$ is Galois.

To see that $\mathrm{Aut}_k(F)$ is cyclic, note that every $\varphi \in \mathrm{Aut}_k(F)$ is determined by $\varphi(\delta)$, which is necessarily a root of $x^m - c$; therefore $\varphi(\delta) = \zeta^i\delta$ for some i, determined up to a multiple of m. This defines an isomorphism of $\mathrm{Aut}_k(F)$ with $\mathbb{Z}/m\mathbb{Z}$, as is immediately verified.

Conversely, assume that $k \subseteq F$ is Galois and $\mathrm{Aut}_k(F)$ is cyclic, generated by φ. The automorphisms φ^i, $i = 0, \ldots, m-1$, are pairwise distinct; thus they are linearly independent over F (Exercise 6.14). Therefore, there exists an $\alpha \in F$ such that[25]

$$\delta := \alpha + \zeta^{-1}\varphi(\alpha) + \cdots + \zeta^{-(m-2)}\varphi^{m-2}(\alpha) + \zeta^{-(m-1)}\varphi^{m-1}(\alpha) \neq 0.$$

[24] We will apply this result in §7.4; for convenience, we will take $\mathrm{char}\, k = 0$ in that application.

[25] This is called a *Lagrange resolvent.*

Note that

$$\varphi(\delta) = \varphi(\alpha) + \zeta^{-1}\varphi^2(\alpha) + \cdots + \zeta^{-(m-2)}\varphi^{m-1}(\alpha) + \zeta^{-(m-1)}\alpha = \zeta\delta$$

(since $\zeta \in k$, so $\varphi(\zeta) = \zeta$). This implies that δ is not fixed by any nonidentity element of $\mathrm{Aut}_k(F)$; that is, $\mathrm{Aut}_{k(\delta)}(F) = \{e\}$; that is, $k(\delta) = F$. Further,

$$\varphi(\delta^m) = (\varphi(\delta))^m = \zeta^m\delta^m = \delta^m;$$

that is, δ^m is fixed by φ, and hence by the whole of $\mathrm{Aut}_k(F)$. Since $k \subseteq F$ is Galois, this implies $\delta^m \in k$ (Theorem 6.9, part (4)). That is, δ satisfies the stated conditions. □

Exercises

6.1. ▷ Prove Lemma 6.3. [§6.1]

6.2. Prove that quadratic extensions are Galois.

6.3. ▷ Let $k \subseteq F$ be a Galois extension, and let E be an intermediate field. Prove that $E \subseteq F$ is a Galois extension. [§6.2]

6.4. ▷ Let $k \subseteq E$ be a finite separable extension. Prove that E may be identified with an intermediate field of a Galois extension $k \subseteq F$ of k.

In fact, prove that there is a *smallest* such extension $k \subseteq F$, in the sense that if $k \subseteq E \subseteq K$, with $k \subseteq K$ Galois, then there exists an embedding of F in K which is the identity on E. (The extension $k \subseteq F$ is the *Galois closure* of the extension $k \subseteq E$. It is clearly uniquely determined up to isomorphism.) [§6.3, 6.5]

6.5. We have proved (Proposition 5.19) that all finite separable extensions $k \subseteq E$ are simple. Let $k \subseteq F$ be a Galois closure of $k \subseteq E$ (Exercise 6.4). For $\alpha \in E$, prove that $E = k(\alpha)$ if and only if α is moved by all $\sigma \in \mathrm{Aut}_k(F) \smallsetminus \mathrm{Aut}_E(F)$.

6.6. • Prove that $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2+\sqrt{2}})$ is Galois, with cyclic Galois group. (Cf. Exercise 1.25.)

- Prove that $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{3+\sqrt{5}})$ is Galois and its Galois group is isomorphic to $(\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/2\mathbb{Z})$.
- Prove that $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{1+\sqrt{2}})$ is *not* Galois, and compute its Galois closure $\mathbb{Q} \subseteq F$. Prove that $\mathrm{Aut}_{\mathbb{Q}}(F) \cong D_8$. (Use Exercise IV.2.16.)

6.7. Let $p > 0$ be prime, and let $d \mid e$ be positive integers, so that there is an extension $\mathbb{F}_{p^d} \subseteq \mathbb{F}_{p^e}$. Prove that $\mathrm{Aut}_{\mathbb{F}_{p^d}} \mathbb{F}_{p^e}$ is cyclic, and describe a generator of this group. (This generalizes Proposition 5.8; use Galois theory.)

6.8. Let $k \subseteq F$ be a Galois extension of degree n, and let E be an intermediate field. Assume that $[E : k]$ is the smallest prime dividing n. Prove $k \subseteq E$ is Galois.

6.9. Let $k \subseteq F$ be a Galois extension of degree 75. Prove that there exists an intermediate field E, with $k \subsetneq E \subsetneq F$, such that the extension $k \subseteq E$ is Galois.

6.10. Let $k \subseteq F$ be a Galois extension and E an intermediate field. Prove that the normalizer of $\mathrm{Aut}_E(F)$ in $\mathrm{Aut}_k(F)$ is the set of $\sigma \in \mathrm{Aut}_k(F)$ such that $\sigma(E) \subseteq E$.

Use this to give an alternative proof of the fact that E is Galois over k if and only if $\mathrm{Aut}_E(F)$ is normal in $\mathrm{Aut}_k(F)$.

6.11. Let $k \subseteq E$ and $E \subseteq F$ be Galois extensions.

- Find an example showing that $k \subseteq F$ is not necessarily Galois.
- Prove that if every $\sigma \in \mathrm{Aut}_k(E)$ is the restriction of an element of $\mathrm{Aut}_k(F)$, then $k \subseteq F$ *is* Galois.

6.12. Find two algebraic extensions $k \subseteq F$, $k \subseteq K$ and embeddings $F \subseteq \overline{k}$, $\sigma_1 : K \subseteq \overline{k}$, $\sigma_2 : K \subseteq \overline{k}$ extending $k \subseteq \overline{k}$, such that the composites $F\sigma_1(K)$, $F\sigma_2(K)$ are *not* isomorphic.

Prove that no such example exists if F and K are Galois over k.

6.13. ▷ Let $k \subseteq F$ and $k \subseteq K$ be Galois extensions, and assume F and K are subfields of a larger field. Prove that $k \subseteq FK$ and $k \subseteq F \cap K$ are both Galois extensions. [§6.4, §7.4]

6.14. ▷ Let $k \subseteq F$ be a field extension, and let $\varphi_1, \ldots, \varphi_m \in \mathrm{Aut}_k(F)$ be pairwise distinct automorphisms. Prove that $\varphi_1, \ldots, \varphi_m$ are linearly independent over F. (Recycle/adapt the hint for Exercise VI.6.15.) [§6.4, 6.16, §IX.7.6]

6.15. Let $k \subseteq F$ be a Galois extension, and let $\alpha \in F$. Prove that

$$N_{k\subseteq F}(\alpha) = \prod_{\sigma\in\mathrm{Aut}_k(F)} \sigma(\alpha), \quad \mathrm{tr}_{k\subseteq F}(\alpha) = \sum_{\sigma\in\mathrm{Aut}_k(F)} \sigma(\alpha).$$

(Exercise 4.19.)

6.16. ▷ Let $k \subseteq F$ be a cyclic Galois extension of degree d, and let φ be a generator of $\mathrm{Aut}_k(F)$. Let $\alpha \in F$ be an element such that $N_{k\subseteq F}(\alpha) = 1$.

- Prove that the automorphisms $\mathrm{id}_F, \varphi, \ldots, \varphi^{d-1}$ are linearly independent over F. (Exercise 6.14.)
- Prove that there exists a $\gamma \in F$ such that

$$\beta := \gamma + \alpha\varphi(\gamma) + \alpha\varphi(\alpha)\varphi^2(\gamma) + \cdots + \alpha\varphi(\alpha)\cdots\varphi^{d-2}(\alpha)\varphi^{d-1}(\gamma) \neq 0.$$

- Prove that $\alpha\varphi(\alpha)\varphi^2(\alpha)\cdots\varphi^{d-1}(\alpha)\varphi^d(\gamma) = \gamma$, and deduce that $\alpha = \beta/\varphi(\beta)$.

Together with the result of Exercise 4.20, the conclusion is that an element α of a cyclic Galois extension as above has norm 1 if *and only if* there exists a β such that $\alpha = \beta/\varphi(\beta)$.

This is *Hilbert's theorem 90* (the 90-th theorem in Hilbert's *Zahlbericht*, a report on the state of number theory at the end of the nineteenth century commissioned by the German Mathematical Society). [6.17, §IX.7.6, IX.7.18]

6.17. Exercise 6.16 should remind the reader of the proof of Proposition 6.19, and for good reasons. Assume that $k \subseteq F$ is a cyclic Galois extension of degree m and that k contains a primitive m-th root ζ of 1. Use Hilbert's theorem 90 to prove

that $F = k(\delta)$ for a δ such that $\delta^m \in k$, thereby recovering one direction of the statement of Proposition 6.19. (Hint: What is $N_{k\subseteq F}(\zeta^{-1})$?)

The case of Hilbert's theorem 90 needed here is due to Kummer.

6.18. Use Hilbert's theorem 90 to find all rational roots a, b of the equation $a^2 + b^2 d = 1$, where d is a positive integer that is not a square. (Hint: $a^2 + b^2 d$ equals $N_{\mathbb{Q}\subseteq\mathbb{Q}(\sqrt{-d})}(a + b\sqrt{-d})$; cf. Exercise 1.12.)

6.19. ¬ Let $k \subseteq F$ be a cyclic Galois extension of degree d, and let φ be a generator of $\mathrm{Aut}_k(F)$. Let $\alpha \in F$ be an element such that $\mathrm{tr}_{k\subseteq F}(\alpha) = 0$. Prove an 'additive version' of Hilbert's theorem 90:

- Prove that there exists a $\gamma \in F$ such that $\mathrm{tr}_{k\subseteq F}(\gamma) \neq 0$.
- Stare at the expression

$$\alpha\varphi(\gamma) + (\alpha + \varphi(\alpha))\varphi^2(\gamma) + \cdots + (\alpha + \varphi(\alpha) + \cdots + \varphi^{d-2}(\alpha))\varphi^{d-1}(\gamma).$$

- Prove that there exists a $\beta \in F$ such that $\alpha = \beta - \varphi(\beta)$.

Together with Exercise 4.20, this says that an element α of a cyclic Galois extension as above has trace 0 if and only if there exists a β such that $\alpha = \beta - \varphi(\beta)$. For an application, see Exercise 7.15. [7.15]

---

7. Short march through applications of Galois theory

Galois theory is a pervasive tool, with wide-ranging applications. We have neither the competence nor room in this book for an adequate survey of applications of the theory, but even the few examples reviewed in this section should suffice to convey its considerable power.

7.1. Fundamental theorem of algebra. Theorem 6.15 and a little group theory are essentially all that is needed[26] to give an algebraic proof of the fundamental theorem of algebra, promised back in §V.5.3.

Theorem 7.1. *$\mathbb{C}$ is algebraically closed.*

Proof. Let $f(x) \in \mathbb{C}[x]$ be a nonconstant polynomial; we have to prove that $f(x)$ has roots in $\mathbb{C}$. Note that if $f(x)$ has no roots in $\mathbb{C}$, then neither does $f(x)\overline{f(x)} \in \mathbb{R}[x]$; that is, we may assume that $f(x)$ has real coefficients. Let F be a splitting field for $f(x)$ over $\mathbb{R}$; embed F in an algebraic closure of $\mathbb{R}$ (we 'don't know yet' that $\mathbb{C}$ is algebraically closed!), and consider the extension

$$\mathbb{R} \subseteq F(i).$$

This extension is Galois: it is the splitting field of the square-free part[27] of $f(x)(x^2+1)$. Let $G = \mathrm{Aut}_{\mathbb{R}}(F(i))$.

[26] We are cheating a little, actually, since the intermediate value theorem will also be used behind the scenes. This is not too surprising: the completeness of $\mathbb{R}$ must enter the proof somewhere.

[27] That is, take each irreducible factor of the polynomial with a power of exactly 1. Over a perfect field, square-free polynomials are clearly separable.

By Sylow I (Theorem IV.2.5), G has a 2-Sylow subgroup H. The index $[G : H]$ is odd, so it corresponds to a finite, separable extension $\mathbb{R} \subseteq E$ with $[E : \mathbb{R}]$ *odd.* However, every polynomial of odd degree over $\mathbb{R}$ has a real root (Exercise V.5.17, which only uses elementary calculus), and it follows that $\mathbb{R} \subseteq E$ is trivial (Exercise 5.23).

Therefore $[G : H] = 1$, proving that G is a 2-group. Since $\mathbb{C} = \mathbb{R}(i) \subseteq F(i)$, the Galois group of the (Galois) extension

$$\mathbb{C} \subseteq F(i)$$

is a subgroup of G; therefore it is a 2-group: $|\operatorname{Aut}_{\mathbb{C}}(F(i))| = 2^n$ for some $n \geq 0$.

Now recall (Proposition IV.2.6) that every group of order p^n, with p prime, contains subgroups of order p^m for all $0 \leq m \leq n$. In particular, if $n \geq 1$, then $\operatorname{Aut}_{\mathbb{C}}(F(i))$ contains a subgroup of order 2^{n-1}, which corresponds to a *quadratic extension of* $\mathbb{C}$ via the Galois correspondence. However, there are *no* quadratic extensions of $\mathbb{C}$, because there are no irreducible quadratic polynomials in $\mathbb{C}[x]$ (Exercise 5.23 again).

This contradiction shows that $n = 0$; that is, $\operatorname{Aut}_{\mathbb{C}}(F(i))$ is necessarily trivial. This proves that $F(i) = \mathbb{C}$ and in particular that $\mathbb{C}$ contains the roots of $f(x)$. □

The argument can be simplified further, bypassing the use of Sylow's theorem at the price of a slight increase in length. But why should we attempt to bypass Sylow's theorem, after the effort put into proving it in Chapter IV?

7.2. Constructibility of regular n-gons. We return one last time to the issue of constructing regular n-gons. Simpler considerations have given us constraints on n that must necessarily be fulfilled for the regular n-gon to be constructible: if $n = p$ is prime, we have found that $p - 1$ must be a power of 2 (§3.3); this was upgraded for any n to the condition that $\phi(n)$ is a power of 2, in §5.2 (and an even more explicit condition is given in Exercise 5.19). But these are all 'negative' results: we have not as yet proved that *if* $\phi(n)$ is a power of 2, *then* the regular n-gon is constructible. This is where Galois theory comes into the picture.

Proposition 7.2. *Let $k \subseteq F$ be a Galois extension, and assume $[F : k] = p^r$ for some prime p and $r \geq 0$. Then there exist intermediate fields*

$$k = E_0 \subseteq E_1 \subseteq E_2 \subseteq \cdots \subseteq E_r = F$$

such that $[E_i : E_{i-1}] = p$ for $i = 1, \ldots, r$.

Proof. As the Galois correspondence is bijective for Galois extensions (Theorem 6.9, part (5)), this statement follows immediately from the fact that a group of order p^r, with p prime, has a complete series of p-subgroups; cf. for example the discussion following the statement of Theorem IV.2.8. □

Theorem 7.3. *The regular n-gon is constructible by straightedge and compass if and only if $\phi(n)$ is a power of 2.*

Proof. As recalled above, we have already established the $\Longrightarrow$ direction.

For the converse, assume $\phi(n) = 2^r$ for some r. The extension $\mathbb{Q} \subseteq \mathbb{Q}(\zeta_n)$ is Galois (it is the splitting field of $\Phi_n(x)$), of order $[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \phi(n) = 2^r$ (Proposition 5.14). Proposition 7.2 shows that the condition given in Theorem 3.4 is satisfied and therefore that ζ_n is constructible, as needed. □

For example, $\phi(17) = 16 = 2^4$; the Galois group of $\mathbb{Q}(\zeta_{17})$ over $\mathbb{Q}$ is isomorphic to $(\mathbb{Z}/17\mathbb{Z})^*$ (by Proposition 5.16), and therefore (Theorem IV.6.10) it is cyclic, of order 16. A generator of $(\mathbb{Z}/17\mathbb{Z})^*$ is $[6]_{17}$; it follows that $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\zeta_{17}))$ is generated by the automorphism σ defined by $\zeta_{17} \mapsto \zeta_{17}^6$. The sequence of subgroups

$$\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\zeta_{17})) = \langle\sigma\rangle \supseteq \langle\sigma^2\rangle \supseteq \langle\sigma^4\rangle \supseteq \langle\sigma^8\rangle \supseteq \{e\}$$

corresponds via the Galois correspondence to a sequence

$$\mathbb{Q} \subseteq \mathbb{Q}(\delta_1) \subseteq \mathbb{Q}(\delta_2) \subseteq \mathbb{Q}(\delta_3) \subseteq \mathbb{Q}(\zeta_{17}).$$

It is instructive to apply what we now know to see how one may determine a generator δ_1 for the first extension, that is, the fixed field of σ^2. Let $\zeta = \zeta_{17}$. As a $\mathbb{Q}$-vector space, $\mathbb{Q}(\zeta)$ is generated by $\zeta, \zeta^2, \ldots, \zeta^{16}$. To find the fixed field of σ^2, write out a general element of $\mathbb{Q}(\zeta)$:

$$\begin{aligned} a_1\zeta + a_2\zeta^2 + a_3\zeta^3 + a_4\zeta^4 + a_5\zeta^5 + a_6\zeta^6 + a_7\zeta^7 + a_8\zeta^8 + a_9\zeta^9 \\ + a_{10}\zeta^{10} + a_{11}\zeta^{11} + a_{12}\zeta^{12} + a_{13}\zeta^{13} + a_{14}\zeta^{14} + a_{15}\zeta^{15} + a_{16}\zeta^{16} \end{aligned}$$

and apply σ^2: $\zeta \mapsto (\zeta^6)^6 = \zeta^{36} = \zeta^2$:

$$\begin{aligned} a_1\zeta^2 + a_2\zeta^4 + a_3\zeta^6 + a_4\zeta^8 + a_5\zeta^{10} + a_6\zeta^{12} + a_7\zeta^{14} + a_8\zeta^{16} + a_9\zeta \\ + a_{10}\zeta^3 + a_{11}\zeta^5 + a_{12}\zeta^7 + a_{13}\zeta^9 + a_{14}\zeta^{11} + a_{15}\zeta^{13} + a_{16}\zeta^{15}. \end{aligned}$$

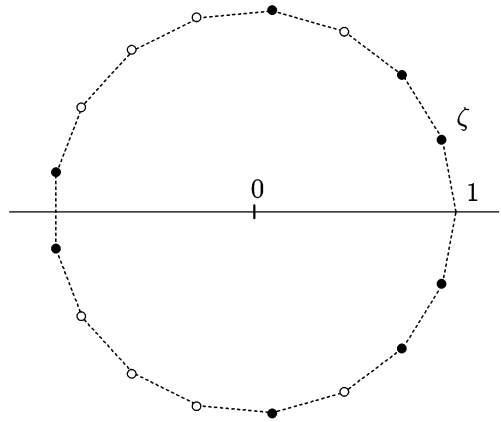
The element is fixed if and only if

$$\begin{aligned} a_1 = a_2 = a_4 = a_8 = a_{16} = a_{15} = a_{13} = a_9, \\ a_3 = a_6 = a_{12} = a_7 = a_{14} = a_{11} = a_5 = a_{10}, \end{aligned}$$

and it follows that the fixed field of σ^2 is generated by

$$\delta_1 = \zeta + \zeta^2 + \zeta^4 + \zeta^8 + \zeta^{16} + \zeta^{15} + \zeta^{13} + \zeta^9$$

and by $\zeta^3 + \zeta^6 + \zeta^{12} + \zeta^7 + \zeta^{14} + \zeta^{11} + \zeta^5 + \zeta^{10} = -1 - \delta_1$ (remember that ζ is a root of $\Phi_{17}(x) = x^{16} + \cdots + x + 1$). Pictorially,



The sum of the roots marked by black dots is δ_1; the white roots add up to $-1-\delta_1$. The theory tells us that δ_1 has degree 2 over $\mathbb{Q}$, so δ_1^2 must be a rational combination of 1 and δ_1. Indeed, pencil and paper and a bit of patience (and the relation $\zeta^{17}=1$) give

$$\begin{aligned}\delta_1^2 &= (\zeta+\zeta^2+\zeta^4+\zeta^8+\zeta^{16}+\zeta^{15}+\zeta^{13}+\zeta^9)^2\\ &= 3(\zeta+\zeta^2+\zeta^4+\zeta^8+\zeta^{16}+\zeta^{15}+\zeta^{13}+\zeta^9)\\ &\qquad +4(\zeta^3+\zeta^6+\zeta^{12}+\zeta^7+\zeta^{14}+\zeta^{11}+\zeta^5+\zeta^{10})+8\\ &= 3\delta_1+4(-1-\delta_1)+8\\ &= -\delta_1+4.\end{aligned}$$

Thus, we find that $\delta_1^2+\delta_1-4=0$. Solving for δ_1 (note that $\delta_1>0$, as the black dots to the right of 0 outweigh those to the left),

$$\delta_1=\frac{-1+\sqrt{17}}{2}.$$

So we could easily construct δ_1. The same procedure applied to the other extensions may be used to produce δ_2, δ_3, and finally ζ_{17}, whose real part is the complicated expression given at the end of §3.3.

7.3. Fundamental theorem on symmetric functions. Let $t_1,\dots,t_n$ be *indeterminates*. The polynomial

$$P_n(x):=(x-t_1)\cdots(x-t_n)\in\mathbb{Z}[t_1,\dots,t_n][x]$$

is universal in the sense that *every* polynomial of degree n with coefficients in (say) an integral domain may be obtained from $P_n(x)$ by suitably specifying $t_1,\dots,t_n$, possibly in a larger ring (Exercise 7.2). The coefficients of the expansion

$$P_n(x)=x^n-s_1(t_1,\dots,t_n)x^{n-1}+\cdots+(-1)^n s_n(t_1,\dots,t_n)$$

are (up to sign) the *elementary symmetric functions* of $t_1, \ldots, t_n$. For example, for $n = 3$,

$$\begin{aligned} s_1(t_1, t_2, t_3) &= t_1 + t_2 + t_3, \\ s_1(t_1, t_2, t_3) &= t_1 t_2 + t_1 t_3 + t_2 t_3, \\ s_1(t_1, t_2, t_3) &= t_1 t_2 t_3. \end{aligned}$$

The functions $s_i(t_1, \ldots, t_n)$ are *symmetric* in the sense that they are invariant under permutations of $t_1, \ldots, t_n$ (because $P_n(x)$ is invariant); they are *elementary* by virtue of the following important result, which has a particularly simple proof thanks to Galois theory.

Theorem 7.4 (Fundamental theorem on symmetric functions). *Let K be a field, and let $\varphi \in K(t_1, \ldots, t_n)$. Then φ is symmetric if and only if it is a rational function (with coefficients in K) of the elementary symmetric functions $s_1, \ldots, s_n$.*

Here we are viewing the s_i's as elements of $K[t_1, \ldots, t_n]$, which we can do unambiguously since $\mathbb{Z}$ is initial in Ring.

Proof. Let $F = K(t_1, \ldots, t_n)$, and let $k = K(s_1, \ldots, s_n)$ be the subfield generated by the elementary symmetric functions over K.

Then F is a splitting field of the separable polynomial $P_n(x)$ over k. In particular $k \subseteq F$ is a Galois extension, and

$$|\operatorname{Aut}_k(F)| = [F : k] \leq n!$$

by Lemma 4.2. The symmetric group S_n acts (faithfully) on F by permuting $t_1, \ldots, t_n$, and this action is the identity on $k = K(s_1, \ldots, s_n)$ since each s_i is symmetric. Thus S_n may be identified with a subgroup of $\operatorname{Aut}_k(F)$; but $|S_n| = n!$, so it follows that $\operatorname{Aut}_k(F) = S_n$.

Since $k \subseteq F$ is Galois, we obtain $k = F^{S_n}$. This says precisely that if $\varphi \in K(t_1, \ldots, t_n)$, then φ is invariant under the S_n action on $t_1, \ldots, t_n$ if and only if $\varphi \in k = K(s_1, \ldots, s_n)$, which is the statement. □

It is not difficult to strengthen the result of Theorem 7.4 and prove that every symmetric *polynomial* is in fact a *polynomial* in the elementary symmetric functions.

Note that the argument given in the proof amounts to the following result, which should be recorded for individual attention:

Lemma 7.5. *Let K be a field, $t_1, \ldots, t_n$ indeterminates, and let $s_1, \ldots, s_n$ be the elementary symmetric functions on $t_1, \ldots, t_n$. Then the extension*

$$K(s_1, \ldots, s_n) \subseteq K(t_1, \ldots, t_n)$$

is Galois, with Galois group S_n.

This result and Cayley's theorem (Theorem II.9.5) immediately yield the following:

Corollary 7.6. *Let G be a finite group. Then there exists a Galois extension $k \subseteq F$ such that $\operatorname{Aut}_k(F) \cong G$.*

Proof. Exercise 7.4. □

It is not known whether k may be chosen to be $\mathbb{Q}$ in Corollary 7.6. The proof hinted at above realizes k as a rather large field, and it is not at all clear how one may 'descend' from this field to $\mathbb{Q}$. This is the *inverse Galois problem*, a subject of current research. The reader can appreciate the difficulty of this problem by noting that even realizing S_n as a Galois group over $\mathbb{Q}$ (for all n) is not so easy, although it is true that for every n there are infinitely many polynomials of degree n in $\mathbb{Z}[x]$ whose splitting fields have Galois group S_n over $\mathbb{Q}$ (see Example 7.20 and Exercise 7.11 for the case $n = p$ prime). If you want a memorable example to carry with you, Schur showed (1931) that the splitting field of the 'truncated exponential'

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}$$

has Galois group S_n over $\mathbb{Q}$ for $n \neq 4$ (and A_n for $n = 4$).

It is known (Shafarevich, 1954) that every finite *solvable* group is the Galois group of a Galois extension over $\mathbb{Q}$.

The inverse Galois problem may be rephrased as asking whether every finite group is a quotient of the Galois group of $\overline{\mathbb{Q}}$ over $\mathbb{Q}$. We have already mentioned that this object (which may be interpreted as the 'fundamental group of $\mathbb{Q}$'; cf. Remark 6.16) is mysterious and extremely interesting[28].

Going back to Lemma 7.5, the alternating group A_n has index 2 in S_n, so it corresponds to a quadratic extension of $K(s_1, \ldots, s_n)$ via the Galois correspondence. It is easy to determine a generator of this extension: it must be a function of $t_1, \ldots, t_n$ which is invariant under the action of all *even* permutations and not invariant under the action of odd permutations. We have used such a function precisely to define even/odd permutations, in §IV.4.3:

$$\Delta = \prod_{1 \leq i < j \leq n} (t_i - t_j).$$

To see that Δ has degree 2 over $K(s_1, \ldots, s_n)$, simply note that the *discriminant*

$$D = \Delta^2 = \prod_{1 \leq i < j \leq n} (t_i - t_j)^2$$

is invariant under the action of all permutations, and therefore it belongs to the field $K(s_1, \ldots, s_n)$ by Theorem 7.4. This proves

Corollary 7.7. *With notation as above, the extension*

$$K(s_1, \ldots, s_n)(\sqrt{D}) \subseteq K(t_1, \ldots, t_n)$$

is Galois, with Galois group A_n.

[28] According to J. S. Milne, the 'most interesting object in mathematics' is the *étale fundamental group of the projective line over* $\mathbb{Q}$ *with the three points* 0, 1, ∞ *removed*. This is an extension of the Galois group of $\overline{\mathbb{Q}}$ over $\mathbb{Q}$.

7.4. Solvability of polynomial equations by radicals. This is possibly the most famous elementary application of Galois theory and one that is very close to Galois' own original motivation.

Quadratic polynomial equations

$$x^2 + bx + c = 0$$

are famously solved (in[29] characteristic $\neq 2$) by the quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2}.$$

Analogous, but more complicated, formulas exist for equations of degree 3 and 4 (Tartaglia/Cardano/Ferrari, around 1540; this generated a priority dispute which degenerated into base personal attacks). Equations of degree 5 and higher resisted the best efforts of mathematicians for centuries. The most ambitious goal, in line with what occurs for degrees 2, 3, and 4, would be to produce a formula for the solutions to the 'general' polynomial equation

$$x^n + a_{n-1}x^{n-2} + \cdots + a_0 = 0$$

in terms of the coefficients a_i and basic operations such as taking roots. This would be a solution 'by radicals' of the equation.

Well, such a formula simply does not exist:

Theorem 7.8. *The general polynomial equation of degree* 5 *or higher admits no solution by radicals.*

We are being intentionally vague about the ground field in which the problem is posed; a 'formula' such as the quadratic formula ought to hold over any field in which it makes sense; for example, the field should not have characteristic 2 in the case of the quadratic formula. An easy way to avoid any such snag is to work in characteristic zero: this has the main technical advantage of ensuring automatic separability of polynomials $x^m - 1$, whose roots define the 'radicals'; cf. Proposition 6.19. Therefore, we will assume in what follows that the ground field has characteristic 0.

The 'general polynomial' is the 'universal' polynomial $P_n(x)$ introduced in §7.3, taken over a chosen ground field K: the general polynomial

$$x^n - s_1 x^{n-1} + \cdots + (-1)^n s_n \in k[x],$$

with $k = K(s_1, \ldots, s_n)$ and s_i elementary symmetric functions in ('indeterminate') roots $t_1, \ldots, t_n$. It is in fact no harder to deal with the more general case of arbitrary irreducible polynomials $f(x) \in k[x]$, over any field k (of characteristic zero); this is what we will do.

A formula by radicals expresses a root t_i of $f(x)$ as an element of an extension obtained from k in the following fashion:

$$k \subseteq k(\delta_1) \subseteq k(\delta_1, \delta_2) \subseteq \cdots \subseteq k(\delta_1, \ldots, \delta_r)$$

where for all $i = 1, \ldots, r$ we have $\delta_i^{m_i} \in k(\delta_1, \ldots, \delta_{i-1})$ for some exponent m_i.

[29] In characteristic 2, one can give a formula in terms of '2-roots', which are roots of the Artin-Schreier equation, and avoid the problematic division by 2; see Exercise 7.16.

Definition 7.9. Extensions of this type are called *radical extensions.* ⌟

Some facts about radical extensions are immediate from the definition: for example, if $k \subseteq F$ and $k \subseteq K$ are two radical extensions and F, K are subfields of a larger field, then $k \subseteq FK$ is radical (Exercise 7.5). Also,

Lemma 7.10. *Every separable radical extension is contained in a Galois radical extension.*

Proof. Let $k \subseteq F$ be a separable radical extension. In particular $k \subseteq F$ is finite and separable, so $F = k(\alpha)$ for some $\alpha \in F$ (Proposition 5.19). Let $p(x)$ be the minimal polynomial of α over k. The splitting field L of $p(x)$ over k may be obtained as the composite of $\sigma(F)$ as σ ranges over the embeddings of F in $\overline{k}$, extending id_k: indeed, this composite is generated over k by all $\sigma(\alpha)$, which range over all roots of $p(x)$. As a composite of radical extensions, L is radical over k. As a splitting field of a separable polynomial, L is Galois over k (Theorem 6.9, part (1)), and we are done. □

Galois theory will allow us to relate solvability by radicals of a polynomial $f(x)$ with a precise statement on the splitting field of $f(x)$. Here is a first formalization of this connection:

Lemma 7.11. *Let k be a field of characteristic 0, and let $f(x) \in k[x]$ be an irreducible polynomial. Then there exists a formula solving $f(x)$ by radicals if and only if the splitting field of $f(x)$ is contained in a Galois radical extension.*

Proof. If the splitting field of $f(x)$ is contained in a radical extension (Galois or not), then the roots may be written as combinations of field operations and radicals, as needed.

For the converse, assume $f(x)$ is solvable by radical. A formula for a root t_1 can be turned into a formula for any other root t_i by applying an element σ_i of (for example) $\mathrm{Aut}_k(\overline{k})$ sending t_1 to t_i (such an automorphism exists since $f(x)$ is irreducible). Thus, if a formula exists for one root, then a formula exists for all roots. The composite of all the corresponding radical extensions gives one, possibly larger, radical extension of k containing all roots of $f(x)$. Therefore, the splitting field of $f(x)$ is contained in a radical extension of k. By Lemma 7.10, this extension may be assumed to be Galois, as needed. □

In a nutshell, we will understand solvability of polynomial equations by radicals if we understand radical Galois extensions. As demanded by the general philosophy underlying Galois theory, we should aim to characterize these extensions in terms of their Galois groups. It turns out that the right condition (modulo technical considerations) is that the Galois group should be *solvable.*

Definition 7.12. A Galois extension $k \subseteq F$ is *solvable* if $\mathrm{Aut}_k(F)$ is a solvable group. ⌟

The reader may want to revisit §IV.3.3 now, for a reminder on solvability in group theory. (We are approaching an explanation for the 'solvable' terminology!)

Just as radical extensions decompose as sequence of cyclic extensions, solvable groups have cyclic composition factors (Proposition IV.3.11). We will say[30] that a radical, resp., solvable, extension $k \subseteq F$ 'has enough roots of 1' if k contains a primitive M-th root of 1 for a common multiple M of the order of the corresponding cyclic factors.

Lemma 7.13. *Let $k \subseteq F$ be a Galois extension, with* $\operatorname{char} k = 0$. *Provided it has enough roots of* 1, *$k \subseteq F$ is radical if and only if it is solvable.*

Proof. Modulo the fundamental theorem of Galois theory, this is a straightforward generalization of Proposition 6.19.

Indeed, assume that $k \subseteq F$ is radical:

$$k \subseteq k(\delta_1) \subseteq \cdots \subseteq k(\delta_1, \ldots, \delta_r) = F$$

with $\delta_i^{m_i} \in k(\delta_1, \ldots, \delta_{i-1})$. By hypothesis, k contains primitive m_i-th roots of 1 for all i. Consider the corresponding series of subgroups of $G = \operatorname{Aut}_k(F(\zeta))$:

$$G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_r = \{e\},$$

with $G_i = \operatorname{Aut}_{k(\delta_1,\ldots,\delta_i)}(F)$. Note that F is Galois over each intermediate field. Each extension

$$k(\delta_1, \ldots, \delta_{i-1}) \subseteq k(\delta_1, \ldots, \delta_i)$$

satisfies the hypothesis of Proposition 6.19, and it follows that it is Galois and cyclic. By the fundamental theorem (Theorem 6.15) each G_i is normal in the previous G_{i-1}, with cyclic quotient, and it follows that G is solvable (Proposition IV.3.11 (ii)).

The same argument in reverse, using the converse implication in Proposition 6.19, proves that solvable extensions with enough roots of 1 are radical. □

How do we account for the possible absence of the needed roots of 1? Surprisingly, this affects the statement in a rather minor way.

Proposition 7.14. *Let k be a field of characteristic* 0, *and let $k \subseteq F$ be a Galois extension. Then $k \subseteq F$ is solvable if and only if it is contained in a Galois radical extension.*

The characteristic 0 in Proposition 7.14 is an overshoot; as in Proposition 6.19, it would suffice to require that $\operatorname{char} k$ does not divide the relevant degrees.

Proof. Assume $\operatorname{Aut}_k(F)$ is solvable. Let M be a common multiple of the order of the cyclic quotients, and let ζ be a primitive M-th root of 1 in an algebraic closure $\overline{k}$ of k. The splitting field $k(\zeta)$ of $x^M - 1$ over k is Galois over k ($x^M - 1$ is separable since k has characteristic 0). By Proposition 6.17, the extension $k(\zeta) \subseteq F(\zeta)$ is also Galois, with Galois group isomorphic to $\operatorname{Aut}_{k(\zeta)\cap F}(F)$. As $k(\zeta) \cap F$ is Galois over k (Exercise 6.13), this subgroup is normal, and it follows (Corollary IV.3.13) that $\operatorname{Aut}_{k(\zeta)}(F(\zeta))$ is solvable.

[30] Warning! This is nonstandard terminology.

Since $k(\zeta) \subseteq F(\zeta)$ has enough roots of 1, Lemma 7.13 implies that it is radical. So is $k \subseteq F(\zeta)$, since $k \subseteq k(\zeta)$ is itself trivially radical. This proves that $k \subseteq F$ is contained in a radical extension.

Conversely, assume that $k \subseteq F$ is Galois and contained in a radical Galois extension $k \subseteq L$. Let M be a common multiple of the order of cyclic factors in this extension, and let ζ be a primitive M-th root of 1. The composite $L(\zeta)$ is Galois and radical over $k(\zeta)$ (Proposition 6.17 and Exercise 7.5); further, it has enough roots of 1.

By Lemma 7.13, $k(\zeta) \subseteq L(\zeta)$ is solvable. It follows that $k \subseteq L(\zeta)$ is solvable. Indeed, this extension is Galois (Exercise 6.13); apply the fundamental theorem to

$$k \subseteq k(\zeta) \subseteq L(\zeta)$$

to establish that $\mathrm{Aut}_k(k(\zeta))$ is isomorphic to the quotient of $\mathrm{Aut}_k(L(\zeta))$ by its normal subgroup $\mathrm{Aut}_{k(\zeta)}(L(\zeta))$. Both $\mathrm{Aut}_{k(\zeta)}(L(\zeta))$ (as shown above) and $\mathrm{Aut}_k(k(\zeta))$ (an abelian group; cf. Example 6.18) are solvable, and it follows that $\mathrm{Aut}_k(L(\zeta))$ is solvable as promised, by Corollary IV.3.13.

Since F is an intermediate field and Galois over k, by the fundamental theorem $\mathrm{Aut}_k(F)$ is a homomorphic image of $\mathrm{Aut}_k(L(\zeta))$. This shows that $\mathrm{Aut}_k(F)$ is a homomorphic image of a solvable group; hence it is itself solvable (Corollary IV.3.13 again), and we are done. □

We may have lost sight of where we were heading; but we are ready to reconnect with the study of solutions to polynomial equations.

Definition 7.15. The *Galois group* $\mathrm{Gal}_k(f(x))$ *of a separable polynomial* $f(x) \in k[x]$ is the Galois group of the splitting field of $f(x)$ over k. ⌟

Corollary 7.16. *Let k be a field of characteristic 0, and let $f(x) \in k[x]$ be an irreducible polynomial. Then $f(x)$ is solvable by radicals if and only if its Galois group is solvable.*

Proof. This is an immediate consequence of Lemma 7.11 and Proposition 7.14. □

Corollary 7.16 is called *Galois' criterion*. Ruffini (1799) and Abel (1824) had previously established that general formulas in radicals for the solutions of equations of degree ≥ 5 do not exist (that is, Theorem 7.8); but it was Galois who identified the precise condition given in Corollary 7.16. Of course, Theorem 7.8 follows immediately from Galois' criterion:

Proof of Theorem 7.8. By Lemma 7.5, the Galois group of the general polynomial of degree n is S_n. The group S_n is not solvable for $n \geq 5$ (Corollary IV.4.21); hence the statement follows from Corollary 7.16. □

In fact, we could now do more: we know that S_3 and S_4 *are* solvable (cf. Exercise IV.3.16); from a composition series with cyclic quotients we could in principle decompose explicitly the splitting field of general polynomials of degree 3 and 4 as radical extensions and as a consequence recover the Tartaglia/Cardano/Ferrari formulas for their solutions.

7.5. Galois groups of polynomials. Impressive as Galois' criterion is, it does not in itself produce a single polynomials of degree (say) 5 over $\mathbb{Q}$ that is not solvable by radicals. Finding such polynomials requires a certain amount of extra work; we will be able to do this by the end of this subsection.

Computing Galois groups of polynomials is in fact a popular sport among algebraists, excelling at which would require much more information than the reader will glean here. We will just list a few straightforward observations.

To begin with, recall that an element of $\mathrm{Aut}_k(F)$ must send roots of a polynomial $f(x) \in k[x]$ to roots of the same polynomial, and if $f(x)$ is irreducible and factors completely in F, then there are automorphisms sending any root of $f(x)$ to any other root (Proposition 1.5, Corollary 1.7). This elementary observation may be rephrased as follows:

Lemma 7.17. *Let $f(x) \in k[x]$ be a separable irreducible polynomial of degree n. Then $\mathrm{Gal}_k(f(x))$ acts transitively on the set of roots of $f(x)$ in $\overline{k}$. In particular, $\mathrm{Gal}_k(f(x))$ may be identified with a* transitive *subgroup of the symmetric group S_n.*

Of course a subgroup of S_n is *transitive* if the corresponding action on $\{\mathbf{1}, \ldots, \mathbf{n}\}$ is transitive; the diligent reader has run across this terminology in Exercise IV.4.12.

The reader will easily produce a statement analogous to Lemma 7.17 in case $f(x)$ is *reducible*. Of course the action is not transitive in this case, and if the factors of $f(x)$ have degrees $n_1, \ldots, n_r$, then $\mathrm{Gal}_k(f(x))$ is contained in a subgroup of S_n isomorphic to $S_{n_1} \times \cdots \times S_{n_r}$.

Lemma 7.17 (or its 'reducible' variations) already give some information. For example, we see that the Galois group of a separable irreducible cubic can only be A_3 or S_3, since these are the only transitive subgroups of S_3. Similarly, the range of possibilities for irreducible polynomials of degree 4 is rather restricted: S_4, A_4, and isomorphic copies of the dihedral group D_8, of $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, and of $\mathbb{Z}/4\mathbb{Z}$ (Exercise 7.8). One approach to the computation of Galois groups of polynomials amounts to defining invariants imposing further restrictions, leading to algorithms deciding which of such possibilities occurs for a given polynomial.

We have already encountered the most important of these invariants: if the roots of a separable polynomial $f(x) \in k[x]$ in its splitting field are $\alpha_1, \ldots, \alpha_n$, the *discriminant* of $f(x)$ is the element $D = \Delta^2$, where

$$\Delta = \prod_{1 \le i < j \le n} (\alpha_i - \alpha_j).$$

Every permutation of the roots fixes D, so D must be fixed by the whole Galois group; therefore, $D \in k$. Odd permutations move Δ (if $\mathrm{char}\, k \neq 2$), and even permutations fix it; therefore Δ is fixed by the Galois group G (in other words, $\Delta \in k$) if and only if $G \subseteq A_n$. This proves

Lemma 7.18. *Let k be a field of characteristic $\neq 2$, and let $f(x) \in k[x]$ be a separable polynomial, with discriminant D. Then the Galois group of $f(x)$ is contained in the alternating group A_n if and only if D is a square in k.*

Example 7.19. Lemma 7.18 and a discriminant computation are all that is needed to compute the Galois group of an irreducible *cubic* polynomial

$$f(x) = x^3 + ax^2 + bx + c.$$

It would be futile to try to remember the discriminant

$$D = a^2b^2 - 4a^3c - 4b^3 + 18abc - 27c^2;$$

but one may remember the trick of shifting x by $a/3$ (in characteristic $\neq 3$), with the effect of killing the coefficient of x^2:

$$f(x - \frac{a}{3}) = x^3 + px + q$$

for suitable p and q. This does not change D (shifting all roots α_i by the same amount has no effect on the differences $\alpha_i - \alpha_j$), yet

$$D = -4p^3 - 27q^2$$

is a little more memorable.

By Lemma 7.18 and the preceding considerations, the Galois group is A_3 if D is a square and S_3 otherwise. For example, $x^3 - 2$ has Galois group S_3 over $\mathbb{Q}$, since $D = -108$ is not a square in $\mathbb{Q}$; $x^3 - 3x + 1$ has discriminant $81 = 9^2$, and therefore it has Galois group $A_3 \cong \mathbb{Z}/3\mathbb{Z}$. ⌟

The reader will have no difficulty locating a discussion of the different possibilities for polynomials of degree 4 and detailed information for higher degree polynomials. We will just highlight the following simple observation.

Example 7.20. Let $f(x) \in \mathbb{Q}[x]$ be an irreducible polynomial of degree p, where p is prime. Assume that $f(x)$ has $p - 2$ real roots and 2 nonreal, complex roots. Then the Galois group of $f(x)$ is S_p.

Indeed, complex conjugation induces an automorphism of the splitting field and acts by interchanging the two nonreal roots, so the Galois group G, as a subgroup of S_p, contains a transposition. On the other hand, the degree of the splitting field (and hence $|G|$) is divisible by p, because it contains a simple extension of order p, obtained by adjoining any one root to $\mathbb{Q}$. Since p is prime, G contains an element of order p by Cauchy's theorem (Theorem IV.2.1); the only elements of order p in S_p are p-cycles, so G contains a p-cycle. It follows that $G = S_p$, by (a simple variation of) Exercise IV.4.7.

For example, the Galois group of $f(x) = x^5 - 5x - 1$ over $\mathbb{Q}$ is S_5, giving a concrete example of a quintic that cannot be solved by radicals (Exercise 7.10).

The reader will use this technique to produce polynomials of every prime degree p in $\mathbb{Z}[x]$ whose Galois group is S_p (Exercise 7.11). ⌟

7.6. Abelian groups as Galois groups over $\mathbb{Q}$. Having mentioned the inverse Galois problem in §7.3, we should point out that the reader is now in the position of proving that every finite *abelian* group may be realized as the Galois group of an extension over $\mathbb{Q}$.

In fact, the reader can prove a much more precise result: every finite abelian group may be realized as the group of some intermediate field of the extension $\mathbb{Q} \subseteq \mathbb{Q}(\zeta_n)$ of $\mathbb{Q}$ in a cyclotomic field. This uses a number-theoretic fact:

For every integer N, there are infinitely many primes p such that $p \equiv 1 \bmod N$.

This is a particular case of *Dirichlet's theorem* (1837), which states that if a, b are positive integers and $\gcd(a, b) = 1$, then there are infinitely many primes of the form $a + nb$ with $n > 0$. The particular case $a = 1, b = N$ needed here was apparently already known to Euler and can be proven by elementary means (in fact, there is a proof using cyclotomic polynomials: see Exercise 5.18). Assuming this fact, argue as follows:

—By the classification theorem (Theorem IV.6.6), every finite abelian group G is isomorphic to a product of cyclic groups

$$(\mathbb{Z}/n_1\mathbb{Z}) \times \cdots \times (\mathbb{Z}/n_r\mathbb{Z}).$$

—By Dirichlet's theorem, we can then choose *distinct* primes p_i such that $p_i \equiv 1 \bmod n_i$.

—Let $n = p_1 \cdots p_r$; by the Chinese Remainder Theorem (Theorem V.6.1)

$$(\mathbb{Z}/n\mathbb{Z})^* \cong (\mathbb{Z}/p_1\mathbb{Z})^* \times \cdots \times (\mathbb{Z}/p_r\mathbb{Z})^*,$$

and the i-th factor on the right-hand side is cyclic of order $p_i - 1$, a multiple of n_i.

—It follows that $(\mathbb{Z}/n\mathbb{Z})^*$ has a subgroup H such that $G \cong (\mathbb{Z}/n\mathbb{Z})^*/H$.

—Since $(\mathbb{Z}/n\mathbb{Z})^* \cong \mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\zeta_n))$ (Proposition 5.16) and cyclotomic fields are Galois over $\mathbb{Q}$, H corresponds to an intermediate field F, $\mathbb{Q} \subseteq F \subseteq \mathbb{Q}(\zeta_n)$.

—Since $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\zeta_n))$ is abelian, H is automatically normal; hence, $\mathbb{Q} \subseteq F$ is Galois, and $\mathrm{Aut}_{\mathbb{Q}}(F) \cong G$, as needed.

Thus, the 'inverse Galois problem' has an easy solution for *abelian groups.*

A much stronger result is true: it can be proven that *every* finite abelian extension of $\mathbb{Q}$ *is contained in some cyclotomic field.* This is the *Kronecker-Weber theorem,* dating back to the second half of the nineteenth century. The natural context for this result is *class field theory* and is well beyond the scope of this book. The reader will be able to verify (Exercise 7.19) that every *quadratic* extension of $\mathbb{Q}$ is contained in some cyclotomic field.

Exercises

7.1. Find explicitly a generator of a quadratic intermediate field of the extension $\mathbb{Q} \subseteq \mathbb{Q}(\zeta_{10})$.

7.2. ▷ Let R be an integral domain, and let $f(x) \in R[x]$ be a polynomial of degree n. Show how to obtain $f(x)$ by specializing the 'universal' polynomial $P_n(x)$ defined in §7.3. [§7.3]

7.3. Prove that the elementary symmetric functions $s_1, \ldots, s_n$ (see §7.3) are algebraically independent. (Hint: Use Exercise 1.28.)

7.4. ▷ Prove that every finite group is isomorphic to the group of automorphisms of some Galois extension. [§7.3]

7.5. ▷ Let $k \subseteq F$ be a radical extension, and let $k \subseteq K$ be any extension; assume F and K are contained in a larger field. Prove that $K \subseteq FK$ is radical. [§7.4]

7.6. Let $f(x) \in \mathbb{Q}[x]$ be an irreducible cubic, and let ρ be a real root of $f(x)$. Prove that the splitting field of $f(x)$ over $\mathbb{Q}$ is $\mathbb{Q}(\rho, \sqrt{D})$, where D is the discriminant of $f(x)$.

7.7. Let $f(x) \in \mathbb{Q}[x]$ be an irreducible cubic with exactly one real root. Prove that the discriminant of $f(x)$ is not a square in $\mathbb{Q}$.

7.8. ▷ (Cf. Exercise IV.4.12.) Find (mathematically or by library search) the lattice of subgroups of S_4, and verify that the transitive subgroups of S_4 are (isomorphic to) S_4, A_4, D_8, $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, and $\mathbb{Z}/4\mathbb{Z}$. [§7.5]

7.9. Compute the Galois group of the polynomial $x^4 - 2$.

7.10. ▷ Prove that the polynomial $x^5 - 5x - 1$ has exactly 3 real roots (this is a calculus exercise!) and is irreducible over $\mathbb{Q}$. Prove that its Galois group is S_5. [§7.5]

7.11. ▷ Let $n > 0$ be an integer. Note that

$$f_n(x) := (x^2 + 2) \cdot x \cdot (x - 2) \cdots (x - 2(n-2)) \cdot (x - 2(n-3))$$

has $n - 2$ rational roots and 2 nonreal, complex roots. Prove that for infinitely many integers q, the polynomial

$$q f_n(x) + 2 \in \mathbb{Z}[x]$$

has $(n-2)$ real roots, 2 nonreal complex roots, and is irreducible over $\mathbb{Q}$. Conclude that for each prime p there are infinitely many polynomials of degree p in $\mathbb{Z}[x]$ whose Galois group is S_p. [§7.3, §7.5]

7.12. ¬ Let $f(x) \in k[x]$ be a separable irreducible polynomial of prime degree p over a field k, and let $\alpha_1, \ldots, \alpha_p$ be the roots of $f(x)$ in its splitting field F. Prove that the Galois group of $f(x)$ contains an element σ of order p, 'cycling' through the roots. [7.13]

7.13. ¬ Let $f(x) \in k[x]$ be a separable irreducible polynomial of prime degree p over a field k. Let α be a root of $f(x)$ in $\overline{k}$, and suppose you can express *another* root of $f(x)$ as a polynomial in α, with coefficients in k. Prove that you can express *all* roots as polynomials in α and that the Galois group of $f(x)$ is $\mathbb{Z}/p\mathbb{Z}$. (Use Exercise 7.12.) [7.14]

7.14. ¬ Let k be a field of characteristic $p > 0$, and let $f(x) = x^p - x - a \in k[x]$. Assume that $f(x)$ has no roots in k. Prove that $f(x)$ is irreducible and that its Galois group is $\mathbb{Z}/p\mathbb{Z}$. (Note that if α is a root of $f(x)$, so is $\alpha + 1$; use Exercises 1.8 and 7.13.)

The splitting field of $f(x)$ is an *Artin-Schreier extension* of $\mathbb{F}_p$. [7.16]

7.15. ¬ Let $k \subseteq F$ be a cyclic Galois extension of degree p, where $\operatorname{char} k = p > 0$. Prove that it is an Artin-Schreier extension. (Hint: What is $\operatorname{tr}_{k \subseteq F}(1)$? Use the additive version of Hilbert's theorem 90, Exercise 6.19.) [6.19]

7.16. ▷ The *Artin-Schreier map* on a field k of positive characteristic p is the function

$$x \mapsto x^p - x.$$

Denote this function by AS, and let $\sqrt[AS]{\cdot}$ denote its inverse (which is only defined up to the addition of an element of $\mathbb{F}_p$: see Exercise 7.14).

Express the solutions of a quadratic equation $x^2 + bx + c = 0$ *in characteristic* 2 in terms of $\sqrt[AS]{\cdot}$, for $b \neq 0$. (For $b \neq 0$, the roots must live in an Artin-Schreier extension of k, since the polynomial is then separable. It is inseparable for $b = 0$; solving the equation in this case amounts to extending k by the unique square root of c.) [§7.4]

7.17. ¬ Let $f(x) \in k[x]$ be a separable irreducible polynomial of degree n over a field k, and let F be its splitting field. Assume $\operatorname{Aut}_k(F) \cong S_n$, and let α be a root of $f(x)$ in F.

- Prove that $\operatorname{Aut}_{k(\alpha)}(F) \cong S_{n-1}$.
- Prove that there are no proper subfields of $k(\alpha)$ properly containing k. (Hint: Exercise IV.4.8.)

[7.18]

7.18. Let $f(x) \in k[x]$ be a separable irreducible polynomial of degree n over a field k, with Galois group S_n. Let α be a root of $f(x)$ in $\overline{k}$, and let $g(x) \in k[x]$ be any nonconstant polynomial of degree $< n$. Prove that α may be expressed as a polynomial in $g(\alpha)$, with coefficients in k. (Use Exercise 7.17.)

7.19. ▷ Let d be a positive integer.

- Prove that the discriminant of a polynomial $f(x) = \prod_{i=1}^{d}(x - \alpha_i)$ equals the product $\pm \prod_{i=1}^{d} f'(\alpha_i)$.
- Prove that the discriminant of $x^d - 1$ is $\pm d^d$, and note that this is a square in $\mathbb{Q}(\zeta_d)$.
- Prove that $\mathbb{Q}(\zeta_{8d})$ contains square roots of d and $-d$.
- Conclude that every *quadratic* extension of $\mathbb{Q}$ may be embedded in a cyclotomic field $\mathbb{Q}(\zeta_n)$, for a suitable n.

(This is a very special case of the Kronecker-Weber theorem.) [§7.6]

Linear algebra, reprise

We come back to linear algebra, with the task of studying slightly more sophisticated constructions than those reviewed in Chapter VI. Our main goals are to discuss some aspects of multilinear algebra, tensor products, and Hom, with special attention devoted to dual modules. Along the way we will meet several other interesting notions, including a first taste of projective and injective modules and of Tor and Ext.

As in Chapter VI, we will attempt to deal with the more general case of R-modules, specializing to vector spaces only when this provides a drastic simplification or conforms to inveterate habits. Covering the material at this level of generality gives us an excellent excuse to complement the preliminary notions described in the very first chapter of this book.

1. Preliminaries, reprise

1.1. Functors. In this book we have given an unusually early introduction to the notion of *category*, motivated by the near omnipresence of universal problems at the very foundation of algebra. The hope was that the unifying viewpoint offered by categories would help the reader in the task of organizing the information carried by the basic constructions we have encountered.

However, so far we have never seriously had to 'go from one category to another', and therefore we have not felt compelled to complete the elementary presentation of categories by discussing more formally how this is done. The careful reader has likely seen it happen between the lines, at least in noninteresting ways: for example, if we strip a ring of its multiplicative structure, we are left with an abelian group; and if a field F is an extension of a field k, then we may view F as a k-vector space: thus, there are evident ways to go from Ring to Ab or from the category Fld_k of extensions of k to k-Vect. More interesting examples are the *group of units* of a ring, the homology of a complex, the automorphism group of a Galois field extension

(and note that this latter 'turns arrows around'; see below), etc. Implicitly, we have in fact encountered many instances of such 'functions between categories'.

The notion of *functor* formalizes these operations. A functor between two categories C, D will 'send objects of C to objects of D', and since categories carry the information of *morphisms*, functors will have to deal with this information as well.

Definition 1.1. Let C, D be two categories. A *covariant functor*

$$\mathscr{F} : \mathsf{C} \to \mathsf{D}$$

is an assignment of an object $\mathscr{F}(A) \in \mathrm{Obj}(\mathsf{D})$ for every $A \in \mathrm{Obj}(\mathsf{C})$ and of a function

$$\mathrm{Hom}_{\mathsf{C}}(A, B) \to \mathrm{Hom}_{\mathsf{D}}(\mathscr{F}(A), \mathscr{F}(B))$$

for every pair of objects A, B in C; this function is also denoted[1] $\mathscr{F}$ and must preserve identities and compositions. That is,

$$\mathscr{F}(1_A) = 1_{\mathscr{F}(A)}$$

$\forall A \in \mathrm{Obj}(\mathsf{C})$, and

$$\mathscr{F}(\beta \circ \alpha) = \mathscr{F}(\beta) \circ \mathscr{F}(\alpha)$$

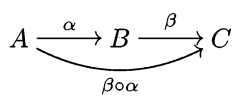
$\forall A, B, C \in \mathrm{Obj}(\mathsf{C})$, $\forall \alpha \in \mathrm{Hom}_{\mathsf{C}}(A, B)$, $\forall \beta \in \mathrm{Hom}_{\mathsf{C}}(B, C)$.

A *contravariant functor* $\mathsf{C} \to \mathsf{D}$ is a covariant functor $\mathsf{C}^{op} \to \mathsf{D}$ from the opposite category. ⌟

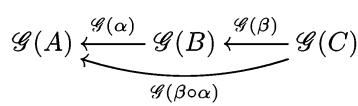
The opposite category C^{op} is obtained from C by 'reversing the arrows'; cf. Exercise I.3.1. The fact that *contravariant* functors $\mathscr{G} : \mathsf{C} \to \mathsf{D}$ (that is, covariant functors $\mathsf{C}^{op} \to \mathsf{D}$) preserve compositions means that $\forall A, B, C \in \mathrm{Obj}(\mathsf{C})$, $\forall \alpha \in \mathrm{Hom}_{\mathsf{C}}(A, B)$, $\forall \beta \in \mathrm{Hom}_{\mathsf{C}}(B, C)$:

$$\mathscr{G}(\beta \circ \alpha) = \mathscr{G}(\alpha) \circ \mathscr{G}(\beta)$$

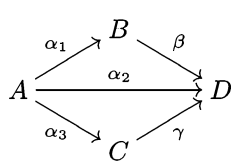
Pictorially,



is sent to

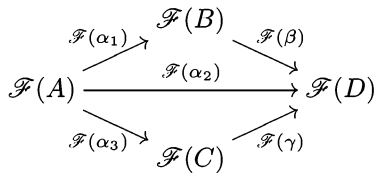


by a *contra*variant functor $\mathscr{G}$. Note the switch in the order of composition. A *co*variant functor 'preserves' arrows, in the sense that every diagram in C, say

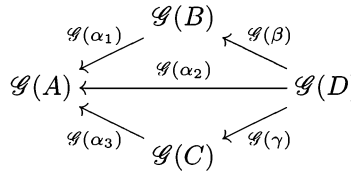


[1] This is a little unfortunate, since the function depends on A and B, but in practice it does not lead to confusion. Some authors prefer $\mathscr{F}_{A,B}$, for clarity.

is mapped by a covariant functor $\mathscr{F} : \mathsf{C} \to \mathsf{D}$ to a diagram with arrows in like directions:



A *contra*variant functor $\mathscr{G} : \mathsf{C} \to \mathsf{D}$ 'reverses' the arrows:



In both cases, *commutative* diagrams are sent to *commutative* diagrams.

In many contexts, contravariant functors are called *presheaves*. Thus, a *presheaf of sets on a category* C is simply a contravariant functor $\mathsf{C} \to \mathsf{Set}$.

If $\mathrm{Hom}_{\mathsf{C}}(A, B)$ carries more structure, we may require functors to preserve this structure. For example, we have seen that for $\mathsf{C} = R\text{-}\mathsf{Mod}$, the category of left-modules over a ring R, $\mathrm{Hom}_{R\text{-}\mathsf{Mod}}(A, B)$ is itself an abelian group (end of §III.5.2); a functor $\mathscr{F} : R\text{-}\mathsf{Mod} \to S\text{-}\mathsf{Mod}$ is *additive* if it preserves this structure, that is, if the corresponding function

$$\mathrm{Hom}_{R\text{-}\mathsf{Mod}}(A, B) \to \mathrm{Hom}_{S\text{-}\mathsf{Mod}}(\mathscr{F}(A), \mathscr{F}(B))$$

is a homomorphism of abelian groups (for all R-modules A, B).

1.2. Examples of functors. In practice, an assignment of objects of a category for objects of another category often comes with a (psychologically) 'natural' companion assignment for morphisms, which is evidently covariant or contravariant. We say that this assignment is 'functorial' if it preserves compositions.

All the simple-minded operations mentioned at the beginning of §1.1 (such as viewing a ring as an abelian group under addition) are covariant functors: for example, the fact that a homomorphism of *rings* is in particular a homomorphism of the underlying *abelian groups* amounts to the statement that labeling each ring with its underlying abelian group defines a covariant functor $\mathsf{Ring} \to \mathsf{Ab}$. Since such functors are obtained by 'forgetting' part of the structure of a given object, they are memorably called *forgetful functors*.

We have already mentioned briefly a few slightly more imaginative examples; here they are again.

Example 1.2. If R is a ring, we have denoted by R^* the *group* of units in R; every ring homomorphism $R \to S$ induces a group homomorphism $R^* \to S^*$, and this assignment is compatible with compositions; therefore this operation defines a covariant functor $\mathsf{Ring} \to \mathsf{Grp}$. ⌟

Example 1.3. The operation $\mathrm{Aut}_k(\_)$ from Galois field extensions to groups is contravariantly functorial. Indeed, if $k \subseteq E \subseteq F$ is viewed as a morphism of two Galois extensions ($k \subseteq E$ to $k \subseteq F$), we have a corresponding group homomorphism

$$\mathrm{Aut}_k(F) \to \mathrm{Aut}_k(E)$$

defined by restriction: $\varphi \mapsto \varphi|_E$; the latter maps E to E by virtue of the Galois condition. The composition of two restrictions is a restriction, so the assignment is indeed functorial. ⌟

Example 1.4. In §III.4.3 we have defined the *spectrum* of a commutative ring R, $\mathrm{Spec}\, R$, as the *set* of prime ideals of R. If R, S are commutative rings and $\varphi : R \to S$ is a ring homomorphism, then the inverse image $\varphi^{-1}(\mathfrak{p})$ of a prime ideal $\mathfrak{p}$ is a prime ideal; thus, φ induces a set-function

$$\varphi^* : \mathrm{Spec}(S) \to \mathrm{Spec}(R)$$

This assignment is clearly functorial, so we can view Spec as a contravariant functor from the category of commutative rings to Set. The reader will assign a topology to $\mathrm{Spec}(R)$ (Exercise 1.7), making this example (even) more interesting. ⌟

Example 1.5. If S is a set, recall (Example I.3.4) that one can construct a category $\hat{\mathsf{S}}$ whose objects are subsets of S and where morphisms correspond to inclusions of subsets. A *presheaf of sets on* S is a contravariant functor $\hat{\mathsf{S}} \to \mathsf{Set}$. The prototypical example consists of the functor which associates to $U \subseteq S$ the set of functions from U to a fixed set; the inclusion $U \subseteq V$ is mapped to the restriction $\rho \mapsto \rho|_U$. More interesting (and very useful) examples may be obtained by considering richer structures. For instance, if T is a *topological space,* one can define a category whose objects are the open sets in T and whose morphisms are inclusions; this leads to the notion of presheaf on a topological space. Standard notions such as '(the groups of) continuous complex-valued functions on open sets of T' are most naturally viewed as the datum of a presheaf of abelian groups on T. Such concrete examples often satisfy additional hypotheses, which make them *sheaves,* but that is a topic for another book. In a sense, studying a branch of geometry (e.g., complex differential geometry) amounts to studying spaces endowed with a suitable sheaf (e.g., the sheaf of complex differentiable functions). ⌟

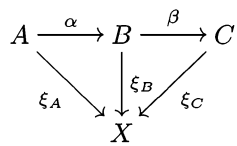
An important class of functors that are available on any category may be obtained as follows, and this will be a source of examples in the next few sections. Let C be a category, and let X be an object of C. Then the assignments

$$A \mapsto \mathrm{Hom}_{\mathsf{C}}(X, A) \quad , \quad A \mapsto \mathrm{Hom}_{\mathsf{C}}(A, X)$$

define, respectively, *covariant* and *contravariant* functors $\mathsf{C} \to \mathsf{Set}$. These are denoted $\mathrm{Hom}_{\mathsf{C}}(X, \_)$, $\mathrm{Hom}_{\mathsf{C}}(\_, X)$, respectively. For example, if

$$A \xrightarrow{\alpha} B \xrightarrow{\beta} C$$

is a diagram in C, stare at



Every $\alpha : A \to B$ determines a function

$$\mathrm{Hom}_{\mathsf{C}}(B, X) \to \mathrm{Hom}_{\mathsf{C}}(A, X)$$

defined by mapping ξ_B to $\xi_A := \xi_B \circ \alpha$, that is, by requiring the triangle on the left to be commutative. The associativity of composition,

$$\xi_C \circ (\beta \circ \alpha) = (\xi_C \circ \beta) \circ \alpha \quad ,$$

expresses precisely the contravariant functoriality of this assignment. The functor $\mathrm{Hom}_{\mathsf{C}}(\_, X)$ is often denoted h_X for short, if C is understood.

The brave reader may contemplate the fact that this technique associates to each *object* X of a category C a *presheaf of sets* h_X on C. This raises natural and important questions, such as how to tell whether a given presheaf of sets $\mathscr{F}$ in fact equals h_X for some object X of C (this makes $\mathscr{F}$ 'representable') or whether an object X may be reconstructed from the corresponding presheaf h_X. For example, if Top is the category of topological spaces and p denotes the final object of Top (that is, the one-point topological space), then $h_X(p)$ is nothing but X, viewed as a set. For this reason, in such 'geometric' contexts (for example, in algebraic geometry) $h_X(A)$ is called the 'set of A-points of X'; h_X is the *functor of points* of X.

1.3. When are two categories 'equivalent'? Essentially every notion of 'isomorphism' encountered so far has boiled down to 'structure-preserving bijection', and the reader may well expect that the natural notion identifying two categories should be drawn from the same model: a functor matching objects of one category precisely with objects of another, preserving the structure of morphisms.

One could certainly introduce such a notion, but it would be exceedingly restrictive. Recall that solutions to universal problems, that is, just about every construction we have run across, are only *defined up to isomorphism* (Proposition I.5.4). Requiring solutions in one context to match exactly with solutions in another would be problematic. The structure of an object in a category is adequately carried by its isomorphism class[2], and a natural notion of 'equivalence' of categories should aim at matching isomorphism classes, rather than individual objects. The *morphisms* are a more essential piece of information; the quality of a functor is first of all measured on how it acts on morphisms.

[2]One should not take this viewpoint too far. It is tempting to think of isomorphic objects in a category as 'the same', but this is also problematic and does not lead (as far as we know) to a workable alternative. For example, some objects in a category may have 'more automorphisms' than others, and this information would be discarded if we simply chopped up the category into isomorphism classes, draconially promoting all isomorphisms to identity morphisms.

Definition 1.6. Let C, D be two categories. A covariant functor $\mathscr{F} : \mathsf{C} \to \mathsf{D}$ is *faithful* if for all objects A, B of C, the induced function

$$\mathrm{Hom}_{\mathsf{C}}(A, B) \to \mathrm{Hom}_{\mathsf{D}}(\mathscr{F}(A), \mathscr{F}(B))$$

is injective; it is *full* if this function is surjective, for all objects A, B. ⌟

'Fully faithful' functors $\mathscr{F} : \mathsf{C} \to \mathsf{D}$ are bijective on morphisms, and it follows easily that they preserve isomorphism classes: $A \cong B$ in C if and only if $\mathscr{F}(A) \cong \mathscr{F}(B)$ in D (Exercise 1.2). For all intents and purposes, a fully faithful functor $\mathscr{F} : \mathsf{C} \to \mathsf{D}$ identifies C with a full[3] subcategory of D, even if 'on objects' $\mathscr{F}$ may be neither injective nor surjective onto that subcategory. This idea is captured by the following definition:

Definition 1.7. A covariant functor $\mathscr{F} : \mathsf{C} \to \mathsf{D}$ is an *equivalence of categories* if it is fully faithful and 'essentially surjective'; that is, $\mathscr{F}$ induces bijections on Hom-sets, and for every object Y of D there is an object X of C such that $\mathscr{F}(X) \cong Y$ in D. ⌟

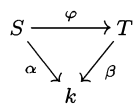
Two categories C, D are *equivalent* if there is a functor $\mathscr{F} : \mathsf{C} \to \mathsf{D}$ satisfying the requirements in Definition 1.7. It is not at all immediate that this is indeed an equivalence relation (why is it symmetric?), but this turns out to be the case. In fact, an equivalence of categories $\mathscr{F}$ has an 'inverse', in a suitably weakened sense: a functor $\mathscr{G} : \mathsf{D} \to \mathsf{C}$ such that there are 'natural isomorphisms' (see §1.5) from $\mathscr{F} \circ \mathscr{G}$ and $\mathscr{G} \circ \mathscr{F}$ to the identity. Since we will not use these notions too extensively, we will let the interested reader research the issue and find proofs.

Example 1.8. (Cf. Exercise I.3.6.) Construct a category by taking the objects to be nonnegative integers and $\mathrm{Hom}(m, n)$ to be the set of $n \times m$ matrices with entries in a field k, with composition defined by product of matrices (and suitable care concerning matrices with no rows or columns). The resulting category is equivalent to the category of finite-dimensional k-vector spaces. Indeed, we obtain a functor from the former to the latter by sending n to the vector space k^n (endowed with the standard basis) and each matrix to the corresponding linear map. This functor is clearly fully faithful, and it is essentially surjective because vector spaces are classified by their dimension. This example makes (more) precise the heuristic considerations at the end of §VI.2.1. ⌟

Example 1.9. Recall (Definition VII.2.19) that the *coordinate ring* of an affine algebraic set over a field K is a reduced, commutative, finite-type K-algebra. We can define the category K-Aff of affine K-algebraic sets by prescribing that the objects be algebraic subsets of some affine K-space and defining[4] $\mathrm{Hom}_{K\text{-}\mathsf{Aff}}(S, T)$ to be $\mathrm{Hom}_{K\text{-}\mathsf{Alg}}(K[T], K[S])$.

[3] See Exercise I.3.8 for the definition of 'full' subcategory.

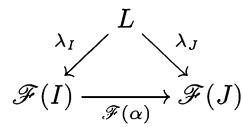
[4] The switch in the order of S and T is reasonable: as the reader has verified in Exercise VII.2.12, $K[S]$ may be interpreted as the K-algebra of 'polynomial functions' on S; any function $\varphi : S \to T$ determines a pull-back of functions, $K[T] \to K[S]$:



Thus, K-Aff is defined in such a way that the functor $K\text{-Aff}^\circ \to K\text{-Alg}$ that maps an affine algebraic set S to its coordinate ring $K[S]$ is an equivalence of the *opposite* category of K-Aff with the subcategory of reduced, commutative, finite-type K-algebra.

1.4. Limits and colimits. The various universal properties encountered along the way in this book are all particular cases of the notion of categorical *limit*, which is worth mentioning explicitly. Let $\mathscr{F} : \mathsf{I} \to \mathsf{C}$ be a covariant functor, where one thinks of I as a category 'of indices'. The *limit* of $\mathscr{F}$ is (if it exists) an object L of C, endowed with morphisms $\lambda_I : L \to \mathscr{F}(I)$ for all objects I of I, satisfying the following properties.

- If $\alpha : I \to J$ is a morphism in I, then $\lambda_J = \mathscr{F}(\alpha) \circ \lambda_I$:



- L is final with respect to this property: that is, if M is another object, endowed with morphisms μ_I, also satisfying the previous requirement, then there exists a unique morphism $M \to L$ making all relevant diagrams commute[5].

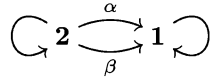
The limit L (if it exists) is unique up to isomorphism, as is every notion defined by a universal property. It is denoted $\varprojlim \mathscr{F}$. The 'left-pointing' arrow reminds us that L stands 'before' all objects of C indexed by I via $\mathscr{F}$ (but it is—up to isomorphisms—the 'last' object that may do so). This notion is also called *inverse*, or *projective*, limit.

Example 1.10 (Products)**.** Let I be the 'discrete category' consisting of two objects $\mathbf{1}$, $\mathbf{2}$, with only identity morphisms[6], and let $\mathscr{A}$ be a functor from I to any category C; let $A_1 = \mathscr{A}(\mathbf{1})$, $A_2 = \mathscr{A}(\mathbf{2})$ be the two objects of C 'indexed' by I. Then $\varprojlim \mathscr{A}$ is nothing but the product of A_1 and A_2 in C, as defined in §I.5.4: a limit exists if and only if a product of A_1 and A_2 exists in C.

We can similarly define the product of any (possibly infinite) family of objects in a category as the limit over the corresponding discrete indexing category, provided of course that this limit exists. ⌟

The limit notion is a little more interesting if the indexing category I carries more structure.

Example 1.11 (Equalizers and kernels)**.** Let I again be a category with two objects $\mathbf{1}$, $\mathbf{2}$, but assume that morphisms look like this:



$\beta \mapsto \alpha := \beta \circ \varphi$. The definition of $\mathrm{Hom}_{K\text{-Aff}}(S,T)$ is concocted so as to ensure that this pull-back operation is a homomorphism of K-algebras.

[5]Any such datum of M and morphisms is called a *cone* for $\mathscr{F}$. Therefore, the limit of $\mathscr{F}$ is a 'universal cone': it is a final object in the (easily defined) category of cones of $\mathscr{F}$.

[6]This latter prescription is what makes the category 'discrete'; cf. Example I.3.3.

That is, add to the discrete category two 'parallel' morphisms α, β from one of the objects to the other. A functor $\mathscr{K} : \mathsf{I} \to \mathsf{C}$ amounts to the choice of two objects A_1, A_2 in C and two parallel morphisms between them. Limits of such functors are called *equalizers*. For a concrete example, assume $\mathsf{C} = R\text{-}\mathsf{Mod}$ is the category of R-modules for some ring R; let $\varphi : A_2 \to A_1$ be a homomorphism, and choose $\mathscr{K}$ as above, with $\mathscr{K}(\alpha) = \varphi$ and $\mathscr{K}(\beta) =$ the zero-morphism. Then $\varprojlim \mathscr{K}$ is nothing but the *kernel* of φ, as the reader will verify (Exercise 1.14). ⌟

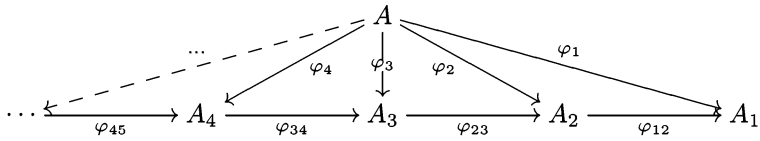
Example 1.12 (Limits over chains)**.** In another typical situation, I may consist of a totally ordered set, for example:

$$\cdots \longrightarrow \mathbf{4} \longrightarrow \mathbf{3} \longrightarrow \mathbf{2} \longrightarrow \mathbf{1}$$

(that is, the objects are $\mathbf{i}$, for all positive integers i, and there is a unique morphism $\mathbf{i} \to \mathbf{j}$ whenever $i \geq j$; we are only drawing the morphisms $\mathbf{j+1} \to \mathbf{j}$). Choosing $\mathscr{F} : \mathsf{I} \to \mathsf{C}$ is equivalent to choosing objects A_i of C for all positive integers i and morphisms $\varphi_{ji} : A_i \to A_j$ for all $i \geq j$, with the requirement that $\varphi_{ii} = 1_{A_i}$, and $\varphi_{kj} \circ \varphi_{ji} = \varphi_{ki}$ for all $i \geq j \geq k$. That is, the choice of $\mathscr{F}$ amounts to the choice of a diagram

$$\cdots \xrightarrow{\varphi_{45}} A_4 \xrightarrow{\varphi_{34}} A_3 \xrightarrow{\varphi_{23}} A_2 \xrightarrow{\varphi_{12}} A_1$$

in C. An inverse limit $\varprojlim \mathscr{F}$ (which may also be denoted $\varprojlim_i A_i$, when the morphisms φ_{ji} are evident from the context) is then an object A endowed with morphisms $\varphi_i : A \to A_i$ such that the whole diagram



commutes and such that any other object satisfying this requirement factors uniquely through A.

Such limits exist in many standard situations. For example, let $\mathsf{C} = R\text{-}\mathsf{Mod}$ be the category of left-modules over a fixed ring R, and let A_i, φ_{ji} be as above.

Claim 1.13. *The limit* $\varprojlim_i A_i$ *exists in* R-Mod.

Proof. The product $\prod_i A_i$ consists of arbitrary sequences $(a_i)_{i>0}$ of elements $a_i \in A_i$. Say that a sequence $(a_i)_{i>0}$ is *coherent* if for all $i > 0$ we have $a_i = \varphi_{i,i+1}(a_{i+1})$. Coherent sequences form an R-submodule A of $\prod_i A_i$; the canonical projections restrict to R-module homomorphisms $\varphi_i : A \to A_i$. The reader will check that A is a limit $\varprojlim_i A_i$ (Exercise 1.15). □

This example easily generalizes to families indexed by more general posets. ⌟

The 'dual notion' to limit is the *colimit* of a functor $\mathscr{F} : \mathsf{I} \to \mathsf{C}$. The colimit is an object C of C, endowed with morphisms $\gamma_i : \mathscr{F}(I) \to C$ for all objects I of $\mathscr{I}$, such that $\gamma_I = \gamma_J \circ \mathscr{F}(\alpha)$ for all $\alpha : I \to J$ and that C is *initial* with respect to this requirement.

Again, we have already encountered several instances of this construction: for example, the *coproduct* of two objects A_1, A_2 of a category is (if it exists) the colimit over the discrete category with two objects, just as their product is the corresponding limit. *Cokernels* are colimits, just as kernels are limits (cf. Example 1.11).

The colimit of $\mathscr{F}$ is (not surprisingly) denoted $\varinjlim \mathscr{F}$ and is called the *direct*, or *injective*, limit of $\mathscr{F}$.

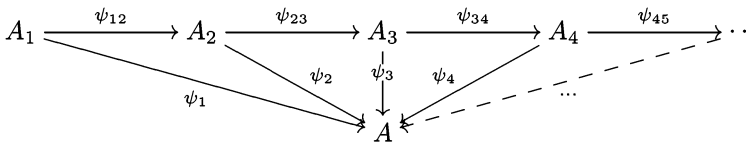
For a typical situation consider again the case of a totally ordered set I, for example:

$$\mathbf{1} \longrightarrow \mathbf{2} \longrightarrow \mathbf{3} \longrightarrow \mathbf{4} \longrightarrow \cdots .$$

A functor $\mathscr{F} : \mathsf{I} \to \mathsf{C}$ consists of the choice of objects and morphisms

$$A_1 \xrightarrow{\psi_{12}} A_2 \xrightarrow{\psi_{23}} A_3 \xrightarrow{\psi_{34}} A_4 \xrightarrow{\psi_{45}} \cdots$$

and the direct limit $\varinjlim_i A_i$ will be an object A with morphisms $\psi_i : A_i \to A$ such that the diagram



commutes and such that A is initial with respect to this requirement.

Example 1.14. If $\mathsf{C} = \mathsf{Set}$ and all the ψ_{ij} are injective, we are talking about a 'nested sequence of sets':

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq A_4 \subseteq \cdots \quad ;$$

the direct limit of this sequence would be the 'infinite union' $\bigcup_i A_i$. ⌟

We have shamelessly relied on the reader's intuition and used this notion already, for example when we constructed the algebraic closure of a field (in §VII.2.1) or whenever we have mentioned polynomial rings with 'infinitely many variables' (e.g., in Example III.6.5). More formally, $\bigcup_i A_i$ consists of equivalence classes of pairs (i, a_i), where $a_i \in A_i$ and (i, a_i) is equivalent to (j, a_j) for $i \leq j$ if $a_j = \psi_{ij}(a_i)$.

In the context of R-modules, one can consider the direct sum $D = \bigoplus_i A_i$ and the submodule $K \subseteq D$ generated by elements of the form $a_j - \psi_{ij}(a_i)$ for all $i \leq j$ (where each A_i is identified with its image in D). Then the quotient D/K satisfies the needed universal property, as the reader will check. In fact, it is no harder to perform these constructions for more general posets (Exercise 1.16). They are somewhat more explicit and better behaved in the case of *directed sets:* partially ordered sets I such that $\forall i, j \in \mathsf{I}$ there exists $k \in \mathsf{I}$ such that $i \leq k$, $j \leq k$.

Several standard constructions rely on a direct limit: for example, *germs* of functions (or, more generally, 'stalks of presheaves') are defined by means of a direct limit.

1.5. Comparing functors. Having introduced functors as the natural notion of 'morphisms between categories', the next natural step is to consider 'morphisms between functors' and other ways to compare two given functors. A detailed description of these notions is not needed in this book, but we want to mention briefly a few concepts and essential remarks, again because they offer a unifying viewpoint.

Definition 1.15. Let C, D be categories, and let $\mathscr{F}$, $\mathscr{G}$ be (say, covariant) functors $\mathsf{C} \to \mathsf{D}$. A *natural transformation*[7] $\mathscr{F} \rightsquigarrow \mathscr{G}$ is the datum of a morphism $\nu_X : \mathscr{F}(X) \to \mathscr{G}(X)$ in D for every object X in C, such that $\forall \alpha : X \to Y$ in C the diagram

$$\begin{array}{ccc} \mathscr{F}(X) & \xrightarrow{\mathscr{F}(\alpha)} & \mathscr{F}(Y) \\ {\scriptstyle \nu_X}\downarrow & & \downarrow{\scriptstyle \nu_Y} \\ \mathscr{G}(X) & \xrightarrow{\mathscr{G}(\alpha)} & \mathscr{G}(Y) \end{array}$$

commutes. A *natural isomorphism* is a natural transformation ν such that ν_X is an isomorphism for every X. ⌟

Natural transformations arise in many contexts, for example when comparing different notions of homology or cohomology in topology. The reader is likely to have run into the Hurewicz homomorphism (from π_1 to H_1); it is a natural transformation. In fact, any statement such as *'there is a natural* (or *canonical*) *homomorphism...* ' is likely hiding a natural transformation between two functors. This technical meaning of the word 'natural' matches, as a rule, its psychological use.

A particularly basic and important example is the notion of *adjoint functors*, which we will mention at a rather informal level, leaving to the inextinguishable reader the task of making it formal by suitable use of natural transformations. Let C, D be categories, and let $\mathscr{F} : \mathsf{C} \to \mathsf{D}$, $\mathscr{G} : \mathsf{D} \to \mathsf{C}$ be functors. We say that $\mathscr{F}$ and $\mathscr{G}$ are *adjoint* (and we say that $\mathscr{G}$ is right-adjoint to $\mathscr{F}$ and $\mathscr{F}$ is left-adjoint to $\mathscr{G}$) if there are *natural* isomorphisms

$$\mathrm{Hom}_{\mathsf{C}}(X, \mathscr{G}(Y)) \xrightarrow{\sim} \mathrm{Hom}_{\mathsf{D}}(\mathscr{F}(X), Y)$$

for all objects X of C and Y of D. (More precisely, there should be a natural isomorphism of 'bifunctors' $\mathsf{C} \times \mathsf{D} \to \mathsf{Set}$: $\mathrm{Hom}_{\mathsf{C}}(\_, \mathscr{G}(\_)) \tilde{\rightsquigarrow} \mathrm{Hom}_{\mathsf{D}}(\mathscr{F}(\_), \_)$.)

Once again, several constructions we have encountered along the way may be recast in these terms, and we will run into more instances of this situation in this chapter.

Example 1.16. The construction of the free group on a given set (§II.5) is concocted so that giving a set-function from a set A to a group G is 'the same as' giving a group homomorphism from $F(A)$ to G (§II.5.2). What this really means is that for all sets A and all groups G there are natural identifications

$$\mathrm{Hom}_{\mathsf{Set}}(A, S(G)) \cong \mathrm{Hom}_{\mathsf{Grp}}(F(A), G) \quad ,$$

[7] The symbol $\Rightarrow$ is more standard than $\rightsquigarrow$, but it looks a little too much like the logical connective $\Longrightarrow$ for our taste.

where $S(G)$ 'forgets' the group structure of G. That is, the functor $F : \mathsf{Set} \to \mathsf{Grp}$ constructing free groups is left-adjoint to the forgetful functor $S : \mathsf{Grp} \to \mathsf{Set}$.

This of course applies to every other construction of 'free' objects we have encountered: the free functor is, as a rule, left-adjoint to the forgetful functor. ⌟

Thus, interesting functors may turn out to be adjoints of harmless-looking functors. This has technical advantages: properties of the interesting ones may be translated into properties of the harmless ones, thereby giving easier proofs of these properties.

In fact, the very fact that a functor has an adjoint will endow that functor with convenient features. We say that $\mathscr{F}$ *is a left-adjoint functor* if it has a right-adjoint, and that $\mathscr{G}$ is a right-adjoint functor if it has a left-adjoint. Some properties of (say) right-adjoint functors may be established without even knowing what the companion left-adjoint functor may be. Here is the prototypical example of this phenomenon.

Lemma 1.17. *Right-adjoint functors commute with limits.*

That is, if $\mathscr{G} : \mathsf{D} \to \mathsf{C}$ has a left-adjoint $\mathscr{F} : \mathsf{C} \to \mathsf{D}$ and $\mathscr{A} : \mathsf{I} \to \mathsf{D}$ is another functor, then there is a canonical isomorphism

$$\mathscr{G}(\varprojlim \mathscr{A}) \xrightarrow{\sim} \varprojlim(\mathscr{G} \circ \mathscr{A})$$

(*if* the limits exist, of course). As every good calculus student should readily understand, this says that right-adjoint functors are *continuous.* (Needless to say, left-adjoint functors turn out to commute with *co*limits and would rightly be termed *cocontinuous.*)

We will not prove Lemma 1.17 in gory detail, but we will endeavor to convince the reader that such statements are easier than they look and just boil down to suitable applications of the universal properties defining the various concepts. By contrast, proving that a specific given functor preserves products (for instance), without appealing to abstract nonsense, may sometimes appear to involve some 'real' work.

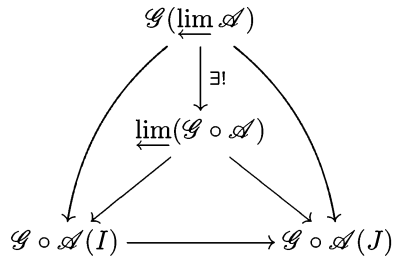
Assume $\mathscr{G} : \mathsf{D} \to \mathsf{C}$ is right-adjoint to $\mathscr{F} : \mathsf{C} \to \mathsf{D}$ and $\mathscr{A} : \mathsf{I} \to \mathsf{D}$ is a given functor. As we have seen, the limit of $\mathscr{A}$ is final subject to fitting in commutative diagrams

$$\begin{array}{ccc} & \varprojlim \mathscr{A} & \\ {\scriptstyle \lambda_I}\swarrow & & \searrow{\scriptstyle \lambda_J} \\ \mathscr{A}(I) & \xrightarrow[\mathscr{A}(\alpha)]{} & \mathscr{A}(J) \end{array}$$

(with hopefully evident notation). Applying $\mathscr{G}$, we get commutative diagrams (in C)

$$\begin{array}{ccc} & \mathscr{G}(\varprojlim \mathscr{A}) & \\ {\scriptstyle \mathscr{G}(\lambda_I)}\swarrow & & \searrow{\scriptstyle \mathscr{G}(\lambda_J)} \\ \mathscr{G} \circ \mathscr{A}(I) & \xrightarrow[\mathscr{G} \circ \mathscr{A}(\alpha)]{} & \mathscr{G} \circ \mathscr{A}(J) \end{array}$$

and hence, by the universal property defining the limit of $\mathscr{G} \circ \mathscr{A}$:



Now, the morphisms (in C)

$$\varprojlim(\mathscr{G} \circ \mathscr{A}) \to \mathscr{G} \circ \mathscr{A}(I)$$

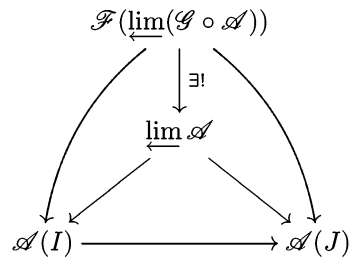
determine, via the adjunction identification

$$\mathrm{Hom}_{\mathsf{C}}(X, \mathscr{G}(Y)) \xrightarrow{\sim} \mathrm{Hom}_{\mathsf{D}}(\mathscr{F}(X), Y) \quad ,$$

morphisms (in D)

$$\mathscr{F}(\varprojlim(\mathscr{G} \circ \mathscr{A})) \to \mathscr{A}(I)$$

By the universal property defining $\varprojlim \mathscr{A}$ we get



Applying adjunction again, the new morphism

$$\mathscr{F}(\varprojlim(\mathscr{G} \circ \mathscr{A})) \to \varprojlim \mathscr{A}$$

determines a unique morphism

$$\varprojlim(\mathscr{G} \circ \mathscr{A}) \to \mathscr{G}(\varprojlim \mathscr{A})$$

Summarizing, we have obtained natural morphisms

$$\mathscr{G}(\varprojlim \mathscr{A}) \to \varprojlim(\mathscr{G} \circ \mathscr{A}) \quad , \quad \varprojlim(\mathscr{G} \circ \mathscr{A}) \to \mathscr{G}(\varprojlim \mathscr{A})$$

The compositions of these are easily checked to be the identity (by virtue of the uniqueness part of various universality properties invoked in the construction), concluding the verification of Lemma 1.17.

What is missing from this outline is the explicit verification of the fact that every needed diagram commutes, which is necessary in order to apply the universal properties as stated. The reader will likely agree that such verifications, while possibly somewhat involved, must be routine.

Lemma 1.17 implies, for example, that right-adjoint functors preserve products and that they must 'preserve kernels' when kernels make sense.

Example 1.18 (Exact functors). A functor is *exact* if it preserves exactness, that is, it sends exact sequences to exact sequences. So far we have only studied exactness in the context of modules over a ring (§III.7.1), so let R, S be rings and let $\mathscr{F} : R\text{-Mod} \to S\text{-Mod}$ be an additive functor $R\text{-Mod} \to S\text{-Mod}$; $\mathscr{F}$ is exact if and only if whenever

$$0 \longrightarrow A \xrightarrow{\varphi} B \xrightarrow{\psi} C \longrightarrow 0$$

is an exact sequence of R-modules, then

$$0 \longrightarrow \mathscr{F}(A) \xrightarrow{\mathscr{F}(\varphi)} \mathscr{F}(B) \xrightarrow{\mathscr{F}(\psi)} \mathscr{F}(C) \longrightarrow 0$$

is an exact sequence of S-modules. It follows easily that the image of every exact complex is exact (Exercise 1.23).

A common and very interesting situation occurs when a functor is *not* exact but preserves 'some' of the exactness of sequences (we will encounter such examples later in this chapter, for example in §2.3). We say that an additive functor $\mathscr{F}$ is *left-exact* if whenever

$$0 \longrightarrow A \xrightarrow{\varphi} B \xrightarrow{\psi} C$$

is exact, then so is

$$0 \longrightarrow \mathscr{F}(A) \xrightarrow{\mathscr{F}(\varphi)} \mathscr{F}(B) \xrightarrow{\mathscr{F}(\psi)} \mathscr{F}(C)$$

It turns out that a left-exact functor gives rise to a whole sequence of 'new', related (*'derived'*) functors; this hugely important phenomenon is studied in homological algebra, and we will come back to it in Chapter IX.

Claim 1.19. *Right-adjoint additive functors $R\text{-Mod} \to S\text{-Mod}$ are left-exact.*

As the reader will verify (Exercise 1.27), this follows from the fact that right-adjoint functors preserve kernels. (Note that the exactness of the sequence amounts to the fact that $\varphi : A \hookrightarrow B$ identifies A with $\ker \psi$.)

Of course the corresponding *co*-statements hold: since left-adjoint functors preserve colimits and cokernel are colimits, it follows that left-adjoint additive functors $\mathscr{G} : R\text{-Mod} \to S\text{-Mod}$ are necessarily *right*-exact, in the sense that if

$$A \xrightarrow{\varphi} B \xrightarrow{\psi} C \longrightarrow 0$$

is exact, then

$$\mathscr{G}(A) \xrightarrow{\mathscr{G}(\varphi)} \mathscr{G}(B) \xrightarrow{\mathscr{G}(\psi)} C \longrightarrow 0$$

is also exact. ⌟

Exercises

1.1. Let $\mathscr{F} : \mathsf{C} \to \mathsf{D}$ be a covariant functor, and assume that both C and D have products. Prove that for all objects A, B of C, there is a unique morphism $\mathscr{F}(A \times B) \to \mathscr{F}(A) \times \mathscr{F}(B)$ such that the relevant diagram involving natural projections commutes.

If D has *co*products (denoted $\amalg$) and $\mathscr{G} : \mathsf{C} \to \mathsf{D}$ is contravariant, prove that there is a unique morphism $\mathscr{G}(A) \amalg \mathscr{G}(B) \to \mathscr{G}(A \times B)$ (again, such that an appropriate diagram commutes).

1.2. ▷ Let $\mathscr{F} : \mathsf{C} \to \mathsf{D}$ be a fully faithful functor. If A, B are objects in C, prove that $A \cong B$ in C if and only if $\mathscr{F}(A) \cong \mathscr{F}(B)$ in D. [§1.3]

1.3. Recall (§II.1) that a group G may be thought of as a groupoid G with a single object. Prove that defining the action of G on an object of a category C is equivalent to defining a functor $\mathsf{G} \to \mathsf{C}$.

1.4. ¬ Let R be a commutative ring, and let $S \subseteq R$ be a *multiplicative subset* in the sense of Exercise V.4.7. Prove that 'localization is a functor': associating with every R-module M the localization $S^{-1}M$ (Exercise V.4.8) and with every R-module homomorphism $\varphi : M \to N$ the naturally induced homomorphism $S^{-1}M \to S^{-1}N$ defines a covariant functor from the category of R-modules to the category of $S^{-1}R$-modules. [1.25]

1.5. For F a field, denote by F^* the group of nonzero elements of F, with multiplication. The assignment $\mathsf{Fld} \to \mathsf{Grp}$ mapping F to F^* and a homomorphism of fields $\varphi : k \to F$ to the restriction $\varphi|_{k^*} : k^* \to F^*$ is clearly a covariant functor.

On the other hand, a homomorphism of fields[8] $k \to F$ is nothing but a field extension $k \subseteq F$. Prove that the assignment $F \mapsto F^*$ on objects, together with the prescription associating with every $k \subseteq F$ the *norm* $N_{k \subseteq F} : F^* \to k^*$ (cf. Exercise VII.1.12), gives a *contravariant* functor $\mathsf{Fld} \to \mathsf{Grp}$. State and prove an analogous statement for the *trace* (cf. Exercise VII.1.13).

1.6. ¬ Formalize the notion of presheaf of abelian groups on a topological space T. If $\mathscr{F}$ is a presheaf on T, elements of $\mathscr{F}(U)$ are called *sections* of $\mathscr{F}$ on U. The homomorphism $\rho_{UV} : \mathscr{F}(U) \to \mathscr{F}(V)$ induced by an inclusion $V \subseteq U$ is called the *restriction map.*

Show that an example of presheaf is obtained by letting $\mathscr{C}(U)$ be the additive abelian group of continuous complex-valued functions on U, with restriction of sections defined by ordinary restriction of functions.

For this presheaf, prove that one can uniquely glue sections agreeing on overlapping open sets. That is, if U and V are open sets and $s_U \in \mathscr{C}(U)$, $s_V \in \mathscr{C}(V)$ agree after restriction to $U \cap V$, prove that there exists a unique $s \in \mathscr{C}(U \cup V)$ such that s restricts to s_U on U and to s_V on V.

This is essentially the condition making $\mathscr{C}$ a *sheaf.* [IX.1.15]

[8] Recall that there are no 'zero-homomorphisms' between fields; cf. §VII.1.1.

1.7. ▷ Define a topology on $\operatorname{Spec} R$ by declaring the closed sets to be the sets $V(I)$, where $I \subseteq \operatorname{Spec} R$ is an ideal and $V(I)$ denotes the set of prime ideals *containing* I.

- Verify that this indeed defines a topology on $\operatorname{Spec} R$. (This is the *Zariski topology* on $\operatorname{Spec} R$.)
- Relate this topology to the Zariski topology defined in §VII.2.3.
- Prove that Spec is then a contravariant functor from the category of commutative rings to the category of topological spaces (where morphisms are continuous functions).

[§1.2]

1.8. Let K be an algebraically closed field, and consider the category K-Aff defined in Example 1.9.

- Denote by h_S the functor $\operatorname{Hom}_{K\text{-}\mathsf{Aff}}(\_, S)$ (as in §1.2), and let $p = \mathbb{A}^0_K$ be a point. Show that there is a natural bijection between S and $h_S(p)$. (Use Exercise VII.2.14.)
- Show how every $\varphi \in \operatorname{Hom}_{K\text{-}\mathsf{Aff}}(S, T)$ determines a function *of sets* $S \to T$.
- If $S \subseteq \mathbb{A}^m_K$, $T \subseteq \mathbb{A}^n_K$, show that the function $S \to T$ determined by a morphism $\varphi \in \operatorname{Hom}_{K\text{-}\mathsf{Aff}}(S, T)$ is the restriction of a 'polynomial function' $\mathbb{A}^m_K \to \mathbb{A}^n_K$. (Part of this exercise is to make sense of what this means!)

1.9. ¬ Let C, D be categories, and assume C to be small. Define a *functor category* D^{C}, whose objects are covariant functors $\mathsf{C} \to \mathsf{D}$ and whose morphisms are natural transformations[9].

Prove that the assignment $X \mapsto h_X := \operatorname{Hom}_{\mathsf{C}}(\_, X)$ (cf. §1.2) defines a covariant functor $\mathsf{C} \to \mathsf{Set}^{\mathsf{C}^\circ}$. (Define the action on morphisms in the natural way.) [1.11, IX.1.11]

1.10. ¬ Let C be a category, X an object of C, and consider the contravariant functor $h_X := \operatorname{Hom}_{\mathsf{C}}(\_, X)$ (cf. §1.2). For every contravariant functor $\mathscr{F} : \mathsf{C} \to \mathsf{Set}$, prove that there is a bijection between the set of natural transformations $h_X \rightsquigarrow \mathscr{F}$ and $\mathscr{F}(X)$, defined as follows. The datum of a natural transformation $h_X \rightsquigarrow \mathscr{F}$ consists of a morphism from $h_X(A) = \operatorname{Hom}_{\mathsf{C}}(A, X)$ to $\mathscr{F}(A)$ for every object A of C. Map h_X to the image of $\mathrm{id}_X \in h_X(X)$ in $\mathscr{F}(X)$. (Hint: Produce an inverse of the specified map. For every $f \in \mathscr{F}(X)$ and every $\varphi \in \operatorname{Hom}_{\mathsf{C}}(A, X)$, how do you construct an element of $\mathscr{F}(A)$?)

This result is called the *Yoneda lemma.* [1.11, IX.2.17]

1.11. (Cf. Exercise 1.9.) Let C be a small category. A contravariant functor $\mathsf{C} \to \mathsf{Set}$ is *representable* if it is naturally isomorphic to a functor h_X (cf. §1.2). In this case, X 'represents' the functor. Prove that C is equivalent to the subcategory of representable functors in $\mathsf{Set}^{\mathsf{C}^\circ}$. (Hint: Yoneda; see Exercise 1.10.)

Thus, every (small) category is equivalent to a subcategory of a functor category.

[9] The reason why C is assumed to be small is that this ensures that the natural transformations between two functors do form a set.

1.12. Let C, D be categories, and let $\mathscr{F} : \mathsf{C} \to \mathsf{D}$, $\mathscr{G} : \mathsf{D} \to \mathsf{C}$ be functors. Prove that $\mathscr{F}$ is left-adjoint to $\mathscr{G}$ if and only if, for every object Y in D, the object $\mathscr{G}(Y)$ represents the functor $h_Y \circ \mathscr{F}$.

1.13. Let Z be the 'Zen' category consisting of no objects and no morphisms. One can contemplate a functor $\mathscr{Z}$ from Z to any category C: no datum whatsoever need be specified. What is $\varprojlim \mathscr{Z}$ (when such an object exists)?

1.14. ▷ Verify that the construction described in Example 1.11 indeed recovers the kernel of a homomorphism of R-modules, as claimed. [§1.4]

1.15. ▷ Verify that the construction given in the proof of Claim 1.13 is an inverse limit, as claimed. [§1.4]

1.16. ▷ Flesh out the sketch of the constructions of colimits in Set and R-Mod given in §1.4, for any indexing poset. In Set, observe that the construction of the colimit is simpler if the poset I is *directed;* that is, if $\forall i, j \in \mathsf{I}$, there exists a $k \in \mathsf{I}$ such that $i \leq k$, $j \leq k$. [§1.4]

1.17. ¬ Let R be a commutative ring, and let $I \subseteq R$ be an ideal. Note that $I^n \subseteq I^m$ if $n \geq m$, and hence we have natural homomorphisms $\varphi_{mn} : R/I^n \to R/I^m$ for $n \geq m$.

- Prove that the inverse limit $\widehat{R}_I := \varprojlim_n R/I^n$ exists as a commutative ring. This is called the *I-adic completion* of R.
- By the universal property of inverse limits, there is a unique homomorphism $R \to \widehat{R}_I$. Prove that the kernel of this homomorphism is $\bigcap_n I^n$.
- Let $I = (x)$ in $R[x]$. Prove that the completion $\widehat{R[x]}_I$ is isomorphic to the power series ring $R[[x]]$ defined in §III.1.3.

[1.18, 1.19]

1.18. Let R be a commutative Noetherian ring, and let $I \subseteq R$ be an ideal. Then $I \cdot \bigcap_n I^n = \bigcap_n I^n$; the reader will prove this in Exercise 4.20.

Assume this, and prove that $\bigcap_n I^n$ equals the set of $r \in R$ such that $(1-a)r = 0$ for some $a \in I$. (Hint: One inclusion is elementary. For the other, use the Nakayama lemma in the form of Exercise VI.3.7. This result is attributed to Krull.)

For example, if I is proper, then $\bigcap_n I^n = (0)$ if R is an integral domain or if it is local. In these cases, the natural map $R \to \widehat{R}_I$ to the I-adic completion is injective (cf. Exercise 1.17).

1.19. ¬ An important example of the construction presented in Exercise 1.17 is the ring $\mathbb{Z}_p$ of *p-adic integers:* this is the limit $\varprojlim_r \mathbb{Z}/p^r\mathbb{Z}$, for a positive prime integer p. The field of fractions of $\mathbb{Z}_p$ is denoted $\mathbb{Q}_p$; elements of $\mathbb{Q}_p$ are called p-adic *numbers.*

- Show that giving a p-adic integer A is equivalent to giving a sequence of integers A_r, $r \geq 1$, such that $0 \leq A_r < p^r$, and that $A_s \equiv A_r \bmod p^s$ if $s \leq r$.
- Equivalently, show that every p-adic integer has a unique infinite expansion $A = a_0 + a_1 \cdot p + a_2 \cdot p^2 + a_3 \cdot p^3 + \cdots$, where $0 \leq a_i \leq p - 1$.

The arithmetic of p-adic integers may be carried out with these expansions in precisely the same way as ordinary arithmetic is carried out with ordinary decimal expansions.

- With notation as in the previous point, prove that $A \in \mathbb{Z}_p$ is invertible if and only if $a_0 \neq 0$.
- Prove that $\mathbb{Z}_p$ is a local domain, with maximal ideal generated by (the image in $\mathbb{Z}_p$ of) p.
- Prove that $\mathbb{Z}_p$ is a DVR (cf. Exercise V.2.19). (There is an evident valuation on $\mathbb{Q}_p$.)

[§II.2.3]

1.20. ¬ If m, n are positive integers and $m \mid n$, then $(n) \subseteq (m)$, and there is an onto ring homomorphism $\mathbb{Z}/n\mathbb{Z} \twoheadrightarrow \mathbb{Z}/m\mathbb{Z}$. The limit ring $\varprojlim_{(n)} \mathbb{Z}/n\mathbb{Z}$ exists and is denoted by $\widehat{\mathbb{Z}}$. Prove that $\widehat{\mathbb{Z}} \cong \mathrm{End}_{\mathsf{Ab}}(\mathbb{Q}/\mathbb{Z})$. (Every $f \in \mathrm{End}_{\mathsf{Ab}}(\mathbb{Q}/\mathbb{Z})$ is determined by $f(\frac{1}{n})$; note that since $n\frac{1}{n} = 1 \equiv 0 \bmod \mathbb{Z}$, $f(\frac{1}{n}) = \frac{g(n)}{n}$ for some integer $g(n)$, which may be chosen so that $0 \leq g(n) < n$. Show that $g(m) \equiv g(n) \mod m$ if $m \mid n$, and think about how elements of $\widehat{\mathbb{Z}}$ may be described.) [1.21]

1.21. Let $\widehat{\mathbb{Z}}$ be as in Exercise 1.20.

- If R is a commutative ring endowed with homomorphisms $R \to \mathbb{Z}/p^r\mathbb{Z}$ for all primes p and all r, compatible with all projections $\mathbb{Z}/p^r\mathbb{Z} \to \mathbb{Z}/p^s\mathbb{Z}$ for $s \leq r$, prove that there are ring homomorphisms $R \to \mathbb{Z}/n\mathbb{Z}$ for all n, compatible with all projections $\mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z}$ for $m \mid n$.
- Deduce that $\widehat{\mathbb{Z}}$ satisfies the universal property for the product of $\mathbb{Z}_p$, as p ranges over all positive prime integers.

It follows that $\prod_p \mathbb{Z}_p \cong \widehat{\mathbb{Z}} \cong \mathrm{End}_{\mathsf{Ab}}(\mathbb{Q}/\mathbb{Z})$.

1.22. Let C be a category, and consider the 'product category' $\mathsf{C} \times \mathsf{C}$ (make sense of such a notion!). There is a 'diagonal' functor associating to each object X of C the pair (X, X) as an object of $\mathsf{C} \times \mathsf{C}$. On the other hand, there may be a 'product functor' $\mathsf{C} \times \mathsf{C} \to \mathsf{C}$, associating to (X, Y) a product $X \times Y$; for example, this is the case in Grp. Convince yourself that the product functor is *right-adjoint* to the diagonal functor. If there is a coproduct functor, verify that it is *left-adjoint* to the diagonal functor.

1.23. ▷ Let R, S be rings. Prove that an additive covariant functor $\mathscr{F} : R\text{-}\mathsf{Mod} \to S\text{-}\mathsf{Mod}$ is exact if and only if $\mathscr{F}(A) \xrightarrow{\mathscr{F}(\varphi)} \mathscr{F}(B) \xrightarrow{\mathscr{F}(\psi)} \mathscr{F}(C)$ is exact in S-Mod whenever $A \xrightarrow{\varphi} B \xrightarrow{\psi} C$ is exact in R-Mod. Deduce that an exact functor sends exact complexes to exact complexes. [§1.5, IX.3.7]

1.24. Let R, S be rings. An additive covariant functor $\mathscr{F} : R\text{-}\mathsf{Mod} \to S\text{-}\mathsf{Mod}$ is *faithfully* exact if '$\mathscr{F}(A) \xrightarrow{\mathscr{F}(\varphi)} \mathscr{F}(B) \xrightarrow{\mathscr{F}(\psi)} \mathscr{F}(C)$ is exact in S-Mod if and only if $A \xrightarrow{\varphi} B \xrightarrow{\psi} C$ is exact in R-Mod'. Prove that an exact functor $\mathscr{F} : R\text{-}\mathsf{Mod} \to$

S-Mod is faithfully exact if and only if $\mathscr{F}(M) \neq 0$ for every nonzero R-module M, if and only if $\mathscr{F}(\varphi) \neq 0$ for every nonzero morphism φ in R-Mod.

1.25. ¬ Prove that localization (Exercise 1.4) is an *exact* functor.

In fact, prove that localization 'preserves homology': if

$$M_\bullet : \cdots \longrightarrow M_{i+1} \xrightarrow{d_{i+1}} M_i \xrightarrow{d_i} M_{i-1} \longrightarrow \cdots$$

is a complex of R-modules and S is a multiplicative subset of R, then the localization $S^{-1}H_i(M_\bullet)$ of the i-th homology of $M_\bullet$ is the i-th homology $H_i(S^{-1}M_\bullet)$ of the localized complex

$$S^{-1}M_\bullet : \cdots \longrightarrow S^{-1}M_{i+1} \xrightarrow{S^{-1}d_{i+1}} S^{-1}M_i \xrightarrow{S^{-1}d_i} S^{-1}M_{i-1} \longrightarrow \cdots$$

[2.12, 2.21, 2.22]

1.26. Prove that localization is faithfully exact in the following sense: let R be a commutative ring, and let

$$(*) \qquad 0 \longrightarrow A \longrightarrow B \longrightarrow C \longrightarrow 0$$

be a sequence of R-modules. Then (*) is exact if and only if the induced sequence of $R_\mathfrak{p}$-modules

$$0 \longrightarrow A_\mathfrak{p} \longrightarrow B_\mathfrak{p} \longrightarrow C_\mathfrak{p} \longrightarrow 0$$

is exact for every prime ideal $\mathfrak{p}$ of R, if and only if it is exact for every maximal ideal $\mathfrak{p}$. (Cf. Exercise V.4.12.)

1.27. ▷ Let R, S be rings. Prove that right-adjoint functors R-Mod $\to$ S-Mod are left-exact and left-adjoint functors are right-exact. [§1.5]

---

2. Tensor products and the Tor functors

In the rest of the chapter we will work in the category R-Mod of modules over a *commutative* ring R. Essentially everything we will see can be upgraded to the noncommutative case without difficulty, but a bit of structure is lost in that case. For example, if R is not commutative, then in the category R-Mod of *left*-R-modules the Hom-sets $\mathrm{Hom}_{R\text{-Mod}}(M, N)$ are 'only' abelian groups (cf. the end of §III.5.2). A tensor product $M \otimes_R N$ can only be defined if M is a right-R-module and N is a left-R-module (in a sense, the two module structures annihilate each other, and what is left is an abelian group). By contrast, in the commutative case we will be able to define $M \otimes_R N$ simply as an R-module. In general, the theory goes through as in the commutative case if the modules carry compatible left- *and* right-module structures, except in questions such as the commutativity of tensors, where it would be unreasonable to expect the commutativity of R to have no bearing. All in all, the commutative case is a little leaner, and (we believe) it suffices in terms of conveying the basic intuition on the general features of the theory.

Thus, R will denote a fixed commutative ring, unless stated otherwise.

2.1. Bilinear maps and the definition of tensor product. If M and N are R-modules, we observed in the distant past (§III.6.1) that $M \oplus N$ serves as both the *product* and *coproduct* of M and N: a situation in which a limit coincides with a colimit. As a set, $M \oplus N$ is just $M \times N$; the R-module structure on $M \oplus N$ is defined by componentwise addition and multiplication by scalars. An R-module homomorphism

$$M \oplus N \to P$$

is determined by R-module homomorphisms $M \to P$ and $N \to P$ (this is what makes $M \oplus N$ into a coproduct).

But there is *another* way to map $M \times N$ to an R-module P, compatibly with the R-module structures.

Definition 2.1. Let M, N, P be R-modules. A function $\varphi : M \times N \to P$ is *R-bilinear* if

- $\forall m \in M$, the function $n \mapsto \varphi(m, n)$ is an R-module homomorphism $N \to P$,
- $\forall n \in N$, the function $m \mapsto \varphi(m, n)$ is an R-module homomorphism $M \to P$.⌟

Thus, if $\varphi : M \times N \to P$ is R-bilinear, then $\forall m \in M$, $\forall n_1, n_2 \in N$, $\forall r_1, r_2 \in R$,

$$\varphi(m, r_1 n_1 + r_2 n_2) = r_1 \varphi(m, n_1) + r_2 \varphi(m, n_2),$$

and similarly for $\varphi(\_, n)$.

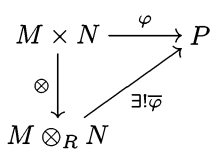
Note that φ itself is *not* linear, even if we view $M \times N$ as the R-module $M \oplus N$, as recalled above. On the other hand, there ought to be a way to deal with R-bilinear maps 'as if' they were R-linear, because such maps abound in the context of R-modules. For example, the very multiplication on R is itself an R-bilinear map

$$R \times R \to R.$$

Our experience with universal properties suggests the natural way to approach this question. What we need is a new R-module $M \otimes_R N$, with an R-bilinear map

$$\otimes : M \times N \to M \otimes_R N,$$

such that *every* R-bilinear map $M \times N \to P$ factors uniquely through this new module $M \otimes_R N$,



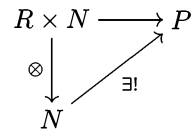
in such a way that the map $\overline{\varphi}$ is a usual R-module homomorphism.

Thus, $M \otimes_R N$ would be the 'best approximation' to $M \times N$ available in R-Mod, if we want to view R-bilinear maps from $M \times N$ as *R-linear*. The module $M \otimes_R N$ is called the *tensor product* of M and N over R. The subscript R is very important: if M and N are modules over two rings R, S, then S-bilinearity is not the same as R-bilinearity, so $M \otimes_R N$ and $M \otimes_S N$ may be completely different objects. In context, it is not unusual to drop the subscript if the base ring is understood, but we do not recommend this practice.

The prescription given above expresses the tensor product as the solution to a universal problem; therefore we know right away that it will be unique up to isomorphism, if it exists (Proposition I.5.4 once more), and we could proceed to study it by systematically using the universal property.

Example 2.2. For all R-modules N, $R \otimes_R N \cong N$.

Indeed, every R-bilinear $R \times N \to P$ factors through N (as is immediately verified):



where $\otimes(r, n) = rn$. By the uniqueness property of universal objects, necessarily $N \cong R \otimes_R N$. ⌟

For another example, it is easy to see that there must be a canonical isomorphism

$$M \otimes_R N \xrightarrow{\sim} N \otimes_R M.$$

Indeed, every R-bilinear $\varphi : M \times N \to P$ may be decomposed as[10]

$$M \times N \longrightarrow N \times M \xrightarrow{\psi} P,$$

with φ the composite, where $\psi(n, m) = \varphi(m, n)$; ψ is also R-bilinear, so it factors uniquely through $N \otimes_R M$. Therefore, φ factors uniquely through $N \otimes_R M$, and this is enough to conclude that there is a canonical isomorphism $N \otimes_R M \cong M \otimes_R N$.

However, such considerations are a little moot unless we establish that $M \otimes_R N$ exists to begin with. This requires a bit of work.

Lemma 2.3. *Tensor products exist in* R-Mod.

Proof. Given R-modules M and N, we construct 'by hand' a module satisfying the universal requirement. Let $F^R(M \times N) = R^{\oplus(M \times N)}$ be the free R-module on $M \times N$ (§III.6.3). This module comes equipped with a *set*-map

$$j : M \times N \to F^R(M \times N),$$

universal with respect to all set-maps from $M \times N$ to any R-module P; the main task is to make this into an *R-bilinear* map. For example, we have to identify elements in $F^R(M \times N)$ of the form $j(m, n_1 + n_2)$ with elements $j(m, n_1) + j(m, n_2)$, etc. Thus, let K be the R-submodule of $F^R(M \times N)$ generated by all elements

$$j(m, r_1 n_1 + r_2 n_2) - r_1 j(m, n_1) - r_2 j(m, n_2)$$

and

$$j(r_1 m_1 + r_2 m_2, n) - r_1 j(m_1, n) - r_2 j(m_2, n)$$

[10] Here is one situation in which the commutativity of R does play a role: if R is not commutative, then this decomposition becomes problematic, even if M and N carry bimodule structures. One can therefore not draw the conclusion $M \otimes_R N \cong N \otimes_R M$ in that case.

as m, m_1, m_2 range in M, n, n_1, n_2 range in N, and r_1, r_2 range in R. Let

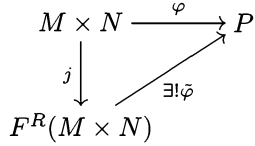
$$M \otimes_R N := \frac{F^R(M \times N)}{K},$$

endowed with the map $\otimes : M \times N \to M \otimes_R N$ obtained by composing j with the natural projection:

$$\otimes : M \times N \xrightarrow{j} F^R(M \times N) \longrightarrow M \otimes_R N = F^R(M \times N)/K$$

The element $\otimes(m, n)$ (that is, the class of $j(m, n)$ modulo K) is denoted $m \otimes n$.

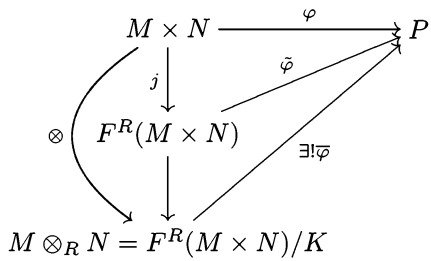
It is evident that $(m, n) \to m \otimes n$ defines an R-bilinear map. We have to check that $M \otimes_R N$ satisfies the universal property, and this is also straightforward. If $\varphi : M \times N \to P$ is any R-bilinear map, we have a unique induced R-linear map $\tilde{\varphi}$ from the free R-module, by the universal property of the latter:



We claim that $\tilde{\varphi}$ restricts to 0 on K. Indeed, to verify this, it suffices to verify that $\tilde{\varphi}$ sends to zero every generator of K, and this follows from the fact that φ is R-bilinear. For example,

$$\begin{aligned}
&\tilde{\varphi}(j(m, r_1 n_1 + r_2 n_2) - r_1 j(m, n_1) - r_2 j(m, n_2)) \\
&\quad = \tilde{\varphi}(j(m, r_1 n_1 + r_2 n_2)) - r_1 \tilde{\varphi}(j(m, n_1)) - r_2 \tilde{\varphi}(j(m, n_2)) \\
&\quad = \varphi(m, r_1 n_1 + r_2 n_2) - r_1 \varphi(m, n_1) - r_2 \varphi(m, n_2) \\
&\quad = 0.
\end{aligned}$$

It follows (by the universal property of quotients!) that $\tilde{\varphi}$ factors uniquely through the quotient by K:



and we are done. □

As is often the case with universal objects, the explicit construction used to prove the existence of $M \otimes_R N$ is almost never invoked. It is however good to keep in mind that elements of $M \otimes_R N$ arise from elements of the free R-module on $M \times N$, and therefore an arbitrary element of $M \otimes_R N$ is a *finite linear combination*

$$(*) \qquad \sum_i r_i (m_i \otimes n_i)$$

with $r_i \in R$, $m_i \in M$, and $n_i \in N$. The R-bilinearity of $\otimes : M \times N \to M \otimes_R N$ amounts to the rules:

$$m \otimes (n_1 + n_2) = m \otimes n_1 + m \otimes n_2,$$
$$(m_1 + m_2) \otimes n = m_1 \otimes n + m_2 \otimes n,$$
$$m \otimes (rn) = (rm) \otimes n = r(m \otimes n),$$

for all $m, m_1, m_2 \in M$, $n_1, n_2, n \in N$, and $r \in R$. In particular, note that the coefficients r_i in (*) are *not* necessary, since they can be absorbed into the corresponding terms $m_i \otimes n_i$:

$$\sum_i r_i(m_i \otimes n_i) = \sum_i (r_i m_i) \otimes n_i.$$

Elements of the form $m \otimes n$ (that is, needing only one summand in the expression) are called *pure tensors*. *Dear reader,* please remember that pure tensors are special: usually, not every element of the tensor product is a pure tensor. See Exercise 2.1 for one situation in which every tensor happens to be pure, and appreciate how special that is.

Pure tensors are nevertheless very useful, as a set of generators for the tensor product. For example, if two homomorphisms $\alpha, \beta : M \otimes_R N \to P$ coincide on *pure* tensors, then $\alpha = \beta$. Frequently, computations involving tensor products are reduced to simple verifications for pure tensors.

2.2. Adjunction with Hom and explicit computations. *The tensor product is left-adjoint to* Hom. Once we parse what this rough statement means, it will be a near triviality; but as we have found out in §1.5, the mere fact that $\otimes_R$ is left-adjoint to *any* functor is enough to draw interesting conclusions about it.

First, we note that every R-module N defines, via $\otimes_R$, a new covariant functor R-Mod $\to$ R-Mod, defined on objects by

$$M \mapsto M \otimes_R N.$$

To see how this works on morphisms, let

$$\alpha : M_1 \to M_2$$

be an R-module homomorphism. Crossing with N and composing with $\otimes$ defines an R-bilinear map

$$M_1 \times N \to M_2 \times N \to M_2 \otimes N,$$

and hence an induced R-linear map

$$\alpha \otimes N : M_1 \otimes N \to M_2 \otimes N.$$

On pure tensors, this map is simply given by $m \otimes n \mapsto \alpha(m) \otimes n$, and functoriality follows immediately: if $\beta : M_0 \to M_1$ is a second homomorphism, then $(\alpha \otimes N) \circ (\beta \otimes N)$ and $(\alpha \circ \beta) \otimes N$ both map pure tensors $m \otimes n$ to $\alpha(\beta(m)) \otimes n$, so they must agree on all tensors.

The adjunction statement given at the beginning of this subsection compares this functor with the covariant functor $P \mapsto \mathrm{Hom}_{R\text{-Mod}}(N, P)$; cf. §1.2. Let's see more precisely how it works.

We have defined $M \otimes_R N$ so that giving an R-linear map $M \otimes_R N \to P$ to an R-module P is 'the same as' giving an R-bilinear map $M \times N \to P$. Now recall the definition of R-bilinear map: $\varphi : M \times N \to P$ is R-bilinear if both $\varphi(m, \_)$ and $\varphi(\_, n)$ are R-linear maps, for all $m \in M$ and $n \in N$. The first part of this prescription says that φ determines a function

$$M \to \mathrm{Hom}_R(N, P);$$

the second part says that this is an R-module homomorphism. Therefore, an R-bilinear map is 'the same as' an element of

$$\mathrm{Hom}_R(M, \mathrm{Hom}_R(N, P)).$$

These simple considerations should be enough to make the following seemingly complicated statement rather natural:

Lemma 2.4. *For all R-modules M, N, P, there is an isomorphism of R-modules*

$$\mathrm{Hom}_R(M, \mathrm{Hom}_R(N, P)) \cong \mathrm{Hom}_R(M \otimes_R N, P).$$

Proof. As noted before the statement, every $\alpha \in \mathrm{Hom}_R(M, \mathrm{Hom}_R(N, P))$ determines an R-bilinear map $\varphi : M \times N \to P$, by

$$(m, n) \mapsto \alpha(m)(n).$$

By the universal property, φ factors uniquely through an R-linear map $\overline{\varphi} : M \otimes_R N \to P$. Therefore, α determines a well-defined element $\overline{\varphi} \in \mathrm{Hom}_R(M \otimes_R N, P)$.

The reader will check (Exercise 2.11) that this map $\alpha \mapsto \overline{\varphi}$ is R-linear and construct an inverse. □

Corollary 2.5. *For every R-module N, the functor $\_ \otimes_R N$ is left-adjoint to the functor* $\mathrm{Hom}_R(N, \_)$.

Proof. The claim is that the isomorphism found in Lemma 2.4 is natural in the sense hinted at, but not fully explained, in §1.5; the interested reader should have no problems checking this naturality. □

By Lemma 1.17 (or rather its co-version), we can conclude that for each R-module N, the functor $\_ \otimes_R N$ preserves colimits, and so does $M \otimes_R \_$, by the basic commutativity of tensor products verified in §2.1. In particular, and this is good material for another Pavlovian reaction,

$$M \otimes_R \_ \text{ and } \_ \otimes_R N \text{ are right-exact functors}$$

(cf. Example 1.18).

These observations have several consequences, which make 'computations' with tensor products more reasonable. Here is a sample:

Corollary 2.6. *For all R-modules M_1, M_2, N,*

$$(M_1 \oplus_R M_2) \otimes N \cong (M_1 \otimes_R N) \oplus (M_2 \otimes_R N).$$

(Moreover, by commutativity, $M \otimes_R (N_1 \oplus N_2) \cong (M \otimes_R N_1) \oplus (M \otimes_R N_2)$ just as well.) Indeed, coproducts are colimits. In fact, $\otimes$ must then commute with *arbitrary* (possibly infinite) direct sums:

$$(\bigoplus_{\alpha \in A} M_\alpha) \otimes_R N \cong \bigoplus_{\alpha \in A} (M_\alpha \otimes_R N).$$

This computes all tensors for *free* R-modules:

Corollary 2.7. *For any two sets A, B:*

$$R^{\oplus A} \otimes_R R^{\oplus B} \cong R^{\oplus A \times B}.$$

Indeed, 'distributing' the direct sum identifies the left-hand side with the direct sum $(R^{\oplus A})^{\oplus B}$, which is isomorphic to the right-hand side (Exercise III.6.5). For *finitely generated* free modules, this simply says that $R^{\oplus m} \otimes R^{\oplus n} \cong R^{\oplus mn}$.

Note that if $e_1, \dots, e_m$ generate M and $f_1, \dots, f_n$ generate N, then the pure tensors $e_i \otimes f_j$ must generate $M \otimes_R N$. In the free case, if the e_i's and f_j's form bases of $R^{\oplus m}$, $R^{\oplus n}$, resp., then the mn elements $e_i \otimes f_j$ must be a basis for $R^{\oplus m} \otimes R^{\oplus n}$. Indeed they generate it; hence they must be linearly independent since this module is free of rank mn. In particular, this is all that can happen if R is a field k and the modules are, therefore, just k-vector spaces (Proposition VI.1.7). Tensor products are more interesting over more general rings.

Corollary 2.8. *For all R-modules N and all ideals I of R,*

$$\frac{R}{I} \otimes_R N \cong \frac{N}{IN}.$$

Indeed, $\_ \otimes_R N$ is right-exact; thus, the exact sequence

$$0 \longrightarrow I \longrightarrow R \longrightarrow \frac{R}{I} \longrightarrow 0$$

induces an exact sequence

$$I \otimes_R N \longrightarrow R \otimes_R N \longrightarrow \frac{R}{I} \otimes_R N \longrightarrow 0.$$

The image of $I \otimes_R N$ in $R \otimes_R N \cong N$ is generated by the image of the pure tensors $a \otimes n$ with $a \in I$, $n \in N$; this is IN. Thus, the second sequence identifies $N/(IN)$ with $(R/I) \otimes_R N$, as needed.

Corollary 2.9. *For all ideals I, J of R,*

$$\frac{R}{I} \otimes_R \frac{R}{J} \cong \frac{R}{I+J}.$$

This follows immediately from Corollary 2.8 and the 'third isomorphism theorem', Proposition III.5.17. Indeed, $IR/J = (I+J)/J$.

Example 2.10. $\mathbb{Z}/m\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/\gcd(m,n)\mathbb{Z}$.

Indeed, $(m) + (n) = (\gcd(m,n))$ in $\mathbb{Z}$. For instance,

$$\frac{\mathbb{Z}}{2\mathbb{Z}} \otimes_{\mathbb{Z}} \frac{\mathbb{Z}}{3\mathbb{Z}} \cong 0,$$

a favorite on qualifying exams (cf. Exercise 2.2). ⌟

Corollary 2.8 is a template example for a basic application of $\otimes$: tensor products may be used to transfer constructions involving R (such as quotienting by an ideal I) to constructions involving R-modules (such as quotienting by a corresponding submodule). There are several instances of this operation; the reader will take a look at *localization* in Exercise 2.5.

2.3. Exactness properties of tensor; flatness. It is important to remember that the tensor product is *not* an exact functor: left-exactness may very well fail. This can already be observed in the sequence appearing in the discussion following Corollary 2.8: for an ideal I of R and an R-module N, the map

$$I \otimes_R N \to N$$

induced by the inclusion $I \subseteq R$ after tensoring by N may not be injective.

Example 2.11. Multiplication by 2 gives an inclusion

$$\mathbb{Z} \xhookrightarrow{\cdot 2} \mathbb{Z}$$

identifying the first copy of $\mathbb{Z}$ with the ideal (2) in the second copy. Tensoring by $\mathbb{Z}/2\mathbb{Z}$ over $\mathbb{Z}$ (and keeping in mind that $R \otimes_R N \cong N$), we get the homomorphism

$$\frac{\mathbb{Z}}{2\mathbb{Z}} \xrightarrow{\cdot 2} \frac{\mathbb{Z}}{2\mathbb{Z}},$$

which sends both [0] and [1] to zero. This is the zero-morphism, and in particular it is not injective. ⌟

On the other hand, if $N \cong R^{\oplus A}$ is *free,* then $\_ \otimes_R N$ *is* exact. Indeed, every inclusion

$$M_1 \subseteq M_2$$

is mapped to $M_1 \otimes_R R^{\oplus A} \to M_2 \otimes_R R^{\oplus A}$, which is identified (via Corollary 2.6) with the *inclusion*

$$M_1^{\oplus A} \subseteq M_2^{\oplus A}.$$

Example 2.12. Since vector spaces are free (Proposition VI.1.7), tensoring is exact in k-Vect: if

$$0 \longrightarrow V_1 \longrightarrow V_2 \longrightarrow V_3 \longrightarrow 0$$

is an exact sequence of k-vector spaces and W is a k-vector space, then the induced sequence

$$0 \longrightarrow V_1 \otimes_k W \longrightarrow V_2 \otimes_k W \longrightarrow V_3 \otimes_k W \longrightarrow 0$$

is exact on both sides. ⌟

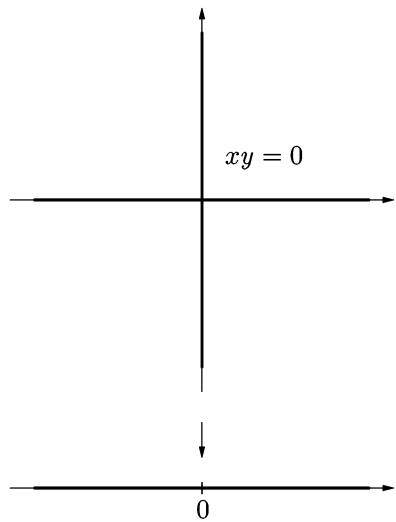
The reader should now wonder whether it is useful to study a condition on an R-module N, guaranteeing that the functor $\_ \otimes_R N$ *is* left-exact as well as right-exact.

Definition 2.13. An R-module N is *flat* if the functor $\_ \otimes_R N$ is exact. ⌟

In the exercises the reader will explore easy properties of this notion and useful equivalent formulations in particular cases.

We have already checked that $\mathbb{Z}/2\mathbb{Z}$ is *not* a flat $\mathbb{Z}$-module, while free modules are flat. Flat modules are hugely important: in algebraic geometry, 'flatness' is the condition expressing the fact that the objects in a family vary 'continuously', preserving certain key invariants.

Example 2.14. Consider the affine algebraic set $\mathscr{V}(xy)$ in the plane $\mathbb{A}^2$ (over a fixed field k) and the 'projection on the first coordinate' $\mathscr{V}(xy) \to \mathbb{A}^1$, $(x, y) \mapsto x$:



In terms of coordinate rings (cf. §VII.2.3), this map corresponds to the homomorphism of k-algebras:

$$k[x] \to \frac{k[x, y]}{(xy)}$$

defined by mapping x to the coset $x + (xy)$ (this will be completely clear to the reader who has worked out Exercise VII.2.12!). This homomorphism defines a $k[x]$-module structure on $k[x, y]/(xy)$, and we can wonder whether the latter is *flat* in the sense of Definition 2.13. From the geometric point of view, clearly something 'not flat' is going on over the point $x = 0$, so we consider the inclusion of the ideal (x) in $k[x]$:

$$k[x] \xhookrightarrow{\cdot x} k[x]$$

Tensoring by $k[x, y]/(xy)$, we obtain

$$\frac{k[x, y]}{(xy)} \xrightarrow{\cdot x} \frac{k[x, y]}{(xy)}$$

which is *not* injective, because it sends to zero the nonzero coset $y + (xy)$. Therefore $k[x, y]/(xy)$ is not flat as a $k[x]$-module.

The term *flat* was inspired precisely by such 'geometric' examples. ⌟

2.4. The Tor functors. The 'failure of exactness' of the functor $\_ \otimes_R N$ is measured by another functor R-Mod $\to$ R-Mod, called $\mathrm{Tor}_1^R(\_, N)$: if N is flat (for example, if it is free), then $\mathrm{Tor}_1^R(M, N) = 0$ for all modules M. In fact (amazingly) if

$$0 \longrightarrow A \longrightarrow B \longrightarrow C \longrightarrow 0$$

is an exact sequence of R-modules, one obtains a new exact sequence after tensoring by any N:

$$\mathrm{Tor}_1^R(C, N) \longrightarrow A \otimes_R N \longrightarrow B \otimes_R N \longrightarrow C \otimes_R N \longrightarrow 0,$$

so if $\mathrm{Tor}_1^R(C, N) = 0$, then the module on the left vanishes; thus *every* short exact sequence ending in C remains exact after tensoring by N in this case. *In fact* (astonishingly) for all N one can continue this sequence with more Tor-modules, obtaining a longer exact complex:

$$\mathrm{Tor}_1^R(A, N) \to \mathrm{Tor}_1^R(B, N) \to \mathrm{Tor}_1^R(C, N) \to A \otimes_R N \to B \otimes_R N \to C \otimes_R N \to 0.$$

This is not the end of the story: the complex may be continued even further by invoking new functors $\mathrm{Tor}_2^R(\_, N)$, $\mathrm{Tor}_3^R(\_, N)$, etc. These are the *derived functors* of tensor. To 'compute' these functors, one may apply the following procedure: given an R-module M, find a free resolution (§VI.4.2)

$$\cdots \longrightarrow R^{\oplus S_2} \longrightarrow R^{\oplus S_1} \longrightarrow R^{\oplus S_0} \longrightarrow M \longrightarrow 0;$$

throw M away, and tensor the free part by N, obtaining a complex $M_\bullet \otimes_R N$:

$$\cdots \longrightarrow N^{\oplus S_2} \longrightarrow N^{\oplus S_1} \longrightarrow N^{\oplus S_0} \longrightarrow 0$$

(recall again that tensor commutes with colimits, hence with direct sums, therefore $R^{\oplus m} \otimes_R N \cong N^{\oplus m}$); then take the *homology* of this complex (cf. §III.7.3). Astoundingly, this *will not depend* (up to isomorphism) on the chosen free resolution, so we can define

$$\mathrm{Tor}_i^R(M, N) := H_i(M_\bullet \otimes N).$$

For example, according to this definition $\mathrm{Tor}_0^R(M, N) \cong M \otimes_R N$ (Exercise 2.14), and $\mathrm{Tor}_i^R(M, N) = 0$ for all $i > 0$ and all M if N is flat (because then tensoring by N is an exact functor, so tensoring the resolution of M returns an exact sequence, thus with no homology). In fact, this proves a remarkable property of the Tor functors: if $\mathrm{Tor}_1^R(M, N) = 0$ for all M, then $\mathrm{Tor}_i^R(M, N) = 0$ for all $i > 0$ for all modules M. Indeed, N is then flat.

At this point you may feel that something is a little out of balance: why focus on the functor $\_ \otimes_R N$, rather than $M \otimes_R \_$? Since $M \otimes_R N$ is canonically isomorphic to $N \otimes_R M$ (in the commutative case; cf. Example 2.2), we could expect the same to apply to every Tor_i^R: $\mathrm{Tor}_i^R(M, N)$ ought to be canonically isomorphic to $\mathrm{Tor}_i^R(N, M)$ for all i. Equivalently, we should be able to compute $\mathrm{Tor}_i^R(M, N)$ as the homology of $M \otimes_R N_\bullet$, where $N_\bullet$ is a free resolution of N. This is indeed the case.

In due time (§§IX.7 and 8) we will prove this and all the other wonderful facts we have stated in this subsection. For now, we are asking the reader to believe that the Tor functors can be defined as we have indicated, and the facts reviewed here

will suffice for simple computations (see for example Exercises 2.15 and 2.17) and applications.

In fact, we know enough about finitely generated modules over PIDs to get a preliminary sense of what is involved in proving such general facts. Recall that we have been able to establish that every finitely generated module M over a PID R has a free resolution of length 1:

$$0 \longrightarrow R^{\oplus m_1} \longrightarrow R^{\oplus m_0} \longrightarrow M \longrightarrow 0\,.$$

This property *characterizes* PIDs (Proposition VI.5.4). If

$$0 \longrightarrow A \longrightarrow B \longrightarrow C \longrightarrow 0$$

is an exact sequence of R-modules, it is not hard to see that one can produce 'compatible' resolutions, in the sense that the rows of the following diagram will be exact as well as the columns:

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & 0 & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & R^{\oplus a_1} & \longrightarrow & R^{\oplus b_1} & \longrightarrow & R^{\oplus c_1} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle \gamma} & & \\
0 & \longrightarrow & R^{\oplus a_0} & \longrightarrow & R^{\oplus b_0} & \longrightarrow & R^{\oplus c_0} & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & A & \longrightarrow & B & \longrightarrow & C & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & 0 & & 0 & & 0 & &
\end{array}$$

(This will be proven in gory detail in §IX.7.) Tensor the two 'free' rows by N; they remain exact (tensoring commutes with direct sums):

$$\begin{array}{ccccccccc}
0 & \longrightarrow & N^{\oplus a_1} & \longrightarrow & N^{\oplus b_1} & \longrightarrow & N^{\oplus c_1} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \alpha\otimes N} & & \downarrow{\scriptstyle \beta\otimes N} & & \downarrow{\scriptstyle \gamma\otimes N} & & \\
0 & \longrightarrow & N^{\oplus a_0} & \longrightarrow & N^{\oplus b_0} & \longrightarrow & N^{\oplus c_0} & \longrightarrow & 0
\end{array}$$

Now the columns (preceded and followed by 0) are precisely the complexes $A_\bullet \otimes_R N$, $B_\bullet \otimes_R N$, $C_\bullet \otimes_R N$ whose homology 'computes' the Tor modules. Applying the snake lemma (Lemma III.7.8; cf. Remark III.7.10) gives the exact sequence

$$\begin{array}{l}
0 \longrightarrow H_1(A_\bullet \otimes_R N) \longrightarrow H_1(B_\bullet \otimes_R N) \longrightarrow H_1(C_\bullet \otimes_R N) \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \overset{\delta}{\longrightarrow} \\
H_0(A_\bullet \otimes_R N) \longrightarrow H_0(B_\bullet \otimes_R N) \longrightarrow H_0(C_\bullet \otimes_R N) \longrightarrow 0,
\end{array}$$

which is precisely the sequence of Tor modules conjured up above,

$$0 \longrightarrow \mathrm{Tor}_1^R(A,N) \longrightarrow \mathrm{Tor}_1^R(B,N) \longrightarrow \mathrm{Tor}_1^R(C,N) \xrightarrow{\ \delta\ } A \otimes_R N \longrightarrow B \otimes_R N \longrightarrow C \otimes_R N \longrightarrow 0$$

with a 0 on the left for good measure (due to the fact that Tor_2^R vanishes if R is a PID; cf. Exercise 2.17).

Note that Tor_i^k vanishes for $i > 0$ if k is a field, as vector spaces are flat, and Tor_i^R vanishes for $i > 1$ if R is a PID (Exercise 2.17). These facts are not surprising, in view of the procedure described above for computing Tor and of the considerations at the end of §VI.5.2: a bound on the length of free resolutions for modules over a ring R will imply a bound on nonzero Tor's. For particularly nice rings (such as the rings corresponding to 'smooth' points in algebraic geometry) this bound agrees with the Krull dimension; but precise results of this sort are beyond the scope of this book.

Exercises

R denotes a fixed commutative ring.

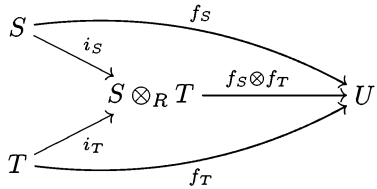
2.1. ▷ Let M, N be R-modules, and assume that N is *cyclic*. Prove that every element of $M \otimes_R N$ may be written as a pure tensor. [§2.1]

2.2. ▷ Prove 'by hand' (that is, without appealing to the right-exactness of tensor) that $\mathbb{Z}/n\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/m\mathbb{Z} \cong 0$ if m, n are relatively prime integers. [§2.2]

2.3. Prove that $R[x_1, \ldots, x_n] \otimes_R R[y_1, \ldots, y_m] \cong R[x_1, \ldots, x_n, y_1, \ldots, y_m]$.

2.4. ¬ Let S, T be commutative R-algebras. Verify the following:

- The tensor product $S \otimes_R T$ has an operation of multiplication, defined on pure tensors by $(s_1 \otimes t_1) \cdot (s_2 \otimes t_2) := s_1 s_2 \otimes t_1 t_2$ and making it into a commutative R-algebra.
- With respect to this structure, there are R-algebra homomorphisms $i_S : S \to S \otimes T$, resp., $i_T : T \to S \otimes T$, defined by $i_S(s) := s \otimes 1$, $i_T(t) := 1 \otimes t$.
- The R-algebra $S \otimes_R T$, with these two structure homomorphisms, is a coproduct of S and T in the category of commutative R-algebras: if U is a commutative R-algebra and $f_S : S \to U$, $f_T : T \to U$ are R-algebra homomorphisms, then there exists a unique R-algebra homomorphism $f_S \otimes f_T$ making the following diagram commute:



In particular, if S and T are simply commutative rings, then $S \otimes_{\mathbb{Z}} T$ is a coproduct of S and T in the category of commutative rings. This settles an issue left open at the end of §III.2.4. [2.10]

2.5. ▷ (Cf. Exercises V.4.7 and V.4.8.) Let S be a multiplicative subset of R, and let M be an R-module. Prove that $S^{-1}M \cong M \otimes_R S^{-1}R$ as R-modules. (Use the universal property of the tensor product.)

Through this isomorphism, $M \otimes_R S^{-1}R$ inherits an $S^{-1}R$-module structure. [§2.2, 2.8, 2.12, 3.4]

2.6. ¬ (Cf. Exercises V.4.7 and V.4.8.) Let S be a multiplicative subset of R, and let M be an R-module.

- Let N be an $S^{-1}R$-module. Prove that $(S^{-1}M) \otimes_{S^{-1}R} N \cong M \otimes_R N$.
- Let A be an R-module. Prove that $(S^{-1}A) \otimes_R M \cong S^{-1}(A \otimes_R M)$.

(Both can be done 'by hand', by analyzing the construction in Lemma 2.3. For example, there is a homomorphism $M \otimes_R N \to (S^{-1}M) \otimes_{S^{-1}R} N$ which is surjective because, with evident notation, $\frac{m}{s} \otimes n = m \otimes \frac{n}{s}$ in $(S^{-1}M) \otimes_{S^{-1}R} N$; checking that it is injective amounts to easy manipulation of the relations defining the two tensor products.

Both isomorphisms will be easy consequences of the associativity of tensor products; cf. Exercise 3.4.) [2.21, 3.4]

2.7. Changing the base ring in a tensor may or may not make a difference:

- Prove that $\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Q} \cong \mathbb{Q} \otimes_{\mathbb{Q}} \mathbb{Q}$.
- Prove that $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C} \ncong \mathbb{C} \otimes_{\mathbb{C}} \mathbb{C}$.

2.8. Let R be an integral domain, with field of fractions K, and let M be a finitely generated R-module. The tensor product $V := M \otimes_R K$ is a K-vector space (Exercise 2.5). Prove that $\dim_K V$ equals the rank of M as an R-module, in the sense of Definition VI.5.5.

2.9. Let G be a finitely generated abelian group of rank r. Prove that $G \otimes_{\mathbb{Z}} \mathbb{Q} \cong \mathbb{Q}^r$. Prove that for infinitely many primes p, $G \otimes_{\mathbb{Z}} (\mathbb{Z}/p\mathbb{Z}) \cong (\mathbb{Z}/p\mathbb{Z})^r$.

2.10. Let $k \subseteq k(\alpha) = F$ be a finite simple field extension. Note that $F \otimes_k F$ has a natural ring structure; cf. Exercise 2.4.

- Prove that α is separable over k if and only if $F \otimes_k F$ is *reduced* as a ring.
- Prove that $k \subseteq F$ is Galois if and only if $F \otimes_k F$ is isomorphic to $F^{[F:k]}$ as a ring.

(Use Corollary 2.8 to 'compute' the tensor. The CRT from §V.6.1 will likely be helpful.)

2.11. ▷ Complete the proof of Lemma 2.4. [§2.2]

2.12. Let S be a multiplicative subset of R (cf. Exercise V.4.7). Prove that $S^{-1}R$ is flat over R. (Hint: Exercises 2.5 and 1.25.)

2.13. Prove that direct sums of flat modules are flat.

2.14. ▷ Prove that, according to the definition given in §2.4, $\mathrm{Tor}_0^R(M,N)$ is isomorphic to $M \otimes_R N$. [§2.4]

2.15. ▷ Prove that for $r \in R$ a non-zero-divisor and N an R-module, the module $\mathrm{Tor}_1^R(R/(r),N)$ is isomorphic to the r-torsion of N, that is, the submodule of elements $n \in N$ such that $rn = 0$ (cf. §VI.4.1). (This is the reason why Tor is called Tor.) [§2.4, 6.21]

2.16. Let I, J be ideals of R. Prove that $\mathrm{Tor}_1^R(R/I, R/J) \cong (I \cap J)/IJ$. (For example, this Tor_1^R vanishes if $I + J = R$, by Lemma V.6.2.) Prove that $\mathrm{Tor}_i^R(R/I, R/J)$ is isomorphic to $\mathrm{Tor}_{i-1}^R(I, R/J)$ for $i > 1$.

2.17. ▷ Let M, N be modules over a PID R. Prove that $\mathrm{Tor}_i^R(M,N) = 0$ for $i \geq 2$. (Assume M, N are finitely generated, for simplicity.) [§2.4]

2.18. Let R be an integral domain. Prove that a cyclic R-module is flat if and only if it is free.

2.19. ¬ The following criterion is quite useful.

- Prove that an R-module M is flat if and only if every monomorphism of R-modules $A \hookrightarrow B$ induces a monomorphism of R-modules $A \otimes_R M \hookrightarrow B \otimes_R M$.
- Prove that it suffices to verify this condition for all *finitely generated* modules B. (Hint: For once, refer back to the construction of tensor products given in Lemma 2.3. An element $\sum_i a_i \otimes m_i \in A \otimes_R M$ goes to zero in $B \otimes_R M$ if the corresponding element $\sum_i (a_i, m_i)$ equals a combination of the relations defining $B \otimes_R M$ in the free R-module $F^R(B \times M)$. This will be an identity involving only finitely many elements of B; hence....)
- Prove that it suffices to verify this condition when $B = R$ and $A = I$ is an ideal of R. (Hint: We may now assume that B is finitely generated. Find submodules B_j such that $A = B_0 \subseteq B_1 \subseteq \cdots \subseteq B_r = B$, with each B_j/B_{j-1} cyclic. Reduce to verifying that $A \otimes_R M$ injects in $B \otimes_R M$ when B/A is cyclic, hence $\cong R/I$ for some ideal I. Conclude by a Tor_1^R argument or—but this requires a little more stamina—by judicious use of the snake lemma.)
- Deduce that an R-module M is flat if and only if the natural homomorphism $I \otimes_R M \to IM$ is an isomorphism for every ideal I of R.

If you believe in Tor's, now you can also show that an R-module M is flat if and only if $\mathrm{Tor}_1^R(R/I, M) = 0$ for all ideals I of R. [2.20]

2.20. Let R be a PID. Prove that an R-module M is flat if and only if it is torsion-free. (If M is finitely generated, the classification theorem of §VI.5.3 makes this particularly easy. Otherwise, use Exercise 2.19.)

Geometrically, this says roughly that an algebraic set fails to be 'flat' over a nonsingular curve if and only if some component of the set is contracted to a point. This phenomenon is displayed in the picture in Example 2.14.

2.21. ¬ (Cf. Exercise V.4.11.) Prove that *flatness is a local property:* an R-module M is flat if and only if $M_{\mathfrak{p}}$ is a flat $R_{\mathfrak{p}}$-module for all prime ideals $\mathfrak{p}$, if and only if $M_{\mathfrak{m}}$ is a flat $R_{\mathfrak{m}}$-module for all maximal ideals $\mathfrak{m}$. (Hint: Use Exercises 1.25 and 2.6.

The $\Longrightarrow$ direction will be straightforward. For the converse, let $A \subseteq B$ be R-modules, and let K be the kernel of the induced homomorphism $A \otimes_R M \to B \otimes_R M$. Prove that the kernel of the localized homomorphism $A_{\mathfrak{m}} \otimes_{R_{\mathfrak{m}}} M_{\mathfrak{m}} \to B_{\mathfrak{m}} \otimes_{R_{\mathfrak{m}}} M_{\mathfrak{m}}$ is isomorphic to $K_{\mathfrak{m}}$, and use Exercise V.4.12.) [2.22]

2.22. ¬ Let M, N be R-modules, and let S be a multiplicative subset of R. Use the definition of Tor given in §2.4 to show $S^{-1}\operatorname{Tor}_i^R(M,N) \cong \operatorname{Tor}_i^{S^{-1}R}(S^{-1}M, S^{-1}N)$. (Use Exercise 1.25.) Use this fact to give a leaner proof that flatness is a local property (Exercise 2.21). [2.25]

2.23. ▷ Let

$$0 \longrightarrow M \longrightarrow N \longrightarrow P \longrightarrow 0$$

be an exact sequence of R-modules, and assume that P is flat.

- Prove that M is flat if and only if N is flat.
- Prove that for all R-modules Q, the induced sequence

$$0 \longrightarrow M \otimes_R Q \longrightarrow N \otimes_R Q \longrightarrow P \otimes_R Q \longrightarrow 0$$

  is exact.

[2.24, §5.4]

2.24. ¬ Let R be a commutative Noetherian local ring with (single) maximal ideal $\mathfrak{m}$, and let M be a finitely generated flat R-module.

- Choose elements $m_1, \ldots, m_r \in M$ whose cosets mod $\mathfrak{m}M$ are a basis of $M/\mathfrak{m}M$ as a vector space over the field $R/\mathfrak{m}$. By Nakayama's lemma, $M = \langle m_1, \ldots, m_r \rangle$ (Exercise VI.3.10).
- Obtain an exact sequence

$$0 \longrightarrow N \longrightarrow R^{\oplus r} \longrightarrow M \longrightarrow 0,$$

  where N is finitely generated.
- Prove that this sequence induces an exact sequence

$$0 \longrightarrow N/\mathfrak{m}N \longrightarrow (R/\mathfrak{m})^{\oplus r} \longrightarrow M/\mathfrak{m}M \longrightarrow 0.$$

  (Use Exercise 2.23.)
- Deduce that $N = 0$. (Nakayama.)
- Conclude that M is free.

Thus, a finitely generated module over a (Noetherian[11]) local ring is flat if and only if it is free. Compare with Exercise VI.5.5. [2.25, 6.8, 6.12]

2.25. Let R be a commutative Noetherian ring, and let M be a finitely generated R-module. Prove that

$$M \text{ is flat} \iff \operatorname{Tor}_1^R(M, R/\mathfrak{m}) = 0 \text{ for every maximal ideal } \mathfrak{m} \text{ of } R\ .$$

[11] The Noetherian hypothesis is actually unnecessary, but it simplifies the proof by allowing the use of Nakayama's lemma.

(Use Exercise 2.21, and refine the argument you used in Exercise 2.24; remember that Tor localizes, by Exercise 2.22. The Noetherian hypothesis is actually unnecessary, but the proofs are harder without it.)

---

3. Base change

We have championed several times the viewpoint that deep properties of a ring R are encoded in the category R-Mod of R-modules; one extreme position is to simply replace R with R-Mod as the main object of study. The question then arises as to how to deal with ring homomorphisms from this point of view, or more generally how the categories R-Mod, S-Mod of modules over two (commutative) rings R, S may relate to each other. The reader should expect this to happen by way of *functors* between the two categories and that the situation at the categorical level will be substantially richer than at the ring level.

3.1. Balanced maps. Before we can survey the basic definitions, we must upgrade our understanding of tensor products. It turns out that $M \otimes_R N$ satisfies a more encompassing universal property than the one examined in §2.1. Let M, N be modules over a commutative ring R, as in §2.1, and let G be an abelian group, i.e., a $\mathbb{Z}$-module.

Definition 3.1. A $\mathbb{Z}$-bilinear map $\varphi : M \times N \to G$ is *R-balanced* if $\forall m \in M$, $\forall n \in N$, $\forall r \in R$,

$$\varphi(rm, n) = \varphi(m, rn).$$ ⌟

If G is an R-module and $\varphi : M \times N \to G$ is R-bilinear, then it is R-balanced[12]. But in general the notion of 'R-balanced' appears to be quite a bit more general, since G is not even required to be an R-module. This may lead the reader to suspect that a solution to the universal problem of factoring balanced maps may be a different gadget than the 'ordinary' tensor product, but we are in luck in this case, and the ordinary tensor product does the universal job for balanced maps as well.

To understand this, recall that we constructed $M \otimes_R N$ as a quotient

$$M \otimes_R N = \frac{R^{\oplus(M \times N)}}{K},$$

where K is generated by the relations necessary to imply that the map

$$M \times N \to R^{\oplus(M \times N)} \to \frac{R^{\oplus(M \times N)}}{K}$$

is R-bilinear. We have observed that every element of $M \otimes_R N$ may be written as a linear combination of pure tensors:

$$\sum_i m_i \otimes n_i;$$

[12] Also note that if R is not commutative and M, resp., N, carries a right-, resp., left-, R module structure, then the notion of 'balanced map' makes sense. This leads to the definition of tensor (as an abelian group) in the noncommutative case.

it follows that the group homomorphism

$$\mathbb{Z}^{\oplus(M\times N)} \twoheadrightarrow M \otimes_R N$$

defined on generators by $(m,n) \mapsto m \otimes n$ is *surjective;* its kernel K_B consists of the combinations

$$\sum_i (m_i, n_i) \in \mathbb{Z}^{\oplus(M\times N)} \quad \text{such that} \quad \sum_i (m_i, n_i) \in K,$$

where the sum on the right is viewed in $R^{\oplus(M\times N)}$. The reader will verify (Exercise 3.1) that K_B is generated by elements of the form

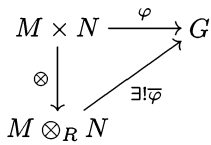
$$\begin{gathered}(m, n_1 + n_2) - (m, n_1) - (m, n_2),\\ (m_1 + m_2, n) - (m_1, n) - (m_2, n),\\ (rm, n) - (m, rn)\end{gathered}$$

(with $m, m_1, m_2 \in M$, $n, n_1, n_2 \in N$, $r \in R$). Therefore, we have an induced isomorphism of abelian groups

$$(*) \qquad \frac{\mathbb{Z}^{\oplus(M\times N)}}{K_B} \cong \frac{R^{\oplus(M\times N)}}{K},$$

which amounts to an alternative description of $M \otimes_R N$. The point of this observation is that the group on the left-hand side of (*) is manifestly a solution to the universal problem of factoring $\mathbb{Z}$-bilinear, R-balanced maps. Therefore, we have proved

Lemma 3.2. *Let R be a commutative ring; let M, N be R-modules, and let G be an abelian group. Then every $\mathbb{Z}$-bilinear, R-balanced map $\varphi : M \times N \to G$ factors through $M \otimes_R N$; that is, there exists a unique group homomorphism $\overline{\varphi} : M \otimes_R N \to G$ such that the diagram*



commutes.

The universal property explored in §2.1 is recovered as the statement that *if* G is an R-module and φ is R-bilinear, *then* the induced group homomorphism $M \otimes_R N \to G$ is in fact an R-linear map.

Remark 3.3. Balanced maps $\varphi : M \times N \to G$ may be defined as soon as M is a *right-R-*module and N is a *left-R-*module, even if R is not commutative: require $\varphi(mr, n) = \varphi(m, rn)$ for all $m \in M$, $n \in N$, $r \in R$. The abelian group defined by the left-hand side of (*) still makes sense and is taken as the definition of the tensor product $M \otimes_R N$; but note that this does not carry an R-module structure in general. This structure is recovered if, e.g., M is a two-sided R-module. ⌟

3.2. Bimodules; adjunction again. The enhanced universal property for the tensor will allow us to upgrade the adjunction formula given in Lemma 2.4. This requires the introduction of yet another notion.

Definition 3.4. Let R, S be two commutative[13] rings. An (R,S)-*bimodule* is an abelian group N endowed with compatible R-module and S-module structures, in the sense that $\forall n \in N$, $\forall r \in R$, $\forall s \in S$,

$$r(sn) = s(rn).$$ ⌟

For example, as R is commutative, every R-module N is an (R,R)-bimodule: $\forall r_1, r_2 \in R$ and $\forall n \in N$,

$$r_1(r_2 n) = (r_1 r_2)n = (r_2 r_1)n = r_2(r_1 n).$$

If M is an R-module and N is an (R,S)-bimodule, then the tensor product $M \otimes_R N$ acquires an S-*module* structure: define the action of $s \in S$ on pure tensors $m \otimes n$ by

$$s(m \otimes n) := m \otimes (sn),$$

and extend to all tensors by linearity. In fact, this gives $M \otimes_R N$ an (R,S)-bimodule structure.

Similarly, if N is an (R,S)-bimodule and P is an S-module, then the abelian group $\mathrm{Hom}_S(N,P)$ is an (R,S)-bimodule: the R-module structure is defined by setting $(r\alpha)(n) = \alpha(rn)$ for all $r \in R$, $n \in N$, and $\alpha \in \mathrm{Hom}_S(N,P)$.

This mess is needed to even make sense of the promised upgrade of adjunction. As with most such results, the proof is not difficult once one understands what the statement says.

Lemma 3.5. *Suppose M is an R-module, N is an (R,S)-bimodule, and P is an S-module. Then there is a canonical isomorphism of abelian groups*

$$\mathrm{Hom}_R(M, \mathrm{Hom}_S(N,P)) \cong \mathrm{Hom}_S(M \otimes_R N, P).$$

Proof. Every element $\alpha \in \mathrm{Hom}_R(M, \mathrm{Hom}_S(N,P))$ determines a map

$$\varphi : M \times N \to P,$$

via $\varphi(m, \_) := \alpha(m)$; φ is clearly $\mathbb{Z}$-bilinear. Further, for all $r \in R$, $m \in M$, $n \in N$:

$$\varphi(rm, n) = \alpha(rm)(n) \overset{1}{=} r\alpha(m)(n) \overset{2}{=} \alpha(m)(rn) = \varphi(m, rn),$$

where $\overset{1}{=}$ holds by the R-linearity of α and $\overset{2}{=}$ holds by the definition of the R-module structure on $\mathrm{Hom}_S(N,P)$. Thus φ is R-balanced. By Lemma 3.2, such a map determines (and is determined by) a homomorphism of abelian groups

$$\overline{\varphi} : M \otimes_R N \to P,$$

[13] Once more, the noncommutative case would be very worthwhile pursuing, but (not without misgivings) we have decided otherwise. In this more general case one requires N to be both a *left-R*-module and a *right-S*-module with the compatibility expressed by $(rn)s = r(ns)$ for all choices of $n \in N$, $r \in R$, $s \in S$. This type of bookkeeping is precisely what is needed in order to extend the theory to the noncommutative case.

such that $\varphi(m,n) = \overline{\varphi}(m \otimes n)$. We claim that $\overline{\varphi}$ is S-linear. Indeed, $\forall s \in S$ and for all pure tensors $m \otimes n$,

$$\begin{aligned}\overline{\varphi}(s(m \otimes n)) &= \overline{\varphi}(m \otimes (sn)) = \varphi(m, sn) = \alpha(m)(sn) = s\alpha(m)(n) = s\varphi(m,n) \\ &= s\overline{\varphi}(m \otimes n),\end{aligned}$$

where we have used the S-linearity of $\alpha(m)$. Thus,

$$\overline{\varphi} \in \mathrm{Hom}_S(M \otimes_R N, P).$$

Tracing the argument backwards, every element of $\mathrm{Hom}_S(M \otimes_R N, P)$ determines an element of $\mathrm{Hom}_R(M, \mathrm{Hom}_S(N,P))$, and these two correspondences are clearly inverses of each other. □

If $R = S$, we recover the adjunction formula of Lemma 2.4; note that in this case the isomorphism is clearly R-linear.

3.3. Restriction and extension of scalars. Coming back to the theme mentioned at the beginning of this section, consider the case in which we have a homomorphism $f : R \to S$ of (commutative) rings. It is natural to look for functors between the categories R-Mod and S-Mod of modules over R, S, respectively. There is a rather simple-minded functor from S-Mod to R-Mod ('restriction of scalars'), while tensor products allow us to define a functor from R-Mod to S-Mod ('extension of scalars'). A third important functor R-Mod $\to S$-Mod may be defined, also 'extending scalars', but for which we do not know a good name.

Restriction of scalars. Let $f : R \to S$ be a ring homomorphism, and let N be an S-module. Recall (§III.5.1) that this means that we have chosen an action of the ring S on the abelian group N, that is, a ring homomorphism

$$\sigma : S \to \mathrm{End}_{\mathsf{Ab}}(N).$$

Composing with f,

$$\sigma \circ f : R \to S \to \mathrm{End}_{\mathsf{Ab}}(N)$$

defines an action *of* R on the abelian group N, and hence an R-module structure on N.

(Even) more explicitly, if $r \in R$ and $n \in N$, define the action of r on n by setting

$$rn := f(r)n.$$

Since S is commutative, this defines in fact an (R,S)-bimodule structure on N. Further, S-linear homomorphisms are in particular R-linear; this assignment is (covariantly) functorial S-Mod $\to R$-Mod.

If f is injective, so that R may be viewed as a subring of S, then all we are doing is viewing N as a module on a 'restricted' range of scalars, hence the terminology. For example, this is how we view a complex vector space as a *real* vector space, in the simplest possible way.

We will denote by f_* this functor S-Mod $\to R$-Mod induced from f by restriction of scalars. Note that f_* is trivially exact, because the kernels and images of a homomorphism of modules are the same regardless of the base ring. In view of the

considerations in Example 1.18, this hints that f_* may have *both* a left-adjoint and a right-adjoint functor, and this will be precisely the case (Proposition 3.6).

Extension of scalars is defined from R-Mod to S-Mod, by associating to an R-module M the tensor product $f^*(M) := M \otimes_R S$, which (as we have seen in §3.2) carries naturally an S-module structure. This association is evidently covariantly functorial. If

$$R^{\oplus B} \to R^{\oplus A} \to M \to 0$$

is a presentation of M (cf. §VI.4.2), tensoring by S gives (by the right-exactness of tensor) a presentation of $f^*(M)$:

$$S^{\oplus B} \to S^{\oplus A} \to M \otimes_R S \to 0.$$

Intuitively, this says that $f^*(M)$ is the module defined by 'the same generators and relations' as M, but with coefficients in S.

The third functor, denoted $f^!$, also acts from R-Mod to S-Mod and is yet another natural way to combine the ingredients we have at our disposal: if M is an R-module, we have pointed out[14] in §3.2 that

$$f^!(M) := \mathrm{Hom}_R(S, M)$$

may be given a natural S-module structure (by setting $s\alpha(s') := \alpha(ss')$). This is again evidently a covariantly functorial prescription.

Proposition 3.6. *Let $f : R \to S$ be a homomorphism of commutative rings. Then, with notation as above, f_* is right-adjoint to f^* and left-adjoint to $f^!$. In particular, f_* is exact, f^* is right-exact, and $f^!$ is left-exact.*

Proof. Let M, resp., N, be an R-module, resp., an S-module. Note that, trivially, $\mathrm{Hom}_S(S, N)$ is canonically isomorphic to N (as an S-module) and to $f_*(N)$ (as an R-module). Thus[15]

$$\begin{aligned}\mathrm{Hom}_R(M, f_*(N)) &\cong \mathrm{Hom}_R(M, \mathrm{Hom}_S(S, N)) \\ &\cong \mathrm{Hom}_S(M \otimes_R S, N) = \mathrm{Hom}_S(f^*(M), N)\end{aligned}$$

where we have used Lemma 3.5. These bijections are canonical[16], proving that f^* is left-adjoint to f_*.

Similarly, there is a canonical isomorphism $N \cong N \otimes_S S$ (Example 2.2); thus $N \otimes_S S \cong f_*(N)$ as R-modules, and for every R-module M

$$\begin{aligned}\mathrm{Hom}_R(f_*(N), M) &\cong \mathrm{Hom}_R(N \otimes_S S, M) \\ &\cong \mathrm{Hom}_S(N, \mathrm{Hom}_R(S, M)) = \mathrm{Hom}_S(N, f^!(M))\end{aligned}$$

again by Lemma 3.5. This shows that $f^!$ is right-adjoint to f_*, concluding the proof. □

[14] The roles of R and S were reversed in §3.2.

[15] These are isomorphisms of abelian groups, and in fact isomorphisms of R-modules if one applies f_* to the Hom_S terms.

[16] The reader should check this....

Remark 3.7 (Warning)**.** Our choice of notation, f_*, etc., is somewhat nonstandard, and the reader should not take it too literally. It is inspired by analogs in the context of sheaf theory over schemes, but some of the properties reviewed above require crucial adjustments in that wider context: for example f_* is *not* exact as a sheaf operation on schemes. ⌟

Exercises

In the following exercises, R, S denote commutative rings.

3.1. ▷ Verify that a combination of pure tensors $\sum_i (m_i \otimes n_i)$ is zero in the tensor product $M \otimes_R N$ if and only if $\sum_i (m_i, n_i) \in \mathbb{Z}^{\oplus(M\times N)}$ is a combination of elements of the form

$$(m, n_1 + n_2) - (m, n_1) - (m, n_2),$$
$$(m_1 + m_2, n) - (m_1, n) - (m_2, n),$$
$$(rm, n) - (m, rn),$$

with $m, m_1, m_2 \in M$, $n, n_1, n_2 \in N$, $r \in R$. [§3.1]

3.2. If $f : R \to S$ is a ring homomorphism and M, N are S-modules (hence R-modules by restriction of scalars), prove that there is a canonical homomorphism of R-modules $M \otimes_R N \to M \otimes_S N$.

3.3. ▷ Let R, S be commutative rings, and let M be an R-module, N an (R, S)-bimodule, and P as S-module. Prove that there is an isomorphism of R-modules

$$M \otimes_R (N \otimes_S P) \cong (M \otimes_R N) \otimes_S P.$$

In this sense, $\otimes$ is 'associative'. [3.4, §4.1]

3.4. ▷ Use the associativity of the tensor product (Exercise 3.3) to prove again the formulas given in Exercise 2.6. (Use Exercise 2.5.) [2.6]

3.5. Let $f : R \to S$ be a ring homomorphism. Prove that $f^!$ commutes with limits, f^* commutes with colimits, and f_* commutes with both. In particular, deduce that these three functors all preserve finite direct sums.

3.6. Let $f : R \to S$ be a ring homomorphism, and let $\varphi : N_1 \to N_2$ be a homomorphism of S-modules. Prove that φ is an isomorphism if and only if $f_*(\varphi)$ is an isomorphism. (Functors with this property are said to be *conservative.*) In fact, prove that f_* is *faithfully* exact: a sequence of S-modules

$$0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$$

is exact *if and only if* the sequence of R-modules

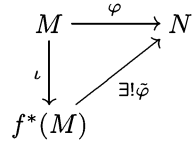
$$0 \longrightarrow f_*(L) \longrightarrow f_*(M) \longrightarrow f_*(N) \longrightarrow 0$$

is exact. In particular, a sequence of R-modules is exact if and only if it is exact as a sequence of abelian groups. (This is completely trivial but useful nonetheless.)

3.7. Let $i : k \subseteq F$ be a finite field extension, and let W be an F-vector space of finite dimension n. Compute the dimension of $i_*(W)$ as a k-vector space (where i_* is restriction of scalars; cf. §3.3).

3.8. Let $i : k \subseteq F$ be a finite field extension, and let V be a k-vector space of dimension n. Compute the dimension of $i^*(V)$ and $i^!(V)$ as F-vector spaces.

3.9. Let $f : R \to S$ be a ring homomorphism, and let M be an R-module. Prove that the extension $f^*(M)$ satisfies the following universal property: if N is an S-module and $\varphi : M \to N$ is an R-linear map, then there exists a unique S-linear map $\tilde{\varphi} : f^*(M) \to S$ making the diagram



commute, where $\iota : M \to f^*(M) = M \otimes_R S$ is defined by $m \mapsto m \otimes 1$. (Thus, $f^*(M)$ is the 'best approximation' to the R-module M in the category of S-modules.)

3.10. Prove the following *projection formula:* if $f : R \to S$ is a ring homomorphism, M is an R-module, and N is an S-module, then $f_*(f^*(M) \otimes_S N) \cong M \otimes_R f_*(N)$ as R-modules.

3.11. Let $f : R \to S$ be a ring homomorphism, and let M be a *flat* R-module. Prove that $f^*(M)$ is a flat S-module.

3.12. In 'geometric' contexts (such as the one hinted at in Remark 3.7), one would actually work with categories which are *opposite* to the category of commutative rings; cf. Example 1.9. A ring homomorphism $f : R \to S$ corresponds to a morphism $f^\circ : S^\circ \to R^\circ$ in the opposite category, and we can simply define ${f^\circ}_*$, etc., to be f_*, etc.

For morphisms $f^\circ : S^\circ \to R^\circ$ and $g^\circ : T^\circ \to S^\circ$ in the opposite category, prove that

- $(f^\circ \circ g^\circ)_* \cong {f^\circ}_* \circ {g^\circ}_*$,
- $(f^\circ \circ g^\circ)^* \cong {g^\circ}^* \circ {f^\circ}^*$,
- $(f^\circ \circ g^\circ)^! \cong {g^\circ}^! \circ {f^\circ}^!$,

where $\cong$ stands for 'naturally isomorphic'. (These are the formulas suggested by the notation: a $*$ in the subscript invariably suggests a basic 'covariance' property of the notation, while modifiers in the superscript usually suggest contravariance. The switch to the opposite category is natural in the algebro-geometric context.)

3.13. Let $p > 0$ be a prime integer, and let $\pi : \mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$ be the natural projection. Compute $\pi^*(A)$ and $\pi^!(A)$ for all finitely generated abelian groups A, as a vector space over $\mathbb{Z}/p\mathbb{Z}$. Compute $\iota^*(A)$ and $\iota^!(A)$ for all finitely generated abelian groups A, where $\iota : \mathbb{Z} \hookrightarrow \mathbb{Q}$ is the natural inclusion.

3.14. Let $f : R \to S$ be an *onto* ring homomorphism; thus, $S \cong R/I$ for some ideal I of R.

- Prove that, for all R-modules M, $f^!(M) \cong \{m \in M \mid \forall a \in I,\ am = 0\}$, while $f^*(M) \cong M/IM$. (Exercise III.7.7 may help.)
- Prove that, for all S-modules N, $f^! f_*(N) \cong N$ and $f^* f_*(N) \cong N$.
- Prove that f_* is fully faithful (Definition 1.6).
- Deduce that if there is an onto homomorphism $R \to S$, then S-Mod is equivalent to a full subcategory of R-Mod.

3.15. Let $f : R \to S$ be a ring homomorphism, and assume that the functor $f_* : S\text{-Mod} \to R\text{-Mod}$ is an equivalence of categories.

- Prove that there is a homomorphism of rings $\overline{g} : S \to \mathrm{End}_{\mathsf{Ab}}(R)$ such that the composition $R \to S \to \mathrm{End}_{\mathsf{Ab}}(R)$ is the homomorphism realizing R as a module over itself (that is, the homomorphism studied in Proposition III.2.7).
- Use the fact that S is commutative to deduce that $\overline{g}(S)$ is isomorphic to R. (Refine the result of Exercise III.2.17.) Deduce that f has a left-inverse $g : S \to R$.
- Therefore, $f_* \circ g_*$ is naturally isomorphic to the identity; in particular, $f_* \circ g_*(S) \cong S$ as an R-module. Prove that this implies that g is injective. (If $a \in \ker g$, prove that a is in the annihilator of $f_* \circ g_*(S)$.)
- Conclude that f is an isomorphism.

Two rings are *Morita equivalent* if their category of left-modules are equivalent. The result of this exercise is a (very) particular case of the fact that two *commutative* rings are Morita equivalent if and only if they are isomorphic. The commutativity is crucial in this statement: for example, it can be shown that any ring R is Morita equivalent to the ring of matrices[17] $\mathcal{M}_{n,n}(R)$, for all $n > 0$.

---

4. Multilinear algebra

4.1. Multilinear, symmetric, alternating maps. *Multilinear* maps may be defined similarly to bilinear maps: if $M_1, \ldots, M_\ell, P$ are R-modules, a function

$$\varphi : M_1 \times \cdots \times M_\ell \to P$$

is *R-multilinear* if it is R-linear in each factor, that is, if the function obtained by arbitrarily fixing all but the i-th component is R-linear in the i-th factor, for $i = 1, \ldots, r$.

Again it is natural to ask whether R-multilinear maps may be turned into R-linear maps: whether there exists an R-module $M_1 \otimes_R \cdots \otimes_R M_\ell$ through which every R-multilinear map must factor. Luckily, this module is already available to us:

Claim 4.1. *Every R-multilinear map $M_1 \times \cdots \times M_\ell \to P$ factors uniquely through*

$$((\cdots(M_1 \otimes_R M_2) \otimes_R \cdots) \otimes_R M_{\ell-1}) \otimes_R M_\ell.$$

[17] The author was once told that $\mathcal{M}_{n,n}(\mathbb{C})$ is 'not seriously noncommutative' since it is Morita equivalent to $\mathbb{C}$, which is commutative.

Indeed, argue inductively: if $\varphi : M_1 \times \cdots \times M_\ell \to P$ is multilinear, then for every $m_\ell \in M_\ell$, the function

$$(m_1, \ldots, m_{\ell-1}) \mapsto \varphi(m_1, \ldots, m_{\ell-1}, m_\ell)$$

is R-multilinear, so it factors through $(\cdots (M_1 \otimes_R M_2) \otimes_R \cdots) \otimes_R M_{\ell-1}$; therefore, φ induces a unique *bilinear* map

$$((\cdots (M_1 \otimes_R M_2) \otimes_R \cdots) \otimes_R M_{\ell-1}) \times M_\ell \to P,$$

and this induces a unique linear map

$$((\cdots (M_1 \otimes_R M_2) \otimes_R \cdots) \otimes_R M_{\ell-1}) \otimes_R M_\ell \to P$$

by the universal property of $\otimes_R$.

Of course the choice of pivoting on the last factor in this argument is arbitrary: any other way to associate the factors would produce a solution to the same universal problem. For $\ell = 3$, it follows in particular that there are canonical isomorphisms

$$(M_1 \otimes_R M_2) \otimes_R M_3 \cong M_1 \otimes_R (M_2 \otimes_R M_3)$$

for all R-modules M_1, M_2, M_3. In this sense, the tensor product is associative[18] in R-Mod, and we are indeed authorized to use the notation $M_1 \otimes_R \cdots \otimes_R M_\ell$. Elements of this module are finite linear combinations

$$\sum_i m_{1,i} \otimes \cdots \otimes m_{\ell,i}$$

of pure tensors. The structure map

$$M_1 \times \cdots \times M_\ell \to M_1 \otimes_R \cdots \otimes_R M_\ell$$

acts as $(m_1, \ldots, m_\ell) \mapsto m_1 \otimes \cdots \otimes m_\ell$. By the multilinearity of this map, we have, e.g.,

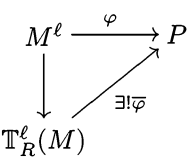
$$(rm_1) \otimes \cdots \otimes m_\ell = m_1 \otimes \cdots \otimes (rm_\ell) = r(m_1 \otimes \cdots \otimes m_\ell).$$

By convention, the tensor product over 0 factors is taken to be the base ring R.

Other important variations on the tensoring theme present themselves when all the factors coincide. We will use the shorthand notation

$$\mathbb{T}^\ell_R(M) := M^{\otimes \ell} = \underbrace{M \otimes_R \cdots \otimes_R M}_{\ell \text{ times}}$$

for the ℓ-fold tensor product of a module M by itself; this is called the (ℓ-th) *tensor power* of M. For every R-module P, every R-multilinear map $M^\ell \to P$ factors uniquely through $\mathbb{T}^\ell_R(M)$:



We set $\mathbb{T}^1_R(M) = M$; by our convention, $\mathbb{T}^0_R(M) = R$.

Now we can impose further restrictions on φ and again study the corresponding universal problems. Popular possibilities are

[18] In fact, the tensor product is associative in a fancier sense; cf. Exercise 3.3.

- φ may be required to be *symmetric;* that is, for all $\sigma \in S_\ell$ and all $m_1, \ldots, m_\ell$, require that
$$\varphi(m_{\sigma(1)}, \ldots, m_{\sigma(\ell)}) = \varphi(m_1, \ldots, m_\ell);$$
- or φ may be required to be *alternating;* that is,
$$\varphi(m_1, \ldots, m_\ell) = 0 \quad \text{whenever } m_i = m_j \text{ for some } i \neq j.$$

The reader may have expected the second prescription to read *for all* $\sigma \in S_\ell$ *and all* $m_1, \ldots, m_\ell$, *require that*

$$\varphi(m_{\sigma(1)}, \ldots, m_{\sigma(\ell)}) = (-1)^\sigma \varphi(m_1, \ldots, m_\ell).$$

The problem with this version is that if the base ring has characteristic 2, then there would be no difference between symmetric and alternating maps. But behold the following:

Lemma 4.2. *Let* $\varphi : M^\ell \to P$ *be an* R*-multilinear function.*

If φ *is alternating, then for all* $\sigma \in S_\ell$, *and all* $m_1, \ldots, m_\ell$,

$$\varphi(m_{\sigma(1)}, \ldots, m_{\sigma(\ell)}) = (-1)^\sigma \varphi(m_1, \ldots, m_\ell).$$

If 2 *is a unit in* R, *the converse holds as well.*

Proof. For the first statement, it suffices to show that interchanging any two factors switches the sign of an alternating function (since transpositions generate the symmetric group). Since the other factors have no effect on this operation, this reduces the question to the case $\ell = 2$. Therefore, we only have to show that if $\varphi(m, m) = 0$ for all $m \in M$, then

$$\varphi(m_2, m_1) = -\varphi(m_1, m_2)$$

for all $m_1, m_2 \in M$. For this, use bilinearity to expand $0 = \varphi(m_1 + m_2, m_1 + m_2)$, and again use the alternating condition.

For the second statement, again we can reduce to the $\ell = 2$ case. For $m_1 = m_2 = m$, $\varphi(m_2, m_1) = -\varphi(m_1, m_2)$ says $\varphi(m, m) = -\varphi(m, m)$; therefore

$$2\varphi(m, m) = 0.$$

If 2 is a unit in R, this implies $\varphi(m, m) = 0$, as needed. □

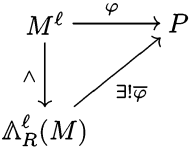
Thus, alternating maps do satisfy the (perhaps) more natural prescription, and this in fact characterizes alternating maps in most cases.

4.2. Symmetric and exterior powers. The symmetric, resp., alternating, conditions come (of course) with companion universal objects: modules $\mathbb{S}^\ell_R(M)$, $\mathbb{\Lambda}^\ell_R(M)$ (*symmetric* and *exterior* power[19], respectively). The construction of these modules follows patterns with which the reader is hopefully thoroughly familiar by now. For example, let $W \subseteq \mathbb{T}^\ell_R(M)$ be the submodule generated by all pure tensors $m_1 \otimes \cdots \otimes m_\ell$ such that $m_i = m_j$ for some $i \neq j$; then

$$\mathbb{\Lambda}^\ell_R(M) := \frac{\mathbb{T}^\ell_R(M)}{W}$$

[19]The module $\mathbb{\Lambda}^\ell_R(M)$ is also called the ℓ-th *wedge* power of M.

satisfies the expected universal property for alternating maps, i.e., every multilinear, alternating map $\varphi : M^\ell \to P$ induces a unique R-linear map $\overline{\varphi}$:



Here, the map $\wedge$ is of course the composition

$$M \times \cdots \times M \to M \otimes \cdots \otimes M \to \frac{M \otimes \cdots \otimes M}{W} = \bigwedge^\ell_R(M),$$

and W is the smallest submodule making the multilinear map $\wedge$ *alternating.* The image of a pure tensor is denoted by using $\wedge$ rather than $\otimes$:

$$m_1 \otimes \cdots \otimes m_\ell \mapsto m_1 \wedge \cdots \wedge m_\ell.$$

The reader will verify (Exercise 4.1) that $\bigwedge^\ell_R(M)$ may indeed be constructed as we have just claimed, and the reader will produce an analogous construction for the symmetric power $\mathbb{S}^\ell_R(M)$.

The module $\bigwedge^\ell_R(M)$ is generated by the pure 'alternating' tensors. By Lemma 4.2,

$$m_2 \wedge m_1 \wedge m_3 = -m_1 \wedge m_2 \wedge m_3$$

(for example), and, of course, $m_1 \wedge \cdots \wedge m_\ell = 0$ if two of the m_i's coincide. It follows that if $e_1, \ldots, e_r$ generate M, then $\bigwedge^\ell_R(M)$ is generated by all

$$e_{i_1} \wedge \cdots \wedge e_{i_\ell}$$

with $1 \le i_1 < \cdots < i_\ell \le r$. In particular, if M is finitely generated, then $\bigwedge^\ell_R(M) = 0$ for $\ell \gg 0$. If M is *free,* we can be very precise:

Lemma 4.3. *Let R be a commutative ring, and let M be a free R-module of rank r. Then $\bigwedge^\ell_R(M)$ is a free R-module of rank $\binom{r}{\ell}$.*

Proof. There are $\binom{r}{\ell}$ sequences of indices $i_1, \ldots, i_\ell$ satisfying $1 \le i_1 < \cdots < i_\ell \le r$, so we just need to show that the generators $e_{i_1} \wedge \cdots \wedge e_{i_\ell}$ are linearly independent.

For a fixed $I = (i_1, \ldots, i_\ell)$ with $1 \le i_1 < \cdots < i_\ell \le r$, consider the map

$$\varphi_I : M^\ell \to R$$

obtained by setting

$$\varphi_I(e_{j_1}, \ldots, e_{j_\ell}) = \begin{cases} (-1)^\sigma & \text{if } \exists \text{ a permutation } \sigma \text{ such that } \sigma(j_k) = i_k, \text{ all } k, \\ 0 & \text{otherwise} \end{cases}$$

and extending by multilinearity. Note that σ is unique if it exists, so the prescribed value of φ_I is well-defined. Also note that since M is free, every element of M is expressed *uniquely* as a combination of the e_i's; hence φ_I is well-defined for all elements of M^ℓ. Further, φ_I is evidently alternating (Exercise 4.5). Thus, it factors uniquely through an R-linear map

$$\overline{\varphi_I} : \bigwedge^\ell_R \to R,$$

with the property that

$$\overline{\varphi_I}(e_{j_1} \wedge \cdots \wedge e_{j_\ell}) = \begin{cases} (-1)^\sigma & \text{if } \exists \text{ a permutation } \sigma \text{ such that } \sigma(j_k) = i_k, \text{ all } k, \\ 0 & \text{otherwise.} \end{cases}$$

Now assume that

$$\sum_{1 \le i_1 < \cdots < i_\ell \le r} \lambda_{i_1 \ldots i_\ell} e_{i_1} \wedge \cdots \wedge e_{i_\ell} = 0;$$

then with $I = (i_1, \ldots, i_\ell)$

$$\lambda_{i_1 \ldots i_\ell} = \overline{\varphi_I} \left(\sum_{1 \le i_1 < \cdots < i_\ell \le r} \lambda_{i_1 \ldots i_\ell} e_{i_1} \wedge \cdots \wedge e_{i_\ell} \right) = \overline{\varphi_I}(0) = 0.$$

This shows that there are no nontrivial linear dependence relations among the generators $e_{i_1} \wedge \cdots \wedge e_{i_\ell}$, as claimed. □

Note that, in particular, it follows that if M is a free R-module of rank r, then

$$\mathbb{\Lambda}^r_R M \cong R;$$

an isomorphism is induced by the map

$$\varphi_{1\ldots r} : M^r \to R$$

obtained by setting

$$\varphi_{1\ldots r}(e_{i_1}, \ldots, e_{i_r}) = \begin{cases} \pm 1 & \text{if the indices } i_1, \ldots, i_r \text{ are distinct,} \\ 0 & \text{otherwise,} \end{cases}$$

where the sign is determined by the permutation ordering the indices. (This is a particular case of the maps φ_I considered in the proof of Lemma 4.3.)

Claim 4.4. *Let A be an $r \times r$ matrix with entries in a field k, and let $a_i \in k^r$ be the i-th column of A. Then with notation as above*

$$\det(A) = \varphi_{1\ldots r}(a_1, \ldots, a_r).$$

This is immediate, since the standard properties of the determinant (obtained in §VI.3.2) prove that it is a multilinear, alternating map of the columns and the determinant of the identity matrix is 1. For this reason, the *top exterior power* of a free module F (that is, $\mathbb{\Lambda}^r_R(F)$ if F has rank r) is called the *determinant* of F, denoted $\det F$.

More generally, the functions φ_I give an explicit isomorphism between $\mathbb{\Lambda}^\ell_k R^r$ and $R^{\binom{r}{\ell}}$.

Example 4.5. For $V = k^4$, $\mathbb{\Lambda}^2_k V$ has dimension $\binom{4}{2} = 6$. On 'pure wedges' $a_1 \wedge a_2$, the isomorphism $\mathbb{\Lambda}^2_k V \to k^6$ works as follows. View the vectors a_1, a_2 as the two columns of a 4×2 matrix A; then send $a_1 \wedge a_2$ to the collection of the six 2×2

minors of A:

$$A = \begin{pmatrix} a_1^1 & a_2^1 \\ a_1^2 & a_2^2 \\ a_1^3 & a_2^3 \\ a_1^4 & a_2^4 \end{pmatrix} \mapsto \begin{pmatrix} a_1^1 a_2^2 - a_1^2 a_2^1 \\ a_1^1 a_2^3 - a_1^3 a_2^1 \\ a_1^1 a_2^4 - a_1^4 a_2^1 \\ a_1^2 a_2^3 - a_1^3 a_2^2 \\ a_1^2 a_2^4 - a_1^4 a_2^2 \\ a_1^3 a_2^4 - a_1^4 a_2^3 \end{pmatrix}.$$

This definition extends by linearity to the whole of $\bigwedge_k^2 V$. ⌟

Similarly to the alternating case, if $e_1, \dots, e_\ell$ generate M, then the *symmetric* power $\mathbb{S}_R^i(M)$ is generated by the (images[20] of) all tensors

$$e_1 \otimes \cdots \otimes e_\ell$$

with $1 \le i_1 \le \cdots \le i_\ell \le r$. In this case all $\mathbb{S}_R^\ell(M)$ are, in general, nonzero.

4.3. Very small detour: Graded algebra. We have obtained tensor, symmetric, exterior powers of a module M for every nonnegative integer ℓ. As it happens, in each case it is useful to consider 'all ℓ at once'. The structures we obtain by doing this are particular cases of a very important notion, which would deserve much more space than we can allow.

A *graded ring* is a (not necessarily commutative) ring S endowed with a decomposition

$$S = \bigoplus_{i \ge 0} S_i$$

of the *abelian group* $(S, +)$ into a direct sum of abelian groups S_i, for nonnegative integers[21] i, such that

$$S_i \cdot S_j \subseteq S_{i+j}.$$

This prescription is parsed as follows: identify S_i with its image in S by the natural map; nonzero elements of S_i are called *homogeneous*, of *degree* i; the condition prescribes that the product of two homogeneous elements of degree i, j is homogeneous, of degree $i + j$ (provided it is not 0).

Example 4.6. The polynomial ring $R[x_1, \dots, x_n]$ over any ring R carries a natural grading, given by the (ordinary) degree of polynomials. We may write

$$R[x_1, \dots, x_n] = R \oplus \langle x_1, \dots, x_n \rangle \oplus \langle x_1^2, x_1 x_2, \dots, x_n^2 \rangle \oplus \cdots. \quad ⌟$$

The reader should have no difficulty providing the notions of graded *modules* and *algebras*. For example, a graded ring $S = \bigoplus_i S_i$ is a graded algebra over a graded ring $R = \bigoplus_i R_i$ if it carries an action of R (making it into a 'conventional' R-algebra) and further $R_i \cdot S_j \subseteq S_{i+j}$. In particular, if $R = R_0$ (so that R is a commutative ring viewed as a graded ring by 'concentrating' it in degree 0), we are just requiring each graded piece of S to be an R-module.

[20] We are not aware of a universally accepted notation to denote 'symmetric' tensors; a sensible choice would be a simple $\cdot$, as in the special case of polynomial rings.

[21] Of course, much more general gradings may be considered, but this is the case that will occur in the applications in this section.

For a concrete example, the commutative ring $R[x_1, \dots, x_n]$ is a graded R-algebra, generated (as an algebra over R) by the piece of degree 1. This is an important template for many interesting situations.

We cannot resist the temptation to mention another example from commutative algebra:

Example 4.7. Let R be a commutative ring, and let $I \subseteq R$ be an ideal. Then the direct sum of R-modules

$$\bigoplus_{\ell \geq 0} I^\ell = R \oplus I \oplus I^2 \oplus I^3 \oplus \cdots$$

has a natural ring structure, since $I^i \cdot I^j \subseteq I^{i+j}$. The resulting graded R-algebra is called the *Rees algebra* of I, $\mathrm{Rees}_R(I)$, and is of fundamental importance in algebraic geometry (as it translates into the notion of *blow-up,* a particularly important class of regular maps). The brave reader will try to explore the Rees algebra of the ideal $I = (x, y)$ in the polynomial ring $k[x, y]$ over a field k (Exercise 4.21), which is the simplest interesting instance of this construction (well, the second simplest: the simplest example may be Exercise III.5.17). ⌟

The additional information carried by the grading on a ring is very substantial, and the theory of graded rings and algebras is extremely rich. For example, *projective* algebraic geometry may be developed by using graded rings along the same lines we sketched in §VII.2.3 for the affine case. As we mentioned at the end of §VII.2.3, this provides a 'globalization' procedure which leads to important technical advantages.

For the (very modest) purpose of this section, the following general remarks are all we need.

Graded rings form a category under ring homomorphisms, but they form a more interesting one if we require the homomorphisms to preserve the grading; that is, if $S = \bigoplus_i S_i$ and $T = \bigoplus_i T_i$ are graded rings, consider ring homomorphisms $\varphi : S \to T$ such that $\varphi(S_i) \subseteq T_i$. Such homomorphisms are called *graded* homomorphisms. Of course, singling out a special class of homomorphisms will then single out a special class of ideals, that is, those ideals which appear as kernels of graded homomorphisms.

Definition 4.8. An ideal I of a graded ring $S = \bigoplus_i S_i$ is *homogeneous* if $I = \bigoplus_i (I \cap S_i)$. ⌟

Lemma 4.9. *Let $S = \bigoplus_i S_i$ be a graded ring and let $I \subseteq S$ be an ideal of S. Then the following are equivalent:*

(i) *I is homogeneous;*

(ii) *if $s \in S$ and $s = \sum_i s_i$ is the decomposition of s into* homogeneous *elements $s_i \in S_i$, then $s \in I \iff s_i \in I$ for all i;*

(iii) *I admits a generating set consisting of homogeneous elements;*

(iv) *I is the kernel of a graded homomorphism.*

Proof. (i) $\iff$ (ii) is the very definition of homogeneous ideal; (ii) $\iff$ (iii) is left to the reader (Exercise 4.10).

(ii) $\implies$ (iv): Assuming (ii) holds, define a grading on S/I by letting the piece of degree i consist of (0 and) the cosets of the elements of degree i in S. This is well-defined by (ii), and it is immediately checked that it induces a graded ring structure on S/I. The ideal I is then the kernel of the graded homomorphism $S \to S/I$, verifying (iv).

(iv) $\implies$ (ii): Let $\varphi : S \to T$ be a graded homomorphism, and assume $I = \ker\varphi$. Let $s = \sum_i s_i \in \ker\varphi$, with $s_i \in S_i$. Then $\sum_i \varphi(s_i) = 0$ in T, and $\varphi_i(s_i)$ is homogeneous of degree i for all i. It follows that $\varphi(s_i) = 0$ for all i, that is, each $a_i \in \ker\varphi = I$, implying (ii). □

Note that a graded homomorphism $\varphi : S \to T$ induces a homomorphism of abelian groups $\varphi_i : S_i \to T_i$ for each i. The ideal $\ker\varphi$ is then the direct sum of all $\ker\varphi_i$.

Example 4.10. The ideal $I = (y - x^2)$ is not homogeneous in the ring $k[x, y]$, if this is given the grading by the usual degree: indeed, $y - x^2 \in I$ while $y \notin I$, contradicting condition (ii) of Lemma 4.9. The ideal $(y, y - x^2)$ *is* homogeneous, since it equals (y, x^2) and both y, x^2 are homogeneous.

The conventional grading of a polynomial ring is not the only option: we could decide to grade $k[x, y]$ by placing constants in degree 0 and assigning degree 2 to the indeterminate x and degree 1 to y. With such a grading, the ideal $(y - x^2)$ *is* homogeneous. ⌟

Projective algebraic geometry (almost) follows the blueprint of affine algebraic geometry reviewed in §VII.2.3, but using only *homogeneous* ideals in $k[x_1, \dots, x_n]$ (taken with the usual grading).

4.4. Tensor algebras. Going back to the context of multilinear algebra, consider again a module M over a commutative ring R. We define the *tensor algebra* of M as the graded R-algebra

$$\mathbb{T}^*_R(M) := \bigoplus_{\ell \geq 0} \mathbb{T}^\ell_R(M) :$$

the multiplication is defined on pure tensors by

$$(m_1 \otimes \cdots \otimes m_i) \cdot (n_1 \otimes \cdots \otimes n_j) := m_1 \otimes \cdots \otimes m_i \otimes n_1 \otimes \cdots \otimes n_j$$

and is extended by linearity. Similarly, we can define the *symmetric algebra* and the *exterior algebra* of M:

$$\mathbb{S}^*_R(M) := \bigoplus_{\ell \geq 0} \mathbb{S}^\ell_R(M), \quad \mathbb{\Lambda}^*_R(M) := \bigoplus_{\ell \geq 0} \mathbb{\Lambda}^\ell_R(M).$$

Remark 4.11. The tensor algebra is not commutative; the symmetric algebra *is* commutative; and the exterior algebra is 'skew-commutative' in the sense that if $\alpha \in \mathbb{\Lambda}^i_R(M)$ and $\beta \in \mathbb{\Lambda}^j_R(M)$, then

$$\alpha \wedge \beta = (-1)^{ij} \beta \wedge \alpha$$

(Exercise 4.16), where we use $\wedge$ to denote the operation in $\mathbb{\Lambda}^*_R(M)$. The reader should recognize this formula from the theory of differential forms; differential forms provide perhaps the most important example of an exterior algebra. ⌟

The definitions of symmetric and exterior powers were concocted so as to yield surjective *graded* homomorphisms of *algebras*

$$\mathbb{T}^*_R(M) \twoheadrightarrow \mathbb{S}^*_R(M),$$
$$\mathbb{T}^*_R(M) \twoheadrightarrow \mathbb{\Lambda}^*_R(M).$$

As we have learned in Lemma 4.9, the kernels of these homomorphisms are certain homogeneous ideals of the tensor algebra; these ideals must then be generated by homogeneous elements.

Lemma 4.12. *Let $I_{\mathbb{S}}, I_{\mathbb{\Lambda}} \subseteq \mathbb{T}^*_R(M)$ be the ideals respectively generated by all elements of the form $(m \otimes n - n \otimes m)$ as $m, n \in M$ and by elements of the form $m \otimes m$ as $m \in M$. Then*

$$\mathbb{S}^*_R(M) \cong \frac{\mathbb{T}^*_R(M)}{I_{\mathbb{S}}}, \quad \mathbb{\Lambda}^*_R(M) \cong \frac{\mathbb{T}^*_R(M)}{I_{\mathbb{\Lambda}}}$$

as graded R-algebras.

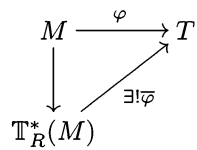
Proof. We have observed in §4.3 that the kernel of a graded homomorphism is the direct sum of the kernels of the induced homomorphisms in each degree; the statement of the lemma then follows easily from the explicit descriptions of the kernels of the canonical projections from the tensor powers to the symmetric and exterior powers, given in §4.2. □

Remark 4.13. In some cases it is also possible to construct subalgebras of $\mathbb{T}^*_R(M)$ isomorphic to $\mathbb{S}^*_R(M)$ and $\mathbb{\Lambda}^*_R(M)$. For example, tensors that are invariant under the action of the symmetric group may be used to give a copy of $\mathbb{S}^*_R(M)$ inside $\mathbb{T}^*_R(M)$: $\frac{1}{2}(m_1 \otimes m_2 + m_2 \otimes m_1)$ would be such a 'symmetric tensor'. However, the introduction of denominators poses distasteful restrictions on the base ring R, and for this reason we prefer the 'quotient' constructions of Lemma 4.12. ⌟

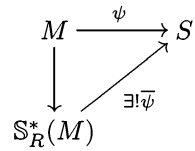
It is now straightforward to establish universal properties satisfied by the algebras defined above. Note that by construction there are canonical isomorphisms of M with $\mathbb{T}^1_R(M)$, $\mathbb{S}^1_R(M)$, $\mathbb{\Lambda}^1_R(M)$, and in particular there are canonical maps

$$M \to \mathbb{T}^*_R(M), \quad M \to \mathbb{S}^*_R(M), \quad M \to \mathbb{\Lambda}^*_R(M).$$

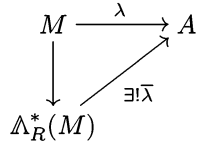
Proposition 4.14. *Let R be a commutative ring, and let M be an R-module. Then for every R-algebra T and every R-module homomorphism $\varphi : M \to T$ there exists a unique homomorphism of R-algebras $\overline{\varphi} : \mathbb{T}^*_R(M) \to T$ making the following diagram commute:*



Proposition 4.15. *Let R be a commutative ring, and let M be an R-module. Then for every* commutative *R-algebra S and every R-module homomorphism $\psi : M \to S$ there exists a unique homomorphism of R-algebras $\overline{\psi} : \mathbb{S}_R^*(M) \to S$ making the following diagram commute:*



Proposition 4.16. *Let R be a commutative ring, and let M be an R-module. Then for every R-algebra A and every R-module homomorphism $\lambda : M \to A$ such that $\lambda(m)^2 = 0\ \forall m \in M$, there exists a unique homomorphism of R-algebras $\overline{\lambda} : \Lambda_R^*(M) \to A$ making the following diagram commute:*



Details may now safely be left to the reader (who may for example establish the first proposition from the universal property of tensor powers and then deduce the second and third from Lemma 4.12).

Example 4.17. The free case is particularly easy to understand. For instance,

$$\mathbb{S}_R^*(R^{\oplus r}) \cong R[x_1, \ldots, x_r].$$

Indeed, the polynomial ring satisfies the appropriate universal property with respect to mapping to commutative rings (cf. §III.2.2 and §III.6.4). Likewise, $\mathbb{T}_R^*(R^{\oplus r})$ should be thought of as a 'noncommutative' polynomial ring, in which the r indeterminates do not commute with each other (cf. §III.6.3). ⌟

Of course the constructions $\mathbb{T}_R^*$, $\mathbb{S}_R^*$, Λ_R^* are all functorial, and their behavior with respect to exact sequences is interesting and important. For example, suppose

$$0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$$

is an exact sequence of R-modules. Then L may be identified with a subset of $M \cong \mathbb{S}_R^1(M)$; hence it defines an ideal $L \cdot \mathbb{S}_R^*(M)$ of the algebra $\mathbb{S}_R^*(M)$. It is not hard to show that the sequence[22]

$$0 \longrightarrow L \cdot \mathbb{S}_R^{*-1}(M) \longrightarrow \mathbb{S}_R^*(M) \longrightarrow \mathbb{S}_R^*(N) \longrightarrow 0$$

is exact. This is often useful in computations (cf. Exercise 4.21). We will leave any further such explorations to the more motivated readers.

[22]The shift $\mathbb{S}^{*-1}$ is introduced in order to preserve degrees in this sequence.

Exercises

In the following exercises, R denotes a commutative ring and k denotes a field.

4.1. ▷ Verify that the module $\Lambda^\ell_R(M)$ constructed in §4.2 does satisfy the universal property for alternating multilinear maps. Construct a module $\mathbb{S}^\ell_R(M)$ satisfying the universal property for symmetric multilinear maps. [§4.2]

4.2. Define the action of $\mathbb{T}^\ell_R$, $\mathbb{S}^\ell_R$, Λ^ℓ_R on R-linear maps, making covariant functors out of these notions.

4.3. Let I be the ideal (x, y) in $k[x, y]$; so every element of I may be written (of course, not uniquely) in the form $fx+gy$ for some polynomials $f, g \in k[x, y]$. Define a function $\varphi : I \times I \to k$ by prescribing

$$\varphi(f_1 x + g_1 y, f_2 x + g_2 y) := f_1(0,0)g_2(0,0) - f_2(0,0)g_1(0,0).$$

- Prove that φ is well-defined.
- Prove that φ is $k[x, y]$-bilinear and alternating.
- Prove that $I\Lambda_{k[x,y]}I \neq 0$.

Note that I has rank 1 as a $k[x, y]$-module; if it were free, its wedge with itself would have to vanish.

4.4. Let F_1 and F_2 be free R-modules of finite rank.

- Construct a 'meaningful' isomorphism $\det(F_1) \otimes \det(F_2) \cong \det(F_1 \oplus F_2)$.
- More generally, prove that

$$\Lambda^r_R(F_1 \oplus F_2) \cong \bigoplus_{i+j=r} (\Lambda^i_R F_1) \otimes_R (\Lambda^j_R F_2).$$

4.5. ▷ Verify that the multilinear map φ_I defined in the proof of Lemma 4.3 is alternating. [§4.2]

4.6. Let V be a vector space, and let $v_1, \ldots, v_\ell \in V$. Prove that $v_1, \ldots, v_\ell$ are linearly independent if and only if $v_1 \wedge \cdots \wedge v_\ell \neq 0$.

4.7. ¬ Let V be a k-vector space, and let $\{v_1, \ldots, v_\ell\}$, $\{w_1, \ldots, w_\ell\}$ be two sets of linearly independent vectors in V. Prove that $\{v_1, \ldots, v_\ell\}$, $\{w_1, \ldots, w_\ell\}$ span the same subspace of V if and only if $v_1 \wedge \cdots \wedge v_\ell$ and $w_1 \wedge \cdots \wedge w_\ell$ are nonzero multiples of each other in $\Lambda^\ell_k V$. (For the interesting direction, if $\langle v_1, \ldots, v_\ell\rangle \neq \langle w_1, \ldots, w_\ell\rangle$, there must be a vector u belonging to the first subspace but not to the second. What can you say about $(v_1 \wedge \cdots \wedge v_\ell) \wedge u$ and $(w_1 \wedge \cdots \wedge w_\ell) \wedge u$ in $\Lambda^*_k V$?)

Deduce that there is an injective function from the Grassmannian of ℓ-dimensional subspaces of V (Exercise VI.2.13) to the projective space $\mathbb{P}(\Lambda^\ell_k V)$: the Grassmannian is identified with the set of 'pure wedges' in the projectivization of the exterior power $\Lambda^\ell_k V$. This is called the *Plücker embedding* of the Grassmannian. [4.8]

4.8. The Plücker embedding described in Exercise 4.7 realizes the Grassmannian $\mathrm{Gr}_k(2,4)$ of 2-dimensional subspaces of k^4 as a subset of the projectivization of $\mathbb{\Lambda}^2_k k^4 \cong k^6$: $\mathrm{Gr}_k(2,4) \subseteq \mathbb{P}^5_k$. Choose projective coordinates (Exercise VII.2.20) $(x_{12} : x_{13} : x_{14} : x_{23} : x_{24} : x_{34})$ in $\mathbb{P}(\mathbb{\Lambda}^2_k k^4)$, listed according to the corresponding minors, as in Example 4.5. Prove that $\mathrm{Gr}_k(2,4)$ is the locus of zeros

$$\mathscr{V}(x_{12}x_{34} - x_{13}x_{24} + x_{14}x_{23})$$

(notation as in Exercise VII.2.21). (Remember that you know how to write every point of $\mathrm{Gr}_k(2,4)$: look back at Exercise VI.2.14.)

Thus, $\mathrm{Gr}_k(2,4)$ may be viewed as a projective algebraic set. In fact, all Grassmannians may be similarly realized as projective algebraic sets (in the sense of Exercise 4.11) via the corresponding Plücker embeddings.

Prove that $\mathrm{Gr}_k(2,4)$ may be covered with six copies of $\mathbb{A}^4_k$. (Recall from Exercise VII.2.20 that $\mathbb{P}^5_k$ may be covered with six copies of $\mathbb{A}^5_k$; prove that the intersection of each with $\mathrm{Gr}_k(2,4)$ may be identified with $\mathbb{A}^4_k$ in a natural way.)

4.9. Assume 2 is a unit in R, and let F be a free R-module of finite rank.

- Define a function $\lambda : \mathbb{\Lambda}^2_R(F) \to \mathbb{T}^2_R(F)$ on a basis $e_i \wedge e_j$, $i < j$, by setting $\lambda(e_i \wedge e_j) = \frac{1}{2}(e_i \otimes e_j - e_j \otimes e_i)$ and extending by linearity. Prove that λ is an injective homomorphism of R-modules and $\lambda(f_1 \wedge f_2) = \frac{1}{2}(f_1 \otimes f_2 - f_2 \otimes f_1)$ for all $f_1, f_2 \in F$.
- Define a function $\sigma : \mathbb{S}^2_R(F) \to \mathbb{T}^2_R(F)$ on a basis $e_i \otimes e_j$, $i \leq j$, by setting $\sigma(e_i \otimes e_j) = \frac{1}{2}(e_i \otimes e_j + e_j \otimes e_i)$ and extending by linearity. Prove that σ is an injective homomorphism of R-modules and $\sigma(f_1 \otimes f_2) = \frac{1}{2}(f_1 \otimes f_2 + f_2 \otimes f_1)$ for all $f_1, f_2 \in F$.
- Prove that λ identifies $\mathbb{\Lambda}^2_R(F)$ with the kernel of the map $\mathbb{T}^2_R(F) \to \mathbb{S}^2_R(F)$ and σ identifies $\mathbb{S}^2_R(F)$ with the kernel of the map $\mathbb{T}^2_R(F) \to \mathbb{\Lambda}^2_R(F)$.

In particular, there is a 'meaningful' isomorphism $F \otimes_R F \cong \mathbb{\Lambda}^2_R(F) \oplus \mathbb{S}^2_R(F)$.

4.10. ▷ Prove the equivalence (ii) $\Longleftrightarrow$ (iii) in Lemma 4.9. [§4.3]

4.11. ¬ Let $I \subseteq k[x_0, \ldots, x_n]$ be a *homogeneous* ideal. Prove that the condition '$F(c_0, \ldots, c_n) = 0$ for all $F \in I$' for a point $(c_0 : \ldots : c_n) \in \mathbb{P}^n_k$ is well-defined: it does not depend on the representative $(c_0, \ldots, c_n)$ chosen for the point $(c_0 : \ldots : c_n)$. (Cf. Exercise VII.2.21, but note that not all F in I are homogeneous.)

Thus, every homogeneous ideal $I \subseteq k[x_0, \ldots, x_n]$ determines a 'projective algebraic set'

$$\mathscr{V}(I) := \{(c_0 : \ldots : c_n) \in \mathbb{P}^n_K \mid (\forall F \in I)\,,\, F(c_0, \ldots, c_n) = 0\}.$$

Note that $\mathscr{V}((x_0, \ldots, x_n)) = \emptyset$. The ideal $(x_0, \ldots, x_n) = \bigoplus_{i>0} k[x_0, \ldots, x_n]_i$ is irreverently called the *irrelevant ideal.* [4.8, 4.12]

4.12. (Cf. Exercise 4.11.) Prove the 'weak homogeneous Nullstellensatz': if k is algebraically closed and $I \subseteq k[x_0, \ldots, x_n]$ is an ideal, then $\mathscr{V}(I) = \emptyset$ if and only if $\sqrt{I}$ is either $k[x_0, \ldots, x_n]$ or the irrelevant ideal $(x_0, \ldots, x_n)$. (Translate this into a question about $\mathbb{A}^{n+1}_k$.)

4.13. Let k be a field of characteristic zero. A *differential operator* in one variable, with polynomial coefficients, is a linear combination

$$(*) \qquad a_0(x) + a_1(x)\partial_x + \cdots + a_r(x)\partial_x^r$$

where $a_i(x) \in k[x]$. Here x acts on a polynomial $f(x) \in k[x]$ by multiplication by x, while ∂_x acts by taking a formal derivative (as in §VII.4.2). With the evident operations, differential operators form a ring, called the (first) *Weyl algebra.* Note that the Weyl algebra is noncommutative: for example,

$$(x\partial_x)f(x) = xf'(x),$$

while

$$(\partial_x x)f(x) = \partial_x(xf(x)) = f(x) + xf'(x).$$

Therefore, $(\partial_x x - x\partial_x)f(x) = f(x)$, or put otherwise

$$(**) \qquad \partial_x x - x\partial_x = 1$$

in the Weyl algebra. Prove that the Weyl algebra is isomorphic to the quotient $\mathbb{T}_k^*(\langle x, y\rangle)/(yx - xy - 1)$. (Use (**) to write any element of the Weyl algebra in the form given in (*). Show that this representation is unique, and deduce that (**) generates all relations among x and ∂_x. Then use the first isomorphism theorem.)

The relation (**) expresses (up to a factor) the basic fact that in quantum mechanics the position and momentum operators do not commute. The Weyl algebra was introduced to study this phenomenon. Modules over rings of differential operators are called *D-modules.*

4.14. Let F be a free R-module of rank r. Prove that $\mathbb{S}_R^\ell(F)$ is free, and compute its rank.

4.15. Let F_1, F_2 be free R-modules of finite rank. Prove that $\mathbb{S}_R^*(F_1 \oplus F_2) \cong \mathbb{S}_R^*(F_1) \otimes_R \mathbb{S}_R^*(F_2)$.

4.16. ▷ Verify the skew commutativity of the exterior algebra, stated in Remark 4.11. [§4.4]

4.17. Let V be a k-vector space of dimension r. Prove that, as a vector space, the exterior algebra $\mathbb{\Lambda}_k^*(V)$ has dimension 2^r.

4.18. ¬ Prove that the Rees algebra $\mathrm{Rees}_R(I)$ of an ideal I is Noetherian if R is Noetherian. [4.19]

4.19. ¬ Let R be a Noetherian ring, I an ideal of R, and consider the Rees algebra $\mathrm{Rees}_R(I) = \bigoplus_{\ell \geq 0} I^\ell$. By Exercise 4.18, $\mathrm{Rees}_R(I)$ is Noetherian.

- For $\ell \geq 0$, let $J_\ell \subseteq I^\ell$ be ideals of R, and view $J := \bigoplus_{\ell \geq 0} J_\ell$ as a sub-R-module of $\mathrm{Rees}_R(I)$. Prove that J is an ideal of $\mathrm{Rees}_R(I)$ if and only if $I^n J_\ell \subseteq J_{\ell+n}$ for all $\ell, n \geq 0$.
- Assume $J := \bigoplus_{\ell \geq 0} J_\ell$ *is* an ideal of $\mathrm{Rees}_R(I)$. Prove that J admits a finite set of homogeneous generators.
- Choose a finite set of homogeneous generators for J, and let s be the largest degree of an element in this set. Prove that $J_{s+1} \subseteq I^{s+1}J_0 + I^s J_1 + \cdots + IJ_s$ (as ideals of R).

- Prove that $J_{s+\ell} = I^\ell J_s$ for all $\ell \geq 0$.

This is a particular case of the 'Artin-Rees' lemma. [4.20]

4.20. ¬ Let R be a Noetherian ring, and let I be an ideal of R. Prove that $I \cdot \bigcap_{n\geq 0} I^n = \bigcap_{n\geq 0} I^n$. (Hint: Exercise 4.19.) [1.18]

4.21. ▷ Let k be a field, and consider the ideal $I = (x, y)$ in the ring $R = k[x, y]$. Prove that the Rees algebra $\bigoplus_{j\geq 0} I^j$ is isomorphic to the quotient of a polynomial ring $k[x, y, s, t]$ by the ideal $(tx - sy)$. Deduce that, in this case, the Rees algebra of I is isomorphic to the symmetric algebra $\mathbb{S}^*_{k[x,y]}(I)$.

(For the last point, use the exact sequence mentioned at the end of §4.4. It will be helpful to have an explicit presentation of I as a $k[x, y]$-module; for this, looking back at Exercise VI.4.15 may help.) [§4.3, §4.4]

4.22. ¬ Let $\underline{a}$ denote a list $a_1, \ldots, a_n$ of elements of R, and let $F \cong R^n$. Rename $\bigwedge^r_R(F)$ by $K_r(\underline{a})$. Define R-module homomorphisms $d_r : K_r(\underline{a}) \to K_{r-1}(\underline{a})$ on bases by setting

$$d(e_{i_1} \wedge \cdots \wedge e_{i_r}) = \sum_{j=1}^{r} (-1)^{j-1} a_{i_j} e_{i_1} \wedge \cdots \wedge \hat{e_{i_j}} \wedge \cdots \wedge e_{i_r},$$

where the hat denotes that the hatted element is omitted.

- Prove that $d_{r-1} \circ d_r = 0$.

Thus, a collection $a_1, \ldots, a_n$ of elements of R determines a complex of R-modules

$$0 \longrightarrow K_n(\underline{a}) \xrightarrow{d_n} \cdots \xrightarrow{d_2} K_1(\underline{a}) \xrightarrow{d_1} K_0(\underline{a}) = R \longrightarrow R/I \longrightarrow 0,$$

where $I = (a_1, \ldots, a_n)$. This is called the *Koszul complex* of $\underline{a}$.

- Check that the complexes constructed in Exercises VI.4.13 and VI.4.14 are Koszul complexes.

As proven for $n = 2, 3$ in Exercises VI.4.13 and VI.4.14, the Koszul complex is exact if $(a_1, \ldots, a_n)$ is a *regular* sequence in R, providing a free resolution for R/I in that case. Try to prove this in general. [VI.4.14]

---

5. Hom and duals

Our rapid overview of tensor products has occasionally brought us into close proximity with the Hom functors, and we will close this chapter by devoting some attention to these functors. As in the case of tensors, we will be preoccupied with adjunction and exactness properties, since concrete computations depend heavily on these properties. We will deal with these properties more carefully in the next (and last) section; in this section we will concentrate on one important special case of Hom, that is, the *duality* functor.

As in the rest of the chapter, R denotes a fixed commutative ring.

5.1. Adjunction again. The careful reader will agree that almost everything we know about the tensor product follows from the fact that it is a left-adjoint functor. This fact is spelled out in Lemma 2.4, whose flip side gives us just as much information concerning the covariant flavor of the Hom functor:

$$L \mapsto \mathrm{Hom}_{R\text{-}\mathsf{Mod}}(N, L).$$

Explicitly, the following is an exact restatement of Corollary 2.5:

Corollary 5.1. *For every R-module N, the functor $\mathrm{Hom}_R(N, \_)$ is right-adjoint to the functor $\_ \otimes_R N$.*

What about the 'contravariant' flavor of Hom:

$$M \mapsto h_N(M) := \mathrm{Hom}_{R\text{-}\mathsf{Mod}}(M, N)$$

(this functor was discussed in general terms in §1.2)?

Proposition 5.2. *For every R-module N, the functor $\mathrm{Hom}_R(\_, N)$ is right-adjoint to itself.*

This statement should be parsed carefully, because $\mathrm{Hom}_R(\_, N)$ is a *contravariant* functor. The statement of Proposition 5.2 may lead us astray into thinking that $\mathrm{Hom}_R(\_, N)$ must also be its own left-adjoint, and this is *not* the case (indeed this would make it a right-exact functor, and we will soon see that $\mathrm{Hom}_R(\_, N)$ is not right-exact in general).

The point is that, by definition of contravariant functor, $h_N = \mathrm{Hom}_R(\_, N)$ should be viewed as a *covariant* functor *from* the opposite category:

$$h_N : R\text{-}\mathsf{Mod}^{op} \to R\text{-}\mathsf{Mod};$$

it can also be viewed as a covariant functor *to* the opposite category,

$$h_N^{op} : R\text{-}\mathsf{Mod} \to R\text{-}\mathsf{Mod}^{op},$$

by simply reversing arrows after the fact rather than before. A more precise statement of Proposition 5.2 is that

$$h_N \text{ is right-adjoint to } h_N^{op}.$$

Proof. Let L, M, N denote R-modules. Recall (cf. the considerations preceding Lemma 2.4) that R-bilinear maps

$$\varphi : L \times M \to N$$

may be identified with R-linear maps

$$L \to \mathrm{Hom}_R(M, N).$$

By the same token, they may be identified with R-linear maps

$$M \to \mathrm{Hom}_R(L, N) :$$

for fixed $m \in M$, the function $\varphi(\_, m)$ is an R-linear map $L \to N$. Tracing these two identifications gives a canonical bijection

$$(*) \qquad \mathrm{Hom}_R(L, \mathrm{Hom}_R(M, N)) \cong \mathrm{Hom}_R(M, \mathrm{Hom}_R(L, N)).$$

We are adhering to the convention that Hom_R stands for $\mathrm{Hom}_{R\text{-Mod}}$. We may view it just as well as $\mathrm{Hom}_{R\text{-Mod}^{op}}$, provided that we reverse arrows; thus, the canonical bijection in (*) may be rewritten as

$$\mathrm{Hom}_{R\text{-Mod}}(L, \mathrm{Hom}_{R\text{-Mod}}(M,N)) \cong \mathrm{Hom}_{R\text{-Mod}^{op}}(\mathrm{Hom}_{R\text{-Mod}^{op}}(N,L), M)$$

or, using the notation introduced before the proof, as

$$\mathrm{Hom}_{R\text{-Mod}}(L, h_N(M)) \cong \mathrm{Hom}_{R\text{-Mod}^{op}}(h_N^{op}(L), M).$$

This says that h_N is right-adjoint to h_N^{op} (the naturality requirement is also immediate, and as usual it is left to the reader). □

As both $\mathrm{Hom}_R(N, \_)$ (by Corollary 5.1) and $\mathrm{Hom}_R(\_, N)$ (by Proposition 5.2) are right-adjoint functors, Lemma 1.17 tells us they commute with limits. This fact must also be parsed carefully, because of the contravariant nature of $\mathrm{Hom}_R(\_, N)$: for example, *products* are limits in R-Mod, but *direct sums* are colimits in R-Mod, hence *limits* in R-Mod^{op}. Thus,

Corollary 5.3. *For every R-module N and every family $\{M_i\}_{i\in I}$ of R-modules,*

$$\mathrm{Hom}_R(N, \prod_{i\in I} M_i) \cong \prod_{i\in I} \mathrm{Hom}_R(N, M_i),$$
$$\mathrm{Hom}_R(\bigoplus_{i\in I} M_i, N) \cong \prod_{i\in I} \mathrm{Hom}_R(M_i, N).$$

Also, as pointed out in Claim 1.19, the fact that both the covariant and contravariant flavors of Hom are right-adjoints has immediate implications for their exactness, adding yet another Pavlovian statement to the list:

$$\mathrm{Hom}_R(M, \_) \text{ and } \mathrm{Hom}_R(\_, N) \text{ are left-exact functors.}$$

By the contravariant nature of $\mathrm{Hom}_R(\_, N)$, the left-exactness of the latter means that if

$$A \longrightarrow B \longrightarrow C \longrightarrow 0$$

is an exact sequence of R-modules, then the induced sequence

$$0 \longrightarrow \mathrm{Hom}_R(C,N) \longrightarrow \mathrm{Hom}_R(B,N) \longrightarrow \mathrm{Hom}_R(A,N)$$

is also exact. The diligent reader has verified this already 'by hand' in the distant past (Exercise III.7.7), without appealing to adjunction. Direct proofs of the left-exactness of both Hom functors are straightforward and instructive.

5.2. Dual modules. We will explore further the exactness (and lack of exactness) of Hom in §6. Before doing that, however, we want to examine one important particular case of the contravariant aspect of the Hom functor.

Definition 5.4. Let M be an R-module. The *dual* $M^\vee$ of M is the R-module $\mathrm{Hom}_R(M, R)$. ⌟

One use of the dual module is to translate Hom computations into $\otimes$ computations, at least in special cases.

Proposition 5.5. *Let M be any R-module, and let F be a free R-module of finite rank. Then*

$$\mathrm{Hom}_R(M, F) \cong M^\vee \otimes_R F.$$

Proof. By hypothesis $F \cong R^{\oplus n} \cong R^n$; hence

$$\mathrm{Hom}_R(M, F) \cong \mathrm{Hom}_R(M, R^n) \cong \mathrm{Hom}(M, R)^n \cong \mathrm{Hom}(M, R) \otimes_R R^n \cong M^\vee \otimes_R F,$$

by Corollary 5.3. □

In fact, note that for all R-modules M, N there is a natural 'evaluation' map

$$\epsilon : M^\vee \otimes_R N \to \mathrm{Hom}_R(M, N)$$

defined on pure tensors by mapping $f \otimes n$ (for $f \in \mathrm{Hom}_R(M, R)$ and $n \in N$) to the R-module homomorphism $M \to N$ given by

$$m \mapsto f(m)n.$$

The reader will check (Exercise 5.5) that ϵ is an isomorphism if N is free of finite rank; this gives a more precise version of Proposition 5.5.

5.3. Duals of free modules. By definition, the 'duality' functor $M \mapsto M^\vee$ is a particular case of the contravariant flavor of Hom; hence it is itself contravariant and commutes with limits. Here is a first, immediate consequence:

Lemma 5.6. *For every family $\{M_i\}$ of R-modules, $(\bigoplus_i M_i)^\vee \cong \prod_i M_i^\vee$.*

Specializing to the case in which $M_i = R$ for all i determines the dual of a free module:

Corollary 5.7. *The dual of a free module is isomorphic to a product of copies of R:*

$$(R^{\oplus S})^\vee \cong R^S.$$

In particular, $(R^n)^\vee \cong R^n$: if F is a free R-module of finite rank, *then $F^\vee \cong F$.*

Proof. This follows from Lemma 5.6 and the fact that $R^\vee = \mathrm{Hom}_R(R, R)$ is isomorphic to R. □

Carefully note the magically disappearing $\oplus$ from the left-hand side to the right-hand side in the statement of Corollary 5.7: direct *sums* become direct *products* through the contravariant Hom (Corollary 5.3). However, a direct product of finitely many modules happens to be isomorphic to their direct sum (as we have known for a long time: Proposition III.6.1), hence the second part of the statement.

Remark 5.8. If S is infinite, R^S is 'much larger' than $R^{\oplus S}$: the first module consists of *all* functions $S \to R$, while the second only retains those that are zero for all but finitely many $s \in S$. ⌟

Remark 5.9. Note that the set S is *not* determined by the isomorphism class of a free module $F = R^{\oplus S}$. We proved in Corollary VI.1.11 (if R is an integral domain) that the *cardinality* of S is determined by the isomorphism class of F, but of course S itself, say as a subset of F, is not. In fact, recall (§VI.1.2) that the choice of a specific isomorphism $F \cong R^{\oplus S}$ is equivalent to the choice of a basis

of F. Now, the isomorphism appearing in Corollary 5.7 requires knowledge of S; indeed, we will verify in a moment (Example 5.11) that this isomorphism, even in the finite case, *does* depend on the choice of a basis. It is not canonical! The reader should endeavor to remember this slogan: *a finite-rank free module*—for example, a finite-dimensional vector space—*is isomorphic to its dual, but not canonically.* ⌟

This remark can be clarified by means of the following notion.

Definition 5.10. Consider the standard basis $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ of R^n. The *dual basis* of $(R^n)^\vee \cong R^n$ consists of $(\check{\mathbf{e}}_1, \dots, \check{\mathbf{e}}_n)$, where $\check{\mathbf{e}}_i \in (R^n)^\vee = \mathrm{Hom}_R(R^n, R)$ is determined by

$$\check{\mathbf{e}}_i(\mathbf{e}_j) = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$ ⌟

The isomorphism $(R^n)^\vee \cong R^n$ of Corollary 5.7 is obtained precisely by matching $\mathbf{e}_i$ and $\check{\mathbf{e}}_i$: this will be crystal clear to the reader who works out Exercise 5.7. In particular, the vectors $\check{\mathbf{e}}_i$ do form a basis of $(R^n)^\vee$.

Example 5.11. To see that these isomorphisms do depend on the choice of the basis, consider the standard basis $(\mathbf{e}_1, \mathbf{e}_2)$ of R^2 and the corresponding dual basis $(\check{\mathbf{e}}_1, \check{\mathbf{e}}_2)$ of $(R^2)^\vee$, and then choose a different basis $(\mathbf{e}'_1, \mathbf{e}'_2)$ for R^2, where $\mathbf{e}'_1 = \mathbf{e}_1$ and $\mathbf{e}'_2 = \mathbf{e}_1 + \mathbf{e}_2$, and let $(\check{\mathbf{e}}'_1, \check{\mathbf{e}}'_2)$ be the corresponding dual basis. By definition

$$\check{\mathbf{e}}'_1(\mathbf{e}'_2) = 0,$$

while

$$\check{\mathbf{e}}_1(\mathbf{e}'_2) = \check{\mathbf{e}}_1(\mathbf{e}_1 + \mathbf{e}_2) = 1 + 0 = 1.$$

Therefore, $\check{\mathbf{e}}'_1 \neq \check{\mathbf{e}}_1$ *even if* $\mathbf{e}'_1 = \mathbf{e}_1$*:* the two isomorphisms $R^2 \cong (R^2)^\vee$ determined by the two bases *are* different. ⌟

The fact that duality leads to noncanonical isomorphisms is somewhat unpleasant, but something magical will happen soon (§5.5): we will recover a *canonical*—that is, independent of any choice—isomorphism (on finite-rank free modules) by applying the duality functor *twice*. Whatever twist is introduced by the duality functor will be untwisted by applying duality *again*.

5.4. Duality and exactness. Before seeing this, let us contemplate another immediate consequence of the fact that duality is a particular case of the contravariant Hom:

Lemma 5.12. *The duality functor is left-exact: every exact sequence*

$$L \longrightarrow M \longrightarrow N \longrightarrow 0$$

of R-modules induces an exact sequence

$$0 \longrightarrow N^\vee \longrightarrow M^\vee \longrightarrow L^\vee.$$

Proof. This is an immediate consequence of the left-exactness of Hom. □

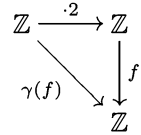
The duality functor is *not* exact in general: for example, taking duals in the exact sequence of abelian groups (a.k.a. $\mathbb{Z}$-modules)

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\cdot 2} \mathbb{Z} \longrightarrow \mathbb{Z}/2\mathbb{Z} \longrightarrow 0$$

gives the sequence

$$(*) \qquad 0 \longrightarrow 0 \longrightarrow \mathbb{Z}^\vee \xrightarrow{\gamma} \mathbb{Z}^\vee \longrightarrow 0$$

since the dual of $\mathbb{Z}/2\mathbb{Z}$ is zero (Exercise 5.6). The map γ in this sequence is the dual of the multiplication by 2. If $f : \mathbb{Z} \to \mathbb{Z}$ is an element of $\mathbb{Z}^\vee$, then $\gamma(f)$ is obtained by composition:



By linearity, $\gamma(f)(n) = f(2n) = 2f(n)$ is even for all $n \in \mathbb{Z}$. In particular, the identity $\mathbb{Z} \to \mathbb{Z}$ (as an element of $\mathbb{Z}^\vee$) is not in the image of γ. Thus γ is not surjective, and the sequence (*) is not exact.

There *are* situations in which duality is exact; we will understand this better after digesting §6, but the following observation will suffice for now.

Proposition 5.13. *Let*

$$0 \longrightarrow M \xrightarrow{\mu} N \xrightarrow{\nu} P \longrightarrow 0$$

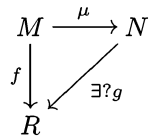
be an exact sequence of R-modules, with P free. *Then the induced sequence*

$$0 \longrightarrow P^\vee \xrightarrow{\nu^\vee} N^\vee \xrightarrow{\mu^\vee} M^\vee \longrightarrow 0$$

is exact.

This is also not new to the diligent readers, as it is a particular case of Exercise III.7.8 and, further, it is a consequence of Exercise 2.23. But here is a direct argument:

Proof. Lemma 5.12 takes care of all but the surjectivity of the map $N^\vee \to M^\vee$ induced from $M \to N$:



The question is whether every R-linear $f : M \to R$ can be extended to an R-linear map $g : N \to R$ so that $f = g \circ \mu$, that is, $f = \mu^\vee(g)$.

As P is free, $P \cong F^R(S) \cong R^{\oplus S}$ for some set S. Choosing (arbitrarily!) preimages in M of the standard basis vectors $\mathbf{e}_s \in R^{\oplus S}$ gives a set-function $S \to M$, extending to an R-linear map $\rho : P \to M$ by the universal property of free modules:

$$0 \longrightarrow M \xrightarrow{\mu} N \underset{\rho}{\overset{\nu}{\rightleftarrows}} P \longrightarrow 0 \,.$$

By construction, $\nu \circ \rho = \mathrm{id}_P$. Now let $n \in N$. Then $\nu(\rho(\nu(n))) = \nu(n)$, giving $\nu(n - \rho(\nu(n))) = 0$. By the exactness of the given sequence, $\exists m \in M$ such that

$$n - \rho(\nu(n)) = \mu(m).$$

The element m is unique, since μ is injective. Setting

$$g(n) := f(m)$$

gives the necessary extension. Indeed, g is immediately checked to be R-linear, and $g(\mu(m)) = f(m)$ by definition. □

In particular,

Corollary 5.14. *The duality functor is exact on vector spaces.*

5.5. Duals and matrices; biduality. Even without the benefit of full exactness, Lemma 5.12 and Corollary 5.7 reduce (in principle) the computation of the dual of any finitely presented module to matrix calculus. If

$$R^n \xrightarrow{\alpha} R^m \longrightarrow M \longrightarrow 0$$

is a presentation of an R-module M, then the dual $M^\vee$ is identified with the kernel of the dual of α:

$$0 \longrightarrow M^\vee \longrightarrow R^m \xrightarrow{\alpha^\vee} R^n.$$

Here we are applying isomorphisms $(R^m)^\vee \cong R^m$, $(R^n)^\vee \cong R^n$ from Corollary 5.7. The map $\alpha^\vee$ is obtained by applying the functor $\mathrm{Hom}_R(\_, R)$ to the map α. Once bases are chosen, α is determined by an $m \times n$ matrix; likewise, $\alpha^\vee$ corresponds to an $n \times m$ matrix. What is the relation between these two matrices? Of course there is only one sensible answer to this question, and it is correct:

Lemma 5.15. *Let A be the matrix representing a linear map $\alpha : R^n \to R^m$ with respect to the standard bases. Then the dual map $\alpha^\vee : (R^m)^\vee \to (R^n)^\vee$ is represented by the transpose of A with respect to the corresponding dual bases.*

The (easy) verification of this fact is left to the reader (Exercise 5.10). The upshot is that *the dual of the cokernel of a matrix A is the kernel of the transpose of A.*

Left-exactness of duality implies restrictions on which modules may appear as duals of other modules:

Proposition 5.16. *Let R be an integral domain, and let M be an R-module. Then $M^\vee$ is torsion-free.*

Proof. There is a surjection $R^{\oplus S} \twoheadrightarrow M$, thus, an exact sequence

$$R^{\oplus T} \to R^{\oplus S} \to M \to 0.$$

Dualizing, $M^\vee$ is realized as the kernel of the induced map $R^S \to R^T$; hence $M^\vee$ may be identified with a submodule of a product R^S. The latter has no nonzero torsion, so neither does $M^\vee$. □

The diligent reader has proved this already in Exercise VI.4.2 and probably by a more direct way than what we did just now, but never mind. In any case, remember that *dualizing kills torsion.* Practice this by working out Exercise 5.11.

Even if not every module is the dual of a module, we may wonder whether an arbitrarily given module M has some relation with a suitably chosen dual, and this is indeed the case: *there is a canonical map to the 'double-dual',*

$$\omega : \quad M \to M^{\vee\vee}.$$

To define this, consider the R-bilinear map

$$M^{\vee} \times M \to R$$

sending (f, m) to $f(m)$. For every $m \in M$, we get[23] an R-linear map $M^{\vee} \to R$, $f \mapsto f(m)$, in other words, an element $\omega(m)$ of $M^{\vee\vee}$. It is immediately checked that ω is R-linear.

By Proposition 5.16, $M^{\vee\vee}$ is torsion-free. In fact, the double-dual construction is a standard way to 'clean up' a module, removing its torsion.

The map ω is interesting even when torsion is not an issue. As we have seen in Corollary 5.7, if F is a finite-rank *free* R-module, then a choice of basis for F determines an isomorphism $F \cong F^{\vee}$; but this does indeed depend on the basis, so it is *not* a canonical isomorphism. Doing it twice,

$$F \cong F^{\vee} \cong F^{\vee\vee},$$

may seem *a priori* bound to make the situation even worse—surely, if we compose two noncanonical maps, then we cannot expect to get something canonical, right? Well, we *do* in this case, since this composition is nothing but the map ω (Exercise 5.13) and ω knows nothing about choices of bases or other such ambiguities. Therefore, our slogan admits the following addendum: *finite-rank free modules are* noncanonically *isomorphic to their duals and* canonically *isomorphic to their double-duals.*

In terms of matrices (and keeping in mind that duality is exact on free modules) this is perhaps not too surprising. After all, taking the transpose of a matrix *twice* simply gives the matrix back; thus, even if making sense of the homomorphism corresponding to the transpose of a matrix may depend on the choice of a basis, surely such a choice is inessential if the transpose is taken twice.

Incidentally, a module M is *reflexive* if the corresponding map ω is an isomorphism. Thus, reflexive modules are necessarily torsion-free and finite-rank free modules are reflexive.

5.6. Duality on vector spaces. In all these considerations, the case of *vector spaces* over a field k is only special in that vector spaces are free. Collecting the foregoing observations (including Exercise 5.1) for this particular case, we have the following:

—Duality is a *contravariant, exact* functor on k-Vect.

[23] The reader with a categorical frame of mind can view this a little more formally by working out Exercise 5.12.

—If V, W are vector spaces, then $(V \oplus W)^\vee \cong V^\vee \oplus W^\vee$.

—If V, W are vector spaces and $\dim V < \infty$ or $\dim W < \infty$, then $\mathrm{Hom}_k(V, W)$ is isomorphic to $V^\vee \otimes_k W$.

—If $\dim_k V < \infty$, a choice of a basis on V determines an isomorphism $V \cong V^\vee$.

—Finite-dimensional vector spaces are reflexive: if $\dim_k V < \infty$, then the canonical map $\omega : V \to V^{\vee\vee}$ is an isomorphism.

Exercises

As usual, R is a fixed commutative ring unless stated otherwise.

5.1. ▷ Prove that if F is a free R-module of finite rank and N is any R-module, then $\mathrm{Hom}_R(F, N) \cong F^\vee \otimes_R N$. [§5.6]

5.2. ¬ Let $\alpha : A \to B$ be a homomorphism of R-modules. Prove that α is an epimorphism if the induced map[24] $\alpha^* : \mathrm{Hom}_R(B, N) \to \mathrm{Hom}_R(A, N)$ is injective for all R-modules N and α is a monomorphism if α^* is surjective for all N.

Prove that the converse to the first statement holds and the converse to the second statement does not hold. However, show that if α admits a left-inverse, then α^* *is* surjective for all N. [5.3, 6.3]

5.3. Prove that a sequence

$$0 \longrightarrow A \xrightarrow{\alpha} B \xrightarrow{\beta} C \longrightarrow 0$$

of R-modules is exact if the induced sequence

$$(*) \qquad 0 \longrightarrow \mathrm{Hom}_R(C, N) \xrightarrow{\beta^*} \mathrm{Hom}_R(B, N) \xrightarrow{\alpha^*} \mathrm{Hom}_R(A, N) \longrightarrow 0$$

is exact for all R-modules N. (You have done most of this already, in Exercise 5.2. To show $\ker\gamma \subseteq \mathrm{im}\,\alpha$, choose $N = B/\mathrm{im}(\alpha)$.) Remember that the converse does not hold, since in general $\mathrm{Hom}_R(\_, N)$ is not exact. What extra hypothesis on α would guarantee the exactness of (*) for all N?

5.4. ¬ Let I be an ideal of R. As $I^2 \subseteq I$, there is a natural restriction map $\mathrm{Hom}_R(I, R/I) \to \mathrm{Hom}_R(I^2, R/I)$. Prove that the image of this map is 0. Prove that $\mathrm{Hom}_R(I/I^2, R/I) \cong \mathrm{Hom}_R(I, R/I)$. (This module is important in algebraic geometry, as it carries the information of a 'normal bundle' in good situations.) [6.20]

5.5. ▷ Prove that the evaluation map $M^\vee \otimes_R F \to \mathrm{Hom}_R(M, F)$ is an isomorphism if F is free of finite rank, providing an alternative proof of Proposition 5.5. [§5.2]

5.6. ▷ Show that $(\mathbb{Z}/2\mathbb{Z})^\vee = \mathrm{Hom}_\mathbb{Z}(\mathbb{Z}/2\mathbb{Z}, \mathbb{Z}) = 0$. [§5.4]

5.7. ▷ Prove 'directly' that $(R^{\oplus S})^\vee \cong R^S$: how does an R-linear map $R^{\oplus S} \to R$ determine a function $S \to R$, and what is the inverse of this correspondence? [§5.3]

[24] Note that the 'set-theoretic' quality of α^* is all that is needed here; in this problem, the R-module structure of Hom_R is irrelevant. Thus, the result also holds for noncommutative R.

5.8. Prove that the datum of an R-linear map $M \to M^\vee$ is equivalent to the datum of an R-bilinear map $M \times M \to R$, and explain why this equivalence can be set up in two ways. If $F = R^n$ is a free R-module of finite rank, determine the bilinear map $F \times F \to R$ corresponding to the isomorphism $F \cong F^\vee$ given in Corollary 5.7.

5.9. An R-bilinear map $\varphi : M \times M \to R$ is *nondegenerate* if the induced maps $M \to M^\vee$ are injective, and it is *nonsingular* if they are isomorphisms. The notions coincide if M is a finite-dimensional vectors space. Prove that the 'standard inner product' in $\mathbb{R}^n$ (defined in Exercise VI.6.18) is nondegenerate.

If M is free of rank n, let $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ be a basis of M, and let $A = (a_{ij})$ be the matrix with entries $a_{ij} = \varphi(\mathbf{e}_i, \mathbf{e}_j)$. Prove that φ is nondegenerate if and only if $\det A$ is nonzero, and it is nonsingular if and only if $\det A$ is a unit. (Cf. Proposition VI.6.5.)

5.10. Prove Lemma 5.15.

5.11. ▷ Let M be a finitely generated module over a PID, of rank r. 'Compute' the dual $M^\vee$. [§5.5]

5.12. ▷ Let M, N be R-modules. Show that there is a canonical bijection

$$\mathrm{Hom}_R(N, M^\vee) \cong \mathrm{Hom}_R(M, N^\vee).$$

Choosing $N = M^\vee$, the left-hand side has a distinguished element, namely the identity $M^\vee \to M^\vee$. Prove that the corresponding element on the right is the map $\omega : M \to M^{\vee\vee}$ defined in §5.2. [§5.5]

5.13. ▷ Let $F \cong R^n$ be a finite-rank free R-module. Verify that the composition of the (noncanonical) isomorphisms $F \cong F^\vee \cong F^{\vee\vee}$ from Corollary 5.7 is the (canonical) isomorphism ω defined in §5.2. [§5.5]

5.14. Let F be a free R-module (of any rank). Prove that the canonical map $F \to F^{\vee\vee}$ is injective. What is the simplest example you know of a module M such that $M \to M^{\vee\vee}$ is not injective?

5.15. ¬ Let V be a vector space, and let $W \subseteq V$ be a subspace. The *annihilator* of W is[25]

$$W^\perp := \{\check{v} \in V^\vee \mid (\forall w \in W),\ \check{v}(w) = 0\}.$$

Prove that $W^\perp$ is a subspace of $V^\vee$. If $\dim V = n$ and $\dim W = r$, prove that $\dim W^\perp = n - r$.

Assuming V is finite dimensional, prove that, under the canonical isomorphism $V^{\vee\vee} \cong V$, $W^{\perp\perp}$ maps isomorphically to W. [5.16, 5.17]

5.16. Let

$$0 \longrightarrow F_1 \longrightarrow F_2 \longrightarrow F_3 \longrightarrow 0$$

be an exact sequence of free R-modules. Viewing F_1 as a submodule of F_2 and extrapolating the notation introduced for vector spaces in Exercise 5.15, prove that $F_1^\perp \cong F_3^\vee$.

[25] The notation is inspired by Exercise VI.6.17, which is a particular case of this problem. Do you see why?

5.17. ¬ Let V be a vector space of dimension n. Prove that there is a natural bijection between the Grassmannian $\mathrm{Gr}(r, V)$ of r-dimensional subspaces of V (cf. Exercise VI.2.13) and the Grassmannian $\mathrm{Gr}(n-r, V^\vee)$ of $(n-r)$-dimensional subspaces of the dual of V. (Use Exercise 5.15.)

In particular, the Grassmannian $\mathrm{Gr}_k(n-1, n)$ has the same structure as the projective space $\mathbb{P}V = \mathrm{Gr}_k(1, n)$. We could in fact *define* the projective space associated to a vector space V of dimension n to be the set of subspaces of 'codimension 1' (that is, dimension $n-1$) in V. There are reasons why this would be preferable[26], but established conventions are what they are. [VI.2.13]

5.18. Let F be a free R-module of finite rank. For any $r \geq 1$, define a multilinear map

$$\delta : \underbrace{F \times \cdots \times F}_{r} \times \underbrace{F^\vee \times \cdots \times F^\vee}_{r} \to R$$

by

$$\delta(v_1, \ldots, v_r, \check{w}_1, \ldots, \check{w}_r) = \det(\check{w}_i(v_j))_{\substack{i=1,\cdots,r \\ j=1,\cdots,r}}.$$

- Prove that δ is multilinear and alternating in the first r and in the last r entries.
- Deduce that δ induces a bilinear map $\tilde{\delta} : \mathbb{\Lambda}^r(F) \times \mathbb{\Lambda}^r(F^\vee) \to R$.
- Prove that $\tilde{\delta}$ induces an isomorphism $(\mathbb{\Lambda}^r(F))^\vee \cong \mathbb{\Lambda}^r(F^\vee)$.

5.19. Let R be a ring, not necessarily commutative, and let M be a *left-R*-module. Prove that $M^\vee$ carries a natural *right-R*-module structure. Prove that $R^\vee$ is isomorphic as a ring to the *opposite* ring R°. (Cf. Exercise III.5.1.)

---

6. Projective and injective modules and the Ext functors

We have seen in §5.1 that the Hom functors are *left*-exact, and not exact in general. Studying this matter further leads us along a path parallel to the one traveled in the case of tensor products. In that case, we singled out a class of special ('flat') modules for which $\otimes$ *is* exact, and we discovered that the lack of exactness of tensor products may be precisely measured in terms of a whole collection of new ('Tor') functors. Precisely the same situation occurs for the covariant and contravariant flavors of Hom, and we will close the chapter by summarizing this story.

This brief overview should suffice for standard applications the reader may encounter. We are also hoping that it will serve as motivation for the next chapter: the wonderful facts that we will present here without proof (on Ext functors, analogous to the facts about Tor that were left without proof in §2.4) could be proven in detail by ad hoc methods, without too much effort; but they find their most natural setting in the wider apparatus of homological algebra and were in fact at

[26] The matter is not moot: it is true that there is a bijection between $\mathbb{P}V$ and $\mathrm{Gr}(n-1, V)$, but this bijection depends on the choice of a basis. Thus we cannot simply identify $\mathbb{P}V$ and $\mathrm{Gr}(n-1, V)$. We *can* identify $\mathbb{P}V$ and $\mathrm{Gr}(n-1, V^\vee)$, and it is sometime useful to do so.

the root of its development. The reader who bears with us throughout Chapter IX will see complete proofs of all the facts advertised here and substantially more.

As in the rest of the chapter, R will be a commutative ring. As elsewhere, the hypothesis of commutativity is essentially unnecessary; it has the convenient consequence that the objects defined below (such as $\mathrm{Ext}^i_R(M,N)$ for two R-modules M and N) are naturally endowed with an R-module structure. If R is not necessarily commutative, they are 'just' abelian groups.

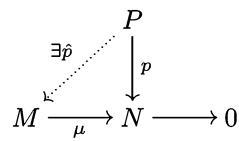
6.1. Projectives and injectives.

Definition 6.1. Let R be a (commutative) ring.

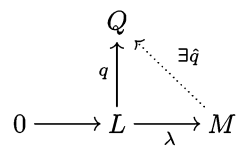
- An R-module P is *projective* if the functor $\mathrm{Hom}_R(P,\_)$ is exact.
- An R-module Q is *injective* if the functor $\mathrm{Hom}_R(\_,Q)$ is exact. ⌟

Since Hom is left-exact in any case, these definitions admit the following straightforward translations:

Lemma 6.2. *An R-module P is* projective *if and only if for all epimorphisms of R-modules $\mu : M \to N$, every R-linear map $p : P \to N$ lifts to an R-linear map $\hat{p} : P \to M$:*



An R-module Q is injective *if and only if for all monomorphisms or R-modules $\lambda : L \to M$, every R-linear map $q : L \to Q$ extends to an R-linear map $\hat{q} : M \to Q$:*



Proof. This is straightforward. Since $\mathrm{Hom}_R(\_,Q)$ is left-exact for all Q, Q is injective if and only if whenever a sequence

$$0 \longrightarrow L \longrightarrow M$$

is exact, then so is the induced sequence

$$\mathrm{Hom}_R(M,Q) \longrightarrow \mathrm{Hom}_R(L,Q) \longrightarrow 0.$$

This translates precisely into the given condition. The argument for projective modules is entirely analogous. □

Example 6.3. Taken as a module over itself, R is projective (because $\mathrm{Hom}_R(R,M)$ is canonically isomorphic to M, for all modules M) but not injective in general (since the duality functor $\mathrm{Hom}_R(\_,R)$ is not exact in general; cf. §5.1). ⌟

A variation on the theme of Lemma 6.2 rephrases the projective/injective conditions in terms of the *splitting* of sequences (cf. the end of §III.7.1). For example, assume that P is projective; then we claim that every exact sequence

$$0 \longrightarrow L \xrightarrow{\lambda} M \xrightarrow{\mu} P \longrightarrow 0$$

splits, in the sense that there is a submodule P' of M such that μ restricts to an isomorphism $P' \to P$ and $M \cong \lambda(L) \oplus P'$. Indeed, since P is projective, then the identity $P \xrightarrow{=} P$ lifts to a homomorphism $\rho : P \to M$, and the reader can then verify that $P' = \rho(P)$ fits the requirement. Loosely speaking, in this situation we can simply replace M by $L \oplus P$.

Similarly, if

$$0 \longrightarrow Q \longrightarrow M \longrightarrow N \longrightarrow 0$$

is exact and Q is *injective*, then the sequence necessarily splits. For example, $\mathbb{Z}$ is *not* an injective $\mathbb{Z}$-module, since the exact sequence

$$0 \longrightarrow \mathbb{Z} \longrightarrow \mathbb{Z} \xrightarrow{\cdot 2} \frac{\mathbb{Z}}{(2)} \longrightarrow 0$$

manifestly does not split.

It is not difficult to show that these properties in fact *characterize* projective and injective modules. With the terminology introduced in §III.7.1, a module P is projective if and only if every epimorphism $M \to P$ is a *split* epimorphism, and a module Q is injective if and only if every monomorphism $Q \to M$ is a *split* monomorphism. Dotting all the i's and crossing all the t's is left to the enterprising reader (Exercise 6.1).

6.2. Projective modules. Lemma 6.2 does not necessarily help in acquiring a feeling for projective or injective modules. We are reasonably happy with projective modules, because of the following characterization:

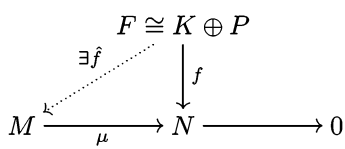
Proposition 6.4. *An R-module P is projective if and only if it is a direct summand of a free module, that is, if and only if there exists a free R-module F, an R-module K, and an isomorphism $K \oplus P \cong F$.*

Proof. Any set S of generators of P determines a surjection of the free module $F = R^{\oplus S}$ onto P and hence an exact sequence

$$0 \longrightarrow K \longrightarrow F \longrightarrow P \longrightarrow 0\,.$$

As observed above, such a sequence necessarily splits if P is projective; thus $F \cong K \oplus P$ in this case, as needed.

For the converse, assume that $K \oplus P \cong F$ for some free R-module F. It is clear that F satisfies the lifting property of Lemma 6.2:



(if $F = F^R(S)$, define $\hat{f}(s)$ to be an arbitrary inverse image of $f(s)$, for all $s \in S$, and extend to F by the universal property of free modules). It follows easily that the lifting property holds for P, by judicious use of the natural maps $P \hookrightarrow K \oplus P \twoheadrightarrow P$. Details are left to the reader. □

For example, free modules are projective. Proposition 6.4 streamlines the verification of simple properties of projective modules:

Corollary 6.5. *Let P_1, P_2 be projective R-modules. Then $P_1 \oplus P_2$ and $P_1 \otimes_R P_2$ are projective. Projective modules are flat.*

Proof. These statements follow easily from Proposition 6.4, the fact that $\otimes$ is distributive with respect to $\oplus$, and the fact that free modules are flat. □

The categorically minded reader should look for alternate proofs of these facts. For example, if P_1 and P_2 are both projective, the functor $\mathrm{Hom}_R(P_1, \mathrm{Hom}_R(P_2, \_))$ is exact; by adjunction, this is the functor $\mathrm{Hom}_R(P_1 \otimes_R P_2, \_)$; since this is exact, $P_1 \otimes_R P_2$ is projective.

A proof via adjunction of the fact that projective modules are flat is straightforward (Exercise 6.7), using the (nontrivial) fact that R-Mod *has enough injectives,* discussed below. As it happens, *finitely generated* flat modules over, e.g., Noetherian rings, are projective; the diligent reader has already done all the work necessary to prove this, in the exercises.

The fact that free modules are projective implies that R-Mod *has enough projectives:* this means that every R-module M admits a *projective resolution,* that is, an exact sequence

$$\cdots \longrightarrow P_3 \longrightarrow P_2 \longrightarrow P_1 \longrightarrow P_0 \longrightarrow M \ .$$

Indeed, a free resolution of M is in particular a projective resolution. Since there are in general 'more' projective modules than free modules, it can in principle be 'easier' to construct a projective resolution. The magic of homological algebra shows that projective resolutions may be used in place of free resolutions in (for example) computing Tor-modules. This will be clear once we go through some homological algebra machinery, in Chapter IX, especially §IX.8.

6.3. Injective modules. We may have inadvertently communicated to the reader the message that every categorical notion leads in a straightforward way to a mirror notion, by 'just reversing arrows'. This would seem to be the case for injective vs. projective modules: the definition of one kind is indeed obtained from the definition of the other by simply reversing arrows in the key diagrams.

However, this is as far as this naive perception can go. A seemingly unavoidable asymmetry of nature manifests itself here, making *injective* modules look more mysterious than projectives. For example, there is no 'easy' description of injective modules in the style of Proposition 6.4 (however, cf. Remark 6.13), and while it is trivially the case that an arbitrary module is surjected upon by a projective (e.g., free) module, the 'mirror' statement for injectives is certainly not trivial.

The problem is that it is a little hard to picture an injective module: while R itself is a trivial example of a projective R-module, at this point the reader would likely find it hard to name a single nonzero injective module over any ring whatsoever. Well, say over any ring that is not a field: if R is a field, then R itself is injective. (Why?) But if R is *not* a field...? Suppose r is a non-zero-divisor in R, and consider the inclusion of the ideal $(r) \cong R$ in R:

$$0 \longrightarrow R \xrightarrow{\cdot r} R\,;$$

extending the identity on R through the multiplication by r,

$$\begin{array}{ccccc} 0 & \longrightarrow & R & \xrightarrow{\cdot r} & R \\ & & = \downarrow & \swarrow_{\exists ?} & \\ & & R & & \end{array}$$

requires the existence of $s \in R$ such that $rs = 1$. It cannot be done unless r is a unit in R.

This tells us that in general R is not injective as an R-module and puts the finger on the 'problem': for Q to be injective, we must in particular be able to extend any map $I \to Q$ from an ideal I of R to a map $R \to Q$. A nice application of Zorn's lemma shows that this property *characterizes* injective modules:

Theorem 6.6. *An R-module Q is injective if and only if every R-linear map $f : I \to Q$, with I an ideal of R, extends to an R-linear map $\hat{f} : R \to Q$.*

This criterion is attributed to Reinhold Baer, who first introduced and studied injective modules (ca. 1940).

Proof. The 'only if' part of the statement is immediate from the definition of injective. To verify the 'if' part, assume Q satisfies the stated extension condition, let $L \subseteq M$ be any inclusion of R-modules, and let $q : L \to Q$ be a given R-linear map:

$$\begin{array}{ccccc} 0 & \longrightarrow & L & \longrightarrow & M \\ & & q \downarrow & \swarrow_{\exists ?\hat{q}} & \\ & & Q & & \end{array}$$

We want to construct an extension $\hat{q} : M \to Q$ of q.

Consider the set E of pairs $(\widetilde{L}, \tilde{q})$, where $\widetilde{L}$ ranges over the submodules of M containing L and $\tilde{q} : \widetilde{L} \to Q$ extends q: $\tilde{q}|_L = q$. Then E is nonempty, since $(L, q) \in E$, and we can define a partial order on E by prescribing that $(\widetilde{L}', \tilde{q}') \preceq (\widetilde{L}'', \tilde{q}'')$ if $L' \subseteq L$ and $\tilde{q}''$ extends $\tilde{q}'$. It is clear that every chain has an upper bound (take the union of the corresponding $\widetilde{L}$), so by Zorn's lemma there exists a maximal extension $(\widetilde{L}, \tilde{q})$, and we are done if we can prove that $\widetilde{L} = M$.

Arguing by contradiction, assume there exists $m \in M$, $m \notin \widetilde{L}$, and consider the proper ideal $I = \{r \in R \,|\, rm \in \widetilde{L}\}$ of R. We can define an R-linear map $f : I \to Q$ by putting

$$f(r) := \tilde{q}(rm);$$

by hypothesis, this map extends to an R-linear map $\hat{f} : R \to Q$. But then we can use this map to extend q beyond $\widetilde{L}$: let $\overline{L} = \widetilde{L} + \langle m \rangle \supsetneq \widetilde{L}$, and define $\overline{q} : \overline{L} \to Q$ by

$$\overline{q}(\ell + rm) := \tilde{q}(\ell) + \hat{f}(r)$$

for $\ell \in \widetilde{L}$, $r \in R$. This is immediately checked to be well-defined and clearly extends $\tilde{q}$. But then $(\overline{L}, \overline{q}) \in E$ and $(\widetilde{L}, \tilde{q}) \prec (\overline{L}, \overline{q})$, contradicting the maximality of $(\widetilde{L}, \tilde{q})$ and concluding the proof. □

Theorem 6.6 completely clarifies the notion of injective modules in the case of PIDs. An R-module D is *divisible* if $rD = D$ for every non-zero-divisor $r \in R$, that is, if we can 'divide' every element of D by every non-zero-divisor of R.

Corollary 6.7. *Let R be a PID. Then an R-module Q is injective if and only if it is divisible.*

Proof. Exercise 6.14. □

Example 6.8. Viewed as abelian groups (i.e., $\mathbb{Z}$-modules), $\mathbb{Q}$ and $\mathbb{Q}/\mathbb{Z}$ are injective. More generally, if D is any divisible abelian group and $K \subseteq D$, then D/K is injective. (Indeed, it is trivially divisible!) ⌟

Going from friendly PIDs like $\mathbb{Z}$ to arbitrary rings may still seem challenging, but here is where some of our previous work pays off. Recall that for every ring homomorphism $f : S \to R$ we have defined in §3.3 a functor[27] $f^! : S\text{-Mod} \to R\text{-Mod}$, by setting $f^!(M) = \mathrm{Hom}_S(R, M)$. This functor *preserves injectives:*

Lemma 6.9. *Let $f : S \to R$ be a homomorphism of commutative rings, and let Q be an injective S-module. Then $f^!(Q)$ is an injective R-module.*

Remark 6.10. Similarly, f^* preserves projectives (Exercise 6.6). ⌟

Proof. By adjunction (Lemma 3.5),

$$\mathrm{Hom}_R(\_, f^!(Q)) \cong \mathrm{Hom}_S(f_*(\_), Q)$$

as functors $R\text{-Mod} \to R\text{-Mod}$. Since f_* is exact (Proposition 3.6) and $\mathrm{Hom}_S(\_, Q)$ is exact by hypothesis, $\mathrm{Hom}_R(\_, f^!(Q))$ must be exact. This is precisely the statement. □

Since every ring R admits a (unique) map $\iota : \mathbb{Z} \to R$, we have a good source of injective objects in R-Mod for every R:

Corollary 6.11. *Let R be a commutative ring. If D is any divisible abelian group, then $\iota^!(D) = \mathrm{Hom}_{\mathbb{Z}}(R, D)$ is injective in R-Mod.*

This result is as close as the author can get to an 'intuitive' feeling for the notion of injective module. For example, at this point we clearly see why 'R-Mod has enough injectives':

[27] Watch out—in §3.3 the homomorphism goes from R to S, so the definition of $f^!$ may look backwards here.

Corollary 6.12. *Let M be an R-module. Then M can be identified with a submodule of an* injective *R-module.*

Proof. We claim that it suffices to show that $\mathbb{Z}$-Mod has enough injectives, Indeed, this will show that there exists a divisible abelian group D such that $M \subseteq D$ (M is in particular an abelian group); since R-linear maps are in particular $\mathbb{Z}$-linear,

$$M \cong \mathrm{Hom}_R(R, M) \hookrightarrow \mathrm{Hom}_{\mathbb{Z}}(R, M) \subseteq \mathrm{Hom}_{\mathbb{Z}}(R, D),$$

and the rightmost module is injective by Corollary 6.11.

Thus, we are reduced to the case $R = \mathbb{Z}$ and $M = A$ an abelian group. There is a surjective homorphism from $\mathbb{Z}^{\oplus A}$ to A and hence an identification

$$A \cong \frac{\mathbb{Z}^{\oplus A}}{K}$$

for some $K \subseteq \mathbb{Z}^{\oplus A}$. Embedding $\mathbb{Z}^{\oplus A}$ in $\mathbb{Q}^{\oplus A}$, we obtain an injective homomorphism

$$A \hookrightarrow \frac{\mathbb{Q}^{\oplus A}}{K},$$

and we are done since $\mathbb{Q}^{\oplus A}/K$ is divisible (Example 6.8). □

As an immediate consequence, every R-module admits an *injective resolution:* an exact sequence

$$0 \longrightarrow M \longrightarrow Q_0 \longrightarrow Q_1 \longrightarrow Q_2 \longrightarrow Q_3 \longrightarrow \cdots$$

in which every Q_i is injective. Indeed, this is obtained by applying Corollary 6.12 to M to construct Q_0 and then applying it again to the cokernel of $M \to Q_0$ to construct Q_1, and so on.

The fact that R-Mod has enough injectives has a generalization to sheaves, which is an essential ingredient in the definition of *sheaf cohomology,* an extremely important tool in modern algebraic geometry.

Remark 6.13. Corollary 6.12 may be refined, proving that every module can in fact be realized as a submodule of a product $\mathrm{Hom}_{\mathbb{Z}}(R, \mathbb{Q}/\mathbb{Z})^S$ (this is injective, by Corollary 6.11 and Exercise 6.15). That is, products of $\iota^!(\mathbb{Q}/\mathbb{Z}) = \mathrm{Hom}_{\mathbb{Z}}(R, \mathbb{Q}/\mathbb{Z})$ play for injectives the role played by free modules (= coproducts of $R = \iota^*(\mathbb{Z})$) for projectives. (Not surprisingly, they are said to be *cofree.*) It follows that every injective R-module is a direct summand of $\mathrm{Hom}_{\mathbb{Z}}(R, \mathbb{Q}/\mathbb{Z})^S$ for some S. ⌟

6.4. The Ext functors. The next step to take now is the construction of tools to 'quantify' the lack of exactness of Hom_R, in the same sense that the functors Tor_i^R quantify the lack of exactness of $\otimes_R$. The corresponding functors are called Ext^i_R. Since there are two functors associated with Hom (the covariant $\mathrm{Hom}_R(M, \_)$ and the contravariant $\mathrm{Hom}_R(\_, N)$), it would be natural to expect two corresponding sequences of functors. Amazingly, the same 'bifunctors' $\mathrm{Ext}^i_R(\_, \_)$ work for both! For all R-modules M and N, we have the following:

- For all $i \geq 0$,
$$\mathrm{Ext}^i_R(M, \_), \quad \mathrm{Ext}^i_R(\_, N)$$
is a covariant, resp., contravariant, functor R-Mod $\to$ R-Mod.

- For $i = 0$,
$$\mathrm{Ext}^0_R(M,N) = \mathrm{Hom}_R(M,N).$$
- For every exact sequence
$$0 \longrightarrow A \longrightarrow B \longrightarrow C \longrightarrow 0$$
of R-modules, there are long exact sequences
$$\begin{array}{l} 0 \longrightarrow \mathrm{Hom}_R(M,A) \longrightarrow \mathrm{Hom}_R(M,B) \longrightarrow \mathrm{Hom}_R(M,C) \\ \qquad\qquad \delta_0 \\ \longrightarrow \mathrm{Ext}^1_R(M,A) \longrightarrow \mathrm{Ext}^1_R(M,B) \longrightarrow \mathrm{Ext}^1_R(M,C) \\ \qquad\qquad \delta_1 \\ \longrightarrow \mathrm{Ext}^2_R(M,A) \longrightarrow \mathrm{Ext}^2_R(M,B) \longrightarrow \mathrm{Ext}^2_R(M,C) \longrightarrow \cdots \end{array}$$
and
$$\begin{array}{l} 0 \longrightarrow \mathrm{Hom}_R(C,N) \longrightarrow \mathrm{Hom}_R(B,N) \longrightarrow \mathrm{Hom}_R(A,N) \\ \qquad\qquad \delta'_0 \\ \longrightarrow \mathrm{Ext}^1_R(C,N) \longrightarrow \mathrm{Ext}^1_R(B,N) \longrightarrow \mathrm{Ext}^1_R(A,N) \\ \qquad\qquad \delta'_1 \\ \longrightarrow \mathrm{Ext}^2_R(C,N) \longrightarrow \mathrm{Ext}^2_R(B,N) \longrightarrow \mathrm{Ext}^2_R(A,N) \longrightarrow \cdots \end{array}$$
for suitable natural connecting morphisms δ_i, δ'_i.

This will be proven in Chapter IX. In any case, just knowing that such wonderful things exist gives us enough information to be able to play with them a little. For example,

—if $\mathrm{Ext}^1_R(P,\_) = 0$, then P is projective;

—if $\mathrm{Ext}^1_R(\_,Q) = 0$, then Q is injective.

Indeed, the exactness of the corresponding $\mathrm{Hom}_R(P,\_)$, resp., $\mathrm{Hom}_R(\_,Q)$, follows immediately from the exact sequences displayed above.

In fact, pushing the envelope a little will tell us much more. Recall the procedure that defines $\mathrm{Tor}^R_i(\_,N)$ from the functor $\_ \otimes_R N$: to compute the value of these functors at a module M,

- find a free resolution of M:
$$\cdots \longrightarrow R^{\oplus S_2} \longrightarrow R^{\oplus S_1} \longrightarrow R^{\oplus S_0} \longrightarrow M \longrightarrow 0;$$
- apply the functor $\_ \otimes_R N$ to the free part, obtaining a complex $M_\bullet \otimes_R N$:
$$\cdots \longrightarrow N^{\oplus S_2} \longrightarrow N^{\oplus S_1} \longrightarrow N^{\oplus S_0} \longrightarrow 0 \qquad ;$$
- define Tor by taking the homology of this complex:
$$\mathrm{Tor}^R_i(M,N) := H_i(M_\bullet \otimes N).$$

We also pointed out that *projective* resolutions may be used in place of free resolutions[28]. This strategy is an example of the general procedure used to 'derive'

[28] In fact, resolutions by *flat* modules suffice; this will be proven in Chapter IX.

functors. Can you dream up how the same strategy may be applied in order to compute Ext-modules? Don't read ahead until you have thought a little about this!

Welcome back. In the case of $\mathrm{Ext}^*_R(M,N)$, there would seem to be two natural and possibly different ways to go:

—One could start from a *projective* resolution of M:

$$\cdots \longrightarrow P_2 \longrightarrow P_1 \longrightarrow P_0 \longrightarrow M \longrightarrow 0;$$

apply the contravariant $\mathrm{Hom}_R(\_,N)$ to the projective part, obtaining a new complex $\mathrm{Hom}_R(M_\bullet,N)$, i.e.,

$$0 \longrightarrow \mathrm{Hom}_R(P_0,N) \longrightarrow \mathrm{Hom}_R(P_1,N) \longrightarrow \mathrm{Hom}_R(P_2,N) \longrightarrow \cdots,$$

and take *co*homology[29], proposing the definition

$$\mathrm{Ext}^i_R(M,N) := H^i(\mathrm{Hom}_R(M_\bullet,N)).$$

—Or one could start from an *injective* resolution of N:

$$0 \longrightarrow N \longrightarrow Q_0 \longrightarrow Q_1 \longrightarrow Q_2 \longrightarrow \cdots;$$

apply the covariant $\mathrm{Hom}_R(M,\_)$ to the injective part, obtaining a new complex $\mathrm{Hom}_R(M,N_\bullet)$, i.e.,

$$0 \longrightarrow \mathrm{Hom}_R(M,Q_0) \longrightarrow \mathrm{Hom}_R(M,Q_1) \longrightarrow \mathrm{Hom}_R(M,Q_2) \longrightarrow \cdots,$$

and again take cohomology, leading to

$$\mathrm{Ext}^i_R(M,N) := H^i(\mathrm{Hom}_R(M,N_\bullet)).$$

These definitions are independent of all choices and agree with each other!

Again, the reader will have to wait for Chapter IX to see a proof that this is the case and that the resulting modules do satisfy the requirements spelled out at the beginning of this subsection. As in the case of Tor, the impatient reader can get a good feel for the needed arguments by working out the case of finitely generated modules over PIDs, making good use of the snake lemma.

The reader who is willing to take our word for now and believe all these beautiful results is already in a position to perform many computations and to understand injective/projective modules (even) better. For example,

Proposition 6.14. *An R-module P is* projective *if and only if* $\mathrm{Ext}^1_R(P,\_) = 0$, *if and only if* $\mathrm{Ext}^i_R(P,\_) = 0$ *for all* $i > 0$.

An R-module Q is injective *if and only if* $\mathrm{Ext}^1_R(\_,Q) = 0$, *if and only if* $\mathrm{Ext}^i_R(\_,Q) = 0$ *for all* $i > 0$.

Proof. The second assertion: we have already noted that Q is injective if $\mathrm{Ext}^1_R(\_,Q) = 0$, and this is trivially the case if $\mathrm{Ext}^i_R(\_,Q) = 0$ for all $i > 0$. So we just have to

[29] Recall from §III.7.1 that the homology of complexes with increasing (rather than decreasing) indices is called *cohomology;* this is usually denoted H^i, with 'upper' indices.

prove that $\mathrm{Ext}^i_R(\_, Q) = 0$ for all $i > 0$ if Q is injective. This is immediate from the second definition given above: if Q is injective, then

$$0 \longrightarrow Q \longrightarrow Q \longrightarrow 0 \longrightarrow 0 \longrightarrow \cdots$$

is an injective resolution of Q; thus, for an R-module M, $\mathrm{Ext}^i_R(M, Q)$ is the cohomology of the complex

$$0 \longrightarrow \mathrm{Hom}_R(M, Q) \longrightarrow 0 \longrightarrow 0 \longrightarrow \cdots$$

giving $\mathrm{Ext}^0_R(M, Q) = \mathrm{Hom}_R(M, Q)$ and $\mathrm{Ext}^i_R(M, Q) = 0$ for $i > 0$ and all M.

The first assertion is proven similarly, using the first definition given above for Ext. □

6.5. $\mathrm{Ext}^*_{\mathbb{Z}}(\mathbf{G}, \mathbb{Z})$. We will attempt to convince the reader that computing Ext modules is not too unreasonable, by discussing the computation of $\mathrm{Ext}^*_{\mathbb{Z}}(G, \mathbb{Z})$ for an arbitrary finitely generated abelian group.

—Since Hom commutes with finite direct sums (Corollary 5.3), so does Ext; this is an easy consequence of the definitions given above.

—By the classification theorem for finitely generated abelian groups, we are reduced to computing $\mathrm{Ext}^i_{\mathbb{Z}}(\mathbb{Z}, \mathbb{Z})$ and $\mathrm{Ext}^i_{\mathbb{Z}}(\mathbb{Z}/m\mathbb{Z}, \mathbb{Z})$, for all $m > 0$ and $i \geq 0$.

—The first module is $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}, \mathbb{Z}) \cong \mathbb{Z}$ for $i = 0$, and it is 0 for $i > 0$: indeed, this follows from Proposition 6.14, since $\mathbb{Z}$ is projective.

—The second module: use the projective resolution

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\cdot m} \mathbb{Z} \longrightarrow \mathbb{Z}/m\mathbb{Z} \longrightarrow 0\ ;$$

this yields that $\mathrm{Ext}^i_{\mathbb{Z}}(\mathbb{Z}/m\mathbb{Z}, \mathbb{Z})$ is the cohomology of the complex

$$0 \longrightarrow \mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}, \mathbb{Z}) \xrightarrow{\cdot m} \mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}, \mathbb{Z}) \longrightarrow 0 \longrightarrow 0 \longrightarrow \cdots,$$

that is,

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\cdot m} \mathbb{Z} \longrightarrow 0 \longrightarrow 0 \longrightarrow \cdots,$$

confirming that $\mathrm{Ext}^0_{\mathbb{Z}}(\mathbb{Z}/m\mathbb{Z}, \mathbb{Z}) = \mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/m\mathbb{Z}, \mathbb{Z}) = 0$ ($\mathbb{Z}$ has no torsion) and computing $\mathrm{Ext}^1_{\mathbb{Z}}(\mathbb{Z}/m\mathbb{Z}, \mathbb{Z}) \cong \mathbb{Z}/m\mathbb{Z}$, with vanishing Ext^2 and higher.

—The conclusion is that $\mathrm{Ext}^0_{\mathbb{Z}}(G, \mathbb{Z})$ picks up the free part of G, while $\mathrm{Ext}^1_{\mathbb{Z}}(G, \mathbb{Z})$ is isomorphic to the torsion part, and $\mathrm{Ext}^i_{\mathbb{Z}}(G, \mathbb{Z}) = 0$ for $i \geq 2$ (as should have been expected; cf. Exercise 6.18).

In particular, if G is a *finite* abelian group, then

$$G \cong \mathrm{Ext}^1_{\mathbb{Z}}(G, \mathbb{Z});$$

as it happens, this isomorphism is not canonical.

Another Ext computation gives a different interpretation of this group. The exact sequence

$$0 \longrightarrow \mathbb{Z} \longrightarrow \mathbb{Q} \longrightarrow \mathbb{Q}/\mathbb{Z} \longrightarrow 0$$

is an injective resolution of $\mathbb{Z}$; thus, $\mathrm{Ext}^*_{\mathbb{Z}}(G, \mathbb{Z})$ is the cohomology of

$$0 \longrightarrow \mathrm{Hom}_{\mathbb{Z}}(G, \mathbb{Q}) \longrightarrow \mathrm{Hom}_{\mathbb{Z}}(G, \mathbb{Q}/\mathbb{Z}) \longrightarrow 0 \longrightarrow \cdots.$$

If G is torsion (for example, if G is finite), it follows that $\mathrm{Ext}^1_{\mathbb{Z}}(G,\mathbb{Z})$ is isomorphic to $\mathrm{Hom}_{\mathbb{Z}}(G,\mathbb{Q}/\mathbb{Z})$.

This group is the *Pontryagin dual* of G, denoted $\widehat{G}$. Thus, we have verified that if G is a finite abelian group, then G is isomorphic (not canonically) to its Pontryagin dual $\widehat{G} = \mathrm{Hom}_{\mathbb{Z}}(G,\mathbb{Q}/\mathbb{Z})$. Of course this fact is not difficult to check directly, but it is immediate from the point of view of Ext.

Finally, the notation Ext is due to the fact that one can construct a *bijection* between the group $\mathrm{Ext}^1_R(M,N)$ and the set of equivalence classes of *extensions* of M by N, that is, of exact sequences

$$0 \longrightarrow M \longrightarrow E \longrightarrow N \longrightarrow 0$$

modulo a suitable isomorphism relation (cf. §IV.5.2). For example, the direct sum $E = M \oplus N$ corresponds to the 0 element of this Ext group. The diligent reader will construct this bijection in the exercises.

Exercises

As usual, R denotes a fixed commutative ring.

6.1. ▷ Prove that an R-module P is projective if and only if every epimorphism $M \to P$ splits and Q is injective if and only if every monomorphism $Q \to M$ splits. [§6.1]

6.2. Prove that the result of Proposition 5.13 holds more generally whenever P is a *projective* module.

6.3. ¬ Prove that an R-linear map $A \to B$ is injective if and only if the induced map $\mathrm{Hom}_R(B,Q) \to \mathrm{Hom}_R(A,Q)$ is surjective for all *injective* modules Q. (Hint: R-Mod has enough injectives.) This strengthens the result of Exercise 5.2. [6.7]

6.4. ¬ Let $\{P_i\}_{i\in I}$ be a family of R-modules. Prove that $\bigoplus_i P_i$ is projective if and only if each P_i is projective. [IX.5.4]

6.5. Prove that the dual of a projective module is projective. Prove that finitely generated projective modules are reflexive (that is, isomorphic to their biduals).

6.6. ¬ If $f : S \to R$ is a ring homomorphism and P is a projective S-module, then $f^*(P)$ is a projective R-module. [§6.3]

6.7. ▷ Give a proof of the fact that projective modules are flat, using adjunction. (Hint: R-Mod has enough injectives, and Exercise 6.3.) [§6.2, 6.8]

6.8. Prove that Exercises 2.24 and 6.7 together imply the result of Exercise VI.5.5.

6.9. Prove that vector spaces are flat, injective, and projective. (Try to do this directly, without invoking Baer's criterion.)

6.10. Prove that a finitely generated module over a PID is free if and only if it is projective if and only if it is flat. (Do this 'by hand', without invoking Exercise 6.12.)

6.11. ¬ Prove that a finitely generated module over a local ring is projective if and only if it is free. (You have already done this, in Exercise VI.5.5.)

Besides PIDs and local rings, there are other classes of rings for which projective and free modules coincide. Topological considerations suggested to Serre that projective modules over a polynomial ring over a field should necessarily be free, but it took two decades to prove that this is indeed the case. [6.12]

6.12. ¬ An R-module M is 'locally free' if $M_{\mathfrak{p}}$ is free as an $R_{\mathfrak{p}}$-module for every prime ideal $\mathfrak{p}$ of R.

Prove that a finitely generated module over a Noetherian ring is locally free if and only if it is projective if and only if it is flat. (Use the results of Exercises 2.24 and 6.11.)

The hypothesis of finite generation is necessary (cf. Exercise 6.13). [6.10]

6.13. ¬ Prove that $\mathbb{Q}$ is flat, but not projective, as a $\mathbb{Z}$-module. [6.12]

6.14. ▷ Prove that a module over a PID is injective if and only if it is divisible. (Use Baer's criterion.) [§6.3]

6.15. ▷ Prove that $Q_1 \oplus Q_2$ is injective if and only if Q_1, Q_2 are both injective. More generally, prove that if $\{Q_i\}_{i\in I}$ is a family of R-modules, then $\prod_i Q_i$ is injective if and only if each Q_i is injective. [§6.3, IX.5.4]

6.16. Prove that $\mathbb{Z} \oplus \mathbb{Q}$ is flat as a $\mathbb{Z}$-module but neither projective nor injective.

6.17. Prove that $\mathrm{Hom}_R(M,N) \cong \mathrm{Ext}^0_R(M,N)$, using any of the definitions provided for $\mathrm{Ext}^i_R(M,N)$.

6.18. ▷ Prove that if R is a PID, then $\mathrm{Ext}^i_R(M,N) = 0$ for all $i \geq 2$ and all R-modules M, N. [§6.5]

6.19. Let r be a non-zero-divisor in R, and let M be an R-module. Compute all $\mathrm{Ext}^i(R/(r), M)$.

6.20. Let I be an ideal of R. Prove that $\mathrm{Ext}^1_R(R/I, R/I) \cong \mathrm{Hom}_R(I/I^2, R/I)$. (Cf. Exercise 5.4. This says that this Ext module essentially computes the normal bundle of an embedding.)

6.21. ¬ In Exercise 2.15 we have seen why Tor is called Tor. Why is Ext called Ext?

Let M, N, E be R-modules. An *extension* of M by N is an exact sequence

$$(*) \qquad 0 \longrightarrow N \longrightarrow E \longrightarrow M \longrightarrow 0.$$

Two extensions are 'isomorphic' if there is a commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & N & \longrightarrow & E & \longrightarrow & M & \longrightarrow & 0 \\ & & \| & & \downarrow & & \| & & \\ 0 & \longrightarrow & N & \longrightarrow & E' & \longrightarrow & M & \longrightarrow & 0 \end{array}$$

linking them. (Note that, by the snake lemma, the middle arrow must then be an isomorphism: cf. Exercise III.7.10.) Extensions that are isomorphic to the standard sequence $0 \to N \to N \oplus M \to M \to 0$ are 'trivial'.

Every extension $\mathscr{E}$ as above determines an element $\epsilon = \epsilon(\mathscr{E}) \in \mathrm{Ext}^1(M,N)$ as follows. The sequence (*) induces a long exact sequence of Ext, i.e.,

$$\cdots \longrightarrow \mathrm{Hom}_R(E,N) \longrightarrow \mathrm{Hom}_R(N,N) \longrightarrow \mathrm{Ext}^1_R(M,N) \longrightarrow \cdots,$$

and we let $\epsilon(\mathscr{E})$ be the image in $\mathrm{Ext}^1_R(M,N)$ of the distinguished element $\mathrm{id}_N \in \mathrm{Hom}_R(N,N)$.

Prove that if $\mathscr{E}$ and $\mathscr{E}'$ are isomorphic extensions, then $\epsilon(\mathscr{E}) = \epsilon(\mathscr{E}')$. Prove that if $\mathscr{E}$ is a trivial extension, then $\epsilon(\mathscr{E}) = 0$. [6.22]

6.22. Exercise 6.21 teaches us that extensions determine elements of Ext^1_R. We get an even sharper statement by constructing an *inverse* to the map ϵ: for every element $\kappa \in \mathrm{Ext}^1_R(M,N)$, we will construct an extension $e(\kappa)$ such that $\epsilon(e(\kappa)) = \kappa$ and such that $e(\epsilon(\mathscr{E}))$ is isomorphic to $\mathscr{E}$.

- Let F be any free module surjecting onto M, and let $i : K \hookrightarrow F$ be the kernel of $\pi : F \twoheadrightarrow M$. Since F is free (hence projective), a piece of the long exact sequence of Ext, i.e.,

  $$\cdots \longrightarrow \mathrm{Hom}_R(K,N) \longrightarrow \mathrm{Ext}^1_R(M,N) \longrightarrow \mathrm{Ext}^1_R(F,N) = 0 \longrightarrow \cdots,$$

  tells us that there exists a homomorphism $k : K \to N$ mapping to the element $\kappa \in \mathrm{Ext}^1_R(M,N)$.
- We now have a monomorphism $(i,k) : K \to F \oplus N$. Let E be the cokernel $(F \oplus N)/K$. Prove that the epimorphism $(\pi, 0) : F \oplus N \to M$ factors through this cokernel, defining an epimorphism $E \twoheadrightarrow M$.
- Prove that the natural monomorphism $N \cong 0 \oplus N \hookrightarrow F \oplus N$ defines a monomorphism $N \hookrightarrow E$, identifying N with the kernel of $E \to M$.
- We let $e(\kappa)$ be the extension

  $$0 \longrightarrow N \longrightarrow E \longrightarrow M \longrightarrow 0$$

  that we have obtained. Prove that different choices in the procedure lead to isomorphic extensions.
- Prove that $e(\epsilon(\mathscr{E}))$ is isomorphic to $\mathscr{E}$ and that $\epsilon(\mathscr{E}(\kappa)) = \kappa$.

The upshot is that there is a natural *bijection* between $\mathrm{Ext}^1_R(M,N)$ and the set of isomorphism classes of extensions of M by N. Hence the name. 'Higher' Ext^i_R ($i > 1$) may also be treated similarly: for example, $\mathrm{Ext}^2_R(M,N)$ can be analyzed in terms of two-step extensions consisting of exact sequences

$$0 \longrightarrow N \longrightarrow E_1 \longrightarrow E_2 \longrightarrow M \longrightarrow 0,$$

where two such extensions are 'isomorphic' if there is a diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & N & \longrightarrow & E_1 & \longrightarrow & E_2 & \longrightarrow & M & \longrightarrow & 0 \\ & & \| & & \downarrow & & \downarrow & & \| & & \\ 0 & \longrightarrow & N & \longrightarrow & E_1' & \longrightarrow & E_2' & \longrightarrow & M & \longrightarrow & 0 \end{array}$$

This approach to Ext is attributed to Nobuo Yoneda.

Homological algebra

The limited exposure to the Tor and Ext functors given in §VIII.2 and §VIII.6 should assist the reader in occasional encounters with these functors. However, it is very worthwhile to go back, dotting all i's and crossing all t's as necessary to prove the wonderful facts responsible for the existence and the behavior of these functors.

This is one motivation for the material covered in this chapter. Tor and Ext are examples of *derived functors,* a machinery that would give us access to several other important constructions, such as group or sheaf cohomology. Exploring any of these applications would take us too far, but understanding the basics of *homological algebra,* including a thorough look at derived functors, is in itself a very worthwhile goal. The material covered here should arm the reader with enough information to be able to absorb any application quickly when the time comes and the opportunity arises. All the wonderful mysteries about Tor and Ext stated in Chapter VIII will be recovered as very particular cases of the results presented in this chapter.

The natural context in which to play this game is that of *abelian categories:* these are categories carrying good notions of kernels and cokernels, with properties to which the reader is accustomed due to long exposure to the category R-Mod of modules over a ring. In fact, elementary (and not so elementary) presentations of homological algebra often take the standpoint that one may as well work with R-Mod; allegedly, nothing of substance is lost by doing this rather than trying to work in the more natural context of abelian categories.

We are not completely convinced. This pedagogical device has strong theoretical backing, provided by the *Freyd-Mitchell embedding theorem:* every small abelian category is equivalent to a full subcategory of the category of left-modules over a (not necessarily commutative) ring. However, we believe that some exposure to purely 'arrow-theoretic' arguments is intrinsically beneficial, and the reader should have the chance to acquire a taste for this type of reasoning. In any case, we dislike the idea of simply asking the reader to believe that 'most of what follows works in a more general setting', without providing some evidence for this fact.

Therefore, we will take a little time to look at abelian categories before getting into homological algebra proper; this will be done in the first two sections of this chapter, which an impatient reader could probably skip over without much harm. While we will not prove the Freyd-Mitchell embedding theorem, we will give a solid justification for the fact that we can indeed perform diagram chases by working with 'elements', in the sense of (for example) checking whether a diagram is commutative or whether a given sequence is exact. Proving this will give us good practice in maneuvering arrow-theoretic arguments; once we are done with it, we can indeed develop homological algebra using conventional 'element-theoretic' arguments, with a relatively clean conscience.

As for homological algebra proper, our general guiding principle will be to hunt for the 'essence of (co)homology': what information in a complex (see §III.7) is really responsible for its homology. This will motivate most constructions in this chapter, including derived functors. A guiding beacon throughout the chapter will be the notion of *derived category:* we will not construct derived categories in full generality, but the reader should nevertheless emerge with a basic understanding of what they are. One reason for this compromise is that we felt that delving into *triangulated* categories would take us too far; they are only mentioned in passing at the end of the chapter. Balancing Tor and Ext will motivate the introduction of double complexes, and these in turn will motivate a quick treatment of spectral sequences, in the last section.

1. (Un)necessary categorical preliminaries

1.1. Undesirable features of otherwise reasonable categories. The category of modules over a ring R is the template example of an *abelian category:* a category for which it makes sense to talk about 'kernels' and 'cokernels', 'complexes', 'exactness', etc., and in which convenient results such as the 'snake lemma' hold. We begin by trying to concretize these thoughts a little.

As motivation for the key definitions, we will first list a few annoying features of some of the categories we have encountered.

- A category may not have initial or final objects; even if they are there, they may not be the same objects (as in Set).
- Products, coproducts may not exist (if the product of $\mathbb{F}_2$ and $\mathbb{F}_3$ existed in Fld, what would its characteristic be?).
- It may make no sense to talk about kernels and cokernels in a category; even when it does make sense, they may not exist (kernels do not necessarily exist in the category of *finitely generated* modules over a ring; cf. Example III.6.5).
- Even when there are kernels and cokernels, monomorphisms need not be kernels (nonnormal subgroups are not kernels: remember §II.7.6), and epimorphisms need not be cokernels (the injection $\mathbb{Z} \to \mathbb{Q}$ is an epimorphism in Ring; see §III.2.3).
- A morphism may be a monomorphism and an epimorphism without being an isomorphism (Ring again, same example).

Homological algebra will deal extensively with kernels, cokernels, exact sequences, and the like: the material covered in §III.7 will be our starting point. That material was developed for the category R-Mod of (left-)modules over a ring, for which none of the annoying things listed above occurs: in R-Mod every epimorphism is a cokernel, a map is an isomorphism if and only if it is a monomorphism and an epimorphism, there is a zero-object (= both initial and final), and so on.

The notions of 'additive' and 'abelian' category extract the few key properties that guarantee that no pathologies such as those listed above may occur. Working in such categories 'feels' quite a bit like working in R-Mod, in the sense that all expected properties of kernels and cokernels hold. The Freyd-Mitchell embedding theorem (see §2.4) will provide a mathematical reason supporting this psychological fact.

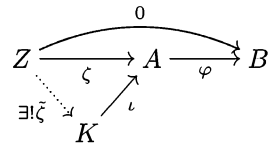
1.2. Additive categories.

Definition 1.1. A category A is *additive* if it has a zero-object, A has both finite products and finite coproducts, and each set of morphisms $\mathrm{Hom}_{\mathsf{A}}(A, B)$ is endowed with an abelian group structure, in such a way that the composition maps are bilinear. A functor between two additive categories is *additive* if it preserves the abelian group structures on Hom-sets. ⌟

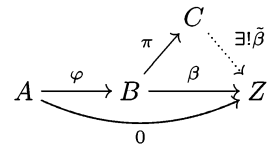
None of the categories mentioned in the list in §1.1 is additive. Even Grp (which has kernels and cokernels, products, etc.) is *not* an additive category: the Hom-sets in Grp do not have a natural abelian group structure. Of course Ab *is* additive, and so are all categories R-Mod.

If a category A is additive, it makes sense to talk about 'zero-morphisms' (which we will denote by 0) and to add or subtract morphisms; two morphisms α, β are equal if and only if $\alpha - \beta = 0$. One can also talk about kernels and cokernels in A, by adopting the categorical definitions as suitable limits and colimits (Example VIII.1.11). Here are those definitions again, expanded for intelligibility:

Definition 1.2. Let $\varphi : A \to B$ be a morphism in an additive category A. A morphism $\iota : K \to A$ is a *kernel* of φ if $\varphi \circ \iota = 0$ and for all morphisms $\zeta : Z \to A$ such that $\varphi \circ \zeta = 0$ there exists a unique $\tilde{\zeta} : Z \to K$ making the diagram



commute. A morphism $\pi : B \to C$ is a *cokernel* of φ if $\pi \circ \varphi = 0$ and for all morphisms $\beta : B \to Z$ such that $\beta \circ \varphi = 0$ there exists a unique $\tilde{\beta} : C \to Z$ making the diagram



commute. ⌟

Here is the same in sound-bite format:

If $\varphi \circ \zeta = 0$, then ζ factors uniquely through $\ker \varphi$.
If $\beta \circ \varphi = 0$, then β factors uniquely through $\operatorname{coker} \varphi$.

We have to get used to the fact that morphisms may have 'many' kernels (or cokernels), uniquely identified with each other by virtue of being answers to a universal question (Proposition I.5.4); it is common to harmlessly abuse language and talk about *the* kernel and *the* cokernel of a morphism. Further, we have to get used to the fact that the kernel $\iota : K \to A$ of a morphism $A \to B$ is a *morphism.* In a category such as Ab we are used to thinking of the kernel as a subobject of A, but this is really nothing but the datum of an inclusion map: the kernel is really that map. In an arbitrary category one cannot talk about 'subobjects' or 'inclusions'; the closest one can get to these notions is *monomorphism.* Similarly, 'surjective' is not really an option; *epimorphism* is the appropriate replacement. Recall (§I.2.6) that $\varphi : A \to B$ is a monomorphism if for all parallel morphisms $\zeta_1, \zeta_2 : Z \to A$,

$$Z \underset{\zeta_2}{\overset{\zeta_1}{\rightrightarrows}} A \xrightarrow{\varphi} B,$$

$\varphi \circ \zeta_1 = \varphi \circ \zeta_2 \implies \zeta_1 = \zeta_2$. It is an epimorphism if for all parallel $\beta_1, \beta_2 : B \to Z$,

$$A \xrightarrow{\varphi} B \underset{\beta_2}{\overset{\beta_1}{\rightrightarrows}} Z,$$

$\beta_1 \circ \varphi = \beta_2 \circ \varphi \implies \beta_1 = \beta_2$. One benefit of working in an additive category is that these definitions simplify a little:

Lemma 1.3. *A morphism $\varphi : A \to B$ in an additive category is a monomorphism if and only if for all $\zeta : Z \to A$,*

$$\varphi \circ \zeta = 0 \implies \zeta = 0.$$

It is an epimorphism if and only if for all $\beta : B \to Z$,

$$\beta \circ \varphi = 0 \implies \beta = 0.$$

Proof. This is simply because two morphisms with the same source and target are equal if and only if their difference in the corresponding Hom-set (which is an abelian group by hypothesis) is 0. □

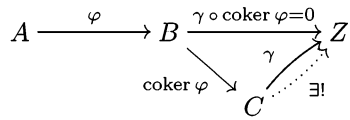
Are kernels necessarily monomorphisms? Yes:

Lemma 1.4. *In any additive category, kernels are monomorphisms and cokernels are epimorphisms.*

We will run into several such 'dual' statements, and we will give the proof for one half, leaving the other half to the reader; as a rule, the two proofs mirror each other closely. There is a good reason for this: the opposite (cf. Exercise I.3.1) of an additive category is additive, and kernels, monomorphisms, etc., in one correspond to cokernels, epimorphisms, etc., in the other. Thus, proving one of these statements for *all* additive categories establishes at the same time the truth of its dual statement.

However, going through the motions necessary to produce stand-alone proofs of the dual statements makes for good practice, and we invite the reader to work out the corresponding exercises at the end of this section. It is invariably the case that these arguments are relatively easy to understand and picture in one's mind but rather awkward to write down precisely. This seems to be a feature of many arrow-theoretic arguments and probably reflects inveterate biases acquired by early exposure to Set.

Proof. Let $\varphi : A \to B$ be a morphism in an additive category A, and let $\operatorname{coker}\varphi : B \to C$ be its cokernel. Let $\gamma : C \to Z$ be a morphism such that $\gamma \circ \operatorname{coker}\varphi = 0$. The composition $(\gamma \circ \operatorname{coker}\varphi) \circ \varphi$ is 0; by definition of cokernel, $\gamma \circ \operatorname{coker}\varphi$ factors *uniquely* through C:



Since $\gamma \circ \operatorname{coker}\varphi = 0 = 0 \circ \operatorname{coker}\varphi$, the uniqueness forces $\gamma = 0$. This proves that $\operatorname{coker}\varphi$ is an epimorphism, by Lemma 1.3.

The proof that kernels are monomorphisms is analogous and is left to the reader (Exercise 1.9). □

Certain limits are guaranteed to exist in an additive category: finite products and coproducts do. On the other hand, kernels and cokernels (which are also limits) do not necessarily exist in an additive category. For example, the category of finitely generated modules over a ring is additive, but it does not have kernels in general (essentially because a submodule of a finitely generated module is not necessarily finitely generated). But as soon as a morphism φ *does* have kernels or cokernels in an additive category, the basic qualities of that morphism can be detected 'as usual' by looking at $\ker\varphi$ and $\operatorname{coker}\varphi$. For example,

Lemma 1.5. *Let $\varphi : A \to B$ be a morphism in an additive category. If φ has a kernel, then φ is a monomorphism if and only if $0 \to A$ is its kernel. If φ has a cokernel, then φ is an epimorphism if and only if $A \to 0$ is its cokernel.*

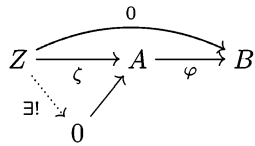
Compare this statement with Proposition III.6.2!

Proof. Let's do kernels this time.

First assume $\varphi : A \to B$ is a monomorphism with a kernel $\iota : K \to A$. In particular, $\iota \circ \varphi : K \to A \to B$ is the zero-morphism; therefore $\iota = 0$ by Lemma 1.3. What about K? If $\zeta : Z \to A$ is any morphism such that the composition $Z \to A \to B$ is 0, then ζ factors through $\iota = 0$ since ι is the kernel of φ. It follows that $\zeta = 0$ and in particular that it factors (uniquely) through $0 \to A$. This proves that $0 \to A$ is a kernel of φ, as stated.

Conversely, assume that $0 \to A$ is a kernel for $\varphi : A \to B$, and let $\zeta : Z \to A$ be a morphism such that $\varphi \circ \zeta = 0$. It follows that ζ factors through $0 \to A$, since

the latter is a kernel for φ:



This implies $\zeta = 0$, proving that φ is a monomorphism.

The statement about epimorphisms and cokernels is left to the reader (Exercise 1.9). □

In view of Lemma 1.5, we should be able to use diagrams

$$0 \to A \to B, \quad A \to B \to 0$$

to signal that $A \to B$ is a monomorphism, resp., an epimorphism: think 'exact'. This is where we are heading, but the fact that kernels and cokernels do not necessarily exist prevents us from effectively doing this in a category that is 'only' additive.

Incidentally, it is common to denote monomorphisms and epimorphisms by suitably decorated arrows; popular choices are $\rightarrowtail$ and $\twoheadrightarrow$, respectively.

1.3. Abelian categories. The moral at this point is that if a morphism in an additive category has kernels and cokernels, then these will behave as expected. But kernels and cokernels do not necessarily exist, and this prevents us from going much further. Also, while (as we have seen) kernels are monomorphisms and cokernels are epimorphisms in an additive category, there is no guarantee that monomorphisms should necessarily be kernels and epimorphisms should be cokernels. In the end, we simply demand these additional features explicitly.

Definition 1.6. An additive category A is *abelian* if kernels and cokernels exist in A; every monomorphism is the kernel of some morphism; and every epimorphism is the cokernel of some morphism. ⌟

As mentioned already, R-Mod is an abelian category, for every ring R. The prototype of an abelian category is Ab: this[1] is why these categories are called *abelian*.

Since kernels are necessarily monomorphisms (by Lemma 1.4), we see that in an *abelian* category we can adopt a mantra entirely analogous to the useful 'kernel $\iff$ submodule' of §III.5.3 vintage: in abelian categories, the slogan would be 'kernel $\iff$ monomorphism' (and similarly for cokernels vs. epimorphisms).

Remark 1.7. Just as it is convenient to think of monomorphisms $A \rightarrowtail B$ as defining A as a 'subobject' of B, it is occasionally convenient to think of epimorphisms as 'quotients': if $\varphi : A \rightarrowtail B$ is a monomorphism, we can use B/A to denote (the target of) $\operatorname{coker}\varphi$. We will have no real use for this notation in this section, but it will come in handy later on. ⌟

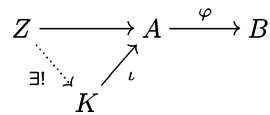
[1] There is nothing particularly 'commutative' about an abelian category.

The very existence of kernels and cokernels links these two notions tightly in an abelian category:

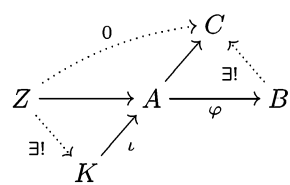
Lemma 1.8. *In an abelian category* A*, every kernel is the kernel of its cokernel; every cokernel is the cokernel of its kernel.*

Proof. We will prove the second half and leave the first half to the reader (Exercise 1.9).

Let $\varphi : A \to B$ be the cokernel of some morphism $Z \to A$; since A is abelian, φ has a kernel $\iota : K \to A$. The composition $Z \to A \to B$ is 0, so $Z \to A$ factors through ι by definition of kernel:



Now let $A \to C$ be a morphism such that the composition $K \to A \to C$ is the zero-morphism; then so is the composition $Z \to A \to C$. Therefore $A \to C$ factors through a unique morphism $B \to C$,



since φ is the cokernel of $Z \to A$. But this shows that $\varphi : A \to B$ satisfies the property defining the cokernel of its kernel $K \to A$, as stated. □

Putting together Lemma 1.8 and Lemma 1.5, we can rephrase Definition 1.6 by listing the following requirements on a category A:

- A is additive;
- kernels and cokernels exist in A;
- if $\varphi : A \to B$ is a morphism whose kernel is 0, then φ is the kernel of its cokernel;
- if $\psi : B \to C$ is a morphism whose cokernel is 0, then ψ is the cokernel of its kernel.

This is a popular equivalent reformulation of the definition of abelian category. The last two requirements should call to mind the exact sequence

$$0 \longrightarrow A \xrightarrow{\varphi} B \xrightarrow{\psi} C \longrightarrow 0$$

familiar from the R-Mod context. The reader can already entertain the sense in which such a sequence can be 'exact' in an abelian category: the third requirement identifies A (or, rather, $A \to B$) with the kernel of ψ, and the fourth one identifies C with the cokernel of φ.

Now suppose that $A \to B$ is sandwiched between zeros:

$$0 \longrightarrow A \rightarrowtail B \longrightarrow 0$$

in an exact sequence. One of our many Pavlovian reactions would make us want to deduce that $A \to B$ is an isomorphism, because this is the case in R-Mod, and this indeed works in any abelian category:

Lemma 1.9. *Let $\varphi : A \to B$ be a morphism in an abelian category* A*, and assume that φ is both a monomorphism and an epimorphism. Then φ is an isomorphism.*

Remark 1.10. This fact does not hold in general additive categories. In fact, the reader will encounter in Exercise 3.4 an example of an additive category with kernels and cokernels (but nevertheless not abelian) and with morphisms that are both monomorphisms and epimorphisms, without being isomorphisms. ⌟

Proof. The kernel of φ exists (because A is abelian), and hence it is $0 \to A$ (by Lemma 1.5, since φ is a monomorphism). Similarly, $B \to 0$ is a cokernel of φ. Further, φ is the cokernel of $0 \to A$ and the kernel of $B \to 0$, by Lemma 1.8.

Now consider the identity $B \to B$:

$$\begin{array}{ccccccc} & & & & B & & \\ & & & & \downarrow{\scriptstyle \mathrm{id}} & & \\ 0 & \longrightarrow & A & \overset{\varphi}{\rightarrowtail} & B & \longrightarrow & 0 \end{array}$$

Since $B \to B \to 0$ is (trivially) the zero morphism and φ is the kernel of $B \to 0$, we obtain a unique morphism $\psi : B \to A$ making the diagram commute:

$$\begin{array}{ccccccc} & & & & B & & \\ & & & {\scriptstyle \exists! \psi} \swarrow & \downarrow{\scriptstyle \mathrm{id}} & & \\ 0 & \longrightarrow & A & \underset{\varphi}{\rightarrowtail} & B & \longrightarrow & 0 \end{array}$$

As $\varphi\psi = \mathrm{id}_B$, this shows that φ has a right-inverse. Similarly, consider the identity $A \to A$, as follows:

$$\begin{array}{ccccccc} & & A & & & & \\ & & \uparrow{\scriptstyle \mathrm{id}} & & & & \\ 0 & \longrightarrow & A & \overset{\varphi}{\rightarrowtail} & B & \longrightarrow & 0 \end{array}$$

The composition $0 \to A \to A$ is the zero morphism, and φ is the cokernel of $0 \to A$, so we have a unique morphism $\eta : B \to A$ making this diagram commute:

$$\begin{array}{ccccccc} & & A & & & & \\ & & \uparrow{\scriptstyle \mathrm{id}} & \nwarrow{\scriptstyle \exists! \eta} & & & \\ 0 & \longrightarrow & A & \underset{\varphi}{\rightarrowtail} & B & \longrightarrow & 0 \end{array}$$

This says $\eta\varphi = \mathrm{id}_A$, so φ has a left-inverse as well.

Thus, φ has both a left-inverse η and a right-inverse ψ. It follows that $\eta = \psi$ is a two-sided inverse of φ and that φ is an isomorphism as promised (cf. Proposition I.4.2). □

The reader should note the arrow-theoretic nature of this argument. In the category R-Mod we could have given the following (possibly) simpler argument: by Proposition III.6.2, monomorphisms are injective and epimorphisms are surjective; therefore φ is bijective, and bijective homomorphisms are isomorphisms by Exercise III.5.12. Such set-theoretic arguments are not an option in an arbitrary abelian category (at least until we develop the material of §2), since the objects of an abelian category are not given as 'sets'. Judicious use of appropriate universal properties accomplishes the same goal and may be argued to convey a 'deeper' sense of why the proven statement is true.

1.4. Products, coproducts, and direct sums. We will denote[2] the product of two objects A, B of an abelian category A by $A \times B$, and we will denote their coproduct by $A \amalg B$. Both exist, since A is additive.

The presence of these objects, and of kernels and cokernels, gives us access to other interesting constructions.

Example 1.11. For instance, *fibered* products (or 'pull-backs') exist in any abelian category, just as in R-Mod (cf. Exercise III.6.10). Consider a diagram

$$\begin{array}{ccc} & & B \\ & & \downarrow{\scriptstyle \psi} \\ A & \xrightarrow[\varphi]{} & C \end{array}$$

in an abelian category. The fibered product of A and B over C is an object $A \times_C B$ with morphisms to A and B, completing the commutative diagram

$$\begin{array}{ccc} A \times_C B & \xrightarrow{\varphi'} & B \\ {\scriptstyle \psi'}\downarrow & \square & \downarrow{\scriptstyle \psi} \\ A & \xrightarrow[\varphi]{} & C \end{array}$$

and final with this property (that is what the small square in the middle of the diagram is supposed to indicate). The fibered product may be constructed in this context as the kernel of the difference of the two morphisms

$$A \times B \underset{\varphi \circ \psi'}{\overset{\psi \circ \varphi'}{\rightrightarrows}} C$$

(just as in the particular case of R-Mod). The reader will prove that these 'fiber squares' preserve kernels, in the sense that $\ker \varphi = \psi' \circ \ker \varphi'$ (Exercise 1.16).

Similarly, fibered *coproducts* (a.k.a. push-outs) may be constructed as cokernels of coproducts and preserve cokernels. ⌟

In fact, in R-Mod such constructions are simplified by the fact that finite products and coproducts coincide, in the sense that the *direct sum* of two modules satisfies both universal requirements for products and coproducts: Proposition III.6.1.

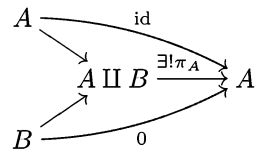
[2] As we will see, the single notation $A \oplus B$ will be an appropriate substitute for both notations.

But the direct sum of two modules M_1, M_2 is defined by giving a module structure to the Cartesian product $M_1 \times M_2$. Again, this strategy cannot be applied in a general abelian category, so we have to come up with an alternative. This gives us another example of the contrast between set-theoretic and arrow-theoretic arguments.

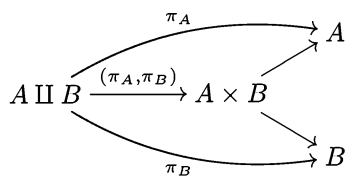
The coproduct $A \amalg B$ is endowed with morphisms

$$A \xrightarrow{i_A} A \amalg B, \quad B \xrightarrow{i_B} A \amalg B\,,$$

which are easily checked to be monomorphisms (do this!). In an arbitrary category, natural morphisms $A \amalg B \to A$ and $A \amalg B \to B$ are not available (example: Set). In an abelian category (and in fact already in an additive category) we do have such morphisms: map $A \to A$ by the identity and $B \to A$ by the zero-morphism to obtain a unique morphism $\pi_A : A \amalg B \to A$, making the following diagram commute:



You can similarly define a morphism $\pi_B : A \amalg B \to B$. It follows (by the universal property of products) that there is a unique morphism from $A \amalg B$ to $A \times B$ making the following diagram commute:



Now we claim that this morphism $(\pi_A, \pi_B) : A \amalg B \to A \times B$ is both a monomorphism and an epimorphism. By Lemma 1.9, this (and a simple induction) will prove

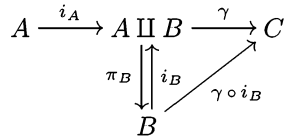
Proposition 1.12. *In an abelian category, finite products and coproducts coincide.*

Proving Proposition 1.12 is a good exercise in manipulating arrows and universal properties. We will quickly review the ingredients going into the proof that the morphism (π_A, π_B) is a *monomorphism*. Similar arguments, proving that the same morphism is an epimorphism, will be left to the reader.

In the process, we recover for these objects all the properties we expect on the basis of our experience with R-Mod. Doing it 'with arrows' requires a bit of practice but is not particularly difficult.

- $\pi_B : A \amalg B \to B$ is the cokernel of $i_A : A \to A \amalg B$:

Indeed, suppose $A \to A \amalg B \to C$ is the zero-morphism; stare at



and keep in mind that π_B is determined by the requirements that $\pi_B \circ i_A = 0$ and $\pi_B \circ i_B = \mathrm{id}_B$. We have to show that γ factors uniquely through π_B. Note that if it factors at all, then the factorization is unique: indeed, if $\gamma = \delta \circ \pi_B$, then

$$\delta = \delta \circ (\pi_B \circ i_B) = (\delta \circ \pi_B) \circ i_B = \gamma \circ i_B.$$

So the only morphism that can work is $\gamma \circ i_B$. Therefore, consider $(\gamma \circ i_B) \circ \pi_B$:

— $(\gamma \circ i_B) \circ \pi_B \circ i_A = 0 = \gamma \circ i_A$ (since $\pi_B \circ i_A = 0$); and

— $(\gamma \circ i_B) \circ \pi_B \circ i_B = \gamma \circ i_B$ (since $\pi_B \circ i_B = \mathrm{id}_B$).

By the universal property of coproducts, it follows that $(\gamma \circ i_B) \circ \pi_B = \gamma$, proving that γ indeed factors uniquely through π_B. This confirms that $\pi_B : A \amalg B \to B$ is the cokernel of i_A.

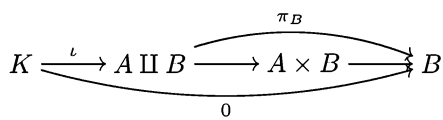
- $i_A : A \to A \amalg B$ is the kernel of $\pi_B : A \amalg B \to B$:

Indeed, i_A is a monomorphism since $A \xrightarrow{i_A} A \amalg B \xrightarrow{\pi_B} A$ is the identity; hence it is a kernel (definition of abelian category!); hence it is the kernel of its cokernel (Lemma 1.8).

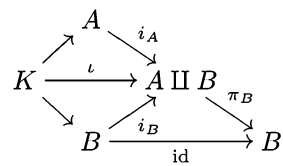
By the same token, $\pi_A : A \amalg B \to A$ is the cokernel of $i_B : B \to A \amalg B$ and $i_B : B \to A \amalg B$ is the kernel of $\pi_A : A \amalg B \to A$.

- The morphism $(\pi_A, \pi_B) : A \amalg B \to A \times B$ constructed above is a monomorphism:

Consider the kernel ι of (π_A, π_B). Composing with the projection to B gives the zero-morphism:



It follows that ι factors through the kernel of π_B, that is, through i_A, and through i_B, by the same token. Therefore we have the following commutative diagram:



The composition $K \longrightarrow A \longrightarrow A \amalg B \longrightarrow B$ is zero (because $\pi_B \circ i_A = 0$); therefore so is the composition $K \longrightarrow B \xrightarrow{\mathrm{id}} B$. That is, $K \to B$ is actually the zero-morphism; *a fortiori*, $\iota = 0$: the kernel of $A \amalg B \to A \times B$ is 0, so this morphism is indeed a monomorphism (by Lemma 1.5).

The same technique (or an appeal to the opposite category) proves that

- the morphism $(\pi_A, \pi_B) : A \amalg B \to A \times B$ constructed above is an epimorphism.

We will happily leave this to the reader (Exercise 1.17).

This will conclude the proof of Proposition 1.12, by Lemma 1.9. □

In view of Proposition 1.12, we can adopt for any abelian category A the convention of identifying $A \amalg B$ and $A \times B$, calling this object (defined up to isomorphism) the *direct sum* of A and B, denoted $A \oplus B$. This object comes endowed with natural morphisms

$$A \rightarrowtail A \oplus B, \quad B \rightarrowtail A \oplus B, \quad A \oplus B \twoheadrightarrow A, \quad A \oplus B \twoheadrightarrow B,$$

satisfying all the properties to which we have grown accustomed.

1.5. Images; canonical decomposition of morphisms. Yet another example of the application of an arrow-theoretic perspective is the existence of 'canonical decompositions' of morphisms in any abelian category, in terms completely analogous to the decompositions studied in Set (§I.2.8), Grp (§II.8.1), Ring (§III.3.3), R-Mod (§III.5.4). In each of these previous examples we could simply argue that the decomposition was the original decomposition obtained in Set, enriched as the case may be with additional structures adapted to the category under examination. This is not an option in the context of an abelian category, and yet the decomposition holds just as well: every morphism may be written as an epimorphism, followed by an isomorphism, followed by a monomorphism. Here is the precise statement:

Theorem 1.13. *Every morphism $\varphi : A \to B$ in an abelian category A may be decomposed as*

$$A \twoheadrightarrow C \xrightarrow[\tilde{\varphi}]{\sim} K \rightarrowtail B, \qquad A \xrightarrow{\ \varphi\ } B$$

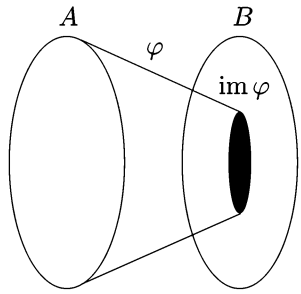
where $A \to C$ is the cokernel of the kernel of φ and $K \to B$ is the kernel of the cokernel of φ. The induced morphism $\tilde{\varphi} : C \to K$ is uniquely determined and is an isomorphism.

This theorem will follow from useful related considerations. Comparing the statement with its set-theoretic counterpart suggests that

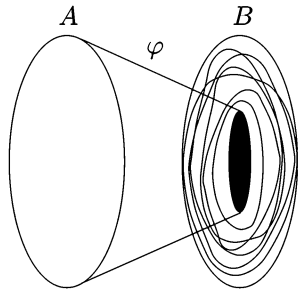
$$\ker(\operatorname{coker}\varphi) : \quad K \rightarrowtail B$$

should play the role of the 'image of φ'. Again note that the set-theoretic definition of 'image' is not an option in the general context we are exploring: unlike groups, rings, etc., the objects of a general abelian category are not 'just' sets endowed with extra structure[3]. But look at your mental image of what the image of a function φ is; ours looks like this:

[3]However, we will soon recover a sense in which images in abelian categories can be 'understood' in set-theoretic terms: see Lemma 2.7.



We can fit many 'subobjects' in B to which A maps via φ:

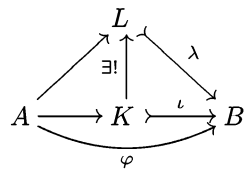


and $\operatorname{im}\varphi$ is the 'smallest' such subobject; that is, it is initial with the property of being a subobject of B through which φ factors. This intuition is precisely captured by $\ker(\operatorname{coker}\varphi)$, once we adapt to thinking of 'subobjects' as 'monomorphisms':

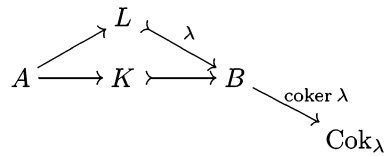
Lemma 1.14. *Let $\varphi : A \to B$ be a morphism in an abelian category, and let $\iota : K \to B$ be the kernel of the cokernel of φ. Then*

- *ι is a monomorphism;*
- *φ factors through ι; and*
- *ι is initial with these properties.*

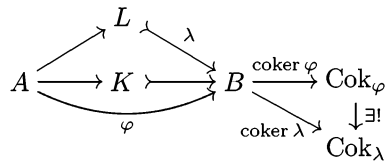
By Lemma 1.4, ι is a monomorphism. It is clear that φ factors through ι: the composition $A \to B \to \operatorname{coker}\varphi$ is the zero-morphism, so there is a naturally induced $A \to K$ by the universal property of kernels. The more interesting part of the statement of Lemma 1.14 asserts that if $\lambda : L \rightarrowtail B$ is any monomorphism through which φ also factors, then ι must factor uniquely through λ:



Proof. Consider $\operatorname{coker}\lambda$:



Since φ factors through λ, the composition $A \to B \to \mathrm{Cok}_\lambda$ is 0; by the universal property of $\operatorname{coker}\varphi$, we have an induced map:



Since $K \to \mathrm{Cok}_\varphi$ is the zero-morphism, this implies that

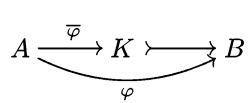
$$K \longrightarrow B \xrightarrow{\operatorname{coker}\lambda} \mathrm{Cok}_\lambda$$

is the zero-morphism. Since $\lambda : L \to B$ is the kernel of its cokernel (Lemma 1.8), it follows that there is a unique morphism $K \to L$ making the diagram commute, as stated. □

Definition 1.15. Let $\varphi : A \to B$ be a morphism in an abelian category. The *image* of φ, denoted $\operatorname{im}\varphi$, is $\ker(\operatorname{coker}\varphi)$. The *coimage* of φ, denoted $\operatorname{coim}\varphi$, is $\operatorname{coker}(\ker\varphi)$. ⌟

Lemma 1.14 motivates the first definition: every monomorphism through which φ factors must factor uniquely through $\operatorname{im}\varphi$. (Of course in this context the source of $\operatorname{im}\varphi$ is only defined up to (unique) isomorphism, as is every solution to a universal problem.) As for the second definition, the reader will formulate the universal property satisfied by $\operatorname{coim}\varphi$ (Exercise 1.19). With the convention mentioned in Remark 1.7 and allowing for some abuse of language, the (target of the) coimage of $\varphi : A \to B$ is the 'quotient' $A/\ker\varphi$.

Our mental image, and surely yours, suggests that if $K \to B$ is the image of φ, then the induced morphism $\overline{\varphi} : A \to K$,



should be an epimorphism. This is indeed the case:

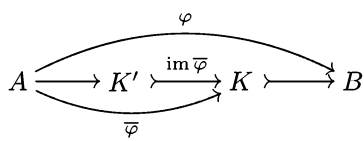
Lemma 1.16. *Let $\varphi : A \to B$ be a morphism in an abelian category, and let $\operatorname{im}\varphi : K \to B$, $\operatorname{coim}\varphi : A \to C$ be its image and coimage, respectively. Then the induced morphisms $A \to K$ and $C \to B$ are, respectively, an epimorphism and a monomorphism.*

Proof. As usual, we will prove half of the statement and leave the other half to the reader (Exercise 1.20.)

To verify that $\overline{\varphi} : A \to K$ is an epimorphism, consider *its* image $K' \to K$:

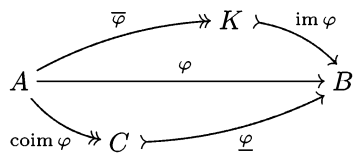
$$K' \overset{\text{im}\,\overline{\varphi}=\ker\operatorname{coker}\overline{\varphi}}{\rightarrowtail} K \overset{\operatorname{coker}\overline{\varphi}}{\twoheadrightarrow} \operatorname{Cok}\,.$$

Recall that every epimorphism is the cokernel of its kernel (Lemma 1.8); in particular, $\operatorname{coker}\overline{\varphi} = \operatorname{coker}(\operatorname{im}\overline{\varphi})$. Now, $K' \to K$ is a monomorphism through which $\overline{\varphi}$ factors:



Therefore, $K' \to B$ is a monomorphism through which φ factors and 'preceding' $\operatorname{im}\varphi : K \to B$. Since the image of φ is initial among such morphisms (as proven in Lemma 1.14), necessarily $\operatorname{im}\overline{\varphi}$ is an isomorphism. It follows that $0 = \operatorname{coker}\operatorname{im}\overline{\varphi} = \operatorname{coker}\overline{\varphi}$, proving that $\overline{\varphi}$ is an epimorphism by Lemma 1.5. □

To summarize the situation, we have the following commutative diagram of epimorphisms and monomorphisms factoring a given morphism φ in an abelian category:



Hopefully we have not lost track of our sought-for 'canonical decomposition':

$$\varphi : \quad A \twoheadrightarrow C \overset{\sim}{\underset{\tilde{\varphi}}{\longrightarrow}} K \rightarrowtail B\,.$$

We now see that this amounts to the statement that *in an abelian category, there is an induced isomorphism linking the coimage and the image of every morphism.*

Proof of Theorem 1.13. A unique morphism $\psi : K \to C$ making the above diagram commute exists, by the universal property of $\operatorname{im}\varphi$ (Lemma 1.14), as well as by the universal property of $\operatorname{coim}\varphi$. Since $\underline{\varphi} \circ \psi = \operatorname{im}\varphi$ is a monomorphism, so is ψ. Since $\psi \circ \overline{\varphi} = \operatorname{coim}\varphi$ is an epimorphism, so is ψ. It follows that ψ is an isomorphism (Lemma 1.9), and letting $\tilde{\varphi} : C \to K$ be the inverse of ψ concludes the proof. □

Loosely speaking, we can say that, in an abelian category, 'image and coimage are naturally isomorphic'. This is imprecise, since image and coimage are themselves morphisms; the isomorphism holds between the target of coimage and the source of image. Of course, for friendly categories such as R-Mod this distinction is immaterial.

Exercises

1.1. Prove that if $\psi \circ \varphi$ is an epimorphism, then ψ is an epimorphism. Prove that if $\psi \circ \varphi$ is a monomorphism, then φ is a monomorphism.

1.2. Let $\varphi : A \to B$ be a morphism in an additive category. Prove that $-\varphi$ is a monomorphism, resp., epimorphism, if and only if φ is.

1.3. ¬ A *preadditive* category is a category in which each Hom-set is endowed with an abelian group structure in such a way that composition maps are bilinear. Prove that a ring is 'the same as' a preadditive category with a single object.

Additive categories are preadditive categories with zero-objects and finite products and coproducts. Note that the notion of an 'additive functor' between *pre*additive categories makes sense. [1.12]

1.4. Let A be an additive category (preadditive would suffice), and let A be an object of A. Show that $\mathrm{End}_{\mathsf{A}}(A)$ has a natural ring structure.

1.5. Let A, B be objects of an additive category A, with zero-object 0. Since 0 is both final and initial, $\mathrm{Hom}_{\mathsf{A}}(A, 0)$ and $\mathrm{Hom}_{\mathsf{A}}(0, B)$ are both singletons, so the image of the composition

$$A \to 0 \to B$$

is a single element e of $\mathrm{Hom}_{\mathsf{A}}(A, B)$. Prove that this is the identity element of the abelian group $\mathrm{Hom}_{\mathsf{A}}(A, B)$. (Hint: Prove $e + e = e$.) In the text, this element is denoted 0, the 'zero-morphism'.

Prove that for every morphism φ in A, $\varphi \circ 0 = 0 \circ \varphi = 0$.

1.6. ▷ Prove that an object A of an additive category A is a zero-object if and only if id_A equals the zero-morphism $A \to A$. [§4.2]

1.7. Give a categorical definition of cokernel in the spirit of the definition for kernel given in Example VIII.1.11, and verify that the ordinary notion of cokernel in R-Mod satisfies this categorical requirement. Verify that your definition agrees with the one given in Definition 1.2.

1.8. Let $\varphi : A \to B$ be a morphism in an additive category A. Prove that $\iota : K \to A$ is a kernel for φ if and only if for all objects Z the induced sequence

$$0 \longrightarrow \mathrm{Hom}_{\mathsf{A}}(Z, K) \longrightarrow \mathrm{Hom}_{\mathsf{A}}(Z, A) \longrightarrow \mathrm{Hom}_{\mathsf{A}}(L, B)$$

is exact. Formulate an analogous result for cokernels.

1.9. ▷ This exercise completes the proofs of results given in the text. The arguments should be constructed so as to mirror the proofs of the companion statements given in the quoted lemmas.

Let A be an additive category.

- Let $\iota : K \to A$ be a kernel in A; prove that ι is a monomorphism. (Cf. Lemma 1.4.)

- Let $\varphi : A \to B$ be a morphism in A. If φ has a cokernel, prove that φ is an epimorphism if and only if $A \to 0$ is its cokernel. (Cf. Lemma 1.5.)
- If A is abelian, prove that every kernel in A is the kernel of its cokernel. (Cf. Lemma 1.8.)

[§1.3]

1.10. ▷ Prove that the opposite A^{op} (cf. Exercise I.3.1) of an abelian category is an abelian category. [§5.3]

1.11. ¬ Let A be an abelian category, and let C be any small category. Prove that the functor category A^{C} (cf. Exercise VIII.1.9) is an abelian category.

For every object X of C, prove that the assignment $\mathscr{F} \mapsto \mathscr{F}(X)$ (with evident action on morphisms of A^{C}, i.e., natural transformations) determines an *exact* functor $\mathscr{X} : \mathsf{A}^{\mathsf{C}} \to \mathsf{A}$. [1.12, 1.14, 1.15, 2.15, 5.7]

1.12. For a ring R, let R denote the preadditive category with a single object determined by R (Exercise 1.3). Prove that the category R-Mod of left-R-modules is equivalent to the full subcategory consisting of *additive* functors in the functor category Ab^{R} (cf. Exercise 1.11). What about the category of *right*-R-modules?

1.13. Let R be a commutative ring, and let R-Mod^f denote the category of *finitely generated* R-modules. Show that R-Mod^f need not be an abelian category; prove that it is, provided R is Noetherian.

1.14. Let T be a topological space. Recall that a *presheaf* on T with values in a category A is a contravariant functor from a certain category associated with T to A (Example VIII.1.5). Define the *category* of A-valued sheaves on T. Prove that presheaves on T with values in an abelian category form an abelian category. (Hint: Exercise 1.11.)

1.15. Consider the presheaf of continuous complex-valued functions $\mathscr{C}$ on a circle S^1 (cf. Exercise VIII.1.6) and the presheaf $\mathscr{C}^*$ of continuous complex-valued functions that do not vanish anywhere along the circle; the first is a sheaf of abelian groups (under $+$), and so is the second (under $\cdot$). Use the exponential map to define a morphism $\exp : \mathscr{C} \to \mathscr{C}^*$.

Prove that exp is *not* an epimorphism of presheaves[4]: show that its cokernel has value 0 over every proper open set of S^1 but is nonzero over S^1.

---

[4] *Watch out!* Both presheaves $\mathscr{C}, \mathscr{C}^*$ satisfy the *sheaf* condition; cf. Exercise VIII.1.6. Sheaves of abelian groups form a full subcategory of the category of presheaves, and this category is also abelian, but the inclusion functor is in general *not* exact. For example, the exponential map $\mathscr{C} \to \mathscr{C}^*$ considered above *is* an epimorphism of *sheaves*.

Also, for every open set U of T, the assignment $\mathscr{F} \mapsto \mathscr{F}(U)$ defines a functor both on the category of presheaves of abelian groups and on the category of sheaves of abelian groups. This functor is exact on presheaves (this is a particular case of Exercise 1.11), but the example given above shows that it is not exact in general on sheaves. This lack of exactness is responsible for the existence of *sheaf cohomology*.

1.16. ▷ Consider the pull-back diagram

$$\begin{array}{ccc} A \times_C B & \xrightarrow{\varphi'} & B \\ {\scriptstyle \psi'}\downarrow & \square & \downarrow{\scriptstyle \psi} \\ A & \xrightarrow[\varphi]{} & C \end{array}$$

in an abelian category; prove that the induced morphism from the source of $\ker \varphi'$ to the source of $\ker \varphi$ is an isomorphism.

State and prove an analogous result for push-outs. [§1.4, §2.2, §6.2]

1.17. ▷ Prove that the morphism $A \amalg B \to A \times B$ defined for objects A, B of an abelian category is an epimorphism. [§1.4]

1.18. Formulate a notion of 'intersection' of two monomorphisms with a common target $A \rightarrowtail Z$, $B \rightarrowtail Z$ in an abelian category. Prove that the intersection of the natural monomorphisms $A \rightarrowtail A \oplus B$, $B \rightarrowtail A \oplus B$ is 0.

1.19. ▷ Formulate the universal property satisfied by the coimage of a morphism in an abelian category. [§1.5]

1.20. ▷ Let $A \to B$ be a morphism in an abelian category, with coimage $A \to C$, and let $C \to B$ be the induced morphism. Prove that $C \to B$ is a monomorphism. (Cf. Lemma 1.16; you should not use Theorem 1.13, since Lemma 1.16 was used in its proof.) [§1.5]

1.21. ¬ Let $\varphi : A \to B$ be a morphism in an abelian category, and assume φ decomposes as an epimorphism π followed by a monomorphism i:

$$A \xrightarrow{\pi} C \xrightarrow{i} B \,.$$

Prove that necessarily $\pi = \operatorname{coim} \varphi$ and $i = \operatorname{im} \varphi$. [§7.4]

---

2. Working in abelian categories

2.1. Exactness in abelian categories. Now that we have a good notion of image in any abelian category, we can formalize the meaning of 'exact sequence'. Consider a sequence of objects and morphisms in an abelian category:

$$\cdots \longrightarrow A \xrightarrow{\varphi} B \xrightarrow{\psi} C \longrightarrow \cdots .$$

The sequence is *exact* at B if

(1) $\psi \circ \varphi = 0$ and

(2) $\operatorname{coker} \varphi \circ \ker \psi = 0$.

The first condition makes the sequence a 'complex'; it tells us that φ factors through $\ker \psi$. As $\ker \psi$ is a monomorphism, the universal property of images (Lemma 1.14) yields a unique factorization of $\operatorname{im} \varphi$ through $\ker \psi$. Similarly, the second condition

tells us that $\ker\psi$ factors through $\ker(\operatorname{coker}\varphi) = \operatorname{im}\varphi$. This implies that $\operatorname{im}\varphi$ and $\ker\psi$ coincide. The conditions defining exactness can therefore be summarized as

- $\operatorname{im}\varphi = \ker\psi$.

Of course this recovers precisely the condition used to define the notion of exactness of a sequence of R-modules, in §III.7.1.

All expected elementary connotations of exactness hold in the context of abelian categories as in R-Mod, as envisioned already in §1.2–1.3. With notation as above, if the sequence is exact at B, then

$$\psi \text{ is a monomorphism if and only if } \varphi \text{ is the zero-morphism}$$

and

$$\varphi \text{ is an epimorphism if and only if } \psi \text{ is the zero-morphism;}$$

the sequence

$$0 \longrightarrow A \xrightarrow{\varphi} B \xrightarrow{\psi} C \longrightarrow 0$$

is exact if and only if φ is a kernel of ψ and ψ is a cokernel of φ, and

$$0 \longrightarrow A \xrightarrow{\varphi} B \longrightarrow 0$$

is exact if and only if φ is an isomorphism.

Example 2.1. There is a natural exact sequence

$$0 \longrightarrow A \longrightarrow A \oplus B \longrightarrow B \longrightarrow 0$$

where $A \oplus B$ is the direct sum defined in §1.4. ⌟

Example 2.2. For a slightly more interesting example, consider a diagram

$$\begin{array}{ccc} D & \xrightarrow{\varphi'} & B \\ {\scriptstyle\psi'}\downarrow & & \downarrow{\scriptstyle\psi} \\ A & \xrightarrow[\varphi]{} & C \end{array}$$

and the associated sequence

$$D \xrightarrow{(\psi',\varphi')} A \oplus B \xrightarrow{(\varphi,-\psi)} C$$

obtained by letting $A \oplus B$ play both roles of product and coproduct. Then

- the diagram is commutative if and only if this sequence is a complex;
- the sequence obtained by adding a 0 to the left,

  $$0 \longrightarrow D \longrightarrow A \oplus B \longrightarrow C\ ,$$

  is exact if and only if D may be identified with the fibered product $A \times_C B$ (cf. Example 1.11);
- likewise, the sequence

  $$D \longrightarrow A \oplus B \longrightarrow C \longrightarrow 0$$

  is exact if and only if C may be identified with the fibered *coproduct* $A \amalg_D B$.

Indeed, the first assertion is trivial; the second and third amount to the explicit construction of fibered products and coproducts mentioned in Example 1.11. ⌟

2.2. The snake lemma, again. The reader is now in the position of making sense in any abelian category of all the statements given in §III.7 for sequences of R-modules; this includes the definition of the *homology* of a complex as a measure of the failure of exactness. These facts will be reviewed in §3. However, note that while we can now *state* the snake lemma (Lemma III.7.8) in any abelian category, the (sketch of the) *proof* given in §III.7 cannot be directly lifted from that context and applied to this one, or so it would seem, since the definition of the 'snaking' homomorphism δ given in §III.7 makes use of 'elements'. Barring more sophisticated alternatives, we should provide an arrow-theoretic alternative for the definition of δ.

This is perhaps not completely immediate.

As practice, note the following:

Lemma 2.3. *Let*

$$\begin{array}{ccc} A \times_C B & \xrightarrow{\varphi'} & B \\ {\scriptstyle \psi'}\downarrow & & \downarrow{\scriptstyle \psi} \\ A & \xrightarrow[\varphi]{} & C \end{array}$$

be a fibered diagram in an abelian category, and assume φ is an epimorphism. Then φ' is also an epimorphism.

It takes seconds to realize that this is true in R-Mod, by chasing elements. An arrow-theoretic proof is considerably subtler.

Proof. First, observe that if $\varphi : A \to C$ is an epimorphism, so is the map $A \oplus B \to C$ considered in Example 2.2. Since epimorphisms are cokernels in an abelian category and cokernels are cokernels of their kernels (Lemma 1.8), we see that $A \oplus B \to C$ is the cokernel of the natural morphism

$$A \times_C B \to A \oplus B.$$

To prove that φ' is an epimorphism, it suffices (Lemma 1.3) to show that if $\zeta : B \to Z$ is a morphism for which $\zeta \circ \varphi' = 0$, then $\zeta = 0$.

For this, consider the morphism

$$A \oplus B \xrightarrow{(0,\zeta)} Z$$

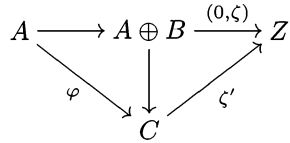
obtained by using the fact that $A \oplus B$ is a coproduct of A and B. The composition

$$A \times_C B \to A \oplus B \to Z$$

agrees with $\zeta \circ \varphi'$, so it is the zero-morphism. By the universal property of cokernels, we have the factorization

$$\begin{array}{ccc} A \oplus B & \xrightarrow{(0,\zeta)} & Z \\ \downarrow & \nearrow_{\zeta'} & \\ C & & \end{array}$$

for a unique morphism ζ'. By the commutativity of



we see that the composition $\zeta' \circ \varphi : A \to C \to Z$ is the zero-morphism. Since φ is an epimorphism, it follows that $\zeta' = 0$. This implies that $(0, \zeta) : A \oplus B \to Z$ is the zero-morphism, and we are done: $\zeta = 0$ as promised. □

Here is a quicker proof (for those who have done their homework): if φ is an epimorphism, then $A \oplus B \to C$ is an epimorphism, so the diagram is a push-out as well as a pull-back (Example 2.2). By Exercise 1.16, $\operatorname{coker} \varphi' = \operatorname{coker} \varphi = 0$; therefore φ' is an epimorphism. The reader should have no difficulty now proving the statement dual to Lemma 2.3, for fibered *coproducts* (Exercise 2.2).

Going back to the snake lemma, start from a commutative diagram linking two exact sequences in an abelian category:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & L_1 & \xrightarrow{\alpha_1} & M_1 & \xrightarrow{\beta_1} & N_1 & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \lambda} & & \downarrow{\scriptstyle \mu} & & \downarrow{\scriptstyle \nu} & & \\ 0 & \longrightarrow & L_0 & \xrightarrow{\alpha_0} & M_0 & \xrightarrow{\beta_0} & N_0 & \longrightarrow & 0 \end{array}$$

The main task is to construct a connecting morphism[5] $\delta : \ker \nu \to \operatorname{coker} \lambda$; once this is done, we should prove the exactness of the sequence displayed in Lemma III.7.8. We will leave this latter step to the more enterprising reader (who may want to wait until we go through §2.3).

The enterprising reader should in fact try to construct δ, as an exercise in coming up with arrow-theoretic proofs replacing well-understood element-theoretic arguments. Feel free to stop reading here and resume when you have tried on your own.

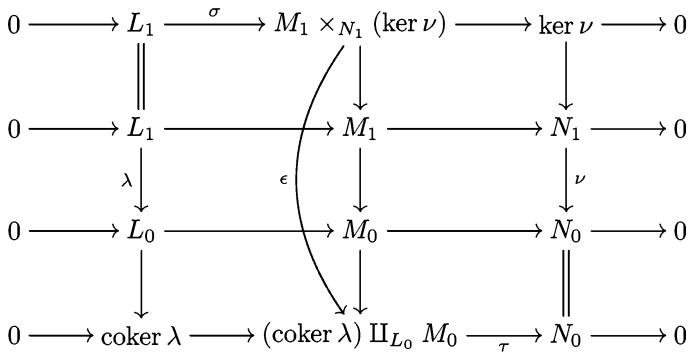
The 'problem' is that the universal properties of kernel and cokernel would seem to guarantee the existence of morphisms *to* $\ker \nu$ and *from* $\operatorname{coker} \lambda$. There's the rub: how on earth are we going to get a morphism *from* $\ker \nu$ *to* $\operatorname{coker} \lambda$? (Last chance to try on your own!)

The answer is to view $\ker \nu$ as a cokernel and $\operatorname{coker} \lambda$ as a kernel. This can be achieved by constructing suitable pull-back, resp., push-out, diagrams:

$$\begin{array}{ccc} M_1 \times_{N_1} (\ker \nu) & \xrightarrow{\beta_1'} & \ker \nu \\ \downarrow & & \downarrow \\ M_1 & \xrightarrow[\beta_1]{} & N_1 \end{array} \qquad , \qquad \begin{array}{ccc} L_0 & \xrightarrow{\alpha_0} & M_0 \\ \downarrow & & \downarrow \\ \operatorname{coker} \lambda & \xrightarrow[\alpha_0']{} & (\operatorname{coker} \lambda) \amalg_{L_0} M_0 \end{array}$$

[5] To avoid unbearably heavy notation, we will write $\ker \nu$, etc., for the *source* of the morphism $\ker \nu$, etc.

By Lemma 2.3, β_1' is an epimorphism since β_1 is an epimorphism, and $\ker \beta_1$, $\ker \beta_1'$ have matching sources by Exercise 1.16. Analogous (dual) statements hold for the second diagram (Exercises 2.2 and 1.16). Putting everything into one commutative diagram,



has exact rows (not columns!). We get a morphism

$$\epsilon : M_1 \times_{N_1} (\ker \nu) \longrightarrow (\operatorname{coker} \lambda) \amalg_{L_0} M_0 \ .$$

Note that

$$\epsilon \circ \sigma = 0, \quad \tau \circ \epsilon = 0$$

by the commutativity of the diagram: indeed, the compositions

$$L_1 \to \operatorname{coker} \lambda, \quad \ker \nu \to N_0$$

are both zero. Since $\ker \nu$ plays the role of cokernel in the top row, $\epsilon \circ \sigma = 0$ implies that ϵ must factor through $\ker \nu$, giving a morphism

$$\epsilon' : \ker \nu \longrightarrow (\operatorname{coker} \lambda) \amalg_{L_0} M_0 \ .$$

Since $\operatorname{coker} \lambda$ plays the role of kernel in the bottom row and $\tau \circ \epsilon' : \ker \nu \to N_0$ is the zero-morphism, ϵ' must factor through $\operatorname{coker} \lambda$, finally yielding

$$\delta : \ker \nu \to \operatorname{coker} \lambda.$$

Writing elements out, the reader can verify that this morphism δ agrees with the connecting morphism δ defined in §III.7.3. Together with maps induced by simpler considerations, we obtain the sequence of morphisms

$$0 \longrightarrow \ker \lambda \longrightarrow \ker \mu \longrightarrow \ker \nu \xrightarrow{\delta} \operatorname{coker} \lambda \longrightarrow \operatorname{coker} \mu \longrightarrow \operatorname{coker} \nu \longrightarrow 0$$

in our abelian category. We should now recommend that the reader prove the exactness of this sequence, by appropriate arrow-theoretic arguments. This would be excellent practice for many such facts that we will encounter when we finally do begin to develop homological algebra.

However, assuming that we have already done this work in §III.7.3, by chasing elements, is it really necessary to go back and redo it with arrows, in order to check the exactness of the sequence in an abelian category?

2.3. Working with 'elements' in a small abelian category. Our personal feeling is that arrow-theoretic arguments are invariably more insightful than their element-theoretic counterparts, even if they are 'harder' as a rule. At the very least, they are more elegant: for example, in the construction of the connecting morphism δ just given above, there was no need to verify that the result was 'well-defined', while this requires an argument in the set-theoretic construction in §III.7.3. Such issues are usually resolved once and for all at the start, when the key universal properties are established, and it is not necessary to belabor them later.

Still, it is true that element-theoretic arguments are usually easier to concoct. Lemma 2.3 appears to be a good example: in terms of elements the statement of this lemma is completely obvious, less so from the diagrammatic point of view. Exercise 2.12 gives another example: providing a purely arrow-theoretic proof of the innocent result stated there is not so easy, while this result is essentially immediate from a more mundane viewpoint. The reader could wonder whether it may not 'be enough' to prove such facts—for example, check the commutativity of a diagram or the exactness of a sequence—by simpler element-theoretic considerations.

This is indeed the case. We will present a construction which would suffice for, e.g., the purpose of completing the proof of the snake lemma by chasing elements rather than arrows. We find this construction instructive and pretty, so we will take the time to go through it in some detail. However, be warned that the Freyd-Mitchell embedding theorem (discussed in §2.4) will obliterate the actual usefulness of the construction, since it provides us with a much stronger result. Thus, we will ask the reader to wade through some material, with the understanding that the result that we will quote thereafter (Theorem 2.9) will make that material immediately obsolete. Our excuse is that while proving Theorem 2.9 is beyond the scope of this book, the construction which we are about to describe is entirely within our reach, and good propaganda for the philosophy underlying the functor of points; this is important in, e.g., algebraic geometry and takes some practice to get used to. In the end, the reader is likely to gain a better understanding of the situation by appreciating a weaker result that we can prove entirely than by leaning on a stronger result to be taken on faith. However, the hurried reader should feel free to skip this material and jump directly to §2.4.

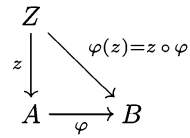
The idea is to define an appropriate notion of an 'element' of an object of an abelian category[6]. The reader should look back at the brief discussion at the end of §VIII.1.2: the 'functor of points' $h_A = \mathrm{Hom}_{\mathsf{A}}(\_, A)$ associates to every object Z the set of morphisms $Z \to A$; in several categories, applying this functor to a final object in A 'recovers' A *as a set.* Working with final objects with this purpose cannot work too well in categories with zero-objects (why?), but the idea underlying the functor of points suggests that we look at morphisms

$$z : Z \to A$$

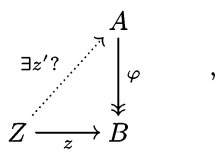
for a fixed object A in an abelian category A as (Z-flavored?!) 'elements' z of A. If $\varphi : A \to B$ is a morphism in A, it makes sense to talk about the 'element' $\varphi(z)$

[6]We will systematically put quotes around this term, i.e., write 'element', to indicate that we are referring to the relatively fancy notion we will introduce. The 'elements' will be ordinary elements of an ordinary (pointed) set determined by the object of the category.

of B: simply compose morphisms, i.e.,



It would now be nice if one could simply check the commutativity of a diagram in A or the exactness of a sequence by using these newly defined 'elements' *as if* they were elements of objects in A. This does not quite work: for example, the simple-minded statement of surjectivity at the level of these 'elements' is *not* equivalent to the notion of epimorphism in A. Indeed, if $\varphi : A \to B$ is an epimorphism, it is not necessarily the case that 'for all elements z of B there exists an element z' of A such that $z = \varphi(z')$': this would amount to saying that every morphism $z : Z \to B$ can be lifted to a morphism $z' : Z \to A$,



,

and this is simply not true in general (for example, in Ab one cannot lift the identity $\mathbb{Z}/2\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$ along the projection $\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$). The problem is really that Hom is not right-exact, so it does not preserve epimorphisms.

Maybe surprisingly, there *is* a way to fix this problem.

As a preliminary note, recall that in the distant past (Example I.3.8) we considered the category Set^* of 'pointed sets', whose objects consist of sets endowed with the choice of a distinguished point (here we will denote by 0 the distinguished point). All the algebraic structures we have encountered are pointed sets, often in more than one way: for example, stripping an abelian group of all its structure except its supporting set and the choice of the identity element leaves us with a pointed set (in other words, this defines an evident forgetful functor $\mathsf{Ab} \to \mathsf{Set}^*$). The category Set^* has a zero-object 0 (the singleton $\{0\}$); hence we can talk of 'zero-morphisms' in Set^*. Further, the category Set^* has enough structure to make sense of the notion of an 'exact sequence of pointed sets': we can stipulate that

$$\cdots \longrightarrow S \xrightarrow{f} T \xrightarrow{g} U \longrightarrow \cdots$$

is 'exact' at T if the (set-theoretic!) image of f equals the 'kernel' of g (that is, the preimage $g^{-1}(0)$). This notion should be taken with a grain of salt: while it is the case that $S \to T \to 0$ is exact at T if and only if $S \to T$ is surjective, it is not true that $0 \to T \to U$ is exact at T in Set^* if and only if $T \to U$ is injective (the only requirement posed by the exactness of $0 \to T \to U$ is that the only element of T mapping to the distinguished element in U is the distinguished element of T; this is weaker than injectivity). The category Set^* is, after all, not additive.

Now we are going to associate to every object A of a (small) abelian category A a pointed set $\hat{A}$. This will in fact give a functor $\mathsf{A} \to \mathsf{Set}^*$ with amazing properties: a morphism will be a monomorphism, resp., an epimorphism, in A *if and only if*

the corresponding function of pointed sets is injective, resp., surjective; the (arrow-theoretic) notions of kernel and image in A will correspond precisely to the (set-theoretic) kernel and image in Set^*; a sequence of objects in A will be exact *if and only if* the corresponding sequence of pointed sets is exact; and it will be possible to verify whether a diagram commutes in A by working in Set^*.

The upshot will be that in a sense we *can* chase a diagram in any small abelian category by 'chasing elements'. For example, working with 'elements' suffices to complete the proof of the snake lemma, in the sense that the exactness (in A) of the sequence obtained with the aid of the connecting morphism defined in §2.2 may be verified by chasing 'elements', since this verifies exactness in Set^*. A proof in R-Mod would transfer word-for-word to give a proof in any small[7] abelian category. This hints that the arrow-theoretic viewpoint that we have tried to defend in the past several subsections, while elegant, is not really strictly necessary; we will come back to this point in §2.4.

The construction can be given as an application of the abstract language developed in §VIII.1; but we promise that the resulting notion will be easy to contemplate, independently of this language. The idea is simply to take the functor of points *up to a relation* identifying the 'element' $Z \to A$ with all 'elements' $W \twoheadrightarrow Z \to A$.

The following abstract nonsense will have a simple-minded explanation. We will (often silently) assume that A is small, to steer clear of any possible set-theoretic subtlety. It is convenient to consider a companion category $\mathsf{A}_{\twoheadleftarrow}$, with the same objects as A but in which

$$\mathrm{Hom}_{\mathsf{A}_{\twoheadleftarrow}}(Z, W) = \{\text{epimorphisms } W \twoheadrightarrow Z \text{ in } \mathsf{A}\}.$$

This is legal, because the composition of two epimorphisms in A is an epimorphism; but note that $\mathsf{A}_{\twoheadleftarrow}$ is no longer additive (the sum of two epimorphisms need not be an epimorphism, so we lose the algebraic structure on Hom). Also note the reversing of arrows; this is in order that the functor that we are about to define be *covariant*.

We will use $\mathsf{A}_{\twoheadleftarrow}$ as 'index' category for a limit. For a fixed object A of A, we define a covariant functor

$$\mathscr{H}_A : \mathsf{A}_{\twoheadleftarrow} \to \mathsf{Set}^*$$

by 'restricting' the functor of points h_A to $\mathsf{A}_{\twoheadleftarrow}$ and preserving the information of the zero-morphism. That is, we set

$$\mathscr{H}_A(Z) = \mathrm{Hom}_{\mathsf{A}}(Z, A)$$

for all objects Z of $\mathsf{A}_{\twoheadleftarrow}$ (that is, of A), viewing $\mathrm{Hom}_{\mathsf{A}}(Z, A)$ as a pointed set by distinguishing the zero-morphism, and define $\mathscr{H}_A$ on morphisms by composition: for a morphism $Z \to W$ in $\mathsf{A}_{\twoheadleftarrow}$ (that is, an epimorphism $\alpha : W \twoheadrightarrow Z$),

$$\begin{array}{ccc} W & \xrightarrow{\alpha} & Z \\ & {\scriptstyle w}\searrow \quad \swarrow{\scriptstyle z} & \\ & A & \end{array} \quad ,$$

[7] If the relevant diagrams only involve finitely many objects, the smallness hypothesis is not restrictive. The interested reader should research this issue; we will shamelessly gloss over it.

define

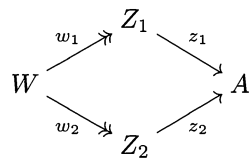
$$\mathscr{H}_A(Z) \to \mathscr{H}_A(W)$$

by sending z to $w = z \circ \alpha$. Of course the zero-morphism is mapped to the zero-morphism, so this is indeed a morphism in Set^*.

Definition 2.4. Let A be an object of a small abelian category A. The pointed set $\hat{A}$ of 'elements of A' is the colimit of $\mathscr{H}_A$:

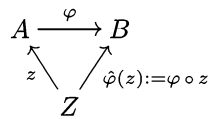
$$\hat{A} := \varinjlim \mathscr{H}_A. \qquad \lrcorner$$

Here is the simple-minded explanation. In the context of pointed sets, colimits may be constructed by joining pointed sets at the distinguished point and then taking a suitable quotient. For the case at hand, this translates into the following: an 'element' of A consists of the choice of a morphism $z : Z \to A$ in A, stipulating that two morphisms $z_1 : Z_1 \to A$, $z_2 : Z_2 \to A$ determine the same 'element' if there are *epi*morphisms $w_1 : W \twoheadrightarrow Z_1$ and $w_2 : W \twoheadrightarrow Z_2$ in A such that the diagram



commutes. This is an equivalence relation, as the reader should check carefully (Exercise 2.5).

Now we want to make the assignment $A \mapsto \hat{A}$ into a covariant functor $\mathsf{A} \to \mathsf{Set}^*$, and there is only one reasonable way to do so: for $\varphi : A \to B$ and $z : Z \to A$, define $\hat{\varphi}(z)$ by composing morphisms, i.e.,



This prescription is (manifestly) compatible with the equivalence relation, so it defines a set-function $\hat{\varphi} : \hat{A} \to \hat{B}$. The image of the zero-morphism is zero, so $\hat{\varphi}$ is a morphism in Set^*, as needed. The covariance property of this assignment is evident. We have obtained the promised functor $\mathsf{A} \to \mathsf{Set}^*$.

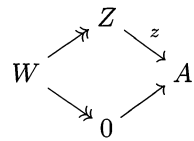
Now we will proceed to verify all the miracles we have advertised. In the following, the symbol $\sim$ will stand for the equivalence relation defined above[8]; thus, two morphisms $z_1 : Z_1 \to A$, $z_1 : Z_2 \to A$ determine the same 'element' of $\hat{A}$ if and only if $z_1 \sim z_2$.

One useful preliminary observation is that the only morphism representing the distinguished 'element' is the zero-morphism and that we can detect whether a morphism is 0 by working in Set^*:

[8]Warning: Later on in this chapter we will define a notion of *homotopy,* and we will recycle the symbol $\sim$ for that notion. The context and the meaning of the symbol will be completely different.

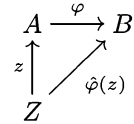
Lemma 2.5. $z \sim 0 \iff z = 0$. *Further, a morphism* $\varphi : A \to B$ *in* A *is* 0 *if and only if* $\hat{\varphi}(z) = 0$ *for all* $z \in \hat{A}$.

Proof. According to the definition given above, $z : Z \to A$ is equivalent to 0 if and only if there is an epimorphism $W \twoheadrightarrow Z$ making the following diagram commute:



Since $W \twoheadrightarrow Z$ is an epimorphism, this diagram commutes if and only if $z = 0$ (Lemma 1.3), as claimed.

For the second statement, consider the situation



If $\varphi = 0$, then so is $\hat{\varphi}(z) = \varphi \circ z$. Conversely, if $\hat{\varphi}(z) = 0$ for all z, then taking $z : Z = A \to A$ to be the identity, we get $\varphi \sim 0$, and hence $\varphi = 0$ by the first part. □

This tells us in particular that we can verify the commutativity of a diagram in A by working in Set^*: because we can check whether two morphisms $\varphi, \psi \in \mathrm{Hom}_{\mathsf{A}}(A, B)$ are equal by verifying that their difference $\varphi - \psi$ equals 0, which we can do[9] at the level of pointed sets (hence, by chasing elements!) by Lemma 2.5.

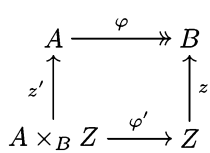
Concerning monomorphisms and epimorphisms,

Lemma 2.6. *Let* $\varphi : A \to B$ *be a morphism in* A. *Then*

- φ *is a monomorphism if and only if* $\hat{\varphi}$ *is injective;*
- φ *is an epimorphism if and only if* $\hat{\varphi}$ *is surjective.*

Proof. As usual, we will propose a division of labor: the reader will prove the first statement (Exercise 2.6), and we will prove the second.

Assume φ is an epimorphism, and let $z : Z \to B$ represent an arbitrary 'element' of $\hat{B}$. Consider the fiber product:



By Lemma 2.3, φ' is an epimorphism. It follows that $\varphi \circ z' \sim z$, that is, $\hat{\varphi}(z') = z$. This shows that $\hat{\varphi}$ is surjective.

[9]Note that we are not saying that $\varphi = \psi$ if and only if $\hat{\varphi}(z) = \hat{\psi}(z)$ for all 'elements' z; this is unfortunately not the case (cf. Exercise 2.10).

Conversely, assume $\hat{\varphi} : \hat{A} \to \hat{B}$ is surjective. In particular, there is an 'element' of $\hat{A}$ mapping to id_B. This 'element' is represented by a morphism $z : Z \to A$, and the condition $\hat{\varphi}(z) = \mathrm{id}_B$ means that there are epimorphisms w_1, w_2, as in the commutative diagram:

$$\begin{array}{ccccc} A & & \xrightarrow{\quad\varphi\quad} & & B \\ {\scriptstyle z}\uparrow & & & & \Vert{\scriptstyle \mathrm{id}_B} \\ Z & \xleftarrow{w_1} & W & \xrightarrow{w_2} & B \end{array}$$

Since $\varphi \circ z \circ w_1 = w_2$ is an epimorphism, it follows that φ is an epimorphism, as needed. □

Next, let's verify that the notions of 'kernel' and 'image' in an abelian category A match precisely their simple-minded counterparts in Set^*.

Lemma 2.7. *With notation as above, let $\varphi : A \to B$ be a morphism in a small abelian category A, and let $\hat{\varphi} : \hat{A} \to \hat{B}$ be the corresponding function of pointed sets. Let $\ker\varphi : K \to A$, resp., $\mathrm{im}\,\varphi : I \to B$, be the kernel and the image of φ, respectively. Then*

- *$\widehat{\ker\varphi}$ identifies $\hat{K}$ with the subset $\hat{\varphi}^{-1}(0)$ of $\hat{A}$;*
- *$\widehat{\mathrm{im}\,\varphi}$ identifies $\hat{I}$ with the image of $\hat{\varphi}$ in $\hat{B}$.*

Proof. These statements are very close to the universal properties satisfied by kernel and image.

The reader will verify the statement about the kernel (Exercise 2.7). For the image, recall that we have a decomposition of φ,

$$\varphi : A \twoheadrightarrow I \overset{\mathrm{im}\,\varphi}{\rightarrowtail} B\,,$$

and that $\mathrm{im}\,\varphi : I \to B$ is a kernel of $\mathrm{coker}\,\varphi$ (this is the definition of 'image' (cf. §1.5); the fact that $A \to I$ is an epimorphism is part of the content of Theorem 1.13).

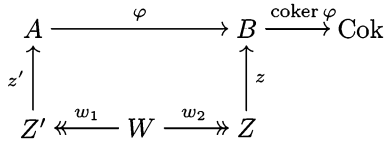
To simplify the notation, let $j = \mathrm{im}\,\varphi$. Then (by Lemma 2.6) $\hat{j}$ is an injective function $\hat{I} \to \hat{B}$, mapping a representative $z : Z \to I$ to $j \circ z : Z \to B$; we have to verify that $\hat{j}(\hat{I}) = \mathrm{im}\,\hat{\varphi}$. To verify that $\hat{j}(z)$ is in the image of $\hat{\varphi}$, consider the fiber product

$$\begin{array}{ccccc} & & \overset{\varphi}{\frown} & & \\ A & \twoheadrightarrow & I & \overset{j}{\rightarrowtail} & B \\ {\scriptstyle z'}\uparrow & & \uparrow{\scriptstyle z} & \nearrow{\scriptstyle \hat{j}(z)} & \\ A \times_I Z & \twoheadrightarrow & Z & & \end{array}$$

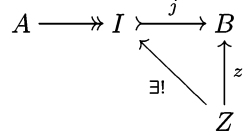
The bottom map is an epimorphism by virtue of Lemma 2.3, since the top map in the fiber square is an epimorphism. This shows that $\varphi \circ z' \sim \hat{j}(z)$; that is, $\hat{j}(z) = \hat{\varphi}(z')$ is in the image of $\hat{\varphi}$, as claimed. This verifies that the image of $\hat{j}$ is contained in the image of $\hat{\varphi}$:

$$\hat{I} \to \text{image of } \hat{\varphi} \subseteq \hat{B}.$$

To show that every 'element' in the image of $\hat{\varphi}$ is obtained in this fashion, let $z : Z \to B$ be in the image of $\hat{\varphi}$. That is, there is a $z' : Z' \to A$ and epimorphisms w_1, w_2 making the following diagram commute:



Note that $\operatorname{coker}\varphi \circ z \circ w_2 = (\operatorname{coker}\varphi \circ \varphi) \circ z' \circ w_1 = 0$; since w_2 is an epimorphism, it follows that $\operatorname{coker}\varphi \circ z = 0$. By the universal property of kernels, this says that z factors uniquely through $\ker\operatorname{coker}\varphi = \operatorname{im}\varphi = j$:



This gives an 'element' of $\hat{I}$ mapping to z, as needed. □

Finally, let's check that the functor is exact:

Proposition 2.8. *Let* A *be a small abelian category. Then a sequence*

$$A \xrightarrow{\varphi} B \xrightarrow{\psi} C$$

in A *is exact if and only if the corresponding sequence*

$$\hat{A} \xrightarrow{\hat{\varphi}} \hat{B} \xrightarrow{\hat{\psi}} \hat{C}$$

is exact in Set^*.

Proof. This now follows immediately from Lemma 2.7: exactness in A means that $\operatorname{im}\varphi = \ker\psi$, and exactness in Set^* means that the image of $\hat{\varphi}$ equals $\hat{\psi}^{-1}(0)$. By Lemma 2.7, these conditions are equivalent. □

2.4. What is missing? Pretty as it is, the construction presented in the previous section has several shortcomings:

- The structure of 'pointed set' is less rigid than, say, that of an abelian group; since Ab is itself an abelian category, it would be more natural to land in Ab than Set^*, if possible[10].
- While we can check that a diagram commutes in A by performing some computation in Set^* (cf. the comments following Lemma 2.5), we have stopped short of claiming that the functor $A \to \hat{A}$ is *faithful,* that is, that it induces *injective* functions $\operatorname{Hom}_{\mathsf{A}}(A, B) \to \operatorname{Hom}_{\mathsf{Set}^*}(\hat{A}, \hat{B})$ for all A, B. In fact, the functor is *not* faithful (Exercise 2.10).

[10]One way to do this would be to take the colimit in Ab rather than Set^*, but we will not pursue this possibility here.

- The functor $\mathsf{A} \to \mathsf{Set}^*$ is not *full;* that is, it is not surjective on Hom-sets. In other words, there will be many functions of pointed sets that are not induced by morphisms in A. Thus, we cannot use 'elements' to *construct* morphisms in A: the arrow-theoretic work performed in §2.2 in order to construct the connecting morphism remains necessary (even if the properties of this morphism can then be verified by using 'elements').

Here another miracle occurs: it is in fact possible to construct *fully faithful, exact* functors from any given (small) abelian category to a category of modules. Here is the precise statement, which will be left without proof in this book:

Theorem 2.9 (Freyd-Mitchell theorem)**.** *Let* A *be a small abelian category. Then there is a fully faithful, exact functor* $\mathsf{A} \to R\text{-}\mathsf{Mod}$ *for a suitable ring* R*.*

Caveat: In this statement R is not necessarily commutative; R-Mod denotes the category of left-R-modules.

For a reminder on the *full & faithful* terminology, see Definition VIII.1.6. Briefly, a functor $\mathscr{F} : \mathsf{A} \to \mathsf{C}$ is fully faithful if it induces a bijection $\mathrm{Hom}_{\mathsf{A}}(A, B) \to \mathrm{Hom}_{\mathsf{C}}(\mathscr{F}(A), \mathscr{F}(B))$ for all objects A, B in A. In practice, such a functor identifies A with a subcategory of C. The exactness (cf. Example VIII.1.18) of the functor $\mathsf{A} \to R\text{-}\mathsf{Mod}$ ensures that kernels and cokernels match in the two environments, and so on (note that the functor is automatically *faithfully* exact; see Exercise 2.16): everything we laboriously checked in §2.3 for our poor man's functor $\mathsf{A} \to \mathsf{Set}^*$ holds for the much fancier functor provided by the Freyd-Mitchell theorem.

This functor *is* fully faithful: this means that one can in fact construct morphisms in an arbitrary (small) abelian category by working with elements. Indeed, this amounts to constructing the appropriate morphisms in the ambient category R-Mod, and fullness guarantees that these morphisms 'already' exist in A.

So, at the end of the day we are really allowed to forget all our arrow-theoretic work. Element-theoretic considerations suffice, even for the purpose of constructing morphisms such as the 'connecting morphism' discussed in §2.2. Monomorphisms really *can* be assimilated with 'subobjects', and 'quotients' in an abelian category really *are* quotients in the ordinary sense.

This is impressive: it says that doing homological algebra in the category of left-modules over a ring is essentially as general as doing it in an arbitrary (small) abelian category, at least for what concerns constructions that do not require the category to have special objects (such considerations will play a role in §5 and following). In the rest of the chapter we will adopt the common compromise of discussing the basic notions of homological algebra in the context of abelian categories while reverting back to element-theoretic considerations whenever this makes life substantially easier. The Freyd-Mitchell theorem allows us to do this unapologetically, and often so would the simpler considerations of §2.3.

Exercises

2.1. In an abelian category, prove that $A \xrightarrow{\varphi} B \xrightarrow{\psi} C$ is exact at B if and only if ψ and $\operatorname{coker}\varphi$ (both of which are morphisms with B as source) have the same kernel, if and only if φ and $\ker\psi$ (both of which are morphisms with B as target) have the same cokernel.

2.2. ▷ Consider a push-out diagram

$$\begin{array}{ccc} D & \xrightarrow{\alpha} & B \\ {\scriptstyle\beta}\downarrow & & \downarrow{\scriptstyle\beta'} \\ A & \xrightarrow[\alpha']{} & A \amalg_D B \end{array}$$

in an abelian category. Prove that if α is a monomorphism, then α' is a monomorphism. (Cf. Lemma 2.3.) [§2.2]

2.3. ¬ Prove the 'four-lemma' (cf. Exercises III.7.12 and III.7.13) in every abelian category. In fact, show that it is only necessary to prove one of the two forms of the lemma, for the other then follows automatically. [III.7.13]

2.4. ¬ Prove the 'short five-lemma' (cf. Exercise III.7.10) in any abelian category: consider a commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & L_1 & \longrightarrow & M_1 & \longrightarrow & N_1 & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle\lambda} & & \downarrow{\scriptstyle\mu} & & \downarrow{\scriptstyle\nu} & & \\ 0 & \longrightarrow & L_0 & \longrightarrow & M_0 & \longrightarrow & N_0 & \longrightarrow & 0 \end{array}$$

with exact rows in any abelian category, and assume that λ and ν are isomorphisms; prove that μ is an isomorphism, by an explicit arrow-theoretic chase of the diagram. [2.11]

2.5. ▷ Prove that the relation used in §2.3 to define the notion of 'element' of an object of an abelian category is indeed an equivalence relation. [§2.3]

2.6. ▷ Prove that a morphism $\varphi : A \to B$ in a small abelian category is a monomorphism if and only if it is injective on 'elements'. (Cf. Lemma 2.6.) [§2.3]

2.7. ▷ Let $\varphi : A \to B$ be a morphism in a small abelian category, and let $\ker\varphi : K \to A$ be a kernel of φ. Prove that the set of 'elements' of K may be identified with $\hat{\varphi}^{-1}(0)$, where $\hat{\varphi} : \hat{A} \to \hat{B}$ is the induced function on the corresponding sets of elements. (Cf. Lemma 2.7.) [§2.3]

2.8. ¬ Let A be an abelian category, and let A be an object of A. Prove that $\mathrm{id}_A : A \to A$ and $-\mathrm{id}_A : A \to A$ determine the same 'element' of A in the sense of §2.3. [2.10]

2.9. Prove that the equivalence relation $\sim$ introduced in §2.3 does not preserve the addition in Hom-sets, in the sense that $\varphi \sim \psi \not\Rightarrow (\varphi + \eta) \sim (\psi + \eta)$.

2.10. ▷ Let A denote an abelian category. Prove that the functor $\mathsf{A} \to \mathsf{Set}^*$, $A \mapsto \hat{A}$ defined in §2.3 is not faithful in general. (Use Exercise 2.8.) [§2.3, §2.4]

2.11. Upgrade the result of Exercise 2.4, by proving the *five-lemma*: if

$$\begin{array}{ccccccccc} K_1 & \longrightarrow & L_1 & \longrightarrow & M_1 & \longrightarrow & N_1 & \longrightarrow & O_1 \\ \downarrow & & \downarrow{\scriptstyle\sim} & & \downarrow{\scriptstyle\mu} & & \downarrow{\scriptstyle\sim} & & \downarrow \\ K_0 & \longrightarrow & L_0 & \longrightarrow & M_0 & \longrightarrow & N_0 & \longrightarrow & O_0 \end{array}$$

is a commutative diagram with exact rows in a (small) abelian category, with monomorphisms, epimorphisms, isomorphisms as indicated, then μ is an isomorphism. Use 'elements' and the material developed in §2.3.

2.12. ▷ Let $\lambda : M \to L$, $\nu : M \to N$ be morphisms in an abelian category:

$$L \xleftarrow{\lambda} M \xrightarrow{\nu} N\,.$$

Prove that[11]

$$\frac{\operatorname{im}\lambda}{\lambda(\ker\nu)} \cong \frac{\operatorname{im}\nu}{\nu(\ker\lambda)}.$$

(Hint: By the considerations in this section, this can be proved by using elements; in fact, by the Freyd-Mitchell theorem it suffices to prove it in R-Mod.) [§2.3, §9.3]

2.13. By refining the construction in §2.3, one can show that every small abelian category can be embedded in Ab. However, it is not always possible to do it *fully*.

Indeed, prove that the abelian category of finite-dimensional $\mathbb{R}$-vector spaces is not equivalent to any *full* subcategory of Ab. (Hint: If it were, the one-dimensional vector space $\mathbb{R}$ would correspond to an abelian group A such that $\operatorname{End}_{\mathsf{Ab}}(A) \cong \operatorname{End}_{\mathbb{R}\text{-}\mathsf{Vect}}(\mathbb{R})$. But look at Exercise VI.1.13.)

2.14. ¬ Let A be an abelian category, and let A' be a full subcategory of A. Prove that A' is abelian and the inclusion functor $\mathsf{A}' \subseteq \mathsf{A}$ is exact (and in particular it preserves kernels and cokernels) if and only if

- A' contains the zero-object of A,
- A' is closed under $\oplus$, and
- A' is closed under kernels and cokernels.

[2.15, 2.17, 3.3]

2.15. ¬ Let A be a small abelian category. Construct a category F of additive contravariant functors $\mathsf{A} \to \mathsf{Ab}$, with natural transformations as morphisms, and show that it is abelian. (Use Exercises 1.11 and 2.14.) [2.17]

2.16. ▷ Let A, B be abelian categories, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be a faithful, exact functor. Prove that it is faithfully exact: a sequence $X \to Y \to Z$ is exact in A *if and only if* the corresponding sequence $\mathscr{F}(X) \to \mathscr{F}(Y) \to \mathscr{F}(Z)$ is exact in B. [§2.4]

[11] Here $\lambda(\ker\nu)$ denotes the image of the restriction of λ to the source of $\ker\nu$, etc.; see Remark 1.7 for 'quotients'.

2.17. Upgrade the Yoneda lemma (Exercise VIII.1.10) to prove that every small abelian category A is equivalent to a full subcategory of the category F of Exercise 2.15, by means of the functor assigning to each object X in A the functor $h_X = \mathrm{Hom}_{\mathsf{A}}(\_, A)$.

Prove that this Yoneda embedding is *left*-exact and 'reflects exactness' in the sense that $X \to Y \to Z$ is exact in A if the corresponding sequence $h_X \to h_Y \to h_Z$ is exact at h_Y (that is, if $h_X(A) \to h_Y(A) \to h_Z(A)$ is exact at $h_Y(A)$ for all A in A).

This is the beginning of a proof of the Freyd-Mitchell theorem. Since each h_X is left-exact, the Yoneda embedding lands in the subcategory L of F consisting of left-exact additive contravariant functors. It turns out that L is abelian in its own right (although its embedding in F is not exact; thus, Exercise 2.14 doesn't help), and the Yoneda embedding of A in L *is* exact. Finally, one proves that L is equivalent to the category of modules over a ring (the endomorphism ring of a 'faithfully projective' object), and the Freyd-Mitchell theorem follows.

3. Complexes and homology, again

After all these preliminary considerations, it is time to begin thinking in earnest about homological algebra. We will be working in a fixed abelian category A; by what we have seen in §2, we are allowed to pretend that the objects of A are ordinary modules over a ring, at least if A is small. Thus we will deal with objects and morphisms with the aid of usual set-theoretic considerations[12]. Until we get to more sophisticated material, the reader will lose little or nothing by taking A to be R-Mod for some ring[13] R.

We will start off with a reminder of what we did in §III.7.1. Once more, as we now know, everything can in fact be done in the more general setting of an abelian category.

3.1. Reminder of basic definitions; general strategy. A chain *complex* $(M_\bullet, d_\bullet)$ in A is a sequence of objects and morphisms,

$$\cdots \xrightarrow{d_{i+2}} M_{i+1} \xrightarrow{d_{i+1}} M_i \xrightarrow{d_i} M_{i-1} \xrightarrow{d_{i-1}} \cdots$$

such that $(\forall i)$, $d_i \circ d_{i+1} = 0$. We can just as well use ascending indices (which are then traditionally written as superscripts),

$$\cdots \xrightarrow{d^{i-2}} M^{i-1} \xrightarrow{d^{i-1}} M^i \xrightarrow{d^i} M^{i+1} \xrightarrow{d^{i+1}} \cdots,$$

and impose $d^i \circ d^{i-1} = 0$, with no change in the mathematics other than in the nomenclature. This is a *co*chain complex $(M^\bullet, d^\bullet)$ (or $M^\bullet$ for short), leading to

[12] In so doing, we will sweep the 'smallness' hypothesis under the carpet. This imprecision should be harmless in any matter concerning only countably many objects of A, such as constructing a diagram or checking that a diagram commutes. We will not deal with anything fancier than this.

[13] In fact the reader may even assume that R is a *commutative* ring; commutativity plays no essential role in the considerations in this chapter.

*co*homology, etc. The morphisms d^i are the *differentials* of the complex. It is not uncommon to call 'd' *all* differentials within sight, decorating them by the name of the complex if necessary to avoid confusion. So, $d^\bullet_{M^\bullet}$ would implicitly denote the differential of the complex $M^\bullet$.

Whether to work 'homologically' or 'cohomologically' is essentially an æsthetic question, until it becomes dictated by the specific context of an application. In §III.7 we chose homology; in this chapter we will choose cohomology: thus, indices will be increasing upper indices. Setting $M_i = M^{-i}$ reconciles the two conventions. Thus, a chain complex $(M_\bullet, d_\bullet)$ may be (and usually is) viewed as the cochain complex $(M^{-\bullet}, d^{-\bullet})$.

The condition $d^i \circ d^{i-1} = 0$ is equivalent to $\operatorname{im} d^{i-1} \subseteq \ker d^i$. The complex is *exact* at M^i if $\operatorname{im} d^{i-1} = \ker d^i$; it is 'exact' if it is exact everywhere. The *i-th cohomology* of a complex $(M^\bullet, d^\bullet)$ measures its 'deviation from exactness' at M^i:

$$H^i(M^\bullet) := \frac{\ker d^i}{\operatorname{im} d^{i-1}}.$$

Setting $H_i(M_\bullet) = H^{-i}(M^{-\bullet})$ gives homology from cohomology, if desired.

Remark 3.1. In an abelian category, the definition of cohomology should be parsed as follows (cf. Remark 1.7). Let $I \to M^i$ be $\operatorname{im} d^{i-1}$, and let $K \to M^i$ be $\ker d^i$; since $M^\bullet$ is a complex, $\operatorname{im} d^{i-1}$ factors through $\ker d^i$, giving a monomorphism $I \to K$; then $H^i(M^\bullet)$ is the cokernel of this morphism. We will use the 'quotient' notation as a good shorthand for this operation. ⌟

The following definition will undergo a few refinements later in the chapter:

Definition 3.2. (Preliminary) A *resolution* of an object A is a complex whose cohomology is concentrated in degree 0 and isomorphic to A. ⌟

To be more precise, the datum of a resolution of A should include a specific isomorphism of the cohomology of the complex with A; a common abuse of language glosses over this point. In fact, this isomorphism in cohomology should be induced by a morphism at the level of complexes, from the given complex to the complex called $\iota(A)$ later in this section. The notion of *quasi-isomorphism* (Definition 4.3) will give us the correct framework in which to talk about resolutions precisely.

It is common to assume that resolutions $M^\bullet$ are 'bounded' above or below, that is, $M^i = 0$ for $i > 0$ or $i < 0$. For example, a complex

$$\cdots \xrightarrow{d^{-3}} M^{-2} \xrightarrow{d^{-2}} M^{-1} \xrightarrow{d^{-1}} M^0 \xrightarrow{d^0} 0 \longrightarrow 0 \longrightarrow \cdots$$

is a resolution of A if $\operatorname{coker} d^{-1} \cong A$, and the complex is exact otherwise. Down the road we will restrict our attention to bounded (-above or -below) resolutions. In a sense, this is not a very serious restriction: any resolution may be truncated (at the price of changing the degree-0 term) in order to produce a bounded resolution (Exercise 3.1). Note that the datum of a bounded-above resolution is equivalent to that of an exact complex

$$\cdots \longrightarrow M_2 \longrightarrow M_1 \longrightarrow M_0 \longrightarrow A \longrightarrow 0\ ;$$

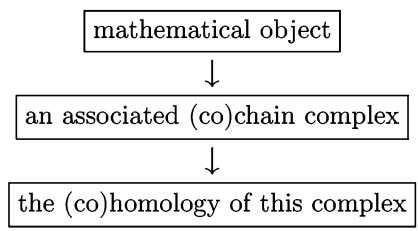
this is the viewpoint we took in our previous encounter with resolutions, in §VI.4.2.

The complex $\iota(A)$,

$$\cdots \longrightarrow 0 \longrightarrow 0 \xrightarrow{d^1} A \xrightarrow{d^0} 0 \longrightarrow 0 \longrightarrow \cdots,$$

is a resolution of A, but not a very interesting one. Resolutions become useful if the objects M^i are 'special' in some sense (usually, A should not be expected to be special): for example, in the context of modules over a ring we could require the objects M^i to be free, flat, projective, etc. We will then say that the resolution itself is free, flat, projective, etc.

There are many contexts in which the following strategy produces interesting invariants:



Both Ext and Tor are examples of this general plan of attack: start with two R-modules M, N; find (for example) a free resolution of M, and tensor it by N, obtaining a chain complex; then the homology of this complex computes the modules $\mathrm{Tor}_i^R(M,N)$ (cf. §VIII.2.4). Algebraic topology is a source of many other examples and was in fact historically the main motivation driving the development of homological algebra. In the 1950's it became clear that this tool would be extremely useful in other fields; for example, it had a tremendous impact on algebraic geometry, through the work of Alexander Grothendieck[14] and others.

Now consider this strategy carefully. In every single useful application, there are in fact *many ways* to associate a complex to a mathematical object: in the example of Tor recalled above, there are infinitely many different free resolutions of a given module; we have in fact claimed (in §VIII.6.4) that resolutions by *projective* modules, or even *flat* modules, would all yield the same Tor modules. In general, there is a gigantic degree of freedom in the choice of the appropriate complex associated to the given mathematical object, and one of our main goals here will be to show that (for a large class of interesting examples) the *(co)homology* of the complex will not depend on this choice.

The inquiring reader should not stop there, however. Precisely because of this large ambiguity, the informational gap from the complex associated to a mathematical entity to its cohomology is large. Could one do better than 'taking cohomology'? The natural approach would be to determine precisely where the ambiguity lies and 'mod out' the choice of a complex by this ambiguity. Whatever one gets from this, it must be enough to determine the cohomology; but there may be a large amount

[14]Grothendieck's 1957 paper "Sur quelques points d'algèbre homologique" was so influential that you can just refer to it by naming the journal in which it appeared. If you say "This is proven in *Tôhoku*", without other qualifiers, most algebraic geometers will understand that you mean this one paper of the thousands published in this journal in the past 50 years.

of interesting information carried by 'complexes modulo ambiguity' which is lost by going 'all the way down' to cohomology.

This is what we are up to: proving that the cohomology of the associated complex is indeed independent of the choices, while keeping an eye out to detect where the key information is really stored.

3.2. The category of complexes. To being this exploration, we assemble cochain complexes in A into a *new* category $\mathsf{C}(\mathsf{A})$:

- $\mathrm{Obj}(\mathsf{C}(\mathsf{A})) = \{\text{cochain complexes in } \mathsf{A}\}$;
- for $M^\bullet = (M^\bullet, d^\bullet_{M\bullet})$, $N^\bullet = (N^\bullet, d^\bullet_{N\bullet})$ cochain complexes, $\mathrm{Hom}_{\mathsf{C}(\mathsf{A})}(M^\bullet, N^\bullet)$ consists of the *commutative*[15] diagrams

$$(*) \qquad \begin{array}{ccccccccc} \cdots & \longrightarrow & M^{i-1} & \xrightarrow{d^{i-1}_{M^\bullet}} & M^i & \xrightarrow{d^i_{M^\bullet}} & M^{i+1} & \longrightarrow & \cdots \\ & & \downarrow{\scriptstyle \alpha^{i-1}} & & \downarrow{\scriptstyle \alpha^i} & & \downarrow{\scriptstyle \alpha^{i+1}} & & \\ \cdots & \longrightarrow & N^{i-1} & \xrightarrow{d^{i-1}_{N^\bullet}} & N^i & \xrightarrow{d^i_{N^\bullet}} & N^{i+1} & \longrightarrow & \cdots \end{array}$$

in A. We will denote by $\alpha^\bullet$ the morphism determined by the collection α^i. Everything that should be checked in order to prove that $\mathsf{C}(\mathsf{A})$ is indeed a category should be evident. The following is also essentially evident, but a little more interesting:

Lemma 3.3. $\mathsf{C}(\mathsf{A})$ *is an abelian category.*

We will leave to the reader the careful verification of this fact (Exercise 3.3). In broad terms, morphisms between two given complexes form an abelian group, essentially because if α^i and $\beta^i : M^i \to N^i$ are both collections of morphisms making the appropriate diagram commute, so is the collection of their sums $\alpha^i + \beta^i$. Finite products and coproducts exist in $\mathsf{C}(\mathsf{A})$ as an easy consequence of the fact that they do in A. As for kernels and cokernels, the commutativity of each square in (*) and the universal properties of kernels and cokernels in A guarantee the existence of morphisms making the larger diagrams

$$\begin{array}{ccc} K^i & \xrightarrow{\exists!} & K^{i+1} \\ \downarrow{\scriptstyle \ker \alpha^i} & & \downarrow{\scriptstyle \ker \alpha^{i+1}} \\ M^i & \xrightarrow{d^i_{M^\bullet}} & M^{i+1} \\ \downarrow{\scriptstyle \alpha^i} & & \downarrow{\scriptstyle \alpha^{i+1}} \\ N^i & \xrightarrow{d^i_{N^\bullet}} & N^{i+1} \\ \downarrow{\scriptstyle \mathrm{coker}\, \alpha^i} & & \downarrow{\scriptstyle \mathrm{coker}\, \alpha^{i+1}} \\ C^i & \xrightarrow{\exists!} & C^{i+1} \end{array}$$

[15] Note the requirement that such diagrams *commute*. It will also be important to consider diagrams that do not satisfy this requirement, but 'morphisms of cochain complexes' do. Actually, a strong case could be made for requiring morphisms of cochain complexes to correspond to *anti*-commutative diagrams, but conventions are what they are and we will adhere to them.

commute. The resulting sequences $K^\bullet$, $C^\bullet$ are immediately checked to be complexes, and the collections $\ker \alpha^i : K^i \to M^i$, $\operatorname{coker} \alpha^i : N^i \to C^i$ give morphisms of complexes $\ker \alpha^\bullet$, $\operatorname{coker} \alpha^\bullet$ satisfying the universal property for kernels and cokernels in $\mathsf{C}(\mathsf{A})$. The reader will check that every monomorphism in $\mathsf{C}(\mathsf{A})$ is a kernel and every epimorphism is a cokernel by using the fact that this is the case in A.

Thus, we have a way to concoct a new abelian category from an old one. The new category $\mathsf{C}(A)$ is a natural starting place to study resolutions of objects in A, and it comes with the same general package of tools available in every abelian category: we can talk about exact sequences of complexes, we have a canonical decomposition of morphisms of complexes, and so on. A sequence of complexes

$$\cdots \longrightarrow L^\bullet \longrightarrow M^\bullet \longrightarrow N^\bullet \longrightarrow \cdots$$

is exact in $\mathsf{C}(\mathsf{A})$ if and only if all sequences

$$\cdots \longrightarrow L^i \longrightarrow M^i \longrightarrow N^i \longrightarrow \cdots$$

are simultaneously exact in A. *Complexes of complexes* will be important later on (§8.2).

Popular variations on the definition of $\mathsf{C}(\mathsf{A})$ require the complexes to be bounded above or below:

- $\mathsf{C}^+(\mathsf{A})$ denote the full[16] subcategory of $\mathsf{C}(\mathsf{A})$ determined by complexes $L^\bullet$ which are bounded *below,* i.e., for which $L^i = 0$ for $i \ll 0$.
- $\mathsf{C}^-(\mathsf{A})$ likewise denotes the full subcategory of bounded-*above* complexes.

These are also abelian categories, and so are further variations such as $\mathsf{C}^{\geq 0}(\mathsf{A})$ (complexes $L^\bullet$ for which $L^i = 0$ for $i < 0$), $\mathsf{C}^{\leq 0}(\mathsf{A})$, and so on. These bounded variants become unavoidable when dealing with resolutions.

It will also be necessary to consider categories of complexes in which the objects are restricted to belonging to a subcategory of A. We will devote particular attention to the full subcategories P and I consisting of *projective,* resp., *injective,* objects in A (see §5.3); thus, $\mathsf{C}^{\leq 0}(\mathsf{P})$, $\mathsf{C}^{\geq 0}(\mathsf{I})$ and other variations on the same themes are fair game[17]. In the generalities that follow we will only deal with $\mathsf{C}(\mathsf{A})$, to keep notation under control. The reader should have no difficulty determining what extends immediately to the different variants of $\mathsf{C}(\mathsf{A})$ and what does not.

There are several functors and other operations available in this new setting; many of them seem almost too simple-minded to mention. For example, for all integers r we have a fully faithful, exact functor

$$\iota_r : \mathsf{A} \longrightarrow \mathsf{C}(\mathsf{A})$$

sending an object A of A to the complex

$$\cdots \longrightarrow 0 \longrightarrow 0 \longrightarrow \underset{\substack{\uparrow \\ \text{degree } r}}{A} \longrightarrow 0 \longrightarrow 0 \longrightarrow \cdots$$

[16]That is, morphisms between two objects in the subcategory agree with morphisms of the same two objects in the ambient category. Cf. Exercise I.3.8.

[17]But note that P and I are not abelian categories!

placing A in degree r and 0 in all other degrees. It is common to identify A with its 'image' in $\mathsf{C}(\mathsf{A})$ via $\iota = \iota_0$. We also have 'shift' functors

$$\mathsf{C}(\mathsf{A}) \longrightarrow \mathsf{C}(\mathsf{A}),$$

$$M^\bullet \longmapsto M[r]^\bullet$$

defined by setting[18] $M[r]^i = M^{i+r}$, $d^i_{M[r]^\bullet} = (-1)^r d^{i+r}_{M^\bullet}$. Complexes can be 'truncated' by replacing all terms of degree $\geq r$ (or $\leq r$) with 0.

These operations turn out to be more important than they may look at first. The importance of the following example is, on the contrary, apparent. We claim that *cohomology is a functor:*

Lemma 3.4. *For every integer i, the assignment*

$$H^i : M^\bullet \longmapsto H^i(M^\bullet)$$

defines an additive covariant functor $\mathsf{C}(\mathsf{A}) \to \mathsf{A}$.

Proof. Of course, the statement means that each H^i induces in a natural (and functorial) way homomorphisms of abelian groups

$$\mathrm{Hom}_{\mathsf{C}(\mathsf{A})}(M^\bullet, N^\bullet) \to \mathrm{Hom}_A(H^i(M^\bullet), H^i(N^\bullet))$$

for all complexes $M^\bullet$, $N^\bullet$. This follows from the commutativity requirement in the definition of morphisms of complexes. Look at the relevant part of the diagram:

$$\begin{array}{ccccc} M^{i-1} & \xrightarrow{d^{i-1}} & M^i & \xrightarrow{d^i} & M^{i+1} \\ \downarrow{\scriptstyle \alpha^{i-1}} & & \downarrow{\scriptstyle \alpha^i} & & \downarrow{\scriptstyle \alpha^{i+1}} \\ N^{i-1} & \xrightarrow{d'^{i-1}} & N^i & \xrightarrow{d'^i} & N^{i+1} \end{array}$$

The composition $d'^i \circ \alpha^i = \alpha^{i+1} \circ d^i$ is 0 on $\ker d^i$ by the commutativity of the square on the right, so the restriction of α^i to $\ker d^i$ factors through $\ker d'^i$. Composing with the projection gives a morphism

$$\ker d^i \to H^i(N^\bullet).$$

This morphism is 0 when restricted to $\operatorname{im} d^{i-1}$, by the commutativity of the square on the left. Therefore α^i induces a morphism

$$H^i(M^\bullet) \to H^i(N^\bullet)$$

in A, as needed. It is clear that this assignment is covariant. □

We can also view cohomology as a functor $\mathsf{C}(A) \to \mathsf{C}(A)$, by placing each cohomology object $H^i(M^\bullet)$ in degree i, connected by zero-morphisms, obtaining a complex $H^\bullet(M^\bullet)$. Note that the cohomology of *this* complex equals the cohomology of $M^\bullet$, that is,

$$H^\bullet(H^\bullet(M^\bullet)) = H^\bullet(M^\bullet).$$

In this sense, taking cohomology feels like performing a 'projection'.

[18] The sign in the differentials is inserted in order to avoid other annoying signs elsewhere; signs do not change the isomorphism class of the complex (Exercise 3.5).

3.3. The long exact cohomology sequence. The one fact we did explore at any level of thoroughness in §III.7.1 is the famous *snake lemma,* Lemma III.7.8, which has come back to entertain us in §2.2. Written cohomologically and with the benefit of the language introduced in §3.2, the snake lemma takes the following form. Consider a commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & L^1 & \longrightarrow & M^1 & \longrightarrow & N^1 & \longrightarrow & 0 \\ & & \uparrow{\scriptstyle \lambda^0} & & \uparrow{\scriptstyle \mu^0} & & \uparrow{\scriptstyle \nu^0} & & \\ 0 & \longrightarrow & L^0 & \longrightarrow & M^0 & \longrightarrow & N^0 & \longrightarrow & 0 \end{array}$$

with exact rows, and view the columns as complexes

$$L^\bullet : \quad \cdots \longrightarrow 0 \longrightarrow L^0 \xrightarrow{\lambda^0} L^1 \longrightarrow 0 \longrightarrow \cdots ,$$

$M^\bullet$, $N^\bullet$. The commutative diagram is then nothing but the datum of a short exact sequences of (very special) complexes:

$$0 \longrightarrow L^\bullet \longrightarrow M^\bullet \longrightarrow N^\bullet \longrightarrow 0 .$$

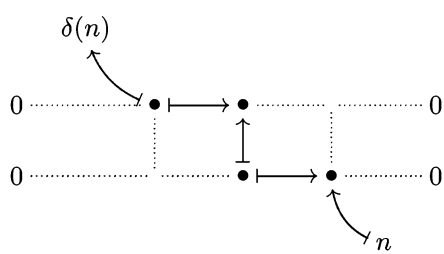
The snake lemma tells us that there is then an exact sequence in A:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & H^0(L^\bullet) & \longrightarrow & H^0(M^\bullet) & \longrightarrow & H^0(N^\bullet) & & \\ & & & & & \overset{\delta}{\searrow} & & & \\ & & H^1(L^\bullet) & \longrightarrow & H^1(M^\bullet) & \longrightarrow & H^1(N^\bullet) & \longrightarrow & 0 . \end{array}$$

Here, H^0, resp., H^1, simply stands for the kernel, resp., cokernel, of the corresponding map: nothing else is going on in these small complexes. The (element-theoretic) definition of the 'connecting' morphism δ is discussed thoroughly in §III.7.3 and can be summarized (*co*homologically speaking) as follows:

- Start with $n \in H^0(N^\bullet) = \ker \nu^0 \subseteq N^0$.
- Choose any preimage m of n in M^0.
- Map m to $\mu^0(m) \in M^1$.
- It is immediately seen that $\mu^0(m)$ maps to 0 in N^1; therefore it determines a unique element $\ell \in L^1$.
- Set $\delta(n) =$ the class of ℓ in $H^1(L^\bullet) = \operatorname{coker}(\lambda^0)$.

Pictorially,



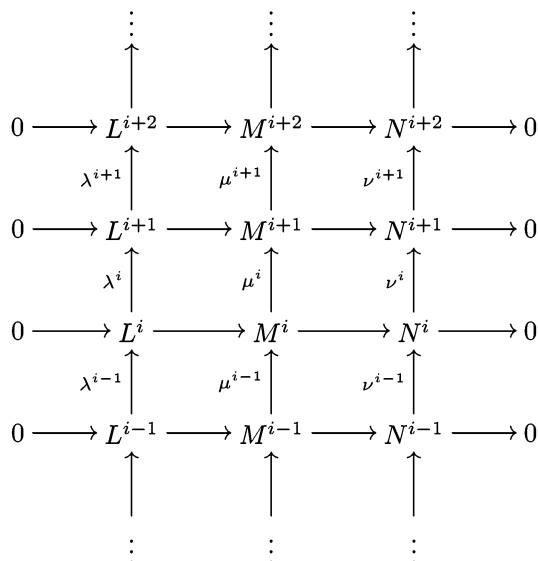
We checked in §III.7.3 that choosing a different preimage for n in M^0, while it may change the element ℓ, does not change its image in $\operatorname{coker}(\lambda_0)$. Thus δ is well-defined. As we have seen in §2.2, it is relatively straightforward to provide an arrow-theoretic

version of this construction; but we have also seen (Theorem 2.9) that we do not *need* to do that, so we will stay with elements in what follows.

Now we will upgrade this construction, in what should come across as a very natural development. In fact, what we are going to do could be absorbed into the snake lemma itself (cf. Exercise 3.10); but we have always found the 'spread out' chase compelling, so here it is. Take *any* short exact sequence in $\mathsf{C}(\mathsf{A})$,

$$0 \longrightarrow L^\bullet \longrightarrow M^\bullet \longrightarrow N^\bullet \longrightarrow 0\ ,$$

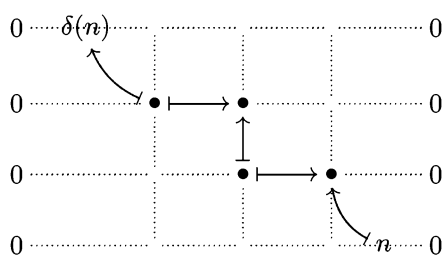
linking three full-blown cochain complexes $L^\bullet$, $M^\bullet$, $N^\bullet$; that is, consider a large commutative diagram



in A, where the rows are exact and the columns are complexes. The connecting morphism constructed for the snake lemma still gives an interesting morphism

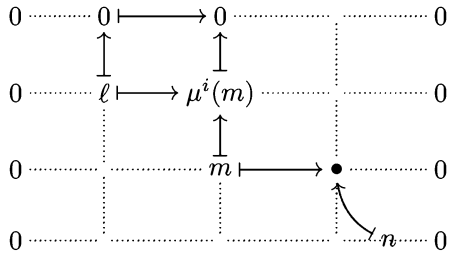
$$\delta : \ker \nu^i \to \operatorname{coker} \lambda^i \cong \frac{L^{i+1}}{\operatorname{im} \lambda^i}\ ,$$

obtained by 'climbing the ladder' in precisely the same way as for the snake lemma:



Where does the image $\delta(n)$ really lie? It is represented by an element ℓ in L^{i+1}, determined by an element $\mu^i(m) \in \operatorname{im} \mu^i$. The image of $\mu^i(m)$ in M^{i+2} must be 0,

because the central column is a complex:



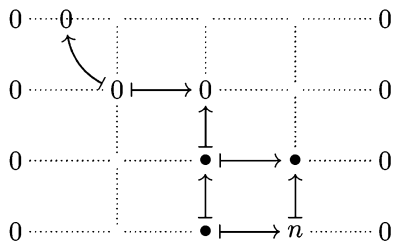
It follows that $\ell \in \ker \lambda^{i+1}$, and therefore we can place $\delta(n)$ in cohomology,

$$\delta(n) \in \frac{\ker \lambda^{i+1}}{\operatorname{im} \lambda_i} = H^{i+1}(L^\bullet),$$

and view δ as a morphism

$$\delta : \ker \nu^i \to H^{i+1}(L^\bullet).$$

Next, the kernel of ν^i contains the image of ν^{i-1}, again because columns are complexes. What is the restriction of δ to $\operatorname{im} \nu^{i-1}$? We claim it is 0. Indeed, if n comes from N^{i-1}, then a preimage in M^i of $\nu^{i-1}(n)$ may be obtained by taking a preimage of n in M^{i-1} and mapping it to M^i; but since the central column is a complex, this preimage maps to 0 in M^{i+1}. It follows that $\delta(n) = 0$ in this case:



Therefore, δ factors through $\operatorname{im} \nu^{i-1}$. Since $\ker \nu^i / \operatorname{im} \nu^{i-1}$ is nothing but the cohomology $H^i(N^\bullet)$, we have obtained a morphism

$$\delta^i : H^i(N^\bullet) \to H^{i+1}(L^\bullet).$$

On the other hand, each H^i is a functor (Lemma 3.4); thus there are natural morphisms

$$H^i(L^\bullet) \longrightarrow H^i(M^\bullet) \longrightarrow H^i(N^\bullet)$$

for each i. Note that completing this diagram with 0 on the left and on the right does *not* give in general a short exact sequence in A (Exercise 3.8). In other words, the i-th cohomology is not an exact functor even if, as we will see, the cohomology sequence *is* exact at $H^i(M^\bullet)$. Understanding the exactness and failure of exactness of the functors H^i is, in a sense, what this discussion is all about.

Together with the connecting morphism, we get a 'long sequence' of objects and morphisms in A,

$$\cdots \longrightarrow H^{i-1}(N^\bullet) \xrightarrow{\delta^{i-1}} H^i(L^\bullet) \longrightarrow H^i(M^\bullet) \longrightarrow H^i(N^\bullet) \xrightarrow{\delta^i} H^{i+1}(L^\bullet) \longrightarrow \cdots,$$

generalizing the sequence for H^0 and H^1 obtained in the snake lemma. The attentive reader must have seen the next result coming all along:

Theorem 3.5 (Long exact cohomology sequence)**.** *The sequence determined as above by a short exact sequence of complexes is an exact sequence.*

The snake lemma is recovered as a particular case of this statement (or conversely, depending on your taste; cf. Exercise 3.10).

Proof. The proof is a diagram chase, which everyone should perform once by oneself in his or her lifetime. So it is mostly left as an exercise for the reader (Exercise 3.9). But we will stress the extent to which H^i is exact 'on the nose': part of the claim in this theorem is that if

$$0 \longrightarrow L^\bullet \xrightarrow{\alpha^\bullet} M^\bullet \xrightarrow{\beta^\bullet} N^\bullet \longrightarrow 0$$

is exact, then the sequence induced by functoriality of H^i

$$0 \longrightarrow H^i(L^\bullet) \xrightarrow{\alpha} H^i(M^\bullet) \xrightarrow{\beta} H^i(N^\bullet) \longrightarrow 0$$

is exact *at* $H^i(M^\bullet)$. This sequence is not exact in general at $H^i(L^\bullet)$ and $H^i(N^\bullet)$; the moral of the long exact cohomology sequence is that the failure of exactness is precisely measured by other cohomology objects, by means of the connecting morphisms.

Let's check exactness at $H^i(M)$. The sequence is a complex, simply because H^i is a functor; thus we only have to verify that $\ker\beta \subseteq \operatorname{im}\alpha$. If $\overline{m}$ is a class in $H^i(M^\bullet)$ such that $\beta(\overline{m}) = 0$ in $H^i(N)$, $\overline{m}$ is represented by an element $m \in \ker\mu^i$ such that $\beta^i(m) \in \operatorname{im}\nu^{i-1}$: $\beta^i(m) = \nu^{i-1}(n)$ for some $n \in N^{i-1}$. There is an $m' \in M^{i-1}$ such that $\beta^{i-1}(m') = n$, and $m - \mu^{i-1}(m')$ represents the same class $\overline{m}$ in $H^i(M)$ as m. By the commutativity of the diagram,

$$\beta^i(m - \mu^{i-1}(m')) = \beta^i(m) - \nu^{i-1}(n) = 0.$$

Therefore there exists $\ell \in L^i$ such that $\alpha^i(\ell) = m - \mu^{i-1}(m')$; then ℓ represents a class $\overline{\ell} \in H^i(L^\bullet)$ such that $\alpha(\overline{\ell}) = \overline{m}$, concluding the verification.

Verifying the exactness at the other places is similarly uninspiring and is left to the reader[19]. □

3.4. Triangles. There is an æsthetically pleasing way to state what we just proved, which turns out to provide a useful viewpoint. Take a short exact sequence of complexes,

$$0 \longrightarrow L^\bullet \longrightarrow M^\bullet \longrightarrow N^\bullet \longrightarrow 0\ ;$$

the element of surprise in the long exact sequence obtained in §3.3 is the fact that we can connect $N^\bullet$ to $L^\bullet$ in an interesting way, after taking cohomology and allowing for a shift. Well, if we are willing to settle for something *less* interesting, we can connect $N^\bullet$ to $L^\bullet$ right away, by simply mapping $N^\bullet$ to 0 *within* $L^\bullet$:

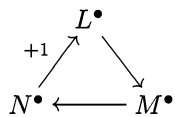
$$N^\bullet \xrightarrow{0} L^\bullet\ .$$

[19]The reader is welcome to look for arrow-theoretic proofs. However, as discussed at length in §2, element-based proofs suffice to prove the statement in any (small) abelian category.

The exactness of the original short exact sequence is then equivalent to the exactness at the center of the three even shorter sequences:

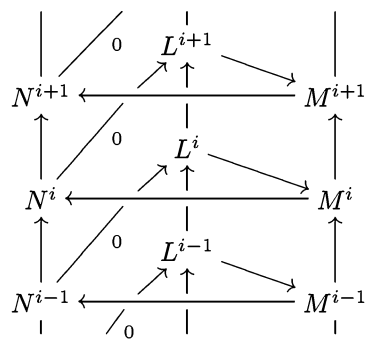
$$N^\bullet \xrightarrow{0} L^\bullet \longrightarrow M^\bullet\,,$$
$$L^\bullet \longrightarrow M^\bullet \longrightarrow N^\bullet\,,$$
$$M^\bullet \longrightarrow N^\bullet \xrightarrow{0} L^\bullet\,.$$

A nice pictorial way to represent this situation is to fold the sequence into a triangle[20], i.e.,



,

which we understand to be exact at its three vertices: this is an *exact triangle* in $\mathsf{C}(\mathsf{A})$. The arrow marked by $+1$ is in this case simply the zero-morphism; the $+1$ records the fact that we are going to view this as a morphism from $N^\bullet$ to the shift $L[1]^\bullet$ of $L^\bullet$ (after all, 0 is 0 in all degrees). Thus, the triangle is shorthand for the impressive commutative 3D-diagram



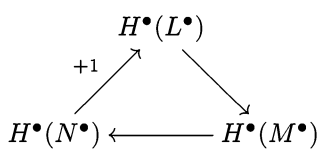
or (understanding the morphisms in the complexes) the single long sequence

$$\cdots \longrightarrow N^{i-1} \xrightarrow{0} L^i \longrightarrow M^i \longrightarrow N^i \xrightarrow{0} L^{i+1} \longrightarrow \cdots\,,$$

whose exactness is equivalent to the exactness of the original short sequence of complexes. In §3.3 we have obtained from this *another* long exact sequence,

$$\cdots \longrightarrow H^{i-1}(N^\bullet) \xrightarrow{\delta^{i-1}} H^i(L^\bullet) \longrightarrow H^i(M^\bullet) \longrightarrow H^i(N^\bullet) \xrightarrow{\delta^i} H^{i+1}(L^\bullet) \longrightarrow \cdots\,,$$

which we fold again into an exact triangle:



[20] A common and easier to typeset notation is $L^\bullet \longrightarrow M^\bullet \longrightarrow N^\bullet \xrightarrow{+1}$

This time the morphism marked +1 is the interesting connecting morphism (while the morphisms in the cohomology complexes are zero; see the end of §3.2).

Thus, the 'long exact cohomology sequence' takes us from a certain exact triangle to another exact triangle. The first triangle arises from a short exact sequence of complexes (the 'connecting morphism' is 0), while the second does not (the connecting morphism is in general nonzero).

Clearly something interesting is going on. Triangles obtained from short exact sequences of complexes in $\mathsf{C}(\mathsf{A})$ appear to be 'special' and give rise to other exact triangles through cohomology.

Actually, the situation may make the reader somewhat uncomfortable. While two sides of the cohomology triangle are obtained by simply applying the cohomology functor to the corresponding sides of the original triangle, the third one is obtained by a different process, involving the connecting morphism. The reader may feel that there should be a mechanism to go from a triangle defined in terms of an exact sequence to its counterpart in cohomology by simply applying the cohomology functor.

This is indeed so. There is an important notion of 'distinguished' triangle, in a category $\mathsf{K}(\mathsf{A})$ closely associated with $\mathsf{C}(\mathsf{A})$ and that we will encounter later[21] (§5). As with the 'special' triangles given above, distinguished triangles lead to long exact sequences in cohomology, and in a somewhat more direct fashion. Distinguished triangles satisfy a number of axioms, which define $\mathsf{K}(\mathsf{A})$ as a *triangulated category*. Exploring these concepts at any level of depth is well beyond the scope of this book, but towards the end of the chapter we will try to clarify these last cryptic remarks (§9.2).

Exercises

3.1. ▷ Let $(M^\bullet, d^\bullet)$ be a resolution of an object A of an abelian category. Verify that there are exact complexes

$$\cdots \xrightarrow{d^{-3}} M^{-2} \xrightarrow{d^{-2}} M^{-1} \longrightarrow \ker d^0 \longrightarrow A \longrightarrow 0 \longrightarrow \cdots,$$

$$\cdots \longrightarrow 0 \longrightarrow A \longrightarrow \operatorname{coker} d^0 \longrightarrow M^1 \xrightarrow{d^1} M^2 \xrightarrow{d^2} \cdots$$

and that replacing A by 0 in these complexes produces new resolutions of A. [§3.1, §6.1]

3.2. ¬ Let I be the category whose objects are the integers and where $\mathrm{Hom}_{\mathsf{I}}(m,n)$ consists of a singleton if $m \leq n$ and is $\emptyset$ otherwise (cf. Example I.3.2). Show that the category $\mathsf{C}(\mathsf{A})$ of cochain complexes is a full subcategory of the abelian category A^{I}. [3.3]

[21] Beware that some authors call the distinguished triangles 'exact'; there is some room for confusion, since the special exact triangles in the present discussion are not 'exact' from this viewpoint.

3.3. ▷ Verify that the category $\mathsf{C}(\mathsf{A})$ of cochain complexes in an abelian category, defined in §3.2, is itself an abelian category. (You can do this either 'by hand' or by applying Exercises 3.2 and 2.14.) [§3.2]

3.4. ¬ Let A be an abelian category. Define a category $\mathsf{Seq}(\mathsf{A})$ whose objects are short exact sequences in A and where morphisms are commutative diagrams

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A_1 & \longrightarrow & B_1 & \longrightarrow & C_1 & \longrightarrow & 0 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 0 & \longrightarrow & A_2 & \longrightarrow & B_2 & \longrightarrow & C_2 & \longrightarrow & 0 \end{array}$$

Thus, $\mathsf{Seq}(\mathsf{A})$ may be viewed as a full subcategory of $\mathsf{C}(\mathsf{A})$.

- Prove that $\mathsf{Seq}(\mathsf{A})$ is an additive category.
- Prove that $\mathsf{Seq}(\mathsf{A})$ has kernels and cokernels. (Watch out: these are *not* the same as in $\mathsf{C}(\mathsf{A})$. Keep the snake lemma in mind.)
- Show that, for every object X of A,

$$\begin{array}{ccccccccc} 0 & \longrightarrow & 0 & \longrightarrow & X & \xrightarrow{\mathrm{id}_X} & X & \longrightarrow & 0 \\ & & \downarrow & & \downarrow{\scriptstyle \mathrm{id}_X} & & \downarrow & & \\ 0 & \longrightarrow & X & \xrightarrow{\mathrm{id}_X} & X & \longrightarrow & 0 & \longrightarrow & 0 \end{array}$$

  is both a monomorphism and an epimorphism in $\mathsf{Seq}(\mathsf{A})$, although it is neither a monomorphism nor an epimorphism in $\mathsf{C}(\mathsf{A})$ if $X \neq 0$.
- Prove that $\mathsf{Seq}(\mathsf{A})$ is not abelian if A has nonzero objects.

[§1.3, 3.11]

3.5. ▷ Let $(L^\bullet, d^\bullet_{L^\bullet})$ be a cochain complex; define new differentials $d'^\bullet_{L^\bullet}$ by arbitrarily changing the sign of $d^i_{L^\bullet}$: $d'^i_{L^\bullet} = \pm d^i_{L^\bullet}$. Show that $(L^\bullet, d'^\bullet_{L^\bullet})$ is a cochain complex isomorphic to $(L^\bullet, d^\bullet_{L^\bullet})$. [§3.2]

3.6. Provide an arrow-theoretic proof of Lemma 3.4 (cf. Remark 3.1).

3.7. Let A, B be abelian categories. An additive functor $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ is *exact* if it maps short exact sequences in A to short exact sequences in B. Prove that exact functors commute with cohomology: if $\mathscr{F}$ is exact and $L^\bullet$ is a cochain complex in A, then $H^\bullet(\mathscr{F}(L^\bullet)) \cong \mathscr{F}(H^\bullet(L^\bullet))$, where $\mathscr{F}(L^\bullet)$ denotes the cochain complex in B obtained by applying $\mathscr{F}$ to all objects and morphisms in $L^\bullet$.

In particular, the image of an exact complex through an exact functor is still an exact complex (cf. Exercise VIII.1.23).

3.8. ▷ For any i, give an example of a short exact sequence of complexes

$$0 \longrightarrow L^\bullet \longrightarrow M^\bullet \longrightarrow N^\bullet \longrightarrow 0$$

such that the sequence

$$0 \longrightarrow H^i(L^\bullet) \longrightarrow H^i(M^\bullet) \longrightarrow H^i(N^\bullet) \longrightarrow 0$$

is not exact at either $H^i(L^\bullet)$ or $H^i(N^\bullet)$. [§7.2]

3.9. ▷ Provide the gory details for the long exact cohomology sequence (Theorem 3.5). [§3.3, 3.10]

3.10. ▷ Here is a way to reduce Exercise 3.9 to the snake lemma itself.

Every complex $(L^\bullet, \lambda^\bullet)$ determines a complex $\ker^\bullet$, with the object[22] $\ker \lambda^{i+1}$ in degree i, connected by zero-morphisms, and a complex $\operatorname{coker}^\bullet$, with the object $\operatorname{coker} \lambda^{i-1}$ in degree i, also connected by zero-morphisms.

Prove that $\lambda^\bullet$ induces a morphism $\overline{\lambda}^\bullet : \operatorname{coker}^\bullet \to \ker^\bullet$ in the abelian category of complexes, such that $\ker \overline{\lambda}^\bullet \cong H^\bullet(L^\bullet)$ and $\operatorname{coker} \overline{\lambda}^\bullet \cong H(L^\bullet)[1]^\bullet$.

Now work out the whole long exact cohomology sequence as a consequence of the snake lemma. (You will use the version of the snake lemma given in Remark III.7.11.) [§3.3, 3.13]

3.11. ¬ Prove that the long exact cohomology sequence is functorial, in the sense that it defines a covariant functor from the category $\mathsf{Seq}(\mathsf{C}(\mathsf{A}))$ of short exact sequences (cf. Exercise 3.4) of complexes to the category of complexes. [7.13]

3.12. Redo Exercise III.7.17.

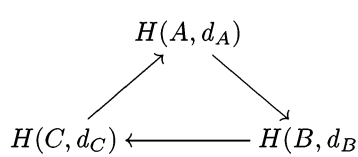
3.13. The viewpoint in Exercise 3.10 admits the following straightforward generalization.

Let A be an abelian category. Define a new category dA, whose objects are pairs (A, d), where A is an object of A and $d : A \to A$ is a morphism (the 'differential') such that $d^2 = 0$. Morphisms $(A, d_A) \to (B, d_B)$ in dA are morphisms $\varphi : A \to B$ commuting with the differentials: $d_B \varphi = \varphi d_A$.

- Prove that dA is an abelian category.
- For any (A, d) in dA, prove that d induces a morphism $\overline{d} : \operatorname{coker} d \to \ker d$.
- Define $H(A, d)$ to be $\ker \overline{d}$. Prove that $\operatorname{coker} \overline{d} \cong H(A, d)$.
- Show that H defines an additive functor $\mathsf{dA} \to \mathsf{A}$ and that this functor is not exact in general.
- However, prove that every short exact sequence

$$0 \longrightarrow (A, d_A) \longrightarrow (B, d_B) \longrightarrow (C, d_C) \longrightarrow 0$$

in dA induces an exact triangle (in the sense of §3.4)



3.14. ▷ Prove that if one of the vertices of a 'special' triangle arising from a short exact sequence of complexes (as in §3.4) is exact, then the other two vertices have isomorphic cohomologies, possibly up to a shift. [§4.1]

[22] Of course ker and coker mean the target and source of the arrows ker and coker, respectively.

3.15. ¬ Define a 'Grothendieck group' and a 'universal Euler characteristic' χ for any abelian category A, in the style of the construction given in §VI.3.4.

Extend χ to all bounded cochain complexes $M^\bullet$ in A, by setting $\chi(M^\bullet) := \sum_i (-1)^i \chi(M^i)$. Prove that $\chi(M^\bullet) = \chi(H^\bullet(M^\bullet))$. For every short exact sequence of bounded complexes $0 \to L^\bullet \to M^\bullet \to N^\bullet \to 0$, prove that $\chi(M^\bullet) = \chi(L^\bullet) + \chi(N^\bullet)$. [4.3, 9.1]

---

4. Cones and homotopies

If $\alpha^\bullet : L^\bullet \to M^\bullet$ is a morphism of cochain complexes, it is natural to try to describe the morphism induced in cohomology:

$$H^\bullet(\alpha^\bullet) : H^\bullet(L^\bullet) \to H^\bullet(M^\bullet).$$

However, we have pointed out that $H^\bullet$ is not an exact functor, and this seems to pose an obstacle to simple-minded strategies aimed at describing (for example) the kernel or cokernel of $H^\bullet(\alpha^\bullet)$ in terms of the kernel or cokernel of $\alpha^\bullet$. The long exact cohomology sequence may be used to compensate for this lack of exactness; we will see how in §4.1. We will also begin exploring a very important condition ('homotopy') guaranteeing that two given morphisms of complexes induce the same morphism in cohomology. Homotopies will be so important that they will lead us later on to change our notion of morphisms between complexes and construct a new 'homotopic category' of complexes (see §5).

4.1. The mapping cone of a morphism. Let $\alpha^\bullet : L^\bullet \to M^\bullet$ be a morphism of cochain complexes. The *mapping cone* $MC(\alpha)^\bullet$ of $\alpha^\bullet$ will allow us to recover $H^\bullet(\alpha^\bullet)$ as the connecting morphism in a long exact cohomology sequence, as constructed in §3.3. In fact, the mapping cone will give us the third vertex of a 'special' triangle (cf. the end of §3.4):

(†)
$$\begin{array}{ccc} & M^\bullet & \\ {\scriptstyle +1}\nearrow & & \searrow \\ L[1]^\bullet & \longleftarrow & MC(\alpha)^\bullet \end{array}$$

The degree-increasing morphism is the zero-morphism here (as in the 'special' triangles encountered in §3.4); but taking cohomology

$$\begin{array}{ccc} & H^\bullet(M^\bullet) & \\ {\scriptstyle +1}\nearrow & & \searrow \\ H^\bullet(L[1]^\bullet) & \longleftarrow & H^\bullet(MC(\alpha)^\bullet) \end{array}$$

we will obtain morphisms

$$H^{i+1}(L^\bullet) = H^i(L[1]^\bullet) \longrightarrow H^{i+1}(M^\bullet)$$

that, we claim, will be nothing but the morphisms induced in cohomology by the given morphism $\alpha^\bullet$. Thus, the morphisms induced in cohomology cannot be extended to exact sequences (that would clearly be asking too much), *but* they can

be extended to triangles arising by applying cohomology to an appropriate exact sequence of complexes.

Incidentally, the triangle (†) displayed above (but with the degree-increasing morphism replaced by $-\alpha^\bullet$) will be the prototype of the 'distinguished triangles' mentioned in §3.4 (cf. §9.2).

Construction of the mapping cone. The objects of $MC(\alpha)^\bullet$ are simply direct sums of objects of $L^\bullet$ and $M^\bullet$:

$$MC(\alpha)^i := L[1]^i \oplus M^i = L^{i+1} \oplus M^i;$$

but the morphisms $d^i_{MC(\alpha)^\bullet} : MC(\alpha)^i \to MC(\alpha)^{i+1}$ are not the 'obvious' ones, but rather

$$d^i_{MC(\alpha)^\bullet}(\ell, m) := (-d^{i+1}_{L^\bullet}(\ell), \alpha^{i+1}(\ell) + d^i_{M^\bullet}(m)).$$

The sign in the first component is inherited from the shifting of $L^\bullet$; thus, the first component is simply the differential of $L[1]^\bullet$. The second component mixes $\alpha^\bullet$ (or rather its shift $\alpha[1]^\bullet$) and the differential of $M^\bullet$. The reader should verify that $d^{i+1}_{MC(\alpha)^\bullet} \circ d^i_{MC(\alpha)^\bullet} = 0$, so that $MC(\alpha)^\bullet$ is indeed a cochain complex (Exercise 4.1); this is where the sign comes in. The mapping cone is a special case of the 'total complex' determined by a double complex, to be explored later (§8.2).

Pictorially, $MC(\alpha)^\bullet$ looks like this:

$$\begin{array}{ccccccccc}
\cdots & \longrightarrow & L^{i+1} & \xrightarrow{-d^{i+1}_{L^\bullet}} & L^{i+2} & \longrightarrow & \cdots \\
 & & \oplus & \searrow^{\alpha^{i+1}} & \oplus & & \\
\cdots & \longrightarrow & M^i & \xrightarrow[d^i_{M^\bullet}]{} & M^{i+1} & \longrightarrow & \cdots
\end{array}$$

There are evident morphisms of complexes $M^\bullet \to MC(\alpha)^\bullet$ and $MC(\alpha)^\bullet \to L[1]^\bullet$, induced by the natural morphisms $M^i \longrightarrow L^{i+1} \oplus M^i \longrightarrow L^{i+1}$. Since the sequences

$$0 \longrightarrow M^i \longrightarrow L[1]^i \oplus M^i \longrightarrow L[1]^i \longrightarrow 0$$

are all exact, the sequence of complexes

$$0 \longrightarrow M^\bullet \longrightarrow MC(\alpha)^\bullet \longrightarrow L[1]^\bullet \longrightarrow 0$$

is exact.

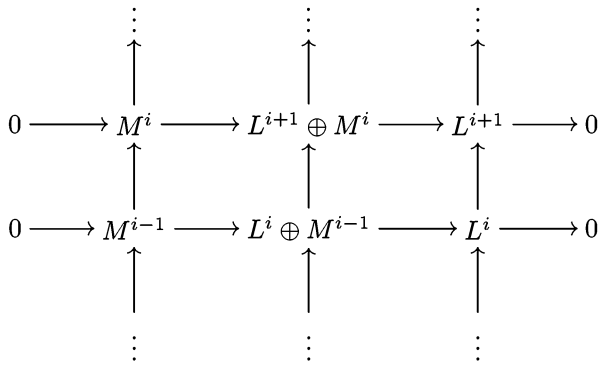
Proposition 4.1. *There is an exact triangle*

$$\begin{array}{ccc}
 & H^\bullet(M^\bullet) & \\
{\scriptstyle +1}\nearrow & & \searrow \\
{\scriptstyle \delta} & & \\
H^\bullet(L[1]^\bullet) & \longleftarrow & H^\bullet(MC(\alpha)^\bullet)
\end{array}$$

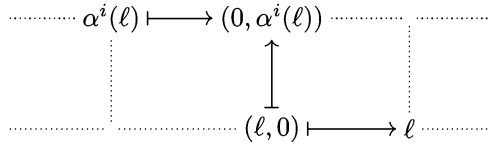
where the connecting morphism δ is the morphism induced by $\alpha^\bullet$ in cohomology.

Proof. The existence of the triangle is a direct consequence of Theorem 3.5; all we have to check is that the connecting morphism indeed agrees with the morphism

induced by $\alpha^\bullet$. Chasing the diagram



with a class in $H^{i-1}(L[1]^\bullet) = H^i(L^\bullet)$ represented by $\ell \in L^i$ (such that $d^i_{L^\bullet}(\ell) = 0$), we find



which verifies the claim. □

Corollary 4.2. *Let $\alpha^\bullet : L^\bullet \to M^\bullet$ be a morphism of cochain complexes. Then the induced morphism $H^\bullet(L^\bullet) \to H^\bullet(M^\bullet)$ is an isomorphism if and only if the mapping cone $MC(\alpha)^\bullet$ is an exact complex.*

(Cf. Exercise 3.14.)

We should mention that mapping cones arose in topology: the mapping cone of a continuous map $f : X \to Y$ is obtained by considering $X \times [0,1]$, identifying $X \times 0$ to a point, and stitching $X \times 1$ to Y by means of f. Analyzing chains of this space leads to the (homological version of the) 'algebraic' mapping cone considered above.

4.2. Quasi-isomorphisms and derived categories. A component of our main strategy consists of understanding when cochain complexes 'have a good reason' to have the same cohomology. Corollary 4.2 provides us with such a reason. The morphisms singled out in that statement have a name:

Definition 4.3. A morphism $\alpha^\bullet$ of cochain complexes is a *quasi-isomorphism* if it induces an isomorphism in cohomology. ⌟

By Corollary 4.2, this condition is equivalent to the exactness of the mapping cone of $\alpha^\bullet$.

Example 4.4. The datum of a *resolution* $M^\bullet$ of an object A of an abelian category A, as in Definition 3.2 and with $M^i = 0$ for $i > 0$, is the same as the datum of a quasi-isomorphism

$$M^\bullet \xrightarrow{\text{q-iso.}} \iota(A)$$

where ι places A in degree 0 and 0 in all other degrees. Here is a larger version of the same diagram:

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & M^{-2} & \xrightarrow{d_{M^\bullet}^{-2}} & M^{-1} & \xrightarrow{d_{M^\bullet}^{-1}} & M^0 & \xrightarrow{d_{M^\bullet}^{0}} & 0 & \longrightarrow & \cdots \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\
\cdots & \longrightarrow & 0 & \longrightarrow & 0 & \longrightarrow & A & \longrightarrow & 0 & \longrightarrow & \cdots
\end{array}$$

where the only nontrivial vertical map is the cokernel of $d_{M^\bullet}^{-1}$. ⌟

Thus, quasi-isomorphisms may be viewed as generalizations of more simple-minded resolutions. Also note that the mapping cone of a resolution as in Example 4.4 is obtained (as the reader should check) by shifting the complex 'one step to the left' and completing it with A, obtaining the *exact* complex:

$$\cdots \longrightarrow M^{-1} \xrightarrow{-d_{M^\bullet}^{-1}} M^0 \xrightarrow{\operatorname{coker} d_{M^\bullet}^{-1}} A \longrightarrow 0 \longrightarrow \cdots .$$

This is nothing but our other point of view on resolutions (as in §VI.4.2).

Remark 4.5. Quasi-isomorphisms with fixed target (resp., source) form a category in a natural way, which we will leave to the reader to make precise. In particular, resolutions in $\mathsf{C}^{\leq 0}(\mathsf{A})$ (resp., $\mathsf{C}^{\geq 0}(\mathsf{A})$) of a given object of A, as in Example 4.4, form a category. ⌟

Wouldn't it be nice if quasi-isomorphisms were *actual* isomorphisms, that is, if they were invertible? In general, they are not. The following example will arise several times to ward off any such hopes.

Example 4.6. In $\mathsf{C}(\mathsf{Ab})$, the morphism of complexes

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & 0 & \longrightarrow & \mathbb{Z} & \xrightarrow{\cdot 2} & \mathbb{Z} & \longrightarrow & 0 & \longrightarrow & \cdots \\
 & & \downarrow & & \downarrow & & \downarrow{\scriptstyle \pi} & & \downarrow & & \\
\cdots & \longrightarrow & 0 & \longrightarrow & 0 & \longrightarrow & \frac{\mathbb{Z}}{2\mathbb{Z}} & \longrightarrow & 0 & \longrightarrow & \cdots
\end{array}$$

where π is the natural projection, is a quasi-isomorphism, but it does not have an inverse since there are no nontrivial homomorphisms $\mathbb{Z}/2\mathbb{Z} \xrightarrow{?} \mathbb{Z}$. ⌟

In fact, this example shows that 'quasi-isomorphism' is not even an equivalence relation (it is not symmetric). This is too bad—if it were, then the 'equivalence class of a complex' would be the most natural candidate for the entity carrying the maximal amount of cohomological information of a complex.

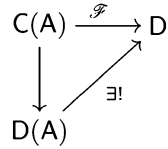
We will stop short of defining a 'quasi-isomorphism relation': in this book, quasi-isomorphism is simply a particular quality of a morphism of cochain complexes, not a relation. In any case recall that, even in the context of sets, taking the quotient by an equivalence relation is not the 'primary' object of interest: the quotient is just the solution to the natural universal problem of studying functions to other sets with identical behavior on 'equivalent' elements. *This* is the primary objective.

Similarly, the primary objective in the present context is not so much to identify together all complexes linked by a quasi-isomorphism, as it is to be able to shuttle information back and forth between such complexes. The natural way to accomplish this is to study environments in which quasi-isomorphisms do have inverses. More precisely, it consists of studying additive functors

$$\mathscr{F} : \mathsf{C}(\mathsf{A}) \longrightarrow \mathsf{D}$$

such that $\mathscr{F}(\rho^\bullet)$ *is an isomorphism for every quasi-isomorphism* $\rho^\bullet$. (Cohomology is such a functor.)

This sounds like yet another universal problem, and the natural line of attack would be to solve it as such, that is, to define (if possible) a category $\mathsf{D}(\mathsf{A})$ endowed with an additive functor $\mathsf{C}(\mathsf{A}) \to \mathsf{D}(\mathsf{A})$ such that quasi-isomorphisms in $\mathsf{C}(\mathsf{A})$ are mapped to isomorphisms in $\mathsf{D}(\mathsf{A})$ and through which every $\mathscr{F}$ as above must factor uniquely:



To be honest, we should take such a statement with a grain of salt: the natural environment in which we should pose this universal problem would be a 'category of categories', and we have not defined such a thing in this book. A category $\mathsf{D}(\mathsf{A})$ as above *does* exist, at least if a certain 'localization' process can be carried out[23]: it is the *derived category* of A. 'Bounded' versions $\mathsf{D}^-(\mathsf{A})$, $\mathsf{D}^+(\mathsf{A})$, and more, are solutions to the corresponding universal problems for appropriately bounded complexes.

These derived categories answer the main question underlying our strategy, of understanding 'what' in a complex really determines its cohomology: the answer amounts to viewing the complex in the appropriate derived category. The construction of $\mathsf{D}(\mathsf{A})$ consists roughly of taking the same objects as $\mathsf{C}(\mathsf{A})$ and formally inverting the quasi-isomorphisms. The details of this construction are somewhat involved and are best left to more advanced texts. But we will be able to capture its essence in this book, at least in lucky cases (when the abelian category has 'enough injectives' or 'enough projectives'; see Definition 5.7).

The derived category has the unfortunate reputation of being an overly abstract notion, and there are fundamental questions about whether it really is the best approach to an abstract study of cohomology. This is partly because the derived category of an abelian category is *not* an abelian category and simple notions such as kernel, cokernel, exact sequences are not available in $\mathsf{D}(\mathsf{A})$. This adds a substantial layer of complication to the theory.

On the other hand, going past these difficulties, one finds that enough structure remains to do much homological algebra: objects in $\mathsf{D}(\mathsf{A})$ have cohomology, and there are 'distinguished triangles' (cf. §3.4) abstracting exact sequences and giving

[23]There may be set-theoretic issues with this process. However, these play no role in the cases we will consider, in which A has 'enough projectives' or 'enough injectives'.

long exact cohomology sequences. The derived category is a *triangulated category,* like the more manageable homotopic category $\mathsf{K}(\mathsf{A})$ that we will soon define. Its uses go beyond the confines of algebra or even mathematics: an approach to the understanding of 'D-branes' in string theory is based on derived categories.

To get a sense of how counterintuitive $\mathsf{D}(\mathsf{A})$ must be, note that the zero-morphism $0 : M^\bullet \to M^\bullet$ may well be a quasi-isomorphism: in fact, this is so precisely when $M^\bullet$ is exact. Well, in any category D as above (and in particular in the derived category $\mathsf{D}(\mathsf{A})$) this zero-morphism remains the zero-morphism, *but* it comes equipped with an inverse when it is a quasi-isomorphism. Thus, in the derived category the zero-morphism on a complex involving nonzero objects may well be invertible! This says that complexes that are nonzero in $\mathsf{C}(\mathsf{A})$ may become zero-objects in the derived category.

Lemma 4.7. *Let* D *be an additive category, and let* $\mathscr{F} : \mathsf{C}(\mathsf{A}) \to \mathsf{D}$ *be an additive functor such that* $\mathscr{F}(\rho^\bullet)$ *is an isomorphism for every quasi-isomorphism* $\rho^\bullet$.

- *Let* $M^\bullet$ *be an exact complex in* $\mathsf{C}(\mathsf{A})$. *Then the complex* $\mathscr{F}(M^\bullet)$, *obtained by applying* $\mathscr{F}$ *to the objects and morphisms of* $M^\bullet$, *is a zero-object in* D.
- *Let* $\alpha^\bullet : L^\bullet \to N^\bullet$ *be a morphism in* $\mathsf{C}(\mathsf{A})$ *that factors through an exact complex,*

$$\begin{array}{ccc} L^\bullet & \xrightarrow{\alpha^\bullet} & N^\bullet \\ & \searrow \quad \nearrow & \\ & M^\bullet & \end{array} \quad ,$$

with $M^\bullet$ *exact. Then* $\mathscr{F}(\alpha^\bullet)$ *is the zero-morphism.*

Proof. The second claim follows from the first, since by applying $\mathscr{F}$ to the given diagram we obtain

$$\begin{array}{ccc} \mathscr{F}(L^\bullet) & \xrightarrow{\mathscr{F}(\alpha^\bullet)} & \mathscr{F}(N^\bullet) \\ & \searrow \quad \nearrow & \\ & \mathscr{F}(M^\bullet) & \end{array} \quad ,$$

and we see that $\mathscr{F}(\alpha^\bullet)$ factors through a zero-object of D.

To verify the first claim, note that since $M^\bullet$ is exact, the zero-morphism: $M^\bullet \to M^\bullet$ is a quasi-isomorphism; hence it is mapped to an invertible morphism by $\mathscr{F}$:

$$\mathscr{F}(M^\bullet) \xrightarrow{\mathscr{F}(0)} \mathscr{F}(M^\bullet) \xrightarrow{\mathscr{F}(0)^{-1}} \mathscr{F}(M^\bullet), \qquad \mathscr{F}(M^\bullet) \xrightarrow{\mathrm{id}_{\mathscr{F}(M^\bullet)}} \mathscr{F}(M^\bullet)$$

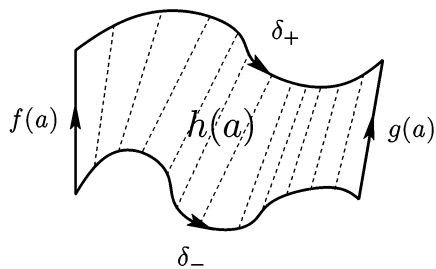
Since $\mathscr{F}$ is additive, $\mathscr{F}(0) = 0$. It follows that $\mathrm{id}_{\mathscr{F}(M^\bullet)} = 0$ and hence that $\mathscr{F}(M^\bullet)$ is a zero-object of D, by Exercise 1.6. □

In the next several sections we will chase the notion of derived category, aiming to understand it rather concretely in particularly favorable circumstances. *We will take it for granted that these categories exist.* A detailed definition of these objects, or a treatment of triangulated categories, is beyond the scope of this book.

4.3. Homotopy. The foregoing considerations put our strategy into focus: we are after constructions that determine cochain complexes 'up to quasi-isomorphism'. However, quasi-isomorphisms appear hard to deal with directly. Thus, we look for more manageable notions that may work as an effective replacement.

Like the mapping cone, this line of approach also owes its origins to topology: in topology, 'homeomorphism' is often too harsh a requirement, while 'homotopy equivalence' is a more malleable but still adequate notion. For example, homotopy equivalent topological spaces have the same homology. Distilling the algebra out of this notion leads to an analog at the level of complexes.

To see how this is done, recall that if f, g are *homotopic* continuous functions between two topological spaces X and Y, then f and g induce the same map on the homology of the spaces. To remind yourself of how this works, look at the picture:



This is supposed to represent the action of a homotopy between f and g on a chain a in X: $h(a)$ is obtained by mapping $a \times [0,1]$ to Y so as to get $f(a)$ when restricting to $a \times \{0\}$ and to get $g(a)$ when restricting to $a \times \{1\}$. Note that $h(a)$ is a chain of dimension 1 higher than the dimension of a, and that the boundary $\partial h(a)$ of this chain consists of $f(a)$, $g(a)$ and of the *restriction of h to the boundary ∂a of a:* taking the boundary 'counterclockwise',

$$\partial h(a) = g(a) - \delta_+ - f(a) + \delta_- = g(a) - f(a) - h(\partial a)$$

(or so it would seem from the picture! As presented here, this is of course at best a plausibility argument; it can be made rigorous, but that is someone else's business). That is,

$$g(a) - f(a) = \partial h(a) + h(\partial a).$$

Since boundaries vanish in homology, $f(a)$ and $g(a)$ will agree in homology.

Here is the translation into algebra of this pleasant geometric situation:

Definition 4.8. A *homotopy* h between two morphisms of cochain complexes

$$\alpha^\bullet, \beta^\bullet : L^\bullet \longrightarrow M^\bullet$$

is a collection of morphisms

$$h^i : L^i \longrightarrow M^{i-1}$$

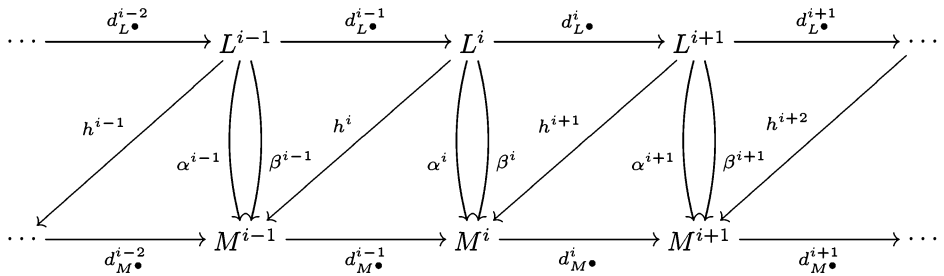
such that $\forall i$

$$\beta^i - \alpha^i = d_{M^\bullet}^{i-1} \circ h^i + h^{i+1} \circ d_{L^\bullet}^i.$$

We say that $\alpha^\bullet$ is *homotopic* to $\beta^\bullet$ and write $\alpha^\bullet \sim \beta^\bullet$ if there is a homotopy between $\alpha^\bullet$ and $\beta^\bullet$. ⌟

We are dealing here with *cochain* complexes; this accounts for the fact that while in the topological situation (where we were interested in *chains* rather than cochains) the homotopy would shift dimensions up, in Definition 4.8 it shifts degrees down.

Aside from its topological motivation, homotopy is not too easy to visualize. The following diagram is *not* assumed to be commutative:



The morphisms h^i do not usually define a morphism of complexes $L^\bullet \to M[-1]^\bullet$: the lozenges in this diagram are not required to commute.

Definition 4.9. A morphism $\alpha^\bullet : L^\bullet \to M^\bullet$ is a *homotopy equivalence* if there is a morphism $\beta^\bullet : M^\bullet \to L^\bullet$ such that $\alpha^\bullet \circ \beta^\bullet \sim 1_{M^\bullet}$ and $\beta^\bullet \circ \alpha^\bullet \sim 1_{L^\bullet}$. The complexes $L^\bullet$, $M^\bullet$ are said to be *homotopy equivalent* if there is a homotopy equivalence $L^\bullet \to M^\bullet$. ⌟

The reader should check that $\sim$ is an equivalence relation and that 'homotopy equivalence of complexes' is also (Exercise 4.4). Further, these relations are clearly compatible with simple operations of morphisms (Exercises 4.5 and 4.6).

Proposition 4.10. *If $\alpha^\bullet, \beta^\bullet : L^\bullet \to M^\bullet$ are homotopic morphisms of complexes, then $\alpha^\bullet$, $\beta^\bullet$ induce the same morphisms on cohomology: $H^\bullet(L^\bullet) \to H^\bullet(M^\bullet)$.*

Proof. Let $\overline{\ell} \in H^i(L^\bullet)$. Then $\overline{\ell}$ is represented by an element $\ell \in \ker(d^i_{L^\bullet})$, and its images in $H^i(M^\bullet)$ under the morphisms induced by $\alpha^\bullet$, $\beta^\bullet$ are represented by

$$\alpha^i(\ell), \quad \beta^i(\ell).$$

Since $\alpha^\bullet$, $\beta^\bullet$ are homotopic, according to Definition 4.8 there are morphisms h^i such that

$$\beta^i(\ell) - \alpha^i(\ell) = d^{i-1}_{M^\bullet}(h^i(\ell)) + h^{i+1}(d^i_{L^\bullet}(\ell)).$$

Since $\ell \in \ker d^i_{L^\bullet}$, the last term vanishes. This shows that

$$\beta^i(\ell) - \alpha^i(\ell) \in \operatorname{im} d^i_{M^\bullet},$$

proving that $\beta^i(\ell) - \alpha^i(\ell)$ vanishes in $H^i(M^\bullet)$, as needed. □

For example, if $\alpha^\bullet$ is homotopic to the identity, then by Corollary 4.2 the cone of $\alpha^\bullet$ must be exact. It is a good exercise to verify this fact directly (Exercise 4.8).

Corollary 4.11. *Homotopy equivalent complexes have isomorphic cohomology.*

Proof. Indeed, morphisms $\alpha^\bullet : L^\bullet \to M^\bullet$, $\beta^\bullet : M^\bullet \to L^\bullet$ such that $\beta^\bullet \circ \alpha^\bullet$ and $\alpha^\bullet \circ \beta^\bullet$ are both homotopic to the identity induce inverse morphisms in cohomology, by Proposition 4.10. □

Remark 4.12. In other words, homotopy equivalences of complexes are quasi-isomorphisms. This does not yet solve the 'problem' that quasi-isomorphisms are not invertible, since Example 4.6 shows that quasi-isomorphisms may well not be invertible even up to homotopy. In other words, 'homotopy equivalence' is a more restrictive notion than 'quasi-isomorphism'. One reason to study complexes of injective/projective modules, as we will do in §5.3, is precisely that quasi-isomorphism and homotopy equivalence *are* equivalent notions for (bounded) complexes of injective or projective modules. Since homotopy equivalence is a ready-made equivalence relation on complexes, this will bypass any technicality needed to make sense of 'quasi-isomorphism' as an equivalence relation. ⌟

Corollary 4.11 is the main technical reason that makes the strategy presented in §3.1 work, in a class of interesting examples. As we will see in due time, in these examples the cohomology of the relevant cochain complexes will end up being independent of the choices, because these choices will produce homotopic complexes.

Another key observation will be that homotopy equivalence is preserved by any additive functor. Suppose A, B are abelian categories and

$$\mathscr{F} : \mathsf{A} \longrightarrow \mathsf{B}$$

is an additive functor (Definition 1.1). Then $\mathscr{F}$ induces an additive functor preserving the grading

$$\mathsf{C}(\mathscr{F}) : \mathsf{C}(\mathsf{A}) \longrightarrow \mathsf{C}(\mathsf{B}).$$

Lemma 4.13. *With $\mathscr{F}$ as above, if $\alpha^\bullet \sim \beta^\bullet$ in $\mathsf{C}(\mathsf{A})$, then $\mathsf{C}(\mathscr{F})(\alpha^\bullet) \sim \mathsf{C}(\mathscr{F})(\beta^\bullet)$ in $\mathsf{C}(\mathsf{B})$, and if $L^\bullet$ and $M^\bullet$ are homotopy equivalent complexes in $\mathsf{C}(\mathsf{A})$, then $\mathsf{C}(\mathscr{F})(L^\bullet)$ and $\mathsf{C}(\mathscr{F})(M^\bullet)$ are homotopy equivalent complexes in $\mathsf{C}(\mathsf{B})$.*

Proof. The second assertion follows from the first. The first is an immediate consequence of the fact that $\mathscr{F}$ is additive. Indeed, if h is a homotopy between $\alpha^\bullet, \beta^\bullet : L^\bullet \to M^\bullet$, then

$$\beta^i - \alpha^i = d^{i-1}_{M^\bullet} \circ h^i + h^{i+1} \circ d^i_{L^\bullet}.$$

Since $\mathscr{F}$ preserves the additive structure on Hom-sets, this implies

$$\mathscr{F}(\beta^i) - \mathscr{F}(\alpha^i) = \mathscr{F}(d^{i-1}_{M^\bullet}) \circ \mathscr{F}(h^i) + \mathscr{F}(h^{i+1}) \circ \mathscr{F}(d^i_{L^\bullet}),$$

showing that the collection of morphisms $\mathscr{F}(h^i)$ gives a homotopy between $\mathscr{F}(\alpha^\bullet)$ and $\mathscr{F}(\beta^\bullet)$. □

Note that the analogous statement does *not* hold for quasi-isomorphisms (Exercise 4.15): an additive functor does not preserve quasi-isomorphisms (while an *exact* functor does).

Lemma 4.13 implies that the conclusion of Corollary 4.11 holds true after applying an additive functor:

Theorem 4.14. *Let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be an additive functor between two abelian categories. If $L^\bullet$, $M^\bullet$ are homotopy equivalent complexes in $\mathsf{C}(\mathsf{A})$, then the cohomology complexes*

$$H^\bullet(\mathscr{F}(L^\bullet)), \quad H^\bullet(\mathscr{F}(M^\bullet))$$

are isomorphic.

The proof of this statement is essentially immediate after all our preparatory work, so it is left to the reader (Exercise 4.16).

We are gracing this statement with theorem status because it is at the root of the strategy we are pursuing. The content of Theorem 4.14 is that any mechanism associating to a mathematical object a cochain complex *determined up to homotopy* will give rise to a slew of interesting invariants: apply your favorite additive functor to any such complex, take cohomology, and Theorem 4.14 guarantees that the result will be independent of the specific chosen complex. This is (part of) the reason underlying the various claims of independence on choices made in §VIII.2.4 and §VIII.6.4, concerning the definition of the Tor and Ext functors, as we will see later in the chapter (§7, especially Examples 7.6 and 7.7).

Exercises

4.1. ▷ Verify that the cone of a morphism of complexes is a complex. [§4.1]

4.2. Let $L^\bullet = \cdots \to 0 \to L \to 0 \to \cdots$ and $M^\bullet = \cdots \to 0 \to M \to 0 \to \cdots$ be two complexes concentrated in degree 0. Giving a morphism $\alpha^\bullet : L^\bullet \to M^\bullet$ is then the same as giving a morphism $\alpha : L \to M$. Describe the mapping cone in this case and its cohomology.

4.3. Let $\alpha^\bullet : L^\bullet \to M^\bullet$ be a morphism of bounded complexes. Note that the mapping cone $MC(\alpha)^\bullet$ of α is also bounded; thus, the three complexes $L^\bullet$, $M^\bullet$, $MC(\alpha)^\bullet$ have a well-defined universal Euler characteristic (see Exercise 3.15). Prove that $\chi(MC(\alpha)^\bullet) = \chi(M^\bullet) - \chi(L^\bullet)$.

4.4. ▷ Prove that the homotopy relation between morphisms of complexes and the relation of homotopy equivalence between complexes are equivalence relations. [§4.3, §6]

4.5. ▷ Let $\alpha_k^\bullet, \beta_k^\bullet : L_k^\bullet \to M_k^\bullet$, $k = 0, 1$, be morphisms of cochain complexes. Assume that $\alpha_0 \sim \beta_0$, $\alpha_1 \sim \beta_1$. Prove that $\alpha_0 \oplus \alpha_1 \sim \beta_0 \oplus \beta_1$ as morphisms $L_0^\bullet \oplus M_0^\bullet \to L_1^\bullet \oplus M_1^\bullet$. [§4.3]

4.6. ▷ Let $\alpha^\bullet, \alpha_0^\bullet, \alpha_1^\bullet : L^\bullet \to M^\bullet$ and $\beta^\bullet, \beta_0^\bullet, \beta_1^\bullet : M^\bullet \to N^\bullet$ be morphisms of cochain complexes.

(i) Assume that $\alpha^\bullet \sim 0$ or $\beta^\bullet \sim 0$. Prove that $\beta^\bullet \circ \alpha^\bullet \sim 0$.

(ii) Assume that $\alpha_0^\bullet \sim \alpha_1^\bullet$ and $\beta_0^\bullet \sim \beta_1^\bullet$. Prove that $\beta_0^\bullet \circ \alpha_0^\bullet \sim \beta_1^\bullet \circ \alpha_1^\bullet$. (Hint: Use (i), while thinking about ideals.)

[§4.3, §5.2, §5.4, §5.5]

4.7. Prove that the equivalence class of morphisms of complexes $L^\bullet \to M^\bullet$ homotopic to a given morphism is parametrized by the set of collections of morphisms $h^i : L^i \to M^{i-1}$, modulo the morphisms of complexes $L[1]^\bullet \to M^\bullet$.

4.8. ▷ Let $\alpha^\bullet$ be a morphism of complexes, homotopic to the identity. Prove directly (without appealing to Corollary 4.2) that the mapping cone of $\alpha^\bullet$ is exact. [§4.3]

4.9. ¬ Let $\alpha^\bullet : L^\bullet \to M^\bullet$ be a morphism of cochain complexes. The *mapping cylinder* $MCyl(\alpha)^\bullet$ is the cochain complex with objects $L^i \oplus L^{i+1} \oplus M^i$ in degree i and differential $d^\bullet_{MCyl(\alpha)^\bullet}$ defined by

$$d^i_{MCyl(\alpha)^\bullet}(\ell, \ell', m) = (d^i_{L^\bullet}(\ell) + \ell', -d^{i+1}_{L^\bullet}(\ell'), \alpha^{i+1}(\ell') + d^i_{M^\bullet}(m)).$$

Verify that $d^{i+1}_{MCyl(\alpha)^\bullet} \circ d^i_{MCyl(\alpha)^\bullet} = 0$ and hence that $MCyl(\alpha)^\bullet$ is indeed a cochain complex. [4.10, 4.11, 4.12, 4.13]

4.10. Topologically speaking, the mapping cylinder of a continuous map $f : X \to Y$ is obtained by considering $X \times [0, 1]$ and gluing $X \times 1$ to Y by means of f; studying (co)chains of this object leads to the algebraic mapping cylinder of Exercise 4.9.

As we learn in topology, two continuous maps $X \to Y$ are homotopic if they may be realized as the restrictions to $X \times 0$, $X \times 1$ of a continuous map from the cylinder $X \times [0, 1]$ to Y. Note that this cylinder is the mapping cylinder of id_X.

Prove that two cochain maps $\alpha^\bullet, \beta^\bullet : L^\bullet \to M^\bullet$ are homotopic in the sense of Definition 4.8 if and only if they extend to a cochain morphism $(-\alpha^\bullet, h^\bullet, \beta^\bullet) : MCyl(\mathrm{id}_{L^\bullet}) \to M^\bullet$.

4.11. ¬ In topology, the mapping cone may be obtained by contracting $X \times 0$ to a point in the mapping cylinder (cf. Exercise 4.9 and the description of the mapping cone given in §4.1).

In the algebraic version, this contraction amounts to mod-ing out the first component of $MCyl(\alpha)^\bullet$. For a cochain morphism $\alpha^\bullet : L^\bullet \to M^\bullet$, prove that there is an exact sequence of cochain complexes

$$0 \longrightarrow L^\bullet \longrightarrow MCyl(\alpha)^\bullet \longrightarrow MC(\alpha)^\bullet \longrightarrow 0\,.$$

[4.14]

4.12. ¬ With notation as in Exercise 4.9 (in particular, for a cochain morphism $\alpha^\bullet : L^\bullet \to M^\bullet$), prove that there is an exact sequence of cochain complexes

$$0 \longrightarrow M^\bullet \longrightarrow MCyl(\alpha)^\bullet \longrightarrow MC(\mathrm{id}_{L^\bullet})^\bullet \longrightarrow 0\,.$$

Deduce that there is a quasi-isomorphism $M^\bullet \to MCyl(\alpha)^\bullet$. [4.13, 4.14]

4.13. ¬ Still with notation as in Exercise 4.9, define maps $\rho^\bullet : M^\bullet \to MCyl(\alpha)^\bullet$ and $\sigma^\bullet : MCyl(\alpha)^\bullet \to M^\bullet$ by

$$\rho^i(m) = (0, 0, m),$$
$$\sigma^i(\ell, \ell', m) = m - \alpha^i(\ell).$$

Prove that $\rho^\bullet$, $\sigma^\bullet$ are both cochain morphisms. Note that $\sigma^\bullet\rho^\bullet = \mathrm{id}_{M^\bullet}$, and prove that $\rho^\bullet\sigma^\bullet$ is homotopic to $\mathrm{id}_{MCyl(\alpha)^\bullet}$. (Hint: $(\ell, \ell', m) \mapsto (0, \ell, 0)$.)

Conclude that $M^\bullet$ and $MCyl(\alpha)^\bullet$ are homotopy equivalent. This strengthens the conclusion of Exercise 4.12. [4.14]

4.14. Combine Exercises 4.11, 4.12, and 4.13 to conclude that, up to homotopy equivalence, a cochain morphism $\alpha^\bullet : L^\bullet \to M^\bullet$ may be replaced with a monomorphism $L^\bullet \to MCyl(\alpha)^\bullet$ that induces the same morphism $H^\bullet(\alpha^\bullet)$ in cohomology (up to the identification $H^\bullet(M^\bullet) \cong H^\bullet(MCyl(\alpha)^\bullet)$) and whose cokernel is the mapping cone $MC(\alpha)^\bullet$.

4.15. ▷ Prove that, in general, additive functors do not preserve quasi-isomorphisms. (Hint: Example 4.6.) Prove that *exact* functors do. (Hint: Use the mapping cone.) [§4.3, §6]

4.16. ▷ Prove Theorem 4.14. [§4.3]

---

5. The homotopic category. Complexes of projectives and injectives

One message we have tried to convey in §4 is that while quasi-isomorphisms are (by definition!) the morphisms of cochain complexes preserving cohomology and therefore may be deemed to 'capture the cohomology of a complex', the stronger notion of *homotopy equivalence* is in fact more natural from the point of view of applications. This is the moral of Theorem 4.14: homotopy equivalent complexes have the same cohomology for a much better reason than complexes linked by a quasi-isomorphism. If our general aim is to understand what it means to make all quasi-isomorphisms invertible (that is, understand the derived category $\mathsf{D}(\mathsf{A})$), we may begin by making homotopy equivalences invertible. This produces a new category, the 'homotopic category' of complexes, that we approach in this section.

We also examine the privileged position of bounded complexes of *projective* and *injective* objects regarding homotopy: for these complexes, quasi-isomorphisms are necessarily homotopy equivalences. Verifying this requires rather technical considerations, but it will be necessary in order to gain some understanding of the derived category, in §6.

5.1. Homotopic maps are identified in the derived category. To see that this 'homotopic category' is a necessary stop on the way from $\mathsf{C}(\mathsf{A})$ to $\mathsf{D}(\mathsf{A})$, go back to our higher-brow considerations at the end of §4.2. Again consider any additive functor

$$\mathscr{F} : \mathsf{C}(\mathsf{A}) \longrightarrow \mathsf{D}$$

transforming quasi-isomorphisms into isomorphisms. The simplest such functor is cohomology; the universal one is the functor from $\mathsf{C}(\mathsf{A})$ to the mystifying derived category $\mathsf{D}(\mathsf{A})$. We have verified in Proposition 4.10 that homotopic maps induce the same morphism in cohomology. Might this not be the case for *any functor* $\mathscr{F}$ as above? It is!

Lemma 5.1. *Let* $\mathscr{F} : \mathsf{C}(\mathsf{A}) \longrightarrow \mathsf{D}$ *be an additive functor such that* $\mathscr{F}(\rho^\bullet)$ *is an isomorphism in* D *for all quasi-isomorphisms* $\rho^\bullet$ *in* $\mathsf{C}(\mathsf{A})$.

Let $\alpha^\bullet, \beta^\bullet : L^\bullet \to M^\bullet$ be homotopic morphisms in $\mathsf{C}(\mathsf{A})$. *Then* $\mathscr{F}(\alpha^\bullet) = \mathscr{F}(\beta^\bullet)$ *in* D.

Proof. Equivalently (as $\mathscr{F}$ is additive), we verify that if $\alpha^\bullet \sim 0$, then $\mathscr{F}(\alpha^\bullet) = 0$. By Lemma 4.7, it suffices to show that $\alpha^\bullet$ factors through an exact complex.

This exact complex is the mapping cone $MC(\mathrm{id}_{L^\bullet})^\bullet$ of the identity $\mathrm{id}_{L^\bullet} : L^\bullet \to L^\bullet$. Recall (§4.1) that

$$MC(\mathrm{id}_{L^\bullet})^i = L^{i+1} \oplus L^i,$$

with differential defined by

$$d^i_{MC(\mathrm{id}_{L^\bullet})}(a, b) = (-d^{i+1}_{L^\bullet}(a), a + d^i_{L^\bullet}(b)).$$

Since the identity is trivially a quasi-isomorphism, $MC(\mathrm{id}_{L^\bullet})$ is an exact complex (Corollary 4.2). The morphism

$$L^i \to MC(\mathrm{id}_{L^\bullet})^i, \quad b \mapsto (0, b)$$

always commutes with differentials, so it defines a morphism of complexes $L^\bullet \to MC(\mathrm{id}_{L^\bullet})^\bullet$. By contrast, a morphism

$$MC(\mathrm{id}_{L^\bullet})^\bullet \to M^\bullet$$

is not always available. Under our hypothesis, however, there is a homotopy $\alpha^\bullet \sim 0$, that is, morphisms

$$h^i : L^i \to M^{i-1}$$

such that $\alpha^i = d^{i-1}_{M^\bullet} \circ h^i + h^{i+1} \circ d^i_{L^\bullet}$. Define morphisms

$$\pi^i : MC(\mathrm{id}_{L^\bullet})^i \to M^i \quad \text{by} \quad (a, b) \mapsto h^{i+1}(a) + \alpha^i(b).$$

It is clear that the composition

$$L^i \to MC(\mathrm{id}_{L^\bullet})^i \to M^i$$

is α^i. All we have to check now is that the collection π^i defines a morphism of complexes: this will show that $\alpha^\bullet$ factors through the exact complex $MC(\mathrm{id}_{L^\bullet})^\bullet$, hence that $\mathscr{F}(\alpha^\bullet) = 0$ (by Lemma 4.7), concluding the proof of this lemma. That is, we have to verify that

$$(*) \qquad d^{i-1}_{M^\bullet} \circ \pi^{i-1} = \pi^i \circ d^{i-1}_{MC(\mathrm{id}_{L^\bullet})^\bullet}$$

for all i. The left-hand side of (*) acts as follows:

$$(a, b) \mapsto h^i(a) + \alpha^{i-1}(b) \mapsto d^{i-1}_{M^\bullet}(h^i(a) + \alpha^{i-1}(b)).$$

The right-hand side acts as

$$(a, b) \mapsto (-d^i_{L^\bullet}(a), a + d^{i-1}_{L^\bullet}(b)) \mapsto -h^{i+1}(d^i_{L^\bullet}(a)) + \alpha^i(a + d^{i-1}_{L^\bullet}(b)).$$

Thus the needed equality is

$$d^{i-1}_{M^\bullet}(h^i(a)) + d^{i-1}_{M^\bullet}(\alpha^{i-1}(b)) = -h^{i+1}(d^i_{L^\bullet}(a)) + \alpha^i(a) + \alpha^i(d^{i-1}_{L^\bullet}(b))$$

or, equivalently,

$$(d^{i-1}_{M^\bullet} \circ h^i + h^{i+1} \circ d^i_{L^\bullet} - \alpha^i)(a) = (\alpha^i \circ d^{i-1}_{L^\bullet} - d^{i-1}_{M^\bullet} \circ \alpha^{i-1})(b)$$

$\forall a \in L^{i+1}$, $\forall b \in L^i$. But both sides are 0: the right-hand side because $\alpha^\bullet$ is a morphism of complexes and the left-hand side because h is a homotopy $\alpha^\bullet \sim 0$. $\square$

Lemma 5.1 tells us that every functor transforming quasi-isomorphisms into isomorphisms must factor through the category obtained from $\mathsf{C}(\mathsf{A})$ by identifying together homotopic morphisms. It is time to define this category.

5.2. Definition of the homotopic category of complexes.

Definition 5.2. Let A be an abelian category. The *homotopy category* $\mathsf{K}(\mathsf{A})$ *of cochain complexes in* A is the category whose objects are the cochain complexes in A (that is, the same objects of $\mathsf{C}(\mathsf{A})$) and whose morphisms are

$$\mathrm{Hom}_{\mathsf{K}(\mathsf{A})}(L^\bullet, M^\bullet) := \mathrm{Hom}_{\mathsf{C}(\mathsf{A})}(L^\bullet, M^\bullet)/\sim$$

where $\sim$ is the homotopy relation. ⌟

Yes, but is this a category? Recall that the homotopy relation $\sim$ respects composition (Exercise 4.6); in other words, the operation of composition

$$\mathrm{Hom}_{\mathsf{C}(\mathsf{A})}(L^\bullet, M^\bullet) \times \mathrm{Hom}_{\mathsf{C}(\mathsf{A})}(M^\bullet, N^\bullet) \longrightarrow \mathrm{Hom}_{\mathsf{C}(\mathsf{A})}(L^\bullet, N^\bullet)$$

does descend to an operation

$$\mathrm{Hom}_{\mathsf{K}(\mathsf{A})}(L^\bullet, M^\bullet) \times \mathrm{Hom}_{\mathsf{K}(\mathsf{A})}(M^\bullet, N^\bullet) \longrightarrow \mathrm{Hom}_{\mathsf{K}(\mathsf{A})}(L^\bullet, N^\bullet).$$

Checking the category axioms is routine and is left to the patient reader.

Bounded variations $\mathsf{K}^-(\mathsf{A})$, $\mathsf{K}^+(\mathsf{A})$, etc., are obtained likewise from $\mathsf{C}^-(\mathsf{A})$, $\mathsf{C}^+(\mathsf{A})$, etc. The considerations which follow will apply to all these versions. The further variations $\mathsf{K}^-(\mathsf{P})$, resp., $\mathsf{K}^+(\mathsf{I})$, in which the complexes will be required to consist of *projective,* resp., *injective,* objects from A (cf. §5.3) will also play a crucially important role.

As to the general structure underlying the homotopic category,

Lemma 5.3. *Let* A *be an abelian category. Then the homotopic category* $\mathsf{K}(\mathsf{A})$ *of complexes is an additive category.*

Proof. Exercise 5.1. □

However, note that, in general, *the homotopic category is not abelian.* Indeed, homotopic maps do not have the same kernel or cokernel in general, so defining these notions becomes problematic. As we already mentioned in §3.4, $\mathsf{K}(\mathsf{A})$ is a *triangulated* category; the 'distinguished triangles' are the triangles arising from the cones of morphisms, as in §4.1. (We will briefly describe the situation in §9.2.)

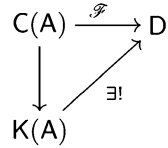
The main motivation for the introduction of the homotopic category is that homotopy equivalences become isomorphisms in $\mathsf{K}(\mathsf{A})$. More precisely, by definition there is a functor

$$\mathsf{C}(\mathsf{A}) \longrightarrow \mathsf{K}(\mathsf{A}),$$

mapping every object to 'itself' and every morphism to its homotopy class. Homotopy equivalences in $\mathsf{C}(\mathsf{A})$ become isomorphisms in $\mathsf{K}(\mathsf{A})$ because the relation $\alpha^\bullet \circ \beta^\bullet \sim \mathrm{id}$ in $\mathsf{C}(\mathsf{A})$ becomes $\alpha^\bullet \circ \beta^\bullet = \mathrm{id}$ in $\mathsf{K}(\mathsf{A})$. The homotopic category is obtained from $\mathsf{C}(\mathsf{A})$ by making all homotopy equivalences invertible on the nose.

We can then reinterpret Lemma 5.1 as the following 'factorization' result:

Proposition 5.4. *Let $\mathscr{F} : \mathsf{C}(\mathsf{A}) \longrightarrow \mathsf{D}$ be an additive functor such that $\mathscr{F}(\rho^\bullet)$ is an isomorphism in D for all quasi-isomorphisms $\rho^\bullet$ in $\mathsf{C}(\mathsf{A})$. Then $\mathscr{F}$ factors uniquely through $\mathsf{K}(\mathsf{A})$:*



Proof. Exercise 5.2. □

In particular, there must be a unique functor from $\mathsf{K}(\mathsf{A})$ to the subtler derived category $\mathsf{D}(\mathsf{A})$. The situation is of course analogous for all bounded versions of these objects: the natural functors from the appropriately bounded complexes to the corresponding derived categories factor uniquely through the corresponding homotopic categories:

$$\begin{array}{ccc} \mathsf{C}^-(\mathsf{A}) & & \mathsf{C}^+(\mathsf{A}) \\ \downarrow & & \downarrow \\ \mathsf{K}^-(\mathsf{A}) & & \mathsf{K}^+(\mathsf{A}) \\ \downarrow & & \downarrow \\ \mathsf{D}^-(\mathsf{A}) & & \mathsf{D}^+(\mathsf{A}) \end{array}$$

A particular case of Proposition 5.4 is the fact that the cohomology functor on $\mathsf{C}(\mathsf{A})$ (resp., $\mathsf{C}^-(\mathsf{A})$, $\mathsf{C}^+(\mathsf{A})$) descends to a functor on $\mathsf{K}(\mathsf{A})$ (resp., $\mathsf{K}^-(\mathsf{A})$, $\mathsf{K}^+(\mathsf{A})$). This fact also follows directly from Proposition 4.10: homotopic maps induce the same morphism in cohomology.

5.3. Complexes of projective and injective objects. We should now ask 'how far' $\mathsf{K}(\mathsf{A})$ may be from $\mathsf{D}(\mathsf{A})$. We have inverted all homotopy equivalences in $\mathsf{K}(\mathsf{A})$, but not all quasi-isomorphisms are homotopy equivalences: keep Example 4.6 in mind. If we want to think of $\mathsf{K}(\mathsf{A})$ as a first approximation to $\mathsf{D}(\mathsf{A})$, we should focus on complexes for which there is essentially no distinction between quasi-isomorphisms and homotopy equivalences.

It turns out that there are such complexes: those consisting of projective or injective objects.

Recall (Definition VIII.6.1) that an R-module M is 'projective' if the functor $\mathrm{Hom}_R(M, \_)$ is exact and it is 'injective' if $\mathrm{Hom}_R(\_, M)$ is exact. We can adopt these definitions in any abelian category:

Definition 5.5. Let A be an abelian category. An object P of A is *projective* if the functor $\mathrm{Hom}_{\mathsf{A}}(P, \_)$ is exact. An object Q is *injective* if the functor $\mathrm{Hom}_{\mathsf{A}}(\_, Q)$ is exact. ⌟

General remarks such as Lemma VIII.6.2 or the comments about splitting of sequences at the end of §VIII.6.1 hold in any abelian category, and a good exercise

for the reader is to collect all such information and verify that it does hold in this new context.

Also note that, at this level of generality, the parts of the theory dealing with injective objects are a faithful mirror of the parts dealing with projective objects. This is of course not a coincidence: the opposite of an abelian category is again abelian (Exercise 1.10), and projectives in one become injectives in the other. Thus, it is really only necessary to prove statements for, say, projectives in arbitrary abelian categories: the corresponding statements for injectives will automatically follow.

This dual state of affairs should not be taken too far: projectives and injectives in a *fixed* abelian category may display very different features. For example, the characterizations worked out in §VIII.6.2 and §VIII.6.3 use the specific properties of the categories R-Mod, so they should not be expected to have a simple counterpart in arbitrary abelian categories.

We will denote by P, resp., I, the full subcategories of A determined by the projective, resp., injective, objects. Of course these are not abelian categories in any interesting case. Note that an abelian category may well have no nontrivial projective or injective objects.

Example 5.6. The category of *finite* abelian groups is abelian (surprise, surprise), but contains no nontrivial projective or injective objects (Exercise 5.5). ⌟

Definition 5.7. An abelian category A has *enough projectives* if for every object A in A there exists a projective object P in A and an epimorphism $P \twoheadrightarrow A$. The category has *enough injectives* if for every object A in A there is an injective object Q in A and a monomorphism $A \rightarrowtail Q$. ⌟

These definitions are not new to the reader, since we ran across them in §VIII.6: in particular, we have already observed that, for every commutative ring R, R-Mod *has enough projectives* (this is not challenging: free modules and their direct summands are projective, Proposition VIII.6.4) and *enough injectives* (this *is* challenging; we verified it in Corollary VIII.6.12).

Example 5.6 shows that we should not expect an abelian category A to have either property. An important case in which one can show that there are enough injectives is the category of *sheaves* of abelian groups over a topological space; this is a key step in the definition of sheaf cohomology as a 'derived functor'. In general, categories of sheaves do not have enough projectives. On the other hand, the category of finitely generated abelian groups has enough projectives but not enough (indeed, no nontrivial) injectives. So it goes.

5.4. Homotopy equivalences vs. quasi-isomorphisms in $\mathsf{K}(\mathsf{A})$. As we already mentioned, we will be interested in certain subcategories of $\mathsf{K}(\mathsf{A})$ determined by complexes of projective or injective objects of A.

Definition 5.8. We will denote by $\mathsf{K}^-(\mathsf{P})$ the full subcategory of $\mathsf{K}(\mathsf{A})$ consisting of bounded-above complexes of projective objects of A, and we will denote by $\mathsf{K}^+(\mathsf{I})$ the full subcategory of $\mathsf{K}(\mathsf{A})$ consisting of bounded-below complexes of injective objects of A. ⌟

The main result of the rest of this section is the following:

Theorem 5.9. *Let* A *be an abelian category, and let* $\alpha^\bullet : P_0^\bullet \to P_1^\bullet$, *resp.,* $\alpha^\bullet : Q_0^\bullet \to Q_1^\bullet$, *be a quasi-isomorphism between bounded-above complexes of projectives, resp., bounded-below complexes of injectives, in* A. *Then* $\alpha^\bullet$ *is a homotopy equivalence.*

Corollary 5.10. *In* $\mathsf{K}^-(\mathsf{P})$ *and* $\mathsf{K}^+(\mathsf{I})$, *(homotopy classes of) quasi-isomorphisms are isomorphisms.*

That is, all quasi-isomorphisms between bounded complexes of projectives or injectives are 'already' inverted in $\mathsf{K}^-(\mathsf{P})$, $\mathsf{K}^+(\mathsf{I})$. Thus, these categories are very close to the corresponding derived categories. In fact, we will see that if A has enough projectives, then $\mathsf{K}^-(\mathsf{P})$ 'suffices' to describe $\mathsf{D}^-(\mathsf{A})$ (in a sense that will be made more precise); similarly, $\mathsf{K}^+(\mathsf{I})$ acts as a concrete realization of $\mathsf{D}^+(\mathsf{A})$ if A has enough injectives.

The proof of Theorem 5.9 will require a few preliminaries, which further clarify the role played by complexes of projective and injective objects with regard to homotopies.

A complicated way of saying that a complex $N^\bullet$ in $\mathsf{C}(\mathsf{A})$ is exact is to assert that the identity map $\mathrm{id}_{N^\bullet}$ and the trivial map 0 induce the same morphism in cohomology, as they would if they were homotopic to each other. It is however easy to construct examples of exact complexes for which the identity is *not* homotopic to 0. As the reader will verify (Exercise 5.11), this has to do with whether the complex splits or not; in general, a complex $N^\bullet$ is said to be 'split exact' if $\mathrm{id}_{N^\bullet}$ is homotopic to 0. Keeping in mind that projective or injective objects 'cause' sequences to split, it may not be too surprising that for bounded exact complexes consisting of projective or injective objects the identity *does* turn out to be homotopic to 0.

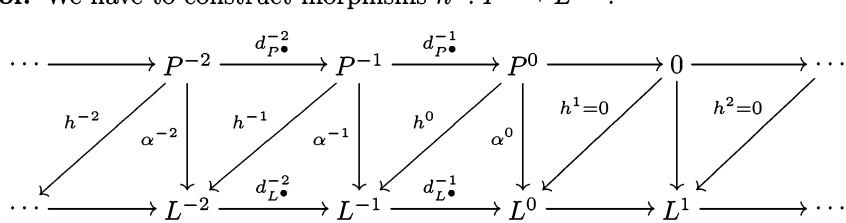
To verify this, we study special conditions that ensure that morphisms are homotopic to 0.

Lemma 5.11. *Let* $P^\bullet$ *be a complex of projective objects of an abelian category* A *such that* $P^i = 0$ *for* $i > 0$, *and let* $L^\bullet$ *be a complex in* $\mathsf{C}(\mathsf{A})$ *such that* $H^i(L^\bullet) = 0$ *for* $i < 0$.

Let $\alpha^\bullet : P^\bullet \to L^\bullet$ *be a morphism inducing the zero-morphism in cohomology. Then* $\alpha^\bullet$ *is homotopic to* 0.

The reader will provide the 'injective' version of this statement, dealing with morphisms from a complex which is exact in positive degree to a complex of injectives $Q^\bullet$ with $Q^i = 0$ for $i < 0$.

Proof. We have to construct morphisms $h^i : P^i \to L^{i-1}$:



such that

$$(*) \qquad \alpha^i = d^{i-1}_{L^\bullet} \circ h^i + h^{i+1} \circ d^i_{P^\bullet}.$$

Of course $h^i = 0$ necessarily for $i > 0$. For $i = 0$, use the fact that the morphism induced by $\alpha^\bullet$ on cohomology is 0; this says that α^0 factors through the image of $d^{-1}_{L^\bullet}$:

$$\begin{array}{ccccc} & & P^0 & & \\ & & \downarrow{\scriptstyle \alpha^0} & & \\ L^{-1} & \xrightarrow{d^{-1}_{L^\bullet}} & \operatorname{im} d^{-1}_{L^\bullet} & \longrightarrow & 0 \end{array}$$

Since P^0 is projective, there exists a morphism $h^0 : P^0 \to L^{-1}$ such that $\alpha^0 = d^{-1}_{L^\bullet} \circ h^0$, as needed.

To define h^{i-1} for $i \leq 0$, proceed inductively and assume that h^i and h^{i+1} satisfying (*) have already been constructed. Note that by (*) (and the fact that $P^\bullet$ is a complex)

$$d^{i-1}_{L^\bullet} \circ h^i \circ d^{i-1}_{P^\bullet} = (\alpha^i - h^{i+1} \circ d^i_{P^\bullet}) \circ d^{i-1}_{P^\bullet} = \alpha^i \circ d^{i-1}_{P^\bullet} - h^{i+1} \circ 0 = \alpha^i \circ d^{i-1}_{P^\bullet} :$$

$$\begin{array}{ccccccc} \cdots & P^{i-1} & \xrightarrow{d^{i-1}_{P^\bullet}} & P^i & \xrightarrow{d^i_{P^\bullet}} & P^{i+1} & \cdots \\ & & {\scriptstyle h^i} \searrow & \downarrow{\scriptstyle \alpha^i} & \swarrow {\scriptstyle h^{i+1}} & & \\ \cdots & L^{i-1} & \xrightarrow{d^{i-1}_{L^\bullet}} & L^i & & & \cdots \end{array}$$

Therefore

$$d^{i-1}_{L^\bullet} \circ (\alpha^{i-1} - h^i \circ d^{i-1}_{P^\bullet}) = d^{i-1}_{L^\bullet} \circ \alpha^{i-1} - \alpha^i \circ d^{i-1}_{P^\bullet} = 0,$$

since $\alpha^\bullet$ is a morphism of cochain complexes. This tells us that $\alpha^{i-1} - h^i \circ d^{i-1}_{P^\bullet}$ factors through $\ker d^{i-1}_{L^\bullet}$. Since $L^\bullet$ is exact at L^{i-1} for $i \leq 0$, $\ker d^{i-1}_{L^\bullet} = \operatorname{im} d^{i-2}_{L^\bullet}$. Therefore, we again have a factorization

$$\begin{array}{ccccc} & & P^{i-1} & & \\ & & \downarrow{\scriptstyle \alpha^{i-1} - h^i \circ d^{i-1}_{P^\bullet}} & & \\ L^{i-2} & \xrightarrow{d^{i-2}_{L^\bullet}} & \operatorname{im} d^{i-2}_{L^\bullet} & \longrightarrow & 0 \end{array}$$

Now P^{i-1} is projective; therefore there exists $h^{i-1} : P^{i-1} \to L^{i-2}$ such that

$$\alpha^{i-1} - h^i \circ d^{i-1}_{P^\bullet} = d^{i-2}_{L^\bullet} \circ h^{i-1},$$

which is precisely what we need for the induction step. □

Corollary 5.12. *Let $P^\bullet$ be a bounded-above cochain complex of projectives of an abelian category A, and let $L^\bullet$ be an* exact *complex in $\mathsf{C}(\mathsf{A})$.*

Then every morphism of complexes $P^\bullet \to L^\bullet$ is homotopic to 0.

This follows immediately from (a harmless shift of) Lemma 5.11, since every morphism to an exact complex has no choice but to induce the zero-morphism in cohomology.

The 'injective' version of Corollary 5.12 is that morphisms *from* an exact complex $L^\bullet$ *to* a bounded-*below* complex $Q^\bullet$ of injectives are necessarily homotopy equivalent to 0.

Note that in all these considerations the complexes $P^\bullet$, $Q^\bullet$ are required to live in the 'bounded' versions $\mathsf{C}^-(\mathsf{P})$, $\mathsf{C}^+(\mathsf{I})$: all objects should vanish in degree sufficiently high or low. This furnishes the 'start' of the inductive construction of the homotopy in the proof of Lemma 5.11. It is an essentially inescapable feature of the theory.

Corollary 5.13. *Let $P^\bullet$ (resp., $Q^\bullet$) be a bounded-above* exact *complex of projectives (resp., a bounded-below* exact *complex of injectives). Then $P^\bullet$ (resp., $Q^\bullet$) is homotopy equivalent to the zero-complex.*

Proof. Exercise 5.12. □

Corollary 5.13 is the first manifestation of the principle captured more fully by Theorem 5.9: we have just verified that, for suitably bounded complexes of projectives or injectives, 'quasi-isomorphic to 0' is the same as 'homotopy equivalent to 0'.

Another consequence of Lemma 5.11 is the following remark, showing that quasi-isomorphisms are 'non-zero-divisors' up to homotopy, with respect to morphisms from complexes of projectives. This will be useful in the next section, for example to show that certain lifts are uniquely defined up to homotopy.

Lemma 5.14. *Let A be an abelian category, and let $\rho^\bullet : L^\bullet \to M^\bullet$ be a quasi-isomorphism in $\mathsf{C}(\mathsf{A})$. Let $P^\bullet$ be a bounded-above complex of projectives, and let $\alpha^\bullet : P^\bullet \to L^\bullet$ be a morphism of cochain complexes such that the composition*

$$P^\bullet \xrightarrow{\alpha^\bullet} L^\bullet \xrightarrow[q\text{-}iso.]{\rho^\bullet} M^\bullet$$

is homotopic to the zero-morphism. Then $\alpha^\bullet \sim 0$.

Proof. Let $h^i : P^i \to M^{i-1}$ define a homotopy between $\rho^\bullet \circ \alpha^\bullet$ and 0, so that $-\rho^i \circ \alpha^i = d^{i-1}_{M^\bullet} \circ h^i + h^{i+1} \circ d^i_{P^\bullet}$. Consider the mapping cone $MC(\rho)^\bullet$ of $\rho^\bullet$ (§4.1), and define morphisms

$$\beta^i = (\alpha^i, h^i) : P^i \to L^i \oplus M^{i-1} = MC(\rho)^{i-1} :$$

$$\begin{array}{ccccccccc}
\cdots & \longrightarrow & P^{i-1} & \xrightarrow{d^{i-1}_{P^\bullet}} & P^i & \xrightarrow{d^i_{P^\bullet}} & P^{i+1} & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle \beta^{i-1}} & & \downarrow{\scriptstyle \beta^i} & & \downarrow{\scriptstyle \beta^{i+1}} & & \\
\cdots & \longrightarrow & L^{i-1} \oplus M^{i-2} & \xrightarrow{-d^{i-2}_{MC(\rho)^\bullet}} & L^i \oplus M^{i-1} & \xrightarrow{-d^{i-1}_{MC(\rho)^\bullet}} & L^{i+1} \oplus M^i & \longrightarrow & \cdots
\end{array}$$

We claim that these give a morphism of cochain complexes $P^\bullet \to MC(\rho)[-1]^\bullet$. Indeed[24],

$$\beta^{i+1} \circ d^i_{P^\bullet} = (\alpha^{i+1} \circ d^i_{P^\bullet}, h^{i+1} \circ d^i_{P^\bullet}) \quad \text{and}$$
$$-d^{i-1}_{MC(\rho)} \circ \beta^i = (d^i_{L^\bullet} \circ \alpha^i, -\rho^i \circ \alpha^i - d^{i-1}_{M^\bullet} \circ h^i);$$

[24]It is time to review how the differential of the mapping cone was defined in §4.1. The negative sign will now come in very handy.

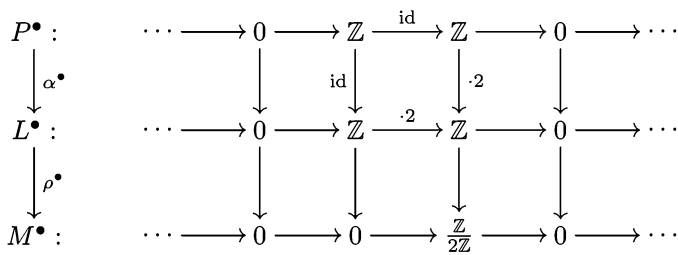
these are equal, by definition of h^i and since $\alpha^\bullet$ is a morphism of complexes.

Now, $\rho^\bullet$ is a quasi-isomorphism by hypothesis. Therefore $MC(\rho)^\bullet$ is an exact complex (Corollary 4.2). Applying Corollary 5.12, it follows that $\beta^\bullet$ is homotopic to zero. But $\alpha^\bullet$ is the composition

$$P^\bullet \xrightarrow{\beta^\bullet} MC(\rho)[-1]^\bullet = L^\bullet \oplus M[-1]^\bullet \longrightarrow L^\bullet \ ,$$

so (Exercise 4.6, part (i)) the fact that $\beta^\bullet \sim 0$ implies that $\alpha^\bullet \sim 0$, concluding the proof. □

Example 5.15. To see that $\alpha^\bullet$ may not be zero on the nose even if $\rho^\bullet \circ \alpha^\bullet = 0$, look back again at Example 4.6:



Here $\rho^\bullet$ is a quasi-isomorphism, and $\rho^\bullet \circ \alpha^\bullet = 0$. According to Lemma 5.14, the (nonzero) morphism $\alpha^\bullet$ is homotopic to 0. (Indeed, a homotopy is immediately visible. What is it?) ⌟

5.5. Proof of Theorem 5.9. Theorem 5.9 will follow from a more general observation: a quasi-isomorphism *to* a complex of projectives, resp., *from* a complex of injectives[25], has a *right,* resp., *left,* homotopy inverse.

Proposition 5.16. *Let* A *be an abelian category, and let* $L^\bullet$ *be a complex in* $\mathsf{C}(\mathsf{A})$. *Let* $P^\bullet$ *in* $\mathsf{C}^-(\mathsf{P})$ *be a bounded-above complex of projectives, and let* $\alpha^\bullet : L^\bullet \to P^\bullet$ *be a quasi-isomorphism. Then there exists a morphism of complexes* $\beta^\bullet : P^\bullet \to L^\bullet$ *such that* $\alpha^\bullet \circ \beta^\bullet$ *is homotopic to* $id_{P^\bullet}$.

(For the injective version, if $Q^\bullet$ is a bounded-below complex of injectives and $\alpha^\bullet : Q^\bullet \to L^\bullet$ is a quasi-isomorphism, then there exists a morphism $\beta^\bullet : L^\bullet \to Q^\bullet$ such that $\beta^\bullet \circ \alpha^\bullet \sim \mathrm{id}_{Q^\bullet}$.)

Proof. Since $\alpha^\bullet : L^\bullet \to P^\bullet$ is a quasi-isomorphism, the mapping cone $MC(\alpha)^\bullet$ of α is an exact complex (Corollary 4.2). Let $\rho^\bullet$ be the morphism of complexes

$$\rho^\bullet = (0, \mathrm{id}_{P^\bullet}) : P^\bullet \to L[1]^\bullet \oplus P^\bullet = MC(\alpha)^\bullet.$$

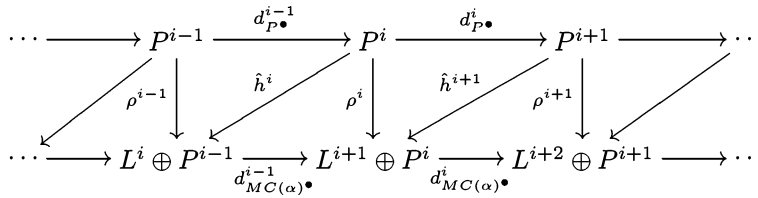
Since $MC(\alpha)^\bullet$ is exact, $\rho^\bullet$ is homotopic to zero by Corollary 5.12. Therefore, there exist morphisms

$$\hat{h}^i \qquad P^i \to MC(\alpha)^{i-1} = L^i \oplus P^{i-1}$$

such that

$$(*) \qquad \rho^i = d^{i-1}_{MC(\alpha)^\bullet} \circ \hat{h}^i + \hat{h}^{i+1} \circ d^i_{P^\bullet} :$$

[25] That is, going 'the wrong way': projectives like being the sources of morphisms, and injectives like to be targets.



Now we unravel what this says. Write $\hat{h}^i$ out in components:

$$\hat{h}^i = (\beta^i, h^i)$$

with $\beta^i : P^i \to L^i$ and $h^i : P^i \to P^{i-1}$. We claim that

(i) the collection $\{\beta^i\}$ is a morphism of cochain complexes: $\beta^{i+1} \circ d^i_{P^\bullet} = d^i_{L^\bullet} \circ \beta^i$ for all i and

(ii) the morphisms h^i give a homotopy $\alpha^\bullet \circ \beta^\bullet \sim \mathrm{id}_{P^\bullet}$.

Indeed, the left-hand side of (*) is

$$\rho^i = (0, \mathrm{id}_{P^i});$$

the right-hand side is

$$d^{i-1}_{MC(\alpha)^\bullet} \circ (\beta^i, h^i) + (\beta^{i+1}, h^{i+1}) \circ d^i_{P^\bullet}$$
$$= (-d^i_{L^\bullet} \circ \beta^i, \alpha^i \circ \beta^i + d^{i-1}_{P^\bullet} \circ h^i) + (\beta^{i+1} \circ d^i_{P^\bullet}, h^{i+1} \circ d^i_{P^\bullet}).$$

It follows that (*) amounts to

$$\begin{cases} \beta^{i+1} \circ d^i_{P^\bullet} - d^i_{L^\bullet} \circ \beta^i = 0, \\ \mathrm{id}_{P^i} - \alpha^i \circ \beta^i = d^{i-1}_{P^\bullet} \circ h^i + h^{i+1} \circ d^i_{P^\bullet}, \end{cases}$$

that is, precisely (i) and (ii).

The morphism of complexes $\beta^\bullet : P^\bullet \to L^\bullet$ is the needed homotopy right-inverse of $\alpha^\bullet$. □

Taking $L^\bullet$ to be a suitably bounded complex of projectives or injectives finally establishes Theorem 5.9:

Proof of Theorem 5.9. We will give the argument for projectives.

By Proposition 5.16, since $P_1^\bullet$ is in $\mathsf{C}^-(\mathsf{P})$ and $\alpha^\bullet : P_0^\bullet \to P_1^\bullet$ is a quasi-isomorphism, there exists a morphism of complexes $\beta^\bullet : P_1^\bullet \to P_0^\bullet$ such that

$$\alpha^\bullet \circ \beta^\bullet \sim \mathrm{id}_{P_1^\bullet}.$$

This implies that $H^\bullet(\alpha^\bullet) \circ H^\bullet(\beta^\bullet) = \mathrm{id}$ (Proposition 4.10). Since $H^\bullet(\alpha^\bullet)$ is invertible, so is $H^\bullet(\beta^\bullet)$: therefore, $\beta^\bullet$ is a quasi-isomorphism. Now $P_0^\bullet$ is also in $\mathsf{C}^-(\mathsf{P})$; hence by Proposition 5.16 there exists a morphism of complexes $\alpha^{\bullet\prime} : P_0^\bullet \to P_1^\bullet$ such that $\beta^\bullet \circ \alpha^{\bullet\prime} \sim \mathrm{id}_{P_0^\bullet}$. Since

$$\alpha^{\bullet\prime} \sim (\alpha^\bullet \circ \beta^\bullet) \circ \alpha^{\bullet\prime} = \alpha^\bullet \circ (\beta^\bullet \circ \alpha^{\bullet\prime}) \sim \alpha^\bullet,$$

(Exercise 4.6), it follows that $\beta^\bullet \circ \alpha^\bullet \sim \mathrm{id}_{P_0^\bullet}$. Thus $\alpha^\bullet$ is a homotopy equivalence, as stated. □

Exercises

5.1. ▷ Prove that the homotopic categories $\mathsf{K}(\mathsf{A})$, $\mathsf{K}^-(\mathsf{A})$, $\mathsf{K}^+(\mathsf{A})$ of an abelian category A are additive. [§5.2]

5.2. ▷ Prove Proposition 5.4. [§5.2]

5.3. Provide a reasonable definition of 'projective' and 'injective' objects in any category, and prove that, according to your definition, every set is both projective and injective in Set.

5.4. ▷ Upgrade Exercises VIII.6.4 and VIII.6.15 to objects of any abelian category. [5.9, 6.3, §7.4]

5.5. ▷ Let F be a nontrivial finite abelian group. Prove that there are exact sequences of finite abelian groups

$$0 \to A_1 \to A_2 \to F \to 0, \quad 0 \to F \to B_1 \to B_2 \to 0$$

which do *not* split.

Deduce that the category of finite abelian groups has no nontrivial projective or injective objects. [§5.3]

5.6. ▷ Let A, B be abelian categories, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$, $\mathscr{G} : \mathsf{B} \to \mathsf{A}$ be additive functors. Assume that $\mathscr{F}$ is left-adjoint to $\mathscr{G}$ and that $\mathscr{G}$ preserves epimorphisms. Prove that if P is a projective object in A, then $\mathscr{F}(P)$ is a projective object in B. Formulate an analogous result for injective objects. [5.8, §7.1]

5.7. ¬ Let A be an abelian category, and let C be a small category. Assume that A contains all products indexed by any set[26]; for example, R-Mod satisfies this condition for every ring R. For an object A of A and a set S, denote by A^S the product of A indexed by S. Note that any set-map $S \to T$ determines a morphism $A^T \to A^S$ in A.

For any object X of C, we will define a functor $\widehat{\mathscr{X}} : \mathsf{A} \to \mathsf{A}^{\mathsf{C}}$; a much more simple-minded functor $\mathscr{X} : \mathsf{A}^{\mathsf{C}} \to \mathsf{A}$ was defined in Exercise 1.11.

For every object A of A, $\widehat{\mathscr{X}}(A)$ must be an object of A^{C}, that is, a functor $\mathsf{C} \to \mathsf{A}$. The value of this functor at the object Y of C is prescribed to be

$$\widehat{\mathscr{X}}(A)Y := A^{\mathrm{Hom}_{\mathsf{C}}(Y,X)}.$$

Every morphism $Y \to Z$ in C determines a function $\mathrm{Hom}_{\mathsf{C}}(Z,X) \to \mathrm{Hom}_{\mathsf{C}}(Y,X)$, and this defines a morphism $\widehat{\mathscr{X}}(A)Y \to \widehat{\mathscr{X}}(A)Z$.

- Prove that this prescription defines $\widehat{\mathscr{X}}(A)$ as a functor $\mathsf{C} \to \mathsf{A}$.
- Prove that $\widehat{\mathscr{X}}$ is a functor $\mathsf{A} \to \mathsf{A}^{\mathsf{C}}$, with the evident action on morphisms.
- Prove that $\widehat{\mathscr{X}}$ is right-adjoint to $\mathscr{X}$.

[26] Recall that 'infinite' products may be defined as limits; see Example VIII.1.10.

(Hint: You have to identify $\mathrm{Hom}_{\mathsf{A}^{\mathsf{C}}}(\mathscr{F}, \widehat{\mathscr{X}}(A))$ with $\mathrm{Hom}_{\mathsf{A}}(\mathscr{F}(X), A)$. Given a natural transformation $\alpha : \mathscr{F} \to \widehat{\mathscr{X}}(A)$, evaluate at X and extract the identity component to obtain a morphism $\mathscr{F}(X) \to A$. Then show that this one piece of information determines the whole natural transformation.) [5.8, 5.9]

5.8. ¬ As in Exercise 5.7, let C be a small category and let A be an abelian category containing all products indexed by sets. Prove that the functor $\widehat{\mathscr{X}}$ preserves injectives for all objects X of C: if Q is an injective object of A, then $\widehat{\mathscr{X}}(Q)$ is injective in A^{C}. (Hint: Exercise 5.6.) [5.9]

5.9. ¬ Let C be a small category, and let A be an abelian category containing all products indexed by sets. Assume that A has enough injectives. Let $\mathscr{F} : \mathsf{C} \to \mathsf{A}$ be a functor. For every object X of C, let Q_X be an injective object of A admitting a monomorphism $\mathscr{F}(X) \rightarrowtail Q_X$.

- With notation as in Exercise 5.7, let $\mathscr{Q} := \prod_{X \in \mathrm{Obj}(\mathsf{C})} \widehat{\mathscr{X}}(Q_X)$ (prove that this product exists in A^{C}).
- Prove that $\mathscr{Q}$ is injective in A^{C}. (Exercises 5.4 and 5.8.)
- Define a morphism $\mathscr{F} \to \mathscr{Q}$, and prove it is a monomorphism. (Use adjunction.)

Therefore, we can conclude that A^{C} has enough injectives if A is an abelian category with enough injectives and A is closed with respect to products indexed by sets.

Prove that if A is an abelian category with enough *projectives* and A is closed with respect to *co*products indexed by sets, then A^{C} has enough projectives. (Hint: $(\mathsf{A}^{\mathsf{C}})^{op} \cong (\mathsf{A}^{op})^{(\mathsf{C}^{op})}$.) [5.10]

5.10. Prove that the category of presheaves of abelian groups on a topological space has enough injectives and enough projectives[27]. (Use Exercise 5.9.)

5.11. ▷ Let $N^\bullet$ be the complex

$$\cdots \longrightarrow 0 \longrightarrow K \longrightarrow M \longrightarrow N \longrightarrow 0 \longrightarrow \cdots .$$

Prove that $N^\bullet$ is exact and splits (in a sense analogous to the one given for modules in §III.7.2) if and only if $\mathrm{id}_{N^\bullet}$ is homotopic to 0 (cf. Proposition III.7.5). Deduce that every additive functor sends split exact sequences to split exact sequences.

More generally, let $N^\bullet$ be any complex. Prove that $\mathrm{id}_{N^\bullet}$ is homotopic to 0 (that is, $N^\bullet$ is 'split exact' as defined in §5.4) if and only if $N^\bullet$ is isomorphic to a complex of the form

$$\cdots \xrightarrow{d^{-3}} M^{-2} \oplus M^{-1} \xrightarrow{d^{-2}} M^{-1} \oplus M^0 \xrightarrow{d^{-1}} M^0 \oplus M^1 \xrightarrow{d^0} M^1 \oplus M^2 \xrightarrow{d^1} \cdots$$

where d^i is 0 on the M^i factor, and $(\mathrm{id}_{M^{i+1}}, 0)$ on the M^{i+1} factor. [§5.4, 6.3, §7.4]

5.12. ▷ Prove that every exact, bounded-above complex of projectives is homotopy equivalent to zero. Prove that every exact, bounded-below complex of injectives is homotopy equivalent to zero. [§5.4]

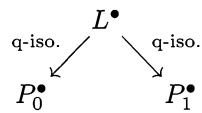
---

[27] The category of *sheaves* of abelian groups on a topological space also has enough injectives, but in general it does not have enough projectives.

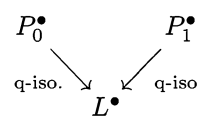
5.13. Let A be an abelian category, and let $\alpha^\bullet : L^\bullet \to M^\bullet$ be a quasi-isomorphism in $\mathsf{C}(\mathsf{A})$. Assume that $\beta^\bullet$ is a homotopy one-sided inverse of $\alpha^\bullet$. Prove that $\beta^\bullet$ is also a quasi-isomorphism.

5.14. ¬ Let $P_0^\bullet$, $P_1^\bullet$ be bounded-above complexes of projective objects in an abelian category A. (You may assume A has enough projectives, if that helps.) Let $L^\bullet$ be a (not necessarily bounded) complex in $\mathsf{C}(\mathsf{A})$.

- Assume that there are quasi-isomorphisms



  Prove that $P_0^\bullet$ and $P_1^\bullet$ are homotopy equivalent.
- Assume that there are quasi-isomorphisms



  Prove that $P_0^\bullet$ and $P_1^\bullet$ are homotopy equivalent. (This may turn out to be challenging. If so, wait until you have covered §6, in particular Theorem 6.6.)

[6.5]

---

6. Projective and injective resolutions and the derived category

Theorem 5.9 is very good news. It tells us that if we are willing and able to change our viewpoint and adopt bounded complexes of projectives or injectives as our natural environment, then we will have an excellent handle on the subtle notion of quasi-isomorphism: because quasi-isomorphisms become homotopy equivalences in that environment and homotopy equivalences are better behaved than arbitrary quasi-isomorphisms. For example, homotopy equivalence of complexes is manifestly an equivalence relation (Exercise 4.4), while quasi-isomorphism is not (see the discussion following Example 4.6). Further, homotopy equivalences are preserved by every additive functor (Theorem 4.14), while general quasi-isomorphisms are not (Exercise 4.15).

The question is therefore how to switch to a homotopic environment and adopt complexes of projectives or injectives as our primary object of study, while keeping intact the essential information carried by the objects of A or even the complexes in $\mathsf{C}(\mathsf{A})$.

This can be done if A has enough projectives or enough injectives; in this situation, we will be able to produce an adequate description of the derived category (§6.3).

6.1. Recovering A. Recall once more that a 'copy' of an abelian category A is available within $\mathsf{C}(\mathsf{A})$, by associating with every object A in A the complex $\iota(A)$ having A in degree 0 and 0 elsewhere. Composing with the functor $\mathsf{C}(\mathsf{A}) \to \mathsf{K}(\mathsf{A})$ realizes A as a full subcategory of $\mathsf{K}(\mathsf{A})$ (Exercise 6.1). We will look for 'more interesting' copies of A in the homotopic category.

Let us begin by recalling (and sharpening) our definition of projective/injective resolutions of an object A.

Definition 6.1. Let A be an object of an abelian category A.

A *projective resolution* of A is a quasi-isomorphism $P^\bullet \to \iota(A)$, where $P^\bullet$ is a complex in $\mathsf{C}^{\leq 0}(\mathsf{P})$.

An *injective resolution* of A is a quasi-isomorphism $\iota(A) \to Q^\bullet$, where $Q^\bullet$ is a complex in $\mathsf{C}^{\geq 0}(\mathsf{I})$. ⌟

By common abuse of language, we will often refer to $P^\bullet$ or $Q^\bullet$ as the resolution, leaving the quasi-isomorphism understood.

Remark 6.2. The terminology is potentially confusing, since it hints that the resolutions *themselves* may be projective/injective as objects of the abelian category $\mathsf{C}(\mathsf{A})$, or of its bounded variations. This is not the case (Exercise 6.3). ⌟

Projective or injective resolutions need not exist: as we pointed out (Example 5.6), an abelian category may have no nontrivial projective or injective objects. It is however clear that such resolutions exist *if* the category has enough projectives/injectives: if A has enough projectives, then given any object A of A there is a projective P_0 with an epimorphism $\pi : P_0 \to A$; and then a projective P_1 with an epimorphism $d^{-1} : P^{-1} \to \ker \pi$; and then a projective P^{-2} with an epimorphism $d^{-2} : P^{-2} \to \ker d^{-1}$; and so on. Letting $d^i_{P^\bullet}$ be the composition $P^i \longrightarrow \ker d^{i+1} \rightarrowtail P^{i+1}$ gives

$$\cdots \longrightarrow P^{-2} \xrightarrow{d^{-2}_{P^\bullet}} P^{-1} \xrightarrow{d^{-1}_{P^\bullet}} P^0 \longrightarrow 0 \longrightarrow \cdots,$$

a projective resolution of A. Just as clearly, if A has enough injectives, then every object A has an injective resolution.

Now consider the category of homotopy classes of quasi-isomorphisms with target $\iota(A)$ (cf. Remark 4.5): this is a 'homotopic category of resolutions' of A. We claim that *projective* resolutions, if they exist, are initial in this category. (Analogously, *injective* resolutions are final in the homotopic category of quasi-isomorphism with *source* $\iota(A)$.)

As the reader knows, this means that projective resolutions are supposed to map uniquely (in $\mathsf{K}(\mathsf{A})$) to *every* resolution of A. The following key lemma proves this and more:

Lemma 6.3. *Let A be an object of an abelian category A. Let $M^\bullet$ be a resolution of A, and let $P^\bullet$ be any complex in $\mathsf{C}^{\leq 0}(\mathsf{P})$. Let*

$$\varphi : H^0(P^\bullet) \to H^0(M^\bullet) = A$$

be an arbitrary morphism. Then

- *there exist morphisms of complexes $\alpha^\bullet : P^\bullet \to M^\bullet$ inducing φ at the level of H^0;*
- *different morphisms satisfying the previous requirement are necessarily homotopy equivalent.*

Of course an analogous statement holds for complexes $Q^\bullet$ in $\mathsf{C}^{\geq 0}(\mathsf{I})$, giving morphisms of cochain complexes $M^\bullet \to Q^\bullet$ inducing a given morphism in H^0.

Proof. We have to define $\alpha^i : P^i \to M^i$ for all i. Since $\alpha^i = 0$ necessarily for $i > 0$, we may as well replace $M^\bullet$ with its truncated version (cf. Exercise 3.1) and then extend both $P^\bullet$ and this complex as follows:

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & P^{-2} & \xrightarrow{d^{-2}_{P^\bullet}} & P^{-1} & \xrightarrow{d^{-1}_{P^\bullet}} & P^0 & \xrightarrow{\pi} & H^0(P^\bullet) & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \alpha^{-2}} & & \downarrow{\scriptstyle \alpha^{-1}} & & \downarrow{\scriptstyle \alpha^0} & & \downarrow{\scriptstyle \varphi} & & \\
\cdots & \longrightarrow & M^{-2} & \xrightarrow{d^{-2}_{M^\bullet}} & M^{-1} & \xrightarrow{d^{-1}_{M^\bullet}} & \ker d^0_{M^\bullet} & \xrightarrow{\mu} & H^0(M^\bullet) = A & \longrightarrow & 0
\end{array}$$

Note that the bottom complex is then *exact*. The morphism φ is given to us, and the task is to define the dotted lifts α^i, $i \leq 0$, so as to obtain a morphism of complexes.

Now, P^0 is projective and maps to A by $\varphi \circ \pi$; this guarantees the existence of a lifting α^0 of φ. (The map $P^0 \to M^0$ to the degree-0 term of the original complex is obtained by following with $\ker d^0_{M^\bullet} \rightarrowtail M^0$.) Next, note that

$$\mu \circ \alpha^0 \circ d^{-1}_{P^\bullet} = \varphi \circ \pi \circ d^{-1}_{P^\bullet} = 0;$$

therefore, $\alpha^0 \circ d^{-1}_{P^\bullet}$ factors through $\ker \mu$. Since the bottom complex is exact, $\ker \mu = \operatorname{im} d^{-1}_{M^\bullet}$. Therefore, we have the diagram

$$\begin{array}{ccccc}
 & & P^{-1} & & \\
 & & \downarrow{\scriptstyle \alpha^0 \circ d^{-1}_{P^\bullet}} & & \\
M^{-1} & \xrightarrow{d^{-1}_{M^\bullet}} & \operatorname{im} d^{-1}_{M^\bullet} & \longrightarrow & 0
\end{array}$$

and a lift α^{-1} then exists since P^{-1} is projective.

The construction of the other α^{-i}, $i > 1$, proceeds inductively in precisely the same way, using at each step the fact that the bottom complex is exact.

This proves the existence of $\alpha^\bullet : P^\bullet \to M^\bullet$. Its uniqueness up to homotopy follows immediately from Lemma 5.11. □

Summarizing, if A contains enough projectives, then every object A of A admits a projective resolution, and further this is initial among all resolutions with target $\iota(A)$. If A has enough injectives, then every object A of A admits an injective resolution, and this is final among all resolutions with source $\iota(A)$.

Recall that initial, resp., final, objects of a category are always unique up to isomorphism (Proposition I.5.4). Here, this says the following:

Proposition 6.4. *Any two projective (resp., injective) resolutions of an object A of an abelian category A are homotopy equivalent.*

(This is also a direct consequence of Lemma 6.3.)

The moral we extract from these considerations is that if an abelian category A has enough (say) projectives, then we can associate with each object A of A an object of $\mathsf{K}^{\leq 0}(\mathsf{P})$, determined up to homotopy. This can in fact be done *functorially,* in the sense that morphisms in A can be lifted to morphisms of corresponding projective resolutions:

Proposition 6.5. *Let A_0, A_1 be objects of an abelian category A, and let $P_i^\bullet$ be a projective resolution of A_i, $i = 0, 1$. Then every morphism $\varphi : A_0 \to A_1$ in A is induced by a morphism $\alpha^\bullet : P_0^\bullet \to P_1^\bullet$, uniquely determined up to homotopy.*

Of course an analogous statement holds for injective resolutions.

Proof. By hypothesis, φ is a morphism $H^0(P_0^\bullet) \to H^0(P_1^\bullet)$. The complex $P_0^\bullet$ consists of projectives, and $P_1^\bullet$ is a resolution of A_1; therefore a lift $\alpha^\bullet$ exists and is unique up to homotopy, as an immediate consequence of Lemma 6.3. □

The assignment provided by Proposition 6.5 is clearly covariant. The bottom line is that the part of A consisting of objects with (say) projective resolutions is equivalent to a full subcategory of $\mathsf{K}^-(\mathsf{P})$. Dear reader, please formalize this assertion, for it is a key point (Exercise 6.2). If A has enough projectives, this gives the promised 'copy' of A within $\mathsf{K}^-(\mathsf{P})$.

Of course the situation is entirely analogous concerning the subcategory of A consisting of objects admitting an injective resolution and $\mathsf{K}^+(\mathsf{I})$.

6.2. From objects to complexes. Suppose that an abelian category A has enough projectives. At this point we hopefully agree (Exercise 6.2) that the homotopic category $\mathsf{K}^-(\mathsf{P})$ contains a full subcategory which is equivalent to A: the (homotopy) category of projective resolutions of objects of A. While *a priori* each object of A corresponds to many projective resolutions, these are all homotopy equivalent to one another (by Proposition 6.4), hence isomorphic in $\mathsf{K}^-(\mathsf{P})$; morphisms in A correspond precisely to homotopy classes of morphisms between projective resolutions (by Proposition 6.5), hence to morphisms in $\mathsf{K}^-(\mathsf{P})$.

In practice, this says that *we can use this subcategory of $\mathsf{K}^-(\mathsf{P})$ as a replacement for the original abelian category A.* We are very close to achieving our goal of identifying the 'essential nature' of cohomology: since homotopy equivalent complexes have the same cohomology (by Theorem 4.14) and additive functors preserve homotopy equivalence, objects of $\mathsf{K}^-(\mathsf{P})$ are ideal carriers of cohomology invariants: while applying an additive functor to objects of A in general destroys cohomological information, applying 'the same functor' to a corresponding object in $\mathsf{K}^-(\mathsf{P})$ preserves that information.

We will follow this lead in the next section; the functors Tor_i and Ext^i encountered in Chapter VIII will arise precisely in this fashion and will stand as our poster examples of *derived functors.*

Of course, exactly the same situation will occur if A has enough injectives, in which case the appropriate replacement for A is found in $\mathsf{K}^+(\mathsf{I})$. We will say a word about the general situation when projectives or injectives are not available, in §9.1.

The reader should now wonder whether the categories $\mathsf{K}^-(\mathsf{P})$, $\mathsf{K}^+(\mathsf{I})$ may in fact be used to extract cohomological information for *any* (bounded) complex, not just for resolutions of a given object of A. As observed in §4.2, resolutions (in $\mathsf{C}^-(\mathsf{A})$ or $\mathsf{C}^+(\mathsf{A})$) are just special cases of quasi-isomorphisms; the natural question is whether *every* (suitably bounded) complex has a 'quasi-isomorphic copy' in $\mathsf{K}^-(\mathsf{P})$ or $\mathsf{K}^+(\mathsf{I})$, again uniquely defined up to isomorphism in the homotopy category (that is, up to homotopy equivalence in $\mathsf{C}(\mathsf{A})$).

This is indeed the case.

Theorem 6.6. *Assume the abelian category A has enough projectives, and let $L^\bullet$ be a complex in $\mathsf{C}^-(\mathsf{A})$. Then there exists a bounded-above complex of projectives $P^\bullet$ and a quasi-isomorphism $P^\bullet \to L^\bullet$, and $P^\bullet$ is uniquely defined up to homotopy equivalence. Further, every morphism $\alpha^\bullet$ of complexes in $\mathsf{C}^-(\mathsf{A})$ lifts to a morphism of the corresponding projective resolutions in $\mathsf{K}^-(\mathsf{P})$, also uniquely determined (up to homotopy).*

The lift obtained in the last part of this statement exists in the homotopy category: any representative in $\mathsf{C}^-(\mathsf{P})$ will be a 'homotopy lift' of $\alpha^\bullet$, in the sense that the relevant diagram will only be guaranteed to commute up to homotopy.

We will call a complex $P^\bullet$ as in the statement a 'projective resolution of $L^\bullet$'; this is a slight abuse of language, since it forgets the specific quasi-isomorphism $P^\bullet \to L^\bullet$.

As always, there is a mirror statement to Theorem 6.6 for injectives, which we will leave to the reader to formulate precisely. It is a particularly good exercise for the reader to construct a stand-alone proof for the injective case.

Proof. The proof of this result is admittedly rather technical, as it involves many of the tools that we have developed.

Construction of $P^\bullet$. We may assume that $L^\bullet$ is in $\mathsf{C}^{\leq 0}(\mathsf{A})$,

$$\cdots \longrightarrow L^{-2} \xrightarrow{d_{L^\bullet}^{-2}} L^{-1} \xrightarrow{d_{L^\bullet}^{-1}} L^0 \longrightarrow 0 \longrightarrow \cdots,$$

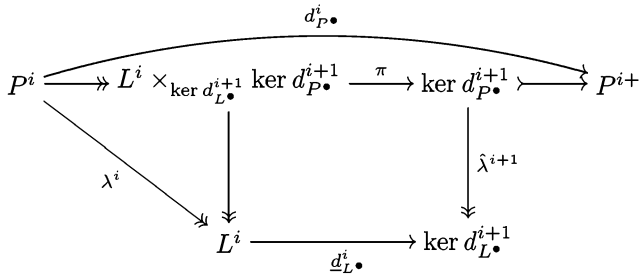
and we are seeking a complex $P^\bullet$ in $\mathsf{C}^{\leq 0}(\mathsf{P})$ and a quasi-isomorphism $\lambda^\bullet : P^\bullet \to L^\bullet$:

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & P^{-2} & \xrightarrow{d_{P^\bullet}^{-2}} & P^{-1} & \xrightarrow{d_{P^\bullet}^{-1}} & P^0 & \longrightarrow & 0 & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle \lambda^{-2}} & & \downarrow{\scriptstyle \lambda^{-1}} & & \downarrow{\scriptstyle \lambda^{0}} & & \downarrow{\scriptstyle \lambda^{1}} & & \\
\cdots & \longrightarrow & L^{-2} & \xrightarrow{d_{L^\bullet}^{-2}} & L^{-1} & \xrightarrow{d_{L^\bullet}^{-1}} & L^0 & \longrightarrow & 0 & \longrightarrow & \cdots
\end{array}$$

The construction will give us a little more: we will obtain a projective $P^\bullet$ and an *epi*morphism $\lambda^\bullet : P^\bullet \to L^\bullet$ and moreover such that each induced morphism $\hat{\lambda}^i : \ker d_{P^\bullet}^i \to \ker d_{L^\bullet}^i$ is also an epimorphism.

For $i > 0$, necessarily $\lambda^i = 0$. Arguing inductively, we may assume we have already constructed a suitable λ^{i+1} and use it to construct λ^i. Here is the diagram

summarizing this inductive step:



Here, $\underline{d}^i_{L^\bullet}$ restricts the target of $d^i_{L^\bullet}$. The square is a pull-back (cf. Examples 1.11 and 2.2). The morphism $\hat\lambda^{i+1}$ on the right is an epimorphism by induction, and it follows that the morphism on the left is also an epimorphism, by Lemma 2.3. A projective object P^i with an epimorphism to the fibered product exists because A has enough projectives; the composition $\lambda^i : P^i \to L^i$ is then also an epimorphism, as claimed. Since the square is a pull-back, the kernel of π maps isomorphically to the kernel of $\underline{d}^i_{L^\bullet}$ (Exercise 1.16); it follows easily that the new induced map $\hat\lambda^i : \ker d^i_{P^\bullet} \to \ker d^i_{L^\bullet}$ is an epimorphism, concluding the inductive step and the construction of the complex $P^\bullet$ and of the morphism $\lambda^\bullet$.

To see that the morphism induced in cohomology by $\lambda^\bullet$ is an isomorphism, look again at the pull-back diagram and note that $H^{i+1}(P^\bullet)$, resp., $H^{i+1}(L^\bullet)$, may be realized as the cokernel of π, resp., $\underline{d}^i_{L^\bullet}$. Since $\hat\lambda^{i+1}$ is an epimorphism, so is the morphism

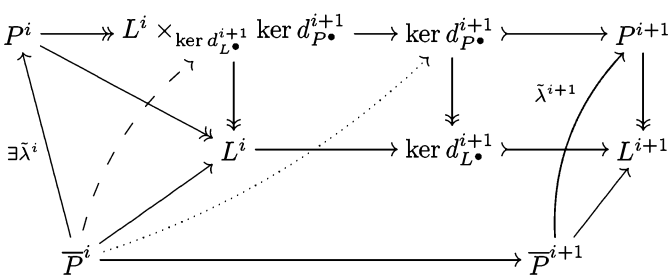
$$(\underline{d}^i_{L^\bullet}, -\hat\lambda^{i+1}) \qquad L^i \oplus \ker d^{i+1}_{P^\bullet} \longrightarrow \ker d^{i+1}_{L^\bullet}.$$

As the diagram is a pull-back, this implies (as observed in Example 2.2) that it is a push-out as well, and (Exercise 1.16 again) it follows that the induced morphism $\operatorname{coker}\pi \to \operatorname{coker}\underline{d}^i_{L^\bullet}$ is an isomorphism. As noted, this is nothing but

$$H^{i+1}(\lambda^\bullet) \qquad H^{i+1}(P^\bullet) \longrightarrow H^{i+1}(L^\bullet),$$

so we are done.

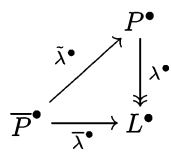
Uniqueness up to homotopy. It suffices to compare an arbitrary $\overline{P}^\bullet$ in $\mathsf{C}^{\leq 0}(\mathsf{P})$ mapping to $L^\bullet$ with the complex $P^\bullet$ constructed in the first part of the proof. Working on the same diagram used above,



we claim that the morphism $\overline{P}^\bullet \to L^\bullet$ lifts to a morphism $\overline{P}^\bullet \to P^\bullet$. Indeed, both complexes are 0 in degree $\gg 0$; thus, for $i \gg 0$, $\overline{P}^i \to P^i$ is necessarily the zero-morphism. Arguing inductively once more, we assume the lift has been constructed in all degrees $> i$, and we construct it in degree i.

The composition $\overline{P}^i \to \overline{P}^{i+1} \to P^{i+1} \to P^{i+2}$ agrees with the composition $\overline{P}^i \to \overline{P}^{i+1} \to \overline{P}^{i+2} \to P^{i+2}$, so it is the zero-morphism. Therefore this composition factors through $\ker d_{P^\bullet}^{i+1}$: this gives the dotted morphism in the diagram. By the universal property of fibered products, we obtain the dashed morphism. Since P^i maps epimorphically to $L^i \times_{\ker d_{L^\bullet}^{i+1}} \ker d_{P^\bullet}^{i+1}$, the needed lift $\tilde{\lambda}^i : \overline{P}^i \to P^i$ exists as $\overline{P}^i$ is projective. The 'outer' diagram is commutative by construction; hence the collection $\tilde{\lambda}^i$ gives a morphism of cochain complexes $\tilde{\lambda}^\bullet$. Note that so far we have not used the hypothesis that $\overline{\lambda}^\bullet$ is a quasi-isomorphism.

We now have the following commutative diagram in $\mathsf{C}^-(\mathsf{A})$:

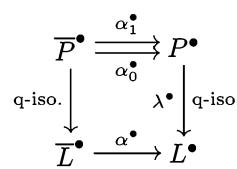


Both $\lambda^\bullet$ and $\overline{\lambda}^\bullet$ are quasi-isomorphisms; therefore so is $\tilde{\lambda}^\bullet$. Theorem 5.9 implies then that $\tilde{\lambda}^\bullet$ is a homotopy equivalence, as needed.

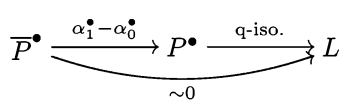
Lifting morphisms. Let $\alpha^\bullet : \overline{L}^\bullet \to L^\bullet$ be any morphism in $\mathsf{C}^-(\mathsf{A})$. The argument we have just given shows that if $\overline{P}^\bullet$ is in $\mathsf{C}^-(\mathsf{P})$ and $L^\bullet$ is in $\mathsf{C}^-(\mathsf{A})$, then *every* morphism $\overline{P}^\bullet \to L^\bullet$ lifts to a morphism $\overline{P}^\bullet \to P^\bullet$, where $P^\bullet$ is the complex constructed in the first part of the proof. In fact, we have now established that there exists a homotopy lift[28] to any complex $P^\bullet$ mapping quasi-isomorphically to $L^\bullet$, since every such complex is homotopy equivalent to the one constructed earlier.

Applying this observation to any complex $\overline{P}^\bullet$ mapping quasi-isomorphically to $\overline{L}^\bullet$ lifts $\alpha^\bullet$ as needed.

Finally, assume that both $\alpha_0^\bullet$, $\alpha_1^\bullet$ lift $\alpha^\bullet$ homotopically:



This means that $\lambda^\bullet \circ (\alpha_1^\bullet - \alpha_0^\bullet)$ is homotopic to 0:



It follows that $\alpha_1^\bullet - \alpha_0^\bullet \sim 0$, by Lemma 5.14, and this concludes the proof. □

[28] That is, the corresponding diagram will commute up to homotopy. In general, a lift in $\mathsf{C}^-(\mathsf{P})$ may not exist; cf. Exercise 6.10.

6.3. Poor man's derived category. We are now in a position to close a relatively large circle of ideas. We claim that if the abelian category A has enough projectives, resp., injectives, then $\mathsf{K}^-(\mathsf{P})$, resp., $\mathsf{K}^+(\mathsf{I})$, is a solution 'up to isomorphisms' of the universal problem presented in §4.2. The precise statement is the following.

By Proposition 5.4, the functor $\mathsf{C}(\mathsf{A}) \to \mathsf{D}(\mathsf{A})$ factors through the homotopic category; this of course holds for the bounded versions of these categories as well. Thus, we have functors $\mathsf{K}^-(\mathsf{A}) \to \mathsf{D}^-(\mathsf{A})$ and $\mathsf{K}^+(\mathsf{A}) \to \mathsf{D}^+(\mathsf{A})$. By restriction, we obtain functors

$$\mathsf{K}^-(\mathsf{P}) \to \mathsf{D}^-(\mathsf{A}), \quad \mathsf{K}^+(\mathsf{I}) \to \mathsf{D}^+(\mathsf{A}).$$

Theorem 6.7. *Let A be an abelian category with enough projectives. Then the functor $\mathsf{K}^-(\mathsf{P}) \to \mathsf{D}^-(\mathsf{A})$ is an equivalence of categories.*

If A has enough injectives, then the functor $\mathsf{K}^+(\mathsf{I}) \to \mathsf{D}^+(\mathsf{A})$ is an equivalence of categories.

Theorem 6.7 is proven in full detail in any more complete treatment of homological algebra. Since we have not actually constructed $\mathsf{D}^-(\mathsf{A})$, we cannot really prove this statement here; but we now know enough to appreciate why Theorem 6.7 should be true, in the sense that $\mathsf{K}^-(\mathsf{P})$, $\mathsf{K}^+(\mathsf{I})$ satisfy the expected universal properties. We will deal with the projective side of the story, since the injective side mirrors it faithfully.

Assume that A has enough projectives. Define a functor $\mathscr{P} : \mathsf{C}^-(\mathsf{A}) \to \mathsf{K}^-(\mathsf{P})$ by associating with every bounded-above complex $L^\bullet$ any projective resolution $\mathscr{P}(L^\bullet)$ of $L^\bullet$ (which exists by Theorem 6.6) and with every morphism $\alpha^\bullet : L^\bullet \to M^\bullet$ in $\mathsf{C}^-(\mathsf{A})$ the morphism $\mathscr{P}(\alpha^\bullet)$ in $\mathsf{K}^-(\mathsf{P})$ lifting $\alpha^\bullet$.

The morphism $\mathscr{P}(\alpha^\bullet)$ is the homotopy class of a homotopy lift of $\alpha^\bullet$; we have proved that any two such lifts are homotopy equivalent. It also follows that $\mathscr{P}$ is indeed a functor: for example, if $\alpha^\bullet : L^\bullet \to M^\bullet$ and $\beta^\bullet : M^\bullet \to N^\bullet$ are morphisms in $\mathsf{C}^-(\mathsf{A})$, then both $\mathscr{P}(\beta^\bullet) \circ \mathscr{P}(\alpha^\bullet)$ and $\mathscr{P}(\beta^\bullet \circ \alpha^\bullet)$ are classes of homotopy lifts of $\beta^\bullet \circ \alpha^\bullet$; therefore $\mathscr{P}(\beta^\bullet) \circ \mathscr{P}(\alpha^\bullet) = \mathscr{P}(\beta^\bullet \circ \alpha^\bullet)$ by uniqueness, as needed.

Note that there is a very large element of arbitrariness in the choice of $\mathscr{P}$: we are not prescribing any particular recipe for choosing a projective resolution rather than another. But all such resolutions are homotopically equivalent, as proven in Theorem 6.6; therefore different choices would only move the resolutions in their isomorphism class in $\mathsf{K}^-(\mathsf{P})$.

Remark 6.8. There is an efficient way to clarify (and defuse) this arbitrariness. Suppose $\mathscr{P}$, $\mathscr{P}'$ are two choices for the resolution functor $\mathsf{C}^-(\mathsf{A}) \to \mathsf{K}^-(\mathsf{P})$. Thus, we have chosen *two* projective resolutions $\mathscr{P}(L^\bullet)$, $\mathscr{P}'(L^\bullet)$ for each object $L^\bullet$ in $\mathsf{C}^-(\mathsf{A})$. By Theorem 6.6, there is a homotopy equivalence between $\mathscr{P}(L^\bullet)$ and $\mathscr{P}'(L^\bullet)$, that is, an isomorphism $\rho_{L^\bullet} : \mathscr{P}(L^\bullet) \to \mathscr{P}'(L^\bullet)$ in $\mathsf{K}^-(\mathsf{P})$; this isomorphism is unique (since it lifts the identity on $L^\bullet$ and lifts are unique up to homotopy). Further, for every morphism of cochain complexes $\alpha^\bullet : L_0^\bullet \to L_1^\bullet$, the

diagram

$$\begin{array}{ccc} \mathscr{P}(L_0^\bullet) & \xrightarrow{\mathscr{P}(\alpha^\bullet)} & \mathscr{P}(L_1^\bullet) \\ \downarrow{\scriptstyle \rho_{L_0^\bullet}} & & \downarrow{\scriptstyle \rho_{L_1^\bullet}} \\ \mathscr{P}'(L_0^\bullet) & \xrightarrow{\mathscr{P}'(\alpha^\bullet)} & \mathscr{P}'(L_1^\bullet) \end{array}$$

commutes: this is again by the uniqueness of lifts up to homotopy, since $\rho_{L_1^\bullet} \circ \mathscr{P}(\alpha^\bullet)$ and $\mathscr{P}'(\alpha^\bullet) \circ \rho_{L_0^\bullet}$ both correspond to lifts of $\alpha^\bullet$ to $\mathscr{P}(L_0^\bullet) \to \mathscr{P}'(L_1^\bullet)$.

This says that *there is a unique natural isomorphism*[29] between any two choices $\mathscr{P}$, $\mathscr{P}'$ of functors of projective resolutions. In a categorical context, this is essentially as good as uniqueness. Thus the arbitrariness in the choice of $\mathscr{P}$ is less dramatic than it may seem at first. ⌟

Choose any functor $\mathscr{P}$ as above. We claim that $\mathscr{P} : \mathsf{C}^-(\mathsf{A}) \to \mathsf{K}^-(\mathsf{P})$ is essentially a solution to the universal problem defining the derived category $\mathsf{D}^-(\mathsf{A})$. Here is a more precise statement:

Theorem 6.9. *With notation as above, we have the following.*

- *If $\rho^\bullet$ is a quasi-isomorphism in $\mathsf{C}^-(\mathsf{A})$, then $\mathscr{P}(\rho^\bullet)$ is an isomorphism in $\mathsf{K}^-(\mathsf{P})$.*
- *Let $\mathscr{F} : \mathsf{C}^-(\mathsf{A}) \longrightarrow \mathsf{D}$ be an additive functor such that $\mathscr{F}(\rho^\bullet)$ is an isomorphism for every quasi-isomorphism $\rho^\bullet$. Then there exists a functor $\widetilde{\mathscr{F}} : \mathsf{K}^-(\mathsf{P}) \to \mathsf{D}$, such that the diagram*

$$\begin{array}{ccc} \mathsf{C}^-(\mathsf{A}) & \xrightarrow{\mathscr{F}} & \mathsf{D} \\ {\scriptstyle \mathscr{P}}\downarrow & \nearrow{\scriptstyle \widetilde{\mathscr{F}}} & \\ \mathsf{K}^-(\mathsf{P}) & & \end{array}$$

commutes up to natural isomorphism. (That is, there is a natural transformation $\nu : \widetilde{\mathscr{F}} \circ \mathscr{P} \rightsquigarrow \mathscr{F}$ such that

$$\nu_{L^\bullet} : \widetilde{\mathscr{F}} \circ \mathscr{P}(L^\bullet) \xrightarrow{\sim} \mathscr{F}(L^\bullet)$$

is an isomorphism for every complex $L^\bullet$ in $\mathsf{C}^-(\mathsf{A})$.) The functor $\widetilde{\mathscr{F}}$ is unique up to natural isomorphism.

The fact that the diagram does not commute on the nose should not be too disturbing. It corresponds to the fact (stated in Theorem 6.7) that there is 'only' an equivalence of categories between $\mathsf{K}^-(\mathsf{P})$ and $\mathsf{D}^-(\mathsf{A})$, not a hard and fast 'isomorphism'. As we already pointed out in §VIII.1.3, this is what should be expected in a categorical context.

[29] 'Natural transformations' were rapidly defined in §VIII.1.5, Definition VIII.1.15. They are the most sensible notion of a morphism between functors. A natural *isomorphism* is a natural transformation that is an isomorphism on each object.

Proof. For the first point, let $\rho^\bullet : L^\bullet \to M^\bullet$ be any morphism in $\mathsf{C}^-(\mathsf{A})$. By Theorem 6.6, $\rho^\bullet$ lifts to a morphism of resolutions: we have a diagram

$$\begin{array}{ccc} \mathscr{P}(L^\bullet) & \xrightarrow{\rho'^\bullet} & \mathscr{P}(M^\bullet) \\ {\scriptstyle \mu_{L^\bullet}}\downarrow & & \downarrow{\scriptstyle \mu_{M^\bullet}} \\ L^\bullet & \xrightarrow{\rho^\bullet} & M^\bullet \end{array}$$

which commutes up to homotopy; here $\mu_{L^\bullet}$, $\mu_{M^\bullet}$ are quasi-isomorphisms, and $\mathscr{P}(\rho^\bullet)$ is the homotopy class of $\rho'^\bullet$. If $\rho^\bullet$ is a quasi-isomorphism, then so is $\rho'^\bullet$ (since the diagram commutes up to homotopy; hence it commutes after taking cohomology); it follows that $\mathscr{P}(\rho^\bullet)$ is an isomorphism in this case, by Corollary 5.10.

To prove the factorization property, define $\widetilde{\mathscr{F}} : \mathsf{K}^-(\mathsf{P}) \to \mathsf{D}$ by setting

$$\widetilde{\mathscr{F}}(P^\bullet) := \mathscr{F}(P^\bullet)$$

for any bounded-above complex $P^\bullet$ of projectives, and define $\widetilde{\mathscr{F}}$ of a morphism in $\mathsf{K}^-(\mathsf{P})$ to be $\mathscr{F}(\alpha^\bullet)$, for any representative $\alpha^\bullet$ of that morphism.

The fact that the action of $\widetilde{\mathscr{F}}$ on morphisms is well-defined is precisely the content of Lemma 5.1: homotopic morphisms have the same image in D. It is immediate that $\widetilde{\mathscr{F}}$ is a functor, and we have to check the statement about commutativity up to isomorphism of the diagram and the uniqueness of $\widetilde{\mathscr{F}}$ up to natural isomorphism.

Then let $L^\bullet$ be a complex in $\mathsf{C}^-(\mathsf{A})$. By construction there is a quasi-isomorphism $\mu_{L^\bullet} : \mathscr{P}(L^\bullet) \to L^\bullet$. Since $\mathscr{F}$ maps quasi-isomorphisms to isomorphisms, the induced morphism

$$\nu_{L^\bullet} : \widetilde{\mathscr{F}}(\mathscr{P}(L^\bullet)) = \mathscr{F}(\mathscr{P}(L^\bullet)) \longrightarrow \mathscr{F}(L^\bullet)$$

is an isomorphism. We just have to verify that the isomorphisms $\nu_{L^\bullet}$ define a natural transformation: that is, that for all morphisms $\alpha^\bullet : L^\bullet \to M^\bullet$ in $\mathsf{C}^-(\mathsf{A})$ the diagram

$$\begin{array}{ccc} \widetilde{\mathscr{F}}(\mathscr{P}(L^\bullet)) & \xrightarrow{\widetilde{\mathscr{F}}(\mathscr{P}(\alpha^\bullet))} & \widetilde{\mathscr{F}}(\mathscr{P}(M^\bullet)) \\ {\scriptstyle \nu_{L^\bullet}}\downarrow & & \downarrow{\scriptstyle \nu_{M^\bullet}} \\ \mathscr{F}(L^\bullet) & \xrightarrow{\mathscr{F}(\alpha^\bullet)} & \mathscr{F}(M^\bullet) \end{array}$$

commutes. This diagram (in D) is obtained by applying $\mathscr{F}$ to the diagram (in $\mathsf{C}(\mathsf{A})$)

$$\begin{array}{ccc} \mathscr{P}(L^\bullet) & \xrightarrow{\alpha'^\bullet} & \mathscr{P}(M^\bullet) \\ {\scriptstyle \mu_{L^\bullet}}\downarrow & & \downarrow{\scriptstyle \mu_{M^\bullet}} \\ L^\bullet & \xrightarrow{\alpha^\bullet} & M^\bullet \end{array}$$

where $\alpha'^\bullet$ is a homotopy lift of $\alpha^\bullet$. This diagram commutes up to homotopy, so the previous one commutes on the nose, again by Lemma 5.1.

To see that $\widetilde{\mathscr{F}}$ is unique up to isomorphism, let $\overline{\mathscr{F}} : \mathsf{K}^-(\mathsf{P}) \to \mathsf{D}$ be another functor satisfying the same property. Thus there is a natural isomorphism $\overline{\nu} : \overline{\mathscr{F}} \circ \mathscr{P} \rightsquigarrow \mathscr{F}$. For $P^\bullet$ a complex in $\mathsf{K}^-(\mathsf{P})$, consider the composition $\rho_{P^\bullet}$:

$$\overline{\mathscr{F}}(P^\bullet) \xrightarrow[\overline{\mathscr{F}}(\mu_{P^\bullet})^{-1}]{\sim} \overline{\mathscr{F}}(\mathscr{P}(P^\bullet)) \xrightarrow[\overline{\nu}_{P^\bullet}]{\sim} \mathscr{F}(P^\bullet) = \widetilde{\mathscr{F}}(P^\bullet)$$

(keep in mind that since $\mu_{P^\bullet} : \mathscr{P}(P^\bullet) \to P^\bullet$ is a quasi-isomorphism, $\mathscr{F}(\mu_{P^\bullet})$ is invertible in D). The reader will verify (Exercise 6.12) that this defines $\rho : \overline{\mathscr{F}} \to \widetilde{\mathscr{F}}$ as a natural isomorphism, concluding the proof. □

If A has enough injectives, we can define a functor $\mathscr{Q} : \mathsf{C}^+(\mathsf{A}) \to \mathsf{K}^+(\mathsf{I})$ by choosing an *injective* resolution for every bounded-below complex. The situation concerning this functor is precisely analogous to the situation just reviewed for the functor $\mathscr{P}$, modulo reversing arrows as appropriate.

Summary. If an abelian category A has enough projectives (resp., enough injectives), then the homotopic category $\mathsf{K}^-(\mathsf{P})$ (resp., $\mathsf{K}^+(\mathsf{I})$) works as a replacement of the derived category $\mathsf{D}^-(\mathsf{A})$ (resp., $\mathsf{D}^+(\mathsf{A})$). As we have argued earlier, viewing a complex in these categories is the most natural way to extract the maximum amount of 'cohomological' information from the complex. The category A itself (or a category equivalent to A) sits as a full subcategory of these categories.

To reiterate the key point, there is a clear advantage in stopping short of taking cohomology. Resolving a complex $L^\bullet$ by a complex $P^\bullet$ of (say) projectives, thanks to Theorem 6.6, gives us an entity which certainly captures the cohomological information of $L^\bullet$ but that can still be manipulated with the entire range of tools available in dealing with complexes. For example, we can apply to $P^\bullet$ any additive functor and still obtain a complex whose cohomology only depends on the original complex $L^\bullet$ (by Theorem 4.14). The cohomology of $L^\bullet$ is only one example of the cohomological invariants that may be extracted from $L^\bullet$; viewing $L^\bullet$ in the derived category gives direct access to all these invariants.

We will capitalize on these considerations in the next section.

Exercises

6.1. ▷ Prove that the full subcategory of $\mathsf{K}(\mathsf{A})$ consisting of complexes concentrated in degree 0 is a copy of the abelian category A. [§6.1]

6.2. ▷ Assuming that A has enough projectives, prove that the full subcategory $\hat{\mathsf{A}}$ of $\mathsf{K}^-(\mathsf{P})$ consisting of complexes with cohomology concentrated in degree 0 is equivalent to A. (Prove that H^0 is fully faithful on this subcategory.) [§6.1, §6.2, 6.13, §7.1]

6.3. ▷ Let A be an abelian category, and let $P^\bullet$ be a complex that is projective *as an object of* $\mathsf{C}(\mathsf{A})$.

- Prove that each P^i is projective in A.

- Prove that there is a morphism of complexes $P^\bullet \to MC(\mathrm{id}_{P^\bullet})[-1]$, such that for each i the corresponding morphism $P^i \to P^i \oplus P^{i-1}$ is of the form (id, h^i).
- Prove that the collection h^i gives a homotopy between $\mathrm{id}_{P^\bullet}$ and 0.
- Conclude that $P^\bullet$ is a split exact complex of projectives.

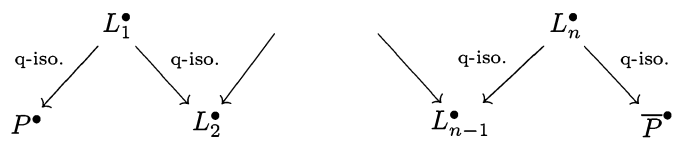
(For example, the complex $\cdots \to 0 \to \mathbb{Z} \xrightarrow{\cdot 2} \mathbb{Z} \to 0 \to \cdots$ is not projective in $\mathsf{C}(\mathsf{Ab})$.)

- Conversely, prove that if $P^\bullet$ is a split exact complex of projectives, then $P^\bullet$ is projective in $\mathsf{C}(\mathsf{A})$. (Hint: Use Exercises 5.11 and 5.4 to reduce to the case $\cdots \to 0 \to P \xrightarrow{\sim} P \to 0 \to \cdots$.)
- If $P^\bullet$ is a bounded-above complex of projectives, prove that $P^\bullet$ is projective as an object of $\mathsf{C}(\mathsf{A})$ if and only if it is exact.

[§6.1]

6.4. ▷ Let A be an abelian category with enough projectives, and let $\mathscr{P} : \mathsf{C}^-(\mathsf{A}) \to \mathsf{K}^-(\mathsf{P})$ be a functor choosing a projective resolution for every bounded-above complex, as in §6.3. Prove that $\mathscr{P}$ induces a functor $\mathsf{K}^-(\mathsf{A}) \to \mathsf{K}^-(\mathsf{P})$, which we will also denote by $\mathscr{P}$. (Use Lemma 5.14.) Next, let $\mathscr{I}$ be the embedding $\mathsf{K}^-(\mathsf{P}) \to \mathsf{K}^-(\mathsf{A})$. Prove that $\mathscr{P} \circ \mathscr{I}$ is naturally *isomorphic* to the identity functor and that there is a natural *transformation* (not an isomorphism in general) $\mathscr{I} \circ \mathscr{P} \rightsquigarrow \mathrm{id}_{\mathsf{K}^-(\mathsf{A})}$. [§7.1, §7.2]

6.5. (Cf. Exercise 5.14.) Let A be an abelian category with enough projectives, and suppose there is a chain of quasi-isomorphisms



in $\mathsf{C}^-(\mathsf{A})$, with $P^\bullet$ and $\overline{P}^\bullet$ in $\mathsf{C}^-(\mathsf{P})$. Prove that $P^\bullet$ and $\overline{P}^\bullet$ are homotopy equivalent.

Formulate and prove a statement in which the morphisms point *to* (rather than *from*) $L_1^\bullet, L_3^\bullet, \ldots, L_n^\bullet$.

6.6. Let A be an abelian category, and assume for simplicity that A has enough projectives. Let $\mu^\bullet : M^\bullet \to N^\bullet$ be a morphism in $\mathsf{C}^-(\mathsf{A})$. Prove that $\mu^\bullet$ induces the zero-morphism in $\mathsf{D}^-(\mathsf{A})$ if and only if there exists a quasi-isomorphism $\lambda^\bullet : L^\bullet \to M^\bullet$ in $\mathsf{C}^-(\mathsf{A})$ such that $\mu^\bullet \circ \lambda^\bullet$ is homotopic to zero.

6.7. For any integer $m > 1$, view the short exact sequence

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\cdot m} \mathbb{Z} \longrightarrow \mathbb{Z}/m\mathbb{Z} \longrightarrow 0$$

as a complex in $\mathsf{C}(\mathsf{Ab})$. Prove that the identity morphism on this complex is not homotopic to zero, but it induces the zero-morphism in the derived category $D^-(\mathsf{Ab})$.

6.8. For any integer $m > 1$, view the short exact sequence

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\cdot m} \mathbb{Z} \longrightarrow \mathbb{Z}/m\mathbb{Z} \longrightarrow 0$$

as a complex in $\mathsf{C}(\mathsf{Ab})$. Find explicitly a projective resolution of this complex produced by the argument at the beginning of the proof of Theorem 6.6.

6.9. ¬ Show that the morphism induced in $\mathsf{D}^-(\mathsf{Ab})$ by the cochain map

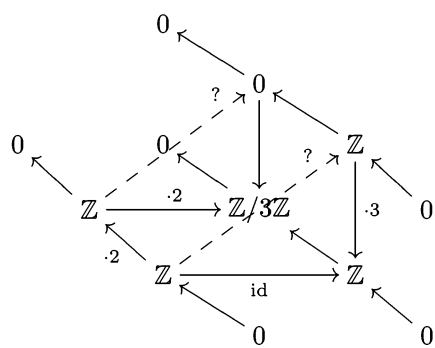
$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathbb{Z} & \xrightarrow{\cdot 2} & \mathbb{Z} & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \cdot 2} & & \downarrow & & \\ 0 & \longrightarrow & \mathbb{Z} & \longrightarrow & \mathbb{Z}/3\mathbb{Z} & \longrightarrow & 0 \end{array}$$

is the zero-morphism, while the morphism induced by

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathbb{Z} & \xrightarrow{\cdot 2} & \mathbb{Z} & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \mathrm{id}} & & \downarrow{\scriptstyle \cdot 2} & & \\ 0 & \longrightarrow & \mathbb{Z} & \longrightarrow & \mathbb{Z}/3\mathbb{Z} & \longrightarrow & 0 \end{array}$$

is not the zero-morphism. Note that both morphisms induce 0 in cohomology. Thus, nonzero morphisms in the derived category may induce the zero-morphism in cohomology. (This is another indication that the derived category carries more information than 'just' cohomology.) [6.10, 6.11]

6.10. ▷ Consider the second morphism given in Exercise 6.9, together with a projective resolution of the bottom complex:



Prove that a 'true' lift as indicated by the dashed arrows does not exist but that one exists up to homotopy (as prescribed by Theorem 6.6). [§6.2]

6.11. Let R be an integral domain, and let r be a nonzero, nonunit element of R. Construct a nonzero morphism in $\mathsf{D}^-(R\text{-}\mathsf{Mod})$ between the complexes

$$\begin{array}{ccccccccccc} \cdots & \longrightarrow & 0 & \longrightarrow & 0 & \longrightarrow & R/(r) & \longrightarrow & 0 & \longrightarrow & \cdots \\ & & \vdots & & \vdots & & \vdots & & \vdots & & \\ \cdots & \longrightarrow & 0 & \longrightarrow & R & \longrightarrow & 0 & \longrightarrow & 0 & \longrightarrow & \cdots \end{array}$$

matching degrees as indicated. (This will be particularly easy if you have worked out Exercise 6.9).

6.12. ▷ Complete the proof that the functor $\widetilde{\mathscr{F}}$ in Theorem 6.9 is unique up to natural isomorphism. [§6.3]

6.13. ▷ (Cf. Exercise 6.2.) Let A be an abelian category with enough projectives, and let $\hat{\mathsf{A}}$ be the subcategory of $\mathsf{K}^-(\mathsf{P})$ whose objects have cohomology concentrated in degree 0. Choosing (arbitrarily) a projective resolution for every object A of A and lifting morphisms as in Proposition 6.5, we obtain a functor $\mathsf{A} \to \hat{\mathsf{A}}$; with notation as in §6.3, this is $\mathscr{P} \circ \iota$. It is clear that $H^0 \circ \mathscr{P} \circ \iota$ is the identity functor on A. Prove that there is a natural transformation from the identity functor on $\hat{\mathsf{A}}$ to $\mathscr{P} \circ \iota \circ H^0$. [§7.1]

---

7. Derived functors

We will now adopt the blanket assumption that A is an abelian category with enough projectives, so that—as explained in §6.3—the homotopic category $\mathsf{K}^-(\mathsf{P})$ is a concrete realization (up to equivalence of categories) of the derived category of the bounded-above complexes in A. Of course everything would go through, with appropriate changes, for bounded-below complexes in the presence of enough injectives. These assumptions are not necessary in order to develop this material, but they simplify it substantially.

7.1. Viewpoint shift. After deriving *categories,* it should not seem too far-fetched to derive *functors.*

We have seen (Exercises 6.2 and 6.13) that, if A has enough projectives, then the full subcategory $\hat{\mathsf{A}}$ of $\mathsf{K}^-(\mathsf{P})$ whose objects are complexes with cohomology concentrated in degree 0 is equivalent to A: H^0 is a fully faithful, surjective functor $\hat{\mathsf{A}} \to \mathsf{A}$. With notation as in §6.3, the functor $\mathscr{P} \circ \iota$ associating with every object of A a projective resolution gives a 'weak' inverse to H^0. As pointed out in §6.3, there are many different possible $\mathscr{P}$; each of them will determine a copy $\mathscr{P} \circ \iota(\mathsf{A})$ of A in $\mathsf{K}^-(\mathsf{P})$.

It may be argued that any such realization of (a category equivalent to) A is 'better' than A itself. Indeed, we have repeatedly stressed that all the cohomological information carried by a complex is ideally captured by viewing that complex in the derived category, and for the complex $\iota(M)$ determined in the simplest way by an object M of A, this simply means replacing $\iota(M)$ with any projective resolution of M, viewed as an object of $\mathsf{K}^-(\mathsf{P})$.

This, however, raises a question: if $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ is a functor and we are serious about replacing A with its counterpart(s) in $\mathsf{D}^-(\mathsf{A})$, then there should be a way to 'reinterpret' $\mathscr{F}$ in this new context: induce in some natural way a 'derived functor' $\mathsf{D}^-(\mathsf{A}) \to \mathsf{D}^-(\mathsf{B})$. One would hope that this functor should carry at least as much information as $\mathscr{F}$ and satisfy better properties.

Now we claim that there indeed *is* an evident way to induce a functor between the realizations of the derived categories in terms of $\mathsf{K}^-(\mathsf{P})$. It is a little less clear what question this functor answers—that is, what kind of universal property it satisfies. The reader is invited to sort all of this out before we do it.

Say that $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ is an additive functor between two abelian categories. It is clear that $\mathscr{F}$ extends to a functor

$$\mathsf{C}(\mathscr{F}) \quad : \quad \mathsf{C}(\mathsf{A}) \to \mathsf{C}(\mathsf{B})$$

(we briefly encountered this functor in §4): the complex $L^\bullet$ is sent to the complex $\mathsf{C}(\mathscr{F})(L^\bullet)$ whose objects are $\mathscr{F}(L^i)$ and whose differentials are $\mathscr{F}(d^i_{L^\bullet})$. Note that the functoriality and additivity of $\mathscr{F}$ imply that

$$\mathscr{F}(d^{i+1}_{L^\bullet}) \circ \mathscr{F}(d^i_{L^\bullet}) = \mathscr{F}(d^{i+1}_{L^\bullet} \circ d^i_{L^\bullet}) = \mathscr{F}(0) = 0,$$

so that $\mathsf{C}(\mathscr{F})(L^\bullet)$ is indeed a complex. It is equally clear that if $\alpha^\bullet$ is a morphism of cochain complexes in $\mathsf{C}(\mathsf{A})$, then the collection $\mathscr{F}(\alpha^i)$ defines a morphism in $\mathsf{C}(\mathsf{B})$, and that this assignment is functorial. Now, we already checked (Lemma 4.13) that homotopic morphisms in $\mathsf{C}(\mathsf{A})$ are sent to homotopic morphisms in $\mathsf{C}(\mathsf{B})$; it follows that $\mathscr{F}$ induces a functor $\mathsf{K}(\mathscr{F}) : \mathsf{K}(\mathsf{A}) \to \mathsf{K}(\mathsf{B})$. The diagram

$$\begin{array}{ccc} \mathsf{A} & \xrightarrow{\mathscr{F}} & \mathsf{B} \\ {\scriptstyle\iota}\downarrow & & \downarrow{\scriptstyle\iota} \\ \mathsf{K}(\mathsf{A}) & \xrightarrow[\mathsf{K}(\mathscr{F})]{} & \mathsf{K}(\mathsf{B}) \end{array}$$

commutes. Also, this operation clearly preserves boundedness conditions.

If now it were the case that additive functors send projective objects to projective objects, then we could restrict $\mathsf{K}(\mathscr{F})$ to $\mathsf{K}^-(\mathsf{P}(\mathsf{A}))$ and land in $\mathsf{K}^-(\mathsf{P}(\mathsf{B}))$ (where $\mathsf{P}(\mathsf{A})$ and $\mathsf{P}(\mathsf{B})$ denote the classes of projective objects in A, resp., B); given our discussion in §6.3, this would give us a natural candidate for the 'derived functor' of $\mathscr{F}$.

However, additive functors do not preserve projectivity or injectivity in general. Why should they? For example, 'tensor' does not preserve projectives[30] (Exercise 7.1). Assuming that B has enough projectives, the natural way to 'fix' this problem is to apply a corresponding functor $\mathscr{P}_\mathsf{B}$ constructed as in §6.3, associating with each complex a projective resolution. By Lemma 5.14, $\mathscr{P}_\mathsf{B}$ sends homotopic morphisms to homotopic morphisms; hence $\mathscr{P}_\mathsf{B}$ also descends to a functor defined on the homotopic category (cf. Exercise 6.4).

Definition 7.1. The *left-derived functor* of $\mathscr{F}$ is the composition

$$\mathsf{K}^-(\mathsf{P}(\mathsf{A})) \xrightarrow{\mathscr{I}_\mathsf{A}} \mathsf{K}^-(\mathsf{A}) \xrightarrow{\mathsf{K}(\mathscr{F})} \mathsf{K}^-(\mathsf{B}) \xrightarrow{\mathscr{P}_\mathsf{B}} \mathsf{K}^-(\mathsf{P}(\mathsf{B}))$$
$$\mathsf{K}^-(\mathsf{P}(\mathsf{A})) \xrightarrow{\qquad\qquad\mathsf{L}\mathscr{F}\qquad\qquad} \mathsf{K}^-(\mathsf{P}(\mathsf{B}))$$

where $\mathscr{I}_\mathsf{A}$ is the 'inclusion' of $\mathsf{K}^-(\mathsf{P}(\mathsf{A}))$ as a full subcategory of $\mathsf{K}^-(\mathsf{A})$. ⌟

It would seem sensible to denote this functor as $\mathsf{D}^-(\mathscr{F})$; but the notation '$\mathsf{L}\mathscr{F}$' is well established, so we will adopt it.

Of course a parallel discussion can be carried out if A, B have enough *injectives*, and it leads to a 'right-derived functor' $\mathsf{D}^+(\mathscr{F})$, pardon me, $\mathsf{R}\mathscr{F}$.

[30] Note that tensor is right-exact, and it is so because it is left-adjoint to another functor; see Corollary VIII.2.5. Functors that are left-adjoints to *right-exact* functors do preserve projectives: Exercise 5.6.

7.2. Universal property of the derived functor. The definition we have given for the derived functor depends on the chosen resolution functor $\mathscr{P}_{\mathsf{B}}$, so that $\mathsf{L}\mathscr{F} : \mathsf{K}^-(\mathsf{P}(\mathsf{A})) \to \mathsf{K}^-(\mathsf{P}(\mathsf{B}))$ is really only defined up to a natural isomorphism. In any case, the derived functor should be thought of as the natural counterpart of $\mathscr{F}$ at the level of the derived categories of A and B: up to equivalences of categories, $\mathsf{L}\mathscr{F}$ acts as $\mathsf{D}^-(\mathsf{A}) \to \mathsf{D}^-(\mathsf{B})$.

Still, it is not too clear exactly what question this definition answers. For example, the derived functor does *not* fit into a commutative diagram analogous to the one displayed above:

$$\begin{array}{ccc} \hat{\mathsf{A}} & \overset{??}{\dashrightarrow} & \hat{\mathsf{B}} \\ \downarrow & & \downarrow \\ \mathsf{K}^-(\mathsf{P}(\mathsf{A})) & \xrightarrow{\mathsf{L}\mathscr{F}} & \mathsf{K}^-(\mathsf{P}(\mathsf{B})) \end{array}$$

Here $\hat{\mathsf{A}}$ and $\hat{\mathsf{B}}$ are the replacements for A, B in the corresponding derived categories, discussed at the beginning of §7.1, that is, the subcategories whose objects have cohomology concentrated in degree 0. The point is that there is no reason why applying $\mathsf{L}\mathscr{F}$ to a complex whose cohomology is concentrated in degree 0 should yield a complex with the same property, unless $\mathscr{F}$ is very special to begin with.

This may be viewed as a nuisance. On the contrary, it is one of the main values of deriving categories and functors. Recasting an additive functors $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ at the level of derived categories, one gets access to interesting new invariants *even* when starting from (the equivalent copy in the derived category of) A itself.

Exactness plays an important role in these considerations:

Example 7.2. Suppose that $\mathscr{F}$ is (additive and) exact. Then we claim that $\hat{\mathsf{A}}$ *is* sent to $\hat{\mathsf{B}}$ by $\mathsf{L}\mathscr{F}$.

Indeed, let $P^\bullet$ be a complex in $\hat{\mathsf{A}}$: $H^i(P^\bullet) = 0$ for $i \neq 0$. The image $\mathsf{L}\mathscr{F}(P^\bullet)$ is obtained by choosing a projective resolution $P^\bullet_{\mathscr{F}(P^\bullet)}$ of $\mathscr{F}(P^\bullet)$. Since quasi-isomorphisms preserve cohomology and exact functors commute with cohomology (Exercise 3.7),

$$H^i(\mathsf{L}\mathscr{F}(P^\bullet)) = H^i(P^\bullet_{\mathscr{F}(P^\bullet)}) \cong H^i(\mathscr{F}(P^\bullet)) \cong \mathscr{F}(H^i(P^\bullet)) = 0$$

for $i \neq 0$. Therefore $\mathsf{L}\mathscr{F}(P^\bullet)$ is an object of $\hat{\mathsf{B}}$.

We also see that $H^0(\mathsf{L}\mathscr{F}(P^\bullet)) \cong \mathscr{F}(H^0(P^\bullet))$ in this case: this says that the restriction of $\mathsf{L}\mathscr{F}$ to $\hat{\mathsf{A}} \to \hat{\mathsf{B}}$ agrees with $\mathscr{F}$ if $\mathscr{F}$ is exact, modulo the equivalences of categories between A, resp., B, and $\hat{\mathsf{A}}$, resp., $\hat{\mathsf{B}}$. ⌟

The moral of Example 7.2 is that deriving *exact* functors is relatively straightforward, and we cannot expect to learn anything new about an exact functor by deriving it. We will mostly be interested in deriving functors that are not exact on the nose but preserve a certain amount of exactness.

Another diagram that may look promising is

$$\begin{array}{ccc} \mathsf{K}^-(\mathsf{A}) & \xrightarrow{\mathsf{K}(\mathscr{F})} & \mathsf{K}^-(\mathsf{B}) \\ {\scriptstyle \mathscr{P}_\mathsf{A}}\downarrow & ?? & \downarrow{\scriptstyle \mathscr{P}_\mathsf{B}} \\ \mathsf{K}^-(\mathsf{P}(\mathsf{A})) & \xrightarrow[\mathsf{L}\mathscr{F}]{} & \mathsf{K}^-(\mathsf{P}(\mathsf{B})) \end{array}$$

This also should not be expected to commute. Note that since most of the functors appearing here are only defined up to natural isomorphism, the best we could hope for is that this diagram commutes up to natural isomorphism: that is, conceivably there could be a natural isomorphism

$$\mathsf{L}\mathscr{F} \circ \mathscr{P}_\mathsf{A} \xrightarrow{\sim} \mathscr{P}_\mathsf{B} \circ \mathsf{K}(\mathscr{F}).$$

Well, in general there isn't (Exercise 7.3). What is $\mathsf{L}\mathscr{F}$ good for, then?

Proposition 7.3. *The left-derived functor $\mathsf{L}\mathscr{F}$ satisfies the following universal property:*

- *There is a natural transformation*
$$\mathsf{L}\mathscr{F} \circ \mathscr{P}_\mathsf{A} \rightsquigarrow \mathscr{P}_\mathsf{B} \circ \mathsf{K}(\mathscr{F});$$
- *for every functor $\mathscr{G} : \mathsf{K}^-(\mathsf{P}(\mathsf{A})) \to \mathsf{K}^-(\mathsf{P}(\mathsf{B}))$ and every natural transformation $\gamma : \mathscr{G} \circ \mathscr{P}_\mathsf{A} \rightsquigarrow \mathscr{P}_\mathsf{B} \circ \mathsf{K}(\mathscr{F})$, there is a unique (up to natural isomorphism) natural transformation $\mathscr{G} \rightsquigarrow \mathsf{L}\mathscr{F}$ inducing a factorization of γ: $\mathscr{G} \circ \mathscr{P}_\mathsf{A} \rightsquigarrow \mathsf{L}\mathscr{F} \circ \mathscr{P}_\mathsf{A} \rightsquigarrow \mathscr{P}_\mathsf{B} \circ \mathsf{K}(\mathscr{F})$.*

Thus, excuse our poor diagram, for it is doing its best to commute, and any choice of a bottom side other than $\mathsf{L}\mathscr{F}$ would make it commute even less.

Proof. If we have done our homework (and in particular Exercise 6.4), then we know that there is a natural transformation

$$\mathscr{I}_\mathsf{A} \circ \mathscr{P}_\mathsf{A} \rightsquigarrow \mathrm{id}_{\mathsf{K}^-(\mathsf{A})};$$

composing on the left by $\mathscr{P}_\mathsf{B} \circ \mathsf{K}(\mathscr{F})$ gives the first point, since $\mathsf{L}\mathscr{F} = \mathscr{P}_\mathsf{B} \circ \mathsf{K}(\mathscr{F}) \circ \mathscr{I}_\mathsf{A}$, by definition.

To verify the second point, note that every natural transformation $\mathscr{G} \rightsquigarrow \mathsf{L}\mathscr{F}$ will induce a natural transformation γ as in the statement, by the first point; on the other hand, every factorization of γ,

$$\mathscr{G} \circ \mathscr{P}_\mathsf{A} \rightsquigarrow \mathsf{L}\mathscr{F} \circ \mathscr{P}_\mathsf{A} \rightsquigarrow \mathscr{P}_\mathsf{B} \circ \mathsf{K}(\mathscr{F}),$$

comes from one and only one (up to natural isomorphism) natural transformation $\mathscr{G} \rightsquigarrow \mathsf{L}\mathscr{F}$, because this can be recovered by composing on the right by $\mathrm{id}_{\mathscr{I}_\mathsf{A}}$ (and again using Exercise 6.4):

$$\mathscr{G} \xrightarrow{\sim} \mathscr{G} \circ \mathscr{P}_\mathsf{A} \circ \mathscr{I}_\mathsf{A} \rightsquigarrow \mathscr{P}_\mathsf{B} \circ \mathsf{K}(\mathscr{F}) \circ \mathscr{I}_\mathsf{A} = \mathsf{L}\mathscr{F}.$$

This concludes the proof. □

The reader should formulate the universal property satisfied by the *right*-derived functor. In this case, not surprisingly, all natural transformations go 'backwards'.

Remark 7.4. The fact that there is a natural transformation $\mathscr{I} \circ \mathscr{P} \rightsquigarrow \mathrm{id}$ (Exercise 6.4), used in the proof of Proposition 7.3, has another interesting consequence. Let A, B, C be abelian categories, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$, $\mathscr{G} : \mathsf{B} \to \mathsf{C}$ be additive functors. Assume that A and B have enough projectives, so that the derived functors $\mathsf{L}\mathscr{F}$ and $\mathsf{L}\mathscr{G}$ are defined, as well as the derived functor $\mathsf{L}(\mathscr{G} \circ \mathscr{F})$. Then there is a natural transformation

$$\mathsf{L}\mathscr{G} \circ \mathsf{L}\mathscr{F} \rightsquigarrow \mathsf{L}(\mathscr{G} \circ \mathscr{F}).$$

Indeed, in the same notational style as above, $\mathsf{L}\mathscr{G} \circ \mathsf{L}\mathscr{F}$ expands to

$$\mathscr{P}_{\mathsf{C}} \circ \mathsf{K}(\mathscr{G}) \circ \mathscr{I}_{\mathsf{B}} \circ \mathscr{P}_{\mathsf{B}} \circ \mathsf{K}(\mathscr{F}) \circ \mathscr{I}_{\mathsf{A}},$$

and contracting $\mathscr{I}_{\mathsf{B}} \circ \mathscr{P}_{\mathsf{B}}$ to $\mathrm{id}_{\mathsf{K}^-(\mathsf{B})}$ takes this to

$$\mathscr{P}_{\mathsf{C}} \circ \mathsf{K}(\mathscr{G}) \circ \mathsf{K}(\mathscr{F}) \circ \mathscr{I}_{\mathsf{A}} = \mathsf{L}(\mathscr{G} \circ \mathscr{F}).$$

If additional conditions are satisfied (for example, if $\mathscr{F}$ preserves projectives), then this natural transformation is a natural isomorphism, so L 'distributes' through compositions in this case (Exercise 7.4).

Towards the end of the chapter the reader will look again at the relation between $\mathsf{L}\mathscr{G} \circ \mathsf{L}\mathscr{F}$ and $\mathsf{L}(\mathscr{G} \circ \mathscr{F})$ from a different point of view, using spectral sequences (Exercises 8.8 and 9.10). ⌟

7.3. Taking cohomology. The content of Proposition 7.3 is again in line with our main strategy: in moving from A to B, the left derived functor $\mathscr{F}$ is the functor that preserves 'as much cohomological information as possible' regarding bounded-above complexes, in the sense that it is the closest one can get to extending $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ to a functor $\mathsf{D}^-(\mathsf{A}) \to \mathsf{D}^-(\mathsf{B})$.

Applying cohomology extracts this information; actually, when deriving on the left, the indexing works better if we take *homology* rather than cohomology. The *i-th left-derived functor of $\mathscr{F}$* is the functor $\mathsf{L}_i\mathscr{F} := H^{-i} \circ \mathsf{L}\mathscr{F}$. Note that

$$\mathsf{L}_i\mathscr{F}(P^\bullet) = H^{-i}(\mathscr{P}_{\mathsf{B}}(\mathsf{C}(\mathscr{F})(P^\bullet))) \cong H^{-i}(\mathsf{C}(\mathscr{F})(P^\bullet)),$$

since $\mathscr{P}_{\mathsf{B}}(\mathsf{C}(\mathscr{F})(P^\bullet))$ is quasi-isomorphic to $\mathsf{C}(\mathscr{F})(P^\bullet)$. Thus, if cohomology is all we are interested in, it is not necessary to find a projective resolution of $\mathsf{C}(\mathscr{F})(P^\bullet)$. (In particular, we can define $\mathsf{L}_i\mathscr{F}$ even if B does not have enough projectives.)

At this point we can come down to earth and define each $\mathsf{L}_i\mathscr{F}$ as a functor $\mathsf{A} \to \mathsf{B}$: we know that this is a somewhat limited scope (because $\mathsf{L}_i\mathscr{F}$ can properly be defined on the much subtler $\mathsf{D}^-(\mathsf{A})$), but it will be the source of the only applications we will really consider at any level of depth, and historically this is how derived functors arose originally.

Definition 7.5. Let A, B be abelian categories, and assume that A has enough projectives. Let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be an additive functor. The *i-th left-derived functor* $\mathsf{L}_i\mathscr{F}$ *of* $\mathscr{F}$ is the functor $\mathsf{A} \to \mathsf{B}$ given by

$$\mathsf{L}_i\mathscr{F} = H^{-i} \circ \mathsf{L}\mathscr{F} \circ \mathscr{P}_{\mathsf{A}} \circ \iota_{\mathsf{A}}.$$

For an object M in A, the complex in $\mathsf{C}(\mathsf{B})$ with $\mathsf{L}_i\mathscr{F}(M)$ in degree $-i$ and with vanishing differentials is denoted by $\mathsf{L}_\bullet\mathscr{F}(M)$. ⌟

Let's spell this out. Given an object M of A, $\mathsf{L}_i\mathscr{F}(M)$ is obtained by finding any projective resolution $P^\bullet_M$ of M, applying the functor $\mathsf{C}(\mathscr{F})$ to $P^\bullet_M$ to obtain a complex in $\mathsf{C}(\mathsf{B})$, and taking the $(-i)$-th cohomology of this complex[31]. Up to isomorphism, the result does not depend on the choice of the projective resolution: this is clear from the path of concepts that led us here, and it is directly implied by

- Proposition 6.4, showing that any two projective resolutions are homotopy equivalent, in conjunction with
- Theorem 4.14, showing that the images of homotopy equivalent complexes via an additive functor have isomorphic cohomology.

Of course we can similarly define the *i-th right-derived functor* $\mathsf{R}^i\mathscr{F}$ of an additive functor $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ and complexes $\mathsf{R}^\bullet\mathscr{F}(M)$, provided that A has enough injectives: concretely, $\mathsf{R}^i\mathscr{F}(M)$ is the cohomology of the image of an injective resolution of M, i.e.,

$$\mathsf{R}^i\mathscr{F} = H^i \circ \mathsf{R}\mathscr{F} \circ \mathscr{Q}_\mathsf{A} \circ \iota_\mathsf{A}.$$

So far, we have always implicitly assumed that $\mathscr{F}$ was a *covariant* functor. *Contravariant* functors $\mathsf{A} \to \mathsf{B}$ should be viewed as covariant functors $\mathsf{A}^{op} \to \mathsf{B}$ (Definition VIII.1.1); the roles of injectives and projectives should therefore be swapped. Thus, the *right*-derived functors of an additive contravariant functor $\mathsf{A} \to \mathsf{B}$ will be defined if A has enough *projectives:* A^{op} will then have enough *injectives,* as needed.

The attentive reader already knows two families of examples of derived functors, both defined from the category R-Mod of modules over a (commutative) ring to itself. Recall that R-Mod has both enough projectives and injectives (as seen in §VIII.6.2 and §VIII.6.3); thus every covariant/contravariant functor R-$\mathsf{Mod} \to R$-Mod can be derived on the left and on the right.

Example 7.6. Every R-module N determines a functor $\_ \otimes_R N$: $M \mapsto M \otimes_R N$ (see §VIII.2.2). The *left*-derived functor of $\_ \otimes_R N$ is denoted $\_ \overset{\mathsf{L}}{\otimes}_R N$ and acts $\mathsf{D}^-(R\text{-}\mathsf{Mod}) \to \mathsf{D}^-(R\text{-}\mathsf{Mod})$. The i-th left-derived functor of $\_ \otimes_R N$, viewed as a functor R-$\mathsf{Mod} \to R$-Mod, is $\mathrm{Tor}_i^R(\_ \otimes_R N)$: indeed, the construction of $\mathrm{Tor}_i^R(M, N)$ given in §VIII.2.4 matches precisely the 'concrete' interpretation of the i-th left-derived functor given above. The reader may note that in §VIII.2.4 we used a *free* resolution of M; free modules are projective, so this was simply a convenient way to choose a projective resolution. The fact that we could use *any* projective resolution of M to compute $\mathrm{Tor}_i^R(M, N)$ was mentioned at the end of §VIII.6.2 and was attributed there to the 'magic of homological algebra'. This piece of magic has now been explained in gory detail.

However, we also mentioned that *flat* resolutions may be used in place of projective resolutions, and *this* piece of magic still needs to be explained. The same applies to the fact, mentioned in §VIII.2.4, that $\mathrm{Tor}_i(M, N)$ may in fact be computed by using a projective resolution *of* N rather than M. Both mysteries will be dispelled in §8. ⌟

[31] Projective resolutions of an object are complexes with nonzero objects only in degree ≤ 0; hence $\mathsf{L}_i\mathscr{F} = 0$ for $i < 0$.

Example 7.7. Similarly, Hom_R admits a right-derived functor $\mathsf{R}\mathrm{Hom}_R$, and its manifestations as the right-derived functors of $\mathrm{Hom}_R(M,\_)$ are the Ext modules: $\mathrm{Ext}^i_R(M,N)$ is (isomorphic to) the i-th cohomology of $\mathrm{Hom}_R(M,N^\bullet)$, where $N^\bullet$ is an *injective* resolution of N. As mentioned in §VIII.6.4, it is *also* the i-th cohomology of $\mathrm{Hom}_R(M^\bullet,N)$, where $M^\bullet$ is a *projective* resolution of M. This projective resolution should really be viewed as an injective resolution in the opposite category $R\text{-}\mathsf{Mod}^{op}$, since the functor $\mathrm{Hom}_R(\_,N)$ is contravariant.

At this point we understand why the results of these operations are independent of the chosen injective/projective resolution. We are, however, not quite ready to verify that the two strategies lead to the same Ext functor; this last point will also be clarified in §8.

Also note that we could define the Ext functors as functors to Ab on any abelian category with enough injectives and/or projectives: any abelian category has left-exact Hom functors[32] to Ab. ⌟

7.4. Long exact sequence of derived functors. The most remarkable property of the functors Tor_i and Ext^i mentioned in Chapter VIII is probably that they 'repair' the lack of exactness of $\otimes$, Hom, respectively, in the sense that they agree with these functors in degree 0 and they fit into long exact sequences. We stated the existence of these exact sequences in §VIII.2.4 and §VIII.6.4, without proof (save for indications in case the base ring R is a PID); now we are ready to understand fully why these sequences exist, in the general context of derived functors.

We will keep assuming that A has enough projectives. We will prove that every short exact sequence

$$0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$$

in A induces a 'long exact sequence of derived functors'; the sequences for Tor and Ext encountered in Chapter VIII will be particular cases. Surely the reader expects this general fact to follow one way or the other from the long exact cohomology sequence (Theorem 3.5); that reader will not be disappointed.

From a more sophisticated perspective, what happens is that derived functors fit the vertices of a 'distinguished triangle' in the derived category: as we mentioned in passing in §4.2, these triangles play the role of exact sequences in the homotopic and derived categories, which do not happen to be abelian. Distinguished triangles give rise to long exact sequences, in much the same way as do exact sequences of complexes in the abelian case explored in §3.3.

Since we do not have the machinery of triangulated categories at our disposal, we have to resort to bringing the action back to the ordinary category of complexes, with the aim of using Theorem 3.5. The key point is therefore the following: assuming that A has enough projectives and that

$$0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$$

is an exact sequence in A, can we arrange for projective resolutions of L, M, N to form an exact sequence in $\mathsf{C}(\mathsf{A})$?

[32] In fact, the presence of injectives or projectives may be bypassed if we adopt the 'Yoneda' viewpoint on Ext.

Yes. This is often called the 'horseshoe lemma', after the shape of the main diagram appearing in its proof.

Lemma 7.8. *Let*

$$(*) \qquad 0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$$

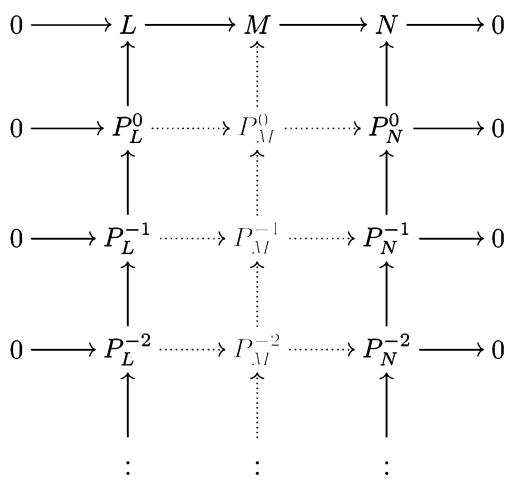
be an exact sequence in an abelian category A *with enough projectives. Assume* $P_L^\bullet$, $P_N^\bullet$ *are projective resolutions of* L, N, *respectively. Then there exists an exact sequence*

$$(**) \qquad 0 \longrightarrow P_L^\bullet \longrightarrow P_M^\bullet \longrightarrow P_N^\bullet \longrightarrow 0$$

where $P_M^\bullet$ *is a projective resolution of* M, *inducing (*) in cohomology.*

By 'inducing (*) in cohomology' we mean that the (not too) long exact cohomology sequence induced by (**) reduces to the H^0 part, since all other cohomology objects of a resolution vanish; identifying the zero-th cohomology of the resolutions with the corresponding objects, this part is nothing but the original short exact sequence (*).

Proof. The hypotheses give us the solid part of the diagram



and our task is to fill in the blanks with projective objects and morphisms so that all rows are exact, and the middle column is a resolution of M. We claim that we can let $P_M^i := P_L^i \oplus P_N^i$; this is projective (Exercise 5.4), and the standard morphisms make each sequence

$$0 \longrightarrow P_L^i \longrightarrow P_L^i \oplus P_N^i \longrightarrow P_N^i \longrightarrow 0$$

exact, so all we have to produce are morphisms giving an exact complex

$$\cdots \longrightarrow P_L^{-2} \oplus P_N^{-2} \longrightarrow P_L^{-1} \oplus P_N^{-1} \longrightarrow P_L^0 \oplus P_N^0 \longrightarrow M \longrightarrow 0$$

fitting in the commutative diagram displayed above.

As always, the construction follows an inductive pattern. Since P_N^0 is projective and $M \to N$ is an epimorphism, the morphism $P_N^0 \to N$ lifts to a morphism

$\beta : P_N^0 \to M$. On the other hand, P_L^0 maps to M via $\alpha : P_L^0 \to L \to M$. Thus, there is a morphism

$$\pi_M := \alpha \oplus \beta \qquad P_L^0 \oplus P_N^0 \longrightarrow M,$$

and the diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & L & \longrightarrow & M & \longrightarrow & N & \longrightarrow & 0 \\ & & \uparrow{\scriptstyle \pi_L} & & \uparrow{\scriptstyle \pi_M} & & \uparrow{\scriptstyle \pi_N} & & \\ 0 & \longrightarrow & P_L^0 & \longrightarrow & P_L^0 \oplus P_N^0 & \longrightarrow & P_N^0 & \longrightarrow & 0 \end{array}$$

commutes by construction (note that the composition $P_L^0 \oplus P_N^0 \to M \to N$ is 0 on the P_L^0 factor, by the exactness of the top row). The morphism π_M is an epimorphism, by an immediate application of the snake lemma.

The snake lemma and the fact that π_L is an epimorphism also imply that the kernels of the vertical maps form an exact sequence, and we have epimorphisms from P_L^{-1} and P_N^{-1} to the corresponding kernels since the columns of the original diagram are exact:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \ker \pi_L & \longrightarrow & \ker \pi_M & \longrightarrow & \ker \pi_N & \longrightarrow & 0 \\ & & \uparrow{\scriptstyle \pi_L^{-1}} & & & & \uparrow{\scriptstyle \pi_N^{-1}} & & \\ & & P_L^{-1} & & & & P_N^{-1} & & \end{array}$$

We let $P_M^{-1} = P_L^{-1} \oplus P_N^{-1}$ and define $\pi_M^{-1} : P_L^{-1} \oplus P_N^{-1} \to \ker \pi_M \to P_L^0 \oplus P_N^0$ by the same mechanism used above:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & L & \longrightarrow & M & \longrightarrow & N & \longrightarrow & 0 \\ & & \uparrow{\scriptstyle \pi_L} & & \uparrow{\scriptstyle \pi_M} & & \uparrow{\scriptstyle \pi_N} & & \\ 0 & \longrightarrow & P_L^0 & \longrightarrow & P_L^0 \oplus P_N^0 & \longrightarrow & P_N^0 & \longrightarrow & 0 \\ & & \uparrow & & \uparrow & & \uparrow & & \\ 0 & \longrightarrow & P_L^{-1} & \longrightarrow & P_L^{-1} \oplus P_N^{-1} & \longrightarrow & P_N^{-1} & \longrightarrow & 0 \end{array}$$

By the snake lemma, $P_L^{-1} \oplus P_N^{-1} \to \ker \pi_M$ is an epimorphism, and this implies exactness of the central column at $P_L^0 \oplus P_N^0$. Continuing inductively constructs $P_M^\bullet$ as required. □

Lemma 7.8 tells us that we can lift exact sequences of objects to exact sequences of projective resolutions. Morally, we would like to say that the functor $\mathscr{P}$ assigning to every object of A a projective resolution in $\mathsf{K}(\mathsf{A})$ is 'exact'; but as $\mathsf{K}(\mathsf{A})$ is not abelian, this is simply not an option. Lemma 7.8 gets as close as possible to such a statement. In fact, from the construction of $P_M^\bullet$ given in the proof, it is easy to show (Exercise 7.5) that $P_M^\bullet$ is the mapping cone of a morphism $\rho^\bullet : P_N[-1]^\bullet \to P_L^\bullet$, so

that we obtain a 'distinguished triangle'

$$\begin{array}{ccc} & P_L^\bullet & \\ {\scriptstyle +1}\nearrow & & \searrow \\ P_N^\bullet & \longleftarrow & MC(\rho)^\bullet = P_M^\bullet \end{array}$$

in the sense mentioned in §4.1. These are the triangles giving the homotopic category the structure of 'triangulated category'. Thus, the functor $\mathscr{P}$ allows us to construct a distinguished triangle in $\mathsf{K}(\mathsf{P})$ starting from a 'special' exact triangle

$$\begin{array}{ccc} & \iota(L) & \\ {\scriptstyle +1}\nearrow & & \searrow \\ \iota(N) & \longleftarrow & \iota(M) \end{array}$$

in the sense of §3.4. The horseshoe lemma shows that this functor preserves as much exactness as is allowed by the context.

We should also note that once short exact sequences are lifted, we can in fact lift every complex; this will be needed later. Here is the precise statement:

Corollary 7.9. *Let*

$$M^\bullet: \qquad \cdots \longrightarrow M^{-3} \longrightarrow M^{-2} \longrightarrow M^{-1} \longrightarrow M^0 \longrightarrow 0$$

be a complex in an abelian category A *with enough projectives. Then there is a complex of complexes:*

$$P_{M^\bullet}^\bullet: \qquad \cdots \longrightarrow P_{M^{-3}}^\bullet \longrightarrow P_{M^{-2}}^\bullet \longrightarrow P_{M^{-1}}^\bullet \longrightarrow P_{M^0}^\bullet \longrightarrow 0$$

such that each $P_{M^i}^\bullet$ *is a projective resolution of* M^i *and inducing* $M^\bullet$ *in cohomology. If* $M^\bullet$ *is exact, then* $P_{M^\bullet}^\bullet$ *may be chosen to be exact.*

Proof. Break up $M^\bullet$ into short exact sequences

$$0 \longrightarrow K^i \longrightarrow M^i \longrightarrow I^{i+1} \longrightarrow 0$$

together with exact sequences

$$0 \longrightarrow I^i \longrightarrow K^i \longrightarrow H^i \longrightarrow 0$$

where K^i is the kernel of $M^i \to M^{i+1}$, I^{i+1} is its image, and H^i is the cohomology at M^i. Choose arbitrary projective resolutions P_{I^i}, P_{H^i} of I^i and H^i, for all i. Then the horseshoe lemma (Lemma 7.8) yields projective resolutions of K^i and short exact sequences

$$0 \longrightarrow P_{I^i}^\bullet \longrightarrow P_{K^i}^\bullet \longrightarrow P_{H^i}^\bullet \longrightarrow 0\,,$$

with evident notation, and then the horseshoe lemma again and the previous sequences give projective resolutions P_{M^i} of M^i and short exact sequences

$$0 \longrightarrow P_{K^i}^\bullet \longrightarrow P_{M^i}^\bullet \longrightarrow P_{I^{i+1}}^\bullet \longrightarrow 0\,.$$

The $P^\bullet_{M^i}$ can now be assembled into a sequence, by using differentials obtained by the compositions

$$P^\bullet_{M^i} \to P^\bullet_{I^{i+1}} \to P^\bullet_{K^{i+1}} \to P^\bullet_{M^{i+1}}.$$

It is clear that one obtains a complex and that $P^\bullet_{K^i}$, resp., $P^\bullet_{I^{i+1}}$, are the kernel, resp., the image, of $P^\bullet_{M^i} \to P^\bullet_{M^{i+1}}$ (Exercise 1.21). The statements follow. (If $M^\bullet$ is exact, then $I^i = K^i$, and the construction gives $P^\bullet_{I^i} = P^\bullet_{K^i}$, so that $P^\bullet_{M^\bullet}$ is exact.) □

Remark 7.10. The proof of Corollary 7.9 shows that the resolution $P^\bullet_{M^\bullet}$, i.e.,

$$\cdots \xrightarrow{d^{-4}} P^\bullet_{M^{-3}} \xrightarrow{d^{-3}} P^\bullet_{M^{-2}} \xrightarrow{d^{-2}} P^\bullet_{M^{-1}} \xrightarrow{d^{-1}} P^\bullet_{M^0} \longrightarrow 0\ ,$$

can in fact be chosen so that $\operatorname{im} d^i$, $\ker d^i$, and the cohomology $\ker d^i / \operatorname{im} d^{i-1}$ are all projective resolutions of the corresponding objects from $M^\bullet$. This is useful in some applications. Resolutions satisfying this property are called *Cartan-Eilenberg* (or 'fully projective') *resolutions.* ⌟

Coming back to the issue at hand, we are one step away from the promised long exact sequence of derived functors. The last ingredient is the following:

Lemma 7.11. *Let* A *be an abelian category, and let*

$$(*) \qquad 0 \longrightarrow L^\bullet \longrightarrow M^\bullet \longrightarrow P^\bullet \longrightarrow 0$$

be an exact sequence of complexes in A*, where* P^i *is projective for all* i*. Let* $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ *be any additive functor of abelian categories. Then the sequence obtained by applying* $\mathscr{F}$ *to (*),*

$$0 \longrightarrow \mathscr{F}(L^\bullet) \longrightarrow \mathscr{F}(M^\bullet) \longrightarrow \mathscr{F}(P^\bullet) \longrightarrow 0\ ,$$

is exact.

Note that we are not asking $\mathscr{F}$ to be exact in any sense.

Proof. Since P^i is projective, the sequence

$$0 \longrightarrow L^i \longrightarrow M^i \longrightarrow P^i \longrightarrow 0$$

splits (see the end of §VIII.6.1). It follows that

$$0 \longrightarrow \mathscr{F}(L^i) \longrightarrow \mathscr{F}(M^i) \longrightarrow \mathscr{F}(P^i) \longrightarrow 0$$

is (split and) exact for all i (Exercise 5.11). Exactness of a sequence of complexes is determined by exactness at each degree, so this proves the statement. □

Now the long exact sequence of right-derived functors is an immediate consequence of the long exact cohomology sequence.

Theorem 7.12. *Let* $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ *be an additive functor of abelian categories, and assume* A *has enough projectives. Every exact sequence*

$$0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$$

in A *induces a long exact sequence*

$$\begin{array}{ccccccc}
\cdots \longrightarrow & \mathsf{L}_2\mathscr{F}(L) & \longrightarrow & \mathsf{L}_2\mathscr{F}(M) & \longrightarrow & \mathsf{L}_2\mathscr{F}(N) & \\
& & \searrow^{\delta_2} & & & & \\
& \mathsf{L}_1\mathscr{F}(L) & \longrightarrow & \mathsf{L}_1\mathscr{F}(M) & \longrightarrow & \mathsf{L}_1\mathscr{F}(N) & \\
& & \searrow^{\delta_1} & & & & \\
& \mathsf{L}_0\mathscr{F}(L) & \longrightarrow & \mathsf{L}_0\mathscr{F}(M) & \longrightarrow & \mathsf{L}_0\mathscr{F}(N) & \longrightarrow 0
\end{array}$$

in B. *Further, this assignment is covariantly functorial.*

The last sentence means that a morphism of short exact sequences will induce a morphism of the corresponding long exact sequences, compatibly with compositions. We will leave the diagram chases necessary to prove this to the reader (Exercise 7.7).

Proof. By Lemma 7.8, the given exact sequence is induced by an exact sequence of projective resolutions

$$0 \longrightarrow P_L^\bullet \longrightarrow P_M^\bullet \longrightarrow P_N^\bullet \longrightarrow 0\ .$$

By Lemma 7.11, the corresponding sequence

$$0 \longrightarrow \mathscr{F}(P_L^\bullet) \longrightarrow \mathscr{F}(P_M^\bullet) \longrightarrow \mathscr{F}(P_N^\bullet) \longrightarrow 0$$

in $\mathsf{C}(\mathsf{B})$ is exact. As the cohomology objects of these complexes are precisely the left-derived functors of $\mathscr{F}$, the long exact cohomology sequence determined by this sequence (by Theorem 3.5) gives a long exact sequence as stated. □

Once more in the style of §3.4, Theorem 7.12 tells us that the triangle

$$\begin{array}{ccc}
 & \iota(L) & \\
{\scriptstyle +1}\nearrow & & \searrow \\
\iota(N) & \longleftarrow & \iota(M)
\end{array}$$

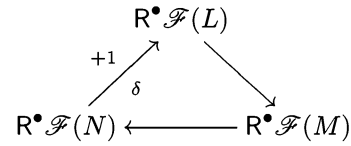
induces a triangle[33]

$$\begin{array}{ccc}
 & \mathsf{L}_\bullet\mathscr{F}(L) & \\
{\scriptstyle -1}\nearrow_{\delta} & & \searrow \\
\mathsf{L}_\bullet\mathscr{F}(N) & \longleftarrow & \mathsf{L}_\bullet\mathscr{F}(M)
\end{array}$$

The vertices of this triangle are complexes in $\mathsf{C}^{\leq 0}(\mathsf{B})$.

[33]The indexing is potentially confusing: the lower • indicates 'homological' indexing, which is opposite to the degree of the corresponding cochain complexes. So $\mathsf{L}_i\mathscr{F}(N)$ corresponds to degree $-i$, which is sent by the northeast arrow to cohomological degree $-i+1$, corresponding to $\mathsf{L}_{i-1}\mathscr{F}(L)$, as it should.

A completely similar situation occurs for right-derived functors: if A has enough injectives, an exact sequence as above induces a triangle



where now the vertices are complexes in $\mathsf{C}^{\geq 0}(\mathsf{B})$.

7.5. Relating $\mathscr{F}$, $\mathsf{L}_i\mathscr{F}$, $\mathsf{R}^i\mathscr{F}$. The reader may have noticed that $\otimes$ was derived *to the left* to obtain Tor, while Hom was derived *to the right* to obtain Ext. The asymmetry mandating this choice lies in the fact that $\otimes$ is right-exact, while Hom is left-exact: any additive functor can be derived to the left or to the right (in the presence of enough projectives, resp., enough injectives), and the derived functors will fit in long exact sequences as proven in Theorem 7.12; but only functors satisfying a measure of exactness can be recovered directly from their derived versions.

To state this fact more precisely, we go back to general considerations for the left-derived functor of an additive functor $\mathscr{F} : \mathsf{A} \to \mathsf{B}$, where A and B are abelian categories and A has enough projectives, and *we now assume that $\mathscr{F}$ is right-exact.*

Proposition 7.13. *Let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be a right-exact additive functor. Then $\mathsf{L}_i\mathscr{F} = 0$ for $i < 0$, and $\mathsf{L}_0\mathscr{F}$ is naturally isomorphic to $\mathscr{F}$.*

Proof. Projective resolutions $P^\bullet$ of an object M of A are in $\mathsf{C}^{\leq 0}(\mathsf{A})$: it follows that $\mathsf{C}(\mathscr{F})(P^\bullet)$ is 0 in positive degree, hence so is its cohomology. Since $H_i = H^{-i}$ (cf. 3.1), the first claim follows.

As for the second, since the sequence

$$P^{-1} \longrightarrow P^0 \longrightarrow M \longrightarrow 0$$

is exact by construction and since $\mathscr{F}$ is right-exact, the sequence

$$\mathscr{F}(P^{-1}) \longrightarrow \mathscr{F}(P^0) \longrightarrow \mathscr{F}(M) \longrightarrow 0$$

is exact. This induces an isomorphism $\nu_L : \mathsf{L}_0\mathscr{F}(M) = H^0(\mathsf{C}(\mathscr{F})(P^\bullet)) \xrightarrow{\sim} \mathscr{F}(M)$. If $\varphi : M_0 \to M_1$ is a morphism in A, by Proposition 6.5 there exists a morphism of projective resolutions $\alpha^\bullet : P_0^\bullet \to P_1^\bullet$ inducing φ: there is a commutative diagram

$$\begin{array}{ccccccccc}
\cdots & \longrightarrow & P_0^{-1} & \longrightarrow & P_0^0 & \longrightarrow & M_0 & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \alpha^{-1}} & & \downarrow{\scriptstyle \alpha^0} & & \downarrow{\scriptstyle \varphi} & & \\
\cdots & \longrightarrow & P_1^{-1} & \longrightarrow & P_1^0 & \longrightarrow & M_1 & \longrightarrow & 0
\end{array}$$

and $H^0(\alpha^\bullet)$ agrees with φ. Applying $\mathscr{F}$ and truncating, we get a commutative diagram

$$\begin{array}{ccccccc}
\mathscr{F}(P_0^{-1}) & \longrightarrow & \mathscr{F}(P_0^0) & \longrightarrow & \mathscr{F}(M_0) & \longrightarrow & 0 \\
\downarrow{\scriptstyle \mathscr{F}(\alpha^{-1})} & & \downarrow{\scriptstyle \mathscr{F}(\alpha^0)} & & \downarrow{\scriptstyle \mathscr{F}(\varphi)} & & \\
\mathscr{F}(P_1^{-1}) & \longrightarrow & \mathscr{F}(P_1^0) & \longrightarrow & \mathscr{F}(M_1) & \longrightarrow & 0
\end{array}$$

with exact rows since $\mathscr{F}$ is right-exact. It follows that the diagram

$$\begin{array}{ccc}
\mathsf{L}_0\mathscr{F}(M_0) & \xrightarrow{\mathsf{L}_0\mathscr{F}(\varphi)} & \mathsf{L}_0\mathscr{F}(M_1) \\
\downarrow{\scriptstyle \nu_{M_0}}\wr & & \wr\downarrow{\scriptstyle \nu_{M_1}} \\
\mathscr{F}(M_0) & \xrightarrow[\mathscr{F}(\varphi)]{} & \mathscr{F}(M_1)
\end{array}$$

commutes, proving that $\nu : \mathsf{L}_0\mathscr{F} \to \mathscr{F}$ is a natural isomorphism. □

Going back to Theorem 7.12, we see that *if $\mathscr{F}$ is right-exact,* then the tail end of the long exact sequence of left-derived functors for $\mathscr{F}$ consists of an application of $\mathscr{F}$ itself. Thus, the situation in this case is the following: starting from a short exact sequence

$$0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$$

in A, we apply $\mathscr{F}$ to obtain an exact sequence

$$\mathscr{F}(L) \longrightarrow \mathscr{F}(M) \longrightarrow \mathscr{F}(N) \longrightarrow 0$$

in which we lost 'the 0 on the left'. The long exact sequence saves the day, continuing the new sequence into an exact complex:

$$\begin{array}{l}
\cdots \longrightarrow \mathsf{L}_2\mathscr{F}(L) \longrightarrow \mathsf{L}_2\mathscr{F}(M) \longrightarrow \mathsf{L}_2\mathscr{F}(N) \\
\qquad\qquad \xrightarrow{\delta_2} \\
\mathsf{L}_1\mathscr{F}(L) \longrightarrow \mathsf{L}_1\mathscr{F}(M) \longrightarrow \mathsf{L}_1\mathscr{F}(N) \\
\qquad\qquad \xrightarrow{\delta_1} \\
\mathscr{F}(L) \longrightarrow \mathscr{F}(M) \longrightarrow \mathscr{F}(N) \longrightarrow 0
\end{array}$$

That is, $\mathsf{L}_1\mathscr{F}$ measures the extent to which $\mathscr{F}$ fails to be left-exact, and the higher $\mathsf{L}_i\mathscr{F}$ give further measures of this failure.

Similarly, $\mathsf{R}^0\mathscr{F}$ is naturally isomorphic to $\mathscr{F}$ if $\mathscr{F}$ is *left*-exact. In this case, applying $\mathscr{F}$ to the original sequence gives the exact sequence

$$0 \longrightarrow \mathscr{F}(L) \longrightarrow \mathscr{F}(M) \longrightarrow \mathscr{F}(N)$$

where we lost the 0 on the right; the job of the right-derived functors of $\mathscr{F}$ is to extend this sequence to an exact complex

$$\begin{array}{ccccccc} 0 & \longrightarrow & \mathscr{F}(L) & \longrightarrow & \mathscr{F}(M) & \longrightarrow & \mathscr{F}(N) \\ & & & \overset{\delta_0}{\searrow} & & & \\ & & \mathsf{R}_1\mathscr{F}(L) & \longrightarrow & \mathsf{R}_1\mathscr{F}(M) & \longrightarrow & \mathsf{R}_1\mathscr{F}(N) \\ & & & \overset{\delta_1}{\searrow} & & & \\ & & \mathsf{R}_2\mathscr{F}(L) & \longrightarrow & \mathsf{R}_2\mathscr{F}(M) & \longrightarrow & \mathsf{R}_2\mathscr{F}(N) \longrightarrow \cdots \end{array}$$

No identification of $\mathscr{F}$ with any of its derived functors should be expected if $\mathscr{F}$ does not satisfy some exactness property. If $\mathscr{F}$ is exact on the nose, then both $\mathsf{L}_0\mathscr{F}$ and $\mathsf{R}^0\mathscr{F}$ agree with $\mathscr{F}$ (up to natural isomorphism); but this is no cause for great excitement, since all other $\mathsf{L}_i\mathscr{F}$, $\mathsf{R}^i\mathscr{F}$ vanish in this case (Exercise 7.8). As already pointed out in Example 7.2, one should not expect to learn anything new from an exact functor by deriving it.

7.6. Example: A little group cohomology. In the beginning of this chapter we pointed out that the general strategy informing the development of homological algebra has numerous applications and we mentioned group and sheaf cohomology as examples. Here are a few words on group cohomology (sheaf cohomology would take us too far).

How do we extract 'cohomological invariants' from a group G? Consider the category G-Mod of abelian groups endowed with a left-G-action, equivalently, the category of left-$\mathbb{Z}[G]$-modules, where $\mathbb{Z}[G]$ is the *group ring* briefly encountered in §III.1.4. Objects of G-Mod may be called G-modules.

For a G-module M, M^G denotes the set of elements that are fixed under the action of G: these are reasonably called the *invariants* of the action. Note that M^G is an abelian group carrying a *trivial* action of G; it is clear that setting $M \to M^G$ defines a covariant functor $\cdot^G : G\text{-}\mathsf{Mod} \to \mathsf{Ab}$. Both G-Mod and Ab are abelian categories, and it takes a moment to realize that $\cdot^G$ is a left-exact functor: the reader should either check this directly or do Exercise 7.16 and then remember Claim VIII.1.19. The reader should in fact contemplate why this functor is not right-exact: if G acts trivially on a coset $[m]$ of a quotient M/L, there is no reason *a priori* why G should act trivially on a representative m.

We would have drawn this conclusion without difficulty before delving into homological algebra. Homological algebra tells us how to address this type of situation and 'quantify' precisely the failure of exactness, by means of an appropriate derived functor.

The i-th right-derived functor of $\cdot^G$ is denoted $H^i(G, \_)$. Therefore, $H^0(G, M) = M^G$, and for every short exact sequence of G-modules

$$0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$$

we obtain (from Theorem 7.12) a long exact sequence of abelian groups:

$$0 \longrightarrow L^G \longrightarrow M^G \longrightarrow N^G \longrightarrow H^1(G, L) \longrightarrow H^1(G, M) \longrightarrow \cdots .$$

Note that $M^G = \mathrm{Hom}_{\mathbb{Z}[G]}(\mathbb{Z}, M)$, where $\mathbb{Z}$ is given the trivial module structure. Thus, $H^i(G, M) = \mathrm{Ext}^i_{\mathbb{Z}[G]}(\mathbb{Z}, M)$; as we know already (and as will finally be verified in §8), this can be computed by an injective resolution of M or by a projective resolution of $\mathbb{Z}$.

Example 7.14. Take $G = \mathbb{Z}$. Then $\mathbb{Z}[G]$ is the ring $\mathbb{Z}[x, x^{-1}]$ of Laurent polynomials. As $\mathbb{Z}[x, x^{-1}]/(1-x) \cong \mathbb{Z}$, with the trivial action (check this!), the complex

$$\cdots \longrightarrow 0 \longrightarrow \mathbb{Z}[x, x^{-1}] \xrightarrow{\cdot(1-x)} \mathbb{Z}[x, x^{-1}] \longrightarrow 0 \longrightarrow \cdots$$

is a free, hence projective, resolution of $\mathbb{Z}$, endowed with the trivial $\mathbb{Z}$-action. Applying the (contravariant) $\mathrm{Hom}_{\mathbb{Z}[x,x^{-1}]}(\_, M)$, we see that $H^\bullet(\mathbb{Z}, M)$ is computed by the cohomology of

$$\cdots \longrightarrow 0 \longrightarrow M \xrightarrow{\cdot(1-x)} M \longrightarrow 0 \longrightarrow \cdots .$$

'Multiplication by x' on M is nothing but the action of the generator of $G = \mathbb{Z}$. We conclude that $H^0(\mathbb{Z}, M) = \ker(M \xrightarrow{\cdot(1-x)} M) = M^{\mathbb{Z}}$ (as we knew already),

$$H^1(\mathbb{Z}, M) = \frac{M}{(1-x)M},$$

and $H^i(\mathbb{Z}, M) = 0$ for all $i < 0$ and $i \geq 2$. ⌟

Example 7.15 (Finite cyclic groups)**.** Let $G = C_m$ be a cyclic group of order m; then $\mathbb{Z}[G] \cong \mathbb{Z}[x]/(x^m - 1)$. Again it is not difficult to produce a projective resolution of $\mathbb{Z}$ (with trivial action) in the category of $\mathbb{Z}[C_m]$-modules: letting $N = 1 + x + \cdots + x^{m-1} = (1 - x^m)/(1 - x)$, the reader will verify that the complex

$$\cdots \xrightarrow{\cdot N} \mathbb{Z}[C_m] \xrightarrow{\cdot(1-x)} \mathbb{Z}[C_m] \xrightarrow{\cdot N} \mathbb{Z}[C_m] \xrightarrow{\cdot(1-x)} \mathbb{Z}[C_m] \longrightarrow 0 \longrightarrow \cdots$$

has cohomology $\cong \mathbb{Z}$, concentrated in degree 0 (i.e., on the last nonzero term). Applying $\mathrm{Hom}_{\mathbb{Z}[C_m]}(\_, M)$, we get the complex

$$\cdots \longrightarrow 0 \longrightarrow M \xrightarrow{\cdot(1-x)} M \xrightarrow{\cdot N} M \xrightarrow{\cdot(1-x)} M \xrightarrow{\cdot N} \cdots$$

and deduce that $H^i(M, C_m) = 0$ for $i < 0$, $H^0(C_m, M) = M^{C_m}$, and for all $i > 0$

$$H^{2i+1}(C_m, M) \cong \frac{\ker(M \xrightarrow{\cdot N} M)}{(1-x)M},$$

$$H^{2i}(C_m, M) \cong \frac{M^{C_m}}{N \cdot M}.$$ ⌟

There is a standard free resolution of $\mathbb{Z}$ over a group ring $\mathbb{Z}[G]$, which leads to a concrete description of group cohomology. For a finite group G, of order m, the beginning of this resolution goes as follows. Consider the complex

$$(\dagger) \qquad \cdots \longrightarrow \mathbb{Z}[G]^{m^2} \xrightarrow{d^{-2}} \mathbb{Z}[G]^m \xrightarrow{d^{-1}} \mathbb{Z}[G] \xrightarrow{d^0} \mathbb{Z} \longrightarrow 0 \longrightarrow \cdots .$$

Here, a basis of $\mathbb{Z}[G]^{m^k}$ over $\mathbb{Z}[G]$ consists of k-tuples $[g_1, \dots, g_k]$ of elements of G. The morphisms in (†) are defined by setting

$$d^{-1}([g]) = 1 \cdot g - 1 \cdot e_G \in \mathbb{Z}[G],$$
$$d^{-2}([g_1, g_2]) = g_1[g_2] - [g_1 g_2] + [g_1],$$

on the bases and extending by $\mathbb{Z}[G]$-linearity, and d^0 sends every $g \in G$ to 1, so that $d^0(\sum_{g \in G} a_g g) = \sum a_g$. (Needless to say, d^{-k} may be defined for all k.) The reader will check that (†) is exact (Exercise 7.17), at least for the part relevant to the discussion that follows. Therefore,

$$\cdots \longrightarrow \mathbb{Z}[G]^{m^2} \xrightarrow{d^{-2}} \mathbb{Z}[G]^m \xrightarrow{d^{-1}} \mathbb{Z}[G] \longrightarrow 0 \longrightarrow \cdots$$

is (the beginning of) a free $\mathbb{Z}[G]$-resolution of $\mathbb{Z}$ (endowed with the trivial G-action), and group cohomology can be computed as the cohomology of the complex obtained by applying $\mathrm{Hom}_{\mathbb{Z}[G]}(\_, M)$ to this resolution:

$$0 \to \mathrm{Hom}_{\mathbb{Z}[G]}(\mathbb{Z}[G], M) \to \mathrm{Hom}_{\mathbb{Z}[G]}(\mathbb{Z}[G]^m, M) \to \mathrm{Hom}_{\mathbb{Z}[G]}(\mathbb{Z}[G]^{m^2}, M) \to \cdots.$$

Here, $\mathrm{Hom}_{\mathbb{Z}[G]}(\mathbb{Z}[G]^{m^k}, M) \cong M^{m^k}$ may be identified with the abelian group of functions $G^k \to M$, since every $\mathbb{Z}[G]$-linear map $\mathbb{Z}[G]^{m^k} \to M$ is determined by its action on a basis. Denote this abelian group by $C^k(G, M)$: this is the group of *k-cochains of G with values in M*.

We can now summarize what we have found as follows:

Proposition 7.16. *Let G be a finite group, and let M be a G-module. Then the group cohomology $H^i(G, M)$ is the cohomology of the cochain complex*

$$0 \longrightarrow C^0(G, M) \xrightarrow{d_G^0} C^1(G, M) \xrightarrow{d_G^1} C^2(G, M) \xrightarrow{d_G^2} \cdots$$

induced by (†).

Tracing definitions, we see that for $a \in C^0(G, M) = M$ and $\alpha : G \to M$ in $C^1(G, M)$,

$$d_G^0(a)(g) = ga - a,$$
$$d_G^1(\alpha)(g_1, g_2) = g_1 \alpha(g_2) - \alpha(g_1 g_2) + \alpha(g_1).$$

Example 7.17. Let G be the Galois group of a finite Galois extension $k \subseteq F$. Then G acts on the multiplicative group F^* of F, and we can view F^* as a G-module.

Claim 7.18. $H^1(G, F^*) = 0$.

Indeed, with notation as above we have $H^1(G, F^*) \cong \ker d_G^1 / \operatorname{im} d_G^0$, and we can compute this quotient explicitly. Let $\alpha \in C^1(G, F^*)$; denote by α_g the image of g in F^* by α. The 'cocycle condition' $\alpha \in \ker d_G^1$ translates into

$$\alpha_{gh} = \alpha_g \cdot g(\alpha_h)$$

for all g, h in G. (We are now writing the operation in F^* multiplicatively and viewing elements of G as automorphisms of F.)

The elements $h \in G$ are pairwise distinct as automorphisms of F, so by Exercise VII.6.14 they must be linearly independent. Therefore, there exists a $\gamma \in F$ such that

$$\beta := \sum_{h \in G} \alpha_h \cdot h(\gamma) \neq 0.$$

Thus $\beta \in F^*$, and for all $g \in G$

$$g(\beta) = \sum_{h \in G} g(\alpha_h) \cdot (gh)(\gamma) \overset{!}{=} \sum_{h \in G} \alpha_g^{-1} \alpha_{gh} \cdot (gh)(\gamma) = \alpha_g^{-1} \beta$$

where the equality marked $\overset{!}{=}$ holds by the cocycle condition. That is, we have found that there exists a $\beta \in F^*$ such that

$$\alpha_g = \frac{\beta}{g(\beta)}$$

for all $g \in G$. But this says precisely that α is in the image of d_G^0, proving that $\ker d_G^1 / \operatorname{im} d_G^0 = 0$, as claimed.

Claim 7.18 is significant: it goes under the name of *Hilbert's theorem 90,* because in the case in which G is the Galois group of a finite *cyclic* extension, it recovers precisely the classical result with this name. The reader had the opportunity to prove this particular case in Exercise VII.6.16 and now will enjoy the chance to verify that indeed Claim 7.18 implies Hilbert's result (Exercise 7.18). ⌟

Group cohomology is a useful tool in algebraic number theory, algebraic topology, and other fields, and so is the companion 'group *homology*' $H_\bullet(G, M)$, defined analogously by the left-derived functors of the functor $\cdot_G$ associating with every G-module M the module

$$M_G := \frac{M}{\langle gm - m \rangle_{g \in G}}$$

(just as the elements of M^G are called *invariants,* elements of M_G are called *coinvariants* of the action). For example, in Example 7.14 we verified that $H^1(\mathbb{Z}, M) = M_{\mathbb{Z}} = H_0(\mathbb{Z}, M)$. The reader should now have no difficulty computing simple examples, such as the homology $H_\bullet(\mathbb{Z}, M)$ (Exercise 7.19).

Exercises

7.1. ▷ Let $\mathscr{T} : \mathsf{Ab} \to \mathsf{Ab}$ be the additive functor defined by tensoring by $\mathbb{Z}/2\mathbb{Z}$: $\mathscr{T}(A) := A \otimes_{\mathbb{Z}} \mathbb{Z}/2\mathbb{Z}$. Prove that no nontrivial projective abelian groups can be written as $\mathscr{T}(P)$, for any abelian group P. [§7.1]

7.2. Let A, B be abelian categories, let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be an additive functor, and let $\mathsf{C}(\mathscr{F}) : \mathsf{C}(\mathsf{A}) \to \mathsf{C}(\mathsf{B})$ be the corresponding functor of complexes. If $\alpha^\bullet : L^\bullet \to M^\bullet$ is a morphism in $\mathsf{C}(\mathsf{A})$, prove that $\mathsf{C}(\mathscr{F})$ sends the mapping cone of $\alpha^\bullet$ to the mapping cone of $\mathsf{C}(\mathscr{F})(\alpha^\bullet)$.

7.3. ▷ With notation as in Proposition 7.3, prove that the natural transformation $\mathsf{L}\mathscr{F} \circ \mathscr{P}_{\mathsf{A}} \rightsquigarrow \mathscr{P}_{\mathsf{B}} \circ \mathsf{K}(\mathscr{F})$ is not necessarily an isomorphism. (Hint: View the quasi-isomorphism in Example 4.6 as a projective resolution in $\mathbb{Z}$-Mod, and take $\mathscr{F} : \mathbb{Z}\text{-Mod} \to (\mathbb{Z}/2\mathbb{Z})\text{-Mod}$ to be $\_ \otimes_{\mathbb{Z}} (\mathbb{Z}/2\mathbb{Z})$.) [§7.2]

7.4. ▷ (Cf. Remark 7.4.) Let A, B, C be abelian categories, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$, $\mathscr{G} : \mathsf{B} \to \mathsf{C}$ be additive functors. Assume that A and B have enough projectives, so that the derived functors $\mathsf{L}\mathscr{F}$ and $\mathsf{L}\mathscr{G}$ are defined, as well as the derived functor $\mathsf{L}(\mathscr{G} \circ \mathscr{F})$. Further, assume that $\mathscr{F}$ maps projective objects to projective objects. Prove that $\mathsf{L}\mathscr{G} \circ \mathsf{L}\mathscr{F}$ and $\mathsf{L}(\mathscr{G} \circ \mathscr{F})$ are naturally isomorphic. [§7.2]

7.5. ▷ With notation as in Lemma 7.8, show that there is a morphism $\rho^\bullet : P_N[-1]^\bullet \to P_L^\bullet$ such that $P_M^\bullet$ is homotopy equivalent to the mapping cone $MC(\rho)^\bullet$. (Hint: Look at the proof of the lemma; use the differentials $P_L^i \oplus P_N^i \to P_L^{i+1} \oplus P_N^{i+1}$ to define morphisms $P_N^i \to P_L^{i+1}$.) [§7.5]

7.6. ¬ Let B, C be abelian categories, and assume B has enough projectives, so that Cartan-Eilenberg resolutions can be constructed, as in Corollary 7.9 (cf. also Remark 7.10). Let $\mathscr{G} : \mathsf{B} \to \mathsf{C}$ be an additive functor, and let $M^\bullet : \cdots M^{-2} \to M^{-1} \to M^0 \to 0$ be a complex in B. Let $P_{M^\bullet}^\bullet : \cdots \to P_{M^{-2}}^\bullet \to P_{M^{-1}}^\bullet \to P_{M^0}^\bullet \to 0$ be a Cartan-Eilenberg resolution of $M^\bullet$. Prove that $\mathscr{G}$ maps the cohomology of this complex isomorphically to the cohomology of the complex $\cdots \to \mathscr{G}(P_{M^{-2}}^\bullet) \to \mathscr{G}(P_{M^{-1}}^\bullet) \to \mathscr{G}(P_{M^0}^\bullet) \to 0$. (Hint: With notation as in the proof of Corollary 7.9, for all i and j we have exact sequences $0 \to P_{K^i}^j \to P_{M^i}^j \to P_{I^{i+1}}^j \to 0$ and $0 \to P_{I^i}^j \to P_{K^i}^j \to P_{H^i}^j \to 0$. Note that these necessarily split.)

Stenographically, $\mathscr{G}(H^q(P_{M^\bullet}^\bullet)) \cong H^q(\mathscr{G}(P_{M^\bullet}^\bullet))$, where cohomology is computed with respect to the M-degree. Also note that, by construction, $H^q(P_{M^\bullet}^\bullet)$ is a projective resolution of $H^q(M^\bullet)$; cf. Remark 7.10. This will be an ingredient in the comparison of derived functors for the composition of two functors by means of spectral sequences; cf. Exercises 8.8 and 9.10. [9.10]

7.7. ▷ Complete the proof of Theorem 7.12 by showing that morphisms of short exact sequences induce morphisms of the corresponding long exact sequences of derived functors. [§7.4]

7.8. ▷ Let A, B be abelian categories, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be an *exact* functor. Assume A has both enough projectives and enough injectives, so that $\mathsf{L}_i\mathscr{F}$ and $\mathsf{R}^i\mathscr{F}$ are both defined. Prove that $\mathsf{L}_i\mathscr{F}$ and $\mathsf{R}^i\mathscr{F}$ are both the zero functor for $i \neq 0$. [§7.5]

7.9. ▷ Let A, B be abelian categories, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be an additive functor. Assume A has enough projectives, so that $\mathsf{L}_i\mathscr{F}$ is defined. Prove that if P is projective, then $\mathsf{L}_0\mathscr{F}(P) \cong \mathscr{F}(P)$ and $\mathsf{L}_i\mathscr{F}(P) = 0$ for $i \neq 0$. [§8.1]

7.10. Let A, B be abelian categories, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be an additive functor. Assume A has enough projectives, so that $\mathsf{L}_i\mathscr{F}$ is defined. Prove that $\mathsf{L}_0(\mathscr{F})$ is a *right-exact* functor, determining the same higher derived functors L_i, $i > 0$, as $\mathscr{F}$.

7.11. ¬ Let A, B be abelian categories, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be an additive functor; assume A has enough projectives. Let

$$0 \longrightarrow K \longrightarrow P \longrightarrow A \longrightarrow 0$$

be a short exact sequence in A, with P projective. Prove that $\mathsf{L}_i\mathscr{F}(A) \cong \mathsf{L}_{i-1}\mathscr{F}(K)$, for all $i > 1$. [7.12]

7.12. ▷ Generalize Exercise 7.11 as follows: with the same notation, assume

$$0 \longrightarrow K \longrightarrow B^{-k} \longrightarrow \cdots \longrightarrow B^{-1} \longrightarrow A \longrightarrow 0$$

is an exact sequence in A, such that $\mathsf{L}_i\mathscr{F}(B^{-j}) = 0$ for $i > 0$, $1 \leq j \leq k$. Prove that $\mathsf{L}_i\mathscr{F}(A) \cong \mathsf{L}_{i-k}\mathscr{F}(K)$, for all $i > k$. [§8.1, 8.11]

7.13. ¬ Let A, B be abelian categories. A collection of functors $\mathscr{T}^i : \mathsf{A} \to \mathsf{B}$, $i \geq 0$, is called a ('cohomological') *δ-functor* if

- every short exact sequence

(†) $$0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$$

in A determines 'connecting morphisms'

$$\delta^i : \mathscr{T}^i(N) \to \mathscr{T}^{i+1}(L)$$

such that the induced sequence

(‡)
$$\begin{array}{l} 0 \longrightarrow \mathscr{T}^0(L) \longrightarrow \mathscr{T}^0(M) \longrightarrow \mathscr{T}^0(N) \\ \qquad\qquad\qquad\qquad\qquad\qquad \xrightarrow{\ \delta^0\ } \\ \mathscr{T}^1(L) \longrightarrow \cdots \longrightarrow \mathscr{T}^i(N) \\ \qquad\qquad\qquad\qquad\qquad\qquad \xrightarrow{\ \delta^i\ } \\ \mathscr{T}^{i+1}(L) \longrightarrow \cdots \end{array}$$

is exact;

- the assignment of a complex (‡) for every short exact sequence (†) is functorial (in the sense specified in Exercise 3.11).

Prove that for every additive functor $\mathscr{F} : \mathsf{A} \to \mathsf{B}$, the derived functors $\mathsf{R}^i\mathscr{F}$, $i \geq 0$, form a cohomological δ-functor. Prove that cohomology itself is a δ-functor from $\mathsf{C}^{\geq 0}(\mathsf{A})$ to A.

Define a notion of 'homological' δ-functor $\{\mathscr{T}_i\}$, and prove that left-derived functors give an example. [7.14, 8.14]

7.14. ¬ 'Morphisms' $\{\mathscr{T}^i\} \to \{\mathscr{S}^i\}$ of δ-functors (Exercise 7.13) are natural transformations of the individual functors $\mathscr{T}^i \rightsquigarrow \mathscr{S}^i$ that induce morphisms of the corresponding long exact sequences (‡), for every short exact sequence (†) (in other words, that preserve the connecting morphisms δ).

A δ-functor $\{\mathscr{T}^i\}$ is *universal* if for every δ-functor $\{\mathscr{S}^i\}$, every natural transformation $\mathscr{T}^0 \rightsquigarrow \mathscr{S}^0$ extends to a unique morphism of δ-functors $\{\mathscr{T}^i\} \to \{\mathscr{S}^i\}$.

Universal δ-functors are of course unique up to isomorphism.

It is not hard (but it involves a fair number of diagram chases) to show that a δ-functor $\{\mathscr{T}^i\}$ is universal if for every $i > 0$ and every object A of A there exists a monomorphism $i : A \to B$ for some B, such that $\mathscr{T}^i(i) = 0$. This makes each $\mathscr{T}^i$, $i > 0$, *effaceable*. Figure out how the proof begins: assume that $\{\mathscr{T}^i\}$ and $\{\mathscr{S}^i\}$ are δ-functors and that $\mathscr{T}^1$ is effaceable; show how to extend a natural transformation

$\mathscr{T}^0 \rightsquigarrow \mathscr{S}^0$ to a natural transformation $\mathscr{T}^1 \rightsquigarrow \mathscr{S}^1$, compatibly with the connecting morphisms. (Hint: If $\mathscr{T}^1$ is effaced by $i : A \to B$, let $C = \operatorname{coker} i$ and consider the exact sequence $0 \to A \to B \to C \to 0$.) [7.15]

7.15. ¬ Assume the result stated in Exercise 7.14. Prove that right-derived functors are universal cohomological δ-functors. (Equivalently, left-derived functors are universal homological δ-functors.) For an abelian category A, Prove that cohomology, viewed as a collection of functors from $\mathsf{C}^{\geq 0}(\mathsf{A})$ to A, is a universal δ-functor. (Hint: To efface $\mathsf{L}^i\mathscr{F}$ for $i > 0$, use injectives. To efface H^i for $i > 0$, stare at the sequence of complexes mentioned right before the statement of Proposition 4.1.)

The upshot is that in order to verify that a given collection of functors agrees with the derived functors of a given (say) left-exact functor, it suffices to verify that they form a cohomological δ-functor and that the effaceability condition holds. This is used, for example, to obtain concrete realizations of sheaf cohomology. [8.14]

7.16. ▷ Prove that the functor $\cdot^G$ defined in §7.6 is right-adjoint to the functor $\mathsf{Ab} \to G\text{-}\mathsf{Mod}$ associating with an abelian group A the same group, endowed with the trivial G-action. Give an analogous interpretation for the functor $\cdot_G$. [§7.6]

7.17. ▷ Consider the complex (†) defined in §7.6. Define maps h^{-1}, h^0, h in the diagram

$$\begin{array}{ccccccccccccc}
\cdots & \longrightarrow & \mathbb{Z}[G]^{m^2} & \xrightarrow{d^{-2}} & \mathbb{Z}[G]^m & \xrightarrow{d^{-1}} & \mathbb{Z}[G] & \xrightarrow{d^0} & \mathbb{Z} & \longrightarrow & 0 & \longrightarrow & \cdots \\
 & & \vdots & \swarrow^{h^{-1}} & \vdots & \swarrow^{h^0} & \vdots & \swarrow^{h} & \vdots & & & & \\
\cdots & \longrightarrow & \mathbb{Z}[G]^{m^2} & \xrightarrow{d^{-2}} & \mathbb{Z}[G]^m & \xrightarrow{d^{-1}} & \mathbb{Z}[G] & \xrightarrow{d^0} & \mathbb{Z} & \longrightarrow & 0 & \longrightarrow & \cdots
\end{array}$$

by

$$h(a) := a \cdot e_G, \quad h^0(g) := [g], \quad h^{-1}([g]) := [e_G, g]$$

(and extending by $\mathbb{Z}[G]$-linearity). Use these maps to verify that the complex (†) is exact at $\mathbb{Z}$, $\mathbb{Z}[G]$, $\mathbb{Z}[G]^m$. [§7.6]

7.18. ▷ Deduce Hilbert's theorem 90 (as stated in Exercise VII.6.16) as a consequence of the vanishing of Claim 7.18. (Hint: Use the result of Example 7.15). [§7.6]

7.19. ▷ Compute the group homology $H_\bullet(\mathbb{Z}, M)$ for any abelian group M carrying a $\mathbb{Z}$-action. [§7.6]

---

8. Double complexes

A few mysteries remain on the table concerning Tor and Ext: the fact that, e.g., *flat* resolutions may be used in place of projective ones in order to compute Tor and the fact that resolving 'either argument' leads to the same functors. For example, $\operatorname{Ext}^i_R(M, N)$ may be computed by using a projective resolution of M or an injective resolution of N.

Both mysteries are (of course) instances of general features of the theory, which we examine in this section. The first can be dispelled with a pleasant inductive

argument that seems worth presenting (§8.1). But everything is clarified substantially by going one step further and considering *double complexes,* as we do in the remaining subsections.

8.1. Resolution by acyclic objects. Let A be an abelian category with enough projectives, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be an additive functor to another abelian category. At this point we know how to construct left-derived functors $\mathsf{L}_i\mathscr{F} : \mathsf{A} \to \mathsf{B}$ of $\mathscr{F}$: in two words, $\mathsf{L}_i\mathscr{F}(M)$ is computed by applying $\mathscr{F}$ to a projective resolution of M and taking (co)homology of the resulting complex.

It follows in particular that $\mathsf{L}_i\mathscr{F}(P) = 0$ for $i \geq 1$ if P is projective (cf. Exercise 7.9): if P is projective, then $\iota(P)$:

$$\cdots \longrightarrow 0 \longrightarrow 0 \longrightarrow P \longrightarrow 0 \longrightarrow \cdots$$

is a projective resolution of P, so $\mathsf{L}_\bullet\mathscr{F}(P)$ is the cohomology of

$$\cdots \longrightarrow 0 \longrightarrow 0 \longrightarrow \mathscr{F}(P) \longrightarrow 0 \longrightarrow \cdots;$$

that is,

$$\mathsf{L}_i\mathscr{F}(P) = \begin{cases} \mathscr{F}(P), & i = 0, \\ 0, & i \neq 0. \end{cases}$$

This says that projective objects are '$\mathscr{F}$-acyclic' with respect to left-derived functors:

Definition 8.1. Let $\mathscr{F}$ be an additive functor. An object M of A is *$\mathscr{F}$-acyclic* (w.r.t. left-derived functors) if $\mathsf{L}_i\mathscr{F}(M) = 0$ for $i \neq 0$. ⌟

An object M is $\mathscr{F}$-acyclic with respect to *right*-derived functors if $\mathsf{R}^i\mathscr{F}(M) = 0$ for $i \neq 0$. In practice, this notion is really useful only for right-exact or left-exact functors; the first are derived on the left, and the second on the right (cf. §7.5), so in context we can just talk about '$\mathscr{F}$-acyclic' objects.

While projectives are acyclic for every right-exact functor, a given functor $\mathscr{F}$ may admit other $\mathscr{F}$-acyclic objects.

Example 8.2. Let R be a commutative ring. Recall (Definition VIII.2.13) that an R-module M is *flat* if $\_ \otimes_R M$ is exact, or equivalently (by the symmetry of $\otimes$) if $M \otimes_R \_$ is flat. Flat modules are acyclic with respect to tensor products, in a very strong sense: if N is flat, then $\mathrm{Tor}_i(M, N) = 0$ for $i \neq 0$ and all M; see §VIII.2.4 to be reminded of why, or—better—prove it anew. Further, since (as we will finally prove in this section) Tor functors may be computed by resolving either argument, we also know that, 'symmetrically', if M is flat, then $\mathrm{Tor}_i(M, N) = 0$ for all $i \neq 0$ and all modules N. Thus, a flat module is $\mathscr{F}$-acyclic for every functor $\mathscr{F}$ defined by $\_ \otimes_R N$, for all modules N. ⌟

Here is the punch line. In §VIII.6.4 we had claimed that *flat* resolutions could be used in place of free or projective resolutions, in order to compute Tor. We are going to verify that *$\mathscr{F}$-acyclic resolutions suffice in order to compute* $\mathsf{L}_i\mathscr{F}$.

Theorem 8.3. *Let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be a right-exact functor of abelian categories, and assume A has enough projectives. Let*

$$A^\bullet : \quad \cdots \longrightarrow A^{-2} \xrightarrow{d_A^{-2}} A^{-1} \xrightarrow{d_A^{-1}} A^0 \longrightarrow 0 \longrightarrow \cdots$$

be a resolution of an object M of A, such that every object A^i is $\mathscr{F}$-acyclic. Then for all i

$$\mathsf{L}_i\mathscr{F}(M) \cong H^{-i}(\mathsf{C}(\mathscr{F})(A^\bullet)).$$

If the objects A^i are projective, the statement is the very definition of $\mathsf{L}_i\mathscr{F}$. The theorem states that $\mathscr{F}$-acyclic objects may be used in place of projective objects in the computation of $\mathsf{L}_i\mathscr{F}$.

Proof. The first two cases follow from explicit computations. To begin with, the sequence

$$A^{-1} \longrightarrow A^0 \longrightarrow M \longrightarrow 0$$

is exact by assumption, and $\mathscr{F}$ is right-exact; thus

$$\mathscr{F}(A^{-1}) \longrightarrow \mathscr{F}(A^0) \longrightarrow \mathscr{F}(M) \longrightarrow 0$$

is exact, and it follows that

$$H^0(\mathsf{C}(\mathscr{F})(A^\bullet)) \cong \mathscr{F}(M) \cong \mathsf{L}_0\mathscr{F}(M).$$

Next, let K be the kernel of $A^0 \to M$; so we have the short exact sequence

$$(*) \qquad 0 \longrightarrow K \longrightarrow A^0 \longrightarrow M \longrightarrow 0\ .$$

Applying Theorem 7.12 (and Proposition 7.13), we obtain a long exact sequence ending with

$$\mathsf{L}_1\mathscr{F}(A^0) \longrightarrow \mathsf{L}_1\mathscr{F}(M) \longrightarrow \mathscr{F}(K) \longrightarrow \mathscr{F}(A^0) \longrightarrow \mathscr{F}(M) \longrightarrow 0\ .$$

The leftmost term is 0 since A^0 is $\mathscr{F}$-acyclic. It follows that

$$\mathsf{L}_1\mathscr{F}(M) \cong \ker(\mathscr{F}(K) \to \mathscr{F}(A^0)) \cong H^{-1}(\mathsf{C}(\mathscr{F})(A^\bullet))\ :$$

for the second $\cong$, apply the result of (the straightforward) Exercise 8.1 to the exact complex

$$\cdots \longrightarrow A^{-1} \longrightarrow A^0 \longrightarrow M \longrightarrow 0 \longrightarrow \cdots\ .$$

Finally, use induction on i. We have just verified that the theorem holds for all objects M and for $i = 0, 1$; given $i > 1$, assume the statement is known for all objects of A and all indices $< i$. Truncating/shifting $A^\bullet$,

$$\cdots \longrightarrow A^{-3} \longrightarrow A^{-2} \longrightarrow A^{-1} \longrightarrow 0 \longrightarrow \cdots$$

(so that A^{-1} is placed in degree 0), gives an $\mathscr{F}$-acyclic resolution of K; by the induction hypothesis, $\mathsf{L}_{i-1}\mathscr{F}(K)$ may be computed by applying $\mathsf{C}(\mathscr{F})$ to this complex and taking cohomology:

$$\mathsf{L}_{i-1}\mathscr{F}(K) \cong H^{-i}(A^\bullet).$$

On the other hand, the other terms in the long exact sequence obtained by applying Theorem 7.12 to (*) give

$$\cdots \longrightarrow \mathsf{L}_i\mathscr{F}(A^0) \longrightarrow \mathsf{L}_i\mathscr{F}(M) \longrightarrow \mathsf{L}_{i-1}\mathscr{F}(K) \longrightarrow \mathsf{L}_{i-1}\mathscr{F}(A^0) \longrightarrow \cdots\ ;$$

since A^0 is $\mathscr{F}$-acyclic and $i > 1$, this shows that

$$\mathsf{L}_i\mathscr{F}(M) \cong \mathsf{L}_{i-1}\mathscr{F}(K)$$

(also cf. Exercise 7.12) and concludes the proof. □

Example 8.4. Let R be a commutative ring, and let

$$F_\bullet : \quad \cdots \longrightarrow F_2 \longrightarrow F_1 \longrightarrow F_0 \longrightarrow 0$$

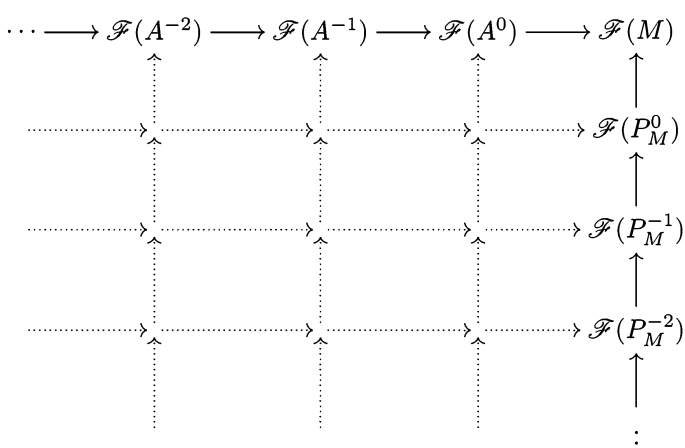
be a resolution of an R-module M by *flat* R-modules. Then for every R-module N,

$$\operatorname{Tor}_i^R(M, N) \cong H_i(F_\bullet \otimes_R N).$$

Indeed, flat modules are acyclic with respect to $\_ \otimes N$ (Example 8.2), so this is now a consequence of Theorem 8.3. ⌟

Theorem 8.3 raises an interesting possibility: since $\mathscr{F}$-acyclic objects suffice in order to compute the derived functors of $\mathscr{F}$ (at least when $\mathscr{F}$ is right-exact), the reader can imagine that there may be situations in which A does *not* have enough projectives, and yet left-derived functors of a functor $\mathscr{F}$ may be defined because A has enough $\mathscr{F}$-acyclic objects (in a suitable sense). This is indeed the case. The reader will run into such examples in more advanced treatments of the subject. (See Exercise 8.14 for a typical situation.)

There is an alternative viewpoint on the question addressed by Theorem 8.3. Let $A^\bullet$ be an $\mathscr{F}$-acyclic resolution of an object M, and let $P^\bullet_M$ be a projective resolution. We will apply $\mathsf{C}(\mathscr{F})$ and place the resulting complexes as sides of an array:



The result of Theorem 8.3 is that taking cohomology of the top row of this diagram gives the same result as taking cohomology of the rightmost column. Might there not be a way to 'interpolate' between these two cohomologies, by cleverly filling in the dotted portion of this diagram?

Double complexes may be used to this effect, as we will see in a moment.

8.2. Complexes of complexes. We warned the reader as far back as §3.2 that we would examine complexes *of complexes*, that is, objects of $\mathsf{C}(\mathsf{C}(\mathsf{A}))$, for an abelian category A. We have run across particular cases: the exact sequences of complexes leading to the long exact cohomology sequence (§3.3) and the construction of the mapping cone (§4.1). The latter case generalizes in a straightforward way, as we will see now.

Keeping with the tradition established earlier in the chapter, we will concentrate on the 'bounded-above' case; the reader should have no difficulty reproducing the bounded-below version of what we will do and should be aware of the fact that the material can be developed to a large extent without boundedness conditions.

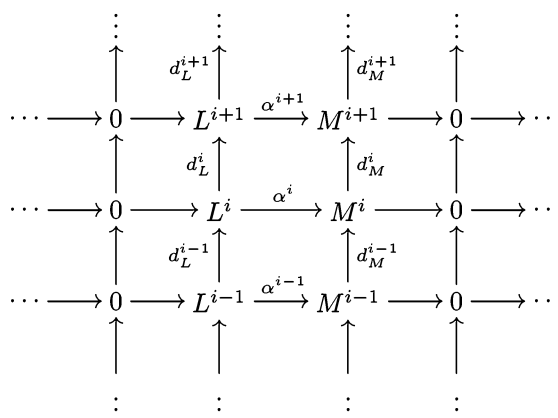
An object of $\mathsf{C}^{\leq 0}(\mathsf{C}^{\leq 0}(\mathsf{A}))$ is a complex

$$\cdots \longrightarrow M^{-3,\bullet} \xrightarrow{d_h^{-3,\bullet}} M^{-2,\bullet} \xrightarrow{d_h^{-2,\bullet}} M^{-1,\bullet} \xrightarrow{d_h^{-1,\bullet}} M^{0,\bullet} \longrightarrow 0 \longrightarrow \cdots$$

in which the object $M^{i,\bullet}$ in degree i is itself a complex in $\mathsf{C}^{\leq 0}(\mathsf{A})$. The subscript h stands, rather unimaginatively, for 'horizontal'; we have $d_h^{i,\bullet} \circ d_h^{i-1,\bullet} = 0$. An example that will motivate the next construction is the particular case in which only two objects are nonzero:

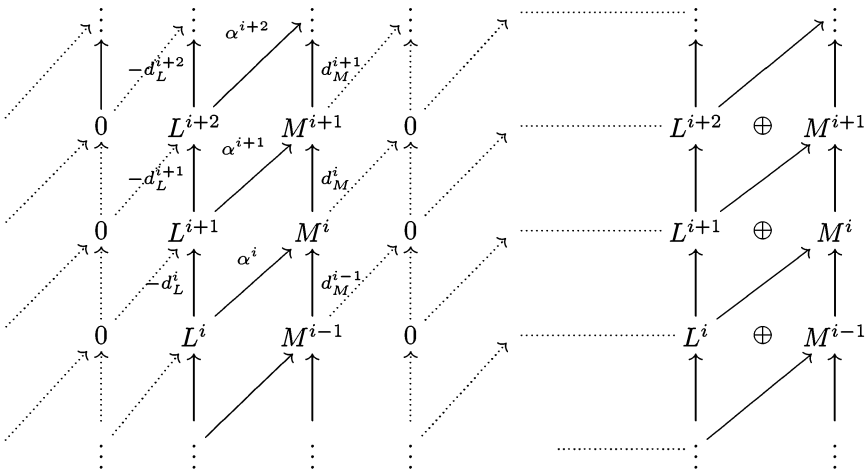
$$(*) \qquad \cdots \longrightarrow 0 \longrightarrow 0 \longrightarrow L^\bullet \xrightarrow{\alpha^\bullet} M^\bullet \longrightarrow 0 \longrightarrow \cdots$$

(and we are considering the particular case in which $L^\bullet$, $M^\bullet$ are bounded above). We will (arbitrarily) place $M^\bullet$ in degree 0 and $L^\bullet$ in degree -1. We can view the information carried by (*) as a commutative diagram:



The mapping cone $MC(\alpha)^\bullet$ constructed in §4.1 is obtained by

- shifting the complex in degree $-i$ by i,
- direct-summing the rows and morphisms of the resulting diagram, collapsing it to a single vertical complex:

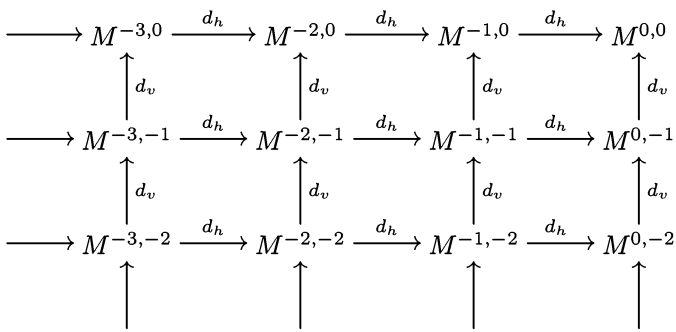


Note that, due to the shift, the diagram on the left has become *anti*commutative in the process. The extra signs are necessary in order that $MC(\alpha)^\bullet$ be a complex.

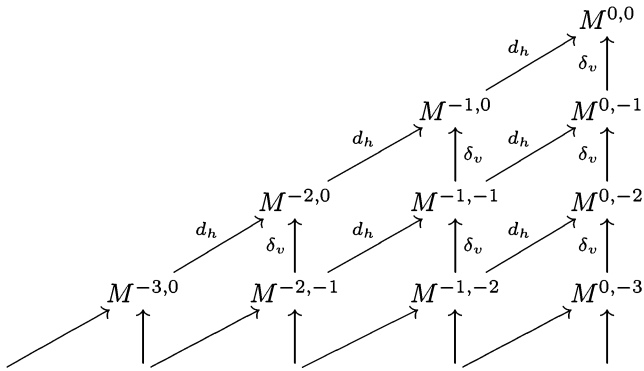
Now we will apply the same procedure to the more general case

$$\cdots \longrightarrow M^{-3,\bullet} \xrightarrow{d_h^{-3,\bullet}} M^{-2,\bullet} \xrightarrow{d_h^{-2,\bullet}} M^{-1,\bullet} \xrightarrow{d_h^{-1,\bullet}} M^{0,\bullet} \longrightarrow 0 \longrightarrow \cdots .$$

This corresponds to the commutative diagram



surrounded by 0 above and to the right (omitted here not to clutter the diagram; the upper indices of the differentials are sacrificed for the same reason); the subscript v stands for 'vertical'. As above, we shift $M^{-i,\bullet}$ by i, with the effect of changing the sign of the differentials in the odd-shifted columns; we will let δ_v denote d_v on even-degree columns and $-d_v$ on odd-degree columns. The resulting diagram is anticommutative:



Taking direct sums of the objects and morphisms on each row, as in the case of the mapping cone, determines a new complex, called the *total complex* $TC(M)^\bullet$ of the given complex of complexes.

Double complexes capture this construction:

Definition 8.5. A *double (cochain) complex* on an abelian category A is an array of objects $M^{i,j}$ of A, $i, j \in \mathbb{Z}$, endowed with morphisms

$$d_h^{i,j} : M^{i,j} \to M^{i+1,j}, \quad \delta_v^{i,j} : M^{i,j} \to M^{i,j+1}$$

such that

$$\begin{cases} d_h^{i,j} \circ d_h^{i-1,j} = 0, \\ \delta_v^{i,j} \circ \delta_v^{i,j-1} = 0, \\ \delta_v^{i+1,j} \circ d_h^{i,j} + d_h^{i,j+1} \circ \delta_v^{i,j} = 0. \end{cases}$$ ⌟

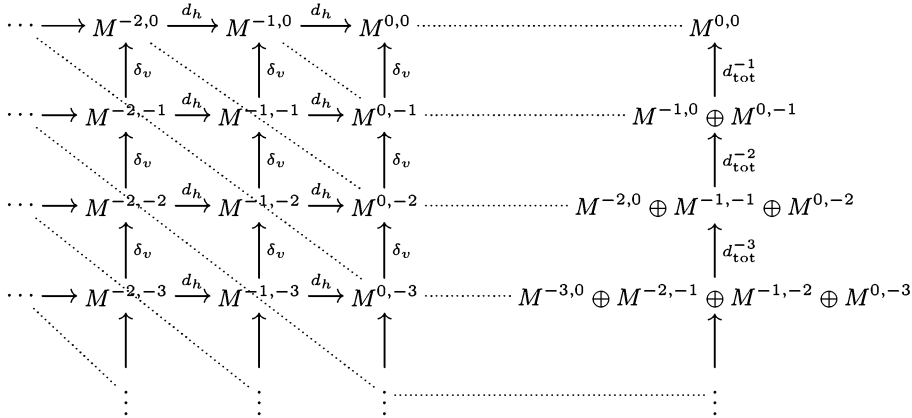
That is, the columns $M^{i,\bullet}$ and the rows $M^{\bullet,j}$ both form complexes, and the whole diagram *anti*commutes. Dropping the position specification, the conditions defining a double complex are summarized in the easier-to-parse prescription

$$\begin{cases} d_h \circ d_h = 0, \\ \delta_v \circ \delta_v = 0, \\ \delta_v \circ d_h + d_h \circ \delta_v = 0. \end{cases}$$

Double complexes, with or without appropriate boundedness conditions (such as requiring $M^{i,j} = 0$ for $i > 0$, $j > 0$, as in the case considered above), form a category in an evident way. It is equally evident that this category is equivalent to the correspondingly bounded category of complexes of complexes: every object of $\mathsf{C}^{\leq 0}(\mathsf{C}^{\leq 0}(\mathsf{A}))$ determines a double complex by setting $\delta_v^{i,j} = (-1)^i d_v^{i,j}$, as above. Changing the signs of odd *rows* rather than columns leads to an isomorphic double complex, Exercise 8.4; in fact, other choices may be convenient depending on the situation.

From this viewpoint, the total complex $TC(M)^\bullet$ (also called the *simple* of the *double* complex $M^{\bullet,\bullet}$) is obtained by direct-summing objects and morphisms of the

double complex along diagonals:



Definition 8.6. The *total complex* $TC(M)^\bullet$ of a double complex $M^{\bullet,\bullet}$ is defined by setting $TC(M)^k := \bigoplus_{i+j=k} M^{i,j}$, with differentials (written stenographically) $d_{\text{tot}} = d_h + \delta_v$. ⌟

Claim 8.7. *$(TC(M)^\bullet, d_{tot}^\bullet)$ is indeed a complex.*

Proof. The question is whether the composition of two consecutive differentials vanishes, as it should. Insisting with our shorthand,

$$d_{\text{tot}} \circ d_{\text{tot}} = (d_h + \delta_v) \circ (d_h + \delta_v) = d_h \circ d_h + d_h \circ \delta_v + \delta_v \circ d_h + \delta_v \circ \delta_v = 0.$$

The reader will formalize this computation (Exercise 8.3). □

It is clear from the definition that $TC(M)^\bullet$ agrees with the mapping cone, when $M^{\bullet,\bullet}$ is concentrated in degrees -1 and 0.

The total complex is defined even without any boundedness hypothesis, but then one has to choose whether to use direct *sums* or direct *products* as objects of the complex. Since only bounded complexes will be needed in the applications we will see, we will happily leave such complications aside and limit the discussion to the bounded case.

Example 8.8. Operations such as Hom_A or $\otimes$ determine double complexes. For a 'first quadrant' example, let $L^\bullet$, resp., $M^\bullet$, be a complex in $\mathsf{C}^{\leq 0}(\mathsf{A})$, resp. $\mathsf{C}^{\geq 0}(\mathsf{A})$:

$$\cdots \longrightarrow L^{-2} \longrightarrow L^{-1} \longrightarrow L^0 \longrightarrow 0 \longrightarrow \cdots$$

$$\cdots \longrightarrow 0 \longrightarrow M^0 \longrightarrow M^1 \longrightarrow M^2 \longrightarrow \cdots .$$

We can consider the complex in $\mathsf{C}^{\geq 0}(\mathsf{C}^{\geq 0})$:

$$\cdots \longrightarrow 0 \longrightarrow \text{Hom}_\mathsf{A}(L^\bullet, M^0) \longrightarrow \text{Hom}_\mathsf{A}(L^\bullet, M^1) \longrightarrow \text{Hom}_\mathsf{A}(L^\bullet, M^2) \longrightarrow \cdots ,$$

that is, in 'grid' form (and omitting zeros to the left and below):

$$\begin{array}{ccccccc}
 & \vdots & & \vdots & & \vdots & \\
 & \uparrow & & \uparrow & & \uparrow & \\
\cdots & \mathrm{Hom}_{\mathsf{A}}(L^{-2},M^0) & \longrightarrow & \mathrm{Hom}_{\mathsf{A}}(L^{-2},M^1) & \longrightarrow & \mathrm{Hom}_{\mathsf{A}}(L^{-2},M^2) & \longrightarrow \cdots \\
 & \uparrow & & \uparrow & & \uparrow & \\
\cdots & \mathrm{Hom}_{\mathsf{A}}(L^{-1},M^0) & \longrightarrow & \mathrm{Hom}_{\mathsf{A}}(L^{-1},M^1) & \longrightarrow & \mathrm{Hom}_{\mathsf{A}}(L^{-1},M^2) & \longrightarrow \cdots \\
 & \uparrow & & \uparrow & & \uparrow & \\
\cdots & \mathrm{Hom}_{\mathsf{A}}(L^{0},M^0) & \longrightarrow & \mathrm{Hom}_{\mathsf{A}}(L^{0},M^1) & \longrightarrow & \mathrm{Hom}_{\mathsf{A}}(L^{0},M^2) & \longrightarrow \cdots
\end{array}$$

We could also consider the complex of complexes

$$\cdots \longrightarrow 0 \longrightarrow \mathrm{Hom}_{\mathsf{A}}(L^0,M^\bullet) \longrightarrow \mathrm{Hom}_{\mathsf{A}}(L^{-1},M^\bullet) \longrightarrow \mathrm{Hom}_{\mathsf{A}}(L^{-2},M^\bullet) \longrightarrow \cdots$$

and this would lead to the grid obtained by flipping the previous one about the main diagonal. The two corresponding total complexes are then clearly isomorphic (Exercise 8.4), and by a slight abuse of language they can both be denoted $TC(\mathrm{Hom}_{\mathsf{A}}(L,M))^\bullet$. The degree-$i$ piece of this total complex, i.e.,

$$\mathrm{Hom}_{\mathsf{A}}(L^{-i},M^0) \oplus \mathrm{Hom}_{\mathsf{A}}(L^{-i+1},M^1) \oplus \cdots \oplus \mathrm{Hom}_{\mathsf{A}}(L^0,M^i),$$

parametrizes degree-preserving morphisms from objects of $L^\bullet$ to objects of $M[i]^\bullet$. These are not cochain morphisms $L^\bullet \to M[i]^\bullet$, since the corresponding diagrams

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & 0 & \longrightarrow & M[i]^{-i} & \longrightarrow & M[i]^{-i+1} & \longrightarrow \cdots \longrightarrow & M[i]^0 & \longrightarrow & M[i]^1 \longrightarrow \cdots \\
 & & \uparrow & & \uparrow & & \uparrow & & \uparrow & & \uparrow \\
\cdots & \longrightarrow & L^{-i-1} & \longrightarrow & L^{-i} & \longrightarrow & L^{-i+1} & \longrightarrow \cdots \longrightarrow & L^0 & \longrightarrow & 0 \longrightarrow \cdots
\end{array}$$

do not satisfy any commutativity hypothesis. The reader will verify (Exercise 8.5) that, with suitable positions:

- the kernel of d^i_{tot} consists of morphisms of cochain complexes $L^\bullet \to M[i]^\bullet$;
- the image of d^{i-1}_{tot} consists of morphisms that are homotopy equivalent to 0; and hence
- the cohomology of $TC(\mathrm{Hom}_{\mathsf{A}}(L,M))^\bullet$ parametrizes morphisms in the homotopy category.

For (a trivial) example, $TC(\mathrm{Hom}_{\mathsf{A}}(L,M))^0 = \mathrm{Hom}_{\mathsf{A}}(L^0,M^0)$, and $d^0_{\mathrm{tot}}(\alpha) = 0$ means that the images $\alpha \circ d^{-1}_{L^\bullet}$ in $\mathrm{Hom}_{\mathsf{A}}(L^{-1},M^0)$ and $d^0_{M^\bullet} \circ \alpha$ in $\mathrm{Hom}_{\mathsf{A}}(L^0,M^1)$ both vanish:

$$\begin{array}{ccccccc}
\longrightarrow & L^{-1} & \xrightarrow{d^{-1}_{L^\bullet}} & L^0 & \longrightarrow & 0 & \longrightarrow \\
 & \downarrow & \searrow^{0} & \downarrow{\scriptstyle \alpha} & \searrow^{0} & \downarrow & \\
\longrightarrow & 0 & \longrightarrow & M^0 & \xrightarrow[d^0_{M^\bullet}]{} & M^1 & \longrightarrow
\end{array}$$

This is precisely the condition needed for α to define a morphism of cochain complexes $L^\bullet \to M^\bullet$. It follows that in this situation $H^0(TC(\mathrm{Hom}_{\mathsf{A}}(L,M))^\bullet)$ is simply $\mathrm{Hom}_{\mathsf{C}(\mathsf{A})}(L^\bullet, M^\bullet) = \mathrm{Hom}_{\mathsf{K}(\mathsf{A})}(L^\bullet, M^\bullet)$ (there are no nontrivial homotopies in this case).

The identification of the i-th cohomology of the total complex with the Hom-set $\mathrm{Hom}_{\mathsf{K}(\mathsf{A})}(L^\bullet, M[i]^\bullet)$ requires care with the signs[34]. ⌟

8.3. Exactness of the total complex. The cohomology of the total complex of a double complex can be computed by an extremely clever device called *spectral sequence,* which we will (briefly) encounter in §9.3. There are standard situations, however, in which one can conclude immediately that the total complex is *exact,* and this is at the root of the applications we will present. The statement is neat and memorable:

Theorem 8.9. *Let* A *be an abelian category, and let*

$$(*) \qquad \cdots \longrightarrow M^{-3,\bullet} \longrightarrow M^{-2,\bullet} \longrightarrow M^{-1,\bullet} \longrightarrow M^{0,\bullet} \longrightarrow 0 \longrightarrow \cdots$$

be a complex in $\mathsf{C}^{\leq 0}(\mathsf{C}^{\leq 0}(\mathsf{A}))$. *Let* $T^\bullet$ *be the corresponding total complex. Then*

- *if (*) is exact, then* $T^\bullet$ *is exact;*
- *if each complex* $M^{i,\bullet}$ *is exact, then* $T^\bullet$ *is exact.*

The proof will yield a more precise statement on the vanishing of individual degrees of the cohomology of the total complex (Exercise 8.6).

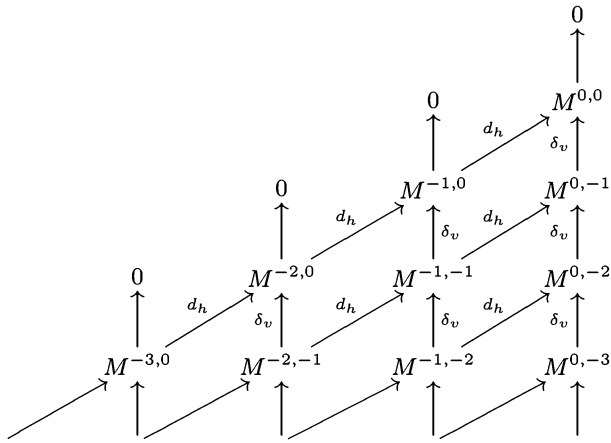
Mutatis mutandis, Theorem 8.9 holds for complexes in $\mathsf{C}^{\geq 0}(\mathsf{C}^{\geq 0}(\mathsf{A}))$, with microscopic adjustments to the proof; we will take this for granted in the applications. In fact, the boundedness hypothesis can be relaxed: one of the two given exactness statements will still hold if only one of the C is bounded. The point is that the proof relies on an inductive argument, which goes through as soon as it has a place to start. The reader is welcome to explore all these ramifications.

In terms of double complexes, Theorem 8.9 says that the total complex of a (suitably bounded) double complex is exact if the rows or the columns of the double complex are exact.

Proof. It is enough to prove the second statement: the first one follows by flipping the double complex corresponding to (*) (cf. Exercise 8.4).

[34] A clever way out of the sign quagmire in this computation is to choose another way to get a double complex out of $\mathrm{Hom}_{\mathsf{A}}(L^\bullet, M^\bullet)$. For example, replacing the vertical differentials $d_v^{i,j}$ by $(-1)^{i+j+1} d_v^{i,j}$ still makes the array anticommutative as needed and irons out annoying sign mismatches elsewhere.

To prove the second statement, we assume that the columns in the staggered (anticommutative) diagram



associated with (*) are exact and that we have an element

$$m = (m^0, m^{-1}, \cdots, m^{-i}) \in M^{-i,0} \oplus M^{-i+1,-1} \oplus \cdots \oplus M^{0,-i}$$

such that $d_{\mathrm{tot}}^{-i}(m) = 0$; we are expected to produce an element

$$n = (n^0, n^{-1}, \cdots, n^{-i-1}) \in M^{-i-1,0} \oplus M^{-i,-1} \oplus M^{-i+1,-2} \oplus \cdots \oplus M^{0,-i-1}$$

such that $d_{\mathrm{tot}}^{-i-1}(n) = m$.

The little induction that accomplishes this is conceptually very simple, but notationally challenging. We believe that viewing it in action in a simple case (we will choose $i = 2$) suffices to convey all that needs to be conveyed. Thus, assume that $m = (m^0, m^{-1}, m^{-2})$ is in the kernel of d_{tot}; that is,

$$\begin{cases} d_h(m^0) + \delta_v(m^{-1}) = 0, \\ d_h(m^{-1}) + \delta_v(m^{-2}) = 0. \end{cases}$$

We choose $n^0 = 0$. Since $M^{-2,\bullet}$ is exact, there exists $n^{-1} \in M^{-2,-1}$ such that $\delta_v(n^{-1}) = m^0$. Since the diagram anticommutes,

$$\delta_v \circ d_h(n^{-1}) = -d_h \circ \delta_v(n^{-1}) = -d_h(m^0) = \delta_v(m^{-1}).$$

Therefore, $m^{-1} - d_h(n^{-1}) \in \ker \delta_v$, and since $M^{-1,\bullet}$ is exact, there exists $n^{-2} \in M^{-1,-2}$ such that $\delta_v(n^{-2}) = m^{-1} - d_h(n^{-1})$. That is,

$$d_h(n^{-1}) + \delta_v(n^{-2}) = m^{-1}.$$

The next step is entirely analogous:

$$\delta_v \circ d_h(n^{-2}) = -d_h \circ \delta_v(n^{-2}) = -d_h(m^{-1} - d_h(n^{-1})) = -d_h(m^{-1}) = \delta_v(m^{-2})$$

(using the fact that $d_h \circ d_h = 0$); therefore $m^{-2} - d_h(n^{-2}) \in \ker \delta_v$, and since $M^{0,\bullet}$ is exact, there exists $n^{-3} \in M^{0,-2}$ such that $\delta_v(n^{-3}) = m^{-2} - d_h(n^{-2})$. That is,

$$d_h(n^{-2}) + \delta_v(n^{-3}) = m^{-2}.$$

With these choices, $d_{\mathrm{tot}}(n)$ equals

$$d_h(n^0) + \delta_v(n^{-1}) + d_h(n^{-1}) + \delta_v(n^{-2}) + d_h(n^{-2}) + \delta_v(n^{-3}) = 0 + m^0 + m^{-1} + m^{-2} = m$$

as needed.

The case for arbitrary i simply requires additional repetitions of the basic steps performed for $i = 2$. □

Example 8.10. Here is a taste of how convenient Theorem 8.9 is. Let $P^\bullet$ be a complex in $\mathsf{C}^{\leq 0}(\mathsf{A})$, where each P^i is projective, and let $L^\bullet$ be an *exact* complex in $\mathsf{C}^{\geq 0}(\mathsf{A})$. Since P^i is projective, $\mathrm{Hom}_{\mathsf{A}}(P^i, L^\bullet)$ is still exact. Therefore, the complex

$$\cdots \longrightarrow 0 \longrightarrow \mathrm{Hom}_{\mathsf{A}}(P^0, L^\bullet) \longrightarrow \mathrm{Hom}_{\mathsf{A}}(P^{-1}, L^\bullet) \longrightarrow \mathrm{Hom}_{\mathsf{A}}(P^{-2}, L^\bullet) \longrightarrow \cdots$$

satisfies the hypothesis of the second statement given in Theorem 8.9 (in the symmetrically bounded version). According to Theorem 8.9, the corresponding total complex is exact; as seen in Example 8.8 (cf. Exercise 8.5), this says that $\mathrm{Hom}_{\mathsf{K}(\mathsf{A})}(P^\bullet, L[i]^\bullet) = 0$ for all i. In other words, *every cochain morphism from a bounded-above complex of projectives to a bounded-below exact complex is homotopic to* 0. This recovers Corollary 5.12, up to easy adjustments, taking care of the boundedness hypothesis on $L^\bullet$. ⌟

8.4. Total complexes and resolutions. Example 8.10 illustrates well the power of results such as Theorem 8.9: the nuts-and-bolts-style arguments used in §5 are bypassed, and statements such as Corollary 5.12 are placed in a clarifying context[35]. The same strategy will allow us to revisit the question examined in §8.1, letting a double complex draw the conclusion of Theorem 8.3. We will then finally be able to go back to the last remaining mystery concerning Tor and Ext, dealing with why we can compute these functors 'by resolving either argument'. This will conclude a large circle of ideas, begun in Chapter VIII.

These applications rely on one more remark concerning double complexes and their simples. The heavy notation poses a substantial obstacle in appreciating what follows; the reader should take comfort in our assurance that notation is the *only* obstacle here—the constructions are actually straightforward.

We have studied in §6.2 the problem of constructing complexes of projectives (for example) that are quasi-isomorphic to a given complex $N^\bullet$: Theorem 6.6 accomplishes this. We have called the corresponding quasi-isomorphism a (projective) *resolution* of $N^\bullet$. Now that we are working in $\mathsf{C}^{\leq 0}(\mathsf{C}^{\leq 0}(\mathsf{A}))$, this looks like an even worse abuse of language: viewing a (bounded-above) complex $N^\bullet$ as an object of the abelian category $\mathsf{A}' = \mathsf{C}^{\leq 0}(\mathsf{A})$, a (bounded-above) *resolution* of $N^\bullet$ ought to be a complex $M^{\bullet,\bullet}$ in $\mathsf{C}^{\leq 0}(\mathsf{A}')$ endowed with a quasi-isomorphism

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & M^{-2,\bullet} & \xrightarrow{d_h} & M^{-1,\bullet} & \xrightarrow{d_h} & M^{0,\bullet} & \longrightarrow & 0 & \longrightarrow & \cdots \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\
\cdots & \longrightarrow & 0 & \longrightarrow & 0 & \longrightarrow & N^\bullet & \longrightarrow & 0 & \longrightarrow & \cdots
\end{array}$$

[35] But we strongly believe in the pedagogic usefulness of first going through the nuts and bolts!

in $\mathsf{C}^{\leq 0}(\mathsf{A}')$. In other words, a resolution of $N^\bullet$ should be a complex $M^{\bullet,\bullet}$ whose cohomology (computed in $\mathsf{C}^{\leq 0}(\mathsf{A}')$) is concentrated in degree 0 and agrees with $N^\bullet$.

We are going to verify that these two notions of 'resolution' of a complex *are* compatible (so no language abuse occurs after all): taking the total complex of a resolution $M^{\bullet,\bullet}$ of $N^\bullet$ (in the 'new' sense of the term) again gives a resolution of $N^\bullet$ (in the 'old' sense).

Before stating this more explicitly, consider more generally any cochain morphism *in*[36] $\mathsf{C}^{\leq 0}(\mathsf{C}^{\leq 0}(\mathsf{A}))$ from $M^{\bullet,\bullet}$ to a complex (of complexes) concentrated in degree 0:

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & M^{-2,\bullet} & \xrightarrow{d_h} & M^{-1,\bullet} & \xrightarrow{d_h} & M^{0,\bullet} & \longrightarrow & 0 & \longrightarrow & \cdots \\
 & & \downarrow & & \downarrow & & \downarrow{\scriptstyle \alpha^\bullet} & & \downarrow & & \\
\cdots & \longrightarrow & 0 & \longrightarrow & 0 & \longrightarrow & N^\bullet & \longrightarrow & 0 & \longrightarrow & \cdots
\end{array}$$

This means that $\alpha^\bullet \circ d_h = 0$, so we can fold this diagram into a single complex of complexes:

$$M_N^{\bullet,\bullet}: \qquad \cdots \longrightarrow M^{-2,\bullet} \xrightarrow{-d_h} M^{-1,\bullet} \xrightarrow{-d_h} M^{0,\bullet} \xrightarrow{\alpha^\bullet} N^\bullet \longrightarrow 0 \longrightarrow \cdots .$$

That is, $M_N^{\bullet,\bullet}$ agrees with the complex obtained by shifting $M^{\bullet,\bullet}$ by one step to the left, with $N^\bullet$ inserted in degree 0. For now, we are making no assumptions on the cohomology of $M^{\bullet,\bullet}$.

We want to compare the various total complexes that can be defined in this situation. We can map the total complex of $M^{\bullet,\bullet}$ to $N^\bullet$,

$$\begin{array}{ccccccc}
 & & & & 0 & \longrightarrow & 0 \\
 & & & & \uparrow & & \uparrow \\
 & & & & M^{0,0} & \longrightarrow & N^0 \\
 & & & \nearrow & \uparrow & & \uparrow \\
 & & M^{-1,0} & \oplus & M^{0,-1} & \longrightarrow & N^{-1} \\
 & \nearrow & \uparrow & \nearrow & \uparrow & & \uparrow \\
M^{-2,0} & \oplus & M^{-1,-1} & \oplus & M^{0,-2} & \longrightarrow & N^{-2} \\
\uparrow & \nearrow & \uparrow & \nearrow & \uparrow & & \uparrow \\
 & & & & & &
\end{array}$$

obtaining a morphism of cochain complexes[37]

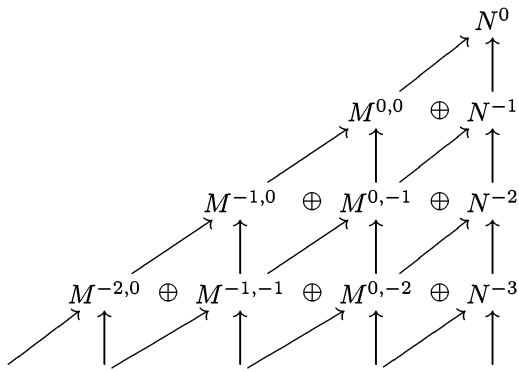
$$TC(\alpha)^\bullet: \qquad TC(M^{\bullet,\bullet}) \longrightarrow N^\bullet .$$

(This morphism is of course just an instance of the evident functoriality of the construction of the total complex.)

[36] Of course a parallel discussion can be carried out in $\mathsf{C}^{\geq 0}(\mathsf{C}^{\geq 0}(\mathsf{A}))$, and with other boundedness possibilities. This is left to the reader as usual.

[37] Note that, e.g., $M^{-1,-1} \to M^{0,-1} \to N^{-1}$ is the zero-morphism (because $\alpha^\bullet \circ d_h = 0$). The morphism from $TC(M^{\bullet,\bullet})$ to $N^\bullet$ only sees $M^{0,\bullet}$, so it clearly *is* a morphism of cochain complexes.

Or, we can shift the M-part of this diagram and view the last display as the construction of the total complex *of* $M_N^{\bullet,\bullet}$:



This involves a sign change in the vertical differentials of the M-part, due to the shift.

Claim 8.11. *The total complex of* $M_N^{\bullet,\bullet}$ *is the mapping cone of* $TC(\alpha)^\bullet$*:*

$$TC(M_N)^\bullet = MC(TC(\alpha))^\bullet.$$

Verifying this claim only amounts to chasing the definitions, so it can safely be left to the reader (Exercise 8.7). The following consequence compares the different notions of 'resolution', as promised a moment ago.

Theorem 8.12. *Let* A *be an abelian category, and denote by* A' *the category* $\mathsf{C}^{\leq 0}(\mathsf{A})$*. Let* $N^\bullet$ *be an object of* A'*, and let*

$$(*) \qquad \cdots \longrightarrow M^{-3,\bullet} \longrightarrow M^{-2,\bullet} \longrightarrow M^{-1,\bullet} \longrightarrow M^{0,\bullet} \longrightarrow 0 \longrightarrow \cdots$$

be a complex in $\mathsf{C}^{\leq 0}(\mathsf{A}')$*, with total complex* $T^\bullet$*.*

- *Assume that (*) is a resolution of* $N^\bullet$ *in* $\mathsf{C}^{\leq}(\mathsf{A}')$*. Then* $T^\bullet$ *is quasi-isomorphic to* $N^\bullet$ *in* $\mathsf{C}^{\leq 0}(\mathsf{A})$ *(that is, it is a resolution of* $N^\bullet$ *in the sense used in* §6.2 *and following).*
- *Assume that the cohomology of each* $M^{i,\bullet}$ *is concentrated in degree* 0*. Then* $T^\bullet$ *is quasi-isomorphic to the complex of cohomology objects induced by (*):*

$$\cdots \longrightarrow H^0(M^{-3,\bullet}) \longrightarrow H^0(M^{-2,\bullet}) \longrightarrow H^0(M^{-1,\bullet}) \longrightarrow H^0(M^{0,\bullet}) \longrightarrow 0 \longrightarrow \cdots .$$

The statement of Theorem 8.12 extends that of Theorem 8.9; its proof is a nearly immediate consequence of the latter. As we mentioned at the beginning of §8.3, these statements are vastly generalized by the machinery of *spectral sequences,* which provide a general framework for the computation of the cohomology of the total complex of a double complex. We will recover Theorem 8.12 once we have learned a bit about spectral sequences in §9.3, but it is not hard (and it is a good exercise) to prove this statement by hand, as we proceed to do.

Proof. The second statement follows from the first, by flipping the corresponding double complex about the main diagonal.

The first statement follows from Theorem 8.9 and Claim 8.11. Indeed, let $M_N^{\bullet,\bullet}$ be the exact complex

$$\cdots \longrightarrow M^{-2,\bullet} \longrightarrow M^{-1,\bullet} \longrightarrow M^{0,\bullet} \longrightarrow N^\bullet \longrightarrow 0 \longrightarrow \cdots;$$

by Theorem 8.9, $TC(M_N)^\bullet$ is exact; by Claim 8.11 this is the mapping cone of the induced morphism $T^\bullet \to N^\bullet$, and it follows (Corollary 4.2) that this morphism is a quasi-isomorphism, as needed. □

To appreciate the usefulness of Theorem 8.12, again in the context of our (hard) work in §6, note that it provides us with a convenient bridge between resolving *objects* (see Lemma 6.3) and resolving *complexes* (as in Theorem 6.6). Indeed, by Corollary 7.9 (which relies on the horseshoe lemma, therefore on resolving individual objects of A) every bounded-above complex $N^\bullet$ in an abelian category with enough projectives may be viewed as the complex induced in H^0 from a complex

$$\cdots \longrightarrow P^{-3,\bullet} \longrightarrow P^{-2,\bullet} \longrightarrow P^{-1,\bullet} \longrightarrow P^{0,\bullet} \longrightarrow 0$$

where $P^{i,\bullet}$ is a projective resolution of N^i . The total complex of $P^{\bullet,\bullet}$ is then a projective resolution of $N^\bullet$, by the second statement in Theorem 8.12. This reproduces the main content of Theorem 6.6, bypassing the nastiest details of the proof given in §6.

As mentioned in Remark 7.10, double complexes of projectives (or, working in $\mathsf{C}^{\geq 0}$, injectives) resolving a given complex (and satisfying the additional condition explained in Remark 7.10) are called *Cartan-Eilenberg resolutions.* Their systematic use can simplify the material we presented in §6 but at the price of dealing early with double complexes.

8.5. Acyclic resolutions again and balancing Tor and Ext. Finally we go back to the last mysteries remaining concerning Tor and Ext. As a warm-up (and another illustration of the power of double complexes), go back to the question we studied in §8.1: we ended that subsection by lining up the complexes obtained by applying an additive functor $\mathscr{F}$ to an $\mathscr{F}$-acyclic resolution $A^\bullet$ of an object M *and* to a projective resolution of the same object, as sides of an array. Corollary 7.9 tells us how to fill the rest of the array: start from

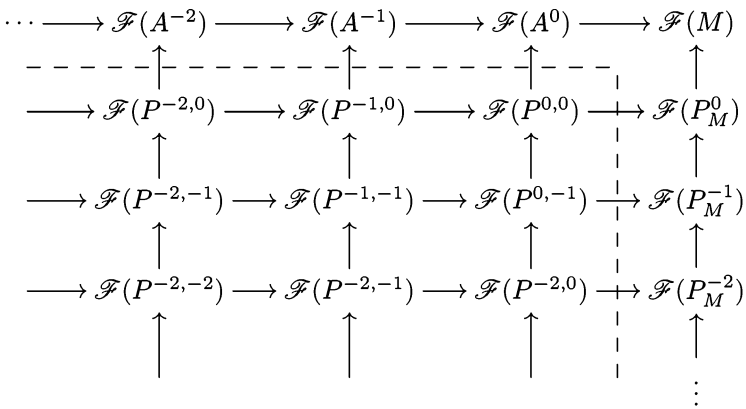
$$(*) \qquad \cdots \longrightarrow A^{-2} \longrightarrow A^{-1} \longrightarrow A^0 \longrightarrow M \longrightarrow 0 \longrightarrow \cdots,$$

which is exact by hypothesis; using Corollary 7.9, view (*) as the complex induced in H^0 by an exact complex

$$(**) \qquad \cdots \longrightarrow P^{-2,\bullet} \longrightarrow P^{-1,\bullet} \longrightarrow P^{0,\bullet} \longrightarrow P_M^\bullet \longrightarrow 0 \longrightarrow \cdots$$

where $P^{i,\bullet}$ is a projective resolution of A^i and $P_M^\bullet$ is a projective resolution of M. Since (**) is exact, the rows of the double complex corresponding to (**) are

projective resolutions of the projective objects P_M^i. Now apply $\mathscr{F}$:



Since each P_M^i is projective and each A^i is $\mathscr{F}$-acyclic, every row and every column of the part of this diagram within the dashed lines is exact. Therefore we can apply Theorem 8.12 to the complex

$$f(P^{\bullet,\bullet}): \qquad \cdots \longrightarrow \mathscr{F}(P^{-2,\bullet}) \longrightarrow \mathscr{F}(P^{-1,\bullet}) \longrightarrow \mathscr{F}(P^{0,\bullet}) \longrightarrow 0 \longrightarrow \cdots$$

and discover that the total complex $TC(\mathscr{F}(P))^\bullet$ is quasi-isomorphic to *both* complexes $\mathsf{C}(\mathscr{F})(A^\bullet)$ and $\mathsf{C}(\mathscr{F})(P_M^\bullet)$. It follows that

$$H^i(\mathsf{C}(\mathscr{F})(A^\bullet)) \cong H^i(\mathsf{C}(\mathscr{F})(P_M^\bullet)) \cong \mathsf{L}_i\mathscr{F}(M),$$

which was the conclusion of Theorem 8.3.

This example illustrates the general strategy of other applications of double complexes: the cohomologies of two complexes are shown to be isomorphic by showing that there is an exact double complex interpolating between them. This strategy finally justifies the claims that Tor and Ext functors can be computed by resolving either one of their arguments.

Theorem 8.13. *Let M, N be modules over a commutative ring R, and let $P_M^\bullet$, resp., $P_N^\bullet$, be projective resolutions of M, resp., N. Then*

$$H^i(P_M^\bullet \otimes_R N) \cong H^i(M \otimes_R P_N^\bullet) \quad (\cong \operatorname{Tor}_i^R(M,N)).$$

The first term was our official definition of Tor, as originally given in §VIII.2.4 (where we used *free* resolutions); see also Example 7.6. Theorem 8.13 makes good on our old promise (also made in §VIII.2.4) to show that the Tor functors could be computed by resolving the second factor rather than the first.

Proof. Apply Theorem 8.12 to the complex

$$(*) \qquad \cdots \longrightarrow P_M^{-2} \otimes_R P_N^\bullet \longrightarrow P_M^{-1} \otimes_R P_N^\bullet \longrightarrow P_M^{-0} \otimes_R P_N^\bullet \longrightarrow 0 \longrightarrow \cdots.$$

Since each P_N^j is projective, the cohomology of this complex is concentrated in degree 0 and equals $M \otimes_R P_N^\bullet$. It follows that

$$TC(P_M^\bullet \otimes_R P_N^\bullet)^\bullet \quad \text{is quasi-isomorphic to} \quad M \otimes_R P_N^\bullet,$$

by the first statement in Theorem 8.12. On the other hand, since each P^i_M is projective, the complex $P^i_M \otimes_R P^\bullet_N$ is a resolution of $P^i_M \otimes N$; that is, the cohomology of each $P^i_M \otimes_R P^\bullet_N$ is concentrated in degree 0 and equals $P^i_M \otimes_R N$. Thus, the complex (*) induces the complex $P^\bullet_M \otimes_R N$ in H^0, and

$$TC(P^\bullet_M \otimes_R P^\bullet_N)^\bullet \quad \text{is quasi-isomorphic to} \quad P^\bullet_M \otimes N,$$

by the second statement in Theorem 8.12. The result follows. □

Theorem 8.14. *Let M, N be modules over a commutative ring R, and let $P^\bullet_M$, resp., $Q^\bullet_N$, be a projective resolution of M, resp., an injective resolution of N. Then*

$$H^i(\mathrm{Hom}_R(P^\bullet_M, N)) \cong H^i(\mathrm{Hom}_R(M, Q^\bullet_N)) \quad (\cong \mathrm{Ext}^i_R(M, N)).$$

The proof of Theorem 8.14 is left to the reader (Exercise 8.9): Example 8.8 and the strategy extensively discussed above will hopefully make this a very easy task.

This statement completes the verification of all claims concerning Tor and Ext made in Chapter VIII, in the sense that the extent to which we have developed the general theory of homological algebra makes those seemingly magical claims now essentially *evident*. Hopefully, the same will apply to other encounters the reader may have with simple applications of homological algebra.

In any case, this seems a fitting place to end the main body of this last chapter. In §9 we will give a brief look at material that really lies beyond the scope of these notes.

Exercises

8.1. ▷ Let A, B be abelian categories, let $A^\bullet$ be an exact complex in $\mathsf{C}(\mathsf{A})$, and let $\mathscr{F}$ be a right-exact functor $\mathsf{A} \to \mathsf{B}$. Let K^i be the kernel of d^i_A. Prove that

$$H^i(\mathsf{C}(\mathscr{F})(A^\bullet)) \cong \ker(\mathscr{F}(K^{i+1}) \to \mathscr{F}(A^{i+1})).$$

[§8.1]

8.2. Let M be an object of an abelian category A with enough projectives, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ be an additive functor. Let $A^\bullet$ be an $\mathscr{F}$-acyclic resolution of M, and let $J^i := \mathrm{im}\, d^i_A$. Prove that

$$\mathsf{L}_i\mathscr{F}(M) \cong \mathsf{L}_1\mathscr{F}(J^{-i+1})$$

for all $i > 1$.

8.3. ▷ Verify carefully that $d_{\mathrm{tot}} \circ d_{\mathrm{tot}} = 0$ (cf. Claim 8.7). [§8.2]

8.4. ▷ We have obtained a double complex from a commutative array of objects $M^{i,j}$ and differentials $d^{\bullet,\bullet}_h$, $d^{\bullet,\bullet}_v$ by 'changing the signs of every other column'. Prove that changing the sign of every other *row* leads to an isomorphic double complex. (In particular, the cohomology of the total complex is essentially unaffected by flipping the original array.) [§8.2, §8.3]

8.5. ▷ With notation as in Example 8.8, prove that the i-th cohomology of the total complex of $\mathrm{Hom}_{\mathsf{A}}(L^\bullet, M^\bullet)$ is isomorphic to $\mathrm{Hom}_{\mathsf{K(A)}}(L^\bullet, M^\bullet[i])$. (Warning: These isomorphisms require careful sign adjustments.) [§8.2, §8.3]

8.6. ▷ Let $M^{\bullet,\bullet}$ be a double complex on an abelian category A, such that (for example) $M^{i,j} = 0$ for $i, j > 0$. Prove that $H^k(TC(M)^\bullet) = 0$ if the rows (or the columns) of $M^{\bullet,\bullet}$ are exact at all $M^{i,j}$ with $i + j = k$. [§8.3]

8.7. ▷ Prove Claim 8.11. [§8.4]

8.8. ¬ Let A, B, C be abelian categories, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$, $\mathscr{G} : \mathsf{B} \to \mathsf{C}$ be additive functors. Assume that A and B have enough projectives, that $\mathscr{F}$ sends projectives of A to $\mathscr{G}$-acyclic objects of B, and that $\mathscr{G}$ is right-exact.

Let A be an object of A, and let $M^\bullet$: $\cdots \to M^{-2} \to M^{-1} \to M^0 \to 0$ be a projective resolution of A. Let $P^\bullet_{\mathscr{F}M^\bullet}$ be a Cartan-Eilenberg resolution of the complex $\mathscr{F}M^\bullet$ (as constructed in Corollary 7.9). Consider the complex

$$(\dagger) \qquad \cdots \longrightarrow \mathscr{G}(P^\bullet_{\mathscr{F}M^{-2}}) \longrightarrow \mathscr{G}(P^\bullet_{\mathscr{F}M^{-1}}) \longrightarrow \mathscr{G}(P^\bullet_{\mathscr{F}M^0}) \longrightarrow 0$$

obtained by applying $\mathscr{G}$ to $P^\bullet_{\mathscr{F}M^\bullet}$. Finally, let $T^\bullet$ be the total complex corresponding to (†).

Prove that $H^n(T^\bullet) \cong \mathsf{L}^n(\mathscr{G} \circ \mathscr{F})(A)$.

(Hint: Prove that the cohomology of each term of (†) is concentrated in degree 0, and apply Theorem 8.12.) [§7.2, 7.6, 9.10]

8.9. ▷ Prove Theorem 8.14: Let M, N be modules over a commutative ring R, and let $P^\bullet_M$, resp., $Q^\bullet_N$, be a projective resolution of M, resp., an injective resolution of N; prove that

$$H^i(\mathrm{Hom}_R(P^\bullet_M, N)) \cong H^i(\mathrm{Hom}_R(M, Q^\bullet_N)),$$

completing the justification for the definition of the Ext functors given in §VIII.6.4. [§8.5]

8.10. If an abelian category A has enough projectives and injectives, we can define functors $\mathrm{Ext}^i_{\mathsf{A}}$ as derived functors of $\mathrm{Hom}_{\mathsf{A}}$, by resolving either component (as we did in R-Mod).

Prove that an object P of A is projective if and only if $\mathrm{Ext}^1_{\mathsf{A}}(P, B) = 0$ for all objects B of A, if and only if $\mathrm{Ext}^i_{\mathsf{A}}(P, B) = 0$ for all objects B of A and all $i \geq 1$.

Prove that an object Q of A is injective if and only if $\mathrm{Ext}^1_{\mathsf{A}}(A, Q) = 0$ for all objects A of A, if and only if $\mathrm{Ext}^i_{\mathsf{A}}(A, Q) = 0$ for all objects A of A and all $i \geq 1$.

8.11. ¬ Let A be an abelian category with enough injectives and projectives. The *projective dimension* of an object A of A is the minimum number $d = \mathrm{pd}(A)$ such that there exists a projective resolution

$$\cdots \longrightarrow P^{-d} \longrightarrow P^{-d+1} \longrightarrow \cdots \longrightarrow P^0 \longrightarrow 0 \longrightarrow \cdots$$

of A, if this number exists, or ∞ if A has no finite projective resolution. Prove that the following are equivalent:

(i) $\mathrm{pd}(A) \leq d$.

(ii) If $0 \to K \to P^{-d+1} \to \cdots \to P^0 \to 0$ is a resolution of A and all P^i are projective, then K is also projective.

(iii) $\mathrm{Ext}^n(A, B) = 0$ for all objects B and all $n > d$.

(iv) $\mathrm{Ext}^{d+1}(A, B) = 0$ for all objects B.

(Hint: (ii) $\Longrightarrow$ (i) $\Longrightarrow$ (iii) $\Longrightarrow$ (iv) $\Longrightarrow$ (ii). For the last implication, use 'dimension shifting': adapt Exercise 7.12.) [8.12, 8.13]

8.12. ¬ Let A be an abelian category with enough injectives and projectives. The *injective dimension* of an object B of A is the minimum number $d = \mathrm{id}(B)$ such that there exists an injective resolution

$$\cdots \longrightarrow 0 \longrightarrow Q^0 \longrightarrow \cdots \longrightarrow Q^{d-1} \longrightarrow Q^d \longrightarrow 0 \longrightarrow \cdots$$

of B, if this number exists, or ∞ if B has no finite injective resolution. Prove that the following are equivalent:

(i) $\mathrm{id}(B) \leq d$.

(ii) If $0 \to Q^0 \to \cdots \to Q^{d-1} \to C \to 0$ is a resolution of B and all Q^i are injective, then C is also injective.

(iii) $\mathrm{Ext}^n(A, B) = 0$ for all objects A and all $n > d$.

(iv) $\mathrm{Ext}^{d+1}(A, B) = 0$ for all objects A.

(See Exercise 8.11.) [8.13]

8.13. (See Exercises 8.11 and 8.12.) Let A be an abelian category with enough injectives and projectives. Prove that the supremum of $\mathrm{pd}(A)$, $A \in \mathrm{Obj}(\mathsf{A})$, equals the supremum of $\mathrm{id}(B)$, $B \in \mathrm{Obj}(\mathsf{A})$, and give an Ext interpretation for this number.

This is called the *global dimension* of A. If R is a ring, the global dimension of R is the global dimension of R-Mod. This invariant is of fundamental importance in commutative algebra. The relation between this number and the Krull dimension of R is subtle and important.

8.14. ▷ Let A, B, C be categories. A *bifunctor* $\mathscr{F} : \mathsf{A} \times \mathsf{B} \to \mathsf{C}$ is an assignment of an object $\mathscr{F}(A, B)$ of C for every object A of A and B of B, such that for all objects A in A, B in B, $\mathscr{F}^A := \mathscr{F}(A, \_)$ is a (say, covariant) functor and $\mathscr{F}_B := \mathscr{F}(\_, B)$ is a (say, contravariant) functor; further, for all morphisms $A_1 \to A_2$ in A, $B_1 \to B_2$ in B, the induced diagram

$$\begin{array}{ccc} \mathscr{F}(A_1, B_1) & \longrightarrow & \mathscr{F}(A_1, B_2) \\ \uparrow & & \uparrow \\ \mathscr{F}(A_2, B_1) & \longrightarrow & \mathscr{F}(A_2, B_2) \end{array}$$

is required to commute. (Arrows in this diagram should be directed according to the covariance of $\mathscr{F}$ in its arguments.)

For example, $\mathrm{Hom}_{\mathsf{A}}(\_, \_)$ is a bifunctor in this sense from $\mathsf{A} \times \mathsf{A}$ to Ab. For R a commutative ring, $\_ \otimes_R \_$ is a bifunctor $R\text{-Mod} \times R\text{-Mod} \to R\text{-Mod}$, covariant in both arguments.

Now assume that A, B, C are abelian categories and that $\mathscr{F} : \mathsf{A} \times \mathsf{B} \to \mathsf{C}$ is a bifunctor, such that $\mathscr{F}^A$ is additive and covariant, $\mathscr{F}_B$ is additive and contravariant, and both are left-exact. Also, assume that B has enough injectives.

Say that an object E of A is *$\mathscr{F}$-exact* if the functor $\mathscr{F}^E$ is exact. An '$\mathscr{F}$-exact resolution' of an object A of A is a resolution in $\mathsf{C}^{\leq 0}(\mathsf{A})$ consisting of $\mathscr{F}$-exact objects. The category A has 'enough $\mathscr{F}$-exact objects' if every object A admits an epimorphism $E \to A$ with $\mathscr{F}$-exact source. Assume that this is the case, so that every object of A admits an $\mathscr{F}$-exact resolution.

Note that we are not assuming that A has enough projectives: in principle, therefore, we should not be able to define a right-derived functor for $\mathscr{F}_B$. We are going to look for an adequate replacement, using the fact that $\mathscr{F}_B$ is one component of a bifunctor.

Since B has enough injectives, the functor $\mathscr{F}^A$ can be right-derived for every object A. Define $\mathsf{T}^i\mathscr{F}_B(A)$ to be $\mathsf{R}^i\mathscr{F}^A(B)$; note that $\mathsf{T}^0\mathscr{F}_B(A) = \mathscr{F}^A(B) = \mathscr{F}(A,B) = \mathscr{F}_B(A)$.

- Prove that every exact sequence $0 \to A_1 \to A_2 \to A_3 \to 0$ in A induces a long exact sequence

$$\begin{array}{l} 0 \longrightarrow \mathscr{F}_B(A_3) \longrightarrow \mathscr{F}_B(A_2) \longrightarrow \mathscr{F}_B(A_1) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \delta^0 \\ \mathsf{T}^1\mathscr{F}_B(A_3) \longrightarrow \cdots \longrightarrow \mathsf{T}^i\mathscr{F}_B(A_1) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \delta^i \\ \mathsf{T}^{i+1}\mathscr{F}_B(A_3) \longrightarrow \cdots \end{array}$$

- Prove that the $\mathsf{T}^i\mathscr{F}_B$ form a δ-functor, in the sense of Exercise 7.13.
- Prove that if E is an $\mathscr{F}$-exact object of A, then $\mathsf{T}^i\mathscr{F}_B(E) = 0$ for $i > 0$, and deduce that the $\mathsf{T}^i\mathscr{F}_B$ form a *universal* δ-functor.
- Let $E^\bullet : \cdots \to E^{-2} \to E^{-1} \to E^0 \to A \to 0$ be an $\mathscr{F}$-exact resolution of A. Then $\mathsf{T}^i\mathscr{F}_B \cong H^i(\mathscr{F}_B(E^\bullet))$. (Hint: Mimic the proof of Theorem 8.3.)

The upshot is that the interplay of the two functors $\mathscr{F}^A$ and $\mathscr{F}_B$ allows us to define $\mathsf{T}^i\mathscr{F}_B$ as if it were a right-derived functor, using T-exact objects in place of projectives. This gives a universal δ-functor, which agrees with the standard right-derived functor when A has enough projectives (by Exercise 7.15 and the uniqueness of universal δ-functors). [§8.1]

---

9. Further topics

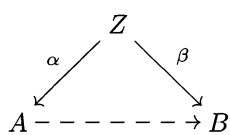
This last section touches upon issues that arise naturally in the context of this chapter and for which a (much) more extensive treatment would be needed. This will hopefully whet the reader's appetite for more and encourage further study of homological algebra beyond the basics covered here. We will be somewhat detailed in §9.3, illustrating the spectral sequence associated to a double complex, since

this can be done reasonably quickly (but delving into any other application of spectral sequences would simply take us too far). For the other themes in this section (derived and triangulated categories) we will really do no more than a little propaganda.

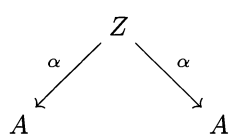
9.1. Derived categories. Derived categories made their appearance in §4.2 and have since served as a guiding concept motivating the natural development of the subject, but we have stopped short of constructing them in general. The closest we have come to a concrete description of the derived category is in the 'poor man's version' of §6.3, which should actually be labeled the 'rich man's version', since it relies on the fortunate presence of enough projectives/injectives. Also, the versions $\mathsf{D}^-(\mathsf{A})$, $\mathsf{D}^+(\mathsf{A})$ considered there are necessarily *bounded,* as they rely on bounded resolutions of objects and complexes in A.

Derived categories may be defined without such restrictions. For an abelian category A, the objects of $\mathsf{D}(\mathsf{A})$ are taken to be the same as in $\mathsf{C}(\mathsf{A})$, that is, cochain complexes; boundedness conditions may be included here, if desired. As for morphisms, remember that the whole point of the derived category is to 'invert quasi-isomorphisms'; the most direct way to do this is essentially the same as the process that allows us to 'invert all nonzero elements of an integral domain', leading to the construction of the field of fractions (see §V.4.2). This process, in a more general version dealing with inverting 'multiplicative subsets' of a ring, is called *localization* (Exercise V.4.7). Categories may be localized as well. For example, the homotopic category $\mathsf{K}(\mathsf{A})$ may be viewed as the localization of $\mathsf{C}(\mathsf{A})$ with respect to homotopy equivalences: these become isomorphisms in $\mathsf{K}(\mathsf{A})$. Localizing $\mathsf{K}(\mathsf{A})$ with respect to homotopy classes of quasi-isomorphisms yields $\mathsf{D}(\mathsf{A})$.

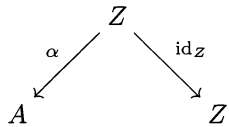
In necessarily oversimplified terms, suppose you want to make a class of morphisms of a category C invertible. In the new category, any 'roof' diagram



with α in that class will determine a morphism $\beta \circ \alpha^{-1} : A \to B$; that is, the dashed arrow will have to exist as a morphism in the new category. These are the analogs of the 'fractions' in the field of fractions; the localization of C may be defined by taking the same objects as in C and setting morphisms in the new category to be compositions of 'roofs', up to a suitable equivalence relation. For example, all roofs



with α in the chosen class should be identified to one another and to id_A in the localized category; the roof

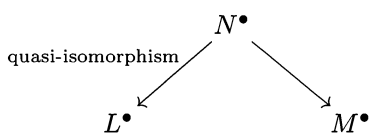


will work as the inverse of α in the localization.

As the reader can imagine, working out the construction in earnest requires dealing with nontrivial technical issues. These are more manageable when the class of morphisms to be inverted is 'localizing' (for example, the composition of two such morphisms should also be a morphism in the class); for example, this condition allows us to represent the composition of two roofs as a roof. To mention one set-theoretic difficulty, one should make sure that morphisms $A \to B$ in the localization still form a *set,* even if the roofs connecting A to B may a priori form a proper *class.*

With all these caveats in mind, the localization process can be carried out on $\mathsf{K}(\mathsf{A})$ with respect to quasi-isomorphisms and does produce a category $\mathsf{D}(\mathsf{A})$ satisfying the universal property for derived categories described in §4.2. If A has (for example) enough projectives, then the bounded version $\mathsf{D}^-(\mathsf{A})$ is equivalent to the homotopy category $\mathsf{K}^-(\mathsf{P})$ of complexes of projectives, as we have endeavored to justify in §6.3. The latter should be viewed as a convenient way to 'compute' the former in favorable circumstances.

Summarizing, a morphism $L^\bullet \to M^\bullet$ in the derived category $\mathsf{D}(\mathsf{A})$ of an abelian category A can be represented as a roof diagram



of (homotopy classes of) morphisms of cochain complexes, modulo a suitable equivalence relation. Note that the construction is applied to $\mathsf{K}(\mathsf{A})$, rather than the simpler-minded $\mathsf{C}(\mathsf{A})$: for one thing, one might as well start from $\mathsf{K}(\mathsf{A})$, since the functor $\mathsf{C}(\mathsf{A}) \to \mathsf{D}(\mathsf{A})$ will have to factor through the homotopic category (Proposition 5.4). In any case it just so happens that, unfortunately, quasi-isomorphisms do not form a localizing class of morphisms in $\mathsf{C}(\mathsf{A})$, while their homotopy classes *are* localizing in $\mathsf{K}(\mathsf{A})$. So it goes.

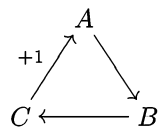
Once we believe that this prescription does define a category—for example, that the composition of two roofs can be represented by a roof and that this composition is associative—then it is essentially clear that $\mathsf{D}(\mathsf{A})$ satisfies the universal property specified in §4.2. However, the real benefits of introducing the derived category lie deeper, for example in its structure as a *triangulated category.* A brief discussion of this concept follows.

The reader should check that the composition of two roofs is indeed a roof (Exercise 9.2); this may be easier after acquiring a little familiarity with distinguished triangles. Also, note that the diagram in Exercise 9.2 does not commute

in $\mathsf{C}(\mathsf{A})$, but it does in $\mathsf{K}(\mathsf{A})$. This is an indication that it would be problematic to go 'directly' from $\mathsf{C}(\mathsf{A})$ to $\mathsf{D}(\mathsf{A})$.

9.2. Triangulated categories. One reason for not treating derived categories to any real extent in this book is that to do so would require us to define precisely the notion of *triangulated category.* The same applies in fact to homotopic categories of complexes. We have pointed out in §5.2 that these categories should not be expected to be (and indeed are not) *abelian,* but they preserve enough structure to make sense of, for example, long exact cohomology sequences. The fact that derived functors ended up having long exact sequences associated to them (§7.4) bears witness to this fact.

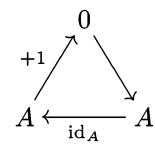
The essential ingredients of a triangulated (additive) category are a 'translation' functor, which in the case of $\mathsf{K}(\mathsf{A})$ or $\mathsf{D}(\mathsf{A})$ is realized as the shift functor $L^\bullet \mapsto L[1]^\bullet$; and a class of diagrams



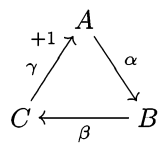
where the morphism labeled '+1' acts as $C \to A[1]$; these diagrams are called *distinguished triangles.* A common alternative notation is

$$A \longrightarrow B \longrightarrow C \xrightarrow{+1} .$$

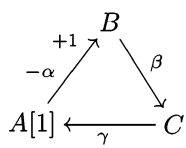
Distinguished triangles are required to satisfy several axioms. Among them,



is required to be distinguished for every A; every morphism $\alpha : A \to B$ must be a side of a distinguished triangle; triangles isomorphic to a distinguished triangle (in the evident sense) must be distinguished; and distinguished triangles can be 'rotated':

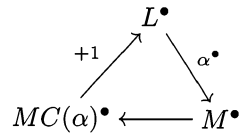


is distinguished if and only if



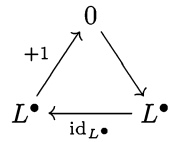
is distinguished. There is more, including an infamous *octahedral axiom* (so named since a popular way to state it invokes an octahedral diagram). The reader will have no difficulties locating this information in the literature.

The category $\mathsf{K}(\mathsf{A})$ is triangulated: the translation functor is the basic shift of a complex, and distinguished triangles are those isomorphic to

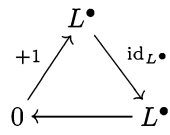


That is, the 'third vertex' of a triangle with an assigned side $\alpha^\bullet : L^\bullet \to M^\bullet$ is the mapping cone of $\alpha^\bullet$, and the unlabeled sides are the natural cochain morphisms $M^\bullet \to MC(\alpha)^\bullet$, $MC(\alpha)\bullet \to L[1]^\bullet$ studied in §4.1.

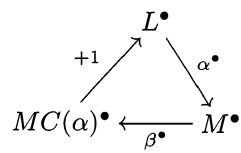
It would be problematic to make the same choice in $\mathsf{C}(\mathsf{A})$, since while



would be distinguished as needed, since $L^\bullet$ is the mapping cone of the zero-morphism $0 \to L^\bullet$, the rotation of this triangle, i.e.,

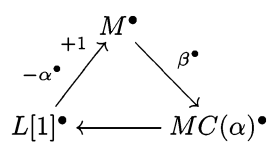


would not seem to be, since the mapping cone of the identity is not 0. *But* the mapping cone of the identity is *homotopically equivalent* to 0 (Exercise 9.3), so the rotated triangle is indeed distinguished in $\mathsf{K}(\mathsf{A})$, as it should be, according to the axioms. More generally, it is easy to see (Exercise 9.4) that if $\alpha^\bullet : L^\bullet \to M^\bullet$ is a cochain morphism, so that we have a distinguished triangle,



then the mapping cone of $\beta^\bullet$ is in fact homotopy equivalent to $L[1]^\bullet$; the rotation

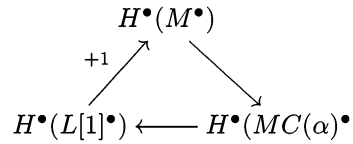
(*)



is distinguished as required, as the reader should check.

Applying the cohomology functor to a distinguished triangle in $\mathsf{K}(\mathsf{A})$ *yields an exact triangle:* indeed, as we have just seen, we may assume that the triangle is (*)

(up to isomorphism in $\mathsf{K}(\mathsf{A})$), and taking cohomology gives the exact triangle



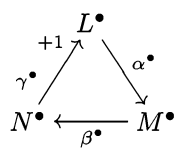
obtained in Proposition 4.1. In general, a *cohomological functor* on a triangulated category is an additive functor to an abelian category, mapping distinguished triangles to exact triangles and hence inducing long exact sequences. Thus, cohomology is a cohomological functor on $\mathsf{K}(\mathsf{A})$; this will not surprise our reader. The functors $\mathrm{Hom}(A, \_)$ and $\mathrm{Hom}(\_, A)$ (to Ab) are cohomological functors on every triangulated category, for every object A.

The derived category $\mathsf{D}(\mathsf{A})$ and its bounded versions are triangulated categories, with distinguished triangles described as above—that is, isomorphic (in $\mathsf{D}(\mathsf{A})$) to the basic triangles determined by mapping cones. The natural functor $\mathsf{K}(\mathsf{A}) \to \mathsf{D}(\mathsf{A})$ is a functor of triangulated categories, in the sense that it preserves the translation functor and it sends distinguished triangles to distinguished triangles.

Isomorphism in the derived category is a less stringent notion than in the homotopic category, so there are 'more' distinguished triangles in $\mathsf{D}(\mathsf{A})$ than in $\mathsf{K}(\mathsf{A})$. For example, if

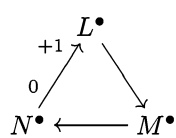
$$0 \longrightarrow L^\bullet \xrightarrow{\alpha^\bullet} M^\bullet \xrightarrow{\beta^\bullet} N^\bullet \longrightarrow 0$$

is a short exact sequence, as above, there is in general no distinguished triangle



in $\mathsf{K}(\mathsf{A})$ (for any choice of $\gamma^\bullet$): indeed, $N^\bullet$ need not[38] be homotopically equivalent to $MC(\alpha)^\bullet$. But such triangles *do* exist in $\mathsf{D}(\mathsf{A})$!

In fact, we are now well-equipped to make (better) sense of the mysterious remarks at the end of §3.4. The 'special triangles'



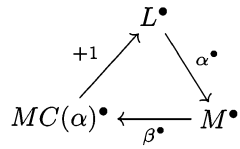
arising as in §3.4 from a short exact sequence

$$0 \longrightarrow L^\bullet \xrightarrow{\alpha^\bullet} M^\bullet \xrightarrow{\beta^\bullet} N^\bullet \longrightarrow 0$$

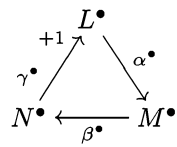
are not distinguished in $\mathsf{K}(\mathsf{A})$, and in general no replacement for the zero-morphism will fix this problem. On the other hand, the reader can now verify (Exercise 9.5) that there is a *quasi-isomorphism* $MC(\alpha)^\bullet \to N^\bullet$: this can be shown by computing

[38] Why? Construct an example showing this.

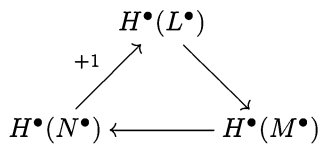
explicitly another mapping cone and applying Corollary 4.2. Thus, in the derived category the standard distinguished triangle



is isomorphic to a triangle



which is therefore distinguished *in* $\mathsf{D}(\mathsf{A})$ (cf. Exercise 9.6). Applying the cohomology functor to this triangle directly gives the exact triangle

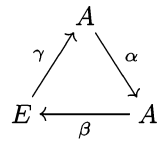


expressing the long exact cohomology sequence as in §3.4, without having to single out the $+1$ morphism for special consideration. In retrospect, the zero-morphism we insisted on using when folding 'special' triangles in §3.4 turns out to be an artifact: placing the triangle in the derived category shows that there is a more informed choice for this morphism. This choice is not available in $\mathsf{C}(\mathsf{A})$ and is not even in $\mathsf{K}(\mathsf{A})$, but it *is* available in $\mathsf{D}(\mathsf{A})$ and leads transparently to the long exact sequence in cohomology.

If $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ is a functor between abelian categories, the induced functor $\mathsf{K}(\mathscr{F}) : \mathsf{K}(\mathsf{A}) \to \mathsf{K}(\mathsf{B})$ is a functor of triangulated categories, and so are the bounded variations $\mathsf{K}^\pm(\mathscr{F}) : \mathsf{K}^\pm(\mathsf{A}) \to \mathsf{K}^\pm(\mathsf{B})$. The derived functors $\mathsf{L}\mathscr{F}$, $\mathsf{R}\mathscr{F} : \mathsf{D}^\pm(\mathsf{A}) \to \mathsf{D}^\pm(\mathsf{B})$ encountered in §7.2 are also functors of triangulated categories. This fact is responsible for the good properties of derived functors, such as the long exact sequences studied in §7.4.

9.3. Spectral sequences. We mentioned in §8.3 that 'spectral sequences' may be used to compute the cohomology of a double complex. There unfortunately does not seem to be an easy entry point in the subject of spectral sequences, which is marred by intrinsic notational complexity. We will limit ourselves to a very scant description of the concept and some information linking it back to §8.4. Our motivation is that, in the literature, references to results such as Theorem 8.12 are often replaced by sentences such as 'by an immediate spectral sequence argument...', and we should try to explain to the reader what this means.

Given that we were thinking about triangles a moment ago, the following construction may be helpful in appreciating the notion of spectral sequence. Let



be an exact triangle[39] in an abelian category; we are placing no assumption on the 'degree' of γ (and indeed, we are not assuming we have defined a translation functor). Let $d : E \to E$ be the composition $\beta \circ \gamma$; since $\gamma \circ \beta = 0$, we have

$$d \circ d = \beta \circ (\gamma \circ \beta) \circ \gamma = 0.$$

Thus we can take the homology of E with respect to d and define

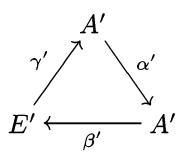
$$E' := \frac{\ker d}{\operatorname{im} d}.$$

Further, we can let $A' := \operatorname{im} \alpha$ and let

- $\alpha' : A' \to A'$ be the restriction of α to A';
- $\beta' : A' \to E'$ be defined by $\beta'(\alpha(a)) := [\beta(a)]$;
- $\gamma' : E' \to A'$ be defined by $\gamma'([e]) := \gamma(e)$,

where $[e]$ denotes the class of $e \in E$ in E'. (Since $\beta \circ \gamma(e) = d(e) = 0$, $\gamma(e) \in \ker \beta = \operatorname{im} \alpha = A'$.) All needed verifications (such as the independence on the representative e of $[e]$ in the third prescription) are straightforward, from the exactness of the given triangle.

Claim 9.1. *The new triangle*



is again exact.

The reader will have no difficulty proving this statement (Exercise 9.7).

The datum of an exact triangle as above is called an *exact couple,* which we find confusing since a triangle has three vertices (but it is true that only two objects are involved here); the new exact triangle (couple) obtained in Claim 9.1 is the *derived couple,* which we find even more confusing since there is no derived category or functor in sight. Exact couples arose in topology (they are due to W. Massey), and the terminology reflects their origin.

Of course we can turn the crank at will and get derived couple after derived couple: let $A_1 = A$, $E_1 = E$, etc., and inductively define $A_{i+1} := A'_i$, $E_{i+1} := E'_i$, etc. For example, β_{i+1} is obtained by 'rewinding' the morphism α a total of i times

[39] That is, $\operatorname{im} \alpha = \ker \beta$, $\operatorname{im} \beta = \ker \gamma$, and $\operatorname{im} \gamma = \ker \alpha$.

and then applying the original morphism β: we can do this since A_{i+1} is the image of α^i.

Thus, one exact triangle produces a whole sequence of triangles and in particular *a sequence of objects E_i, each endowed with a differential d_i, such that $E_{i+1} = \ker d_i / \operatorname{im} d_i$.*

Definition 9.2. A *spectral sequence* $\{(E_i, d_i)\}_{i=1,2,\dots}$ is a sequence of objects E_i and morphisms $d_i : E_i \to E_i$ in an abelian category, such that $d_i \circ d_i = 0$ and $E_{i+1} \cong \ker d_i / \operatorname{im} d_i$. ⌟

Thus, exact couples are a way to produce spectral sequences. There is a sense in which we can turn the crank 'infinitely many times'; in fact, this can be done with any spectral sequence. To see this, let

$$Z_1 = E_1, \quad B_1 = 0,$$

so that $E_1 \cong Z_1/B_1$; inductively, assume we have defined $Z_i \subseteq Z_1, B_i \subseteq Z_i$ so that $E_i \cong Z_i/B_i$, and let $\overline{Z}_{i+1} = \ker d_i$, $\overline{B}_{i+1} = \operatorname{im} d_i$; define Z_{i+1}, B_{i+1} as the corresponding subobjects of Z_i. Then

$$E_{i+1} \cong \frac{\overline{Z}_{i+1}}{\overline{B}_{i+1}} \cong \frac{Z_{i+1}}{B_{i+1}},$$

realizing E_{i+1} as a *subquotient*[40] of E_1, and

$$B_1 \subseteq \cdots \subseteq B_i \subseteq B_{i+1} \subseteq \cdots \subseteq Z_{i+1} \subseteq Z_i \subseteq \cdots \subseteq Z_1.$$

Define[41]

$$B_\infty := \bigcup_i B_i, \quad Z_\infty := \bigcap_i Z_i, \quad E_\infty := \frac{Z_\infty}{B_\infty}.$$

This 'ultimate' subquotient E_∞ is the *limit* of the spectral sequence; it is common to say that the spectral sequence E_r *abuts* to E_∞. By definition, if $d_r = d_{r+1} = \cdots = 0$, then $Z_\infty = Z_r$ and $B_\infty = B_r$, so that $E_\infty \cong E_r$; in this case we say that the sequence *collapses* at E_r.

The reader will verify (Exercise 9.8) that if the spectral sequence arises from an exact couple, as above, then

$$Z_\infty = \gamma^{-1}\Big(\bigcap_r \operatorname{im} \alpha^r\Big), \quad B_\infty = \beta\Big(\bigcup_r \ker \alpha^r\Big),$$

so the limit E_∞ has a concrete realization in this case.

Typically, a spectral sequence is as interesting as its limit, especially if some fortunate circumstance causes its collapse very soon; it is not too uncommon to have situations in which $E_\infty = E_2$. If all goes well, this translates into a useful relation between the given 'input' E_1 and the interesting 'output' E_∞.

---

[40] That is, as a quotient of a subobject.

[41] Here we are implicitly working in a category R-Mod of modules over a ring, containing the given abelian category A (invoking the Freyd-Mitchell theorem). The infinite unions and intersections used here make sense in R-Mod; the limits and the corresponding E_∞ are therefore defined as objects in R-Mod and in general not as objects of A. But if, e.g., the sequence collapses at E_r, then $B_\infty = B_r$, $Z_\infty = Z_r$, and E_∞ exists in A. In practice, this difficulty will play no role in the considerations that follow.

Where is my double complex? We promised that spectral sequences could be used to 'compute' the cohomology of the total complex of a double complex. We will now see how a double complex gives rise to an exact couple and hence to a spectral sequence. This is a particular case of a useful mechanism producing an exact sequence from a *filtration.*

A 'descending filtration' of M consists of a sequence of subobjects:

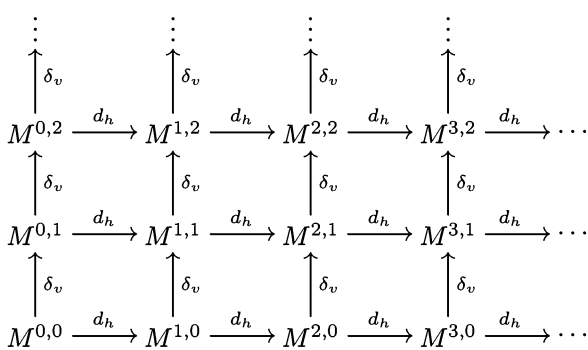
$$M \supseteq \cdots \supseteq M_m \supseteq M_{m+1} \supseteq \cdots .$$

For example, the *series* of subgroups of a group G considered in §IV.3.1 are filtrations of G. If M is an object of an abelian category, a filtration as above determines an associated graded object

$$\mathrm{gr}(M) := \bigoplus_m \mathrm{gr}(M)_m, \quad \mathrm{gr}(M)_m := \frac{M_m}{M_{m+1}};$$

We will assume this is an object of the same category for simplicity, although this is not necessarily the case (abelian categories are not necessarily closed under infinite direct sums); since only finitely many objects will be involved in any specific computation, this plays no role.

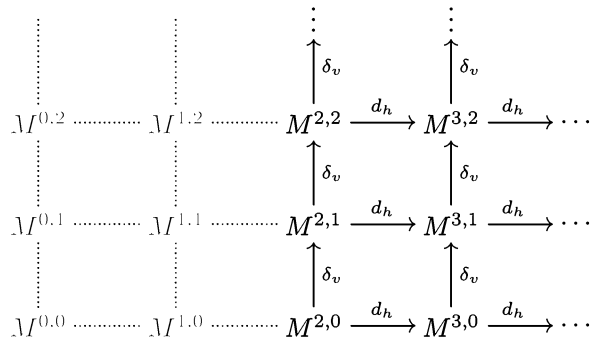
Start with a double complex $M^{\bullet,\bullet}$; we will assume that $M^{i,j} = 0$ for $i < 0$, $j < 0$:



As always, zero-objects to the left and below the given part of the diagram are implicit; we are assuming (as in Definition 8.5) that the diagram *anticommutes,* so that the differential of the total complex $TC(M)^\bullet$ is simply $d_h + \delta_v$. Analogous results can be obtained if $M^{i,j} = 0$ for $i > 0$, $j > 0$ and in fact under more relaxed hypotheses; such considerations are left to the reader.

Let $T^\bullet = TC(M)^\bullet$ denote the total complex: therefore, $T^k = \bigoplus_{i+j=k} M^{i,j}$. This complex admits (at least) two filtrations: a *horizontal* filtration and a *vertical* filtration. We will focus on the vertical one, and again it is understood that a parallel discussion would hold for the horizontal one. The vertical filtration $T^\bullet_m$ is

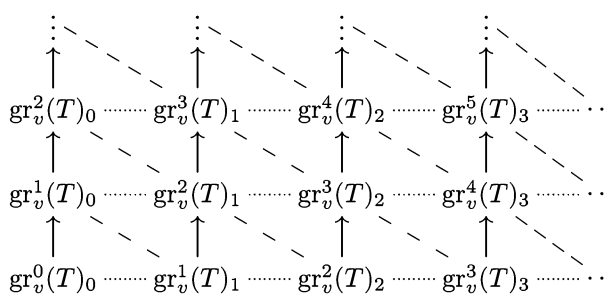
defined by chopping off the terms to the left of a vertical bar $i = m$ in the diagram



That is, we set

$$T^k_m := \bigoplus_{\substack{i+j=k \\ i \geq m}} M^{i,j},$$

still with differential $d_h + \delta_v$. We will denote the corresponding graded object by $\mathrm{gr}^\bullet_v(T)$, to record the fact that this arises from the vertical filtration. (The $\bullet$ reminds us that this is still a complex.) Explicitly, the term of 'filtration degree' m in $\mathrm{gr}^k_v(T)$ is $\mathrm{gr}^k_v(T)_m = T^k_m / T^k_{m+1} = \bigoplus_{i \geq m} M^{i,k-i} / \bigoplus_{i \geq m+1} M^{i,k-i} \cong M^{m,k-m}$. The differential $d_h + \delta_v$ induces the differential δ_v on this graded piece, since d_h is zero modulo the next piece of the filtration. The modules on the original array can therefore be labeled as follows:



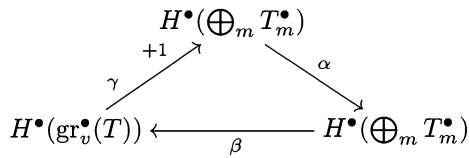
with dashes connecting pieces with the same degree in $\mathrm{gr}^\bullet_v(T)$.

The filtration on $T^\bullet$ also determines a filtration on the cohomology of $T^\bullet$: we can take $H^\bullet(T^\bullet)_m$ to be the image in $H^\bullet(T^\bullet)$ of $H^\bullet(T^\bullet_m)$. Thus, we also have a graded object $\mathrm{gr}_v H^\bullet(T^\bullet)$. *The relation between* $H^\bullet(\mathrm{gr}^\bullet_v(T))$ *and* $\mathrm{gr}_v H^\bullet(T^\bullet)$ *is subtle:* this *relation is what spectral sequences will help us understand.*

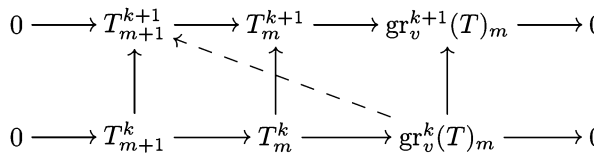
The monomorphisms $T^\bullet_{m+1} \subseteq T^\bullet_m$ define a monomorphism $\bigoplus_m T^\bullet_m \to \bigoplus_m T^\bullet_m$, of which $\mathrm{gr}^\bullet_v(T)$ is the cokernel: we have an exact sequence

$$(\dagger) \qquad 0 \longrightarrow \bigoplus_m T^\bullet_m \longrightarrow \bigoplus_m T^\bullet_m \longrightarrow \mathrm{gr}^\bullet_v(T) \longrightarrow 0\,.$$

As we know since §3.3, this determines an exact triangle (couple!)



Note that α decreases the filtration degree by 1, leaving the cohomology degree k unchanged; β leaves both unchanged; and γ increases both by 1. This is particularly clear if one draws a small slice of the sequence (†), sufficient to see the connecting morphism in action on a piece of given filtration and cohomology degrees:



Definition 9.3. The *spectral sequence of the double complex* $M^{\bullet,\bullet}$ (with respect to the vertical filtration) is the spectral sequence determined by this exact couple. ⌟

We will denote this spectral sequence by ${}_vE$, with the absurdly positioned index v recording that ${}_vE$ arises from the *vertical* filtration. The horizontal filtration leads likewise to a spectral sequence ${}_hE$.

All terms ${}_vE_r$ of the spectral sequence are subquotients of the first one, ${}_vE_1 = H^\bullet(\mathrm{gr}_v^\bullet(T))$.

Remark 9.4. The differential $d_r = \beta_r \circ \gamma_r$ acts by increasing the cohomological degree k by 1 and the filtration degree m by r: indeed, γ_r is simply induced by γ, so it increases both k and m by 1 (as observed above); β_r is obtained by applying β after undoing $(r-1)$ times the effect of α (as pointed out after Claim 9.1), so it does not change k but increases m by $(r-1)$. ⌟

Interest in the sequence rests on the interpretation for its limit in the statement that follows, which also summarizes the situation.

Theorem 9.5. *Let $M^{\bullet,\bullet}$ be a double complex in an abelian category. Assume that $M^{i,j} = 0$ for $i < 0$, $j < 0$, and let $T^\bullet$ be the total complex of $M^{\bullet,\bullet}$. Then, with notation as above, there exists a spectral sequence $\{({}_vE_i, d_i)\}_i$ such that*

$$ {}_vE_1 \cong H^\bullet(\mathrm{gr}_v^\bullet(T)), \quad {}_vE_\infty \cong \mathrm{gr}_v H^\bullet(T^\bullet). $$

In other words, 'turning the crank' moves the gr from inside the cohomology to outside of it. The limit ${}_vE_\infty$ does not quite compute the cohomology of the total complex, as we glibly announced in §8.3, but it computes the graded object determined by a filtration on the cohomology of the total complex, and this is good enough for many applications.

Proof. The limit ${}_vE_\infty$ can be computed directly from the exact couple (Exercise 9.8), as a subquotient of ${}_vE_1$: it is the quotient Z_∞/B_∞, where

$$Z_\infty = \gamma^{-1}(\bigcap_r \operatorname{im}\alpha^r), \quad B_\infty = \beta(\bigcup_r \ker\alpha^r).$$

The computation is streamlined by focusing on a specific degree k for the cohomology and m for the grading. The corresponding part in ${}_vE_1$ is $H^k(T^\bullet_m/T^\bullet_{m+1})$, and α acts as $H^k(T^\bullet_m) \to H^k(T^\bullet_{m-1})$. Note that

$$\text{if } r \gg 0, \text{ then } \quad H^k(T^\bullet_{m+r}) = 0, \quad H^k(T^\bullet_{m-r}) = H^k(T^\bullet),$$

under our boundedness hypothesis: indeed, $T^j_i = 0$ if $i > j$, and $T^\bullet_i = T^\bullet$ for $i \le 0$. By the first of these observations, the piece sent by α^r to $H^k(T^\bullet_m)$ is 0 for $r \gg 0$, and it follows that $\bigcap_r \operatorname{im}\alpha^r = 0$. Therefore,

$$Z_\infty = \ker\gamma = \operatorname{im}\beta.$$

By the second observation, $\ker\alpha^r$ stabilizes to $\ker(H^k(T^\bullet_m) \to H^k(T^\bullet))$ for each k, hence

$$B_\infty = \beta(\bigoplus_m \ker(H^\bullet(T^\bullet_m) \to H^\bullet(T^\bullet))).$$

Denote by $\nu_m : H^\bullet(T^\bullet_m) \to H^\bullet(T^\bullet)$ the morphisms induced in cohomology by the monomorphisms $T^\bullet_m \to T^\bullet$ and by $\nu = \bigoplus \nu_m$ their direct sum. We have morphisms

$$H^\bullet(\mathrm{gr}^\bullet_v(T)) \xleftarrow{\beta} \textstyle\bigoplus_m H^\bullet(T^\bullet_m) \xrightarrow{\nu} \bigoplus_m H^\bullet(T^\bullet)\ ,$$

and we have obtained

$${}_vE_\infty \cong \frac{\operatorname{im}\beta}{\beta(\ker\nu)}.$$

By the innocent Exercise 2.12,

$$\frac{\operatorname{im}\beta}{\beta(\ker\nu)} \cong \frac{\operatorname{im}\nu}{\nu(\ker\beta)};$$

by the exactness of the original couple,

$$\ker\beta = \operatorname{im}\alpha = \bigoplus_m \operatorname{im}(H^\bullet(T^\bullet_{m+1}) \to H^\bullet(T^\bullet_m)),$$

hence

$$\nu(\ker\beta) = \bigoplus_m \operatorname{im}\nu_{m+1}.$$

Therefore

$${}_vE_\infty \cong \bigoplus_m \frac{\operatorname{im}\nu_m}{\operatorname{im}\nu_{m+1}} = \mathrm{gr}_v(H^\bullet(T^\bullet))$$

as stated. □

Theorem 9.5 is a substantial generalization of Theorems 8.9 and 8.12. To see that it implies Theorem 8.12, note again that

$$\mathrm{gr}^k_v(T) = \bigoplus_m \frac{T^k_m}{T^k_{m+1}} \cong \bigoplus_m M^{m,k-m},$$

with its 'vertical' differential δ_v. The term ${}_vE_1$ of the spectral sequence is the cohomology of this complex, that is, the cohomology of the columns of the original double complex, suitably shifted. If the cohomology of each column $M^{i,\bullet}$ is concentrated in degree 0, then ${}_vE_1$ is simply the complex

$$(*) \qquad \cdots \longrightarrow 0 \longrightarrow H^0(M^{0,\bullet}) \longrightarrow H^0(M^{1,\bullet}) \longrightarrow H^0(M^{2,\bullet}) \longrightarrow \cdots;$$

hence ${}_vE_2$ is the cohomology of this complex, and higher differentials of the spectral sequence are 0 by degree considerations. It follows that the sequence collapses at ${}_vE_2$, and by Theorem 9.5 this says that $\mathrm{gr}_v(H^\bullet(T^\bullet))$ is isomorphic to the cohomology of (*). In this case we clearly have $\mathrm{gr}_v(H^\bullet(T^\bullet)) \cong H^\bullet(T^\bullet)$, so the conclusion reproduces the corresponding statement in Theorem 8.12. This is what is meant by the incantation *'by an immediate spectral sequence argument'* we mentioned in the beginning of this subsection.

This reasoning and other applications of the spectral sequence of a double complex are easier to understand if the various degrees and shifts are visualized in a more compelling way. View the original double complex

$$\begin{array}{ccccccc}
\vdots & & \vdots & & \vdots & & \vdots \\
\uparrow & & \uparrow & & \uparrow & & \uparrow \\
M^{0,2} & \cdots\cdots & M^{1,2} & \cdots\cdots & M^{2,2} & \cdots\cdots & M^{3,2} \cdots\cdots \cdots \\
\uparrow & & \uparrow & & \uparrow & & \uparrow \\
M^{0,1} & \cdots\cdots & M^{1,1} & \cdots\cdots & M^{2,1} & \cdots\cdots & M^{3,1} \cdots\cdots \cdots \\
\uparrow & & \uparrow & & \uparrow & & \uparrow \\
M^{0,0} & \cdots\cdots & M^{1,0} & \cdots\cdots & M^{2,0} & \cdots\cdots & M^{3,0} \cdots\cdots \cdots
\end{array}$$

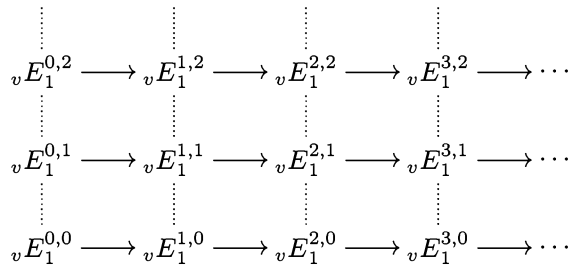
as an 'E_0' term in the sequence. The horizontal differentials are dotted out to highlight the fact that ${}_vE_1$ is obtained by taking the cohomology of the columns; this gives

$$\begin{array}{ccccccc}
\vdots & & \vdots & & \vdots & & \vdots \\
H^2(T_0^\bullet/T_1^\bullet) & \longrightarrow & H^3(T_1^\bullet/T_2^\bullet) & \longrightarrow & H^4(T_2^\bullet/T_3^\bullet) & \longrightarrow & H^5(T_3^\bullet/T_4^\bullet) \longrightarrow \cdots \\
\vdots & & \vdots & & \vdots & & \vdots \\
H^1(T_0^\bullet/T_1^\bullet) & \longrightarrow & H^2(T_1^\bullet/T_2^\bullet) & \longrightarrow & H^3(T_2^\bullet/T_3^\bullet) & \longrightarrow & H^4(T_3^\bullet/T_4^\bullet) \longrightarrow \cdots \\
\vdots & & \vdots & & \vdots & & \vdots \\
H^0(T_0^\bullet/T_1^\bullet) & \longrightarrow & H^1(T_1^\bullet/T_2^\bullet) & \longrightarrow & H^2(T_2^\bullet/T_3^\bullet) & \longrightarrow & H^3(T_3^\bullet/T_4^\bullet) \longrightarrow \cdots
\end{array}$$

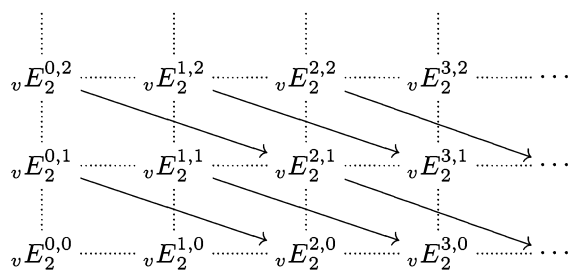
For example, the part in degree 2 of $H^3(\mathrm{gr}_v^\bullet(T))$ is the cohomology of the second column $M^{2,\bullet}$ computed at $\mathrm{gr}_v^3(T)_2 = M^{2,1}$, so we placed the corresponding object $H^3(T_2^\bullet/T_3^\bullet)$ in position $(2,1)$. The differential of ${}_vE_1$ increases both the cohomological degree and the filtration degree by 1, as indicated. Such diagrams are somewhat hard to parse. To simplify dealing with the indices, it is common to assign degrees (i,j) to the terms in ${}_vE_1$ according to the indices of the objects $M^{i,j} = M^{m,k-m}$ at which the cohomology is computed; that is, set

$${}_vE_1^{i,j} = H^{i+j}(T_i^\bullet/T_{i+1}^\bullet).$$

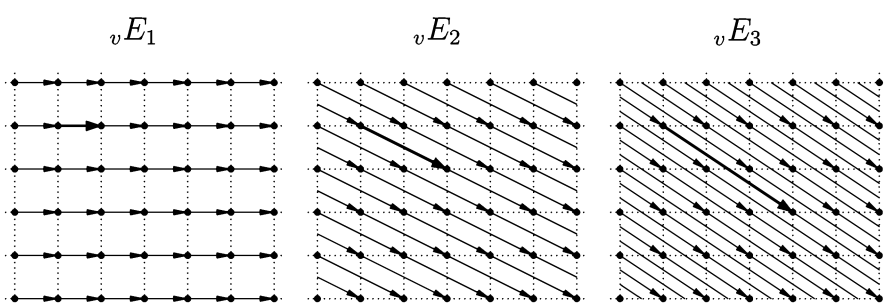
Then the same array is drawn as



The next term ${}_vE_2$ is obtained by taking the cohomology of ${}_vE_1$. Its differential acts by increasing m by 2 and k by 1, that is, i by 2 and j by -1. We can draw this as



In general, ${}_vE_r$ is the cohomology of ${}_vE_{r-1}$, and its differential increases m by r and k by 1 (Remark 9.4); that is, $(i,j) \mapsto (i+r, j-(r-1))$. Pictorially, differentials at different stages act like this:



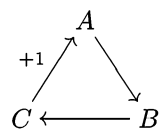
and so on. This picture of a 'rotating' differential may be the image that most people have in mind when they think of a spectral sequence. The argument sketched above, deriving Theorem 8.12 from Theorem 9.5, is probably clearer from this point of view: if the cohomology of the columns is concentrated in degree 0, then ${}_vE_1$ is concentrated along the bottom row of the array. It is then clear that all differentials d_i for $i \geq 2$ will have to be zero, since either their source or their target is zero. Thus the sequence collapses at ${}_vE_2$ and yields the cohomology of the total complex by Theorem 9.5. The other situation considered in Theorem 8.12 can be analyzed by using the sequence ${}_hE_i$ obtained from the horizontal filtration.

Of course, spectral sequences are not limited to these simple applications. The *Grothendieck spectral sequence* 'computes' the derived functor of the composition of two functors[42]: for example, if $\mathscr{F} : \mathsf{A} \to \mathsf{B}$ and $\mathscr{G} : \mathsf{B} \to \mathsf{C}$ are two right-exact functors and $\mathscr{F}$ sends projectives to projectives, then there is a spectral sequence whose E_2 term collects the compositions $\mathsf{L}_p\mathscr{G} \circ \mathsf{L}_q\mathscr{F}$ and that abuts to $\mathsf{L}_{p+q}(\mathscr{G} \circ \mathscr{F})$. For instance, if $f : R \to S$ is a homomorphism of commutative rings (so that S may be viewed as an R-module), A is an R-module, and B is an S-module (and hence an R-module, via f), then there is a 'change-of-ring spectral sequence' $\operatorname{Tor}_p^S(\operatorname{Tor}_q^R(A,S),B) \Rightarrow \operatorname{Tor}_\bullet^R(A,B)$. In topology, the *Serre spectral sequence* can be used to compute the homology of a fibration in terms of the homology of the base and of the fiber. It would be completely futile for us to attempt to do any justice to the range of applications of spectral sequences: the set of mathematicians X such that there is an important X-spectral sequence includes (but is not limited to) Adams, Atiyah, Barratt, Bloch, Bockstein, Bousfield, Cartan, Connes, Eilenberg, Federer, Frölicher, Green, Grothendieck, Hirzebruch, Hochschild, Hodge, Hurewicz, van Kampen, Kan, Kunneth, Leray, Lichtenbaum, Lyndon, May, Miller, Milnor, Moore, Novikov, de Rham, Quillen, Rothenberg, Serre, Steenrod, ...

Many important spectral sequences may be derived as particular instances of the Grothendieck spectral sequence mentioned above. The reader will have the pleasure of seeing how this spectral sequence is put together by working out Exercise 9.10.

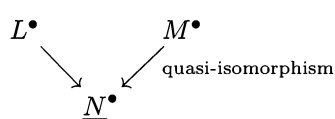
Exercises

9.1. Let A be an abelian category. Since the objects of $\mathsf{K}(\mathsf{A})$ and $\mathsf{D}(\mathsf{A})$ are simply cochain complexes, a 'universal Euler characteristic' χ is defined for bounded complexes in these categories (see Exercise 3.15). Prove that if



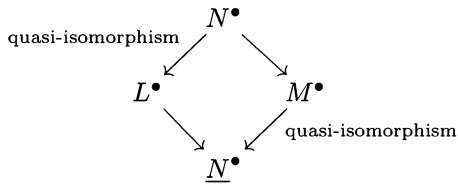
is a distinguished triangle in $\mathsf{K}(\mathsf{A})$ or $\mathsf{D}(\mathsf{A})$, then $\chi(B) = \chi(A) + \chi(C)$.

9.2. ▷ Let A be an abelian category, and suppose two complexes $L^\bullet$, $N^\bullet$ are connected by an 'upside-down roof' (a 'trough'?)



[42]For a different viewpoint on this question, see Remark 7.4.

Prove that they are also connected by a regular roof: there exists a complex $N^\bullet$ and morphisms $N^\bullet \to L^\bullet$ and $N^\bullet \to M^\bullet$ such that the diagram



commutes in $\mathsf{K}(\mathsf{A})$. Deduce that the composition of two roofs may be consolidated into a single roof in $\mathsf{K}(\mathsf{A})$. (Hint: Let $\alpha^\bullet : L^\bullet \to \underline{N}^\bullet$, $\beta : M^\bullet \to \underline{N}^\bullet$ be the given morphisms, with β a quasi-isomorphism. Consider the composition $\gamma^\bullet : L^\bullet \to \underline{N}^\bullet \to MC(\beta)^\bullet$, and let $N^\bullet := MC(\gamma)[-1]^\bullet$. We have $N^i = L^i \oplus M^i \oplus \underline{N}^{i-1}$, with morphisms to L^i, M^i given, respectively, by $(\ell, m, n) \mapsto \ell$, $(\ell, m, n) \mapsto -m$; these define morphisms of cochain complexes. The morphisms $N^i \to \underline{N}^{i-1}$ defined by $(\ell, m, n) \mapsto -n$ define the homotopy showing that the diagram commutes in $\mathsf{K}(\mathsf{A})$. To verify that $N^\bullet \to L^\bullet$ is a quasi-isomorphism, use the fact that $MC(\beta)^\bullet$ is exact, and rotate a triangle.) [§9.1]

9.3. ▷ Let $\alpha^\bullet : L^\bullet \to M^\bullet$ be an isomorphism of cochain complexes. Prove that $MC(\alpha)^\bullet$ is homotopy equivalent to 0. [§9.2, 9.4]

9.4. ▷ Let $\alpha^\bullet : L^\bullet \to M^\bullet$ be a cochain morphism, and let $\beta^\bullet : M^\bullet \to MC(\alpha)^\bullet$ be the natural morphism. Prove that $MC(\beta)^\bullet$ is homotopy equivalent to $L[1]^\bullet$. (Hint: Do Exercise 9.3 first.) [§9.2]

9.5. ▷ Let

$$0 \longrightarrow L^\bullet \xrightarrow{\alpha^\bullet} M^\bullet \xrightarrow{\beta^\bullet} N^\bullet \longrightarrow 0$$

be a short exact sequence of cochain complexes on an abelian category.

- Prove that there is a cochain morphism $\gamma^\bullet : MC(\alpha)^\bullet \to N^\bullet$ through which $\beta^\bullet : M^\bullet \to N^\bullet$ factors.
- Prove that $MC(\gamma)^\bullet$ is an exact complex. (Hint: Chase elements.)
- Conclude that $MC(\alpha)^\bullet$ is quasi-isomorphic to $N^\bullet$.

[§9.2]

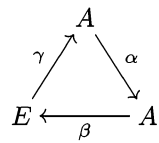
9.6. ▷ Let

$$0 \longrightarrow L^\bullet \xrightarrow{\alpha^\bullet} M^\bullet \xrightarrow{\beta^\bullet} N^\bullet \longrightarrow 0$$

be a short exact sequence of cochain complexes on an abelian category A. Prove that there is a morphism $N^\bullet \to L[1]^\bullet$ in the derived category $\mathsf{D}(\mathsf{A})$, inducing the connecting morphism $H^i(N^\bullet) \to H^{i+1}(L^\bullet)$ in cohomology. [§9.2]

9.7. ▷ Prove Claim 9.1. [§9.3]

9.8. ▷ Let $(E_r, d_r)_{r=1,2,\dots}$ be a spectral sequence arising from an exact couple



Prove that the limit E_∞ is isomorphic to $\gamma^{-1}(\bigcap_r \operatorname{im}\alpha^r)/\beta(\bigcup_r \ker\alpha^r)$. (Hint: In fact, $Z_{r+1} = \gamma^{-1}(\operatorname{im}\alpha^r)$ and $B_{r+1} = \beta(\ker\alpha^r)$.) [§9.3]

9.9. Let ${}_vE_r$ be the spectral sequence of a first-quadrant double complex from the vertical filtration, as in §9.3. Prove that ${}_vE_r^{i,j} = {}_vE_\infty^{i,j}$ for $r > j+1$.

9.10. ▷ Let A, B, C be abelian categories, and let $\mathscr{F} : \mathsf{A} \to \mathsf{B}$, $\mathscr{G} : \mathsf{B} \to \mathsf{C}$ be additive functors. Assume that A and B have enough projectives, that $\mathscr{F}$ sends projectives of A to $\mathscr{G}$-acyclic objects of B, and that $\mathscr{G}$ is right-exact.

Let A be an object of A, and let $M^\bullet$: $\cdots \to M^{-2} \to M^{-1} \to M^0 \to 0$ be a projective resolution of A. Let $P^\bullet_{\mathscr{F}M^\bullet}$ be a Cartan-Eilenberg resolution of the complex $\mathscr{F}M^\bullet$ (as constructed in Corollary 7.9). Consider the complex

$$(\dagger) \qquad \cdots \longrightarrow \mathscr{G}(P^\bullet_{\mathscr{F}M^{-2}}) \longrightarrow \mathscr{G}(P^\bullet_{\mathscr{F}M^{-1}}) \longrightarrow \mathscr{G}(P^\bullet_{\mathscr{F}M^0}) \longrightarrow 0$$

obtained by applying $\mathscr{G}$ to $P^\bullet_{\mathscr{F}M^\bullet}$.

This is precisely the set-up of Exercise 8.8. Now take the double complex associated with (†), and construct the spectral sequence corresponding to the horizontal filtration. Prove that the E_2 term of this sequence has terms

$${}_hE_2^{p,q} = \mathsf{L}_p\mathscr{G}(\mathsf{L}_q\mathscr{F}(A)).$$

(Use Exercise 7.6 to compute the ${}_hE_1$ term.)

This is the Grothendieck spectral sequence mentioned at the end of §9.3. Prove that it abuts to $\mathsf{L}_\bullet(\mathscr{G}\circ\mathscr{F})(A)$.

(Use the 'horizontal' version of Theorem 9.5, together with the computation of the cohomology of the total complex from Exercise 8.8.) [§7.2, 7.6, §9.3]

Index

A^n, 38
$\mathrm{Ann}_R(M)$, 342
$\mathrm{Ass}_R(M)$, 347
$\mathrm{Aut}_k(F)$, 390
$\mathrm{Aut}_{\mathsf{C}}(A)$, 29
B^A, 9
$\mathbb{C}$, 2
$\mathsf{C}(\mathsf{A})$, 594
$(c_0 : \ldots : c_n)$, 416
$\mathsf{D}(\mathsf{A})$, 609
D^{C}, 497
D_{2n}, 52
$\dim_k V$, 311
$\mathrm{End}_{\mathsf{C}}(A)$, 19
f^*, 519
f_*, 518
F_{sep}, 439
$\mathbb{F}_q$, 442
$G * H$, 62
G/H, 93, 102
$\mathrm{GL}_n(\mathbb{R})$, 43
$\mathrm{Gal}_k(f(x))$, 477
$\mathbb{H}$, 128
$H\backslash G$, 104
$\mathrm{Hom}_{\mathsf{C}}(A, B)$, 19
I, 620
$\mathrm{id}(B)$, 679
$\mathrm{Inn}(G)$, 86
$\mathsf{K}(\mathsf{A})$, 618
$\mathsf{K}^+(\mathsf{I})$, 620
$\mathsf{K}^-(\mathsf{P})$, 620
$k(\alpha)$, 387
$k(\alpha_1, \ldots, \alpha_n)$, 393
K-Aff, 488
$\mathsf{L}\mathscr{F}$, 642
Λ_R^*, 529
$\varinjlim$, 491
$\varprojlim$, 489
$MC(\alpha)^\bullet$, 605, 606
$MCyl(\alpha)^\bullet$, 615
μ_n, 445
$\mathbb{N}$, 2
$N_{k\subseteq F}(\alpha)$, 398
P, 620
$\mathscr{P}(S)$, 4
$\mathbb{P}V$, 326
$\mathrm{pd}(A)$, 678
$\mathbb{Q}$, 2
$\mathbb{Q}(\sqrt{d})$, 154
Q_8, 128
$\mathbb{Q}_p$, 498
$\mathbb{R}$, 2
$\mathsf{R}\mathscr{F}$, 642
$R\langle A\rangle$, 169
$R_{\mathfrak{p}}$, 278
$\mathrm{Rees}_R(I)$, 528
$\mathrm{rk}_R M$, 311
$\mathrm{SL}_2(\mathbb{Z})$, 87
$S^{-1}R$, 277
$\mathbb{S}_R^*$, 529
$\mathsf{Seq}(\mathsf{A})$, 603
$TC(M)^\bullet$, 667, 668
$\mathbb{T}_R^*$, 523
$\mathrm{Tor}_R(M)$, 340
$\mathrm{tr}(A)$, 362
$\mathrm{tr}_{k\subseteq F}(\alpha)$, 398
${}_vE_1^{i,j}$, 693
$\mathbb{Z}$, 2
ζ_n, 445
$\widehat{\mathbb{Z}}$, 499
$\mathbb{Z}_p$, 498

$\exists$, 2
$\forall$, 3
$\gg$, 275
$\hookrightarrow$, 11
$\rightarrowtail$, 564
$\twoheadrightarrow$, 11, 564
$\leadsto$, *see* natural transformation
2^A, *see* power set
$\sqrt{I}$, *see* radical of an ideal
$\oplus$, *see* direct sum
$\otimes$, *see* tensor product

a.c.c., *see* ascending chain condition
Abel, Niels Henrik, 477
Abel-Ruffini theorem, 477
abelian category, 156, 559, 564
abelian group, 45, 62
  finite, classification, 234
  finitely generated, 82, 234, 327
abstract nonsense, 18
abuts (spectral sequence), 688
action, 64
  effective, 109
  faithful, 109
  free, 109
  of a group, 108
  of a ring, 156
  orbit of, 110
  transitive, 110
acyclic object, 662
additive
  category, 561
  functor, 485, 561, 613
adjoint functors, 492, 498, 642
  preserve limits/colimits, 493
adjoint matrix, 331, 369
affine
  algebraic set, 408
  space, 406
algebra, 159
  commutative, 159
  division, 159
  finite, 171, 405
  finite type, 171, 405
  finitely generated, 171
  free commutative, 168
  Rees, 164, 528
algebraic
  closure, 285, 400, 445
  element of an extension, 391
  geometry, 406
  number, 394
algebraic closure, 285, 400
  of finite fields, 445
algebraic set
  affine, 408
  projective, 533
algebraically closed
  $\mathbb{C}$ is, 152, 286, 468
algebraically independent set, 399
Alperin, Roger, 114
alternating group, 220, 473
  A_n is simple for $n \geq 5$, 222
  conjugacy in, 221
  is generated by 3-cycles, 223
alternating multilinear map, 524
annihilator, 544
  ideal of a module, 342
Archimedes, 428
Argand, Jean-Robert, 286
Artin, Emil, 401
  Artinian ring, 250
Artin-Rees lemma, 535
Artin-Schreier
  extension, 481
  map, 482
Artinian ring, 250
ascending chain condition, 244
  for principal ideals, 248, 253, 275
associate elements in a ring, 246
associated
  graded object, 689
  prime, 347
associative law, 42
automorphism, 29, 49, 67, 192
  inner, 69, 86
  of a field extension, 390
axiom of choice, 6, 262, 265, 307, 403

Baer, Reinhold
  criterion, 549
Baker, Alan, 302
balanced bilinear map, 515
Banach, Stefan, 262
Banach-Tarski paradox, 262
basis, 306
  ordered, 308
  standard, 316
  transcendence, 400
Betti numbers, 336
bifunctor, 679
bijective, 11
bilinear map, 501
  nondegenerate, 544
  nonsingular, 544
bimodule, 517
binary operation, 42
Birkhoff-Vandiver theorem, 453
block matrix, 322
Boole, George, 144
Boolean ring, 144, 156, 300
Burnside, William
  theorem, 214

calculus, 433
cancellation, 45, 46, 121, 122
  holds in integral domains, 122
canonical decomposition
  in Grp, 97
  in Ring, 141
  in R-Mod, 162
  in Set, 15
  in abelian categories, 570
canonical form
  Jordan, 379
  rational, 375
Cartan-Eilenberg resolution, 651, 675
category, 18
  Ab, 62, 68
  Fld, 385
  G-Mod, 655
  Grp, 58
  G-Set, 111
  k-Vect, 158
  R-Alg, 159
  Ring, 129
  R-Mod, 158
  Set, 20
  Set*, 24
  Seq(A), 603
  abelian, 156, 559, 564
  additive, 561
  comma, 24
  coslice, 24
  derived, 609, 681
  discrete, 21, 489
  equivalence of, 487
  of affine algebraic sets, 488
  of complexes, 594
  of functors, 497
  opposite, 26, 38
  preadditive, 574
  slice, 24
  small, 18, 22
  subcategory
    full, 26, 595
  triangulated, 602, 610, 618, 650, 683
  with coproducts, 37
  with products, 36
Cauchy, Augustin Louis
  theorem, 103, 195
Cayley, Arthur
  graph, 74, 106
  theorem, 110, 472
Cayley-Hamilton theorem, 365, 369, 376
center
  of a group, 189
  of a ring, 137, 159
centralizer, 137, 190, 191
chain, 261
change of basis, 319
characteristic
  of a field, 386, 435
  of a ring, 141
characteristic ideal of a module, 356
characteristic polynomial, 362
  is a multiple of the minimal polynomial, 376
Chinese remainder theorem (CRT), 235, 291
class formula, 187, 190
classification
  of finite abelian groups, 234
  of finitely generated abelian groups, 82, 327
  of finitely generated modules over PIDs, 349
cochain
  complex, 591
  of a group, 657
coefficient, 125
cofactors of matrices, 331
cofree module, 551
cohomological functor, 685
cohomology, 174, 178, 553, 592
  as functor, 596
  long exact sequence, 600
  of groups, 655
coimage, 572
  isomorphic to image, 573
coinvariant of a group action, 658
cokernel, 104
  categorical definition, 491, 561
  in Ab, 104
  in Ring, 137
  in R-Mod, 167
  universal property, 104, 166
colimit, 490, 498
column space of a matrix, 333
comma category, 24
commutative, 45
  ring, 121
commutator, 210
  of two subsets, 226
companion matrix, 374
completion of a ring, 498
complex, 174, 335, 591
  cochain, 591
  double, 667
  exact, 175, 592
  in an abelian category, 576
  of complexes, 665
  split exact, 621, 627
  total, 606
composite field, 393, 438
composition
  factors, 207, 313
  series, 206, 313

cone
  mapping, 605, 623, 624
  of a functor, 489
congruence, 54
conjugacy class, 190
  in S_n, 216
conjugation, 110, 189
connecting morphism, 180, 580
conservative functor, 520
constant, 125
constructible
  geometric figures, 417
  numbers, 421
  regular polygons, 469
content of a polynomial, 269
contravariant functor, 484
coordinate ring of an algebraic set, 413
coproduct, 36
  fibered, 39, 173
  in an abelian category, 567
  of abelian groups, 62
  of algebras, 511
  of groups, 62, 100
  of modules, 165
  of rings, 133, 512
  universal property, 36
cosets of a subgroup, 91
coslice category, 24
covariant functor, 484
Cramer, Gabriel
  rule, 328, 332, 365
cycle, 215
  notation, 215
cyclic group, 54, 67, 103
  $\mathbb{Z}/p\mathbb{Z}^*$ is a, 67, 240
  infinite, 67, 70
  subgroup of, 82
  the multiplicative group of a finite field is a, 239, 442
cyclic module, 174, 341
cyclotomic
  field, 448, 480
  polynomial, 67, 289, 426, 446
cylinder
  mapping, 615

D-brane, 610
D-module, 534
d.c.c., *see* descending chain condition
Dedekind, Richard, 256, 298
degree, 125
  in a graded ring, 527
  inseparable, 439, 440
  separable, 436
delta functor, 660, 680
derivative of a polynomial, 433
derived
  couple, 687
  functor, 642, 645
    composition, 645, 695
    left, 645
    right, 646
derived category, 609, 681
  poor man's, 635
derived series of a group, 211
descending chain condition, 250, 314
determinant
  of a matrix, 328
  of a module, 526
  of an endomorphism, 361
diagonalizable
  linear transformation, 368, 380
  real symmetric matrices are, 381
diagram, 10
  chase, 180
  commutative, 10, 19
  in a category, 20
differential
  form, 530
  of a complex, 592
  operator, 534
dihedral group, 52, 225
  presentation, 57, 106
dimension
  global, 679
  injective, 679
  Krull, of a ring, 153, 250, 354, 679
  of a vector space, 311
  projective, 678
direct limit, 402, 491
direct product, 61, 226
direct sum, 62, 76, 77, 164
  in an abelian category, 570
directed set, 491, 498
Dirichlet, Peter Gustav Lejeune
  theorem on primes in arithmetic progressions, 454, 480
discrete valuation ring, 260, 278
discriminant, 473, 478, 482
disjoint union, 4, 16
distinguished triangle, 606, 683
divisible
  module, 550
division
  algebra, 159
  ring, 123, 128
double complex, 667
dual
  basis, 539
  double, 542
  group, 241
  kills torsion, 542
  module, 537
  Pontryagin, 555

vector space, 537
DVR, 260, 278, 416

echelon form, 325
of a matrix, 325
effaceable functor, 660
effective action, 109
eigenspace, 367, 379
eigenvalue of a linear transformation, 365
multiplicity of, 366, 367, 379
eigenvector, 367
Eisenstein, Ferdinand Gotthold Max
criterion, 288
element, 1
of an object of an abelian category, 584
elementary
divisors, 237, 354
operations, 320
symmetric functions, 472
endomorphism, 19, 119, 134, 359
characteristic polynomial of an, 362
determinant of an, 361
trace of an, 362
enough
injectives, 550, 620
projectives, 548, 620
epimorphism, 29
in an additive/abelian category, 563
of groups, 104
of modules, 167
of rings, 132, 398
of sets, 14
split, 178, 547
equalizers, 490
equivalent categories, 487
equivariant function, 111
essentially surjective functor, 488
Euclid, 261
Euclidean algorithm, 257
Euclidean domain, 255, 323, 341, 347
$\Longrightarrow$ PID, 256
Euler, Leonhard, 298
ϕ-function, 58, 70, 87, 107, 301
characteristic
of a complex, 335
universal, 337, 605, 695
theorem, 107
exact
couple, 687
functor, 495, 588, 603, 613, 643
faithfully, 499
left- or right-, 495, 653
sequence, 175, 228
in an abelian category, 576
of Ext, 552
of pointed sets, 582
of Tor, 509
short, 176
split, 177, 229, 621, 627
triangle, 601, 684
exactness property
of $\otimes$, 507
of Hom, 537
of $\mathbb{T}^*_R$, $\mathbb{S}^*_R$, $\mathbb{\Lambda}^*_R$, 531
of adjoint functors, 495
of functors, 495
of the duality functor, 539, 541
exponential map, 64
Ext, 551, 647
long exact sequence, 552
may be computed by resolving either argument, 553, 677
why it is called Ext, 556
extension
algebraic, 391
Artin-Schreier, 481
degree of an, 386
field, 163, 283
finite, 386
finitely generated, 393
Galois, 458
normal, 431
of groups, 228
of modules, 184, 555, 556
of scalars, 518
quadratic, 422
contained in cyclotomic field, 482
radical, 475
separable, 436
simple, 387, 449
solvable, 475
split, 177, 229, 231
exterior
algebra, 529
power, 524

factorial ring (UFD), 248
factorization into irreducibles, 248
unique, 248
faithful
action, 109
faithfully exact, 499
functor, 488, 587
Feit-Thompson theorem, 212, 214
Fermat, Pierre de
last theorem, 280
little theorem, 103, 107, 439
primes, 427, 454
theorem on sums of squares, 298, 299
Ferrers diagram, 216
fiber, 13
square, 567
fibered products and coproducts, 39, 173
in an abelian category, 567, 577

field, 123
  algebraic closure, 285, 400, 401, 445
  algebraically closed, 152, 285
  characteristic of, 386, 435
  composite, 393
  cyclotomic, 448, 480
  extension, 163, 283, 386
    algebraic, 391
    degree of a, 386
    finite, 386
    finite $\Longrightarrow$ algebraic, 391
    finitely generated, 393
    Galois, 458
    normal, 431
    quadratic, 422, 482
    radical, 475
    separable, 436
    simple, 387, 449
    solvable, 475
  finite, 441
  fixed, 455
  Galois, 442
  intermediate, 392
  of fractions, 270
  of quotients, 270
  of rational functions, 273
  perfect, 435
  prime subfield of, 279, 386
  splitting, 429
filtration, 689
finite field, 441
finite separable extensions are simple, 450
five-lemma, 185, 590
  short, 184, 589
fixed field, 455
flag of subspaces, 383
flat
  criterion for flatness, 513, 514
  free $\Longrightarrow$, 508
  is a local property, 513
  module, 507, 552, 662
  modules over local rings are free, 514
  projective $\Longrightarrow$, 548
forgetful functor, 485
  right-adjoint to free, 493
four-lemma, 184, 589
Fränkel, Abraham, 1
free
  abelian group, 75
  action, 109
  algebra, 167
  functor, left-adjoint to forgetful, 493
  group, 70
    universal property, 71
  locally, 556
  module, 167, 305
    basis of, 306
    homomorphisms of, 314
    is flat, 508
    is projective, 548
  product, 62
  resolution of a module, 343
freshman's dream, 439
Freyd-Mitchell embedding theorem, 156, 559, 588, 591
Frobenius, Georg
  homomorphism, 435
full
  functor, 488, 588
  subcategory, 26, 595
function, 9
  composition, 10
  identity, 9
  set-function, 19
functor, 61
  δ-functor, 660, 680
  additive, 485, 561, 613
  adjoint, 492, 498, 642
  bifunctor, 679
  category, 497
  cohomological, 685
  cohomology, 596
  cone of a, 489
  conservative, 520
  continuous, 493
  contravariant, 484
  covariant, 484
  derived, 509, 642, 645
  effaceable, 660
  equivalence of categories, 488
  essentially surjective, 488
  exact, 495, 588, 603, 613, 643
  Ext, 551
  faithful, 488, 587
  forgetful, 485, 493
  full, 488, 588
  fully faithful, 488
  natural
    isomorphism, 492
    transformation, 492, 636
  of points, 487, 581
  representable, 487, 497
  Tor, 509
fundamental theorem
  of algebra, 286, 468
  of arithmetic, 255
  on symmetric functions, 471

Galois, Evariste
  closure, 466
  criterion, 477
  extension, 391, 454, 458, 657
  field, 442
  group of a polynomial, 477

group of an extension, 459
inverse problem, 473
theory, 454
Gauss, Carl Friedrich, 286, 302
elimination, 322
lemma, 270, 273
Gaussian elimination, 322, 347, 376
over Euclidean domains, 323
Gaussian integers, 294
gcd, 252
general linear group, 321
germ of a rational function, 415
global dimension, 679
golden ratio, 425
graded
algebra, 527
homomorphisms, 528
module, 413, 527
ring, 413, 527
Gram-Schmidt process, 371
graph, 9
Grassmannian, 326, 545
Grayson, Dan, 266
greatest common divisor, 252
Grothendieck, Alexander, 413
group, 337, 605
spectral sequence, 695, 697
group, 29, 42
p-group, 188, 190
abelian, 45, 62
action, 64, 108
coinvariant, 658
invariants, 655
alternating, 220, 473
automorphism
inner, 86, 193
center of a, 189
class formula, 190
cohomology, 655
cyclic, 54, 67, 82, 103
described by generators and relations, 99
dihedral, 52, 225
dual, 241
finitely generated, 82
finitely presented, 99
free, 70
free abelian, 75
general linear, 321
generated by a subset, 52, 81
Grothendieck, 337, 605
homology, 658
homomorphism, 59
infinite cyclic, 67, 70
modular, 63, 95, 114
nilpotent, 213, 233
object, 115
of order p^2, 191
of order p^2q, 214
of order p^3, 240
of order pq, 201, 232
of order 8, 204, 240
of permutations, 50
of units in a ring, 123
opposite, 113
orthogonal, 370
presentation, 99
quaternionic, 88, 128, 234
quotient, by a normal subgroup, 93
quotient, by an equivalence relation, 90
ring, 127, 655
simple, 196, 205
solvable, 211, 475
symmetric, 50, 478
trivial, 42, 61
unitary, 370
groupoid, 29, 41

Hamilton, William Rowan, 128
Hasse, Helmut, 256
Hauptidealsatz, 259
Hausdorff, Felix
maximal principle, 264
height of a prime ideal, 259
hermitian
matrix, 371
product, 370
Hilbert, David
basis theorem, 172, 245, 407
Grand Hotel, 183
Nullstellensatz, 153, 171, 285, 400, 404
'theorem 90', 467, 658
Zahlbericht, 467
Hodges, Wilfrid, 265
Hom
functors, 486
is left-exact, not right-exact, 537
is right-adjoint, 536
homogeneous
element of a graded ring, 527
ideal, 528
homological algebra, 174, 495, 559
homology and cohomology, 178, 335, 592
computing Ext, 553
computing Tor, 509
homology of groups, 658
homomorphism
as Ext^0, 552
of field extensions, 386
of fields, 385
of groups, 59
of rings, 129
homomorphism of groups, 53
homotopy
between morphisms of complexes, 611

equivalence of complexes, 612
homotopy category of complexes, 618, 669
horseshoe lemma, 648
Hurewicz, Witold, 492
hydrogen atom, 366

IBN property, 310, 313
icosahedral group, 223
ideal, 138
$\iff$ kernel, 141
annihilator, 342
characteristic, 356
finitely generated, 145
homogeneous, 528
intersection, 146
irrelevant, 533
maximal, 150
prime, 150
minimal, 250, 267, 348
principal, 145
product, 146, 269
radical, 409
sum, 145
identity element, 42
image
isomorphic to coimage, 573
of a morphism, in an abelian category, 572
indeterminate, 125
induction, 265
principle of, 265
transfinite, 263
injective, 29
R-Mod has enough injectives, 548, 550
abelian group, 550
dimension, 679
enough injectives, 550, 620
function, 11, 12, 14
module, 546
object of an abelian category, 619
resolution, 629
resolution of a module, 551
inner product, 370
inseparable
degree, 439, 440
element of an extension, 436
polynomial, 433
purely, 439
integral domain, 122
Invariant Basis Number property, 310
invariant factors, 237, 354
invariants of a group action, 655
inverse, 42
is unique, 43
inverse Galois problem, 473
inverse limit, 489
irreducible, 247
$\implies$ prime in UFDs, 253
algebraic set, 415
factor, 251
module, 163
irrelevant ideal, 533
isomorphism, 27
$\iff$ monomorphism and epimorphism, in abelian categories, 566
of sets, 11
quasi-, 607
theorems, 97, 101, 142, 162

Jacobi, Carl Gustav Jacob
identity, 312
Jacobson, Nathan, 120
radical, 267
Jordan block, 378
Jordan canonical form, 379
Jordan-Hölder theorem, 206

kernel, 80, 132
$\iff$ ideal, in Ring, 141
$\iff$ monomorphism, in an abelian category, 564
$\iff$ normal subgroup, in Grp, 95
$\iff$ submodule, in R-Mod, 161
categorical definition, 490, 561
universal property, 81, 166
Koszul, Jean-Louis
complex, 348, 535
Kronecker-Weber theorem, 480, 482
Krull, Wolfgang, 498
dimension, 153, 250, 354
Hauptidealsatz, 259
theorem, 265
Kummer, Ernst, 468
theory, 465

Lagrange, Joseph Louis
four-square theorem, 299, 303
interpolation, 290
resolvent, 465
theorem, 103
Lamé, Gabriel, 280
lattice
of intermediate fields of an extension, 460
of subgroups, 84
of a quotient, 100
Laurent polynomial, 127, 656
leading coefficient, 147, 245
lemma
Artin-Rees, 535
five-, 185, 590
four-, 184, 589
Gauss's, 270
horseshoe, 648
Nakayama's, 164, 339, 357, 514

nine-, 185
Schur's, 163
snake, 179, 510, 553, 579, 597
Yoneda, 497, 591
Zorn's, 262, 265, 307, 311, 403, 549
Lichiardopol, Nicolas, 453
Lie, Sophus
algebra, 127, 312
bracket, 312
limit, 489
direct, 445, 491
injective, 491
inverse, 489
projective, 489
linear
map, 158
maps, similar, 361
polynomial, 284
transformation, 359
linearly dependent/independent, 306
Liouville, Joseph
theorem, 152
local
parameter, in a DVR, 260
ring, 278, 339, 357, 514, 556
localization, 270, 415, 681
as a functor, 496
is exact, 500
of a module, 277, 496
of a ring, 277
locally factorial, 279
locally free, 556
$\iff$ flat, 556
$\iff$ projective, 556
long exact sequence
in (co)homology, 600
of (co)homology, 180
of derived functors, 647, 651
of Ext, 552
of Tor, 509
Lüroth, Jacob
theorem, 400

Mac Lane, Saunders, 120
mapping
cone, 605, 623, 624, 665
in topology, 607
cylinder, 615
in topology, 615
Marcolli, Matilde, 419
Massey, William, 687
matrices, 43, 315
adjoint, 331
block, 322
cofactors of, 331
companion, 374
corresponding to homomorphisms, 316
diagonalizable, 368, 380
echelon form, 325
elementary, 321
elementary operations, 320
equivalent, 320
hermitian, 371
inverse, through determinants, 332
inverse, through Gaussian elimination, 338
Jordan canonical form of, 379
minor of, 330
normal, 383
of change of basis, 319
orthogonal, 370
rational canonical form of, 375
ring of, 121, 316
self-adjoint, 371
similar, 360
symmetric, 371
transpose, 329, 377, 541
triangular, 86
unitary, 370
maximal ideal, 150
every proper ideal is contained in a, 265
McKay, James, 195
Milne, James S., 473
minimal polynomial
divides the characteristic polynomial, 376
of a linear transformation, 365, 397
of an element of a field extension, 388
minor of a matrix, 330
Möbius transformation, 114
modular group, 63, 95, 114
module, 157
$\mathbb{Z}$-module $\iff$ abelian group, 158
algebra, 159
associated to a linear transformation, 372
bimodule, 517
cofree, 551
divisible, 550
dual, 537
finite, 171
finite length, 314
finitely generated, 169, 342
over a PID, 240, 349, 354
finitely presented, 342
flat, 507
free, 167, 306
injective, 546
irreducible, 163
localization, 277, 496
Noetherian, 170
over a group, 655
projective, 173, 546
quotient, 161
reflexive, 542
simple, 163, 174

submodule, 160
torsion, 340, 355
torsion-free, 340
monic polynomial, 147
monoid, 120, 126, 169
monomorphism, 29
in an additive/abelian category, 563
of groups, 84
of modules, 167
of rings, 132
of sets, 14
split, 178, 547
Morita, Kiiti
equivalent rings, 522
morphism, 18
trivial, 64
multilinear map, 522
alternating, 524
symmetric, 524
multiplication table, 48
multiplicative subset, 277
multiplicatively closed set, 277
multiplicity
of a root of a polynomial, 281
of an eigenvalue, algebraic, 366, 379
of an eigenvalue, geometric, 367, 379
multiset, 2, 10, 27, 207, 251

Nakayama, Tadashi
lemma, 164, 339, 357, 514
natural
isomorphism, 492
transformation, 492
natural projection, 15
natural transformation, 636
nilpotent, 128, 144, 279, 382, 409
group, 213, 233
nilradical, 144, 155, 267, 409
and associated primes, 348
nine-lemma, 185
Noether, Emmy
normalization theorem, 416
Noetherian
module, 170
ring, 145, 244, 267
non-zero-divisor, 122
nondegenerate bilinear map, 544
nonsingular bilinear map, 544
norm
in $\mathbb{H}$, 136
in $\mathbb{Q}(\sqrt{d})$, 154, 260
in $\mathbb{Z}[\sqrt{-2}]$, 301
in $\mathbb{Z}[\sqrt{-5}]$, 251
in $\mathbb{Z}[i]$, 295
of an element of a field extension, 398, 440, 467
is transitive, 440
normal
bundle, 543, 556
extension, 431
subgroup, 88
$\iff$ kernel, 95
normalizer, 191
nullity, 334
Nullstellensatz, 153, 171, 404
strong, 410
weak, 409
homogeneous, 533

object
final, 31
initial, 31
terminal, 31, 33
zero-, 61, 64, 159, 561
octahedral axiom, 683
operator, 359
opposite category, 26, 38, 484, 620
orbit, 110
order
of a group, 47
of a power series, 250
of an element, 46
ordered basis, 308
ordered pairs, 5
orthogonal
complement, 370
group, 370
vectors, 370
orthonormal set of vectors, 370

$\mathrm{PSL}_2(\mathbb{Z})$, 95, 114
p-adic
integers, 498
numbers, 498
partial fractions, 276
partition, 7
of a positive integer, 216, 224, 240
Pavlovian reaction, 97, 176, 505, 537, 566
perfect field, 435
permutation, 50
conjugate, 217
cycle notation, 215
even, odd, 219
type of, 216
PID, 145, 254
$\implies$ UFD, 254
characterization of, 259, 344, 347, 353, 358
classification of finitely generated modules over a, 240, 354
divisible over a $\iff$ injective, 550
pivot, 325
platonic solids, 57
Plücker, Julius

embedding, 532
pointed set, 24
polynomial, 125
constant, 125
content of a, 269
cyclotomic, 67, 289, 426, 446
derivative of a, 433
function, 131, 282
inseparable, 433
irreducible over a UFD, 275
Laurent, 127, 656
monic, 147
primitive, 268
root, 281
separable, 433
very primitive, 268
with noncommuting indeterminates, 169, 531
Pontryagin, Lev Semenovich
dual, 555
poset = partially ordered set, 261
power
series, 126, 498
order, 250
ring, Euclidean domain, 259
ring, Noetherian, 250
ring, not necessarily a UFD, 273
set, 4, 18, 127
preadditive category, 574
presentation
of a group, 99
of a module, 342
presheaf, 485, 575
on a topological space, 486
prime
$\Longrightarrow$ irreducible, 247
associated, of a module, 347
element in an integral domain, 247
ideal, 150
minimal, 250, 267, 348
nonzero ideals are maximal in a PID, 151
subfield, 279, 386
primitive
polynomial, 268
very, 268
prime divisor, 453
root of 1, 446
principal ideal domain, *see* PID
principle of induction, 263
product, 5, 489
direct, 61, 226
fibered, 39, 173
free, 62
in an abelian category, 567
of groups, 61
of modules, 165
of rings, 133
semidirect, 230
universal property, 35
projection, 172
formula, 521
projective
R-Mod has enough projectives, 548
and splitting sequences, 547
dimension, 678
enough projectives, 548, 620
flat $\Longrightarrow$, in some cases, 548
module, 173, 546
modules over local rings are free, 358, 556
object of an abelian category, 619
resolution, 629
resolution of a module, 548
projective space, 326, 413, 416
algebraic set in, 417, 533
coordinates in, 416
pull-back, 173
in an abelian category, 567
pure tensor, 504
purely inseparable, 439
push-out, 173
in an abelian category, 567

quantifier, 2
quasi-isomorphism, 592, 607
quaternion, 128, 136, 234, 302
integral, 302
norm, 136
quaternionic group, 88, 128, 234
quotient, 4, 7
group, 90, 93
in an abelian category, 564, 588
module, 161
ring, 140
universal property, 33

Rabinowitsch, J. L.
trick, 410
radical of an ideal, 409
rank
of a linear map, 334
of a matrix, 334
of a module, 311, 354
rational
canonical form, 375
function, 387
root test, 282
rational function, 273
reduced ring, 144, 412, 512
Rees, David
algebra, 164, 528
refinement
of a normal series, 209
reflexive module, 542, 555

regular polygon, constructibility of a, 418, 424, 426, 449, 469
regular sequence, 348, 535
relation, 6
  compatible with a group structure, 90
  equivalence, 7
  order, 261
representable functor, 497
resolution, 509, 592, 607, 629
  Cartan-Eilenberg, 651, 675
  fully projective, 651
  injective, 551, 629
  of a complex, 632, 672
  of a module, 343
  projective, 548, 629
restriction
  of scalars, 518
  of sections of a presheaf, 496
ring, 119
  Artinian, 250
  Boolean, 144, 156, 300
  center of a, 137, 159
  characteristic of, 141
  commutative, 121
  discrete valuation, 260, 278
  division, 123, 128
  domain with factorizations, 248
  Euclidean domain, 255
  factorial, 248
  graded, 413, 527
  homomorphism, 129
  ideal of, 138
  integral domain, 122
  Krull dimension of a, 153, 250, 354, 679
  local, 278, 339, 357, 514, 556
  localization, 270, 277, 415
  Noetherian, 145, 244, 267
  of Laurent polynomials, 127
  of matrices, 121
  of power series, 126, 250, 259, 273, 498
  of quaternions, 128, 136, 302
  PID, 145
  polynomial, 124
  quotient, 140
  reduced, 144, 412, 512
  simple, 144
  subring, 132
  UFD, 248
  zero-, 121
rng, 120, 139
roof diagram, 681, 682
root
  of 1, 445
  of a polynomial, 281
  primitive, 446
row space of a matrix, 333
Ruffini, Paolo, 477
$SO_3(\mathbb{R})$, 85
  is a quotient of $SU_2(\mathbb{C})$, 106
$SU_2(\mathbb{C})$, 106
  and quaternions, 136, 234
  is simply connected, 86
scalar, 308
  extension, 518
  restriction, 518
scheme, 413
Schreier's theorem, 209
Schubert, Hermann
  cells, 326
Schur, Issai, 473
  decomposition, 383
  form, 383
  lemma, 163
section, 13
  of a presheaf, 496
self-adjoint matrix, 371
semidirect product, 230
semigroup, 126
separable
  closure, 439
  degree, 436
  element of an extension, 436
  extension, 436
  polynomial, 433
series of subgroups, 205, 689
  abelian, 211
  composition, 206
  cyclic, 211
  derived, 211
  equivalent, 207
  normal, 205
Serre, Jean-Pierre
  problem, 556
set, 1
  -function, 19
  indexed, 10
  multiset, 10
  of parts, 4
  pointed, 24
  power, 4, 18, 127
Shafarevich, Igor, 473
sheaf, 486, 496, 575
  cohomology, 575
shift functor, 596
short exact sequence, 176
  of complexes, 597
similar
  endomorphisms, 361
  matrices, 360
simple
  group, 196, 205
    of order 60, 205, 222
  module, 163, 174
  of a double complex, 667

ring, 144
singleton, 2, 24
slice category, 24
Smith normal form, 324
snake lemma, 179, 510, 553, 579, 597
solvability of polynomial equations, 474
solvable
  extension, 475
  for $n \geq 5$, S_n is not, 224
  group, 211, 475
Spec R, 151, 413, 486
spectral
  decomposition, 367, 380
  theorem, 383
spectral sequence, 670, 686
  Grothendieck, 695, 697
  of a double complex, 691
  picture, 694
spectrum
  maximal, 155
  of a ring, 151, 413, 486
  of an operator, 365
split
  epimorphism, 178, 547
  exact complex, 621, 627
  exact sequence, 177
  monomorphism, 178, 547
splitting
  field, 429
  sequence, 177, 229, 547
stabilizer, 111, 187
standard basis, 316
Stark, Harold, 302
Steenrod, Norman, 18
straightedge-and-compass constructions, 417
subgroup
  p-Sylow, 196
  centralizer, 190, 191
  characteristic, 202
  commutator, 83, 96, 210, 226
  cosets of, 91
  cyclic, 82
  generated by a subset, 81
  index of, 102
  normal, 88
  normalizer, 191
  of a free group, 83
  of cyclic group, 82
  stabilizer, 111, 187
  transitive, of S_n, 225
submodule, 160
  $\Longleftrightarrow$ kernel, 161
  generated by a subset, 169
subobject classifier, 27
subring, 132
surjective function, 12
Sylow theorems
  first, 196
  second, 198
  third, 199
symmetric
  algebra, 529
  functions
    elementary, 472
    fundamental theorem, 471
  group, 50, 214, 478
    conjugacy in, 217
    is generated by transpositions, 219
  matrix, 371
  multilinear map, 524
  power, 524
symmetry, 52
systems of linear equations, 327

Tarski, Alfred, 262
tensor
  algebra, 529
  as Tor_0, 509
  by free modules is exact, 507
  commutes with direct sums, 505
  is associative, 523
  is left-adjoint to Hom, 505, 517
  is right-exact, not left-exact, 505, 507
  power, 523
  product, 501
  pure, 504
theorem
  Abel-Ruffini, 477
  Baer's criterion, 549
  Birkhoff-Vandiver, 453
  Burnside's, 214
  Cauchy's, 195
    abelian case, 107
  Cayley's, 110, 472
    for rings, 135
  Cayley-Hamilton, 365, 376
  characterizations of Galois extensions, 457
  Chinese remainder, 291
  classification of finite abelian groups, 237, 480
  classification of finitely generated modules over a PID, 354
  cohomology of total complexes, 674
  constructibility of the regular n-gon, 469
  derived functors are computed by acyclic resolutions, 663
  Dirichlet's, 454, 480
  Euler's, 107
  exactness of the total complex, 670
  Ext may be computed by resolving either argument, 677
  Feit-Thompson, 212, 214

Fermat's
  last, 280
  little, 103, 107, 439
  on sums of squares, 298
first isomorphism
  for groups, 97
  for modules, 162
  for rings, 142
fixed point, 188
Freyd-Mitchell embedding, 156, 559, 588, 591
fundamental
  of algebra, 285, 468
  of arithmetic, 255
  of Galois theory, 460, 461
  on symmetric functions, 471
Hilbert's
  '90', 467, 482, 658
  basis, 172, 245, 407
  Nullstellensatz, 153, 171, 405, 409, 410
Jordan-Hölder, 206, 313
Kronecker-Weber, 480, 482
Krull
  Hauptidealsatz, 259
Krull's, 265
Lagrange's, 103
Lagrange's four-square, 299, 303
Liouville's, 152
long exact sequence
  in cohomology, 600
  of derived functors, 651
Lüroth's, 400
Noether's normalization, 416
of nonsolvability by radicals, 474
on constructibility by straightedge and compass, 422
realization of the derived category, 635
Schreier's, 209
second isomorphism
  for groups, 101
  for modules, 162
  for rings, 142
spectral, for normal operators, 383
Sylow's
  first, 196
  second, 198
  third, 199
third isomorphism
  for groups, 101
  for modules, 162
  for rings, 142
Tor may be computed by resolving either argument, 676
Wedderburn's (little), 124, 137, 204, 441, 453
well-ordering, 263
Wilson's, 70, 225
Tor, 509, 646
  long exact sequence, 509
  may be computed by resolving either argument, 509, 676
  why it is called Tor, 513
torsion
  element, 340
  module, 340, 355
  submodule, 340
total complex
  conditions for exactness, 670
  of a double complex, 668
total complex of a double complex, 606
totient function, 87
trace
  of a matrix, 362
  of an element of a field extension, 398, 440, 467
    is transitive, 440
  of an endomorphism, 362
transcendence
  basis, 400
  degree, 400
transcendental
  π is, 394, 418, 426
  element of an extension, 391
  number, 394
  purely, extension, 400
transitive
  action, 110
  subgroup of S_n, 225, 478
transitivity of norm and trace, 440
transpose of a matrix, 329, 377, 541
transposition, 219
triangle
  distinguished, 602, 606, 650, 683
  exact, 601, 684
triangular matrix, 86
triangulated category, 602, 610, 618, 650, 683

UFD, 248
  $R[x]$ is if R is, 273
  $\mathbb{Z}[x]$ is a, 276
  characterization of, 253
  every PID is a, 254
  is not a local property, 279
  locally factorial, 279
  primes of height 1 are principal in a, 259
unique factorization domain, *see* UFD
unit, 123
unitary group, 370
universal
  identity, 332
  property, 31, 489
upper bound, 262

valuation
  Dedekind-Hasse, 256
  discrete, 260
  Euclidean, 256
variety, 407, 415
vector, 308, 316
  bundle, 358
  space, 158, 308
    basis, 306
    dimension, 311
    dual, 537
Venn, John
  diagram, 4
versal, 284, 388

Wedderburn, Joseph
  (little) theorem, 124, 137, 204, 441, 453
wedge power, 524
Weil, André
  conjectures, 369
well-defined, 16
well-ordering, 262
  principle, 83, 262
  theorem, 263
Weyl, Hermann
  algebra, 534
Wilson, John
  theorem, 70, 225
Witt, Ernst, 453
word, 72
word problem, 100
woset = well-ordered set, 262

Yoneda, Nobuo
  Ext, 557
  lemma, 497
Young diagram, 216

Zariski, Oscar
  topology, 408, 497
Zermelo, Ernst, 1
Zermelo-Fränkel, 1, 262
zero-divisor, 122
  and associated primes, 348
  non-, 122
zero-morphism, in an abelian category, 561, 574
zero-object, 61, 64, 159, 561
Zorn, Max
  lemma, 262, 265, 307, 311, 403, 549

Titles in This Series

104 **Paolo Aluffi,** Algebra: Chapter 0, 2009

103 **Branko Grünbaum,** Configurations of points and lines, 2009

102 **Mark A. Pinsky,** Introduction to Fourier analysis and wavelets, 2009

101 **Ward Cheney and Will Light,** A course in approximation theory, 2009

100 **I. Martin Isaacs,** Algebra: A graduate course, 2009

99 **Gerald Teschl,** Mathematical methods in quantum mechanics: With applications to Schrödinger operators, 2009

98 **Alexander I. Bobenko and Yuri B. Suris,** Discrete differential geometry: Integrable structure, 2008

97 **David C. Ullrich,** Complex made simple, 2008

96 **N. V. Krylov,** Lectures on elliptic and parabolic equations in Sobolev spaces, 2008

95 **Leon A. Takhtajan,** Quantum mechanics for mathematicians, 2008

94 **James E. Humphreys,** Representations of semisimple Lie algebras in the BGG category $\mathcal{O}$, 2008

93 **Peter W. Michor,** Topics in differential geometry, 2008

92 **I. Martin Isaacs,** Finite group theory, 2008

91 **Louis Halle Rowen,** Graduate algebra: Noncommutative view, 2008

90 **Larry J. Gerstein,** Basic quadratic forms, 2008

89 **Anthony Bonato,** A course on the web graph, 2008

88 **Nathanial P. Brown and Narutaka Ozawa,** C*-algebras and finite-dimensional approximations, 2008

87 **Srikanth B. Iyengar, Graham J. Leuschke, Anton Leykin, Claudia Miller, Ezra Miller, Anurag K. Singh, and Uli Walther,** Twenty-four hours of local cohomology, 2007

86 **Yulij Ilyashenko and Sergei Yakovenko,** Lectures on analytic differential equations, 2007

85 **John M. Alongi and Gail S. Nelson,** Recurrence and topology, 2007

84 **Charalambos D. Aliprantis and Rabee Tourky,** Cones and duality, 2007

83 **Wolfgang Ebeling,** Functions of several complex variables and their singularities (translated by Philip G. Spain), 2007

82 **Serge Alinhac and Patrick Gérard,** Pseudo-differential operators and the Nash–Moser theorem (translated by Stephen S. Wilson), 2007

81 **V. V. Prasolov,** Elements of homology theory, 2007

80 **Davar Khoshnevisan,** Probability, 2007

79 **William Stein,** Modular forms, a computational approach (with an appendix by Paul E. Gunnells), 2007

78 **Harry Dym,** Linear algebra in action, 2007

77 **Bennett Chow, Peng Lu, and Lei Ni,** Hamilton's Ricci flow, 2006

76 **Michael E. Taylor,** Measure theory and integration, 2006

75 **Peter D. Miller,** Applied asymptotic analysis, 2006

74 **V. V. Prasolov,** Elements of combinatorial and differential topology, 2006

73 **Louis Halle Rowen,** Graduate algebra: Commutative view, 2006

72 **R. J. Williams,** Introduction the the mathematics of finance, 2006

71 **S. P. Novikov and I. A. Taimanov,** Modern geometric structures and fields, 2006

70 **Seán Dineen,** Probability theory in finance, 2005

69 **Sebastián Montiel and Antonio Ros,** Curves and surfaces, 2005

68 **Luis Caffarelli and Sandro Salsa,** A geometric approach to free boundary problems, 2005

67 **T.Y. Lam,** Introduction to quadratic forms over fields, 2004

66 **Yuli Eidelman, Vitali Milman, and Antonis Tsolomitis,** Functional analysis, An introduction, 2004

65 **S. Ramanan,** Global calculus, 2004

64 **A. A. Kirillov,** Lectures on the orbit method, 2004

63 **Steven Dale Cutkosky,** Resolution of singularities, 2004

62 **T. W. Körner,** A companion to analysis: A second first and first second course in analysis, 2004

61 **Thomas A. Ivey and J. M. Landsberg,** Cartan for beginners: Differential geometry via moving frames and exterior differential systems, 2003

60 **Alberto Candel and Lawrence Conlon,** Foliations II, 2003

59 **Steven H. Weintraub,** Representation theory of finite groups: algebra and arithmetic, 2003

58 **Cédric Villani,** Topics in optimal transportation, 2003

57 **Robert Plato,** Concise numerical mathematics, 2003

56 **E. B. Vinberg,** A course in algebra, 2003

55 **C. Herbert Clemens,** A scrapbook of complex curve theory, second edition, 2003

54 **Alexander Barvinok,** A course in convexity, 2002

53 **Henryk Iwaniec,** Spectral methods of automorphic forms, 2002

52 **Ilka Agricola and Thomas Friedrich,** Global analysis: Differential forms in analysis, geometry and physics, 2002

51 **Y. A. Abramovich and C. D. Aliprantis,** Problems in operator theory, 2002

50 **Y. A. Abramovich and C. D. Aliprantis,** An invitation to operator theory, 2002

49 **John R. Harper,** Secondary cohomology operations, 2002

48 **Y. Eliashberg and N. Mishachev,** Introduction to the h-principle, 2002

47 **A. Yu. Kitaev, A. H. Shen, and M. N. Vyalyi,** Classical and quantum computation, 2002

46 **Joseph L. Taylor,** Several complex variables with connections to algebraic geometry and Lie groups, 2002

45 **Inder K. Rana,** An introduction to measure and integration, second edition, 2002

44 **Jim Agler and John E. McCarthy,** Pick interpolation and Hilbert function spaces, 2002

43 **N. V. Krylov,** Introduction to the theory of random processes, 2002

42 **Jin Hong and Seok-Jin Kang,** Introduction to quantum groups and crystal bases, 2002

41 **Georgi V. Smirnov,** Introduction to the theory of differential inclusions, 2002

40 **Robert E. Greene and Steven G. Krantz,** Function theory of one complex variable, third edition, 2006

39 **Larry C. Grove,** Classical groups and geometric algebra, 2002

38 **Elton P. Hsu,** Stochastic analysis on manifolds, 2002

37 **Hershel M. Farkas and Irwin Kra,** Theta constants, Riemann surfaces and the modular group, 2001

36 **Martin Schechter,** Principles of functional analysis, second edition, 2002

35 **James F. Davis and Paul Kirk,** Lecture notes in algebraic topology, 2001

34 **Sigurdur Helgason,** Differential geometry, Lie groups, and symmetric spaces, 2001

33 **Dmitri Burago, Yuri Burago, and Sergei Ivanov,** A course in metric geometry, 2001

For a complete list of titles in this series, visit the
AMS Bookstore at **www.ams.org/bookstore/**.



Courtesy of Ryals Lee/FSU Photo Lab.

Algebra: Chapter 0 is a self-contained introduction to the main topics of algebra, suitable for a first sequence on the subject at the beginning graduate or upper undergraduate level. The primary distinguishing feature of the book, compared to standard textbooks in algebra, is the early introduction of categories, used as a unifying theme in the presentation of the main topics. A second feature consists of an emphasis on homological algebra: basic notions on complexes are presented as soon as modules have been introduced, and an extensive last chapter on homological algebra can form the basis for a follow-up introductory course on the subject. Approximately 1,000 exercises both provide adequate practice to consolidate the understanding of the main body of the text and offer the opportunity to explore many other topics, including applications to number theory and algebraic geometry. This will allow instructors to adapt the textbook to their specific choice of topics and provide the independent reader with a richer exposure to algebra. Many exercises include substantial hints, and navigation of the topics is facilitated by an extensive index and by hundreds of cross-references.

ISBN 978-0-8218-4781-7

9 780821 847817

GSM/104

For additional information and updates on this book, visit

www.ams.org/bookpages/gsm-104

AMS *on the* Web
www.ams.org