Mathematics from Categories

From counting to higher toposes

Anonymous

Contents

I Sets, maps, and universal properties 1

1 Functions, orders, and composition 1
- 1.1 Arrows and their laws . . . 1
- 1.2 Order and paths . . . 2
- 1.3 Inverses and opposite categories . . . 3
- 1.4 Cancellation in place of elements . . . 4

2 Counting with sums and products 6
- 2.1 Pairs and the universal product . . . 6
- 2.2 Alternatives and the universal coproduct . . . 7
- 2.3 The empty cases and finite counting . . . 8
- 2.4 Distributivity and the binomial theorem . . . 9
- 2.5 The same properties in an order . . . 10
- 2.6 Inclusion–exclusion on the subset lattice . . . 10

3 Solving equations and forming quotients 12
- 3.1 Equalizers . . . 12
- 3.2 Pullbacks . . . 13
- 3.3 Fibers, intersections, and the diagonal . . . 14
- 3.4 Coequalizers and generated equivalences . . . 15
- 3.5 Pushouts and gluing . . . 16
- 3.6 Finite limits and colimits . . . 17

4 Recovering objects from their maps 20
- 4.1 Functors . . . 20
- 4.2 Naturality . . . 21
- 4.3 Presheaves and representable functors . . . 21
- 4.4 The Yoneda lemma . . . 22
- 4.5 Recovering objects and arrows . . . 24
- 4.6 Products as representing objects . . . 25
- 4.7 Representations for maps out of an object . . . 25

5 Adjunctions and free constructions 27
- 5.1 Words and their universal property . . . 27
- 5.2 The hom-set bijection . . . 29
- 5.3 Unit, counit, and the triangle identities . . . 30
- 5.4 Constructing a left adjoint from universal maps . . . 32
- 5.5 Generators and relations . . . 32
- 5.6 Character pairings and Pontryagin duality . . . 39

II Numbers and symmetry 41

6 Natural numbers from recursion 41
- 6.1 Zero, successor, and the recursion map . . . 41

6.2 Induction from initiality . . . 42
6.3 Recursion with parameters . . . 42
6.4 Addition and cancellation . . . 43
6.5 Multiplication and distributivity . . . 44
6.6 The initial algebra for one more step . . . 44

7 Integers from group completion **46**
7.1 Addition and formal differences . . . 46
7.2 The group and its universal map . . . 47
7.3 Naturality and cancellation . . . 48
7.4 Normal forms for the integers . . . 50
7.5 Multiplying formal differences . . . 51
7.6 A completion that forgets a flag . . . 51

8 Groups from operations and equations **53**
8.1 The operations of a group . . . 53
8.2 The free group as a quotient of terms . . . 54
8.3 Group presentations . . . 55
8.4 Cyclic and triangle symmetries . . . 56
8.5 Groups as product-preserving models . . . 57
8.6 Finite characters and Pontryagin duality . . . 59

9 Symmetry from group actions **61**
9.1 Actions as functors . . . 61
9.2 Orbits, stabilizers, and cosets . . . 62
9.3 The action groupoid . . . 63
9.4 Fixed points as a limit and orbits as a colimit . . . 64
9.5 Burnside's Lemma . . . 65
9.6 Four positions and two colors . . . 66
9.7 Cyclic necklaces and dihedral bracelets . . . 67
9.8 From group actions to algebra actions . . . 68

III Rings, modules, and linear algebra 74

10 Rings and algebras from generators and relations **74**
10.1 Unital rings and their scalars . . . 74
10.2 The free associative algebra . . . 75
10.3 Commutative polynomials from finite multisets . . . 76
10.4 Imposing equations by a quotient . . . 77
10.5 A presentation for upper triangular matrices . . . 78

11 Ideals, quotients, and localization **81**
11.1 Ideals and the equations of a quotient . . . 81
11.2 Independent quotient conditions . . . 82
11.3 Constructing fractions . . . 84
11.4 Adjoining inverse elements by relations . . . 85
11.5 Rational numbers and a zero-divisor example . . . 86

12 Ring and algebra actions on modules **87**
12.1 Module maps and regular actions . . . 87
12.2 Two commuting scalar actions . . . 88
12.3 An algebra action specified by an operator . . . 88
12.4 Free modules from arbitrary assignments . . . 89
12.5 Direct sums and biproducts . . . 90

13 Linear maps and module presentations **92**
13.1 Matrices and chosen bases . . . 92
13.2 Kernels and cokernels . . . 92
13.3 Image and coimage . . . 94
13.4 Exactness and rank–nullity . . . 95
13.5 Generators, relations, and maps out . . . 96
13.6 An integer presentation in two coordinates . . . 97
13.7 Counting spanning trees by a determinant . . . 98

14 Tensor products, duals, and determinants **99**
14.1 A module representing bilinear maps . . . 99
14.2 Changing inputs and regrouping tensors . . . 100
14.3 Tensor bases and torsion . . . 101
14.4 Duals and linear maps . . . 102
14.5 Pontryagin duality for finite modules . . . 103
14.6 Alternating maps and the top exterior power . . . 104
14.7 Determinants from functoriality . . . 105

15 Restriction and extension of scalars **106**
15.1 Restriction and extension . . . 106
15.2 Coextension of scalars . . . 107
15.3 Successive extensions . . . 109
15.4 Changing the coefficients of a presentation . . . 109
15.5 Inverting one scalar . . . 111

IV Space, analysis, and probability **112**

16 Topologies from universal maps **112**
16.1 Topologies ordered by their open sets . . . 112
16.2 Initial topologies . . . 113
16.3 Final topologies . . . 114
16.4 Quotients and gluing . . . 115
16.5 A space that classifies open sets . . . 117
16.6 Pontryagin duality for locally compact abelian groups . . . 117

17 Spaces from their open sets **120**
17.1 Frames and their maps . . . 120
17.2 Points of a frame . . . 121
17.3 The opens–points adjunction . . . 122
17.4 Spatial locales and sober spaces . . . 123
17.5 An algebra of opens with no points . . . 124

18 Real numbers from completion **126**
18.1 Cauchy sequences and rational bounds . . . 126
18.2 An irrational number by bisection . . . 128
18.3 Completing a metric space . . . 128
18.4 The universal extension . . . 129
18.5 Operations and the choice of distance . . . 131
18.6 Equalizers and continuous maps . . . 131

19 Differentiation from the tangent functor **133**
19.1 The linear term of a map . . . 133
19.2 The tangent functor and its projection . . . 134
19.3 Products and the pullback of covectors . . . 135

19.4 Velocities as algebra maps . . . 136
19.5 A two-parameter model . . . 138

20 Probability from stochastic maps **139**
20.1 Finite distributions and substitution . . . 139
20.2 Stochastic maps and matrix multiplication . . . 140
20.3 Independent channels and joint outputs . . . 141
20.4 Two factorizations and Bayes' rule . . . 142
20.5 From finite sums to kernels . . . 144
20.6 Combinatorics of walks on a cycle . . . 144

V Logic, semantics, and quantification 146

21 Propositions from subobjects **146**
21.1 Subobjects and their order . . . 146
21.2 Truth and conjunction . . . 147
21.3 Changing the context . . . 148
21.4 Equality and a calculation in context . . . 148
21.5 Boolean operations in sets . . . 149
21.6 Monotone predicates and Boolean antichains . . . 150
21.7 The scope of a classifier . . . 151

22 The logic–semantics and structure–semantics adjunctions **152**
22.1 Satisfaction and semantic closure . . . 152
22.2 Equations as a category of operations . . . 153
22.3 Models and natural operations . . . 155
22.4 The structure–semantics adjunction . . . 156
22.5 Recovering the theory from its models . . . 157
22.6 Pontryagin duality and linear satisfaction . . . 159

23 Internal logic from elementary toposes **160**
23.1 A universal predicate . . . 160
23.2 Conjunction and implication . . . 161
23.3 Testing a family of truth values . . . 161
23.4 Falsity and disjunction from the universal test . . . 162
23.5 Truth at two stages . . . 163
23.6 Implication and negation in context . . . 164

24 Quantifiers from pullback **166**
24.1 The two adjoints of inverse image . . . 166
24.2 Substituting parameters under a quantifier . . . 167
24.3 Universal quantification in a topos . . . 168
24.4 Existential quantification in a topos . . . 169
24.5 Graphs, arbitrary maps, and images . . . 170

25 Extending diagrams by universality **172**
25.1 Restriction and its adjoints . . . 172
25.2 Which arrows contribute at an object? . . . 173
25.3 Collapsing a diagram and filling a missing corner . . . 175
25.4 Quantifiers and families on discrete categories . . . 176

VI Type theory and proof 178

26 Simply typed syntax and free cartesian closed categories 178
26.1 Function types from exponentials . . . 178
26.2 Terms in context . . . 179
26.3 Equations and their interpretation . . . 180
26.4 The syntactic cartesian closed category . . . 180
26.5 A calculation with a function returning a pair . . . 182
26.6 Pontryagin duality and a finite character type . . . 182

27 Dependent types from families 184
27.1 Contexts and their extensions . . . 184
27.2 Dependent pairs . . . 185
27.3 Dependent functions . . . 186
27.4 Binding after substitution . . . 187
27.5 Families in other categories . . . 188
27.6 Base change in slices . . . 188

28 Inductive types from initial algebras 190
28.1 Constructors and recursion . . . 190
28.2 Folds on lists . . . 191
28.3 Dependent induction as a section . . . 192
28.4 Well-founded trees . . . 193
28.5 Counting binary trees by their constructors . . . 195
28.6 Induction for trees and change of parameters . . . 196

29 Identity types and universes 197
29.1 Identity and its elimination rule . . . 197
29.2 Transport and path operations . . . 198
29.3 Equality in the set model . . . 199
29.4 Identity witnesses in a groupoid . . . 199
29.5 Universes as families of types . . . 201
29.6 The comparison with equivalences . . . 201

30 Proofs as terms 203
30.1 Assumptions and implication . . . 203
30.2 Conjunction, disjunction, and their cases . . . 204
30.3 When truth forgets a proof . . . 205
30.4 Quantifiers and the information in witnesses . . . 206
30.5 A proof that constructs a function . . . 207

VII Computation and programming languages 208

31 Program syntax and compiler correctness 208
31.1 The initial algebra of expressions . . . 208
31.2 Evaluation and substitution . . . 209
31.3 Stack programs as composable arrows . . . 210
31.4 Correctness with a continuation . . . 210
31.5 An execution trace . . . 212
31.6 Stack height and semantic equality . . . 212
31.7 Pontryagin duality and fast Fourier programs . . . 213

32 Recursion from least fixed points 215
32.1 Ordering partial answers . . . 215
32.2 A domain of functions . . . 216

32.3 Taking the least solution . . . 216
32.4 Recursion in a context . . . 217
32.5 Moving a map through recursion . . . 218
32.6 A loop and its finite runs . . . 219

33 Computational effects from monads 221
33.1 Unit and multiplication from an adjunction . . . 221
33.2 Sequencing computations . . . 222
33.3 Exceptions as extra results . . . 223
33.4 Reading and updating state . . . 223
33.5 Which state is visible on failure? . . . 224

34 Automata and streams from coalgebras 226
34.1 States and behavior maps . . . 226
34.2 Streams as a final coalgebra . . . 227
34.3 Languages as a final coalgebra . . . 228
34.4 Coinduction by a relation . . . 229
34.5 The reachable quotient . . . 230
34.6 Combinatorics of automata and transfer matrices . . . 232

35 Computability and undecidability 233
35.1 Partial maps as computations . . . 233
35.2 A universal evaluator . . . 234
35.3 Specializing a program . . . 235
35.4 The diagonal obstruction . . . 236
35.5 Extensional properties of programs . . . 237
35.6 Programs as witnesses . . . 238

36 Realizability and the effective topos 239
36.1 One program for every instance . . . 239
36.2 Substitution and quantification . . . 240
36.3 Assemblies and tracked functions . . . 241
36.4 Images and quotient maps of assemblies . . . 242
36.5 Adjoining quotients . . . 243
36.6 The computable maps on natural numbers . . . 244

VIII Machine learning and statistics 246

37 Data, predictors, and empirical risk 246
37.1 A sample and its empirical law . . . 246
37.2 Multisets and pooling . . . 247
37.3 Averaging the loss . . . 247
37.4 Minimizing empirical risk . . . 249
37.5 Population risk and conditional decisions . . . 250

38 Regression and kernel methods 251
38.1 The adjoint of an evaluation map . . . 251
38.2 Least squares as an orthogonal projection . . . 252
38.3 Ridge regression . . . 252
38.4 Feature vectors and the representer calculation . . . 253
38.5 Kernels as inner-product data . . . 254
38.6 Tangent features of a parametrized model . . . 255
38.7 Pontryagin duality and stationary kernels . . . 256

39 Neural networks from parameterized maps **258**
39.1 A category of parameterized maps . . . 258
39.2 Parallel maps and shared parameters . . . 259
39.3 Affine layers and an exact network . . . 260
39.4 Attention as a weighted evaluation . . . 261

40 Backpropagation from reverse differentiation **263**
40.1 The reverse arrow of a smooth map . . . 263
40.2 Products and the cotangent of copying . . . 264
40.3 The reverse sweep on a finite graph . . . 265
40.4 A worked network with shared parameters . . . 266
40.5 Differentials and gradient updates . . . 267

41 Optimization as a state update **268**
41.1 Turning a differential into a step . . . 268
41.2 From a model decrease to a loss decrease . . . 269
41.3 Composing updates with memory . . . 270
41.4 Stochastic steps and their variance . . . 270

42 Bayesian learning from stochastic kernels **272**
42.1 An experiment and its posterior . . . 272
42.2 Learning from a sequence . . . 273
42.3 Posterior predictive laws and decisions . . . 274
42.4 Learning after coarsening the data . . . 275
42.5 A conjugate update for a continuum of coins . . . 276

43 Generalization from sampling assumptions **278**
43.1 Sampling and the law of a score . . . 278
43.2 A bound for one predictor . . . 279
43.3 Selecting from a finite class . . . 280
43.4 Identical marginals and shared draws . . . 281
43.5 Validation after training . . . 282
43.6 Changing the population law . . . 282
43.7 Combinatorics of shattering and learning . . . 283

IX Sheaves, descent, schemes, and classifying toposes **285**

44 Sheaves from compatible local data **285**
44.1 Restriction and gluing . . . 285
44.2 The stalk at a point . . . 286
44.3 A space of germs . . . 287
44.4 Sheafification and its universal property . . . 288
44.5 Local square roots on a circle . . . 289

45 Grothendieck toposes from sites **291**
45.1 Sieves and their pullbacks . . . 291
45.2 The sheaf test is an extension property . . . 292
45.3 Sheafification by refining covers . . . 293
45.4 Why sheaves form a topos . . . 295
45.5 Maps between toposes . . . 296

46 Bundles from descent **297**
46.1 Reindexing a family . . . 297
46.2 Descent along a surjection of sets . . . 297
46.3 Bundle data on an open cover . . . 298

46.4 Transition functions and a twisted line . . . 300
46.5 Transporting descent data . . . 301
46.6 Counting colorings by descent . . . 302

47 Schemes from functors of points **303**
47.1 Testing equations in a ring . . . 303
47.2 Prime ideals and distinguished opens . . . 303
47.3 The sheaf of local fractions . . . 304
47.4 Affine maps and base change . . . 306
47.5 Gluing affine charts . . . 307
47.6 Affine tests determine maps . . . 308

48 Galois theory from finite étale schemes **309**
48.1 Finite algebras that split over a closure . . . 309
48.2 The fiber functor and the Galois action . . . 310
48.3 Recovering fields and their maps . . . 311
48.4 Restriction and factorization over finite fields . . . 313
48.5 Reconstructing the group from the functor . . . 314
48.6 Pontryagin duality for profinite abelian groups . . . 315

49 Classifying toposes from geometric theories **317**
49.1 Formulas that travel by inverse image . . . 317
49.2 The universal model and its direction . . . 318
49.3 The site of definable objects . . . 319
49.4 A topos classifying a single proposition . . . 320
49.5 What the classifier records . . . 322

X Homology, cohomology, and derived categories **323**

50 Exactness from kernels and cokernels **323**
50.1 An abelian category . . . 323
50.2 Short exact sequences and splitting . . . 324
50.3 The connecting map . . . 324
50.4 The snake lemma . . . 325
50.5 Exactness of functors . . . 326

51 Homology from chain complexes **328**
51.1 Cycles, boundaries, and maps . . . 328
51.2 What a chain homotopy preserves . . . 329
51.3 The connecting map . . . 330
51.4 Exactness across the degrees . . . 331
51.5 Counting graph cycles with an incidence complex . . . 332
51.6 Changing coefficients . . . 333
51.7 Cochain indexing . . . 334

52 Derived categories from localization **335**
52.1 Complexes, homotopies, and quasi-isomorphisms . . . 335
52.2 Shifts and cones . . . 336
52.3 Projective complexes compute maps . . . 336
52.4 Localization and roofs . . . 338
52.5 Exact triangles . . . 339
52.6 What derived tensor remembers . . . 340
52.7 Pontryagin duality on complexes with finite cohomology . . . 341

53 Sheaf cohomology from global sections **343**
53.1 Global sections and local lifts . . . 343
53.2 Injectives from stalks . . . 344
53.3 Deriving global sections . . . 345
53.4 The connecting homomorphism . . . 346
53.5 Computing on a finite cycle . . . 347
53.6 The Mayer–Vietoris sequence . . . 348
53.7 Twisting sheaves on the projective line . . . 348

XI Homotopy, higher categories, and spectra **350**

54 Fundamental groupoids from paths **350**
54.1 From continuous paths to arrows . . . 350
54.2 Maps, homotopies, and basepoints . . . 351
54.3 Winding around a circle . . . 352
54.4 Gluing across a disconnected overlap . . . 353
54.5 What passes to the groupoid . . . 355

55 Coherence in two dimensions **356**
55.1 Two compositions of transformations . . . 356
55.2 Composing spans by pullback . . . 357
55.3 Comparing parenthesizations and units . . . 358
55.4 The pentagon and the triangle . . . 359
55.5 Counting finite span witnesses . . . 360
55.6 From equations to comparison cells . . . 361

56 The infinity-category of spaces **362**
56.1 The maps to invert . . . 362
56.2 Simplices and horn filling . . . 363
56.3 Mapping spaces and coherent composition . . . 364
56.4 Localization and comparison with topology . . . 366
56.5 The Möbius function as an Euler characteristic . . . 367
56.6 The fundamental groupoid as truncation . . . 367

57 Homotopy limits and colimits **370**
57.1 Universality in mapping spaces . . . 370
57.2 Pullbacks from paths with prescribed endpoints . . . 371
57.3 Fibers, loops, and base change . . . 372
57.4 Pushouts from cylinders on the overlap . . . 373
57.5 Suspension and the loop adjunction . . . 373

58 Spectra from stabilization **375**
58.1 Suspension and loops . . . 375
58.2 Spectra as compatible deloopings . . . 375
58.3 From a space to its suspension spectrum . . . 376
58.4 Why exactness takes the form of a square . . . 377
58.5 The stable homotopy groups of a sphere . . . 378
58.6 Mapping cones as stable cofibers . . . 379
58.7 Pontryagin duality for finite homotopy groups . . . 380

XII Higher toposes, type theory, and derived geometry 382

59 Higher toposes from sheaves of spaces 382
59.1 Testing descent by a covering sieve . . . 382
59.2 Why a cover produces a Čech diagram . . . 383
59.3 Gluing a space of sections on a finite site . . . 384
59.4 Sheafification and the higher topos axioms . . . 385
59.5 What truncation recovers . . . 386
59.6 Pontryagin duality as an internal character sheaf . . . 387
59.7 Hyperdescent and hypercompletion . . . 388

60 Stacks and moduli from groupoid-valued descent 390
60.1 Descent data and their isomorphisms . . . 390
60.2 A descent condition for 1-types . . . 391
60.3 The stack of torsors . . . 392
60.4 Classifying covers of the circle . . . 393
60.5 Quotient stacks and stabilizers . . . 394
60.6 Counting objects with higher symmetries . . . 395
60.7 What a scheme quotient remembers . . . 395

61 Homotopy type theory from spaces and fibrations 397
61.1 Families over a context . . . 397
61.2 Dependent pairs and dependent functions . . . 398
61.3 Dependent binding commutes with substitution . . . 399
61.4 Identity by a path object . . . 399
61.5 Univalence and the universal family . . . 401
61.6 The circle as a higher inductive type . . . 402

62 Derived geometry from homotopy pullbacks 404
62.1 Rings with higher homotopy . . . 404
62.2 Testing an intersection by maps from affines . . . 405
62.3 Free resolutions and higher intersection groups . . . 406
62.4 Three intersections over a field . . . 407
62.5 Infinitesimal tests and the cotangent complex . . . 407
62.6 Localizing and gluing derived affines . . . 408

Appendices 410

A Diagram notation and duality 410
A.1 Reading the routes . . . 410
A.2 Reversing a category . . . 411
A.3 Dual universal properties . . . 411
A.4 Variance and two directions of reversal . . . 412
A.5 Commuting up to specified homotopy . . . 413

B Size, universes, and smallness assumptions 415
B.1 A universe for the data . . . 415
B.2 Small categories and small hom sets . . . 416
B.3 Functor categories and presheaves . . . 416
B.4 Small diagrams in limits and Kan extension . . . 417
B.5 A solution set for an adjoint . . . 418
B.6 Spaces, universes of types, and localization . . . 418

C Worked proofs 420
C.1 Taking mates across adjunctions . . . 420

C.2 Identifying the base-change map 421
C.3 The norm on invariants and coinvariants 422
C.4 A derived intersection in several coordinates 423

D Index of universal properties and their diagrams 425
D.1 Maps, limits, and colimits 425
D.2 Algebra and symmetry 426
D.3 Topology, analysis, and learning 427
D.4 Logic and type formation 428
D.5 Recursion and behavior 428
D.6 Local objects and derived algebra 429
D.7 Spaces and higher geometry 429

E Index of symbols and notation 431

F Index of concepts and definitions 438

Bibliography 453

Part I Sets, maps, and universal properties

CHAPTER 1

Functions, orders, and composition

A function $f\colon A \to B$ has a source A and a target B. From a second function $g\colon B \to C$ we obtain $g \circ f\colon A \to C$ by applying f and then g. The source and target tell us whether this composition is defined. For any function $h\colon C \to D$, both ways of grouping the three steps send $a \in A$ to $h(g(f(a)))$.

There are other ways to compose arrows. In an order, two successive comparisons give a third, as $2 \mid 6$ and $6 \mid 30$ give $2 \mid 30$. In a directed graph, a path ending where another begins can be followed by that second path. Functions, comparisons, and paths all have sources, targets, and an identity step that leaves composition unchanged. A category describes these common rules. See Leinster [1, § 1.1] for the category axioms and their basic examples.

1.1 Arrows and their laws

Definition 1.1. A *category* $\mathcal{C}$ has objects and, for each pair of objects A, B, a collection $\mathrm{Hom}_{\mathcal{C}}(A, B)$ of arrows from A to B. It assigns an identity $\mathrm{id}_A\colon A \to A$ to each object and a composite $g \circ f\colon A \to C$ to every pair $f\colon A \to B$, $g\colon B \to C$. The following equations hold whenever the arrows have the indicated types.

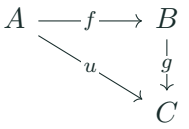
$$(h \circ g) \circ f = h \circ (g \circ f) \quad (h\colon C \to D),$$
$$f \circ \mathrm{id}_A = f = \mathrm{id}_B \circ f \quad (f\colon A \to B).$$

An arrow is also called a *morphism*.

For a fixed pair A, B we compare arrows by equality in $\mathrm{Hom}_{\mathcal{C}}(A, B)$, also written $\mathcal{C}(A, B)$. The identities are forced by the two identity equations. Indeed, if $e, e'\colon A \to A$ both act as identities, then $e \circ e' = e$ and $e \circ e' = e'$. Associativity allows us to speak of the composite of a path of any finite length, with the identity as the composite of a length-zero path.

Example 1.2 (Sets and functions)**.** The category **Set** has sets as objects and functions as arrows. Composition and identities are ordinary function composition and $a \mapsto a$. The functions between any two fixed sets form a set. A category is *locally small* when this holds for the arrows between every fixed pair of objects. The category **FinSet** uses finite sets as its objects and all functions between them.

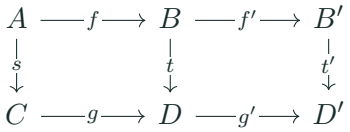
In a diagram, a path is read in the direction of its arrows. The following triangle *commutes* when the two paths from A to C have equal composites.



$$u = g \circ f.$$

For instance, write $[n] = \{0, 1, \ldots, n-1\}$ when $n > 0$, and $[0] = \varnothing$. Let $r\colon [6] \to [3]$ take the remainder modulo 3, and let $s\colon [3] \to [2]$ take the remainder modulo 2. Then $s \circ r$ has values $0, 1, 0, 0, 1, 0$ in order. Taking u to have this list makes the triangle commute. Changing the value of u at 3 changes the composite represented by its edge, so the two paths can be distinguished at that input.

A square presents two length-two paths with common endpoints. In the diagram below the two squares share the arrow $t\colon B \to D$.



If both squares commute, the left square gives $t \circ f = g \circ s$ and the right square gives $t' \circ f' = g' \circ t$. These equations give equality along the outside boundary.

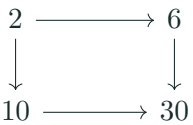
$$t' \circ (f' \circ f) = (t' \circ f') \circ f = (g' \circ t) \circ f = g' \circ (t \circ f) = (g' \circ g) \circ s.$$

This calculation is diagram pasting. Each equality follows from one of the given squares or from associativity.

1.2 Order and paths

Every partially ordered set $(P, \leq)$ gives a category. Its objects are the elements of P, and there is exactly one arrow $a \to b$ when $a \leq b$. Reflexivity gives the identities, and transitivity gives composition. Since there is at most one arrow between two fixed objects, parallel paths always have equal composites. Such categories are called *thin*.

For the positive integers ordered by divisibility, $m \to n$ means $m \mid n$. The two paths in



both give the divisibility arrow from 2 to 30. The intermediate numbers 6 and 10 give two proofs of divisibility by transitivity. The category records the resulting

comparison as one arrow. For the order by inclusion on subsets of a fixed set S, an arrow $U \to V$ expresses $U \subseteq V$. An identity expresses $U \subseteq U$, and composition is transitivity of inclusion.

A directed graph gives another category by specifying which paths count as arrows. A directed graph has vertices and edges, and each edge has a source and a target vertex. Its *path category* has the vertices as objects. An arrow from A to B is a finite list of successive edges starting at A and ending at B. Each vertex has an empty list as its identity, and composition concatenates lists. Concatenation is associative and concatenating an empty list leaves a list equal to itself, so the category laws hold.

In the graph with edges $A \xrightarrow{f} B \xrightarrow{g} C$ and $A \xrightarrow{u} C$, the resulting path category has the two distinct arrows (f, g) and (u) from A to C. They have different lengths. The composite is $(g) \circ (f) = (f, g)$. Composition applies its right-hand factor first, and the list records the edges in the order traversed. Assign the vertices to objects of a category and the edges to arrows between the corresponding objects. The assignments give equal values to these paths exactly when the image triangle commutes. In Chapter 4 we will see how the assignment extends to every path by composition.

1.3 Inverses and opposite categories

Definition 1.3. An arrow $f\colon A \to B$ is an *isomorphism* if there is an arrow $g\colon B \to A$ with $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$.

The inverse is unique. If g and h are both inverses of f, then

$$g = g \circ \mathrm{id}_B = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{id}_A \circ h = h.$$

In **Set**, inverse maps are inverse functions. The equation $g \circ f = \mathrm{id}_A$ makes f injective by allowing application of g to $f(a) = f(a')$. The equation $f \circ g = \mathrm{id}_B$ makes f surjective, since each b has preimage $g(b)$. Conversely, a bijection defines an inverse by assigning each b its unique preimage.

In a poset category, an isomorphism gives both $a \leq b$ and $b \leq a$. Antisymmetry gives $a = b$, and equality supplies the identity isomorphism. Thus the same criterion on arrows detects bijections between sets and equality of elements of a poset. Isomorphism preserves the object as seen through its arrows. For instance, $f\colon A \to B$ invertible makes $u \mapsto f \circ u$ a bijection from maps $T \to A$ to maps $T \to B$, with inverse $v \mapsto f^{-1} \circ v$.

For any category $\mathcal{C}$, its *opposite* $\mathcal{C}^{\mathrm{op}}$ has the same objects and an arrow $f^{\mathrm{op}}\colon B \to A$ for each $f\colon A \to B$ in $\mathcal{C}$. An identity becomes an identity and

$$(g \circ f)^{\mathrm{op}} = f^{\mathrm{op}} \circ g^{\mathrm{op}}.$$

Associativity and the identity laws follow by reversing the corresponding equations in $\mathcal{C}$. The opposite of a poset category comes from the reverse order. A square commutes in $\mathcal{C}$ exactly when its reversed square commutes in $\mathcal{C}^{\mathrm{op}}$.

Remark 1.4 (Duality principle). A theorem proved for every category applies in particular to $\mathcal{C}^{\mathrm{op}}$. Translating that instance into $\mathcal{C}$ reverses each arrow and reverses the order of each composite. The translated statement, with its hypotheses also reversed, is the *dual* of the original theorem.

For instance, a construction specified by maps into an object has a dual construction specified by maps out of an object.

1.4 Cancellation in place of elements

A map from a singleton $1 = \{*\}$ into A selects an element. One way to test injectivity of a function $m\colon A \to B$ is to make two such selections and ask whether equality after applying m forces the original selections to agree. A map $B \to \{0,1\}$ records a subset by its indicator. To test whether $e\colon A \to B$ covers B, compare two such indicators that agree on its image. Any point outside the image lets one indicator differ from the other there. The following definitions allow arbitrary domains and codomains for these test maps.

Definition 1.5. An arrow $m\colon A \to B$ is a *monomorphism* if for every T and every $u, v\colon T \to A$, the equation $m \circ u = m \circ v$ implies $u = v$. An arrow $e\colon A \to B$ is an *epimorphism* if for every T and every $r, s\colon B \to T$, the equation $r \circ e = s \circ e$ implies $r = s$.

The tests use cancellation on opposite sides. Reversing arrows turns the monomorphism test into the epimorphism test. An isomorphism passes both tests because its inverse can be composed on the appropriate side of an equation.

Proposition 1.6. *In* **Set**, *monomorphisms are precisely the injections and epimorphisms are precisely the surjections.*

Proof. If m is injective, $m(u(t)) = m(v(t))$ implies $u(t) = v(t)$ for each $t \in T$, so m is monic. Conversely, for any $a, a' \in A$ with $m(a) = m(a')$, let u, v be the maps from a singleton to A with values a, a'. Monicity gives $u = v$, and hence $a = a'$.

If e is surjective, every $b \in B$ is $e(a)$ for some $a \in A$. Thus equality of the composites $r \circ e$ and $s \circ e$ gives $r(b) = s(b)$ at every b. Conversely, write $I = e[A] \subseteq B$. Let $r\colon B \to \{0,1\}$ be constantly zero, and let s take value zero on I and value one on $B \setminus I$. We have $r \circ e = s \circ e$, so epicity gives $r = s$. This makes s constantly zero, hence $I = B$. □

Remark 1.7 (Cancellation and inverses). In **Set**, an arrow that is monic and epic is invertible. The divisibility category supplies a counterexample to this implication. In a thin category every arrow is monic and epic, since the arrows being compared in either test already agree. In particular $2 \to 6$ passes both cancellation tests. An inverse would imply $2 = 6$, by antisymmetry.

Example 1.8 (Factoring a function through its image)**.** Let $u\colon [6] \to [2]$ be the composite $s \circ r$ above. The map r is surjective onto $[3]$ and s is surjective onto $[2]$, so u is epic. If the target is enlarged by composing with the inclusion $i\colon [2] \hookrightarrow [3]$, the new function $i \circ u$ has image $\{0, 1\}$. It factors as a surjection to this image followed by its inclusion. The same construction works for a function $k\colon A \to B$, using the corestriction $A \to k[A]$ and the inclusion $k[A] \hookrightarrow B$. Thus every function factors as an epimorphism followed by a monomorphism in **Set**.

The test maps $T \to A$ and $B \to T$ compare arrows by composition. In **Set**, a map $T \to A$ selects an element for each $t \in T$, giving a T-indexed family. Chapter 2 specifies a product by pairing test maps with a common source. The dual test specifies a sum by combining maps with a common target.

CHAPTER 2

Counting with sums and products

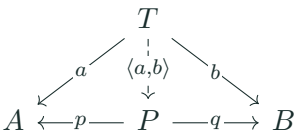
Choosing a shirt and a hat requires two choices. Choosing either a shirt or a hat requires one choice and a record of which kind it was. In **Set** these operations give the set of pairs and the tagged disjoint union. Their categorical descriptions specify how a map is determined by other maps. The two descriptions are dual, and they give different answers to the question of how to count choices.

2.1 Pairs and the universal product

Definition 2.1. A *product* of objects A, B in a category is an object P and two arrows $p\colon P \to A$, $q\colon P \to B$ with the following property. For every T, $a\colon T \to A$, and $b\colon T \to B$, there is a unique arrow $\langle a, b\rangle\colon T \to P$ such that

$$p \circ \langle a, b\rangle = a, \qquad q \circ \langle a, b\rangle = b.$$

The data are the two solid projections out of P. Once a pair of maps out of T is supplied, the condition asserts a unique dashed arrow making the two triangles commute.



The object together with its projections is written $A \times B$ when a product has been chosen. Given a map into the product, one recovers its two components by composition with p, q. Given the components, one recovers the map by pairing. In a locally small category this is a bijection

$$\mathrm{Hom}_{\mathcal{C}}(T, A \times B) \;\cong\; \mathrm{Hom}_{\mathcal{C}}(T, A) \times \mathrm{Hom}_{\mathcal{C}}(T, B),$$

where the product on the right is an ordinary product of sets.

Proposition 2.2. *In* **Set**, *the set of pairs*

$$P = \{(a, b) \mid a \in A,\ b \in B\}, \qquad p(a, b) = a, \quad q(a, b) = b$$

is a product of A, B.

Proof. For $a_0 \colon T \to A$ and $b_0 \colon T \to B$, the formula $t \mapsto (a_0(t), b_0(t))$ gives a function into P with the required components. Any other such function sends t to a pair whose first component is $a_0(t)$ and whose second component is $b_0(t)$. Equality of both components implies equality of the function values at each t. $\square$

Proposition 2.3 (Uniqueness with projections)**.** *For two products (P, p, q) and (P', p', q') of A, B, there is a unique isomorphism $u \colon P \to P'$ such that $p' \circ u = p$ and $q' \circ u = q$.*

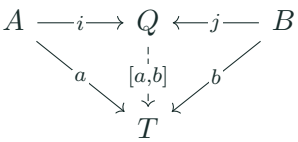
Proof. Apply the property of P' with $T = P$ to obtain u, and the property of P with $T = P'$ to obtain $v \colon P' \to P$. Both $v \circ u$ and id_P have components p, q, so $v \circ u = \mathrm{id}_P$. Both $u \circ v$ and $\mathrm{id}_{P'}$ have components p', q', so $u \circ v = \mathrm{id}_{P'}$. The same uniqueness property that supplied u gives uniqueness of this isomorphism. $\square$

The distinction between an object and its structure maps matters. In **Set** two carriers of the same cardinality can have many bijections between them. The universal property singles out the bijection commuting with the projections. Whenever both $(A \times B) \times C$ and $A \times (B \times C)$ exist, they represent a triple of component maps. A pair of nested applications of the product property proves this for each bracketing. The same argument as in Proposition 2.3 then gives the canonical reassociation. In **Set** it sends $((a, b), c)$ to $(a, (b, c))$.

2.2 Alternatives and the universal coproduct

Definition 2.4. A *coproduct* of objects A, B in a category is an object Q with maps $i \colon A \to Q$, $j \colon B \to Q$ satisfying the following property. For every T and maps $a \colon A \to T$, $b \colon B \to T$, there is a unique map $[a, b] \colon Q \to T$ with $[a, b] \circ i = a$ and $[a, b] \circ j = b$.

Reversing all the arrows in the product diagram gives the coproduct diagram. The dashed map out of Q is determined by its restrictions.



The coproduct is written $A \amalg B$. Applying Proposition 2.3 in the opposite category gives uniqueness up to a unique isomorphism commuting with the two coproduct maps.

Proposition 2.5. *In* **Set**, *the set of tagged elements*

$$Q = (\{0\} \times A) \cup (\{1\} \times B)$$

is a coproduct of A, B with maps $i(x) = (0, x)$, $j(y) = (1, y)$.

Proof. Every tagged element has exactly one of the two tags. Consequently the formula $[a, b](0, x) = a(x)$, $[a, b](1, y) = b(y)$ defines the required function, and the two equations prescribe every value. □

For example, if $A = B = \{x\}$, this construction gives two elements, $(0, x)$ and $(1, x)$. It distinguishes a choice from the first set from a choice from the second even when the untagged values agree.

The same specifications apply to a family $(A_i)_{i \in I}$ indexed by a set I. A product $\prod_{i \in I} A_i$ has projections to every A_i. For each T, composing with the projections gives a bijection from maps into the product to families of maps $T \to A_i$. Dually, a coproduct $\coprod_{i \in I} A_i$ has a map from every A_i. Composing with these maps gives a bijection from maps out of the coproduct to families of maps $A_i \to T$. In **Set**, the product consists of tuples whose ith entries lie in A_i, and the coproduct consists of pairs (i, a) with $a \in A_i$. The projections select entries and the coproduct maps attach tags. A family with a finite index set gives finite products and coproducts in this sense, including the empty cases.

2.3 The empty cases and finite counting

Definition 2.6. An object 1 is *terminal* if every object has exactly one map to 1. An object 0 is *initial* if it has exactly one map to every object.

The empty family of component maps out of T has one possible value, the empty list. Thus a map $T \to 1$ represents that empty family, and 1 is the product of the empty family of objects. Dually 0 is the empty coproduct. The singleton is terminal in **Set**, because every point of T must go to the one point. The empty set is initial, because a function with empty source has an empty graph and a prescribed target.

Maps $T \to A \times 1$ consist of a map to A and the uniquely available map to 1. It follows that A, equipped with its identity and its map to 1, already has this product property. The canonical isomorphism $A \times 1 \cong A$ sends $(a, *)$ to a in **Set**. Dually $A \amalg 0 \cong A$. This gives the identity laws for counting pairs and alternatives, including the single empty tuple and the zero possible choices from an empty list of alternatives.

If $|A| = a$ and $|B| = b$ are finite, each of the a possible first components has b possible second components, giving $|A \times B| = ab$. The disjoint tag classes of a coproduct have sizes a, b, giving $|A \amalg B| = a + b$. Composing bijections of inputs gives a bijection of products and of coproducts by applying the bijections componentwise. Thus these operations descend to isomorphism classes of finite sets.

2.4 Distributivity and the binomial theorem

Let $i\colon B \to B \amalg C$ and $j\colon C \to B \amalg C$ be the coproduct maps. In any category with binary products and coproducts, the two universal properties give a canonical comparison arrow

$$[\mathrm{id}_A \times i, \mathrm{id}_A \times j]\colon (A \times B) \amalg (A \times C) \longrightarrow A \times (B \amalg C).$$

The product arrow $\mathrm{id}_A \times i$ has components π_A and $i \circ \pi_B$, where π_A, π_B project from $A \times B$. The arrow $\mathrm{id}_A \times j$ is defined in the same way from $A \times C$. The coproduct property combines the two arrows into the comparison.

Proposition 2.7. *In* **Set**, *the comparison arrow is a bijection.*

Proof. With the set constructions, the comparison sends $(0, (a, b))$ to $(a, (0, b))$ and $(1, (a, c))$ to $(a, (1, c))$. To define an inverse, read the tag in the second coordinate and move it outside the pair. Each composite recovers the input tag and its data. □

The inverse uses the tags available in **Set**. When the comparison is invertible for every triple of objects, products distribute over coproducts. The chapter on simple types explains why this holds in a cartesian closed category with binary coproducts. There, $A \times -$ is a left adjoint and preserves coproducts.

For finite sets, taking cardinalities gives $a(b + c) = ab + ac$, where $a = |A|$, $b = |B|$, and $c = |C|$. Iterating the same decomposition gives a binomial count. Write $Y^X = \mathbf{Set}(X, Y)$ for the set of functions from X to Y, regarded as X-indexed tuples in Y. In particular, $Y^{\varnothing}$ consists of the one empty function because $\varnothing$ is initial. The notation Y^n uses the index set $I = \{1, \ldots, n\}$, which is empty when $n = 0$. For $S \subseteq I$, write $S^c = I \setminus S$.

An element of $(A \amalg B)^n$ specifies a subset $S \subseteq I$ where its tags come from A. For a fixed S, the remaining data consist of maps $S \to A$ and $S^c \to B$. Conversely these maps recover the tagged tuple by reading the coordinates in order. This gives a bijection

$$(A \amalg B)^n \;\cong\; \coprod_{S \subseteq I} A^S \times B^{S^c}.$$

Write $\binom{n}{k}$ for the number of k-element subsets of I. Grouping the summands by $|S| = k$ yields

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

For $n = 0$ this counts the empty tuple, with $a^0 = b^0 = 1$ even when the corresponding set of choices is empty. As a concrete case, $a = 2$, $b = 1$, and $n = 3$ give $1 + 6 + 12 + 8 = 27$ tuples. The 12 counts the three tag patterns with two A-coordinates and the four choices for values in those coordinates.

2.5 The same properties in an order

For a poset viewed as a category, a pair of maps $T \to A$ and $T \to B$ says that T is a lower bound of A, B. The product condition says that these lower bounds are precisely the elements with an arrow to P. With its projections, P is therefore the greatest lower bound. Dually a coproduct is a least upper bound. In the inclusion order of subsets of a fixed set, the product and coproduct are intersection and union. Terminal and initial objects are the entire set and the empty subset.

In the divisibility order on positive integers, a product of 12 and 18 is 6. It divides both numbers and every common divisor divides $18 - 12 = 6$. Their coproduct is 36. It is a multiple of both numbers, and a common multiple has the form $12a = 18b$, giving $2a = 3b$ and hence $3 \mid a$. Thus every common multiple is divisible by 36. These computations recover the greatest common divisor and the least common multiple from the same definitions that gave pairs and tagged alternatives.

The number 1 is initial because it divides every positive integer. A terminal object would have to be a multiple of every positive integer, and each candidate n fails divisibility by $n + 1$. If the order is extended to include 0, that number is terminal because every positive integer divides 0. In this extended order the numbers 1 and 0 play the roles of the categorical 0 and 1, respectively.

2.6 Inclusion–exclusion on the subset lattice

Let $A_i \subseteq X$, for i in a finite set I, be conditions on elements of a finite set X. For $S \subseteq I$, set $X_S = \bigcap_{i \in S} A_i$, with $X_\varnothing = X$. If $S \subseteq T$, there is an inclusion $X_T \hookrightarrow X_S$. Thus $S \mapsto X_S$ gives a functor from the opposite of the subset order of I to the subset order of X. Joining two sets of conditions gives a product of their solution subsets over X.

$$\begin{array}{ccc} X_{S \cup T} & \hookrightarrow & X_T \\ \downarrow & & \downarrow \\ X_S & \hookrightarrow & X \end{array}$$

A map to X whose image lies in both X_S and X_T factors uniquely through their intersection. This gives the pullback property used in Chapter 3.

Let $h(S) = |X_S|$, and let $g(S)$ count the elements satisfying exactly the conditions in S. Partitioning X_S by the complete sets of conditions its elements satisfy gives

$$h(S) = \sum_{T \supseteq S} g(T).$$

The subset order determines the inverse of this counting operation.

$$g(S) = \sum_{T \supseteq S} (-1)^{|T \setminus S|} h(T).$$

To see this, substitute the expression for $h(T)$. The coefficient of $g(U)$ is the sum of $(-1)^{|T \setminus S|}$ over $S \subseteq T \subseteq U$. Choosing T amounts to choosing a subset of $U \setminus S$, so the coefficient is 1 for $U = S$ and $(1-1)^{|U \setminus S|} = 0$ otherwise. At $S = \varnothing$ the inverse gives the *inclusion–exclusion formula* for the number of elements avoiding every condition [70, § 2.1].

Take X to be the functions from five labeled positions to three labeled colors. For each color i, let A_i consist of the functions that omit i. A function in X_S may use any of the $3 - |S|$ remaining colors at each position, so $h(S) = (3 - |S|)^5$. The number of functions using all three colors is

$$3^5 - \binom{3}{1} 2^5 + \binom{3}{2} 1^5 - \binom{3}{3} 0^5 = 150.$$

Each term uses a product count inside a single intersection. A function that omits $k > 0$ colors contributes $(1-1)^k = 0$ to the sum.

In Chapter 3, maps out of a test object T carry equations. An equalizer parametrizes the maps $T \to X$ whose two composites to a fixed target agree. A pullback parametrizes pairs whose composites agree, refining the description of a product. Reversing the arrows gives tests by maps into a target, from which we construct quotients and glue sets along shared data.

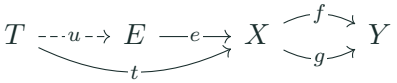
CHAPTER 3

Solving equations and forming quotients

An equation $f(x) = g(x)$ singles out certain elements of a set X. It also singles out certain families of elements of X. A family indexed by T is a map $t\colon T \to X$, and every member of the family solves the equation precisely when $f \circ t = g \circ t$. Describing solutions by this equality of maps gives a construction that makes sense in any category. The same method works when the two sides of an equation have different domains, and duality turns a condition on maps into an object into a condition on maps out of it.

3.1 Equalizers

Definition 3.1. Let $f, g\colon X \to Y$ be parallel maps. An *equalizer* of f and g is a map $e\colon E \to X$ satisfying $f \circ e = g \circ e$ with the following property. For every $t\colon T \to X$ satisfying $f \circ t = g \circ t$, there is a unique $u\colon T \to E$ with $e \circ u = t$.



The solid map t is given. The dashed map u is the unique factorization promised by the definition. If $e \circ a = e \circ b$ for maps $a, b\colon T \to E$, both a and b factor the same map $t = e \circ a$. Uniqueness gives $a = b$. Thus every equalizer is a monomorphism. This is the categorical counterpart of the inclusion of a solution set into the domain of an equation.

Proposition 3.2. *In* **Set**, *the equalizer of $f, g\colon X \to Y$ is the inclusion of $\{x \in X \mid f(x) = g(x)\}$ into X.*

Proof. Let $E = \{x \in X \mid f(x) = g(x)\}$ and let e be its inclusion. The maps $f \circ e$ and $g \circ e$ agree by construction. If $f \circ t = g \circ t$, then $t(w) \in E$ for each $w \in T$. Define $u(w) = t(w)$ with codomain E. This gives $e \circ u = t$. Every factorization has these same values because e is an inclusion. □

Example 3.3. Take $f, g\colon \mathbb{Z} \to \mathbb{Z}$ with $f(x) = x^2$ and $g(x) = 2x + 3$. The equation becomes $(x - 3)(x + 1) = 0$. Since a product of integers vanishes if

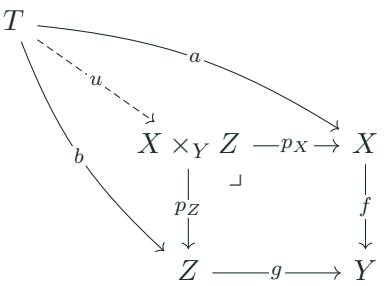
and only if one factor vanishes, the equalizer is $\{-1, 3\} \hookrightarrow \mathbb{Z}$. For any set T, a map $T \to \mathbb{Z}$ solves the equation at all indices exactly when it arises from a map $T \to \{-1, 3\}$. The latter map records one of the two roots at each index.

The universal property also compares equalizers constructed in different ways. If $e' \colon E' \to X$ is another equalizer of the same pair, e' factors through e, and e factors through e'. The composite on E and the identity on E both factor e through e, so they agree. The argument with E' gives the other identity. The resulting isomorphism is uniquely determined by the requirement that it commute with the maps to X.

3.2 Pullbacks

Suppose $f \colon X \to Y$ and $g \colon Z \to Y$. The equation $f(x) = g(z)$ now relates an input from X to an input from Z. A family of solutions indexed by T consists of maps $a \colon T \to X$ and $b \colon T \to Z$ that satisfy $f \circ a = g \circ b$. A pullback represents these families.

Definition 3.4. A *pullback* of f and g is an object P with maps $p_X \colon P \to X$ and $p_Z \colon P \to Z$ satisfying $f \circ p_X = g \circ p_Z$. For every pair (a, b) as above there must be a unique map $u \colon T \to P$ with $p_X \circ u = a$ and $p_Z \circ u = b$. We write $P = X \times_Y Z$ for a chosen pullback.



The square expresses agreement of the two composites from P to Y. The corner mark identifies it as a pullback square. The outer maps a, b agree after composition with f, g. The dashed arrow is the unique map through which both outer maps factor via the indicated projections.

Proposition 3.5. *In* **Set**, *the set*

$$X \times_Y Z = \{(x, z) \in X \times Z \mid f(x) = g(z)\},$$

with its coordinate projections, is a pullback.

Proof. The projections agree after applying f and g because every pair in the displayed set satisfies the defining equality. If a and b are a compatible pair of maps from T, define $u(w) = (a(w), b(w))$. Compatibility ensures that this pair lies in the pullback. Its coordinates give the required factorizations, and the coordinates determine each pair uniquely. □

Example 3.6. Let $X = \{0, 1, 2, 3\}$, $Z = \{0, 1, 2\}$, and $Y = \{0, 1\}$. Let $f(x)$ be the remainder of x modulo 2 and $g(z)$ the remainder of $z + 1$ modulo 2. The coordinates must have opposite parity. Grouping pairs by their common image gives

Common image	Pairs in the pullback
0	$(0, 1),\ (2, 1)$
1	$(1, 0),\ (1, 2),\ (3, 0),\ (3, 2).$

For finite sets this method yields a count before listing the pairs. For each $y \in Y$, exactly $|f^{-1}(\{y\})|\,|g^{-1}(\{y\})|$ pairs have common image y. The groups for distinct y are disjoint. Here the count is $2 \cdot 1 + 2 \cdot 2 = 6$.

Proposition 3.7 (Pasting pullbacks)**.** *Let $P = X \times_Y Z$ be the pullback of $f\colon X \to Y$ and $g\colon Z \to Y$, and let $h\colon W \to Z$. If $Q = P \times_Z W$ is formed using p_Z and h, the induced maps $Q \to X$ and $Q \to W$ exhibit Q as a pullback of f and $g \circ h$.*

$$\begin{array}{ccccc} Q & \xrightarrow{r} & P & \xrightarrow{p_X} & X \\ \downarrow{\scriptstyle s} & \lrcorner & \downarrow{\scriptstyle p_Z} & \lrcorner & \downarrow{\scriptstyle f} \\ W & \xrightarrow{h} & Z & \xrightarrow{g} & Y \end{array}$$

Proof. Write $r\colon Q \to P$ and $s\colon Q \to W$ for the projections. The two pullback equations give $f \circ p_X \circ r = g \circ p_Z \circ r = g \circ h \circ s$. Suppose $a\colon T \to X$ and $b\colon T \to W$ satisfy $f \circ a = g \circ h \circ b$. There is a unique $c\colon T \to P$ with projections a and $h \circ b$. There is then a unique $d\colon T \to Q$ with projections c and b. A map to Q with induced projections a and b must have projection c to P by the first pullback property, and must then equal d by the second. □

For sets, the resulting identification sends $((x, z), w)$ to (x, w). The condition $z = h(w)$ recovers the dropped coordinate. The proposition explains why successive pullbacks have this effect in any category with the required pullbacks.

3.3 Fibers, intersections, and the diagonal

In **Set**, the *fiber* of $f\colon X \to Y$ over $y \in Y$ is the pullback along the map $1 \to Y$ from a singleton that picks out y. Its set construction consists of pairs $(x, *)$ with

$f(x) = y$. Projection to x is a bijection onto $f^{-1}(\{y\})$, with inverse $x \mapsto (x, *)$. More generally, pulling back along the inclusion of a subset $B \subseteq Y$ yields the inverse image $f^{-1}(B)$. In that case the inverse to the projection onto $f^{-1}(B)$ sends x to $(x, f(x))$.

The pullback of two inclusions $A \hookrightarrow X$ and $B \hookrightarrow X$ consists of pairs (a, b) whose images in X coincide. The correspondence $x \mapsto (x, x)$ is therefore a bijection from $A \cap B$ to the pullback. For the parity map in the previous example, the fiber over 0 is $\{0, 2\}$. Restricting it to the subset $\{1, 2, 3\}$ gives the intersection $\{2\}$.

Whenever $Y \times Y$ exists, an equalizer of two maps into Y can be expressed as a pullback. The diagonal $\Delta_Y \colon Y \to Y \times Y$ is the map whose two projections are the identity on Y. For a pair $f, g \colon X \to Y$, form the map $\langle f, g \rangle \colon X \to Y \times Y$ by pairing them.

$$\begin{array}{ccc} E & \xrightarrow{\;e\;} & X \\ \downarrow{\scriptstyle d} & & \downarrow{\scriptstyle \langle f,g \rangle} \\ Y & \xrightarrow{\;\Delta_Y\;} & Y \times Y \end{array}$$

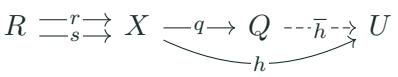
If the displayed square is a pullback, its commutativity gives $f \circ e = d = g \circ e$ by projection onto each factor. When $t \colon T \to X$ satisfies $f \circ t = g \circ t$, take $s = f \circ t$. Pairing gives $\langle f, g \rangle \circ t = \Delta_Y \circ s$. The pullback gives a unique lift $u \colon T \to E$ with $e \circ u = t$ and $d \circ u = s$. Any map with $e \circ u = t$ also satisfies $d \circ u = f \circ e \circ u = s$, so e has the equalizer property. Conversely, if e is an equalizer, set $d = f \circ e = g \circ e$. A pair t, s compatible with the bottom and right arrows satisfies $f \circ t = s = g \circ t$ by projection. The unique equalizer factorization u of t also gives $d \circ u = f \circ t = s$, establishing the pullback property.

Products and equalizers also suffice to construct pullbacks. With $f \colon X \to Y$ and $g \colon Z \to Y$, take the equalizer $e \colon P \to X \times Z$ of $f \circ \pi_X$ and $g \circ \pi_Z$, where π_X, π_Z are the product projections. Set $p_X = \pi_X \circ e$ and $p_Z = \pi_Z \circ e$. For compatible a, b from T, the product gives a unique $\langle a, b \rangle \colon T \to X \times Z$. Compatibility makes this map factor uniquely through e. A map to P with projections a, b must give this factorization by the uniqueness of pairing, which proves the pullback property.

3.4 Coequalizers and generated equivalences

For a pair $r, s \colon R \to X$, an equalizer asks where r and s already agree on R. Reversing the arrows in the universal property asks how to make their images agree in a quotient of X.

Definition 3.8. A *coequalizer* of $r, s \colon R \to X$ is a map $q \colon X \to Q$ satisfying $q \circ r = q \circ s$. Every $h \colon X \to U$ with $h \circ r = h \circ s$ must factor uniquely through q.



Every coequalizer is an epimorphism. Indeed, if $a \circ q = b \circ q$ then a and b are both factorizations of that composite through q, so they agree. This is the dual of the monomorphism argument for equalizers.

Proposition 3.9. *In* **Set**, *the coequalizer of* $r, s \colon R \to X$ *is the quotient map* $X \to X/\!\sim$ *for the equivalence relation generated by all pairs* $(r(w), s(w))$ *with* $w \in R$.

Proof. Declare $x \sim x'$ when a finite chain joins x to x' in which each step is one of the prescribed pairs in either direction. A chain of length zero gives reflexivity. Reversing a chain gives symmetry, and concatenating chains gives transitivity. Every equivalence relation containing the prescribed pairs contains the endpoints of each chain, so this is the least such relation.

Write $q(x) = [x]$ for the map to equivalence classes. Every pair prescribed by r and s lies in a single class, hence $q \circ r = q \circ s$. For a map h that agrees on the prescribed pairs, equality propagates along every chain. The assignment $\overline{h}([x]) = h(x)$ is therefore well-defined and satisfies $\overline{h} \circ q = h$. Since every equivalence class is the image of a representative, $\overline{h}$ is unique. □

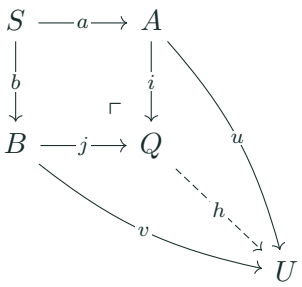
Example 3.10. Let $X = \{0, 1, 2, 3, 4\}$, let $R = \{0, 1, 2\}$, and choose $r(w) = w$ and $s(w) = w + 2$. The pairs identify 0 with 2, 1 with 3, and 2 with 4. Parity is constant along every chain, and any two elements of a given parity are joined by the displayed steps. The quotient consists of the classes $\{0, 2, 4\}$ and $\{1, 3\}$. Taking the remainder modulo 2 gives the same coequalizer with its two classes labeled 0 and 1.

Every equivalence relation $L \subseteq X \times X$ can itself serve as the source of the two maps that present a quotient. Use the coordinate projections $L \rightrightarrows X$. Since L is already an equivalence relation, the generated relation is exactly L, and the resulting coequalizer is $X \to X/L$.

3.5 Pushouts and gluing

The dual of a pullback is a pushout. It asks how to join two objects whose maps from a common source must agree after gluing.

Definition 3.11. Given $a \colon S \to A$ and $b \colon S \to B$, a *pushout* has maps $i \colon A \to Q$ and $j \colon B \to Q$ with $i \circ a = j \circ b$. For every $u \colon A \to U$ and $v \colon B \to U$ with $u \circ a = v \circ b$, there is a unique map $h \colon Q \to U$ with $h \circ i = u$ and $h \circ j = v$.



Every compatible pair u, v determines the dashed arrow h. The pushout is written $A \amalg_S B$, with the maps from S understood.

Proposition 3.12. *Let $a\colon S \to A$ and $b\colon S \to B$. Suppose $A \amalg B$ exists, with coproduct maps ι_A, ι_B. If $q\colon A \amalg B \to Q$ is a coequalizer of the pair $\iota_A \circ a, \iota_B \circ b$, then $q \circ \iota_A$ and $q \circ \iota_B$ make Q a pushout of a, b.*

Proof. Set $i = q \circ \iota_A$ and $j = q \circ \iota_B$. The coequalizer equation is $i \circ a = j \circ b$. Given u, v as in the definition, the coproduct gives a unique map $k\colon A \amalg B \to U$ with $k \circ \iota_A = u$ and $k \circ \iota_B = v$. Its composites with the parallel pair are $u \circ a$ and $v \circ b$, which agree. The coequalizer property gives a unique $h\colon Q \to U$ with $h \circ q = k$, and it satisfies $h \circ i = u$ and $h \circ j = v$. Every map h with those two composites must give $h \circ q = k$ by the coproduct property, so the required mediator is unique. □

Example 3.13. In **Set**, take disjoint copies of A and B and identify the pairs selected by S. For $A = \{a_0, a_1\}$, $B = \{b_0, b_1\}$, let S be a singleton selecting a_1 and b_0. The pushout has classes $\{a_0\}, \{a_1, b_0\}, \{b_1\}$. A map from it to U assigns one value per class, which makes its restrictions to A and B agree at the glued pair.

3.6 Finite limits and colimits

Products specify families of maps, and equalizers impose equations on maps. Together they construct the universal family of maps into an entire finite diagram. Here the *index category* $\mathcal{I}$ records the shape of the diagram. A diagram assigns to each object i of $\mathcal{I}$ an object D_i of $\mathcal{C}$, and to each arrow $\alpha\colon i \to j$ a map $D_\alpha\colon D_i \to D_j$. It preserves identities and composition. Chapter 4 calls such an assignment a functor. The diagram is *finite* when $\mathcal{I}$ has finitely many objects and finitely many arrows.

Definition 3.14. A *cone* from T into a diagram D is a family of maps $t_i\colon T \to D_i$ with $D_\alpha \circ t_i = t_j$ for each $\alpha\colon i \to j$. A *limit* of D consists of an object L with a cone (ℓ_i) through which every cone (t_i) factors by a unique $u\colon T \to L$ satisfying $\ell_i \circ u = t_i$ for all i. A *cocone* and *colimit* reverse these arrows.

A product is a limit of a diagram with separate objects and only their identity arrows. The empty diagram has a limit precisely when there is a terminal object. A cone into the diagram of two parallel arrows is determined by an equalizing map into their source. A cone into $X \to Y \leftarrow Z$ is determined by its compatible maps into X, Z. These give the equalizer and pullback as limits. Their duals give the coequalizer and pushout as colimits.

Proposition 3.15. *If a category has a terminal object, all binary products, and all equalizers, it has limits of all finite diagrams.*

Proof. Finite products exist under the first two hypotheses. Start with the terminal object for the empty product and take a binary product each time an object is added. Pairing a map to the product already constructed with a map to the new object gives the required family property at each step.

For a finite diagram D, form two products

$$P = \prod_{i \in \mathrm{Ob}(\mathcal{I})} D_i, \qquad B = \prod_{(\alpha\colon i \to j) \in \mathrm{Mor}(\mathcal{I})} D_j,$$

where $\mathrm{Ob}(\mathcal{I})$ and $\mathrm{Mor}(\mathcal{I})$ are the sets of objects and arrows of $\mathcal{I}$. Write $p_i \colon P \to D_i$ for the first projections and $b_\alpha \colon B \to D_j$ for the second. Define parallel maps $r, s \colon P \to B$ by their components

$$b_\alpha \circ r = D_\alpha \circ p_i, \qquad b_\alpha \circ s = p_j \qquad (\alpha \colon i \to j).$$

Take their equalizer $e \colon L \to P$. The composites $\ell_i = p_i \circ e$ form a cone because $r \circ e = s \circ e$ gives $D_\alpha \circ \ell_i = \ell_j$ at every component.

Given a cone (t_i) from T, the product P gives a unique $t \colon T \to P$ with $p_i \circ t = t_i$. The cone equations make the components of $r \circ t$ and $s \circ t$ agree, so these maps to B are equal. The equalizer gives a unique $u \colon T \to L$ with $e \circ u = t$. By uniqueness of the map into P, any arrow with the required composites $\ell_i \circ u = t_i$ factors t this way. This proves the limit property. □

Applying the argument in the opposite category shows that an initial object, binary coproducts, and coequalizers suffice for finite colimits. See Leinster [1, §§ 5.1–5.2] for both constructions. The dual construction coequalizes two parallel maps between coproducts. For each index arrow $\alpha \colon i \to j$, their common source has one summand D_i. The two maps from that summand into the coproduct of the D_i use the coproduct map for i and the composite of D_α with the coproduct map for j.

For a fixed pair f, g, the pullback problem assigns to each test object T the compatible pairs of maps from T. A map $T' \to T$ sends such a pair to a compatible pair from T' by precomposition. The same rule sends a cone from T

into any fixed diagram to a cone from T'. Chapter 4 packages these assignments as presheaves in a locally small category. The limit L represents the presheaf of cones, since one map $T \to L$ specifies the whole compatible family from T.

CHAPTER 4

Recovering objects from their maps

A map from a singleton to a set X chooses an element of X. A map from $\{0, 1\}$ to X chooses an ordered pair, with repetition allowed. Precomposing with the inclusion that selects 0 returns the first coordinate. Precomposing with the transposition of 0 and 1 swaps the coordinates. The sets of maps record the elements and pairs. The precomposition operations record their relationships.

In an arbitrary category a map $u\colon T \to X$ is called a *generalized element* of X with domain T. The Yoneda lemma makes this viewpoint precise. It identifies every rule for assigning data to generalized elements that respects change of domain. The special case in which the assigned data are themselves generalized elements recovers maps between objects.

4.1 Functors

Definition 4.1. A *functor* $F\colon \mathcal{C} \to \mathcal{D}$ sends each object X of $\mathcal{C}$ to an object $F(X)$ of $\mathcal{D}$ and each map $f\colon X \to Y$ to a map $F(f)\colon F(X) \to F(Y)$. It satisfies

$$F(\mathrm{id}_X) = \mathrm{id}_{F(X)}, \qquad F(g \circ f) = F(g) \circ F(f).$$

A monotone function between partially ordered sets is a functor between their associated categories. For $x \leq y$, the arrow $x \to y$ has an image precisely when $F(x) \leq F(y)$. The identity and composition laws hold because each ordered pair of objects admits at most one arrow. Thus order preservation is exactly the functor condition in this example.

An assignment of vertices and edges as in Chapter 1 extends uniquely to a functor from the path category. Send a finite path to the composite of the arrows assigned to its edges, and send the empty path at each vertex to the identity of the assigned object. Associativity makes this rule preserve concatenation. Conversely, a functor extending the assignment must send each path to that composite because the path is the concatenation of its individual edges.

For another example, fix a set S. The rule $P_S(X) = X \times S$ on sets, with $P_S(f)(x, s) = (f(x), s)$, is a functor $P_S\colon \mathbf{Set} \to \mathbf{Set}$. On a pair (x, s), the image of id_X returns (x, s). The image of a composite $g \circ f$ returns $(g(f(x)), s)$, which is also the result of applying $P_S(f)$ and then $P_S(g)$. The specification on arrows supplies the composition law that connects all the sets $X \times S$.

4.2 Naturality

Definition 4.2. For functors $F, G\colon \mathcal{C} \to \mathcal{D}$, a *natural transformation* $\eta\colon F \Rightarrow G$ has a component $\eta_X\colon F(X) \to G(X)$ at each object X of $\mathcal{C}$. For every map $f\colon X \to Y$, its components satisfy

$$\begin{array}{ccc} F(X) & \xrightarrow{\eta_X} & G(X) \\ {\scriptstyle F(f)}\downarrow & & \downarrow{\scriptstyle G(f)} \\ F(Y) & \xrightarrow{\eta_Y} & G(Y). \end{array}$$

In the square, applying the rule at X and then moving along f gives the same result as moving along f first and applying the rule at Y. For instance, the first projections $p_X\colon X \times S \to X$ define a natural transformation from P_S to the identity functor on **Set**. Both paths in the square take (x, s) to $f(x)$.

The same calculation can be used to determine what a natural rule can depend on. Suppose $q\colon P_S \Rightarrow \mathrm{Id}_{\mathbf{Set}}$ is any natural transformation. Take a singleton $1 = \{*\}$. For every $x \in X$, let $a_x\colon 1 \to X$ send $*$ to x. For $s \in S$, naturality at a_x gives

$$q_X(x, s) = q_X\big(P_S(a_x)(*, s)\big) = a_x(q_1(*, s)) = x.$$

If $X \times S$ is empty the two maps agree there as well, by uniqueness of the empty function. It follows that q is the first projection. A square involving a small test domain has determined the entire family.

Natural transformations compose by composing their components. If $\theta\colon G \Rightarrow H$, then

$$H(f) \circ \theta_X \circ \eta_X = \theta_Y \circ G(f) \circ \eta_X = \theta_Y \circ \eta_Y \circ F(f),$$

so the composite satisfies naturality. The identity components give an identity transformation. A natural transformation is a *natural isomorphism* precisely when every component is an isomorphism. Indeed, if all components are invertible, the naturality equation rearranges to $F(f) \circ \eta_X^{-1} = \eta_Y^{-1} \circ G(f)$. Thus the componentwise inverses form the natural inverse. Conversely, any natural inverse supplies an inverse at every component.

4.3 Presheaves and representable functors

In the opposite category $\mathcal{C}^{\mathrm{op}}$, an arrow $v\colon U \to T$ of $\mathcal{C}$ becomes an arrow from T to U. A *presheaf* on $\mathcal{C}$ is a functor $F\colon \mathcal{C}^{\mathrm{op}} \to \mathbf{Set}$. We write $F(v)\colon F(T) \to F(U)$ using the original arrow v. If $w\colon V \to U$, the functor law in this notation reads

$$F(v \circ w) = F(w) \circ F(v).$$

For example, let $\mathcal{C}$ be the poset of subsets of a fixed set D, ordered by inclusion. For a fixed set R, assign to a subset T the set of functions $T \to R$. The map

assigned to an inclusion $U \subseteq T$ restricts functions to U. The identity and composition laws say that restriction to the full domain leaves a function fixed and successive restrictions equal the composite restriction.

For a locally small category and an object A, define the presheaf $h_A = \mathcal{C}(-, A)$ by

$$h_A(T) = \mathcal{C}(T, A), \qquad h_A(v)(u) = u \circ v \quad (v\colon U \to T,\ u\colon T \to A).$$

The identity and associativity of composition give the presheaf laws. A presheaf F is *represented* by A if there is a natural isomorphism $h_A \cong F$. The isomorphism gives a bijection at each object and identifies the operations of changing the domain.

The constructions of Chapters 2 and 3 specify presheaves by describing the families they classify. For the product of A and B, the presheaf is

$$F_{A,B}(T) = \mathcal{C}(T, A) \times \mathcal{C}(T, B), \qquad F_{A,B}(v)(a, b) = (a \circ v, b \circ v)$$

for $v\colon U \to T$. The product $P = A \times B$ represents $F_{A,B}$. Pairing gives the inverse of the map from $\mathcal{C}(T, P)$ to pairs of maps, and associativity gives $(p \circ u) \circ v = p \circ (u \circ v)$ for a projection p and $u\colon T \to P$. Hence the bijections are natural.

For parallel maps $f, g\colon X \to Y$, the equalizer presheaf has values

$$F(T) = \{t \in \mathcal{C}(T, X) \mid f \circ t = g \circ t\}.$$

Precomposition preserves the equation, since $f \circ (t \circ v)$ and $g \circ (t \circ v)$ are the composites of $f \circ t$ and $g \circ t$ with v. An equalizer $e\colon E \to X$ represents this presheaf, by the bijection $u \mapsto e \circ u$ from $\mathcal{C}(T, E)$. This bijection commutes with precomposition by associativity.

For a pullback of $f\colon X \to Y$ and $g\colon Z \to Y$, use the presheaf of compatible pairs

$$F(T) = \{(a, b) \in \mathcal{C}(T, X) \times \mathcal{C}(T, Z) \mid f \circ a = g \circ b\}.$$

Precomposition on both entries preserves compatibility, and the pullback $X \times_Y Z$ represents F by composing with its two projections. In **Set**, taking $Z = 1$ and $g(*) = y$ turns each such pair into a T-indexed family in X with constant image y. The representing object is then the fiber $f^{-1}(\{y\})$.

4.4 The Yoneda lemma

Theorem 4.3 (Yoneda lemma)**.** *Let $F\colon \mathcal{C}^{\mathrm{op}} \to \mathbf{Set}$ be a presheaf on a locally small category and let A be an object of $\mathcal{C}$. There is a bijection*

$$\mathrm{Nat}(h_A, F) \longrightarrow F(A), \qquad \eta \longmapsto \eta_A(\mathrm{id}_A).$$

Here $\mathrm{Nat}(h_A, F)$ consists of the natural transformations from h_A to F.

Proof. For a natural transformation $\eta\colon h_A \Rightarrow F$, write $a = \eta_A(\mathrm{id}_A)$. A map $u\colon T \to A$ gives this naturality square.

$$\begin{array}{ccc} \mathcal{C}(A,A) & \xrightarrow{\eta_A} & F(A) \\ {\scriptstyle h_A(u)}\downarrow & & \downarrow{\scriptstyle F(u)} \\ \mathcal{C}(T,A) & \xrightarrow{\eta_T} & F(T). \end{array}$$

The left side takes id_A to u. Following both paths from id_A gives $\eta_T(u) = F(u)(a)$. Thus evaluation at id_A determines every component.

Given $a \in F(A)$, define $\eta^a_T(u) = F(u)(a)$. For $v\colon U \to T$ and $u\colon T \to A$,

$$\eta^a_U(h_A(v)(u)) = F(u \circ v)(a) = F(v)(F(u)(a)) = F(v)(\eta^a_T(u)).$$

This is the naturality equation. At $T = A$ and $u = \mathrm{id}_A$, the formula returns $F(\mathrm{id}_A)(a) = a$. The two operations are inverse. □

Riehl [2, § 2.2] treats both variances of this correspondence and the resulting embeddings.

4.4.1 Naturality of the correspondence

The bijection itself respects two kinds of change. If $\theta\colon F \Rightarrow G$, composing $\eta\colon h_A \Rightarrow F$ with θ takes its corresponding element $a \in F(A)$ to $\theta_A(a)$, since $(\theta\circ\eta)_A(\mathrm{id}_A) = \theta_A(\eta_A(\mathrm{id}_A))$. If $k\colon A' \to A$, postcomposition with k gives a natural map $h_k\colon h_{A'} \Rightarrow h_A$ because postcomposition commutes with precomposition by associativity. The composite $\eta \circ h_k$ has corresponding element

$$(\eta \circ h_k)_{A'}(\mathrm{id}_{A'}) = \eta_{A'}(k) = F(k)(a).$$

Thus evaluation commutes with applying a transformation to the presheaf and with restricting the domain along a map in $\mathcal{C}$. This is what it means for the bijection to be natural in F and A.

4.4.2 Universal elements

The proof also identifies exactly which data represent F. For a pair (A, a) with $a \in F(A)$, the natural map from h_A to F sends a generalized element $u\colon T \to A$ to $F(u)(a)$. This map exhibits a representation if and only if for every T and $x \in F(T)$ there is exactly one $u\colon T \to A$ satisfying $F(u)(a) = x$. When this holds, a is called a *universal element* of F.

For $F = h_A$, the identity natural isomorphism gives the universal element $\mathrm{id}_A \in h_A(A)$. The product representation at $P = A \times B$ sends id_P to the

two projections $(p, q) \in F_{A,B}(P)$. The equalizer representation sends id_E to $e\colon E \to X$, a solution of $f \circ e = g \circ e$ in $F(E)$. The universal element of the pullback presheaf is the pair of pullback projections, whose compatibility is the commutativity of the pullback square.

4.5 Recovering objects and arrows

Apply the theorem to $F = h_B$. The correspondence becomes

$$\mathrm{Nat}(h_A, h_B) \cong \mathcal{C}(A, B).$$

A map $f\colon A \to B$ corresponds to the transformation h_f whose component sends $u\colon T \to A$ to $f \circ u\colon T \to B$. Precomposition changes the test domain T, while postcomposition with f changes the target object. Associativity makes these two actions commute.

The assignment $A \mapsto h_A$, $f \mapsto h_f$ is a covariant functor $\mathcal{C} \to [\mathcal{C}^{\mathrm{op}}, \mathbf{Set}]$. The brackets denote the category of presheaves and natural transformations. Each h_A is contravariant in its test object T, while the assignment to A is covariant because $f\colon A \to B$ acts by postcomposition. Identities and composition are preserved since $\mathrm{id}_A \circ u = u$ and $g \circ (f \circ u) = (g \circ f) \circ u$. When $\mathcal{C}$ is large, the presheaf category is formed in a larger set-theoretic universe containing $\mathcal{C}$. A functor is *full* when its map on every hom-set is surjective and *faithful* when it is injective. The correspondence above proves that this functor, the *Yoneda embedding*, has both properties.

Corollary 4.4. *Objects A and B are isomorphic precisely when their presheaves of generalized elements h_A and h_B are naturally isomorphic.*

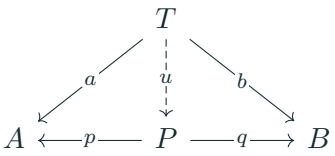
Proof. An isomorphism in $\mathcal{C}$ gives a natural isomorphism by applying the functor. In the other direction, write a natural isomorphism and its inverse as h_f and h_g using fullness. The identity equations give $h_{g \circ f} = h_{\mathrm{id}_A}$ and $h_{f \circ g} = h_{\mathrm{id}_B}$. Faithfulness then gives $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$. □

In a poset P regarded as a category, $h_a(t)$ is a singleton for $t \leq a$ and is empty otherwise. The transformation corresponding to $a \leq b$ expresses inclusion of the lower sets of a and b. Conversely, that inclusion places a in the lower set of b, so $a \leq b$. In this setting the recovery of order from natural transformations can be read directly from the components.

When two objects represent the same presheaf, the representing data specify which isomorphism relates them. Let $\eta\colon h_A \xrightarrow{\sim} F$ and $\theta\colon h_B \xrightarrow{\sim} F$ have universal elements a and b. The natural isomorphism $\theta^{-1} \circ \eta$ corresponds to an isomorphism $f\colon A \to B$. Evaluating $\theta \circ h_f = \eta$ at id_A gives $F(f)(b) = a$. The universal property of b, applied in domain A, makes f the unique map with this property.

4.6 Products as representing objects

Suppose P represents the product presheaf from Section 4.3, with natural isomorphism $\eta\colon h_P \Rightarrow F_{A,B}$. The universal element $\eta_P(\mathrm{id}_P) = (p, q)$ consists of maps $p\colon P \to A$ and $q\colon P \to B$. The Yoneda formula sends any $u\colon T \to P$ to $(p \circ u, q \circ u)$. Bijectivity says that for every pair (a, b) there is a unique u making both triangles commute.



The dashed factorization is the defining property of a product, so the representing object and its universal element recover both P and its projections.

4.7 Representations for maps out of an object

Representations also describe rules for maps out of an object. A functor $G\colon \mathcal{C} \to \mathbf{Set}$ is a presheaf on $\mathcal{C}^{\mathrm{op}}$, where the representable at D is $\mathcal{C}^{\mathrm{op}}(-, D) = \mathcal{C}(D, -)$. In this variance a map $v\colon T \to U$ acts on $u\colon D \to T$ by postcomposition, sending it to $v \circ u\colon D \to U$. Varying the representing object D reverses arrows. A map $k\colon D \to D'$ acts by precomposition and defines $\mathcal{C}(D', -) \Rightarrow \mathcal{C}(D, -)$, with $u \mapsto u \circ k$ on maps out of D'. Thus $\mathcal{C}(D, T)$ is contravariant in its first argument and covariant in its second. Applying the Yoneda lemma to the opposite category gives

$$\mathrm{Nat}(\mathcal{C}(D, -), G) \cong G(D).$$

The transformation corresponding to $d \in G(D)$ sends $u\colon D \to T$ to $G(u)(d)$. This transformation is an isomorphism precisely when each $x \in G(T)$ comes from a unique such u. In that case d is a universal element for the covariant functor G.

For a calculation in $\mathbf{Set}$, take $D = \{0, 1\}$. Evaluation at 0 and 1 identifies a map $D \to X$ with an ordered pair in X^2, and the identification commutes with applying any $X \to Y$ to the pair. Under this identification the identity on D corresponds to $(0, 1)$. The covariant form of Yoneda says that a natural transformation from the ordered-pair functor to itself is determined by the pair it gives on $(0, 1)$ in D. If that pair is (i, j), then naturality at $a\colon D \to X$, where $a(0) = x_0$ and $a(1) = x_1$, forces

$$(x_0, x_1) \longmapsto (a(i), a(j)) = (x_i, x_j).$$

Each choice of (i, j) gives a natural rule by the same formula, since a function applied coordinatewise commutes with coordinate selection. There are exactly

four rules. They are the identity, the swap, and the two rules that repeat a single coordinate.

Consider also the category **Grp** of groups and homomorphisms. A group has an associative multiplication, an identity element, and a multiplicative inverse for each element. Its underlying-set functor $U\colon \mathbf{Grp} \to \mathbf{Set}$ sends a homomorphism to its underlying function. A homomorphism from the additive group $\mathbb{Z}$ to a group G is determined by its value a at 1, since its value at n must be a^n. Conversely, $n \mapsto a^n$ is a homomorphism because $a^{m+n} = a^m a^n$. Evaluation at 1 thus identifies $\mathbf{Grp}(\mathbb{Z}, G)$ with $U(G)$. These identifications are natural in G because postcomposition with a homomorphism carries the image of 1 to its new image.

The covariant Yoneda correspondence now parametrizes all natural transformations $U \Rightarrow U$ by integers. For $n \in \mathbb{Z}$, the corresponding rule sends each $a \in U(G)$ to a^n. Indeed, if a is the image of 1 under $f\colon \mathbb{Z} \to G$, naturality forces the image of a under the rule to be $f(n) = a^n$. Different integers give different rules, as can be seen by evaluating at 1 in $\mathbb{Z}$ itself.

The group $\mathbb{Z}$ represents $G \mapsto \mathbf{Set}(\{*\}, U(G))$ because it is free on one generator. What happens if $\{*\}$ is replaced by an arbitrary alphabet A? In Chapter 5, the corresponding problem for monoids is to represent $M \mapsto \mathbf{Set}(A, U(M))$ for each A, where U now takes the underlying set of a monoid. The representing objects form the free monoid adjunction.

CHAPTER 5

Adjunctions and free constructions

An assignment of symbols to matrices extends to an interpretation of words. Read the symbols from left to right and multiply the corresponding matrices in that order. The extension respects concatenation, and agreement on the one-letter words determines it on every word. The same construction works with any associative multiplication and unit. Its uniqueness gives a bijection between assignments of symbols and homomorphisms out of the set of words. This is a representation in the sense of Chapter 4, now carried out for every alphabet.

5.1 Words and their universal property

Definition 5.1. A *monoid* is a set M with an associative multiplication $(m, n) \mapsto mn$ and a two-sided unit 1_M. A homomorphism $h\colon M \to N$ preserves multiplication and unit. Monoids and their homomorphisms form the category **Mon**.

Associativity gives $(mn)p = m(np)$, while the unit laws give $1_M m = m = m1_M$. Preservation of these operations is stable under function composition, and each identity function preserves both operations. There is a *forgetful functor* $U\colon \mathbf{Mon} \to \mathbf{Set}$. It sends a monoid to its underlying set and a homomorphism to its underlying function. For instance, $U(\mathbb{N}, +, 0) = \mathbb{N}$.

For a set A, called the alphabet, write A^* for the disjoint union of A^n over all integers $n \geq 0$. An element is a finite word $[a_1, \ldots, a_n]$, and the unique word of length zero is $[]$. Concatenation appends one list of entries to another. It is associative because either parenthesization gives the same ordered list of entries. The empty word is the unit, so A^* is a monoid. The function

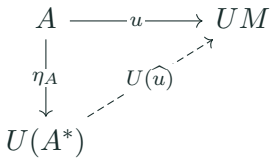
$$\eta_A\colon A \longrightarrow U(A^*), \qquad a \longmapsto [a]$$

inserts each symbol as a one-letter word. In particular, the empty alphabet gives a monoid with one element, its empty word.

Proposition 5.2 (Free monoid on an alphabet)**.** *For every monoid M and function $u\colon A \to UM$, there is a unique homomorphism $\widehat{u}\colon A^* \to M$ satisfying $U(\widehat{u}) \circ \eta_A = u$.*

For fixed A, the proposition identifies A^* as a representative of the covariant functor $M \mapsto \mathbf{Set}(A, UM)$. The map η_A is its universal element in the sense of

Section 4.4.2. The dashed arrow in this diagram in **Set** is the underlying function of the resulting homomorphism.



Proof. Define $\widehat{u}([]) = 1_M$. Evaluate a nonempty word recursively by $\widehat{u}(w[a]) = \widehat{u}(w)u(a)$, where $w[a]$ denotes appending the letter a to w. This defines one value on every word since each nonempty word has a unique last letter and preceding word. In particular, $\widehat{u}([a]) = 1_M u(a) = u(a)$.

We prove $\widehat{u}(wv) = \widehat{u}(w)\widehat{u}(v)$ by induction on the length of v, with w arbitrary. For $v = []$ this is the right unit law. If the identity holds for v, then associativity gives

$$\begin{aligned}\widehat{u}(w(v[a])) &= \widehat{u}((wv)[a]) \\ &= \widehat{u}(wv)u(a) \\ &= (\widehat{u}(w)\widehat{u}(v))u(a) \\ &= \widehat{u}(w)\widehat{u}(v[a]).\end{aligned}$$

Thus $\widehat{u}$ is a homomorphism. Any homomorphism with the required values on one-letter words sends $[]$ to 1_M and evaluates $w[a]$ by multiplying its value on w by $u(a)$. Induction on word length forces the values just constructed. □

Example 5.3 (Ordered products of matrices)**.** Let $A = \{a, b\}$ and let M be the monoid of two-by-two integer matrices with standard matrix multiplication $(P, Q) \mapsto PQ$. The word $[a, b]$ places the factor for a on the left. As a linear map on column vectors, PQ applies Q first. Set

$$u(a) = P = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad u(b) = Q = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Then the extension gives

$$\widehat{u}([a, b]) = PQ = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \widehat{u}([b, a]) = QP = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

In particular, the two possible orders give distinct values. For $[a, b, a]$ the product is $PQP = \begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}$, and the empty word evaluates to the identity matrix. A longer word can be split into any two consecutive pieces. Evaluating each piece and then multiplying their values gives the value of the whole word.

The construction of A^* extends to a functor $F\colon \mathbf{Set} \to \mathbf{Mon}$. For $f\colon A \to A'$, define $F(f)$ by applying f to every entry of a word. This preserves the empty word and concatenation. Applying two functions in succession to the entries agrees with applying their composite, and applying the identity function leaves every entry fixed. Therefore F preserves composition and identities.

5.2 The hom-set bijection

We often drop parentheses in object expressions such as $FA = F(A)$ and $UFA = U(F(A))$. In a composition of functors, the rightmost functor acts first. We write parentheses on morphism applications such as $F(a)$ when writing the arrows explicitly.

The free monoid property gives a bijection

$$\mathbf{Mon}(FA, M) \longrightarrow \mathbf{Set}(A, UM), \qquad h \longmapsto U(h) \circ \eta_A.$$

The inverse sends an assignment u to $\widehat{u}$. Changing the alphabet by precomposition changes the restriction by the same precomposition. Changing the target monoid by postcomposition changes the restriction by the underlying function of that homomorphism. These compatibility laws are part of an adjunction.

Definition 5.4. Let $\mathcal{C}$ and $\mathcal{D}$ be locally small categories and let $F\colon \mathcal{C} \to \mathcal{D}$ and $U\colon \mathcal{D} \to \mathcal{C}$ be functors. An *adjunction* $F \dashv U$ is a family of bijections

$$\Phi_{A,B}\colon \mathcal{D}(FA, B) \xrightarrow{\sim} \mathcal{C}(A, UB)$$

natural in A and B. For $a\colon A' \to A$ in $\mathcal{C}$, $b\colon B \to B'$ in $\mathcal{D}$, and $h\colon FA \to B$, the naturality condition is

$$\Phi_{A',B'}(b \circ h \circ F(a)) = U(b) \circ \Phi_{A,B}(h) \circ a. \tag{5.1}$$

The functor F is the *left adjoint* and U the *right adjoint*.

In the naturality equation, maps act by precomposition in the first variable and by postcomposition in the second. Its square makes both directions explicit.

$$\begin{array}{ccc} \mathcal{D}(FA, B) & \xrightarrow{\Phi_{A,B}} & \mathcal{C}(A, UB) \\ \downarrow{\scriptstyle b\circ(-)\circ F(a)} & & \downarrow{\scriptstyle U(b)\circ(-)\circ a} \\ \mathcal{D}(FA', B') & \xrightarrow{\Phi_{A',B'}} & \mathcal{C}(A', UB') \end{array}$$

To check naturality for the free monoid, take $f\colon A' \to A$, a homomorphism $g\colon M \to N$, and $h\colon FA \to M$. The underlying function of $F(f)$ sends $[a']$ to $[f(a')]$, so $U(F(f)) \circ \eta_{A'} = \eta_A \circ f$. Restricting $g \circ h \circ F(f)$ to one-letter words therefore gives $U(g) \circ (U(h) \circ \eta_A) \circ f$, as required. This proves $F \dashv U$ for the free-monoid and forgetful functors.

Example 5.5 (Images and inverse images)**.** Order the power sets $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ by inclusion, and regard each as a category. For a function $f\colon X \to Y$, direct image $S \mapsto f(S)$ and inverse image $T \mapsto f^{-1}(T)$ are monotone. There is a unique arrow $f(S) \to T$ exactly when $f(S) \subseteq T$. This is equivalent to $S \subseteq f^{-1}(T)$ since each $x \in S$ contributes the element $f(x)$ to $f(S)$. The two hom-sets therefore have the same cardinality, either zero or one. Each codomain in the naturality square has at most one element, so the two functions around the square agree. Hence direct image is left adjoint to inverse image.

Example 5.6 (Paths and the diagonal)**.** Work with small directed graphs and small categories in a fixed universe. The path construction takes a directed graph G to a category $\mathrm{Path}(G)$. Write $U_{\mathrm{gr}}D$ for the graph with the objects of a category D as vertices and all its arrows as edges. A graph map $G \to U_{\mathrm{gr}}D$ preserves sources and targets. By Chapter 4, it extends uniquely to a functor $\mathrm{Path}(G) \to D$. The bijection commutes with relabeling the source graph and applying a functor to the target. Thus the path construction is left adjoint to U_{gr}.

Suppose $\mathcal{C}$ is locally small and has chosen binary products and coproducts. The diagonal functor $\Delta\colon \mathcal{C} \to \mathcal{C} \times \mathcal{C}$ sends T to (T, T) and a map to the pair of two copies of that map. The product and coproduct properties give

$$(\mathcal{C} \times \mathcal{C})\big(\Delta T, (A, B)\big) \cong \mathcal{C}(T, A \times B),$$
$$\mathcal{C}(A \amalg B, T) \cong (\mathcal{C} \times \mathcal{C})\big((A, B), \Delta T\big).$$

Products act on a pair of arrows by applying the arrows to the two components. Coproducts act by applying them on the two corresponding summands. These rules commute with taking components or restrictions, so the bijections are natural. The result is $\amalg \dashv \Delta \dashv \times$.

5.3 Unit, counit, and the triangle identities

An adjunction specifies two maps by applying its bijections to identities. For A in $\mathcal{C}$ and B in $\mathcal{D}$, define

$$\eta_A = \Phi_{A,FA}(\mathrm{id}_{FA})\colon A \to UFA, \qquad \varepsilon_B = \Phi^{-1}_{UB,B}(\mathrm{id}_{UB})\colon FUB \to B.$$

The families η and ε are called the *unit* and *counit*. Naturality of Φ applied to $h\colon FA \to B$ by postcomposition at id_{FA} gives $\Phi_{A,B}(h) = U(h) \circ \eta_A$. For $k\colon A \to UB$, precomposing at ε_B gives

$$\Phi_{A,B}(\varepsilon_B \circ F(k)) = \Phi_{UB,B}(\varepsilon_B) \circ k = k.$$

Thus the forward and inverse bijections are

$$h \longmapsto U(h) \circ \eta_A, \qquad k \longmapsto \varepsilon_B \circ F(k). \tag{5.2}$$

The unit and counit are natural transformations. For $a\colon A \to A'$, the map $\Phi_{A,FA'}(F(a))$ equals $U(F(a))\circ\eta_A$ by postcomposition at id_{FA}. By precomposition at $\mathrm{id}_{FA'}$ it equals $\eta_{A'} \circ a$. This proves naturality of η. For $b\colon B \to B'$, each of $b \circ \varepsilon_B$ and $\varepsilon_{B'} \circ F(U(b))$ has image $U(b)$ under $\Phi_{UB,B'}$ by (5.1). Injectivity gives their equality, the naturality of ε.

Transposing η_A back gives id_{FA}, and transposing ε_B forward gives id_{UB}. By (5.2), these statements are the *triangle identities*

$$\varepsilon_{FA} \circ F(\eta_A) = \mathrm{id}_{FA}, \qquad U(\varepsilon_B) \circ \eta_{UB} = \mathrm{id}_{UB}\,. \tag{5.3}$$

For the free monoid, the first triangle describes a word of letters passing to a word of one-letter words and then being concatenated.

$$\begin{array}{ccc} FA & \xrightarrow{F(\eta_A)} & FUFA \\ & \searrow^{\mathrm{id}_{FA}} & \downarrow{\scriptstyle \varepsilon_{FA}} \\ & & FA \end{array}$$

Here $\varepsilon_M\colon FUM \to M$ evaluates a word of elements of M by multiplying them. In the second triangle, inserting an element as a one-letter word and evaluating returns that element.

$$\begin{array}{ccc} UB & \xrightarrow{\eta_{UB}} & UFUB \\ & \searrow^{\mathrm{id}_{UB}} & \downarrow{\scriptstyle U(\varepsilon_B)} \\ & & UB \end{array}$$

For the free monoid, put $B = FA = A^*$ in the second triangle. On $[a, b]$, the underlying function of $F(\eta_A)$ acts letter by letter and gives $[[a], [b]]$. The unit η_{UFA} treats the whole word as one element and gives $[[a, b]]$. The first result has two outer letters and the second has one. In each case evaluation concatenates the inner words and returns $[a, b]$.

Proposition 5.7. *Natural transformations $\eta\colon \mathrm{Id}_{\mathcal{C}} \to UF$ and $\varepsilon\colon FU \to \mathrm{Id}_{\mathcal{D}}$ satisfying (5.3) determine an adjunction.*

Proof. Use the maps in (5.2). Starting with $h\colon FA \to B$, the composite sends h to

$$\varepsilon_B \circ F(U(h) \circ \eta_A) = h \circ \varepsilon_{FA} \circ F(\eta_A) = h,$$

where the first equality uses functoriality of F and naturality of ε. Starting with $k\colon A \to UB$, the other composite gives

$$U(\varepsilon_B \circ F(k)) \circ \eta_A = U(\varepsilon_B) \circ \eta_{UB} \circ k = k,$$

by functoriality of U and naturality of η. Thus the maps are inverse. Naturality in the codomain follows from functoriality of U. For a change of domain $a\colon A' \to A$, the identity $U(F(a)) \circ \eta_{A'} = \eta_A \circ a$ gives naturality in the domain. □

5.4 Constructing a left adjoint from universal maps

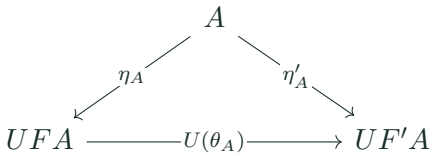
For a given functor $U\colon \mathcal{D} \to \mathcal{C}$, it suffices to construct one universal map for each object A of $\mathcal{C}$. Specifically, suppose we choose FA in $\mathcal{D}$ and $\eta_A\colon A \to UFA$ with the property that any $k\colon A \to UB$ has a unique factorization $k = U(h) \circ \eta_A$ by $h\colon FA \to B$. This makes η_A a universal element for the functor $B \mapsto \mathcal{C}(A, UB)$ on $\mathcal{D}$, so FA represents this functor. The free-monoid proposition constructs one such representative for each A.

For a map $a\colon A \to A'$, take $F(a)\colon FA \to FA'$ to be the unique map satisfying $U(F(a)) \circ \eta_A = \eta_{A'} \circ a$. For $a = \mathrm{id}_A$, the identity of FA has the required factorization, so $F(\mathrm{id}_A) = \mathrm{id}_{FA}$. For composable $a\colon A \to A'$ and $a'\colon A' \to A''$, the composite $F(a') \circ F(a)$ satisfies

$$U(F(a') \circ F(a)) \circ \eta_A = U(F(a')) \circ \eta_{A'} \circ a = \eta_{A''} \circ a' \circ a.$$

Uniqueness gives $F(a' \circ a) = F(a') \circ F(a)$. The restriction bijections $h \mapsto U(h) \circ \eta_A$ are natural by the defining equation for $F(a)$ and functoriality of U. This constructs the entire left adjoint from the universal maps [2, § 4.2].

The comparison of two left adjoints starts with the uniqueness of representing objects and their universal elements from Chapter 4. Suppose $\eta'_A\colon A \to UF'A$ has the same universal property. Applying the property of η_A to η'_A produces a unique $\theta_A\colon FA \to F'A$ with $U(\theta_A) \circ \eta_A = \eta'_A$. Reversing the roles produces a unique $\psi_A\colon F'A \to FA$ with $U(\psi_A) \circ \eta'_A = \eta_A$. The resulting diagram compares the two universal maps.



The composite $\psi_A \circ \theta_A$ extends η_A, and so does id_{FA}. Uniqueness gives $\psi_A \circ \theta_A = \mathrm{id}_{FA}$. The reverse composite extends η'_A and is $\mathrm{id}_{F'A}$ by the same argument. Thus θ_A is the unique isomorphism compatible with the two universal maps.

When both families of universal maps are chosen for every A, these isomorphisms are natural. For $a\colon A \to A_1$, the maps $F'(a) \circ \theta_A$ and $\theta_{A_1} \circ F(a)$ from FA to $F'A_1$ both extend $\eta'_{A_1} \circ a$. For the first map use $U(\theta_A) \circ \eta_A = \eta'_A$ followed by naturality of η'. For the second use naturality of η followed by $U(\theta_{A_1}) \circ \eta_{A_1} = \eta'_{A_1}$. Uniqueness gives $F'(a) \circ \theta_A = \theta_{A_1} \circ F(a)$, so the left adjoints are naturally isomorphic.

5.5 Generators and relations

A presentation prescribes generators and equations. A map out of the presented object is uniquely determined by an assignment of the generators that satisfies the

equations. In the examples below, the target and the meaning of an assignment vary. Each construction proves the same type of uniqueness by exhibiting the extending map.

Example 5.8 (Words as normal forms for product expressions)**.** Construct a set of finite expressions from symbols $a \in A$, a constant 1, and a binary operation written (st) on expressions s and t. Let $\sim$ be the smallest equivalence relation on expressions compatible with the binary operation and containing

$$((st)r) \sim (s(tr)), \qquad (1s) \sim s, \qquad (s1) \sim s$$

for all expressions r, s, t. It exists because the intersection of all relations with those properties again has those properties. Compatibility makes $[s][t] = [(st)]$ a well-defined product on equivalence classes. The displayed relations give associativity and make $[1]$ its unit.

Given an assignment $u\colon A \to UM$, evaluate a symbol a at $u(a)$, evaluate 1 at 1_M, and evaluate (st) by multiplying the values of s and t. Equality of evaluations is an equivalence relation compatible with multiplication and containing the three kinds of pairs above. Therefore evaluation factors through the quotient. Its factor is a homomorphism, and induction on expressions shows that every homomorphism extending u has these values. The quotient has the free-monoid universal property.

The unique compatible isomorphism from the quotient to A^* sends each expression to the word obtained by reading its symbols in order and treating 1 as the empty word. Consequently two expressions represent the same class exactly when those words coincide. For example, $((a1)(bc))$ and $(a((b1)c))$ both represent $[a, b, c]$. The universal property thus gives a test for equality in the quotient.

5.5.1 Presenting a ring and a module

A commutative unital ring has an abelian group under addition and a commutative monoid under multiplication, with multiplication distributing over addition. For such a ring R, an R-algebra here is a commutative unital ring A equipped with a unital ring map $\rho\colon R \to A$. A map of R-algebras preserves addition, multiplication, and the unit, and commutes with the designated maps out of R.

The polynomial algebra $R[t]$ is free on one generator t. Given $a \in A$, the value of every polynomial under an R-algebra map sending t to a is forced to be

$$\mathrm{ev}_a\Big(\sum_i c_i t^i\Big) = \sum_i \rho(c_i) a^i.$$

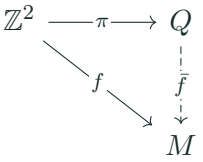
This rule defines a map because polynomial addition is coefficientwise and the distributive law gives the product of the two evaluated sums.

To present an element subject to the equation $t^2 = t$, form $P = R[t]/(t^2 - t)$. Two polynomials give the same class in P exactly when their difference is a multiple of $t^2 - t$. The multiples are closed under addition, negation, and multiplication by arbitrary polynomials, so the polynomial ring operations descend to the classes. The generator $[t]$ satisfies $[t]^2 = [t]$. For $a \in A$ with $a^2 = a$, evaluation sends each multiple of $t^2 - t$ to zero. It therefore descends to a unique R-algebra map $P \to A$ sending $[t]$ to a. Conversely, any such map sends $[t]$ to an idempotent, an element whose square equals itself. Thus P presents one idempotent over R. For $R = \mathbb{Z}$ and $A = \mathbb{Z}/6\mathbb{Z}$, the possible images of $[t]$ are $0, 1, 3, 4$. Indeed, $a(a-1)$ evaluated at the six residues gives $0, 0, 2, 0, 0, 2$ modulo 6. This gives four maps $P \to A$.

A $\mathbb{Z}$-module is an abelian group with integer scalar multiplication given by repeated addition and additive inverses. The module $\mathbb{Z}^2$ is free on $e_1 = (1, 0)$ and $e_2 = (0, 1)$. For $n \geq 1$, impose the relation $ne_1 = 0$ to obtain

$$Q = \mathbb{Z}^2/\mathbb{Z}(ne_1) \cong (\mathbb{Z}/n\mathbb{Z}) \oplus \mathbb{Z}.$$

The isomorphism sends the class of (r, s) to $(\bar{r}, s)$. To assign e_1, e_2 to u, v in a module M, first define $f(r, s) = ru + sv$. This map descends along the quotient π precisely when $nu = 0$, since two representatives differ by an integer multiple of ne_1. This makes π the coequalizer of the two maps $\mathbb{Z} \rightrightarrows \mathbb{Z}^2$ given by $k \mapsto kne_1$ and $k \mapsto 0$.



The induced map is forced by $\bar{f}(\bar{r}, s) = ru + sv$. For $n = 2$ and $M = \mathbb{Z}/6\mathbb{Z}$, there are two choices of u, the residues 0 and 3, and six choices of v. The module presentation therefore gives twelve homomorphisms $Q \to M$.

5.5.2 Presenting a propositional theory

Take two proposition symbols p, q and the premise $p \Rightarrow q$. Classical conjunction, disjunction, and negation are interpreted by $\wedge, \vee, \neg$ in a Boolean algebra. A Boolean algebra is a bounded distributive lattice with complements. Thus each pair has a meet $\wedge$ and a join $\vee$, there are least and greatest elements $0, 1$, and $\neg x$ satisfies $x \wedge \neg x = 0$ and $x \vee \neg x = 1$. In a powerset, these operations are intersection, union, complement, the empty set, and the whole set. A homomorphism preserves the two operations on pairs, complements, and both constants.

For p interpreted as a and q as b, the premise requires $\neg a \vee b = 1$. Meeting this equation with a gives $a \wedge b = a$. Conversely, if $a \wedge b = a$, then $a \leq b$, so

$\neg a \vee b \geq \neg a \vee a = 1$. The premise can therefore be imposed as the relation $p \wedge q = p$.

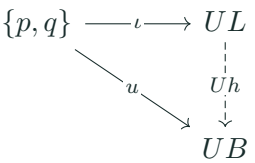
Write v_{ij} for the two-valued assignment with $p = i$ and $q = j$, with 1 denoting truth. The assignments that satisfy the premise form the set $S = \{v_{00}, v_{01}, v_{11}\}$. In the powerset Boolean algebra $L = \mathcal{P}(S)$, take

$$P = \{v_{11}\}, \qquad Q = \{v_{01}, v_{11}\}.$$

These are the truth sets of p and q, so $P \wedge Q = P$.

Proposition 5.9 (Free presentation by an implication)**.** *For every Boolean algebra B and $a, b \in B$ with $a \leq b$, there is a unique Boolean algebra homomorphism $h\colon L \to B$ such that $h(P) = a$ and $h(Q) = b$.*

The diagram is in **Set**, with U taking a Boolean algebra to its underlying set. The map ι sends p, q to P, Q, and u sends them to a, b. The dashed map is the underlying function of the unique homomorphism.



Proof. Assign the singleton truth sets $\{v_{00}\}, \{v_{01}\}, \{v_{11}\}$ the elements $e_{00} = \neg b$, $e_{01} = \neg a \wedge b$, and $e_{11} = a$, respectively. Their pairwise meets are zero. The first has zero meet with each of the others because both lie below b. The second and third have zero meet by the complement law for a. Since $a \leq b$, distributivity gives $(\neg a \wedge b) \vee a = (\neg a \vee a) \wedge b = b$. Thus the three elements join to 1.

For $E \subseteq S$, define $h(E) = \bigvee_{v_{ij} \in E} e_{ij}$, where the empty join is 0. This preserves union by the join laws and preserves intersection by distributivity and the zero pairwise meets. The images of E and $S \setminus E$ have meet 0 and join 1, so h preserves complements and both constants. By construction $h(P) = a$ and $h(Q) = b$.

Every homomorphism taking P, Q to a, b has the stated values on the singletons, since $\{v_{00}\} = \neg Q$, $\{v_{01}\} = Q \wedge \neg P$ and $\{v_{11}\} = P$. It must preserve finite joins, so these values determine $h(E)$ for all E. □

The proposition characterizes L by its maps out. A homomorphism from L to B amounts to a permissible assignment of the generators p, q in B. As with the free monoid, postcomposition with a homomorphism of targets acts on the assignment by applying that homomorphism to the values of the generators.

The eight subsets of S give all the formula classes. Every subset is a union of the singletons expressed in the proof. To compare two formulas, evaluate their truth sets in L. If these sets agree, their images under every permissible

assignment agree by the proposition. If they differ, a valuation belonging to exactly one set distinguishes the formulas. For example, $p \wedge \neg q$ has empty truth set under the premise, while $q \Rightarrow p$ has truth set $\{v_{00}, v_{11}\}$ and fails at v_{01}.

For the quotient construction, let $V = S \cup \{v_{10}\}$ be the four assignments before imposing the premise. The same join construction represents arbitrary assignments a, b, using the four pairwise disjoint elements formed by meeting a choice of $a, \neg a$ with a choice of $b, \neg b$. Thus $\mathcal{P}(V)$ is free on p, q before imposing the relation. The quotient map $\mathcal{P}(V) \to \mathcal{P}(S)$ is intersection with S. It identifies subsets that agree on S, sending the singleton $\{v_{10}\}$, the truth set of $p \wedge \neg q$, to 0. The quotient is also a coequalizer. From the free Boolean algebra on one symbol, obtained by the one-variable truth-table construction, take the two homomorphisms to $\mathcal{P}(V)$ assigning that symbol the truth sets of $p \wedge q$ and of p, respectively. A map out of $\mathcal{P}(V)$ equalizes them precisely when its assignment satisfies $a \wedge b = a$. The proposition then gives the unique factorization through $\mathcal{P}(S)$.

5.5.3 The logic–semantics adjunction

A premise determines the valuations in which it is true. Conversely, a class of valuations determines the formulas true throughout that class. These two operations form an adjunction already in the propositional language on p, q.

Let Form be the set of formulas in these variables and V the set of four valuations v_{ij} from the preceding presentation. Write $v \models \varphi$ when the valuation v makes φ true. A theory $T \subseteq \mathrm{Form}$ is a set of premises, and a class of models $K \subseteq V$ is a set of valuations. Define

$$\mathrm{Mod}(T) = \{v \in V \mid v \models \varphi \text{ for every } \varphi \in T\},$$
$$\mathrm{Th}(K) = \{\varphi \in \mathrm{Form} \mid v \models \varphi \text{ for every } v \in K\}.$$

Adding premises shrinks the model class, and adding valuations shrinks the set of formulas true in all of them. Give $\mathcal{P}(\mathrm{Form})$ its inclusion order and $\mathcal{P}(V)^{\mathrm{op}}$ the reverse inclusion order. The two assignments are monotone and define functors

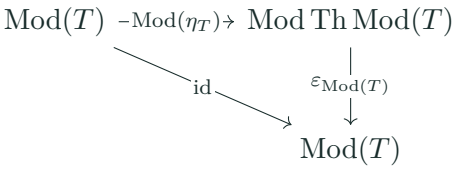
$$\mathrm{Mod}\colon \mathcal{P}(\mathrm{Form}) \longrightarrow \mathcal{P}(V)^{\mathrm{op}}, \qquad \mathrm{Th}\colon \mathcal{P}(V)^{\mathrm{op}} \longrightarrow \mathcal{P}(\mathrm{Form}).$$

For the adjunction $\mathrm{Mod} \dashv \mathrm{Th}$, the hom-set bijection reduces to the equivalence

$$\mathrm{Mod}(T) \leq_{\mathrm{op}} K \quad \Longleftrightarrow \quad K \subseteq \mathrm{Mod}(T) \quad \Longleftrightarrow \quad T \subseteq \mathrm{Th}(K).$$

The last two conditions both say that every $v \in K$ satisfies every $\varphi \in T$. When they hold, each hom-set is a singleton. Otherwise both hom-sets are empty. Every hom-set involved has at most one element, so the naturality squares commute. The proof works for any fixed sets of sentences and interpretations with a satisfaction relation between them.

The unit η_T is the arrow given by the inclusion $T \subseteq \mathrm{Th}(\mathrm{Mod}(T))$. The counit is the arrow $\varepsilon_K \colon \mathrm{Mod}(\mathrm{Th}(K)) \to K$ in the reverse order. In ordinary inclusion it says $K \subseteq \mathrm{Mod}(\mathrm{Th}(K))$, since each valuation in K satisfies every formula true throughout K. The triangle below is in $\mathcal{P}(V)^{\mathrm{op}}$.



The top and right arrows give mutual inclusions, so T and $\mathrm{Th}(\mathrm{Mod}(T))$ have the same models. Applying Th to this equality proves that the monotone operation $T \mapsto \mathrm{Th}(\mathrm{Mod}(T))$ is idempotent. Together with the unit, this makes it a closure operation. It closes a theory under *semantic consequence*, the formulas true in every model of the premises.

For $T = \{p \Rightarrow q\}$, where $p \Rightarrow q$ abbreviates $\neg p \vee q$, the model class is $S = \{v_{00}, v_{01}, v_{11}\}$. Its theory contains $\neg q \Rightarrow \neg p$, which is true at all three valuations. The valuation v_{01} refutes $q \Rightarrow p$, so the converse implication lies outside the closure. In the Boolean presentation above, a formula denotes the subset of S on which it is true. Its denotation is all of S exactly when the formula belongs to $\mathrm{Th}(S)$. Two denotations coincide exactly when the equivalence of their formulas belongs to $\mathrm{Th}(S)$. An inconsistent premise set such as $\{p, \neg p\}$ has no models, and its closure is all of Form by the condition defining $\mathrm{Th}(\varnothing)$.

For a sound and complete proof calculus, this semantic closure equals the set of formally derivable consequences of the premises. The satisfaction calculation above establishes the adjunction independently of the choice of calculus. Speirs [3] gives the corresponding Galois connection between first-order theories and classes of structures.

There is a further question when models have operations. A family of maps $U(B)^n \to U(B)$ that commutes with every model homomorphism defines an n-ary natural operation. The group powers in Chapter 4 give the unary case for groups. The later chapter on logic and semantics develops Lawvere's structure–semantics adjunction [4], which relates operations of this kind to interpretations of algebraic theories in a category of models.

5.5.4 Product types from generators and equations

Start with two base types A, B. Form new types using the unit type 1 and binary products $S \times T$. A context Γ lists variables with their types. A judgment $\Gamma \vdash t \colon S$ says that t is a term of type S using variables from Γ. A variable has its declared type, and $*$ is a term of type 1. Terms $u \colon S$ and $v \colon T$ in a common

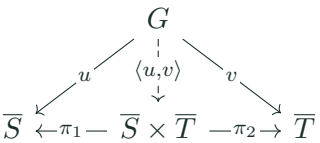
context form a pair $\langle u, v\rangle \colon S \times T$. A term $w \colon S \times T$ has projections $\pi_1 w \colon S$ and $\pi_2 w \colon T$.

For terms in a common context impose the product equations

$$\pi_1\langle u, v\rangle = u, \quad \pi_2\langle u, v\rangle = v,$$
$$\langle \pi_1 w, \pi_2 w\rangle = w, \qquad t = * \quad \text{if } t \colon 1.$$

The first two are the β-equations for products. They compute a projection of a pair. The third is the η-equation. It reconstructs a term from its two projections. The last equation makes the term of unit type unique. Equality of terms is the least equivalence compatible with the term operations and containing these equations at every permitted instance.

To interpret this presentation in a category with finite products, choose objects $\overline{A}, \overline{B}$ for the two base types. Interpret 1 as a terminal object, interpret type formation by products, and interpret the context as the product G of the objects assigned to its variables. A term $\Gamma \vdash t \colon S$ then gives a map $G \to \overline{S}$. Variables give context projections, $*$ gives the terminal map, and pairing and projection give the corresponding operations on arrows. For example, the map for a pair is the unique dashed arrow making both triangles commute.



The commuting triangles give the two β-equations. Uniqueness of the dashed arrow gives η, and terminality gives the equation for $*$. Since equalities are preserved by each operation, the interpretation respects every equality generated by the presentation.

Consider the judgment

$$x \colon A,\ y \colon B,\ z \colon A \ \vdash\ \langle z, \pi_2\langle x, y\rangle\rangle \colon A \times B.$$

The second β-equation reduces this term to $\langle z, y\rangle$. In this context the possible normal forms at type $A \times B$ are exactly $\langle x, y\rangle$ and $\langle z, y\rangle$. To see why, normalize inductively on the term in any context of base-type variables. A variable or $*$ is already a normal form. To normalize a pair, normalize both entries and pair the results. To normalize a projection, first normalize its argument. The argument has product type, so its normal form is a pair. The corresponding β-equation selects its normalized entry. Thus a normal form has the shape of its type, with a variable of matching type at each base-type leaf and $*$ at each unit leaf. The two forms in the example are distinct. In **Set**, interpret A and B by $\{0, 1\}$ and evaluate in the environment $(x, y, z) = (0, 0, 1)$. The forms interpret as $(0, 0)$ and $(1, 0)$, respectively.

The categorical presentation uses contexts of base-type variables as objects, recorded by their lists of types. An arrow $\Gamma \to \Delta$ selects, for each output position in Δ, an input position of the same type in Γ. This description follows from normalization, applied to the base-type term in each output slot. Any two different selections differ at some slot and can be distinguished at that slot by assigning values 0 and 1 to the differing inputs in **Set**. Composition substitutes a selection into a selection, and the identity selects each position itself. Concatenation of lists gives the product, since any selection for a concatenated output gives a unique selection for each of its two parts. The empty list is terminal because the empty selection is unique.

This is a free finite-product category on the two base types. Given $\overline{A}, \overline{B}$ as above, a product-preserving functor sends each context to the product of its assigned objects. Its value on a selection is forced by its components, which are the indicated projections. These maps preserve composition by the β-equations and identities by product uniqueness. Any such functor has canonical comparison isomorphisms to this interpretation, given at each context by its product projections. The same projections make the comparisons natural on all selections and determine them uniquely once the comparisons on A, B are fixed to be the identities.

5.6 Character pairings and Pontryagin duality

Let G, H be abelian groups, written additively, and let $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ be the circle group. For this section take only the group structure of $\mathbb{T}$. An algebraic character of G is a homomorphism $G \to \mathbb{T}$. The characters form a group $DG = \mathrm{Hom}_{\mathbf{Ab}}(G, \mathbb{T})$ by pointwise addition. A homomorphism $f\colon G \to G'$ gives $Df\colon DG' \to DG$ by precomposition.

Any homomorphism $u\colon G \to DH$ defines a pairing $b(g,h) = u(g)(h)$, additive in each variable. Conversely, a pairing additive in each variable gives such a u, because $h \mapsto b(g,h)$ is a character and $g \mapsto b(g,-)$ preserves addition. Reading the variables in the other order gives the same correspondence with $H \to DG$. Thus

$$\mathrm{Hom}_{\mathbf{Ab}}(G, DH) \cong \mathrm{Bilin}(G, H; \mathbb{T}) \cong \mathrm{Hom}_{\mathbf{Ab}}(H, DG).$$

Precomposition on either variable commutes with both correspondences. This gives $D \dashv D^{\mathrm{op}}$, where the left adjoint goes from $\mathbf{Ab}$ to $\mathbf{Ab}^{\mathrm{op}}$. Both functors use the same object assignment and reverse maps by precomposition. The identity of DG corresponds to evaluation

$$\eta_G\colon G \longrightarrow DDG, \qquad \eta_G(g)(\chi) = \chi(g).$$

For example, evaluation at 1 identifies $D\mathbb{Z}$ with $\mathbb{T}$, since the character with value t at 1 sends m to mt. Section 8.6 proves biduality on finite abelian

groups. Section 16.6 equips character groups with a topology and states the full Pontryagin duality theorem.

For a one-letter alphabet, each natural number is the length of exactly one word. Under this identification, concatenation adds lengths and the empty word has length zero. Giving the letter a value m in any monoid therefore determines a homomorphism that sends n to m^n. Appending the one letter corresponds to the successor operation. The next chapter uses a universal property of zero and successor to derive recursion, induction, and the arithmetic operations on the natural numbers.

Part II Numbers and symmetry

CHAPTER 6

Natural numbers from recursion

Start with a point x_0 in a set X and a step $t\colon X \to X$. After zero steps the state is x_0. After one more step it is the image under t of the preceding state. To interpret this rule as a map on the natural numbers, the domain needs a zero and an operation of taking one more step. Its universal property says that a choice of x_0, t in any target admits exactly one interpretation. That same property lets us specify arithmetic operations by recursion.

6.1 Zero, successor, and the recursion map

Let $\mathcal{C}$ have a terminal object 1. A point of X means a map $1 \to X$. The starting point and step give a triple (X, x_0, t). A map $f\colon X \to Y$ of triples must preserve the starting point and commute with the step. Composites and identities preserve both equations, so these triples and their maps form a category.

Definition 6.1. A *natural numbers object* in the zero-and-successor sense is a triple $(N, 0, s)$ initial in the category of these triples. Thus for every $x_0\colon 1 \to X$ and $t\colon X \to X$ there is a unique map $u\colon N \to X$ with $u \circ 0 = x_0$ and $u \circ s = t \circ u$.

Both squares in the diagram must commute for the same map u.

$$\begin{array}{ccccc} 1 & \xrightarrow{0} & N & \xrightarrow{s} & N \\ \downarrow{\scriptstyle \mathrm{id}_1} & & \downarrow{\scriptstyle u} & & \downarrow{\scriptstyle u} \\ 1 & \xrightarrow{x_0} & X & \xrightarrow{t} & X \end{array}$$

Both halves of the specification matter. In **Set**, choosing a map that commutes with the step gives the recursion rule. Its value at zero fixes the starting state. An initial object is unique up to a unique isomorphism in its category. Here the isomorphism preserves the designated zero and successor. Indeed, for two such triples initiality gives one map in each direction, and the two composites are identities by the uniqueness of initial maps.

Proposition 6.2. *In* **Set**, *the finite words on a one-letter alphabet form a natural numbers object, with zero the empty word and successor the operation of appending the letter.*

Proof. Send the empty word to $x_0(*)$, where $*$ is the point of 1, and each word of length n to $t^n(x_0(*))$. Here t^n is the n-fold composite from Chapter 1. The formula preserves zero and appending a letter corresponds to applying t. Every word is obtained by appending its letters to the empty word. The two equations therefore force its image to be the same iterate. □

The set construction uses finite lists, introduced in Chapter 5. Their lengths give the conventional set $\mathbb{N} = \{0, 1, 2, \ldots\}$. We write 0 for its distinguished element and $s(n)$ for the successor of n. In constructing this model we used finite iteration in the surrounding set theory. The next argument derives induction on the resulting object from its universal property. That property makes the recursion map independent of which carrier realizes the numbers. For another account of this property in the categorical axioms for sets, see Leinster [5].

6.2 Induction from initiality

Suppose $S \subseteq \mathbb{N}$ contains 0 and is closed under successor. The restriction of successor to S, together with its element 0, makes S a triple with an inclusion $i\colon S \to \mathbb{N}$. Initiality gives a map $r\colon \mathbb{N} \to S$ of triples. Then $i \circ r$ preserves zero and successor, and so does $\mathrm{id}_{\mathbb{N}}$. Initiality forces $i \circ r = \mathrm{id}_{\mathbb{N}}$. Thus every n is the image of the element $r(n) \in S$, so $S = \mathbb{N}$.

Apply this argument to the subset of numbers satisfying a proposition $P(n)$. A proof of $P(0)$ puts zero in the subset. A proof of $P(n) \Rightarrow P(s(n))$ makes the subset closed under successor. The conclusion is $P(n)$ for every n. This is induction. The argument also applies to a monomorphism $i\colon S \to N$ of triples in an arbitrary category with a natural numbers object. Initiality again gives $i \circ r = \mathrm{id}_N$, and cancellation by i in $i \circ (r \circ i) = i \circ \mathrm{id}_S$ gives $r \circ i = \mathrm{id}_S$.

6.3 Recursion with parameters

Let A be a set of parameters. For a starting map $p\colon A \to X$ and a step $k\colon A \times X \to X$, the recursion holds $a \in A$ fixed while it iterates. The equations are

$$R(a, 0) = p(a), \qquad R(a, s(n)) = k(a, R(a, n)).$$

To see existence and uniqueness together, apply the natural numbers property to the set X^A of functions $A \to X$. Its starting point is p, and its step sends a function f to the function $a \mapsto k(a, f(a))$. The resulting map $\mathbb{N} \to X^A$ gives $R\colon A \times \mathbb{N} \to X$ by evaluation at a. Conversely, every R satisfying the two equations gives such a recursion map into X^A by taking its functions of a. The

step equation is the statement that the following square commutes.

$$\begin{array}{ccc} A \times \mathbb{N} & \xrightarrow{\langle \pi_A, R \rangle} & A \times X \\ \downarrow{\scriptstyle \mathrm{id}_A \times s} & & \downarrow{\scriptstyle k} \\ A \times \mathbb{N} & \xrightarrow{\quad R \quad} & X \end{array}$$

If a step needs the iteration number, keep that number as part of its state. For example, use the initial state $(0, 0)$ and the step $(i, a) \mapsto (s(i), a + s(i))$ on $\mathbb{N} \times \mathbb{N}$, once addition has been defined. The first projection is n at the n-th step by induction. The second starts at zero and adds the new value of the first at each step. Thus the first five states are

$$(0, 0), \quad (1, 1), \quad (2, 3), \quad (3, 6), \quad (4, 10).$$

This is a recursion for the triangular numbers, implemented by a fixed endomorphism of the enlarged state set.

6.4 Addition and cancellation

In parameterized recursion take $A = X = \mathbb{N}$, $p(m) = m$, and $k(m, x) = s(x)$. The resulting map is called addition. It is characterized by

$$m + 0 = m, \qquad m + s(n) = s(m + n). \tag{6.1}$$

The two equations make $m + n$ the result of starting at m and iterating s exactly n times. The first few values of $n \mapsto 2 + n$ are $2, 3, 4, 5$.

Proposition 6.3. *Addition on $\mathbb{N}$ is associative and commutative, has identity 0, and is cancellative.*

Proof. By recursion, $n \mapsto 0 + n$ and $n \mapsto n$ both start at zero and commute with successor, so $0 + n = n$. Also for fixed m the two functions $n \mapsto s(m) + n$ and $n \mapsto s(m + n)$ both start at $s(m)$ and commute with successor. Hence $s(m) + n = s(m + n)$.

Fix m. The functions $n \mapsto m+n$ and $n \mapsto n+m$ start at m. The first commutes with successor by (6.1), and the second does by the preceding identity. Uniqueness gives commutativity. Similarly, for fixed m, n the maps $k \mapsto (m + n) + k$ and $k \mapsto m + (n + k)$ start at $m + n$ and commute with successor, giving associativity.

Successor on one-letter words is injective, since deleting their last letter recovers the input word. If $m + k = m' + k$, then $s^k(m) = s^k(m')$, so injectivity of the k-fold composite gives $m = m'$. Commutativity gives cancellation on the other side. □

6.5 Multiplication and distributivity

Use parameter m with starting value 0 and step $x \mapsto x + m$. The resulting multiplication satisfies

$$m \cdot 0 = 0, \qquad m \cdot s(n) = m \cdot n + m. \tag{6.2}$$

For $m = 3$, the successive outputs are $0, 3, 6, 9, 12, \ldots$. The law $m \cdot (n + k) = m \cdot n + m \cdot k$ follows by induction on k. At $k = 0$, both sides reduce to $m \cdot n$. For the successor step, use $n + s(k) = s(n + k)$, then

$$\begin{aligned} m \cdot (n + s(k)) &= m \cdot (n + k) + m \\ &= (m \cdot n + m \cdot k) + m \\ &= m \cdot n + m \cdot s(k). \end{aligned}$$

To establish commutativity first note $0 \cdot n = 0$ by induction on n. Establish $s(m) \cdot n = m \cdot n + n$ by the same induction. The base case uses $m \cdot 0 = 0$. For the step, the left-hand side is $s(m) \cdot n + s(m)$, which by the induction hypothesis and $s(m) = m + 1$ is $(m \cdot n + n) + (m + 1)$. Reordering the summands gives $(m \cdot n + m) + (n + 1)$, the value $m \cdot s(n) + s(n)$. Here 1 denotes $s(0)$, and $n + 1 = s(n)$ follows from the addition equations. For fixed m, both $m \cdot n$ and $n \cdot m$ now start at zero and advance by adding m, so uniqueness gives commutativity. Commutativity gives the distributive law in the other variable.

With $1 = s(0)$ we have $m \cdot 1 = m$ by (6.2), and $1 \cdot m = m$ by commutativity. Finally, induction on k gives $a \cdot (b \cdot k) = (a \cdot b) \cdot k$. Both sides are zero at zero. At $s(k)$ the left-hand side is $a \cdot (b \cdot k + b) = a \cdot (b \cdot k) + a \cdot b$ by distributivity, which by induction is the right-hand side at $s(k)$. Thus the recursively defined operations give the usual commutative semiring structure on $\mathbb{N}$. A semiring has a commutative additive monoid, a multiplicative monoid, distributivity, and absorption by zero under multiplication.

6.6 The initial algebra for one more step

Suppose $\mathcal{C}$ also has the coproduct $1 \amalg X$ for every X. Write $H(X) = 1 \amalg X$. On maps $f \colon X \to Y$, use the identity on the first summand and f on the second to give $H(f)$. The coproduct property shows that this preserves identities and composition, since they agree on each of the two summands. An *H-algebra* is an object X with a structure map $a \colon H(X) \to X$. A map of algebras $f \colon (X, a) \to (Y, b)$ satisfies $f \circ a = b \circ H(f)$. By the coproduct property, specifying a amounts to specifying a point and a step. By restricting the equation for f to the two summands, one obtains precisely the two equations for maps of triples. So a natural numbers object is an initial H-algebra.

Proposition 6.4 (Lambek's lemma)**.** *For any endofunctor H, the structure map $a\colon H(N) \to N$ of an initial H-algebra is an isomorphism.*

Proof. The object $H(N)$ carries the structure $H(a)$. There is a unique algebra map $v\colon N \to H(N)$ by initiality. The map a is an algebra map in the reverse direction, since the required equation is $a \circ H(a) = a \circ H(a)$. Thus $a \circ v = \mathrm{id}_N$ by initiality. The equation for v gives $v \circ a = H(a) \circ H(v) = H(a \circ v) = \mathrm{id}_{H(N)}$. □

For $H(X) = 1 \amalg X$ this gives $1 \amalg N \cong N$, respecting the maps of the two summands into N. In **Set** it says that each element is either zero or the successor of exactly one element. Deleting the last letter of a nonempty word recovers its predecessor. In **FinSet** a natural numbers object would be finite and would have cardinality $1 + |N| = |N|$, which is impossible. Initiality specifies the object in its category. The category supplies the objects that can realize it.

The next chapter asks for a different universal map out of the additive monoid $\mathbb{N}$. Its targets have additive inverses, so sending n and m to their images determines the value of a formal difference. The recursion and cancellation laws just proved give its arithmetic interpretation.

CHAPTER 7

Integers from group completion

Natural numbers record how many objects are present. A change between two counts carries different information. Going from three objects to seven adds four, while going from seven to three removes four. In an additive group both changes are elements, and adding them gives zero. These two changes can be recorded as pairs $(7, 3)$ and $(3, 7)$ before subtraction is available. The aim is to turn such pairs into elements of a group by giving a precise rule for when they describe the same difference.

The rule comes from a universal property. Suppose a system of counts can already be mapped additively into some group. There the pair (a, b) must have the value of a minus the value of b. We will construct a single group through which every map of this kind factors. The construction uses the quotient method of Chapter 3, and the factorization property will give an adjunction of the kind developed in Part I.

7.1 Addition and formal differences

Definition 7.1. A *commutative monoid* is a set M with an associative and commutative operation $+$ and an element 0 satisfying $0 + a = a$ for every $a \in M$. It is an *abelian group* if every a has an additive inverse b with $a + b = 0$. A homomorphism of commutative monoids preserves addition and zero.

In an abelian group the inverse is unique. If b, c are both inverses of a, then $b = b + (a + c) = (b + a) + c = c$. We write $-a$ for this inverse and $a - b$ for $a + (-b)$. A monoid homomorphism between abelian groups preserves inverses because it sends the equation $a + (-a) = 0$ to the corresponding equation in the target.

The additive monoid $\mathbb{N}$ is our main example. Another is the set $\{0, 1\}$ with $x+y = \max(x, y)$, where 0 is still the identity. In the second example $1+0 = 1+1$. If we map this equation into a group and cancel the image of 1, equality of the images of 0 and 1 follows. Thus passing to a group can identify elements that were distinct in the monoid.

Definition 7.2. For a commutative monoid M, declare two pairs in $M \times M$ equivalent by the rule

$$(a, b) \sim (c, d) \iff \text{there exists } u \in M \text{ with } a + d + u = c + b + u.$$

The element u is a witness to the equality after adding a common term to both sides.

The equation comes from evaluating formal differences under an additive map $j\colon M \to H$ into an abelian group. If the two pairs have such a witness, then

$$j(a) + j(d) + j(u) = j(c) + j(b) + j(u).$$

Cancelling $j(u)$ and subtracting $j(b) + j(d)$ gives $j(a) - j(b) = j(c) - j(d)$. The common term also makes it possible to prove transitivity directly in M, where cancellation may fail.

Proposition 7.3. *The relation $\sim$ is an equivalence relation and is compatible with coordinatewise addition.*

Proof. For reflexivity use $u = 0$. For symmetry use the same witness and reverse the equality. For transitivity suppose

$$a + d + u = c + b + u, \qquad c + f + v = e + d + v.$$

Adding $f + v$ to the first equation and $b + u$ to the second gives, after regrouping,

$$a + f + (d + u + v) = e + b + (d + u + v).$$

Thus $d + u + v$ is a witness for $(a, b) \sim (e, f)$.

If $(a, b) \sim (c, d)$ by u and $(a', b') \sim (c', d')$ by v, add the two witnessing equations. Associativity and commutativity give

$$(a + a') + (d + d') + (u + v) = (c + c') + (b + b') + (u + v).$$

This is exactly the condition for the coordinatewise sums to be equivalent. □

7.2 The group and its universal map

Let $K(M) = (M \times M)/\!\sim$, and write $[a, b]$ for the class of (a, b). Proposition 7.3 lets us set

$$[a, b] + [c, d] = [a + c, b + d].$$

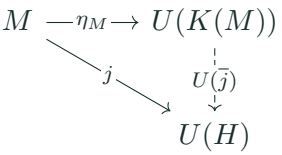
The operation is associative and commutative because coordinatewise addition is, and its zero is $[0, 0]$. Each class has inverse $[b, a]$, since $[a + b, b + a] = [0, 0]$ with witness 0. It follows that $K(M)$ is an abelian group, called the *group completion* or *Grothendieck group* of M. This construction also appears as the starting point of algebraic K-theory [6].

Write $U(H)$ for the underlying monoid of an abelian group H. The map $\eta_M \colon M \to U(K(M))$ sending a to $[a, 0]$ preserves addition and zero. Every element of $K(M)$ has the form

$$[a, b] = \eta_M(a) - \eta_M(b).$$

This equation both determines the value of a group homomorphism on a class and supplies the candidate for constructing that homomorphism.

Theorem 7.4 (Universal property of group completion)**.** *Let H be an abelian group. Every monoid homomorphism $j \colon M \to U(H)$ extends uniquely to a group homomorphism $\overline{j} \colon K(M) \to H$ with $U(\overline{j}) \circ \eta_M = j$.*



Proof. Set $\overline{j}([a, b]) = j(a) - j(b)$. The calculation in Section 7.1 proves that equivalent pairs have equal images. In the abelian group H the sum of the images of $[a, b]$ and $[c, d]$ is $j(a + c) - j(b + d)$, so $\overline{j}$ preserves addition. It preserves zero and has $\overline{j}([a, 0]) = j(a)$. The expression of every class as $\eta_M(a) - \eta_M(b)$ forces the same value for every extension. □

The equivalence classes now have a characterization in terms of all possible target groups. If every additive map into a group gives two pairs the same difference, apply the condition to η_M. The pairs then give the same element of $K(M)$. Conversely the universal property shows that every additive map gives equal values to pairs in the same class.

The universal property also controls any alternate construction. Suppose a group G and a map $i \colon M \to U(G)$ have the property of Theorem 7.4. The property of the pair $(K(M), \eta_M)$ gives a homomorphism $K(M) \to G$ extending i. The property of (G, i) gives a homomorphism $G \to K(M)$ extending η_M. Each composite preserves the original map out of M, so uniqueness makes each composite the identity. This gives a uniquely specified isomorphism under M, meaning an isomorphism that intertwines the two maps from M.

7.3 Naturality and cancellation

A homomorphism $k \colon M \to N$ of commutative monoids induces

$$K(k) \colon K(M) \longrightarrow K(N), \qquad [a, b] \longmapsto [k(a), k(b)].$$

Applying k to an equality witnessed by u produces an equality witnessed by $k(u)$. Thus the displayed formula respects the classes, and preservation of addition and zero follows coordinatewise. Applying two homomorphisms in succession gives the same result as applying their composite. Identity maps act by identities. Hence K is a functor, and η is natural by the following square.

$$\begin{array}{ccc} M & \xrightarrow{\eta_M} & U(K(M)) \\ \downarrow{\scriptstyle k} & & \downarrow{\scriptstyle U(K(k))} \\ N & \xrightarrow{\eta_N} & U(K(N)) \end{array}$$

Write **CMon** and **Ab** for the categories of commutative monoids and abelian groups with their homomorphisms. Theorem 7.4 gives a bijection

$$\mathrm{Hom}_{\mathbf{Ab}}(K(M), H) \;\cong\; \mathrm{Hom}_{\mathbf{CMon}}(M, U(H)).$$

Its map from left to right is restriction along η_M. If we change the source by $k\colon M' \to M$, extending $j \circ k$ gives $\overline{j} \circ K(k)$, as either map sends $[a, b]$ to $j(k(a)) - j(k(b))$. If we change the target by a group homomorphism $v\colon H \to H'$, extending $U(v) \circ j$ gives $v \circ \overline{j}$, since v preserves subtraction. The bijection is therefore natural in both variables, and $K \dashv U$.

A monoid is *cancellative* when $a + u = c + u$ implies $a = c$ for all a, c, u. The map η_M is injective exactly in this case. Indeed $\eta_M(a) = \eta_M(c)$ means that the pairs $(a, 0)$ and $(c, 0)$ are equivalent, so it means there is a witness u with $a + u = c + u$. If the monoid is cancellative, the general equivalence simplifies as well. It becomes $(a, b) \sim (c, d)$ precisely when $a + d = c + b$, because the witness can be cancelled in one direction and chosen as 0 in the other.

The completion can therefore be understood in two steps. First identify elements a, c precisely when a common term makes them equal. This is an equivalence relation compatible with addition, since it is equality of the images under η_M. Its quotient is a cancellative monoid. Indeed, equality of the classes of $a + u$ and $c + u$ gives a witness v with $a + u + v = c + u + v$. The common term $u + v$ then identifies a, c in the quotient. The quotient embeds into $K(M)$ by sending the class of a to $\eta_M(a)$, and every element of $K(M)$ is a difference of two elements of this image. This factors η_M into a quotient followed by an inclusion of a cancellative monoid into an additive group.

Example 7.5. An abelian group H is already complete for this construction. Applying Theorem 7.4 to $\mathrm{id}_{U(H)}$ gives a homomorphism $K(U(H)) \to H$ sending $[a, b]$ to $a - b$. The map $h \mapsto [h, 0]$ is a group homomorphism by the monoid homomorphism property. One composite sends h to $h - 0 = h$. The other sends $[a, b]$ to $[a - b, 0] = [a, b]$, as the equality $(a - b) + b = a$ witnesses. Thus a

second application of completion recovers the same group up to its specified isomorphism.

7.4 Normal forms for the integers

Proposition 6.3 shows that addition on $\mathbb{N}$ is cancellative. The injectivity criterion for group completion therefore embeds $\mathbb{N}$ into $K(\mathbb{N})$ by $a \mapsto [a, 0]$.

Proposition 7.6. *Every element of $K(\mathbb{N})$ has exactly one form $[n, 0]$ for $n \in \mathbb{N}$ or $[0, m]$ for $m \in \mathbb{N}$ with $m \neq 0$.*

Proof. For any $a, b \in \mathbb{N}$, there is $d \in \mathbb{N}$ with $a = b + d$ or $b = a + d$. To see this, induct on b with a arbitrary. When $b = 0$ take $d = a$. For $b = s(b')$, the case $a = 0$ gives $b = a + b$. Otherwise $a = s(a')$, and we compare a' with b' using the inductive hypothesis. Applying s to either resulting equality gives the desired equality for a, b, since $s(x + d) = s(x) + d$ by commutativity and the recursion equations for addition.

If $a = b + d$, then the cancellation criterion gives $[a, b] = [d, 0]$. If $b = a + d$, it gives $[a, b] = [0, d]$, which is the negative form when $d \neq 0$ and gives $[0, 0]$ when $d = 0$. Two forms of the same kind are equal only when their displayed natural numbers agree, by the same criterion. Equality between $[n, 0]$ and $[0, m]$ requires $n + m = 0$. When $m \neq 0$, the number m is a successor and $n + m$ is therefore a successor. This rules out that equality. □

We write n for $[n, 0]$ and $-m$ for $[0, m]$ when $m \neq 0$. The resulting abelian group is $\mathbb{Z}$. Addition of two forms of the first kind adds their natural numbers. Addition of two negative forms adds their magnitudes and keeps the negative sign. For mixed signs, $[a, 0] + [0, b] = [a, b]$, which is reduced to its unique form by comparing a and b. For example,

$$[7, 3] + [2, 8] = [9, 11] = [0, 2].$$

This gives $4 + (-6) = -2$ directly from the quotient arithmetic. The designation of a form as positive, zero, or negative follows from the two unique forms and the distinguished zero.

The abelian group $\mathbb{Z}$ is free on one generator. For an abelian group H, choose an arbitrary element h. Recursion gives a function $j_h \colon \mathbb{N} \to H$ with $j_h(0) = 0$ and $j_h(s(n)) = j_h(n) + h$. To see that it preserves addition, fix m and induct on n in $j_h(m + n) = j_h(m) + j_h(n)$. At 0 both sides are $j_h(m)$. At $s(n)$, rewrite $m + s(n)$ as $s(m + n)$ and add h to both sides of the equality for n. Thus j_h is an additive map taking 1 to h. Since $s(n) = n + 1$, every additive map with that value at 1 obeys the same recursion and equals j_h. The universal property of

group completion gives exactly one group homomorphism $\mathbb{Z} \to H$ taking 1 to h. It sends the formal difference $[a, b]$ to $j_h(a) - j_h(b)$.

Finite counting gives a concrete instance of the same construction. Choose the set $[n] = \{0, \ldots, n-1\}$ as the representative of size n, including the empty set at $n = 0$. Taking the size of a disjoint union as in Chapter 2 identifies the monoid of sizes with $(\mathbb{N}, +, 0)$. A pair of finite sets (A, B) then specifies $[|A|, |B|]$ in the group completion. Adding the same finite set to both entries leaves this class unchanged. The resulting signed count is universal for turning addition of disjoint sizes into an operation with inverses.

7.5 Multiplying formal differences

Multiplication on $\mathbb{N}$ preserves addition and zero in each variable. We use group completion twice to extend it to formal differences. Write $\iota(n) = [n, 0]$. For each $a \in \mathbb{N}$, the additive map $b \mapsto \iota(ab)$ extends uniquely to a group homomorphism $L_a \colon \mathbb{Z} \to \mathbb{Z}$. For $a, c \in \mathbb{N}$, the functions L_{a+c} and $L_a + L_c$ are homomorphisms that agree on $\iota(\mathbb{N})$, by distributivity on the natural numbers. They are equal by uniqueness. The same reasoning makes L_0 the zero homomorphism. Thus for each $y \in \mathbb{Z}$, the rule $a \mapsto L_a(y)$ is additive and extends to a group homomorphism in the first variable. Call its value at x the product xy.

For $x = [a, b]$, the product is $L_a(y) - L_b(y)$, so it is additive in y as well. Evaluating this on $y = [c, d]$ yields

$$[a, b]\,[c, d] = [ac + bd,\ ad + bc].$$

For example, $[0, 5][0, 2] = [10, 0]$, giving $(-5)(-2) = 10$. The formula defines an operation on classes because it was obtained by extending maps on the completed groups. It is the unique operation additive in both variables with $\iota(a)\iota(b) = \iota(ab)$, since these values determine the product on differences.

Commutativity follows because $(x, y) \mapsto xy$ and $(x, y) \mapsto yx$ are additive separately and agree when both inputs lie in $\iota(\mathbb{N})$. Likewise, $(xy)z$ and $x(yz)$ are additive separately in three variables and agree on triples of natural numbers. Extending equality to differences one variable at a time proves associativity. Multiplication by $\iota(1)$ on each side agrees with the identity on $\iota(\mathbb{N})$ and hence on all of $\mathbb{Z}$. Separate additivity gives distributivity. These laws make $\mathbb{Z}$ a commutative ring with unit.

7.6 A completion that forgets a flag

Consider $M = \mathbb{N} \times \{0, 1\}$ with the operation

$$(n, \epsilon) + (m, \delta) = (n + m, \max(\epsilon, \delta)).$$

The second coordinate records whether a flag is set. Adding $(0,1)$ to either $(n,0)$ or $(n,1)$ gives $(n,1)$, so their images in the completion agree. The first coordinate still distinguishes n from m when $n \neq m$, since addition in it is cancellative.

More generally, the projection $\pi\colon M \to \mathbb{N}$ induces

$$\kappa\colon K(M) \longrightarrow \mathbb{Z}, \qquad [(n,\epsilon),(m,\delta)] \longmapsto [n,m].$$

It is surjective by taking both flags to be 0. If the images of $[(n,\epsilon),(m,\delta)]$ and $[(n',\epsilon'),(m',\delta')]$ are equal, then $n+m' = n'+m$. The witness $(0,1)$ realizes the corresponding equation in M because both second coordinates become 1. Thus the input classes are equal and κ is an isomorphism.

$$\begin{array}{ccc} M & \xrightarrow{\eta_M} & U(K(M)) \\ \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle U(\kappa)} \\ \mathbb{N} & \xrightarrow{\eta_{\mathbb{N}}} & U(\mathbb{Z}) \end{array}$$

The square records exactly what survives in this example. Under any map into an abelian group, the value of (n,ϵ) depends only on n and is determined by the value of $(1,0)$. Chapter 8 turns to groups presented by operations and relations, with commutativity among the relations one can impose.

CHAPTER 8

Groups from operations and equations

A word in an alphabet can be interpreted in any monoid by multiplying the values of its letters. In a group, a generator can also occur with an inverse. An expression can involve multiplication, inversion, and the unit, with parentheses recording how it was built. The group equations tell us which expressions become equal. Taking exactly the consequences of these equations gives the free group. Adding equations gives presentations such as those for the symmetries of an equilateral triangle.

8.1 The operations of a group

Definition 8.1. A *group* is a monoid G in which every x has a two-sided inverse x^{-1}. Its three operations are multiplication $m\colon G \times G \to G$, inversion $i\colon G \to G$, and the unit $e\colon 1 \to G$, where $1 = \{*\}$ and $e(*) = 1_G$. A group homomorphism preserves these operations. Groups and their homomorphisms form the category **Grp**.

In multiplicative notation the axioms are the equations

$$(xy)z = x(yz), \qquad 1_G x = x = x 1_G, \qquad x x^{-1} = 1_G = x^{-1} x.$$

Each equation is required for all values of its variables. The first inverse law, for instance, says that multiplication of the pair $(x, i(x))$ agrees with the constant function with value 1_G.

$$\begin{array}{ccc} G & \xrightarrow{\langle \mathrm{id}_G, i \rangle} & G \times G \\ \downarrow{\scriptstyle !} & & \downarrow{\scriptstyle m} \\ 1 & \xrightarrow{\quad e \quad} & G \end{array}$$

Here $!\colon G \to 1$ is the unique function to the singleton. Swapping the entries of the pair gives the second inverse law. These equations determine inversion uniquely. If $xy = 1_G$ and $zx = 1_G$, then $y = (zx)y = z(xy) = z$. In particular the inverse of xy is $y^{-1}x^{-1}$, since multiplication in either order gives the unit. A map preserving multiplication and unit preserves inverse as well, because the image of an inverse satisfies both inverse equations for the image of the element.

For an integer k, write x^k for a product of k copies of x when $k \geq 0$, with $x^0 = 1_G$, and for $(x^{-1})^{-k}$ when $k < 0$. These definitions give $x^{k+1} = x^k x$ for every integer k, using cancellation when $k < 0$. Both functions $n \mapsto x^m x^n$ and $n \mapsto x^{m+n}$ have value x^m at $n = 0$ and obey that recurrence. The recurrence can be run backwards by multiplying by x^{-1}, so the functions agree for all integers. Thus $x^m x^n = x^{m+n}$ and $(x^m)^{-1} = x^{-m}$.

8.2 The free group as a quotient of terms

Fix a set A. The set $\mathsf{T}(A)$ of *terms* consists of finite expressions built using four constructors. There is a generator term for each $a \in A$, a constant term $\mathbf{e}$, a term $i(t)$ for each term t, and a term $m(s,t)$ for each pair of terms s, t. At this stage the constructors record the shape of the expression. For example the terms $m(m(a,b),c)$ and $m(a,m(b,c))$ have different shapes. A recursive definition on these shapes can specify a value at the generator and constant terms and then use the values of the immediate subterms at the two other constructors. Every term has finite depth and a unique outer constructor, so this gives exactly one value per term.

An equivalence relation $\sim$ on an algebra with these three operations is a *congruence* if $s \sim t$ implies $i(s) \sim i(t)$ and $s \sim s'$, $t \sim t'$ imply $m(s,t) \sim m(s',t')$. Its classes carry well-defined operations, for instance $m([s],[t]) = [m(s,t)]$. On $\mathsf{T}(A)$, require a congruence to contain the two sides of every group axiom after substituting terms for its variables. The relation equating all terms is one such congruence. Their intersection is an equivalence relation, since reflexivity, symmetry, and transitivity hold in each one. Compatibility with the operations and containment of the required pairs also hold in the intersection. Thus the intersection $\sim_{\mathrm{grp}}$ is the least congruence imposing the group equations.

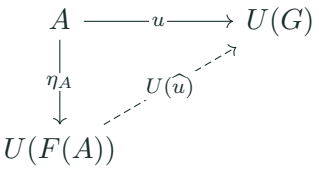
Put $F(A) = \mathsf{T}(A)/{\sim_{\mathrm{grp}}}$. The induced operations satisfy the group equations on all classes, since representatives of any tuple of classes give a term substitution of the required kind. There is a function $\eta_A \colon A \to U(F(A))$ sending a generator to its class, where $U \colon \mathbf{Grp} \to \mathbf{Set}$ takes underlying sets.

Proposition 8.2 (Free group on a set)**.** *Given a function $u \colon A \to U(G)$, there is a unique homomorphism $\widehat{u} \colon F(A) \to G$ with $U(\widehat{u}) \circ \eta_A = u$.*

Proof. Evaluate each term in G recursively. Use $u(a)$ at a generator, 1_G at the constant, inverse at $i(t)$, and multiplication at $m(s,t)$. Equality of evaluations is a congruence because multiplication and inversion are functions. It contains the substituted group equations because G satisfies the axioms. Hence the evaluation descends to $F(A)$, and the rules at the constructors make the induced map a homomorphism. Conversely, any homomorphism agreeing with u on generators

must evaluate the constant and each successive constructor by the same rules. Induction on the depth of terms proves uniqueness. □

The factorization in the following diagram takes place in **Set**. The dashed map comes from the unique group homomorphism.



Thus $F(A)$ represents the functor $G \mapsto \mathbf{Set}(A, U(G))$, in the sense of Chapter 4. A function $v\colon A \to B$ extends to $F(v)\colon F(A) \to F(B)$ by the assignment $\eta_B \circ v$. Uniqueness gives the identity and composition laws for F. Postcomposing $\widehat{u}$ with a homomorphism changes its assignment by postcomposition. Precomposing with $F(v)$ changes the assignment by precomposition with v when the assignment has domain B. These two facts give naturality of the representing bijections and the adjunction $F \dashv U$.

Equality in the free group has an equivalent test in terms of evaluation. If two terms have equal classes, every assignment into every group evaluates them equally by the proposition. Conversely, if all assignments evaluate them equally, take $G = F(A)$ and use the assignment η_A. Evaluation with this assignment returns the class of each term, so their classes agree. For example two distinct generator symbols give distinct classes, since assigning them $0, 1$ in the additive group $\mathbb{Z}$ distinguishes their evaluations.

8.3 Group presentations

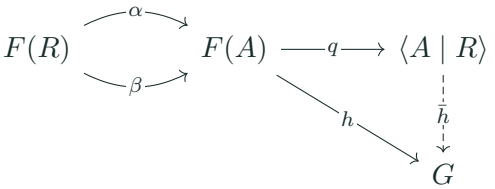
Let R be a set of prescribed equations $s = t$ between terms in $\mathsf{T}(A)$. Each side first has a value in the free group $F(A)$. Quotient $F(A)$ by the least congruence that identifies the two values for each equation. This congruence exists by the intersection argument. The quotient group is denoted $\langle A \mid R \rangle$ and is called the *group presented by A and R*. Every element is the value of a term in the classes of the generators.

Proposition 8.3 (Presentations represent constrained assignments)**.** *Homomorphisms $\langle A \mid R \rangle \to G$ correspond to functions $A \to U(G)$ under which the two sides of every relation in R evaluate to the same element. This correspondence is natural in G.*

Proof. Given an assignment, first extend it to $h\colon F(A) \to G$ by Proposition 8.2. Its relation of equality under h is a congruence. It contains all prescribed pairs

precisely when the assignment satisfies the equations. In this case it contains the least such congruence, so h descends uniquely to a homomorphism on the classes. Any homomorphism on the quotient sends equal classes to the same element and therefore gives a satisfying assignment. The two operations are inverse by uniqueness of the extension, and postcomposition with a homomorphism commutes with evaluation. □

Each equation supplies a pair of values in $F(A)$, which define two assignments from the set R to $U(F(A))$. Their extensions are homomorphisms $\alpha, \beta \colon F(R) \to F(A)$. A homomorphism out of $F(A)$ coequalizes α, β exactly when it identifies the two values of each equation, since R freely generates $F(R)$. Consequently the quotient arrow q is a coequalizer in **Grp**, with the same factorization pattern as in Chapter 3.



There is also a description in terms of normal subgroups. A subgroup N of G contains 1_G and is closed under multiplication and inversion. It is normal if $xnx^{-1} \in N$ for all $n \in N$ and $x \in G$. For a group congruence, the class N of 1_G is a subgroup by compatibility with multiplication and inverse. It is normal because $n \sim 1_G$ implies $xnx^{-1} \sim x1_Gx^{-1} = 1_G$. Multiplication by y^{-1} gives $x \sim y$ precisely when $xy^{-1} \in N$. Conversely, this condition defines a congruence for any normal subgroup. Reflexivity follows from $xx^{-1} = 1_G$, symmetry from closure under inversion, and transitivity from $(xy^{-1})(yz^{-1}) = xz^{-1}$. For compatibility with multiplication, if $xx'^{-1}, yy'^{-1} \in N$, then

$$(xy)(x'y')^{-1} = (x(yy'^{-1})x^{-1})(xx'^{-1}) \in N.$$

For compatibility with inversion, $xy^{-1} \in N$ implies $x^{-1}y = x^{-1}(xy^{-1})^{-1}x \in N$. In a presentation the least congruence therefore corresponds to the smallest normal subgroup containing the values of st^{-1} for $s = t$ in R. This describes the usual quotient by relators [46, § 1.2].

8.4 Cyclic and triangle symmetries

Example 8.4 (One generator of order n)**.** For $n \geq 1$, let $C_n = \langle r \mid r^n = 1 \rangle$, with r also denoting its class. The powers of r contain the unit and are closed under product and inverse by the power laws. Since every element is a term in r, these

powers exhaust C_n. The relation $r^n = 1$ gives $r^{k+n} = r^k r^n = r^k$ for every integer k. Taking remainders modulo n reduces every power to one of $r^0, \dots, r^{n-1}$.

To separate the n candidates, use the additive group $\mathbb{Z}/n\mathbb{Z}$ of residues modulo n. Addition and negation descend to residues since differences divisible by n remain so under addition and negation. The residue $\bar{1}$ satisfies $n\bar{1} = \bar{0}$, so the presentation gives a homomorphism $C_n \to \mathbb{Z}/n\mathbb{Z}$ sending r to $\bar{1}$. Its values on the candidates are the distinct residues $\bar{0}, \dots, \overline{n-1}$. The map is therefore an isomorphism. Proposition 8.3 also gives

$$\mathrm{Hom}_{\mathbf{Grp}}(C_n, G) \cong \{g \in G \mid g^n = 1_G\}.$$

For instance $C_2 \to C_6$ has two possible homomorphisms, corresponding to residues 0 and 3, since $2k \equiv 0 \pmod 6$ exactly when 3 divides k.

Example 8.5 (A presentation of the symmetric group on three letters)**.** Let S_3 be the group of bijections of $\{0, 1, 2\}$, with the rightmost permutation applied first. In cycle notation, take $R = (0\,1\,2)$ and $S = (1\,2)$. Thus R rotates the three labels and S exchanges 1 and 2 while fixing 0. Applying SRS to $0, 1, 2$ gives $2, 0, 1$, respectively, so $SRS = R^{-1}$. Also $R^3 = S^2 = 1$. We obtain a homomorphism $\pi\colon P \to S_3$ sending r to R and s to S, where

$$P = \langle r, s \mid r^3 = 1,\ s^2 = 1,\ srs = r^{-1} \rangle.$$

In P, conjugation by s has the form $x \mapsto sxs$, since $s^{-1} = s$. It preserves products by the identity $(sxs)(sys) = s(xy)s$, and it preserves inverses as a homomorphism. Hence $sr^k s = r^{-k}$ for all integers k. It follows that pairs of the form $r^i s^j$ multiply by the rule

$$(r^i s^j)(r^k s^\ell) = r^{i+(-1)^j k} s^{j+\ell}.$$

The exponents reduce modulo 3 and 2 respectively. The inverse of $r^i s^j$ is $s^{-j} r^{-i} = r^{-(-1)^j i} s^j$, with the same reductions. The forms $r^i s^j$ with $0 \le i < 3$ and $0 \le j < 2$ thus contain the unit and both generators and are closed under all operations. Every element of P has one of these forms.

The images $R^i S^j$ send 0 to i, which separates forms with different i. For the same i, left composition with R^{-i} reduces comparison to S^j, separating the two choices of j. Thus the six forms are distinct. There are $3 \cdot 2 \cdot 1 = 6$ bijections of the three labels, so π is an isomorphism. For a sample product, the rule gives $(r^2 s)(rs) = r$, while $(rs)(r^2 s) = r^2$. The presentation makes both computations possible using just the defining relations.

8.5 Groups as product-preserving models

A group specifies operations on one set and equations between composites of those operations. The same description can be made categorical by collecting

the formal composites into a category with finite products. This is the functorial semantics of algebraic theories [4].

Define $\mathcal{T}_{\mathrm{grp}}$ to have objects $n \geq 0$, where n stands for a list of variables $x_1, \ldots, x_n$. An arrow $n \to m$ is an m-tuple of terms in those n variables modulo the group equations. More precisely, its entries are elements of $F(\{x_1, \ldots, x_n\})$. The identity at n is the tuple of its variables. To compose $t\colon n \to m$ and $u\colon m \to k$, substitute the m entries of t into every entry of u. Substitution is well-defined on classes because a tuple t extends to a homomorphism out of the free group on the m variables. It takes each variable to the specified entry and preserves all the operations. Substitution by the variables gives the identity homomorphism. Successive substitutions and substitution of the composed tuple agree on each variable, so the free-group property gives associativity.

The product of m and n is $m+n$. The projections take the first m and the last n variables respectively. Splitting an $(m+n)$-tuple of terms along this boundary gives precisely an m-tuple and an n-tuple in the same variables, which is the product universal property. The empty tuple gives the unique arrow to 0. In particular 0 is terminal and each n is a power of the one-variable object 1.

A group G gives a functor $n \mapsto U(G)^n$ by evaluating each entry of a tuple on a list of elements. Evaluation respects substitution, as follows by induction on the term being substituted into. The functor preserves finite products because tuple splitting gives the canonical bijection $U(G)^{m+n} \cong U(G)^m \times U(G)^n$, and $U(G)^0$ is a singleton.

Conversely, let $M\colon \mathcal{T}_{\mathrm{grp}} \to \mathbf{Set}$ preserve finite products. Put $G = M(1)$. Product preservation identifies $M(n)$ canonically with G^n by the images of the variable projections. The constant term $0 \to 1$, inverse term $1 \to 1$, and product term $2 \to 1$ give the unit, inversion, and multiplication on G. The group equations hold because their sides are equal arrows in $\mathcal{T}_{\mathrm{grp}}$. Every term is assembled from the projections and these three operations by substitution and tuple formation. Thus M evaluates each term using the operations just recovered, and the canonical identifications $M(n) \cong G^n$ are natural in n.

For example, the arrow $2 \to 1$ with term $m(i(x_2), x_1)$ evaluates as $(g_1, g_2) \mapsto g_2^{-1} g_1$. Pairing this arrow with the second projection yields the arrow $2 \to 2$ with entries $(x_2, m(i(x_2), x_1))$. Composing that pair with the multiplication arrow gives the first projection, because the group equations give $x_2(x_2^{-1}x_1) = (x_2x_2^{-1})x_1 = x_1$. Every product-preserving functor carries this equality of arrows to the corresponding equality of functions on pairs.

Natural transformations between the resulting models recover group homomorphisms. Naturality at the projections determines each component at n by applying the component at 1 to all n coordinates. Naturality at the three operation arrows says that this component preserves the operations. Conversely, a

homomorphism on the underlying groups commutes coordinatewise with evaluation of every term, by induction on terms. Groups and homomorphisms are therefore equivalent to finite-product-preserving functors $\mathcal{T}_{\mathrm{grp}} \to \mathbf{Set}$ and their natural transformations.

The functor description can also start from a fixed group G. Make a category BG with one object and with elements of G as arrows, composing by multiplication. A functor $BG \to \mathbf{Set}$ gives a set X and a bijection of X for every $g \in G$, because it sends inverse arrows to inverse functions. Its composition law says that these bijections give a group homomorphism to the group of permutations of X. For $G = S_3$, the presentation above gives such functors by specifying the permutations assigned to r, s and checking their three relations.

8.6 Finite characters and Pontryagin duality

Use $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ as in Section 5.6. Giving a character of $C_n = \mathbb{Z}/n\mathbb{Z}$ is equivalent to choosing the value t of its generator subject to $nt = 0$. The n choices are $k/n + \mathbb{Z}$, for $k \in C_n$. The resulting characters $\chi_k(x) = kx/n + \mathbb{Z}$ satisfy $\chi_k + \chi_\ell = \chi_{k+\ell}$. Evaluation gives the pairing

$$C_n \times \widehat{C}_n \longrightarrow \mathbb{T}, \qquad (x, \chi_k) \longmapsto kx/n + \mathbb{Z}.$$

In particular, $\chi_1(x) = 0$ only for $x = 0$. The evaluation map to the double dual is therefore injective and is bijective since both groups have order n.

For finite abelian G, H, restricting a character of $G \times H$ to the two factors gives the characters χ, ψ, and its value at (g, h) is $\chi(g) + \psi(h)$. Thus the dual of a finite product is the product of the duals. The decomposition of a finite abelian group as a product of cyclic groups [7, § 13.1] now gives $|\widehat{G}| = |G|$ and enough characters to distinguish all its elements. It follows that

$$G \xrightarrow{\eta_G} \widehat{\widehat{G}}$$

is an isomorphism for every finite abelian G. Naturality follows from evaluation: if $f\colon G \to H$ and $\chi \in \widehat{H}$, then $\chi(f(g)) = (\chi \circ f)(g)$. This gives a dual equivalence on finite abelian groups.

For calculations with complex values, identify the circle with the unit complex numbers by $t \mapsto e^{2\pi i t}$. In this notation the cyclic characters are $x \mapsto e^{2\pi i k x/n}$. For any nontrivial complex character χ of a finite abelian group, choose a with $\chi(a) \neq 1$. Translation by a permutes the group, so $\sum_x \chi(x) = \chi(a) \sum_x \chi(x)$, and that sum is zero. Applying this to $\chi\overline{\psi}$ gives

$$\frac{1}{|G|} \sum_{x \in G} \chi(x)\overline{\psi(x)} = \begin{cases} 1 & \chi = \psi, \\ 0 & \chi \neq \psi. \end{cases}$$

There are $|G|$ orthogonal characters in the $|G|$-dimensional space of complex functions on G, hence they form a basis. This counting argument supplies the Fourier coordinates used in Sections 20.6, 31.7, and 38.7.

CHAPTER 9

Symmetry from group actions

Paint each position of a square with one of two colors. A quarter turn moves the colors to new positions, and four successive quarter turns restore the original assignment. Some assignments return after a single turn or a half turn. The same group of rotations acts on every assignment, but the rotations that preserve a given assignment depend on its pattern. Counting the distinct patterns up to rotation requires both pieces of information.

A group action specifies how each symmetry moves each member of a set. It gives a category whose arrows record the individual symmetries and the members they move. Counting the arrows that return to their sources in two ways produces Burnside's lemma.

9.1 Actions as functors

A group G defines a category BG with one object $\star$. An arrow $\star \to \star$ is an element of G. The identity arrow is the group identity e, and composing the arrow g followed by the arrow h gives hg. The group axioms give the category axioms, and g^{-1} supplies the inverse arrow of g.

Definition 9.1. A *left action* of G on a set X is a function $G \times X \to X$, written $(g, x) \mapsto gx$, satisfying $ex = x$ and $h(gx) = (hg)x$ for all $g, h \in G$ and $x \in X$. A function $f\colon X \to Y$ between sets with left G-actions is *equivariant* if $f(gx) = gf(x)$ for all g, x.

The action is precisely a functor $F\colon BG \to \mathbf{Set}$. Its value $F(\star)$ is X, and $F(g)\colon X \to X$ sends x to gx. The two action equations amount to preservation of identity and composition. In particular, each $F(g)$ is a permutation because $F(g^{-1})$ is its inverse. Conversely, any such functor gives an action by evaluating $F(g)$ at each x. These two constructions recover one another.

For functors $F, F'\colon BG \to \mathbf{Set}$ with object values X, Y, a natural transformation has a single component $f\colon X \to Y$. Its naturality square for each g is

$$\begin{array}{ccc} X & \xrightarrow{F(g)} & X \\ \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle f} \\ Y & \xrightarrow{F'(g)} & Y. \end{array}$$

This square commutes exactly when applying f after the action of g gives the same result as applying the action of g after f. Thus the objects and arrows of the functor category $\mathbf{Set}^{BG}$ are the actions and the equivariant maps. Composition of equivariant maps is ordinary function composition.

For example, the group S_3 of Chapter 8 acts on $\{1,2,3\}$ by permutations, with $(\sigma\tau)i = \sigma(\tau(i))$. It also acts on the power set of $\{1,2,3\}$ by taking direct images. The map $i \mapsto \{i\}$ is equivariant because the image of a singleton under a permutation is the singleton of the image. The cardinality map to $\{0,1,2,3\}$ is equivariant when the target has the trivial action, in which every group element acts by the identity. The square expresses the invariance of cardinality under relabeling.

A homomorphism $\varphi\colon K \to G$ induces a functor $B\varphi\colon BK \to BG$ by applying φ to the arrows. The composite $F \circ B\varphi$ gives a K-action on X in which k acts by $\varphi(k)$. When K is a subgroup, this is restriction to the symmetries in K. For example, the quarter turn acts on the four positions of a square by $0 \mapsto 1 \mapsto 2 \mapsto 3 \mapsto 0$. Restricting this action to the two-element subgroup generated by the half turn leaves the two independent pairs $\{0,2\}$ and $\{1,3\}$. Precomposition also takes each equivariant map for G to an equivariant map for K, since its naturality square already holds for every $\varphi(k)$.

9.2 Orbits, stabilizers, and cosets

Definition 9.2. For $x \in X$, the *orbit* of x is $Gx = \{gx \mid g \in G\}$. The *stabilizer* of x is $G_x = \{g \in G \mid gx = x\}$.

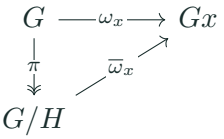
Belonging to the same orbit is an equivalence relation. The identity gives reflexivity, the inverse of a symmetry gives symmetry of the relation, and multiplication gives transitivity. The orbits therefore partition X. The stabilizer G_x is a subgroup of G. It contains e, the composite of two symmetries fixing x fixes x, and applying g^{-1} to $gx = x$ shows that g^{-1} fixes x whenever g does.

Proposition 9.3 (Orbit–stabilizer)**.** *For a left action and $x \in X$, the orbit map $g \mapsto gx$ induces an equivariant bijection from the set of left cosets G/G_x to Gx. If G is finite, then*

$$|Gx|\,|G_x| = |G|.$$

Proof. Put $H = G_x$. The value gx is constant on the left coset $gH = \{gh \mid h \in H\}$ because $(gh)x = g(hx) = gx$. Conversely, if $gx = g'x$, applying $(g')^{-1}$ gives $(g')^{-1}gx = x$, so $(g')^{-1}g \in H$ and $gH = g'H$. Therefore the fibers of $g \mapsto gx$ are exactly the left cosets of H. The orbit consists of the values attained by this

map, so the induced map $\overline{\omega}_x$ is bijective.



Left multiplication gives an action on the cosets by $k(gH) = (kg)H$. The definition of $\overline{\omega}_x$ gives $\overline{\omega}_x((kg)H) = (kg)x = k\overline{\omega}_x(gH)$, so it is equivariant. When G is finite, multiplication by g is a bijection $H \to gH$, with inverse multiplication by g^{-1}. Counting the disjoint cosets therefore gives $|G| = |G/H|\,|H| = |Gx|\,|G_x|$. □

In the permutation action of S_3 on $\{1, 2, 3\}$, the stabilizer of 1 is $H = \{e, (2\,3)\}$. A permutation fixing 1 is determined by its permutation of $2, 3$. The three cosets $H, (1\,2)H, (1\,3)H$ have orbit values $1, 2, 3$, respectively. Each value is reached by two permutations, so the orbit has size $6/2 = 3$. Stabilizers at different points of one orbit are related by conjugation. For $y = gx$, the equation $hy = y$ is equivalent to $(g^{-1}hg)x = x$, giving $G_y = gG_xg^{-1}$.

An action is *transitive* if the set is nonempty and has a single orbit. Every subgroup H gives a transitive action on G/H by left multiplication. Each coset gH is obtained by acting on H with g, and the stabilizer of H is H since $gH = H$ holds precisely for $g \in H$. Conversely, the orbit–stabilizer bijection shows that every transitive action with a chosen point x is isomorphic to G/G_x, with G_x as the chosen coset.

For actions with the point left unspecified, the isomorphism class corresponds to a conjugacy class of subgroups. Indeed, an equivariant bijection from G/H to G/K takes H to some gK and identifies their stabilizers. The stabilizer of gK is gKg^{-1}, so $H = gKg^{-1}$. Conversely, if $H = gKg^{-1}$, the transitive action on G/K with chosen point gK is isomorphic to G/H by the same orbit–stabilizer bijection.

9.3 The action groupoid

The *action groupoid* $G \ltimes X$ has objects $x \in X$. An arrow is a pair (g, x), with source $s(g, x) = x$ and target $t(g, x) = gx$. If the next arrow is (h, gx), the composite is (hg, x). The arrow (e, x) is the identity at x, and (g^{-1}, gx) inverts (g, x). Composition is associative by associativity of group multiplication. A category in which every arrow is invertible is called a *groupoid*.

The objects reachable by an arrow from x are exactly Gx. The arrows from x to itself, with their composition, form the group G_x by identifying (g, x) with g. In the example of S_3, both e and $(2\,3)$ give arrows from 1 to 1, and composition

retains their distinction. The orbit relation records reachability, while the action groupoid also records the symmetry that realizes an arrow.

Let $E = \{(g, x) \in G \times X \mid gx = x\}$ be the set of all loops in the action groupoid. It is a pullback along the diagonal $\Delta_X \colon X \to X \times X$, where $\Delta_X(x) = (x, x)$.

$$\begin{array}{ccc} E & \hookrightarrow & G \times X \\ \downarrow{\scriptstyle p_X} & \lrcorner & \downarrow{\scriptstyle (s,t)} \\ X & \xrightarrow{\Delta_X} & X \times X. \end{array}$$

The equality with a diagonal pair requires $s(g, x) = t(g, x) = x$, so the lower coordinate of the pullback is determined by (g, x). The top inclusion followed by projection to G gives a second map $p_G \colon E \to G$. Together these maps record the point and symmetry of a loop.

9.4 Fixed points as a limit and orbits as a colimit

The functor $F \colon BG \to \mathbf{Set}$ has one object in its diagram and an arrow $F(g)$ for every g. A cone with vertex T is specified by a map $a \colon T \to X$ satisfying $F(g) \circ a = a$ for every g. At each input of a this is the condition that the image is fixed by all of G. Write

$$X^G = \{x \in X \mid gx = x \text{ for every } g \in G\}.$$

Every such a therefore factors uniquely through the inclusion $X^G \hookrightarrow X$. The inclusion itself satisfies all the cone equations, so it exhibits the limit of F. Its elements are the compatible choices of a point in the diagram.

A cocone with vertex Y is specified by $b \colon X \to Y$ satisfying $b \circ F(g) = b$ for every g. This says $b(gx) = b(x)$, so b is constant on each orbit. Let X/G denote the set of orbits, with orbit map $q(x) = Gx$. The formula $\overline{b}(Gx) = b(x)$ then defines a function $X/G \to Y$. It is well defined by orbit constancy and unique because every orbit is $q(x)$ for some x.

$$\begin{array}{ccccc} G \times X & \overset{s}{\underset{t}{\rightrightarrows}} & X & \xrightarrow{q} & X/G \\ & & & \searrow^{b} & \downarrow{\scriptstyle \overline{b}} \\ & & & & Y. \end{array}$$

The condition $b \circ s = b \circ t$ is the cocone condition on b. Thus q is both the colimit map for F and the coequalizer of the source and target maps of its action groupoid. An orbit is the class of points identified by every invariant map. The fixed points, by the limit description, are the points whose choices already satisfy every compatibility equation.

The two test properties give adjunctions with the trivial-action functor. Let ΔY be the set Y on which every element of G acts as the identity. An equivariant map $X \to \Delta Y$ is exactly an invariant map, so its factor through X/G gives a bijection

$$\mathrm{Hom}_{\mathbf{Set}^{BG}}(X, \Delta Y) \cong \mathrm{Hom}_{\mathbf{Set}}(X/G, Y).$$

An equivariant map $\Delta Y \to X$ is exactly a map from Y whose values are fixed, giving

$$\mathrm{Hom}_{\mathbf{Set}^{BG}}(\Delta Y, X) \cong \mathrm{Hom}_{\mathbf{Set}}(Y, X^G).$$

An equivariant map $f\colon X \to X'$ induces a map on orbits by $Gx \mapsto Gf(x)$ and a map on fixed points by restriction. The two descriptions commute with composition and identity maps, so orbits and fixed points are functors. The bijections are natural because precomposing or postcomposing the tested maps gives the same factorizations on either side. In the language of adjunctions, $(-)/G \dashv \Delta \dashv (-)^G$.

9.5 Burnside's Lemma

For any $g \in G$, write $X^g = \{x \in X \mid gx = x\}$. The set X^g tests fixation by one element at a time, and $X^G = \bigcap_{g \in G} X^g$.

Theorem 9.4 (Burnside's lemma)**.** *If a finite group G acts on a finite set X, then*

$$|X/G| = \frac{1}{|G|} \sum_{g \in G} |X^g|.$$

Proof. Count the same set $E = \{(g, x) \mid gx = x\}$ using its two projections. For a given g, the fiber of p_G over g consists of (g, x) with $x \in X^g$. For a given x, the fiber of p_X over x consists of (g, x) with $g \in G_x$. Each projection partitions E into its fibers, so

$$\sum_{g \in G} |X^g| = |E| = \sum_{x \in X} |G_x|.$$

Now group the terms in the right-hand sum by orbits. If O is an orbit, Proposition 9.3 gives $|G_x| = |G|/|O|$ for every $x \in O$. The contribution of this orbit is

$$\sum_{x \in O} |G_x| = |O| \frac{|G|}{|O|} = |G|.$$

The orbit classes partition X, so the sum is $|X/G|\,|G|$. A group has at least its identity, and division by $|G|$ gives the formula. For $X = \varnothing$, all terms in the sum over G vanish and the orbit set is empty. □

Burnside's lemma computes the size of a colimit. Its terms $|X^g|$ come from simpler limit tests, namely the equalizers of $F(g)$ and id_X. The orbit–stabilizer calculation is what makes each orbit contribute one copy of $|G|$ to the total number of loops. For a group-theoretic treatment of the counting lemma, see Judson [7, § 14.3].

9.6 Four positions and two colors

Number the positions around a square by $i \in \mathbb{Z}/4\mathbb{Z}$. A coloring is a function $c\colon \mathbb{Z}/4\mathbb{Z} \to \{0,1\}$. The cyclic group C_4 is generated by the quarter turn r sending position i to $i+1$. A rotated coloring has at position i the color that was at position $i-k$ before the rotation r^k. Thus the action is

$$(r^k c)(i) = c(i-k).$$

Applying r^ℓ after r^k shifts the color source by $-\ell - k$, as required by the action law.

A coloring is fixed by r^k exactly when it is constant on each cycle of the position permutation $i \mapsto i+k$. For each cycle there are two independent choices of color. The cycle decompositions and resulting counts are

rotation	cycles on positions	fixed colorings
e	$(0)(1)(2)(3)$	$2^4 = 16$
r	$(0\,1\,2\,3)$	$2^1 = 2$
r^2	$(0\,2)(1\,3)$	$2^2 = 4$
r^3	$(0\,3\,2\,1)$	$2^1 = 2$.

For instance, the half turn fixes the patterns (a,b,a,b) with $a,b \in \{0,1\}$, while a quarter turn requires all four entries to agree. Burnside's lemma gives

$$|X/C_4| = \frac{16+2+4+2}{4} = 6.$$

The six orbits have representatives $0000, 1111, 1000, 0111, 1100, 1010$, where positions are read in the order $0,1,2,3$. The first two have constant colors. The next two have exactly one 1 or exactly one 0. Of the patterns with two 1's, one orbit has adjacent 1's and the other has opposite 1's. For example, the orbit of 1000 is the cycle

$$\begin{array}{ccc} 1000 & \xrightarrow{\;r\;} & 0100 \\ {\scriptstyle r}\uparrow & & \downarrow{\scriptstyle r} \\ 0001 & \xleftarrow{\;r\;} & 0010. \end{array}$$

Both constant patterns have orbit size one. The patterns $1000, 0111, 1100$ have orbit size four, and 1010 has orbit size two. The six sizes sum to 16, accounting for every coloring.

The same derivation counts colorings of $n \geq 1$ cyclic positions with $q \geq 1$ labeled colors. The shift by k has cycles of length $n/\gcd(n,k)$. To see this, put $d = \gcd(n,k)$. The first return occurs at the least positive m for which n divides mk. Since n/d and k/d are coprime, this divisibility is equivalent to $n/d \mid m$. There are therefore d cycles, including the n singleton cycles for $k = 0$. Coloring the cycles independently and applying Burnside's lemma gives the number of rotation classes

$$\frac{1}{n}\sum_{k=0}^{n-1} q^{\gcd(n,k)}.$$

9.7 Cyclic necklaces and dihedral bracelets

A *necklace* is a rotation class of colorings of n cyclic positions. Write $N_n(q)$ for the number of necklaces with q labeled colors. For a divisor $d \mid n$, the rotations of order d are the shifts by $(n/d)j$ with $0 \leq j < d$ and $\gcd(j,d) = 1$. Write $\varphi(d)$ for their number, including $\varphi(1) = 1$. Each such shift has n/d cycles, so grouping the preceding Burnside sum by rotation order gives

$$N_n(q) = \frac{1}{n}\sum_{d \mid n} \varphi(d) q^{n/d}.$$

A *bracelet* allows both rotations and reflections. For $n \geq 3$, the dihedral group of the n-gon consists of the n rotations and the n reflections $s_k(i) = k - i$ on $\mathbb{Z}/n\mathbb{Z}$. A reflection conjugates each rotation to its inverse, so the rotation subgroup is normal and a reflection induces an involution on the set of necklaces. The bracelet set is the coequalizer of this involution and the identity. Every dihedrally invariant map from colorings first factors through the necklace set, then factors uniquely through the coequalizer.

A fixed position of s_k solves $2i = k$ modulo n. If there are f fixed positions, all remaining positions fall into pairs. Thus there are $f + (n - f)/2 = (n + f)/2$ cycles, each allowing one independent choice of color. For odd n, every reflection has $f = 1$ because 2 is invertible modulo n. For even n, the $n/2$ even values of k give $f = 2$, and the $n/2$ odd values give $f = 0$. Burnside's lemma gives the bracelet count

$$B_n(q) = \frac{1}{2n}\left(\sum_{d \mid n} \varphi(d) q^{n/d} + \begin{cases} n q^{(n+1)/2}, & n \text{ odd}, \\ \frac{n}{2}\left(q^{n/2+1} + q^{n/2}\right), & n \text{ even}. \end{cases}\right)$$

The divisor sum comes from the rotations, and the two cases give the total fixed colorings contributed by the reflections.

Take six positions and two colors. The rotation sum is $64 + 8 + 8 + 4 = 84$ from the divisors $1, 2, 3, 6$, giving 14 necklaces. The three reflections with fixed

vertices and the three without them contribute $3 \cdot 16 + 3 \cdot 8 = 72$. There are $(84 + 72)/12 = 13$ bracelets. The involution therefore exchanges two necklace classes and fixes the other twelve. The exchanged classes are represented by 001011 and its reversal 110100. Indeed, the six rotations of 001011 are $001011, 010110, 101100, 011001, 110010, 100101$.

9.8 From group actions to algebra actions

A group action sends each group element to a permutation. A ring action must interpret addition of scalars as well as multiplication. That adds structure to the object on which the scalars act. The same four-position coloring example will give a linear action and show how a single element of its group algebra encodes the orbit count.

9.8.1 Ring and algebra actions

The one-object description of an action extends to a ring R with identity by retaining addition on the arrows. The corresponding objects acted on are abelian groups. For an abelian group M, let $\mathrm{End}_{\mathbf{Ab}}(M)$ be the ring of additive endomorphisms of M. Addition is pointwise, multiplication is composition, and the multiplicative identity is id_M. Composition distributes over pointwise addition because the endomorphisms preserve addition.

A left R-module is an abelian group M with a scalar action $(r, x) \mapsto rx$ satisfying

$$(r+s)x = rx + sx, \quad r(x+y) = rx + ry,$$
$$(rs)x = r(sx), \qquad 1_R x = x.$$

The map $x \mapsto rx$ is additive by the second equation, and it takes zero to zero by cancellation in M. Write this endomorphism as $\rho_M(r)$. The other equations give, respectively,

$$\rho_M(r+s) = \rho_M(r) + \rho_M(s), \qquad \rho_M(rs) = \rho_M(r) \circ \rho_M(s), \qquad \rho_M(1_R) = \mathrm{id}_M .$$

Thus $\rho_M \colon R \to \mathrm{End}_{\mathbf{Ab}}(M)$ is a unital ring homomorphism. Conversely, evaluating such a homomorphism by $rx = \rho_M(r)(x)$ gives the scalar action equations, and these constructions recover each other.

Let BR have one object $\star$, with $BR(\star, \star) = R$ and $r \circ s = rs$. It is a *preadditive category*, meaning that each hom-set is an abelian group and composition is additive in each variable. Here both properties follow from the ring laws. The category $\mathbf{Ab}$ of abelian groups has the same structure on its hom-sets by pointwise addition. An *additive functor* preserves sums and zero in each hom-set along with the ordinary functor laws. Thus an additive functor $BR \to \mathbf{Ab}$ is precisely the data M and ρ_M above.

A natural transformation between two such functors has one component, an additive map $u\colon M \to N$. Its naturality at $r \in R$ is expressed by the following diagram in **Ab**.

$$\begin{array}{ccc} M & \xrightarrow{\rho_M(r)} & M \\ \downarrow{\scriptstyle u} & & \downarrow{\scriptstyle u} \\ N & \xrightarrow{\rho_N(r)} & N. \end{array}$$

The square says that $u(rx) = ru(x)$. Thus an intertwiner of the actions is exactly an R-linear map. This identifies the category of left R-modules with additive functors out of BR and natural transformations. For a right action, set $\rho(r)(x) = xr$. The identity $\rho(r)\rho(s)(x) = (xs)r = x(sr)$ exhibits a homomorphism from the opposite ring R^{op}, whose multiplication is reversed. Right R-modules are consequently additive functors $(BR)^{\mathrm{op}} \to \mathbf{Ab}$.

For a commutative base ring k, a unital k-algebra A has a unital scalar map $k \to Z(A)$, where the center $Z(A)$ consists of elements commuting with every element of A. On a given k-module M, a left A-action extending its scalars corresponds to a unital k-algebra map

$$A \longrightarrow \mathrm{End}_k(M).$$

Indeed, the action of $a \in A$ commutes with the action of each scalar by centrality. If $\kappa\colon k \to Z(A)$ is the structural map, then $a(cx) = (a\kappa(c))x = (\kappa(c)a)x = c(ax)$, so the action of a is k-linear. The scalar map on the endomorphism algebra sends $c \in k$ to $c\,\mathrm{id}_M$, and compatibility with this map gives the required scalar action back on restriction.

The polynomial presentation of Section 5.5.1 gives a concrete case. An action of $k[t]$ on M is determined by the k-linear endomorphism T assigned to t. It acts by

$$\left(\sum_i c_i t^i\right) x = \sum_i c_i T^i(x).$$

Scalar endomorphisms commute with T, and $T^iT^j = T^{i+j}$, so this formula respects polynomial multiplication as well as addition and the unit. The action descends to $k[t]/(t^2 - t)$ precisely when $T^2 = T$. Descending forces the equation because the polynomial $t^2 - t$ acts by zero. Conversely, when $T^2 = T$, every multiple of $t^2 - t$ acts by zero, so the operator associated to a polynomial depends only on its class.

An idempotent T gives a direct-sum decomposition $M = \mathrm{im}(T) \oplus \ker(T)$. Every x is the sum of Tx in the image and $x - Tx$ in the kernel. An element z in the image satisfies $Tz = z$, so membership in the kernel as well forces $z = 0$. For example $T(x, y) = (x, 0)$ on k^2 makes t act as identity on the first coordinate and zero on the second. Conversely, an ordered decomposition $M = M_1 \oplus M_0$ gives an action of $k[t]/(t^2 - t)$ by taking T to be projection onto M_1 along M_0.

A k-linear map $u\colon M \to N$ intertwines two polynomial actions with operators T and S exactly when $uT = Su$. The condition is necessary by evaluation at t and sufficient because it implies $uT^i = S^i u$ for each $i \geq 0$. For idempotent operators, this is equivalent to sending the image summand to the image summand and the kernel summand to the kernel summand. Indeed, T acts as identity on its image and zero on its kernel, and the corresponding statement holds for S. The generator and its relation thus determine both the decomposition and the maps that respect it.

A group can also act on a k-algebra B by algebra automorphisms. In that case each g carries products bc to $(g \cdot b)(g \cdot c)$ and fixes the unit. The elements fixed by every g therefore form a subalgebra B^G, since the automorphisms also preserve addition and k-scalars.

9.8.2 Orbits under units

Write $R^\times$ for the group of units of a ring R, the elements with a two-sided multiplicative inverse. In a ring action on M, an element $u \in R^\times$ acts by an automorphism of the underlying abelian group, with inverse the action of u^{-1}. Thus a finite subgroup $H \subseteq R^\times$ gives a group action to which Burnside's lemma applies on any finite H-invariant subset of M.

Example 9.5 (The units acting on six residues)**.** Take $R = M = \mathbb{Z}/6\mathbb{Z}$ with the action by multiplication. Its unit group is $\{1, 5\}$. Multiplication by 1 fixes all six residues, and multiplication by $5 = -1$ fixes precisely 0 and 3, the solutions of $2x = 0$. Burnside's lemma gives

$$|M/R^\times| = \frac{6 + 2}{2} = 4.$$

The four orbits are $\{0\}, \{3\}, \{1, 5\}, \{2, 4\}$. The full scalar action determines a reachability relation. From m one can reach any point of $Rm = \{rm : r \in R\}$, the cyclic submodule generated by m. Zero is reachable from every point, and from zero one reaches only zero. Here $R1 = R$, while $R0 = \{0\}$. The two examples of reachability show how cyclic submodules record the full scalar action.

9.8.3 Group algebras and linear actions

Let k be a commutative ring with unit. The *group algebra* $k[G]$ is the free k-module with one basis element $[g]$ for each $g \in G$. Its elements are finite sums $\sum_g a_g[g]$. Multiplication extends $[g][h] = [gh]$ by k-bilinearity. On each triple of basis elements, associativity follows from $(gh)\ell = g(h\ell)$ in G. Both ways of multiplying three elements are k-linear in each argument, so equality on basis triples gives equality on all triples. The unit is $[e]$, and $a \mapsto a[e]$ gives the central

scalar action of k. For the group algebra and its representations over a field, see Etingof and others [8, §§ 2.2–2.3].

A k-linear action of G on a k-module V is a group homomorphism $\rho\colon G \to \mathrm{Aut}_k(V)$. Linearly extending $[g] \mapsto \rho(g)$ gives a map $\widetilde{\rho}\colon k[G] \to \mathrm{End}_k(V)$. The square commutes as a diagram of multiplicative monoids.

$$\begin{array}{ccc} G & \xrightarrow{\ \rho\ } & \mathrm{Aut}_k(V) \\ {\scriptstyle g\mapsto[g]}\downarrow & & \downarrow \\ k[G] & \xrightarrow{\ \widetilde{\rho}\ } & \mathrm{End}_k(V) \end{array}$$

In fact $\widetilde{\rho}$ is a k-algebra homomorphism. For $a = \sum_g a_g[g]$, $b = \sum_h b_h[h]$, and $v \in V$,

$$\widetilde{\rho}(a)\big(\widetilde{\rho}(b)(v)\big) = \sum_{g,h} a_g b_h \rho(g)\rho(h)(v) = \sum_{g,h} a_g b_h \rho(gh)(v) = \widetilde{\rho}(ab)(v).$$

The unit maps to $\rho(e) = \mathrm{id}_V$. Therefore $av = \widetilde{\rho}(a)(v)$ makes V a left $k[G]$-module. Conversely, a left $k[G]$-module restricts to a k-module along $k \to k[G]$. Acting by $[g]$ gives a k-linear automorphism with inverse the action of $[g^{-1}]$, and $[g][h] = [gh]$ ensures the group action law. These constructions are inverse because the module action is determined by its action on the basis elements.

A k-linear map $f\colon V \to W$ is an *intertwiner* if $f \circ \rho_V(g) = \rho_W(g) \circ f$ for each g. Such an f commutes with the action of every sum $\sum_g a_g[g]$ by k-linearity. Conversely, a $k[G]$-module map is k-linear by restriction and intertwines the action of every $[g]$. The correspondences preserve identity maps and composition, giving an equivalence of the category of k-linear G-actions with the category of left $k[G]$-modules.

The extension argument works with any unital k-algebra B in place of $\mathrm{End}_k(V)$. Each algebra map out of $k[G]$ sends $[g]$ to a unit because $[g^{-1}]$ is its inverse. Each group homomorphism $G \to B^\times$ extends uniquely by k-linearity, and the same product calculation gives an algebra map. These assignments are natural in both G and B. Thus the construction is an adjunction

$$\mathrm{Hom}_{\mathbf{Alg}_k}(k[G], B) \;\cong\; \mathrm{Hom}_{\mathbf{Grp}}(G, B^\times),$$

where $\mathbf{Alg}_k$ is the category of associative unital k-algebras.

For $C_4 = \langle r \rangle$, the assignment $t \mapsto [r]$ gives $k[C_4] \cong k[t]/(t^4 - 1)$. The relation holds because $r^4 = e$. It reduces every polynomial class to a linear combination of $1, t, t^2, t^3$. The map sends these four classes to $[e], [r], [r^2], [r^3]$, the given free basis of $k[C_4]$. This proves surjectivity and makes the kernel zero. An action of C_4 on a k-module can be specified by one k-linear operator T satisfying $T^4 = \mathrm{id}_V$. Its inverse is T^3. For example, the choice $T = -\mathrm{id}_V$ on a rational line defines an action in which each quarter turn changes a vector's sign.

9.8.4 Averaging and a trace proof of Burnside's lemma

Any G-set X gives a free k-module $k[X]$ on its elements, with g acting on basis vectors by $e_x \mapsto e_{gx}$. This is a $k[G]$-module by the preceding construction. A $k[G]$-linear map from $k[X]$ to M corresponds to an equivariant assignment $x \mapsto m_x$ in the underlying G-set of M. The unique linear extension of the assignment is equivariant exactly when it is so on the basis. This gives another free–forgetful adjunction.

Suppose G and its set X are finite. Form the rational vector space $V = \mathbb{Q}[X]$ with basis $(e_x)_{x \in X}$. The permutations of X act linearly by $\rho(g)e_x = e_{gx}$. Consider the operator

$$P = \frac{1}{|G|} \sum_{g \in G} \rho(g).$$

For every $h \in G$, multiplication by h permutes the summation indices, giving $\rho(h)P = P$. Hence the image of P lies in the subspace V^G of vectors fixed by every $\rho(h)$. For $v \in V^G$, every summand sends v to itself, so $Pv = v$. Consequently $P^2 = P$ and $\operatorname{im} P = V^G$.

The operator is the action of the element $a_G = |G|^{-1} \sum_{g \in G} [g]$ in $\mathbb{Q}[G]$. Left or right multiplication by $[h]$ permutes its terms, so $[h]a_G = a_G = a_G[h]$ for every h. Averaging these equalities gives $a_G^2 = a_G$. Thus one central idempotent of the group algebra acts as the averaging projection on every rational G-module. The map $\mathbb{Q}[t]/(t^2 - t) \to \mathbb{Q}[G]$ sending t to a_G makes this an instance of the idempotent presentation in Section 9.8.1. The same averaging construction works over any commutative ring in which $|G|$ is a unit.

A vector $\sum_x a_x e_x$ is invariant exactly when its coefficients are constant on each orbit. Thus the sums $\sum_{x \in O} e_x$, one for each orbit O, span V^G. They are linearly independent because the orbits have disjoint supports. It follows that $\dim V^G = |X/G|$. The decomposition $v = Pv + (v - Pv)$ writes V as $\operatorname{im} P \oplus \ker P$. Indeed, $P(v - Pv) = 0$, and P acts as the identity on its image, so the summands meet only in 0. Trace is the sum of the diagonal entries, and it is unchanged under change of basis because $\operatorname{tr}(AB) = \sum_{i,j} A_{ij} B_{ji} = \operatorname{tr}(BA)$. In a basis adapted to this decomposition, the trace of P equals $\dim V^G$.

In the original basis, the xth diagonal entry of $\rho(g)$ is 1 when $gx = x$ and 0 otherwise. Thus $\operatorname{tr}(\rho(g)) = |X^g|$. Computing $\operatorname{tr}(P)$ by linearity now gives a second proof of Theorem 9.4.

$$|X/G| = \operatorname{tr}(P) = \frac{1}{|G|} \sum_{g \in G} |X^g|.$$

For the sixteen binary colorings of the square from Section 9.6, the operators for e, r, r^2, r^3 have traces $16, 2, 4, 2$, respectively. The averaging projection has

trace 6 and its image has the six orbit sums as a basis. On two individual basis vectors it acts by

$$Pe_{1000} = \tfrac{1}{4}(e_{1000} + e_{0100} + e_{0010} + e_{0001}),$$
$$Pe_{1010} = \tfrac{1}{2}(e_{1010} + e_{0101}).$$

The alternating coloring returns after a half turn, so each of its two orbit members occurs twice in the four-term average.

9.8.5 Actions as maps from a tensor product

Let A be a unital algebra over a commutative ring k and M a k-module. The tensor product $A \otimes_k M$ represents k-bilinear maps out of $A \times M$. Using this universal property from Section 14.1, the rule $(a, m) \mapsto a \cdot m$ is a single k-linear map $\alpha\colon A \otimes_k M \to M$. Let $\mu\colon A \otimes_k A \to A$ be multiplication and $\eta\colon k \to A$ the unit. The associativity law says that the following square commutes, with the two parenthesizations of three tensor factors identified by their canonical isomorphism.

$$\begin{array}{ccc} A \otimes_k A \otimes_k M & \xrightarrow{\mu \otimes \mathrm{id}_M} & A \otimes_k M \\ \downarrow{\scriptstyle \mathrm{id}_A \otimes \alpha} & & \downarrow{\scriptstyle \alpha} \\ A \otimes_k M & \xrightarrow{\quad\alpha\quad} & M \end{array}$$

On $a \otimes b \otimes m$, the paths give $(ab) \cdot m$ and $a \cdot (b \cdot m)$. The unit law says that $\alpha \circ (\eta \otimes \mathrm{id}_M)$ is the map $k \otimes_k M \to M$ sending $c \otimes m$ to cm. The square and unit law determine the usual action laws because the pure tensors generate their respective tensor products.

This describes a module as an action of a monoid object in the category of k-modules with tensor product. Part III develops the same structure using additive functors, then uses generators and relations to present modules and tensor products to change scalars.

CHAPTER 10

Rings and algebras from generators and relations

A ring expression such as $3xy - 2yx + 1$ combines two structures. The products xy and yx record two distinct orders of the letters, and the whole expression is a sum with integer coefficients. The free associative algebra keeps that order in its monomials and uses distributivity to multiply their sums. Equations between expressions then give quotients. The resulting universal properties specify which assignments of the generators extend to maps out of each algebra.

10.1 Unital rings and their scalars

Definition 10.1. A *unital ring* A is an abelian group under addition, equipped with an associative multiplication and a multiplicative unit 1_A. Multiplication distributes over addition in each argument. A homomorphism of unital rings preserves addition, multiplication, and the multiplicative unit. Write **Ring** for their category.

The one-element ring is included. Each map $x \mapsto ax$ or $x \mapsto xa$ for a fixed a is additive by distributivity. An additive map takes zero to zero, by applying it to $0 + 0$ and cancelling. It also takes additive inverses to additive inverses, by applying it to $x + (-x)$. In particular $0a = a0 = 0$, and the negative signs in a product follow from additivity.

Proposition 10.2 (The initial unital ring)**.** *For every unital ring A, there is a unique unital ring map $\lambda_A \colon \mathbb{Z} \to A$. It sends n to $n1_A$, and its image is central in A.*

Proof. The additive group of the integers is freely generated by 1, as in Chapter 7, so $n \mapsto n1_A$ is the only additive map sending 1 to 1_A. For $a \in A$, left and right multiplication by a are additive and send 1_A to a. Applied to the integer multiples of 1_A, they give $(n1_A)a = na = a(n1_A)$. Taking $a = m1_A$ shows $(n1_A)(m1_A) = (nm)1_A$, so λ_A preserves multiplication. □

Let k be a commutative unital ring. The *center* $Z(A)$ of A is the subring of elements that commute with every element of A. A *unital associative k-algebra* is

a unital ring A with a specified unital map $\lambda_A\colon k \to Z(A)$. A map of k-algebras is a unital ring map that commutes with these scalar maps. Write k-**Alg** for their category. The structural map $k \to A$ is the unique k-algebra map from k itself, so k is initial in k-**Alg**. For $k = \mathbb{Z}$, every ring already has the unique scalar map from Proposition 10.2. Every unital ring map commutes with these maps because both composites from $\mathbb{Z}$ send 1 to the unit. Thus the category of $\mathbb{Z}$-algebras gives back **Ring**.

The scalar action on A is $ca = \lambda_A(c)a$. This makes A a k-module. Centrality and associativity give $(ca)b = c(ab) = a(cb)$, and distributivity gives additivity of the product in each argument. In particular, multiplication is k-bilinear. A k-algebra whose underlying ring is commutative is a commutative k-algebra, as in Section 5.5.1.

10.2 The free associative algebra

For a set S, let S^* be the monoid of finite words in S, with concatenation and empty word ϵ. Define $k\langle S\rangle$ to consist of formal finite sums $\sum_{w\in S^*} c_w b_w$, with coefficients in k, so that only finitely many c_w are nonzero. Here b_w denotes the basis expression for the word w. Addition is coefficientwise and scalars multiply coefficients. Define the product by

$$\left(\sum_u a_u b_u\right)\left(\sum_v c_v b_v\right) = \sum_{u,v} a_u c_v b_{uv}.$$

The right side is a finite sum with contributions to the same word added together. Multiplying any three basis expressions gives b_{uvw} in either grouping by associativity of concatenation. Expansion of finite sums therefore proves associativity in general. Expansion also gives both distributive laws. The empty word represents the unit and $c \in k$ acts by the central element cb_ϵ, since appending an empty word on either side leaves a word unchanged. These operations give a k-algebra.

Proposition 10.3 (Free associative k-algebra)**.** *For every k-algebra A, each function $u\colon S \to U(A)$ extends uniquely to a k-algebra map $\widehat{u}\colon k\langle S\rangle \to A$ sending $b_{[s]}$ to $u(s)$. Here U is the underlying-set functor, and $b_{[s]}$ is the basis expression of the one-letter word $[s]$.*

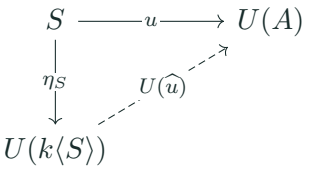
Proof. For a word $w = [s_1, \dots, s_n]$, put $u_w = u(s_1)\cdots u(s_n)$, with $u_\epsilon = 1_A$. Define

$$\widehat{u}\left(\sum_w c_w b_w\right) = \sum_w \lambda_A(c_w) u_w.$$

The formula preserves addition, the unit, and scalars. It also preserves products because $u_{wv} = u_w u_v$ and the coefficients $\lambda_A(c)$ commute with all the u_w. Any

extension has the displayed value on every word by multiplicativity and therefore has the displayed value on every finite sum. □

Writing $\eta_S \colon S \to U(k\langle S\rangle)$ for the map to the one-letter basis expressions, this gives the factorization in **Set** below.



Postcomposition with a map of k-algebras carries this unique extension to the extension of the postcomposed assignment. A function $v \colon S \to S'$ gives the map of free algebras extending $\eta_{S'} \circ v$. Precomposing with this map restricts an assignment on S' along v, again by uniqueness. Thus the factorization is natural in both variables and gives an adjunction with U, as in Chapter 5.

10.3 Commutative polynomials from finite multisets

When the chosen images of the generators commute pairwise, a word can be evaluated from its multiplicities. A finite multiset on S is a function $d \colon S \to \mathbb{N}$ that is zero outside a finite subset. These multisets form a commutative monoid under pointwise addition. For each d write X^d for a basis expression. The polynomial algebra $k[S]$ consists of finite k-linear combinations of these expressions, with

$$X^d X^{d'} = X^{d+d'}, \qquad 1 = X^0.$$

The product is extended by the same finite-sum rule as for words. Associativity and commutativity follow from the corresponding identities for multiset addition. A generator s maps to X^{δ_s}, where δ_s is 1 at s and 0 elsewhere.

For a commuting family $a_s \in A$, an extension must send X^d to the product of $a_s^{d(s)}$ over the finite support of d. The value is independent of the order because the factors commute pairwise. Combining the two products for d, d' gives the product for $d + d'$, so extension by finite sums gives a k-algebra map. The powers of the images and scalar coefficients force this map uniquely. Conversely, every map from $k[S]$ has commuting images, since the generators commute in $k[S]$. In the category of commutative k-algebras this condition holds for every family, so $k[S]$ is free on S there.

There is a surjection $k\langle S\rangle \to k[S]$ taking a basis word to the monomial for its multiset of letters. It is multiplicative because concatenation adds multiplicities, and it is surjective because a finite multiset can be listed as a word. For example, with $S = \{x, y\}$ the words xyx and xxy both map to x^2y. Their common value

under any assignment of commuting elements explains the identification in the target.

10.4 Imposing equations by a quotient

An equivalence relation on an algebra is an *algebra congruence* if it respects addition, additive inverse, and multiplication. The constants, including the scalars, pass to their classes. A *two-sided ideal* I of A is an additive subgroup with $aI \subseteq I$ and $Ia \subseteq I$ for every $a \in A$. Congruences correspond to two-sided ideals by taking the class of zero. Indeed, for a congruence the class of zero is closed under sums and negatives, and multiplying a member on either side sends its class to zero. Adding $-b$ gives $a \sim b$ precisely when $a - b \sim 0$. Conversely, a two-sided ideal gives an equivalence relation by $a \sim b$ if $a - b \in I$. The additive subgroup property gives compatibility with addition and negation. For multiplication, $a - a' \in I$ and $b - b' \in I$ imply

$$ab - a'b' = (a - a')b + a'(b - b') \in I.$$

Thus the operations on classes give a k-algebra A/I, with scalar map induced from A. The quotient arrow is a k-algebra map.

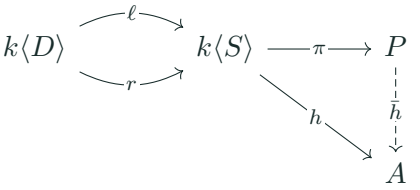
Take a set D indexing equations $p_d = q_d$ with $p_d, q_d \in k\langle S\rangle$. Their *presentation* is the quotient P of $k\langle S\rangle$ by the least congruence containing all these pairs. Equivalently, divide by the ideal I generated by the differences $p_d - q_d$. This ideal consists of the finite sums of expressions $a(p_d - q_d)b$ for a, b in the free algebra. Such sums are closed under addition and negation, and multiplication on either side keeps them in the set. The unit shows that every prescribed difference is included. An ideal containing the differences must contain every such sum, which proves minimality.

Proposition 10.4 (Maps from a presented algebra)**.** *Maps $P \to A$ of k-algebras correspond naturally in A to assignments $u\colon S \to U(A)$ for which $\widehat{u}(p_d) = \widehat{u}(q_d)$ for every $d \in D$.*

Proof. An assignment extends uniquely to the free algebra. If every specified pair has equal value, the formula for the generated ideal shows that its elements go to zero. Sending a class to the value of a representative is then well-defined and gives the unique factorization through P. Conversely, a map out of P gives a satisfying assignment because the two sides of each equation have equal classes. The constructions are inverse by the free and quotient properties. Postcomposition carries assignments to the corresponding new assignments. □

Extend the assignments $d \mapsto p_d$ and $d \mapsto q_d$ to maps $\ell, r\colon k\langle D\rangle \rightrightarrows k\langle S\rangle$. A map h coequalizes ℓ, r exactly when it identifies the assigned pairs of values on

D. The quotient map π is therefore their coequalizer in k-**Alg**.



For example, the quotient of $k\langle S\rangle$ imposing $st = ts$ for every $s, t \in S$ represents assignments of pairwise commuting elements. The construction of $k[S]$ represents the same assignments. Uniqueness of representing objects identifies the two algebras by the map carrying each generator to itself.

10.4.1 A presentation of a given algebra

Let A be any k-algebra. Take one generator x_a for each element a of its underlying set and impose the relations

$$x_a + x_b = x_{a+b}, \qquad x_a x_b = x_{ab}, \qquad x_{\lambda_A(c)} = c1 \quad (c \in k).$$

Write P_A for the presented algebra. The assignment $x_a \mapsto a$ satisfies the equations, so it gives a map $\pi\colon P_A \to A$. The assignment in the other direction sends a to the class of x_a. The first two families of relations make this function preserve sums and products. The scalar relations give preservation of zero, unit, and all scalars, so it is a k-algebra map $\sigma\colon A \to P_A$.

The composite $\pi \circ \sigma$ fixes every element a of A. The composite $\sigma \circ \pi$ fixes each generator of P_A, and every element is obtained from those generators by products, sums, and scalars. Therefore the composites are identities. This presentation uses the entire operation table of A, while the matrix example below recovers a ring from two generators and two relations.

10.5 A presentation for upper triangular matrices

Consider the k-algebra of upper triangular two-by-two matrices, denoted B. Its elements are sums $aI_2 + bE + cW$, where I_2 is the identity matrix and

$$E = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad W = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

The formula $E^2 = E$ says that E is a projection, and $W = EW(I_2 - E)$ says that the input and output of W occupy specific summands. These two equations present B.

Proposition 10.5. *The algebra presented by generators e, w and relations $e^2 = e$ and $w = ew(1-e)$ is isomorphic to B, with e, w sent to E, W respectively. Every element has a unique expression $a1 + be + cw$ with $a, b, c \in k$.*

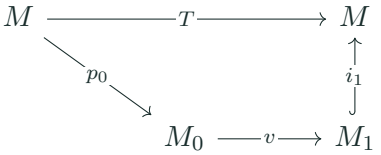
Proof. In the presented algebra, multiply $w = ew(1-e)$ on the left by e. The relation $e^2 = e$ gives $ew = w$. Multiplying it on the right by e gives $we = 0$, using $(1-e)e = 0$. Then $w^2 = w(ew) = (we)w = 0$. The k-linear combinations of $1, e, w$ are closed under multiplication by these equations and contain the unit and both generators. Induction on word length shows that they contain every word, and hence they give all elements of the presented algebra.

Matrix multiplication gives $E^2 = E$, $EW = W$, and $WE = 0$. Consequently $EW(I_2 - E) = W$, so the presentation produces a homomorphism to B. On the proposed expressions the map has the formula

$$a1 + be + cw \longmapsto \begin{pmatrix} a+b & c \\ 0 & a \end{pmatrix}.$$

The lower diagonal entry determines a, the upper diagonal entry minus a determines b, and the off-diagonal entry determines c. This proves uniqueness of the expressions and injectivity of the map. Every upper triangular matrix occurs by choosing the three coefficients in this way, so the map is an isomorphism. □

The presentation describes actions of B on a k-module M in the sense of Section 9.8.1. For representation terminology over a field, see [8, Section 2.3]. The endomorphism P assigned to e is idempotent. By the image–kernel splitting proved in Section 9.8.1, write $M_1 = \operatorname{im}(P)$, $M_0 = \ker(P)$, and $M = M_1 \oplus M_0$. The equation for the endomorphism T assigned to w reads $T = PT(1-P)$. It says that T vanishes on M_1 and maps M_0 into M_1. Thus T is determined by its restriction $v\colon M_0 \to M_1$.



Conversely, given a decomposition $M = M_1 \oplus M_0$ and a k-linear map $v\colon M_0 \to M_1$, let P be projection onto M_1 and put $T = i_1 v p_0$. Here i_1 is inclusion of M_1 and p_0 is projection onto M_0. Then $P^2 = P$, $Pi_1 = i_1$, and $p_0(1-P) = p_0$, so the relations hold and specify a unique B-action. The relation on w therefore allows precisely one freely chosen linear map between the two summands.

Maps of B-modules also have a description in terms of the summands. Commuting with the action of e is equivalent to preserving the image and kernel summands. A k-linear map with this property has components $u_1\colon M_1 \to N_1$

and $u_0 \colon M_0 \to N_0$. Commutation with the action of w reduces to the following square on M_0, since the w actions are zero on M_1 and N_1.

$$\begin{array}{ccc} M_0 & \xrightarrow{v} & M_1 \\ \downarrow{\scriptstyle u_0} & & \downarrow{\scriptstyle u_1} \\ N_0 & \xrightarrow{v'} & N_1 \end{array}$$

Since e, w generate B as a k-algebra, commutation with those two actions gives commutation with the action of every matrix in B.

In a presentation the ideal of differences lists everything that must map to zero. More generally, the elements a homomorphism sends to zero form its kernel ideal. Two inputs have equal images precisely when their difference belongs to this ideal, connecting presentations to the quotient and image constructions for ring maps.

CHAPTER 11

Ideals, quotients, and localization

A ring map can impose new equations on familiar elements. Reduction modulo 6 sends 6 to zero. Inclusion into the rational numbers makes 6 invertible. Each operation has a universal description. A quotient accepts every map that kills a prescribed collection of elements. A localization accepts every map that sends a prescribed collection to units. These descriptions let us calculate the resulting rings and compare different constructions of them.

Throughout this chapter rings are commutative and unital, and ring maps preserve the unit. We allow the zero ring, whose one element is both zero and the unit. An element s is a *unit* if it has a multiplicative inverse.

11.1 Ideals and the equations of a quotient

Definition 11.1. An *ideal* I of a ring R is an additive subgroup of R such that $rx \in I$ for $r \in R$ and $x \in I$. A *ring congruence* is an equivalence relation on R compatible with addition, negation, and multiplication.

To impose $x = 0$ in the image of a ring homomorphism is also to impose $-x = 0$ and $rx = 0$ there for every r. The elements that satisfy such equations are therefore governed by an ideal. For example, the kernel of a ring map $f\colon R \to A$, defined by $f(r) = 0$, is an ideal since f preserves addition, negation, and multiplication.

Proposition 11.2. *The ring congruences on R correspond bijectively to its ideals. The congruence determined by I is $x \sim_I y$ when $x - y \in I$.*

Proof. Every unital ring is a $\mathbb{Z}$-algebra by Proposition 10.2. Section 10.4 identifies algebra congruences with two-sided ideals by taking the class of zero. In a commutative ring the left ideal and two-sided ideal conditions agree, so the correspondence gives the stated rule. □

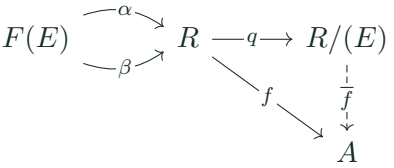
The quotient R/I is the set of $\sim_I$-classes with operations induced from R. Compatibility makes the operations independent of representatives. Applying each ring axiom to representatives and passing to classes verifies it in the quotient. The map $q\colon R \to R/I$, $r \mapsto [r]$, is therefore a surjective ring map, with kernel I.

Proposition 11.3 (Universal property of a quotient)**.** *A ring map $f\colon R \to A$ factors uniquely through $q\colon R \to R/I$ if $I \subseteq \ker f$. Every map that factors through q has this property.*

Proof. If $x - y \in I \subseteq \ker f$, then $f(x) - f(y) = 0$. Hence the formula $\overline{f}([x]) = f(x)$ is well-defined. It preserves the ring operations because those operations are induced from R. Surjectivity of q gives uniqueness, and $q(x) = 0$ for $x \in I$ gives the converse. □

Given a subset $E \subseteq R$, the ideal (E) it generates consists of finite sums $\sum_i r_i e_i$ with $e_i \in E$. The empty sum is zero, addition concatenates sums, negation changes signs of coefficients, and multiplication by a ring element multiplies each coefficient by that element. The set is thus an ideal. It contains E by taking coefficient 1, and every ideal containing E must contain all the sums.

The quotient by (E) can be presented as a coequalizer. Let $F(E)$ be the free commutative ring on the set of labels for elements of E. There are two maps $\alpha, \beta\colon F(E) \to R$, sending the generator labelled e to e and to 0, respectively.

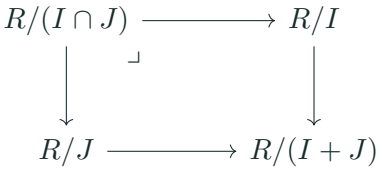


A map f coequalizes α, β exactly when it sends every e to zero, since the two composites are determined by their values on the free generators. Killing every e is equivalent to killing (E), so Proposition 11.3 gives the unique dashed map. The coequalizer here is in the category of rings and uses the same diagram as the set construction in Chapter 3.

11.2 Independent quotient conditions

For ideals $I, J \subseteq R$, their sum $I + J$ consists of all $a + b$ with $a \in I$ and $b \in J$. The ideal laws for I, J give the ideal laws for $I + J$. The ideals are *comaximal* when $I + J = R$, which is equivalent to the existence of $a \in I, b \in J$ with $a + b = 1$. Indeed any ideal containing 1 contains every element of R.

Proposition 11.4. *For any ideals I, J of R, the reductions in the following square exhibit $R/(I \cap J)$ as a pullback in the category of rings.*



Proof. Pairs of residues with the same image in $R/(I+J)$ form a subring of $R/I \times R/J$. Indeed the condition is equality under two ring maps, so it holds for $0, 1$ and is preserved by sums, negatives, and products. This ring of compatible pairs is the pullback. Given ring maps into R/I and R/J with equal composites into $R/(I+J)$, pairing their values gives a ring map into the subring. The coordinate projections determine it uniquely.

Reduction on both coordinates defines a map from $R/(I \cap J)$ into this pullback. Two classes have the same images precisely when the difference of their representatives belongs to both ideals, so the map is injective. If a compatible pair is represented by x and y, then $x - y = a + b$ for some $a \in I, b \in J$. The element $x - a = y + b$ represents the pair, proving surjectivity. □

Theorem 11.5 (Chinese remainder theorem for two ideals)**.** *If I, J are comaximal ideals of R, reduction on both coordinates gives a ring isomorphism*

$$R/(I \cap J) \longrightarrow R/I \times R/J.$$

Moreover $I \cap J = IJ$, where IJ is the ideal of finite sums of products xy with $x \in I$ and $y \in J$.

Proof. By comaximality, $R/(I+J)$ is the zero ring. All pairs of residues agree there, so Proposition 11.4 identifies $R/(I \cap J)$ with the whole product. An explicit simultaneous representative for x modulo I and y modulo J is $bx + ay$, with $a \in I, b \in J, a + b = 1$. Modulo I we have $a = 0, b = 1$, so it gives x. Modulo J we have $a = 1, b = 0$, so it gives y.

Every product xy with $x \in I, y \in J$ lies in both ideals, and so does every finite sum of such products. Conversely, for $z \in I \cap J$ the equality $z = za + zb$ expresses z as a sum of products of the required form. Hence $I \cap J = IJ$. □

Example 11.6. Take $I = 2\mathbb{Z}$ and $J = 3\mathbb{Z}$. We can use $a = -2$, $b = 3$ because $-2 + 3 = 1$. The intersection is $6\mathbb{Z}$ by the theorem, and a pair (x mod 2, y mod 3) lifts to $3x - 2y$ mod 6. For instance $(1, 2)$ lifts to -1 mod 6, which is the class of 5.

The classes of 3 and 4 modulo 6 correspond to $(1, 0)$ and $(0, 1)$. They are idempotent, their product is zero, and their sum is one. Multiplying by these two idempotents selects the two residue coordinates separately.

Compatibility can impose a genuine restriction. With $I = 4\mathbb{Z}$ and $J = 6\mathbb{Z}$, their sum is $2\mathbb{Z}$ since all elements of the sum are even and $6 - 4 = 2$. Their intersection is $12\mathbb{Z}$. The pullback description says that a residue modulo 4 and one modulo 6 come from a common integer precisely when they have the same parity. The pair (3 mod 4, 1 mod 6) comes from 7 mod 12. The pair (0 mod 4, 1 mod 6) has distinct images modulo 2.

11.3 Constructing fractions

A *multiplicative set* S in R contains 1 and is closed under multiplication. We allow $0 \in S$. We seek a ring in which every $s \in S$ has an inverse. A proposed fraction is a pair $(a, s) \in R \times S$. We use the relation

$$(a, s) \sim (a', s') \quad \iff \quad \text{there exists } u \in S \text{ with } u(as' - a's) = 0.$$

Under a ring map $f\colon R \to A$ that sends S to units, the equation on the right implies $f(a)f(s)^{-1} = f(a')f(s')^{-1}$. Apply f and multiply by the inverses of $f(u)$, $f(s)$, and $f(s')$. The witness u lets the fraction relation record equations that become true after denominators are inverted.

Proposition 11.7. *The relation $\sim$ is an equivalence relation. Its classes, written a/s, form a ring $S^{-1}R$ with*

$$\frac{a}{s} + \frac{b}{t} = \frac{at + bs}{st}, \qquad \frac{a}{s}\frac{b}{t} = \frac{ab}{st}, \qquad -\frac{a}{s} = \frac{-a}{s}.$$

The map $\lambda\colon R \to S^{-1}R$, $a \mapsto a/1$, is a ring map and each $\lambda(s)$, $s \in S$, is a unit.

Proof. Reflexivity uses witness 1, and symmetry negates the equation. For transitivity suppose $u(as' - a's) = 0$ and $v(a's'' - a''s') = 0$. The identity

$$s'(as'' - a''s) = s''(as' - a's) + s(a's'' - a''s')$$

shows that $uvs' \in S$ witnesses $(a, s) \sim (a'', s'')$.

To verify addition and multiplication on classes, suppose u witnesses $(a, s) \sim (a', s')$ and v witnesses $(b, t) \sim (b', t')$. The cross differences for the proposed sum and product are respectively

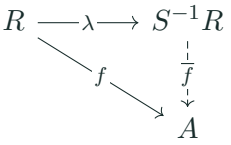
$$\begin{aligned}(at + bs)s't' - (a't' + b's')st &= (as' - a's)tt' + (bt' - b't)ss',\\ ab\,s't' - a'b'\,st &= (as' - a's)bt' + (bt' - b't)a's.\end{aligned}$$

Both expressions vanish after multiplication by uv. Negation changes the sign of a cross difference. Thus all three operations are well-defined.

Both ways of adding three fractions $a/s, b/t, c/w$ give $(atw + bsw + cst)/(stw)$, and both ways of multiplying them give $abc/(stw)$. Addition and multiplication commute by their formulas. The identities are $0/1$ and $1/1$, and the proposed negative sums with a/s to $0/s^2 = 0/1$. Multiplying a numerator and its denominator by the same member of S leaves its class unchanged, by cross equality. For distributivity, $(a/s)(b/t + c/w)$ gives $a(bw + ct)/(stw)$, while $(a/s)(b/t) + (a/s)(c/w)$ gives the same expression with numerator and denominator multiplied by s. The two classes agree.

The formulas at denominator 1 prove that λ preserves the ring operations and identities. The product of $s/1$ and $1/s$ is $s/s = 1/1$, so each $\lambda(s)$ is a unit. □

Theorem 11.8 (Universal property of localization)**.** *For every ring map $f\colon R \to A$ sending each member of S to a unit, there is a unique ring map $\overline{f}\colon S^{-1}R \to A$ satisfying $\overline{f} \circ \lambda = f$.*



Proof. Use $\overline{f}(a/s) = f(a)f(s)^{-1}$. The fraction relation ensures that this depends only on the class, as shown above. Since $f(st)^{-1} = f(s)^{-1}f(t)^{-1}$, the sum and product formulas give the sum and product of the images. The formula sends $0/1$ and $1/1$ to 0 and 1, and its composite with λ is f. Every fraction is $\lambda(a)\lambda(s)^{-1}$, so every extension must have the displayed value. □

The localization map is therefore initial among maps from R that invert S, with morphisms given by commuting triangles of ring maps. Its kernel can be read directly from the relation.

$$\ker \lambda = \{a \in R \mid ua = 0 \text{ for some } u \in S\}.$$

Indeed $a/1 = 0/1$ gives exactly that condition. For $0 \in S$ the witness 0 makes all fractions equivalent, yielding the zero ring. For an integral domain and a multiplicative set of nonzero elements, the kernel is zero. These algebraic and universal descriptions of localization also apply in the presence of zero divisors [9, Section 10.9, Tag 00CM].

11.4 Adjoining inverse elements by relations

Localization can also be computed using a free R-algebra and a quotient. For each $s \in S$, introduce a variable x_s. Form the polynomial ring $R[x_s \mid s \in S]$ and quotient by the ideal J generated by the equations $sx_s - 1 = 0$. Polynomials use only finitely many variables at a time, so this free algebra is defined for an arbitrary set S.

A ring map from the polynomial ring into A is determined by a map $f\colon R \to A$ and values $y_s \in A$ assigned to the variables. By Proposition 11.3, it descends to the quotient precisely when $f(s)y_s = 1$ for each $s \in S$. Such an assignment exists when f inverts S, and then each value is uniquely determined as $f(s)^{-1}$. The quotient has the universal property of Theorem 11.8.

The two universal maps give an explicit isomorphism with the fraction construction. It sends the variable x_s to $1/s$ and sends a/s in the opposite direction to the class of ax_s. Both composites extend the original map from R, so uniqueness in the universal property makes them identities.

11.5 Rational numbers and a zero-divisor example

In the integer ring the product of two nonzero numbers is nonzero, since its magnitude is a product of positive natural numbers. Such a product is a sum of at least one positive term. Take $S = \mathbb{Z} \setminus \{0\}$. Now every witness can be cancelled from the cross difference. Fractions agree precisely by cross multiplication, and the formulas give ordinary rational arithmetic. The map from $\mathbb{Z}$ is injective by the kernel formula, so 0 and 1 are distinct. Every nonzero fraction a/s has inverse s/a, giving the field $\mathbb{Q}$. A *field* is a nonzero commutative ring in which each nonzero element is a unit. The universal property of $\mathbb{Q}$ extends every map out of $\mathbb{Z}$ sending nonzero integers to units. For example the formula for addition gives

$$\frac{3}{4} + \frac{5}{6} = \frac{3 \cdot 6 + 5 \cdot 4}{4 \cdot 6} = \frac{38}{24} = \frac{19}{12}.$$

The last equality is an equality of classes by cross multiplication.

For comparison, take $R = \mathbb{Z}/6\mathbb{Z}$ and invert the class of 2. In the following calculation numerals denote classes in R when they occur in fractions over R. The powers of 2 form $S = \{1, 2, 4\}$. In $S^{-1}R$ we have $3/1 = 0/1$ because $2 \cdot 3 = 0$ in R. Also $1/2 = 2/1$, since the cross difference $1 - 4 = -3$ is annihilated by 2. The witness is needed here because the cross difference is a nonzero class in R.

Reduction modulo 3 gives a ring map $\rho\colon R \to \mathbb{Z}/3\mathbb{Z}$ sending each member of S to a unit. It extends to $\overline{\rho}\colon S^{-1}R \to \mathbb{Z}/3\mathbb{Z}$. In the other direction, λ kills the ideal generated by 3. That ideal is the kernel of reduction, since the classes vanishing modulo 3 are 0 and 3. The quotient property therefore gives $g\colon \mathbb{Z}/3\mathbb{Z} \to S^{-1}R$ with $g \circ \rho = \lambda$. The composite $\overline{\rho} \circ g$ agrees with the identity after precomposition with the surjection ρ, so it is the identity. The composite $g \circ \overline{\rho}$ agrees with the identity after precomposition with λ, so uniqueness in Theorem 11.8 makes it the identity as well. Thus inverting 2 gives the ring $\mathbb{Z}/3\mathbb{Z}$.

The Chinese remainder coordinates explain the calculation. In $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ the element 2 has coordinates $(0, 2)$. Making it invertible forces the first idempotent $(1, 0)$ to vanish, since its product with $(0, 2)$ is zero. The second coordinate already supplies the inverse and survives intact. Later, change of scalars will use the map $R \to S^{-1}R$ to make the same designated elements act invertibly on linear data.

CHAPTER 12

Ring and algebra actions on modules

Chapter 9 described a module as an action of a ring on an abelian group. We now ask what those actions and their maps can construct. The action of R on itself represents the choice of one element of a module. Taking a sum of copies of that action represents a choice of one element for each generator. The result is the free module, which supplies the starting object for any presentation by linear relations.

12.1 Module maps and regular actions

Write R-**Mod** for the category of unital left R-modules. Its arrows are additive maps $f\colon M \to N$ with $f(rx) = rf(x)$ for every scalar r and element x. In the description as an additive functor $BR \to \mathbf{Ab}$, this equation is naturality at the arrow r of BR. It ensures that the composite of two module maps is a module map. The set $\mathrm{Hom}_R(M, N)$ of such maps is an abelian group under pointwise addition, since $(f + g)(rx) = r(f(x) + g(x))$. Composition distributes over this addition in both variables. Thus the module category is preadditive, like **Ab** in Section 9.8.1.

The *left regular module* is R with action given by multiplication on the left. For any left module M, evaluation at 1_R gives a bijection

$$\mathrm{Hom}_R(R, M) \longrightarrow M.$$

Given $m \in M$, the corresponding map is $r \mapsto rm$. This rule respects addition and the scalar action by the module laws. Every module map φ has $\varphi(r) = \varphi(r1_R) = r\varphi(1_R)$, proving uniqueness. Postcomposing with a map $M \to N$ sends the representing element m to its image in N, so the bijection is natural in the module. In the language of Chapter 4, the covariant functor that takes a module to its set of elements is represented by R, with universal element 1_R.

Taking $M = R$ identifies each left-module endomorphism with right multiplication by an element. Write $T_b(x) = xb$. Then

$$(T_b \circ T_c)(x) = (xc)b = x(cb), \qquad T_b + T_c = T_{b+c}.$$

The resulting ring isomorphism is $\mathrm{End}_R(R) \cong R^{\mathrm{op}}$. The order reversal comes from applying the right multiplications in succession. The left scalar action instead

gives the map $a \mapsto L_a$, with $L_a(x) = ax$, into the ring of additive endomorphisms. It determines a from $L_a(1_R) = a$. A left multiplication L_a is R-linear exactly when a is central. Indeed, R-linearity would give $ax = xL_a(1_R) = xa$ for all x, and this equation identifies L_a with the linear map T_a.

12.2 Two commuting scalar actions

A right S-module structure on M is a left action of S^{op}. An *R-S-bimodule* is an abelian group with left R- and right S-actions satisfying

$$r(xs) = (rx)s \qquad (r \in R,\ s \in S,\ x \in M).$$

Equivalently, the two ring maps $R \to \mathrm{End}_{\mathbf{Ab}}(M)$ and $S^{\mathrm{op}} \to \mathrm{End}_{\mathbf{Ab}}(M)$ have commuting images. Writing $L_r(x) = rx$ and $T_s(x) = xs$, the compatibility is the following square of additive maps.

$$\begin{array}{ccc} M & \xrightarrow{\ L_r\ } & M \\ {\scriptstyle T_s}\downarrow & & \downarrow{\scriptstyle T_s} \\ M & \xrightarrow{\ L_r\ } & M \end{array}$$

This says each right operator is an R-linear endomorphism, so the right action is equivalently a unital ring map $S^{\mathrm{op}} \to \mathrm{End}_R(M)$.

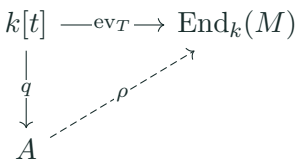
Example 12.1 (Rectangular matrices as a bimodule)**.** For a commutative ring k, take the additive group of m-by-n matrices over k. Square matrices of size m act by multiplication on the left, and square matrices of size n act by multiplication on the right. The dimensions make each product rectangular of the original size. The identities $A(XB) = (AX)B$ and $I_m X I_n = X$ give the commuting actions and their units. In particular a column of height m carries a left action of the m-by-m matrix ring and a commuting right scalar action of k.

The order convention will recur when we change scalars. A tensor product pairs a right action on one factor with a left action on the other. The bimodule condition says that a further action on the first factor commutes with the identifications used in forming that tensor product.

12.3 An algebra action specified by an operator

Let k be commutative and let M be a k-module. The polynomial algebra $k[t]$ acts on M by choosing a k-linear endomorphism T to be the action of t. The polynomial $\sum_i c_i t^i$ then acts as $\sum_i c_i T^i$. The free associative algebra on one generator is $k[t]$, since its words are the powers t^i. Its universal property therefore

applies to the endomorphism algebra. Impose the equation $t^2 = 0$ and write $A = k[t]/(t^2)$. By the universal property of a quotient, the action factors through A precisely when $T^2 = 0$.



For two such actions, with operators T_M, T_N, a k-linear map $u\colon M \to N$ is A-linear precisely when $uT_M = T_N u$. The necessity is the scalar compatibility at t. For sufficiency, compatibility extends to every power and every k-linear combination of powers. Thus a generator and relation describe the actions, and one intertwining equation describes their maps.

Example 12.2 (A two-step action)**.** On k^2, take $T(x, y) = (y, 0)$. Then $T^2 = 0$. In the coordinates (x, y), multiplication by t on A has exactly this form under $(x, y) \mapsto y + xt$. This is an isomorphism of k-modules because reducing modulo t^2 gives a unique expression $y + xt$. The coordinate swap in the formula puts the t-coefficient first.

Every k-linear map on k^2 is a matrix, whose columns are its values on the two coordinate vectors. If $U = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, then

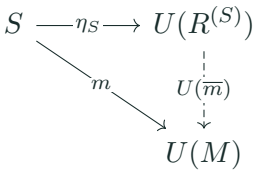
$$UT = \begin{pmatrix} 0 & a \\ 0 & c \end{pmatrix}, \qquad TU = \begin{pmatrix} c & d \\ 0 & 0 \end{pmatrix}.$$

Equality gives $c = 0$ and $d = a$. The resulting endomorphisms are $a\,\mathrm{id} + bT$. Composition multiplies the corresponding classes $a + bt$ in A. This calculation recovers the regular module endomorphism ring in concrete coordinates. For this commutative algebra, A^{op} and A coincide.

12.4 Free modules from arbitrary assignments

For a set S, define $R^{(S)}$ to be the set of functions $S \to R$ of finite support, with addition and left scalar multiplication pointwise. The support of a function is the set where it takes a nonzero value. For $s \in S$, let e_s be 1_R at s and 0 elsewhere. Every vector can be written as a finite sum $\sum_s r_s e_s$, and its coordinates determine the coefficients.

Proposition 12.3 (The free module property)**.** *For every set S and left R-module M, an arbitrary assignment $m\colon S \to U(M)$ extends uniquely to an R-linear map $\overline{m}\colon R^{(S)} \to M$ with $\overline{m}(e_s) = m(s)$. Here U takes a module to its underlying set.*



$\eta_S(s) = e_s.$

Proof. There is one candidate, given by $\overline{m}(\sum_s r_s e_s) = \sum_s r_s m(s)$. The sums are finite, and zero coefficients contribute zero. For addition, take the union of the two supports and add coordinates before applying the rule. The distributive law gives the sum of the two evaluations. For a scalar r, the rule evaluates $\sum_s (r r_s) e_s$ as $\sum_s (r r_s) m(s) = r \sum_s r_s m(s)$, so the map is linear. It takes e_s to $m(s)$, and linearity forces its value on every vector. □

For a function $a\colon S \to T$, extend $s \mapsto e_{a(s)}$ to a linear map $F(a)\colon R^{(S)} \to R^{(T)}$. Sending basis vectors through two such maps agrees with applying the composite function and extending. The identity map acts by the identity on the generators. Thus $F(S) = R^{(S)}$ defines a functor. Restricting a map out of $F(S)$ to its generators gives

$$\mathrm{Hom}_R(F(S), M) \cong \mathbf{Set}(S, U(M)).$$

Changing S by a function precomposes the assignment, and changing M by a linear map postcomposes with its underlying function. The bijection commutes with both changes. This proves $F \dashv U$.

Example 12.4 (Combining two generators)**.** Let a send the two elements of $\{p, q\}$ to the sole element of $\{s\}$. Then $F(a)$ sends both e_p and e_q to e_s, so it takes $r e_p + r' e_q$ to $(r + r') e_s$. Identifying generators by a map of sets adds the corresponding coefficients under free linear extension.

12.5 Direct sums and biproducts

The product of modules M_i consists of all tuples with the componentwise operations. The projection maps determine a linear map into a product, since prescribing the component maps gives a unique map to the tuples and each component is linear. The *direct sum* $\bigoplus_i M_i$ consists of tuples with finite support and has one map $j_i\colon M_i \to \bigoplus_i M_i$ that puts an element in its own coordinate. A family $f_i\colon M_i \to N$ extends to the sum by $(m_i) \mapsto \sum_i f_i(m_i)$. The sum is finite, and the formula is unique because each tuple is a finite sum of the $j_i(m_i)$. It is linear by linearity of the f_i. Thus the direct sum is a coproduct. In particular $R^{(S)} = \bigoplus_{s \in S} R$.

For finitely many modules the product and direct sum are the same tuple module. In the binary case the four maps are

$$M \underset{p_M}{\overset{j_M}{\rightleftarrows}} M \oplus N \underset{j_N}{\overset{p_N}{\rightleftarrows}} N$$

On their own coordinates the composites are identities, $p_M j_M = \mathrm{id}_M$ and $p_N j_N = \mathrm{id}_N$. On the other coordinates the composites are zero, $p_N j_M = 0$ and $p_M j_N = 0$. The equality $j_M p_M + j_N p_N = \mathrm{id}_{M \oplus N}$ says that one recovers a tuple by adding its coordinate tuples. An object carrying both these universal properties with these maps is called a *biproduct*. The zero module is the empty biproduct. It has exactly one map to and from every module.

The same identity expresses a linear map between finite direct sums in components. For $f\colon \bigoplus_i M_i \to \bigoplus_j N_j$, the component from M_i to N_j is $p_j f j_i$. If g follows f, inserting $\sum_j j_j p_j$ between them shows that the components of gf are sums of composites of components. When every summand is R and R is commutative, this is ordinary matrix multiplication. The next chapter uses such maps to impose linear relations and to compute the modules they present.

CHAPTER 13

Linear maps and module presentations

A list of linear formulas can specify either equations to solve or relations to impose. For a linear map $f\colon M \to N$, the kernel consists of the inputs sent to zero. The cokernel is a quotient of the target in which the entire image of f becomes zero. Their universal properties are the equalizer and coequalizer conditions of Chapter 3 applied to f and the zero map. Taken together, they explain how a list of generators and relations determines a module.

13.1 Matrices and chosen bases

Let R be a commutative ring. Write the elements of R^n as columns and let e_j denote the basis column whose j-th entry is 1. A linear map $f\colon R^n \to R^m$ is specified by the images of these basis elements. If $f(e_j) = \sum_{i=1}^{m} a_{ij} e'_i$ in the chosen basis of the target, then $A = (a_{ij})$ represents f. In coordinates,

$$f(x) = Ax, \qquad (Ax)_i = \sum_{j=1}^{n} a_{ij} x_j.$$

Indeed, each input is $\sum_j x_j e_j$, so linearity gives the formula. Conversely the formula defines a linear map for every matrix of size m by n. Composing linear maps multiplies their matrices in the same order as function composition. To see this, apply the second matrix to each column of the first, which gives the columns of the composite.

For example, the matrix $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ sends (x, y) to $(x + 2y, 3x + 4y)$. The first equation $x + 2y = 0$ expresses a condition on an input of this map. The first column $(1, 3)$ can instead be used as a relation between two generators in its target. These are different uses of the same matrix, and the kernel and cokernel distinguish their roles.

13.2 Kernels and cokernels

Let $f\colon M \to N$ be a map of left R-modules, allowing R to be noncommutative in this section. For every pair of modules there is a zero map, the homomorphism

with constant value zero. The subsets

$$\ker f = \{x \in M \mid f(x) = 0\}, \qquad \operatorname{im} f = \{f(x) \mid x \in M\}$$

are submodules. For the kernel this follows because f preserves sums and scalar multiples. For the image, $f(x) + f(x') = f(x + x')$ and $rf(x) = f(rx)$ give witnesses for closure under the two operations. Both subsets contain zero and are closed under additive inverses.

For any submodule $L \subseteq N$, put $y \sim y'$ when $y - y' \in L$. This is an equivalence relation by the additive subgroup laws. Addition and scalar multiplication respect the relation, since the difference of two sums of equivalent representatives is a sum in L, and $r(y - y') \in L$. Thus the set N/L of equivalence classes inherits a module structure. Write $y + L$ for the class of y. The quotient map $y \mapsto y + L$ is linear and surjective, with kernel L because $y + L = 0 + L$ precisely when $y \in L$.

Proposition 13.1. *The inclusion $i\colon \ker f \to M$ is an equalizer of f and 0. The quotient $\pi\colon N \to N/\operatorname{im} f$ is a coequalizer of f and 0.*

$$\ker f \xhookrightarrow{\ i\ } M \underset{0}{\overset{f}{\rightrightarrows}} N \xrightarrow{\ \pi\ } N/\operatorname{im} f$$

Proof. The inclusion satisfies $f \circ i = 0$. If $h\colon T \to M$ satisfies $f \circ h = 0$, each $h(t)$ belongs to $\ker f$. The same formula with codomain $\ker f$ defines a linear map $T \to \ker f$ because the operations there are those of M. It is the unique map whose composite with i is h.

For the quotient, $\pi \circ f = 0$ because $f(x)$ lies in $\operatorname{im} f$. Suppose $k\colon N \to T$ satisfies $k \circ f = 0$. Then k kills every element of $\operatorname{im} f$. The assignment $y + \operatorname{im} f \mapsto k(y)$ is well-defined because equivalent representatives differ by an element killed by k. The assignment preserves the module operations and defines a factorization through π. Surjectivity of π makes this factorization unique. □

The quotient object so characterized is written $\operatorname{coker} f$. The same two constructions also describe an arbitrary parallel pair $f, g\colon M \to N$. The difference $f - g$ is linear, with value $f(x) - g(x)$ at x. Then $f \circ h = g \circ h$ is equivalent to $(f - g) \circ h = 0$, and $k \circ f = k \circ g$ is equivalent to $k \circ (f - g) = 0$. Consequently the equalizer of the pair is $\ker(f - g)$ and the coequalizer is $\operatorname{coker}(f - g)$.

In the category of R-modules, monomorphisms are precisely the injective maps. An injective map cancels by comparing the values of the two maps being composed with it. Conversely, suppose f is monic and $x \in \ker f$. The map $R \to M$ given by $r \mapsto rx$ and the zero map have the same composite with f. They agree by monicity, and evaluating at 1 gives $x = 0$. Thus the kernel is

zero. If $f(x) = f(x')$, the difference $x - x'$ lies in the zero kernel, which proves injectivity.

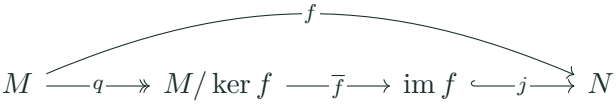
Epimorphisms are precisely the surjective maps. A surjection cancels because two maps agreeing on its image agree on the whole codomain. If f is epic, its cokernel projection π must equal the zero map, since $\pi \circ f = 0 \circ f$. Each $y \in N$ therefore has zero class in $N/\operatorname{im} f$, giving $y \in \operatorname{im} f$. The image is all of N, as required.

13.3 Image and coimage

Every map factors through its image as a surjection followed by an inclusion. A second factorization begins with the quotient $M/\ker f$. This first quotient is called the *coimage* of f, since it is the cokernel of the kernel inclusion. Dually the kernel of the cokernel projection is the image. The two constructions meet at the canonical map

$$\overline{f}\colon M/\ker f \longrightarrow \operatorname{im} f, \qquad \overline{f}(x + \ker f) = f(x).$$

Equivalence of representatives here is exactly equality of their outputs under f, since $x - x' \in \ker f$ if and only if $f(x) - f(x') = 0$. Thus $\overline{f}$ is well-defined and injective. Each element of $\operatorname{im} f$ has a representative by definition, so $\overline{f}$ is surjective. It preserves the module operations, and its inverse does as well because a bijective linear map carries an operation on its inputs to the corresponding operation on its outputs. This proves the image–coimage isomorphism.



The image is the least submodule of N through which f factors. In terms of composition, suppose $f = j' \circ g$ with $j'\colon J \to N$ monic. Each element of $\operatorname{im} f$ lies in the image of j', because $f(x) = j'(g(x))$. There is a unique map $h\colon \operatorname{im} f \to J$ sending $f(x)$ to $g(x)$. If two inputs have the same image under f, the two values of g agree by injectivity of j'. Thus h is well-defined, and it is linear because sums and scalar multiples of representatives are representatives of the corresponding operations. It satisfies $j' \circ h = j$, and this equation forces h by monicity.

The factorization respects commutative squares. If $a\colon M \to M'$ and $b\colon N \to N'$ satisfy $b \circ f = f' \circ a$, then a sends $\ker f$ into $\ker f'$, and b sends $\operatorname{im} f$ into $\operatorname{im} f'$. Thus a induces a map on coimages and b restricts to a map on images. The resulting square with horizontal maps $\overline{f}, \overline{f'}$ commutes by evaluation on a representative x of a class. Both paths give $b(f(x)) = f'(a(x))$. So the isomorphism is compatible with changes of source and target that preserve the original linear map.

13.4 Exactness and rank–nullity

The sequence $L \xrightarrow{u} M \xrightarrow{v} N$ is *exact at* M when $\operatorname{im} u = \ker v$. The inclusion $\operatorname{im} u \subseteq \ker v$ is precisely the equation $v \circ u = 0$. The reverse inclusion says that each v-zero in M is the image of some element of L. A *short exact sequence*

$$0 \longrightarrow L \xrightarrow{i} M \xrightarrow{p} N \longrightarrow 0$$

is exact at L, M, N, where 0 denotes the zero module. At L, exactness says that i is injective, since the image of the map from 0 is zero. At N, it says that p is surjective, since the next map kills all of N. In particular, the image–coimage factorization of $f\colon M \to N$ gives short exact sequences

$$0 \longrightarrow \ker f \longrightarrow M \longrightarrow \operatorname{im} f \longrightarrow 0,$$
$$0 \longrightarrow \operatorname{im} f \longrightarrow N \longrightarrow \operatorname{coker} f \longrightarrow 0.$$

Over a field k, a short exact sequence of vector spaces with finite-dimensional middle term can be split using a basis of its last term. Indeed, N is finite-dimensional because the images of a finite spanning set of M span N. For each basis vector n_j of N, choose a lift m_j in M with $p(m_j) = n_j$. Extend $n_j \mapsto m_j$ linearly to a map $\sigma\colon N \to M$. Then $p \circ \sigma = \mathrm{id}_N$ because the two maps agree on the basis. Such a σ is called a section of p. Every $m \in M$ has a decomposition

$$m = i(\ell) + \sigma(n), \qquad n = p(m).$$

The difference $m - \sigma(p(m))$ lies in $\ker p = \operatorname{im} i$. It therefore supplies ℓ, and injectivity of i makes ℓ unique. The map $(\ell, n) \mapsto i(\ell) + \sigma(n)$ is a linear isomorphism $L \oplus N \to M$. Here the finite direct sum $L \oplus N$ has pairs with componentwise operations. Taking bases of the summands gives $\dim M = \dim L + \dim N$.

Apply this decomposition to the first short exact sequence for a linear map f whose domain M is finite-dimensional. The nullity is the dimension of the kernel, and the rank is the dimension of the image. Thus

$$\dim M = \dim \ker f + \dim \operatorname{im} f.$$

Example 13.2. For $f\colon k^3 \to k^2$ given by $f(x, y, z) = (x + y, y + z)$, the formula $\sigma(u, v) = (u, 0, v)$ gives a section. Solving the two kernel equations gives exactly the vectors $(-t, t, -t)$. The decomposition along the kernel and the lifted copy of k^2 is therefore

$$(x, y, z) = (-y, y, -y) + (x + y, 0, y + z).$$

The summands are individually characterized by the value of $f(x, y, z)$ and the remaining difference. Their dimensions are one and two, giving $3 = 1 + 2$.

The proof obtains the dimension identity from exactness and the ability to choose a section. Over the integers, consider the exact sequence

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\times 2} \mathbb{Z} \longrightarrow \mathbb{Z}/2\mathbb{Z} \longrightarrow 0$$

A section of its quotient map would send the class of 1 to an integer killed by 2. The only such integer is zero, whose class is zero, a contradiction.

13.5 Generators, relations, and maps out

A presentation of an R-module M consists of a set S of generators, a set T indexing relations, and a diagram

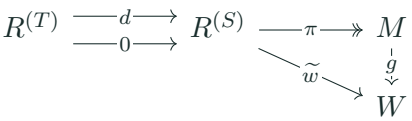
$$R^{(T)} \xrightarrow{d} R^{(S)} \xrightarrow{\pi} M \longrightarrow 0$$

in which π exhibits M as the cokernel of d. Here $R^{(S)}$ is the free module on S, with each element a finite R-linear combination of its basis vectors e_s. Thus the column $d(e_t) = \sum_{s \in S} r_{st} e_s$ specifies a relation with finite support. In M, that column becomes zero. The submodule $\operatorname{im} d$ is generated by the columns, since linearity gives $d(\sum_t a_t e_t) = \sum_t a_t d(e_t)$. This accounts for all consequences of the named relations under addition and scalar multiplication.

A map $g\colon M \to W$ amounts to a choice of elements $w_s \in W$ satisfying

$$\sum_{s \in S} r_{st} w_s = 0 \qquad (t \in T).$$

To prove the correspondence, extend the assignment $s \mapsto w_s$ uniquely to a linear map $\widetilde{w}\colon R^{(S)} \to W$ by freeness. The displayed equations are precisely $\widetilde{w}(d(e_t)) = 0$ for all t, which by linearity is equivalent to $\widetilde{w} \circ d = 0$. The cokernel property now gives a unique g with $g \circ \pi = \widetilde{w}$. Conversely a map g gives back the assignment $w_s = g(\pi(e_s))$. These correspondences are inverse and commute with postcomposition by a map out of W.



Every module admits such a presentation. Take S to be the underlying set of M and extend $e_s \mapsto s$ linearly. The resulting π is surjective. Take T to be the underlying set of $\ker \pi$, and send the basis vector indexed by $t \in T$ to $t \in R^{(S)}$. Its linear extension d has image contained in the kernel because the kernel is a submodule. Every element of the kernel is the image of its own basis vector under d, so the two submodules coincide. The image–coimage isomorphism for π identifies $R^{(S)}/\operatorname{im} d$ with M. A presentation with both sets finite is called a finite presentation.

13.6 An integer presentation in two coordinates

Example 13.3 (Two integer relations)**.** Take the two generators of $\mathbb{Z}^2$ and impose the relations $2e_1 + 4e_2 = 0$ and $4e_1 + 2e_2 = 0$. The presentation map has these relations as its columns, so it is represented by

$$A = \begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix}, \qquad M = \operatorname{coker} A.$$

Subtract twice the first row from the second, subtract twice the first column from the second, and negate the second row. The successive matrices are

$$\begin{pmatrix} 2 & 4 \\ 0 & -6 \end{pmatrix}, \qquad \begin{pmatrix} 2 & 0 \\ 0 & -6 \end{pmatrix}, \qquad D = \begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix}.$$

The row operations multiply on the left by $U = \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix}$. The column operation multiplies on the right by $V = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$. Thus $D = UAV$. The matrix U is its own inverse, and the inverse of V has upper right entry 2. In particular, each operation is invertible over $\mathbb{Z}$.

$$\begin{array}{ccc} \mathbb{Z}^2 & \xrightarrow{\ A\ } & \mathbb{Z}^2 \\ \downarrow{\scriptstyle V^{-1}} & & \downarrow{\scriptstyle U} \\ \mathbb{Z}^2 & \xrightarrow{\ D\ } & \mathbb{Z}^2 \end{array}$$

Surjectivity of V gives $\operatorname{im} D = U(\operatorname{im} A)$. Consequently $[x] \mapsto [Ux]$ identifies the two cokernels, with inverse induced by U^{-1}. The image of D consists of the pairs in $2\mathbb{Z} \oplus 6\mathbb{Z}$. Reduction modulo 2 in the first coordinate and modulo 6 in the second has this image as its kernel, so the presentation gives

$$M \cong \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/6\mathbb{Z}.$$

Column operations change the generating list of relations and preserve its span. Row operations change coordinates in the module of generators, and the induced quotient map uses that change.

For a map into $W = \mathbb{Z}/12\mathbb{Z}$, let a, b be the images of the two diagonal generators. The conditions are $2a = 0$ and $6b = 0$. Thus a is 0 or 6, and b is any even residue. To return to the original generators, apply U to their basis columns. The resulting images are $a + 2b$ and $-b$, respectively. Substitution gives $2(a + 2b) + 4(-b) = 2a = 0$ and $4(a + 2b) + 2(-b) = 4a + 6b = 0$, verifying the original relations. There are twelve assignments. The invertible change of coordinates and the universal property account for every map to W.

13.7 Counting spanning trees by a determinant

For a finite undirected simple graph with vertex set V and edge set E, a *spanning tree* is a connected, acyclic choice of edges on all vertices. Suppose V is nonempty, orient every edge, and form the incidence map $\partial\colon \mathbb{Z}^E \to \mathbb{Z}^V$, sending an edge basis vector to the target basis vector minus the source basis vector. Each column has coordinate sum zero, so ∂ factors through the kernel of the coordinate-sum map $\mathbb{Z}^V \to \mathbb{Z}$. Deleting the coordinate at a chosen root r identifies that kernel with $\mathbb{Z}^{V\setminus\{r\}}$, since the deleted coordinate is forced by the others. Put $k = |V| - 1$, and call the resulting k-row incidence matrix B. The determinant of BB^{T} counts the spanning trees.

To see how a determinant counts discrete choices, expand each column of BB^{T} as a linear combination of the columns of B. The determinant is multilinear and alternating in its columns, as described in Chapter 14. A tuple that uses the same edge twice contributes zero. For a fixed subset $S \subseteq E$ of size k, the surviving contributions come from orderings of S. Summing their signed coefficients gives the determinant of the transpose of the submatrix B_S on those columns. The Cauchy–Binet expansion in this case is

$$\det(BB^{\mathsf{T}}) = \sum_{\substack{S \subseteq E \\ |S| = k}} \det(B_S)^2.$$

If S contains a cycle, the signed sum of its cycle-edge columns vanishes, so $\det(B_S) = 0$. If S has no cycle, its k edges make it a tree. Each edge of a forest joins two previously separate components, and k such joins leave one component. If the tree has at least two vertices, choose a leaf other than r. Its row in B_S has just one nonzero entry, $+1$ or -1. Expanding along that row reduces the determinant up to sign to the same calculation for the leaf-deleted tree. A one-vertex tree has the empty determinant 1, so $\det(B_S)^2 = 1$ for every spanning tree. Equivalently the spanning trees are the subsets of k edges whose incidence columns form an integer basis of the coordinate-sum kernel. This proves the matrix–tree theorem [68, Lecture 28] and explains why the answer is independent of orientations and root.

For the complete graph on four vertices, BB^{T} has diagonal entries 3 and off-diagonal entries -1. It acts by 1 on the all-ones line and by 4 on the two-dimensional space of vectors whose coordinates sum to zero. Its determinant is $1 \cdot 4^2 = 16$. The same calculation for the complete graph on $n \geq 2$ vertices gives n^{n-2} spanning trees.

CHAPTER 14

Tensor products, duals, and determinants

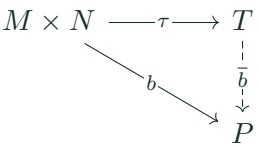
The entries of a matrix product are linear in the entries of each matrix when the other matrix is held fixed. The pairing of a vector with a linear functional has the same property. Tensor products turn such rules with two inputs into linear maps from a single module, using presentations as in Chapter 13. Exterior powers add equations to the tensor construction and give a module on which an endomorphism acts by its determinant.

Throughout this chapter R is a commutative ring with unit, and all modules and linear maps are over R unless a different ring is specified.

14.1 A module representing bilinear maps

Definition 14.1. A map $b\colon M \times N \to P$ is *bilinear* if each of the two maps obtained by fixing one input is R-linear. A *tensor product* of M and N is a module T and a bilinear map $\tau\colon M \times N \to T$ such that every bilinear map $b\colon M \times N \to P$ factors uniquely as $\overline{b} \circ \tau$ for a linear map $\overline{b}\colon T \to P$.

The diagram uses the underlying sets. The dashed arrow is linear, and the other two arrows are bilinear maps of pairs.



In terms of maps out of T, this says that composition with τ gives a bijection

$$\mathrm{Hom}_R(T, P) \cong \mathrm{Bil}_R(M, N; P),$$

where the right side is the set of bilinear maps to P. The bijection commutes with a linear map $P \to P'$ by associativity of composition. Thus T represents the rule assigning to each target the bilinear maps into that target.

Proposition 14.2. *Every pair of modules M, N has a tensor product.*

Proof. Let F be the free R-module on the set $M \times N$. Write $[x, y]$ for the basis element indexed by (x, y). Let L be the submodule generated by all elements of the following forms, as x, x' range over M, y, y' range over N, and r ranges over R.

$$[x + x', y] - [x, y] - [x', y], \qquad [x, y + y'] - [x, y] - [x, y'],$$
$$[rx, y] - r[x, y], \qquad [x, ry] - r[x, y].$$

Write $M \otimes_R N = F/L$ and $x \otimes y$ for the class of $[x, y]$. The relations make $(x, y) \mapsto x \otimes y$ bilinear. They include the balancing equality

$$(rx) \otimes y = r(x \otimes y) = x \otimes (ry).$$

Given b as in the definition, the assignments $[x, y] \mapsto b(x, y)$ extend uniquely to a linear map $F \to P$ by freeness. This map sends all the displayed generators of L to zero by bilinearity. It therefore descends uniquely to a linear map $F/L \to P$ with the prescribed values on $x \otimes y$. □

Elements $x \otimes y$ are called *pure tensors*. Since the elements $[x, y]$ generate F, every element of the tensor product is a finite linear combination of pure tensors. Accordingly, two linear maps from $M \otimes_R N$ agree if their values agree on pure tensors. This gives a direct test for the identities below.

14.2 Changing inputs and regrouping tensors

For linear maps $f\colon M \to M'$ and $g\colon N \to N'$, the rule $(x, y) \mapsto f(x) \otimes g(y)$ is bilinear. It determines a linear map $f \otimes g\colon M \otimes_R N \to M' \otimes_R N'$. Applying identities to both inputs fixes each pure tensor. Applying a second pair of maps sends $x \otimes y$ to the tensor of the two composites applied to x and y. These computations give the identity and composition laws for a functor of the two input modules.

Multiplication of a scalar on a vector gives a bilinear map $R \times M \to M$, inducing $\lambda_M\colon R \otimes_R M \to M$ with $\lambda_M(r \otimes x) = rx$. The linear map $x \mapsto 1 \otimes x$ is inverse. The composite on M sends x to x, and the composite on $R \otimes_R M$ sends a generating tensor $r \otimes x$ to $1 \otimes (rx) = r \otimes x$. Swapping inputs gives $M \otimes_R N \cong N \otimes_R M$ by bilinearity. Swapping twice fixes each generator, verifying the inverse and giving the right unit isomorphism $M \otimes_R R \cong M$ as well.

There is a canonical reassociation

$$\alpha_{M,N,L}\colon (M \otimes_R N) \otimes_R L \longrightarrow M \otimes_R (N \otimes_R L)$$

sending $(x \otimes y) \otimes z$ to $x \otimes (y \otimes z)$. Given a map $b\colon M \times N \times L \to P$ that is linear in each input, fix z. The map on the first two inputs induces a linear map $b_z\colon M \otimes_R N \to P$. The equalities $b_{z+z'} = b_z + b_{z'}$ and $b_{rz} = r b_z$ hold on pure

tensors by linearity in the last input, so they hold on all of $M \otimes_R N$. Consequently $(u, z) \mapsto b_z(u)$ is bilinear and induces a linear map $(M \otimes_R N) \otimes_R L \to P$. Its values on the generators $(x \otimes y) \otimes z$ are exactly $b(x, y, z)$. Conversely, these values give a trilinear map for each linear map from $(M \otimes_R N) \otimes_R L$. The argument with the last two inputs grouped first gives the same property for $M \otimes_R (N \otimes_R L)$. The maps $\alpha_{M,N,L}$ and its reverse follow from these two factorizations. Their composites fix the tensors of triples, which generate both bracketings.

The construction respects changes of inputs. For $f\colon M \to M'$, $g\colon N \to N'$, and $h\colon L \to L'$, both routes in the following square send a tensor of x, y, z to the same tensor of $f(x), g(y), h(z)$.

$$\begin{array}{ccc} (M \otimes_R N) \otimes_R L & \xrightarrow{\alpha_{M,N,L}} & M \otimes_R (N \otimes_R L) \\ \downarrow{\scriptstyle (f\otimes g)\otimes h} & & \downarrow{\scriptstyle f\otimes(g\otimes h)} \\ (M' \otimes_R N') \otimes_R L' & \xrightarrow{\alpha_{M',N',L'}} & M' \otimes_R (N' \otimes_R L') \end{array}$$

For four modules, each sequence of these reassociations from one bracketing to another takes the tensor of any four given vectors to the tensor of those same vectors in the new bracketing. Such tensors generate by repeated bilinearity, so all the sequences with the same endpoints give the same map. This is the coherence that allows parentheses to be suppressed in a tensor of several modules.

14.3 Tensor bases and torsion

Suppose M and N are free of finite ranks m, n, with bases $(e_i)_{i=1}^m$ and $(f_j)_{j=1}^n$. Expanding each input in its basis gives

$$\left(\sum_i x_i e_i\right) \otimes \left(\sum_j y_j f_j\right) = \sum_{i,j} x_i y_j (e_i \otimes f_j).$$

Thus the $e_i \otimes f_j$ span. The rule assigning the tuple $(x_i y_j)_{i,j}$ to the pair of vectors is bilinear with values in R^{mn}. Its linear extension sends $e_i \otimes f_j$ to the standard basis vector at (i, j). Applying it to a finite linear relation among these tensors shows that every coefficient is zero. Hence the displayed spanning family is a basis with mn vectors.

For $M = N = \mathbb{Q}^2$, the four coordinates indexed by i, j can be arranged in a matrix. A pure tensor has entries $x_i y_j$, so all of its columns are scalar multiples of the first input vector (x_1, x_2). The tensor $e_1 \otimes f_1 + e_2 \otimes f_2$ has the identity matrix as its coordinate matrix. Its two columns are independent, and the displayed sum expresses it using the minimum of two pure tensors. Thus the coordinate model also distinguishes individual tensors of vectors from their linear combinations.

A computation with torsion shows how to use the representing property even when bases are unavailable. Let $m, n > 0$ be integers and let $d = \gcd(m, n)$. A bilinear map

$$b\colon \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z} \longrightarrow P$$

to an abelian group is determined by $p = b(\overline{1}, \overline{1})$. Linearity gives $b(\overline{a}, \overline{c}) = acp$. Since each relation $m\overline{1} = 0$ or $n\overline{1} = 0$ holds in its respective factor, $mp = np = 0$. To express d as a linear combination, take the least positive integer combination d_0 of m, n among $1, \dots, m$. The list contains m itself. Dividing m and n by d_0 gives remainders that are also integer combinations. A positive remainder would be smaller than d_0, so both are zero. Thus d_0 divides both inputs, and every common divisor divides d_0. Hence $d = d_0$, which gives $dp = 0$. Conversely $dp = 0$ implies $mp = np = 0$, and the formula acp then depends only on the residue classes of a, c. It defines a bilinear map.

The map $(\overline{a}, \overline{c}) \mapsto \overline{ac}$ into $\mathbb{Z}/d\mathbb{Z}$ is itself bilinear. Each $p \in P$ killed by d determines exactly one homomorphism from $\mathbb{Z}/d\mathbb{Z}$ to P, by sending $\overline{1}$ to p. Postcomposing the displayed bilinear map with that homomorphism gives precisely the bilinear map attached to p. This factors every bilinear map uniquely, so $\mathbb{Z}/d\mathbb{Z}$ has the tensor-product property. The resulting isomorphism is

$$(\mathbb{Z}/m\mathbb{Z}) \otimes_{\mathbb{Z}} (\mathbb{Z}/n\mathbb{Z}) \cong \mathbb{Z}/d\mathbb{Z}, \qquad \overline{a} \otimes \overline{c} \longmapsto \overline{ac}.$$

For example, with $m = 12$ and $n = 18$, the tensor $\overline{4} \otimes \overline{5}$ corresponds to $\overline{2}$ in $\mathbb{Z}/6\mathbb{Z}$, because $4 \cdot 5 = 20$. For coprime m, n, the result is the zero module.

14.4 Duals and linear maps

The *dual* of M is $M^\vee = \operatorname{Hom}_R(M, R)$, with pointwise addition and scalar multiplication. The rule $(\ell, x) \mapsto \ell(x)$ is bilinear and so defines the evaluation map $M^\vee \otimes_R M \to R$. Evaluation also gives a linear map

$$M \longrightarrow M^{\vee\vee}, \qquad x \longmapsto (\ell \mapsto \ell(x)).$$

When M is finite free, the coordinate functions $e_i^\vee$ form its dual basis. Indeed, a functional ℓ has the expansion $\sum_i \ell(e_i) e_i^\vee$, as both sides agree on all e_i. This also proves linear independence by evaluation at each e_i. Under the corresponding bases on M and $M^{\vee\vee}$, the double-dual map sends each basis vector to its corresponding basis vector. It is therefore an isomorphism.

For any N, the bilinear rule

$$(\ell, y) \longmapsto (x \mapsto \ell(x) y)$$

induces a linear map $M^\vee \otimes_R N \to \operatorname{Hom}_R(M, N)$. If M is finite free, an inverse is

$$u \longmapsto \sum_i e_i^\vee \otimes u(e_i).$$

Indeed, the sum acts on $x = \sum_i x_i e_i$ by sending it to $\sum_i x_i u(e_i) = u(x)$. In the other direction, starting with $\ell \otimes y$ gives the sum $\sum_i e_i^\vee \otimes \ell(e_i) y = \ell \otimes y$ by balancing and the dual basis expansion. The pure tensors generate the domain, proving that the two linear maps are inverse.

A map $a\colon M' \to M$ acts on the dual by $\ell \mapsto \ell \circ a$, reversing the arrow. If $b\colon N \to N'$ is linear, the tensor $(\ell \circ a) \otimes b(y)$ then acts by $x \mapsto \ell(a(x))b(y)$. This is the result of precomposing the linear map represented by $\ell \otimes y$ with a and postcomposing it with b. This compatibility follows on all tensors by linearity. The map $M^\vee \otimes_R N \to \mathrm{Hom}_R(M, N)$ is therefore independent of bases and natural in both modules.

With $M = N = R^2$, the tensor $e_i^\vee \otimes e_j$ sends e_i to e_j and sends the other basis vector to zero. Its matrix has a single 1 in row j, column i. Thus the matrix $\left(\begin{smallmatrix} 2 & 3 \\ -1 & 4 \end{smallmatrix}\right)$ over $\mathbb{Z}$ corresponds to

$$2e_1^\vee \otimes e_1 - e_1^\vee \otimes e_2 + 3e_2^\vee \otimes e_1 + 4e_2^\vee \otimes e_2.$$

The coefficients record how the two column vectors of the matrix are built from the chosen output basis.

14.5 Pontryagin duality for finite modules

The additive group of a finite module also has a character dual. For this construction let R be any ring with unit and let M be a finite left R-module. Write $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ additively and put

$$D(M) = \mathrm{Hom}_{\mathbb{Z}}(M, \mathbb{T}), \qquad (\chi \cdot r)(m) = \chi(rm).$$

The scalar action makes $D(M)$ a right R-module. In fact, $((\chi \cdot r) \cdot s)(m) = \chi(r(sm)) = (\chi \cdot (rs))(m)$. Precomposition by an R-linear map is right R-linear on characters. The construction with the sides reversed sends a right module to a left module. Evaluation $m \mapsto (\chi \mapsto \chi(m))$ is left R-linear because evaluation at rm sends χ to $(\chi \cdot r)(m)$. Evaluation is an isomorphism of additive groups by Section 8.6, so it is also a module isomorphism. This gives a contravariant equivalence between finite left and finite right R-modules.

For a right R-module P, the tensor–Hom bijection takes the form

$$\mathrm{Hom}_{\mathbb{Z}}(P \otimes_R M, \mathbb{T}) \;\cong\; \mathrm{Hom}_{\text{Mod-}R}(P, D(M)).$$

A character b on the tensor product gives $p \mapsto (m \mapsto b(p \otimes m))$. This is right R-linear because $pr \otimes m = p \otimes rm$. Conversely, a right-linear map $P \to D(M)$ gives an additive, balanced pairing $P \times M \to \mathbb{T}$ by evaluation. The tensor product factors the pairing uniquely, and the pure tensors determine the resulting character.

For $R = \mathbb{Z}$ and $M = \mathbb{Z}/4\mathbb{Z}$, any map $M \to \mathbb{Z}$ sends $\overline{1}$ to an integer killed by 4, hence to zero. A character may send $\overline{1}$ to any of $0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ modulo $\mathbb{Z}$. This gives $D(M) \cong \mathbb{Z}/4\mathbb{Z}$ and shows how the choice of target detects torsion. The topology needed for character duality beyond finite groups is developed in Section 16.6.

14.6 Alternating maps and the top exterior power

An n-linear map $b\colon M^n \to P$ is *alternating* if $b(x_1, \dots, x_n) = 0$ whenever two of its arguments are equal. Swapping two arguments changes the sign of the value. For the two slots in question, put $u + v$ in both and expand by multilinearity. The two terms with repeated arguments vanish, so the terms with (u, v) and (v, u) sum to zero. This proof uses the repeated-argument equation directly and holds in characteristic two. For example, over the field $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$ the bilinear rule $(u, v) \mapsto uv$ satisfies the sign equation because multiplication commutes and $1 = -1$. Its value at $(1, 1)$ is 1, so the repeated-argument equation has additional content in this characteristic.

The *nth exterior power* $\bigwedge^n M$ represents alternating n-linear maps out of M^n. To construct it, take the free module on all ordered n-tuples of elements of M. Quotient by the relations expressing linearity in each slot and by the generators indexed by tuples with repeated entries. Write $x_1 \wedge \cdots \wedge x_n$ for the class of a tuple. The quotient relations make the map sending a tuple to its class alternating and multilinear. For any alternating n-linear map to P, extend its values from the tuple generators to the free module. Each defining relation is sent to zero, so the extension factors through the quotient. Its values on the wedge generators determine the factor uniquely. We set $\bigwedge^0 M = R$, since a map of zero arguments amounts to a choice of one element of the target.

Let M be free of rank $n \geq 1$ with basis $e_1, \dots, e_n$. The element $\omega = e_1 \wedge \cdots \wedge e_n$ spans $\bigwedge^n M$. Expanding each argument of a wedge in the basis produces a linear combination of wedges of n basis vectors. Every term with repeated basis vectors vanishes. The remaining terms use all the basis vectors once and can be reordered to ω up to a sign.

To see that ω is a basis element, write x_{ij} for the ith coordinate of the jth vector x_j. The coordinate formula

$$D(x_1, \dots, x_n) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} x_{\sigma(j), j}$$

is n-linear because each product is linear in the coordinates of each individual vector. If columns i, j agree, the terms indexed by σ and by $\sigma \circ (i\,j)$ have equal coordinate products and opposite signs. They cancel in pairs. Here the sign is determined by the parity of inversions in σ. An adjacent swap changes parity, and a swap of arbitrary positions is a composite of an odd number of adjacent swaps.

Hence D is alternating. When $x_j = e_j$ for every j, the identity permutation contributes 1 and every other term is zero. The induced map $\overline{D}\colon \bigwedge^n M \to R$ therefore sends ω to 1. It is inverse to $r \mapsto r\omega$, which is surjective by the spanning calculation. Thus $\bigwedge^n M$ is free of rank one.

14.7 Determinants from functoriality

A linear map $f\colon M \to N$ gives $\bigwedge^n f\colon \bigwedge^n M \to \bigwedge^n N$ by sending a wedge of vectors to the wedge of their images. The alternating property holds for the image rule because equal inputs to f give equal outputs. The representing property gives the linear map on exterior powers. On the generating wedges this rule composes by applying the composite to each vector, and the identity fixes each wedge. Thus $\bigwedge^n$ is a functor.

For an endomorphism f of a free module M of rank n, the map $\bigwedge^n f$ acts on the rank-one module $\bigwedge^n M$ by a scalar. Its scalar is unique, since applying the map to a generator determines its value on every element. We define this scalar to be $\det(f)$. Each basis of M gives a generator of $\bigwedge^n M$, and multiplication by the scalar has the same description with respect to every generator. In the coordinates used above, the defining square is

$$\begin{array}{ccc} \bigwedge^n M & \xrightarrow{\bigwedge^n f} & \bigwedge^n M \\ \downarrow{\scriptstyle \overline{D}} & & \downarrow{\scriptstyle \overline{D}} \\ R & \xrightarrow{r \mapsto \det(f) r} & R. \end{array}$$

Applying $\overline{D}$ to $f(e_1) \wedge \cdots \wedge f(e_n)$ gives the coordinate formula for the determinant of the matrix of f. Functoriality gives $\bigwedge^n (g \circ f) = (\bigwedge^n g) \circ (\bigwedge^n f)$. Since the composite of scalar actions multiplies their scalars, $\det(g \circ f) = \det(g)\det(f)$. The determinant of the identity equals 1. This also covers rank zero by taking $\bigwedge^0 f = \mathrm{id}_R$.

For the integer matrix $A = \left(\begin{smallmatrix} 2 & 3 \\ -1 & 4 \end{smallmatrix}\right)$ from the dual calculation, the action on $e_1 \wedge e_2$ is

$$(2e_1 - e_2) \wedge (3e_1 + 4e_2) = 8e_1 \wedge e_2 - 3e_2 \wedge e_1 = 11 e_1 \wedge e_2.$$

Let $B = \left(\begin{smallmatrix} 1 & 2 \\ 0 & 3 \end{smallmatrix}\right)$. It sends $e_1 \wedge e_2$ to $e_1 \wedge (2e_1 + 3e_2) = 3e_1 \wedge e_2$. The composite AB therefore has determinant 33, in agreement with $AB = \left(\begin{smallmatrix} 2 & 13 \\ -1 & 10 \end{smallmatrix}\right)$ and the scalar $2 \cdot 10 - 13 \cdot (-1)$.

The tensor construction also changes the ring of scalars in the next chapter. A map of commutative rings $\varphi\colon R \to S$ makes S an R-module. In $S \otimes_R M$, the balancing relation $(s\varphi(r)) \otimes x = s \otimes (rx)$ makes an R-scalar on x agree with its image acting on the S-factor. This is the relation needed to extend an R-linear map into an S-module by an S-linear map.

CHAPTER 15

Restriction and extension of scalars

A system of relations with integer coefficients can be interpreted over the rationals or reduced modulo a prime. Both operations take a matrix to the matrix obtained by changing its coefficients. The resulting module is characterized by a test on maps. It receives a map from the original module so that every compatible map to a module over the new ring extends uniquely.

Let $f\colon R \to S$ be a homomorphism of commutative unital rings. Throughout the chapter S itself is an R-module by $r \cdot s = f(r)s$. We write $\mathrm{Hom}_R(M, N)$ for the set of R-linear maps $M \to N$ and use the analogous notation over S.

15.1 Restriction and extension

An S-module N already has an action of R. Define N_R to have the same additive group as N and the action

$$r \cdot n = f(r)n.$$

The laws of this action follow from the ring map laws for f and the module laws over S. An S-linear map $v\colon N \to N'$ preserves multiplication by each $f(r)$, so the same map gives an R-linear map $v_R\colon N_R \to N'_R$. This operation leaves the identity and composition of the underlying maps intact. It is a functor from S-modules to R-modules called *restriction of scalars*.

To go in the other direction, start with an R-module M. Its *extension of scalars* is the tensor product $M_S = S \otimes_R M$ with the action

$$s_0 \cdot (s \otimes m) = (s_0 s) \otimes m.$$

Multiplication by s_0 on S is R-linear. Thus this action is well-defined on the tensor product, as follows by applying the tensor universal property to $(s, m) \mapsto (s_0 s) \otimes m$. The module laws hold on pure tensors by the ring laws of S and hence on the module they generate. An R-linear map $u\colon M' \to M$ induces an S-linear map $u_S\colon M'_S \to M_S$ with $u_S(s \otimes m') = s \otimes u(m')$. Checking these formulas on pure tensors gives the functor laws.

There is a canonical map $\eta_M\colon M \to (M_S)_R$ given by $\eta_M(m) = 1 \otimes m$. It is additive. Its compatibility with R follows from balancing in the tensor product,

$$1 \otimes rm = f(r) \otimes m = f(r)(1 \otimes m).$$

Composing with η_M tests any proposed extension to an S-linear map.

Proposition 15.1 (Extension of scalars is left adjoint to restriction)**.** *For every R-module M and S-module N, there is a natural bijection*

$$\mathrm{Hom}_S(S \otimes_R M, N) \cong \mathrm{Hom}_R(M, N_R).$$

Proof. Given $g\colon M \to N_R$, use the map $(s, m) \mapsto sg(m)$. It is additive in both variables. Multiplication by r on either input multiplies the result by $f(r)$, since

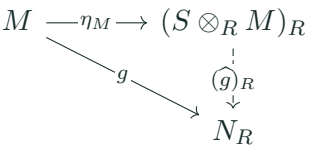
$$(sf(r))g(m) = f(r)(sg(m)) = sg(rm).$$

The tensor property for this R-bilinear map thus gives $\widehat{g}\colon S \otimes_R M \to N_R$ with $\widehat{g}(s \otimes m) = sg(m)$. On pure tensors, multiplying the first factor by s_0 multiplies the result by s_0. This proves S-linearity since the pure tensors generate.

In the reverse direction send $h\colon S \otimes_R M \to N$ to $m \mapsto h(1 \otimes m)$. This is R-linear by the property of η_M. Starting with g, the two steps return g because $\widehat{g}(1 \otimes m) = g(m)$. Starting with h, note that $h(s \otimes m) = sh(1 \otimes m)$ by S-linearity. Consequently the two steps return h on every pure tensor and hence on the whole module.

For $u\colon M' \to M$ and $v\colon N \to N'$, both $\widehat{v_R \circ g \circ u}(s \otimes m')$ and $(v \circ \widehat{g} \circ u_S)(s \otimes m')$ are $s\, v(g(u(m')))$. This proves naturality in both variables. □

The diagram takes place over R. Its dashed arrow is the restriction of the unique S-linear map extending g.



The unit of the adjunction is η_M. Its counit at N is $S \otimes_R N_R \to N$, $s \otimes n \mapsto sn$, obtained by extending the identity of N_R. The Stacks Project treats this base change construction [9, Section 10.14, Tag 05G3].

15.2 Coextension of scalars

The same restriction functor has a right adjoint. For an R-module M, take the R-linear maps out of S,

$$C_f(M) = \mathrm{Hom}_R(S, M).$$

Pointwise addition gives an abelian group. Define an S-action on it by the formula

$$(s \cdot h)(t) = h(ts).$$

Multiplication by s on S is R-linear, so $s \cdot h$ is an R-linear map. Additivity in s follows from the distributive law in S and additivity of h. Additivity in h follows from pointwise addition. The unit acts as the identity, and

$$(s_1 \cdot (s_2 \cdot h))(t) = h((ts_1)s_2) = ((s_1 s_2) \cdot h)(t).$$

This gives the S-module $C_f(M)$, called the *coextension of scalars*. Postcomposition with an R-linear map $w\colon M \to M'$ is S-linear because $w(h(ts)) = (w \circ h)(ts)$, and it respects identities and composition. Thus coextension is a functor.

Proposition 15.2 (Restriction is left adjoint to coextension)**.** *For every S-module N and R-module M, there is a natural bijection*

$$\mathrm{Hom}_S(N, C_f(M)) \cong \mathrm{Hom}_R(N_R, M).$$

Proof. Given $l\colon N_R \to M$, set $\widetilde{l}(n)(t) = l(tn)$. As a function of t, it is additive because the action and l are additive. For $r \in R$ it sends $f(r)t$ to $l(r \cdot (tn)) = rl(tn)$, proving R-linearity in t. As a function of n, it is additive and satisfies

$$\widetilde{l}(sn)(t) = l((ts)n) = (s \cdot \widetilde{l}(n))(t).$$

Thus $\widetilde{l}\colon N \to C_f(M)$ is S-linear.

Given $j\colon N \to C_f(M)$, take $n \mapsto j(n)(1)$. Evaluation at 1 is R-linear on $C_f(M)_R$, since

$$(f(r) \cdot h)(1) = h(f(r)) = rh(1).$$

The two constructions return l since $\widetilde{l}(n)(1) = l(n)$. They also return j, since $j(tn)(1) = (t \cdot j(n))(1) = j(n)(t)$. For $v\colon N' \to N$ over S and $w\colon M \to M'$ over R, the coextension of $w \circ l \circ v_R$ takes n' to $t \mapsto w(l(t\, v(n')))$. The same formula comes from precomposing $\widetilde{l}$ by v and postcomposing by $C_f(w)$. This gives naturality. □

Here the unit at N sends n to $t \mapsto tn$. The counit at M is evaluation at 1 on $C_f(M)_R$. Writing the two adjunctions together puts restriction in the middle.

$$(S \otimes_R -) \ \dashv \ (N \mapsto N_R) \ \dashv \ \mathrm{Hom}_R(S, -).$$

Example 15.3 (Scalars modulo an ideal)**.** Take $S = R/I$ for an ideal $I \subseteq R$. Let IM be the submodule of finite sums of terms im with $i \in I$, $m \in M$. The module M/IM carries an R/I-action because each element of I sends every class to zero. An R-linear map to N_R, with N an R/I-module, kills IM and hence factors through M/IM. Its factor is R/I-linear because the action on classes is induced from R. This is the extension test, so $(R/I) \otimes_R M \cong M/IM$.

For coextension, an R-linear map $R/I \to M$ is determined by its value m at $\bar{1}$. This value satisfies $im = 0$ for every $i \in I$, and each such element gives a well-defined map $\bar{r} \mapsto rm$. Under evaluation at $\bar{1}$, the R/I-action on the maps corresponds to $\bar{r} \cdot m = rm$. Write $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ for a prime p. Over $\mathbb{Z}$ with $I = 2\mathbb{Z}$, extension of $\mathbb{Z}$ gives $\mathbb{F}_2$ and coextension of $\mathbb{Z}$ gives the zero module.

15.3 Successive extensions

Let $g\colon S \to T$ be another homomorphism of commutative unital rings. A T-module can first be restricted to S and then to R. The resulting action of $r \in R$ is multiplication by $g(f(r))$, so it gives the restriction along $g \circ f$. Extension in the other direction has a corresponding isomorphism

$$\Phi\colon T \otimes_S (S \otimes_R M) \longrightarrow T \otimes_R M, \qquad t \otimes (s \otimes m) \longmapsto tg(s) \otimes m.$$

Here T is an R-module along $g \circ f$ when it occurs in the tensor product on the right. For each t, the formula $(s, m) \mapsto tg(s) \otimes m$ is R-bilinear. In particular, changing s to $f(r)s$ has the same effect as changing m to rm, by the balancing relation on the right. This gives the inner map out of $S \otimes_R M$. It is additive in t, and multiplication by s_0 on $S \otimes_R M$ has the same effect as multiplication by $g(s_0)$ on t. It therefore gives Φ by the tensor property over S, and the formula shows that Φ is T-linear.

The inverse sends $t \otimes m$ to $t \otimes (1 \otimes m)$. The formula is R-bilinear because $tg(f(r)) \otimes (1 \otimes m) = t \otimes (f(r) \otimes m) = t \otimes (1 \otimes rm)$. One composite fixes $t \otimes m$ by $g(1) = 1$. The other sends $t \otimes (s \otimes m)$ to $tg(s) \otimes (1 \otimes m) = t \otimes (s \otimes m)$. These tensors generate the respective modules.

Let $\eta^f, \eta^g, \eta^{g\circ f}$ be the extension units for the three ring maps. The isomorphism compares the two-step unit with the direct unit. Regard all displayed modules and maps as R-linear.

$$\begin{array}{ccc} M & \xrightarrow{\eta^f_M} & S \otimes_R M \\ & \searrow^{\eta^{g\circ f}_M} & \downarrow{\scriptstyle \Phi \circ \eta^g_{S\otimes_R M}} \\ & & T \otimes_R M \end{array}$$

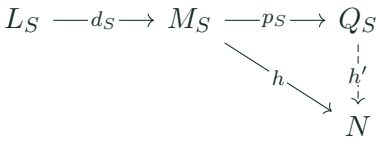
Indeed, the upper route sends m first to $1 \otimes m$ and then to $1 \otimes (1 \otimes m)$ before applying Φ. Thus a map that extends M from R to T has the same restriction to M when obtained by the two successive extension properties.

15.4 Changing the coefficients of a presentation

Suppose $d\colon L \to M$ is R-linear and $p\colon M \to Q$ is its cokernel map. The universal condition for p is that it kills d and that every map killing d factors uniquely through p. The maps d_S, p_S obtained by extension of scalars satisfy the same condition over S.

Indeed, $p_S d_S = 0$ since $pd = 0$. Given an S-linear map $h\colon M_S \to N$ with $hd_S = 0$, its adjoint map $g\colon M \to N_R$ kills d by naturality of the adjunction. There is a unique $g'\colon Q \to N_R$ with $g'p = g$. The adjoint map $h'\colon Q_S \to N$ is

the unique map with $h' p_S = h$, again by naturality.



So Q_S is the cokernel of d_S.

For a finite free module, the canonical map $S \otimes_R R^n \to S^n$ sends $s \otimes (r_1, \dots, r_n)$ to $(sf(r_1), \dots, sf(r_n))$. The formula is additive in both variables and balanced over R, and the inverse sends $(s_i)_i$ to $\sum_i s_i \otimes e_i$, with e_i the standard basis of R^n. The two maps are inverse on the standard basis on one side and on pure tensors on the other. A matrix $A = (a_{ij})$ representing $d\colon R^n \to R^m$ now represents d_S by $(f(a_{ij}))$. The jth basis vector of S^n corresponds to $1 \otimes e_j$, whose image is

$$1 \otimes d(e_j) = \sum_i f(a_{ij}) \otimes e_i.$$

Thus if the original presentation is $R^n \xrightarrow{A} R^m \to Q$, the changed presentation is $S^n \xrightarrow{f(A)} S^m \to Q_S$.

Example 15.4 (One matrix over three fields)**.** In Example 13.3 the integer matrix

$$A = \begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix}$$

has cokernel $Q \cong \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/6\mathbb{Z}$. After extension to $\mathbb{Q}$, both entries in its diagonal form $\operatorname{diag}(2, 6)$ are units, so the matrix is invertible. Hence $\mathbb{Q} \otimes_{\mathbb{Z}} Q = 0$. Here a nonzero abelian group disappears when the scalars include the inverses of the integers annihilating its generators.

The residue rings $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$ and $\mathbb{F}_3 = \mathbb{Z}/3\mathbb{Z}$ are fields because each nonzero class has a multiplicative inverse. Over $\mathbb{F}_2$ the matrix has all entries zero, so its cokernel is $\mathbb{F}_2^2$. Over $\mathbb{F}_3$ the columns are $(2, 1)$ and $(1, 2) = 2(2, 1)$. Their span is the kernel of $(x, y) \mapsto x + y$. This map is onto $\mathbb{F}_3$, so the cokernel over $\mathbb{F}_3$ has dimension one. The three answers also follow from the two diagonal relations $2e_1 = 0$ and $6e_2 = 0$, interpreted over each field.

Changing coefficients may also change the solutions of a homogeneous system. Take the map $d\colon \mathbb{Z} \to \mathbb{Z}$ given by multiplication by 2. Its integer kernel is zero. Its cokernel is $\mathbb{Z}/2\mathbb{Z}$, so the presentation calculation gives $\mathbb{F}_2 \otimes_{\mathbb{Z}} \operatorname{coker} d \cong \mathbb{F}_2$. The changed map $d_{\mathbb{F}_2}\colon \mathbb{F}_2 \to \mathbb{F}_2$ is zero, with cokernel and kernel both equal to $\mathbb{F}_2$. In particular its kernel is larger than the extension of the integer kernel. The same change of coefficients has preserved the universal property for maps *out of* the cokernel and created solutions of the equation $2x = 0$. The kernel and the cokernel therefore record different effects of the same operation on the matrix.

15.5 Inverting one scalar

Let $a \in R$ and take $S = R[1/a]$, the localization from Chapter 11. On M_S, multiplication by a has inverse multiplication by $1/a$. The extension is characterized by the maps that factor through η_M. Suppose multiplication by a is an invertible operator on an R-module P. The scalar operators on P, together with the inverse of multiplication by a, generate a commutative subring of $\mathrm{End}_R(P)$. Indeed, scalar operators commute with every R-linear map. The universal property of $R[1/a]$ extends the scalar map into this subring, giving an S-action on P. The action is unique since every fraction r/a^n must act as multiplication by r followed by the nth power of the inverse.

An R-linear map between such modules commutes with multiplication by a and therefore also with its inverse. It is automatically S-linear for these actions. The extension adjunction now says that every R-linear map $M \to P$ into a module with invertible a factors uniquely through η_M by an R-linear map from $(M_S)_R$. So the module version of localization has the same form as the ring version. The map η_M is universal for making the prescribed operation invertible.

If $a^n m = 0$ for some $n \geq 0$, then $\eta_M(m) = 1 \otimes m = a^{-n} \otimes a^n m = 0$. For the integer example, take $S = \mathbb{Z}[1/2]$. Extending the two diagonal relations gives

$$S \otimes_{\mathbb{Z}} Q \cong S/(2S) \oplus S/(6S) \cong S/(3S).$$

The first summand vanishes because 2 is a unit, and the ideals $6S$ and $3S$ coincide for the same reason. The map $r/2^n \mapsto \bar{r}\,\bar{2}^{-n}$ into $\mathbb{F}_3$ is well-defined by the localization property because 2 is invertible in $\mathbb{F}_3$. It is onto, and its kernel consists of the fractions with $3 \mid r$, which form $3S$. Thus extension to S leaves a module isomorphic to $\mathbb{F}_3$. Inverting 2 removes the summand of order 2 and the order-2 part of the summand of order 6.

Part IV uses the same test by maps into objects with a prescribed property to characterize the completion of a metric space.

Part IV Space, analysis, and probability

CHAPTER 16

Topologies from universal maps

A subset of a space inherits a topology from its inclusion. A product gets its topology from the coordinate projections, while a quotient gets its topology from a map that identifies points [10]. The direction of these maps determines how to construct the topology. Maps already specified out of a set give an initial topology, and maps already specified into it give a final topology. The two constructions lift the limits and colimits of sets to spaces.

16.1 Topologies ordered by their open sets

Definition 16.1. A *topology* on a set S is a collection τ of subsets that contains $\varnothing, S$ and is closed under arbitrary unions and finite intersections. The members of τ are *open*. A map $f\colon (S,\sigma) \to (T,\tau)$ is *continuous* if $f^{-1}(V) \in \sigma$ for each $V \in \tau$.

Inverse images take identities to identities and satisfy $(g \circ f)^{-1}(W) = f^{-1}(g^{-1}(W))$. Thus spaces and continuous maps form the category **Top**, with a forgetful functor $U\colon \mathbf{Top} \to \mathbf{Set}$ that retains the underlying sets and functions. Write $\mathcal{O}(X)$ for the topology of a space X.

On a fixed set, order topologies by inclusion. The topology σ is *finer* than τ if $\tau \subseteq \sigma$. Equivalently, the identity function is continuous from (S,σ) to (S,τ). Finer topologies therefore provide more open inverse images for maps out of a space. For maps into a space, coarser topologies provide fewer open sets to test.

Intersections of any family of topologies on S are topologies. It follows that for any collection $\mathcal{B}$ of subsets of S there is a least topology containing $\mathcal{B}$, the intersection of all topologies containing it. The same topology consists of arbitrary unions of finite intersections of members of $\mathcal{B}$. The empty intersection gives S, the empty union gives $\varnothing$, and distributing one finite intersection over arbitrary unions verifies that the displayed construction gives a topology. Any topology containing $\mathcal{B}$ must contain all the resulting opens.

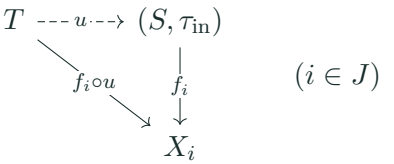
The finest topology on S is the *discrete* topology $\mathcal{P}(S)$. The coarsest is the *indiscrete* topology $\{\varnothing, S\}$. The corresponding spaces $D(S)$ and $I(S)$ extend to functors $\mathbf{Set} \to \mathbf{Top}$. Every function $S \to U(X)$ is continuous as a map $D(S) \to X$, and every function $U(X) \to S$ is continuous as a map $X \to I(S)$.

These bijections keep the underlying functions and commute with composition. Hence $D \dashv U \dashv I$ by Chapter 5.

16.2 Initial topologies

Let S be a set, let $(X_i)_{i \in J}$ be a set-indexed family of spaces, and fix functions $f_i \colon S \to U(X_i)$.

Proposition 16.2 (Initial structure)**.** *There is a coarsest topology on S for which all f_i are continuous. With that topology, a function $u \colon U(T) \to S$ from a space T is continuous if and only if every composite $f_i \circ u$ is continuous.*



Proof. The sets $f_i^{-1}(V)$, for $i \in J$ and $V \in \mathcal{O}(X_i)$, must be open in any topology making every f_i continuous. Take the topology they generate and denote it τ_{in}. If all $f_i \circ u$ are continuous, the inverse image under u of each of these generating sets is open in T. Inverse image preserves arbitrary unions and finite intersections, so u pulls back every open set to an open set. Conversely, if u is continuous, so is each composite. □

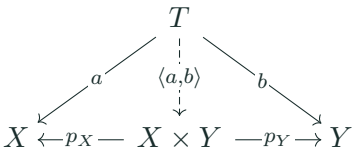
Equivalently, the forgetful map gives a bijection

$$\mathrm{Hom}_{\mathbf{Top}}(T, (S, \tau_{\text{in}})) \cong \{u \in \mathrm{Hom}_{\mathbf{Set}}(U(T), S) \mid f_i \circ u \text{ is continuous for every } i\}.$$

It is natural in T by precomposition. Taking a second topology with this property and testing with its identity function gives a continuous identity into (S, τ_{in}). Testing the other way gives continuity of the inverse identity, so the topology is unique on the prescribed set. When J is empty, this property gives the indiscrete topology.

Example 16.3 (Subspaces and pullbacks)**.** For an inclusion $A \hookrightarrow U(X)$, the sets $A \cap V$ with V open in X already make a topology. This is the *subspace* topology. In particular, a map into A is continuous exactly when its composite into X is continuous. For any continuous $f \colon X \to Y$ and $g \colon Z \to Y$, give the set of pairs $X \times_Y Z$ from Chapter 3 the initial topology for its two projections. The projections agree over Y. Any two compatible continuous maps from T to X, Z determine a unique function to the set of pairs, and the initial property makes it continuous. Thus the set of pairs with this topology is the pullback in **Top**.

Example 16.4 (Products)**.** The set product equipped with the initial topology for all projections is the categorical product of spaces. A tuple of continuous maps $T \to X_i$ has a unique pairing into the set product, and its continuity follows from those of its coordinates. This gives the same universal property as in Chapter 2.



For two factors the finite intersections of generating sets reduce to rectangles $V \times W$, where V, W are open in the respective factors. For a general family the basic opens impose open conditions on finitely many coordinates at a time.

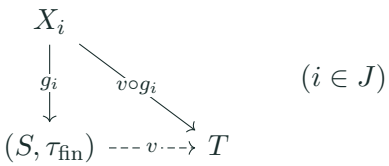
This finiteness follows from the universal property. For instance, let $C = \{0, 1\}$ be discrete and consider the infinite product $C^{\mathbb{N}}$. Every basic open containing the all-zero tuple contains a second tuple obtained by setting an unrestricted coordinate to 1. Hence the singleton of the all-zero tuple is not open. Giving the set product the discrete topology would make its projections continuous, but it would make the identity function from the actual product discontinuous. That function must be the unique pairing of the projections.

The same argument lifts every small limit. Take the set limit of the diagram obtained by applying U, and give it the initial topology for the projections in its limiting cone. Every compatible cone of continuous maps factors through the set limit. Its factorization is continuous because its composites with the limit projections are the maps of the given cone.

16.3 Final topologies

Now fix a set S, spaces $(X_i)_{i \in J}$ and functions $g_i \colon U(X_i) \to S$. The construction reverses the continuity test.

Proposition 16.5 (Final structure)**.** *There is a finest topology on S for which all g_i are continuous. With that topology, a function $v \colon S \to U(T)$ to a space T is continuous if and only if every composite $v \circ g_i$ is continuous.*



Proof. Define τ_{fin} to consist of those subsets $V \subseteq S$ with $g_i^{-1}(V)$ open in X_i for every i. Inverse images preserve arbitrary unions and finite intersections, so this is a topology. The definition makes all g_i continuous and contains every possible open in a topology having that property. If each $v \circ g_i$ is continuous and W is open in T, each $g_i^{-1}(v^{-1}(W))$ is open. Hence $v^{-1}(W) \in \tau_{\text{fin}}$. The converse follows by composing continuous maps. □

This time the forgetful map gives the natural bijection

$$\mathrm{Hom}_{\mathbf{Top}}((S, \tau_{\text{fin}}), T) \cong \{v \in \mathrm{Hom}_{\mathbf{Set}}(S, U(T)) \mid v \circ g_i \text{ is continuous for every } i\}.$$

Naturality in T is by postcomposition. Testing with identity functions again characterizes the topology uniquely. When J is empty, every subset passes the test and the final topology is discrete.

Example 16.6 (Coproducts and colimits)**.** The disjoint union $\coprod_i U(X_i)$ with the final topology for its injections is the coproduct of the spaces. An open set is exactly a union of opens taken separately in the summands. A family of maps $X_i \to T$ determines a function out of their disjoint union, and that function is continuous precisely when all the given maps are. In the same way, equip the set colimit of any small diagram with the final topology of its cocone maps. The set colimit factors every compatible cocone uniquely, and the final property makes every such factorization continuous.

For a single function $f\colon S \to T$, these constructions can be placed in an adjunction between posets of topologies. Let $\mathrm{Topol}(S)$ denote the poset of topologies on S. For τ on T, the initial topology $f^*\tau$ consists of the inverse images of opens of τ, since the inverse images already form a topology. For σ on S, let $f_{\text{fin}}\sigma$ be the final topology on T. Both assignments are monotone, and

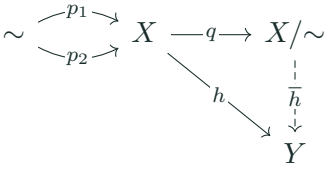
$$f^*\tau \subseteq \sigma \quad \Longleftrightarrow \quad f\colon (S, \sigma) \to (T, \tau) \text{ is continuous} \quad \Longleftrightarrow \quad \tau \subseteq f_{\text{fin}}\sigma.$$

Thus $f^* \dashv f_{\text{fin}}$ in the inclusion order. Here the objects of the posets are entire topologies.

16.4 Quotients and gluing

Let $\sim$ be an equivalence relation on the underlying set of a space X, and write $q\colon U(X) \to U(X)/\!\sim$ for the map to equivalence classes. The *quotient topology* is the final topology for q. A subset of the quotient is open exactly when the union of its classes is open in X. A continuous map $h\colon X \to Y$ that is constant on each class induces a unique function on the set of classes. The final property makes it

continuous. The quotient is thus the coequalizer in **Top** of the two projections from the relation, given its subspace topology in $X \times X$.



A bijective continuous map can fail to identify the topologies. For example, the identity from a two-point discrete space to an indiscrete one is continuous, and its inverse fails the open-set test. A bijective map that gives its codomain the final topology is a homeomorphism. The quotient criterion thus supplies exactly the continuity of the inverse when one identifies a quotient with a concrete space.

Example 16.7 (Gluing the ends of an interval)**.** Take $[0, 4]$ with its usual topology and identify 0 and 4. Write Q for the resulting quotient. Since any map from the discrete space $\{0, 4\}$ is continuous, the quotient property presents Q as the following pushout.

$$\begin{array}{ccc} \{0,4\} & \hookrightarrow & [0,4] \\ \downarrow & & \downarrow{\scriptstyle q} \\ 1 & \longrightarrow & Q \end{array}$$

Indeed a compatible pair of maps out of $[0, 4]$ and 1 is exactly a map out of the interval that takes the same value at both endpoints, together with that forced value at 1.

To identify the space concretely, let B be the boundary of $[0, 1]^2$, with its subspace topology. Travel along its four edges at constant speed, starting from $(0, 0)$. The points reached at times $0, 1, 2, 3, 4$ are

$$(0,0), \quad (1,0), \quad (1,1), \quad (0,1), \quad (0,0).$$

On each segment $[j, j + 1]$ take the affine map to the corresponding edge. These maps are homeomorphisms on their segments, with continuous inverses given by the changing coordinate or its reversal.

If a space is covered by finitely many closed subspaces, its topology is final for their inclusions. Indeed, a set whose intersection with each member of the cover is relatively closed is a finite union of closed sets in the whole space. Taking complements proves the open-set version. Apply this fact to the four closed segments of the interval and to the four closed edges of B. The travel map $p\colon [0, 4] \to B$ is continuous by the final property for the segments. If $p^{-1}(W)$ is open, then its intersection with each segment is open in that segment. The homeomorphisms on segments make W relatively open in each edge, so W is

open in B. Thus B has the final topology for p. The fibers are precisely the equivalence classes defining Q. Consequently the induced bijection $Q \to B$ is a homeomorphism, since a subset on either side is open exactly when its inverse image in the interval is open.

16.5 A space that classifies open sets

The *Sierpiński space* $\mathbb{S}$ has underlying set $\{0,1\}$ and open sets $\varnothing, \{1\}, \{0,1\}$. For a subset V of a space X, let $\chi_V \colon X \to \mathbb{S}$ be the function taking value 1 precisely on V. It is continuous exactly when V is open, since $\{1\}$ is the only open with a nontrivial inverse-image condition. Every function into $\{0,1\}$ is the characteristic function of its inverse image of 1. This gives a bijection

$$\mathcal{O}(X) \cong \mathrm{Hom}_{\mathbf{Top}}(X, \mathbb{S}), \qquad V \longmapsto \chi_V.$$

For a continuous map $f \colon Y \to X$, the identity $\chi_{f^{-1}(V)} = \chi_V \circ f$ shows that the bijection is natural. It exhibits the contravariant open-set functor as representable in the sense of Chapter 4.

When V is open, the classifier map also describes its inclusion by a pullback. The right map in the square picks the point 1.

$$\begin{array}{ccc} V & \hookrightarrow & X \\ \downarrow & \lrcorner & \downarrow{\scriptstyle \chi_V} \\ 1 & \xrightarrow{1} & \mathbb{S} \end{array}$$

The underlying set of the pullback is identified with V, and its initial topology for the projections is the subspace topology since the map into the one-point space brings no additional open conditions. Pullback along f then replaces the open inclusion of V by the open inclusion of $f^{-1}(V)$.

16.6 Pontryagin duality for locally compact abelian groups

A Hausdorff topological abelian group is a Hausdorff space with continuous abelian group operations. It is *locally compact* if each point has a compact neighborhood. Write **LCA** for the category of these groups and continuous homomorphisms. The circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ has its quotient topology. Define the *Pontryagin dual* of G by

$$\widehat{G} = \mathrm{Hom}_{\mathrm{cts}}(G, \mathbb{T}).$$

These are homomorphisms of topological groups. Give the group $\widehat{G}$ the compact-open topology, whose subbasic opens require that a character send a specified

compact subset of G into a specified open subset of $\mathbb{T}$. In particular, a neighborhood basis at the zero character consists of the sets

$$[K, U] = \{\chi : \chi(K) \subseteq U\},$$

with K compact and U an open neighborhood of zero. Precomposition $\widehat{f}(\chi) = \chi \circ f$ is continuous, since $f(K)$ is compact. As in Section 5.6, evaluating a point g against characters defines $\eta_G(g)(\chi) = \chi(g)$. Each such evaluation is continuous on $\widehat{G}$ by taking $K = \{g\}$.

Theorem 16.8 (Pontryagin–van Kampen duality)**.** *For every $G \in$ **LCA**, its dual $\widehat{G}$ belongs to* **LCA** *and $\eta_G \colon G \to \widehat{\widehat{G}}$ is an isomorphism of topological groups. The character functor is a contravariant equivalence on* **LCA***.*

Morris proves the topological assertion using the structure of locally compact groups, and Roeder gives a proof through exact sequences and limits [60, 61]. To see how the theorem gives the categorical assertion, let $f \colon G \to H$. Both routes in the following square send g, evaluated at $\chi \in \widehat{H}$, to $\chi(f(g))$.

$$\begin{array}{ccc} G & \xrightarrow{\eta_G} & \widehat{\widehat{G}} \\ \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle \widehat{\widehat{f}}} \\ H & \xrightarrow{\eta_H} & \widehat{\widehat{H}}. \end{array}$$

These natural isomorphisms make $G \mapsto \widehat{G}$ a quasi-inverse of its opposite functor. For example, a map $u \colon \widehat{H} \to \widehat{G}$ comes from the map $\eta_H^{-1} \widehat{u} \eta_G \colon G \to H$. Evaluation at a character checks this formula directly.

16.6.1 Compact and discrete groups

If G is discrete, the compact subsets are the finite subsets. The compact-open topology on $\widehat{G}$ is therefore the subspace topology from $\mathbb{T}^G$ with the product topology. The character equations $\chi(x + y) = \chi(x) + \chi(y)$ are closed conditions, so $\widehat{G}$ is compact. If G is compact, take an open arc U about zero in $\mathbb{T}$ containing no nonzero subgroup, for instance the image of $(-1/4, 1/4)$. Any nonzero element in that arc has an integer multiple outside it. Hence $[G, U]$ contains only the zero character, so $\widehat{G}$ is discrete. Biduality gives the converses and identifies compact abelian groups and discrete abelian groups contravariantly.

The discrete group $\mathbb{Z}$ has dual $\mathbb{T}$ by evaluation at 1, since the character with value t at 1 is specified coordinatewise by continuous maps $t \mapsto mt$. Under $\eta_{\mathbb{Z}}$, the integer m corresponds to the character $t \mapsto mt$ of $\mathbb{T}$. The real line has dual $\mathbb{R}$ under $\xi \mapsto (x \mapsto \xi x + \mathbb{Z})$. To see the group identification, lift a continuous

character through $\mathbb{R} \to \mathbb{T}$, choosing $u(0) = 0$. The path-lifting construction of Section 54.3 applies on each bounded interval containing 0. Uniqueness makes the lifts agree on overlaps. The difference $u(x+y) - u(x) - u(y)$ is integer-valued and continuous on the connected plane, so it is zero. Continuity then gives $u(x) = xu(1)$ by first checking rational x. The parameterization is continuous for uniform convergence on compact sets. Conversely, if the character with parameter ξ sends $[-1, 1]$ into the image of $(-\epsilon, \epsilon)$ for $0 < \epsilon < 1/4$, its lift ξx stays in that interval. In particular $|\xi| < \epsilon$, which proves continuity of the inverse.

16.6.2 Subgroups, quotients, and Fourier coordinates

If H is a closed subgroup of G, put $H^\perp = \{\chi \in \widehat{G} : \chi|_H = 0\}$. The quotient $G \to G/H$ identifies $\widehat{G/H}$ with $H^\perp$, because a character descends uniquely precisely when it vanishes on H. In **LCA**, the inclusion of H is the kernel of the quotient map, and that quotient map is its cokernel. The equivalence exchanges these two universal properties. A kernel is a closed embedding, and a cokernel is the open quotient by the closure of the image. We obtain

$$0 \longrightarrow \widehat{G/H} \longrightarrow \widehat{G} \longrightarrow \widehat{H} \longrightarrow 0$$

with the subspace and quotient topologies [60, 62]. In particular, every continuous character of H extends to G. Thus $\widehat{\mathbb{Z}/m\mathbb{Z}} = \mathbb{T}[m]$ also follows by dualizing $0 \to \mathbb{Z} \xrightarrow{m} \mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \to 0$, for $m \geq 1$. Multiplication by m on the two copies of the circle is the dual arrow, and its kernel consists of the m-torsion points.

For harmonic analysis write $e(t) = e^{2\pi i t}$. Choose a Haar measure dx on G, which is invariant under translation. For $f \in L^1(G, dx)$, set

$$\mathcal{F}f(\chi) = \int_G f(x)\overline{e(\chi(x))}\, dx.$$

The variable χ is a point of the dual group. Translating $f(x)$ to $f(x-a)$ multiplies its Fourier transform by $\overline{e(\chi(a))}$, by the change of variable $x = y + a$. With the corresponding normalization of Haar measure on the dual group, the Plancherel theorem gives a unitary map $L^2(G) \to L^2(\widehat{G})$ that agrees with the integral on $L^1(G) \cap L^2(G)$ [63]. For a finite group with uniform probability measure the integral is a normalized sum. The dual has counting measure, and unitarity follows from the orthonormal character basis of Section 8.6. On the circle with probability Haar measure it gives Fourier series indexed by $\mathbb{Z}$. The finite case will turn a count of walks into products of character values in Section 20.6.

CHAPTER 17

Spaces from their open sets

A continuous map acts on open sets by inverse image. Inverse image preserves unions and finite intersections, and these two operations specify the algebraic structure that opens carry. We can begin with an abstract algebra of opens, define its points, and ask whether those points recover the algebra. Beginning with a space raises the other question. Which spaces can be recovered from their opens? The two questions refer to opposite sides of an adjunction.

17.1 Frames and their maps

A *complete lattice* is a partially ordered set in which each set of elements has a least upper bound and a greatest lower bound. We write these bounds as joins and meets, with 0 the empty join and 1 the empty meet. A *frame* is a complete lattice L satisfying

$$a \wedge \bigvee_{i \in I} b_i = \bigvee_{i \in I} (a \wedge b_i)$$

for every $a \in L$ and every family $(b_i)_{i \in I}$. A frame map preserves arbitrary joins and finite meets, including 0 and 1.

The opens of a space X, ordered by inclusion, form a frame $\mathcal{O}(X)$. A join is a union and a finite meet is an intersection. The meet of an arbitrary family of opens is the interior of their intersection, the largest open contained in it. The frame law is the distributivity of intersection over union. The need for an interior in an infinite meet explains the difference between arbitrary joins and finite meets in the definition of a frame map. For example, inclusion of the point 0 into $\mathbb{R}$ pulls every $(-1/n, 1/n)$ back to the whole point, whereas the meet of these intervals in $\mathcal{O}(\mathbb{R})$ is empty.

For a continuous map $f\colon X \to Y$, the inverse image $f^*\colon \mathcal{O}(Y) \to \mathcal{O}(X)$ is a frame map. For composable maps f, g, inverse image gives $(g \circ f)^* = f^* \circ g^*$. Let **Frm** be the category of frames and frame maps. Its opposite $\mathbf{Loc} = \mathbf{Frm}^{\mathrm{op}}$ is the category of *locales*. We use the same symbol for a locale and its underlying frame of opens. A map of locales $u\colon L \to M$ means a frame map $u^*\colon M \to L$. With this convention, forming opens is a covariant functor

$$\Omega\colon \mathbf{Top} \longrightarrow \mathbf{Loc}, \qquad \Omega X = \mathcal{O}(X), \qquad (\Omega f)^* = f^*.$$

The direction agrees with the direction of the continuous map when the opens are regarded as a locale. Picado and Pultr [11] develop the theory of frames and locales from this change of viewpoint.

Finite meets and arbitrary joins also determine an implication operation. For $a, b \in L$, put

$$a \Rightarrow b = \bigvee \{c \in L \mid a \wedge c \leq b\}.$$

Distributivity shows that meeting this join with a still gives an element below b. Thus

$$a \wedge c \leq b \quad \Longleftrightarrow \quad c \leq (a \Rightarrow b).$$

This is an adjunction between two order-preserving maps on the frame. For opens U, V of a space, it gives $U \Rightarrow V = \operatorname{int}((X \setminus U) \cup V)$. In particular the largest open disjoint from U is the interior of its complement. On the real line the open ray $(0, \infty)$ and its largest disjoint open ray $(-\infty, 0)$ have union $\mathbb{R} \setminus \{0\}$. The two opens therefore fail to exhaust the space. We will return to implication when deriving logic from order.

17.2 Points of a frame

The one-point space has frame of opens $\mathbf{2} = \{0 < 1\}$. A *point* of a frame L is a frame map $p \colon L \to \mathbf{2}$, or equivalently a locale map $\mathbf{2} \to L$. For a space X and a point $x \in X$, evaluation at x is such a frame map. It takes an open U to 1 exactly when $x \in U$.

There is an equivalent description using the opens at which p takes value 1. A *completely prime filter* $P \subseteq L$ is upward closed, contains 1, omits 0, and is closed under binary meets. It also satisfies

$$\bigvee_{i \in I} a_i \in P \quad \Longrightarrow \quad a_i \in P \text{ for some } i \in I.$$

These are precisely the conditions on $P = p^{-1}(1)$. Indeed, upward closure gives the reverse implication for joins and gives membership of both factors when a meet belongs to P. The remaining filter conditions give the other direction for meets and specify the values at 0 and 1. Taking the characteristic function of P therefore recovers p.

Write $\operatorname{Pt}(L)$ for the set of points and put

$$\widehat{a} = \{p \in \operatorname{Pt}(L) \mid p(a) = 1\}.$$

Preservation of joins and finite meets by each p gives

$$\widehat{\bigvee_i a_i} = \bigcup_i \widehat{a_i}, \qquad \widehat{a \wedge b} = \widehat{a} \cap \widehat{b}, \qquad \widehat{0} = \varnothing, \quad \widehat{1} = \operatorname{Pt}(L).$$

Consequently the sets $\widehat{a}$ already form a topology. For a locale map $u\colon L \to M$, define $\mathrm{Pt}(u)(p) = p \circ u^*$. The inverse image of $\widehat{b} \subseteq \mathrm{Pt}(M)$ under this map is $\widehat{u^*(b)}$, so the map is continuous. Precomposition with inverse-image maps respects identities and composition. This defines $\mathrm{Pt}\colon \mathbf{Loc} \to \mathbf{Top}$.

Example 17.1 (The Sierpiński space)**.** The frame $\mathcal{O}(\mathbb{S})$ of the Sierpiński space has the three elements $0 < a < 1$, with $a = \{1\}$. A point must take $0, 1$ to $0, 1$, respectively. Either value of $p(a)$ gives a frame map, since these maps are monotone between finite chains and preserve the endpoints. Calling the points p_0, p_1 according to that value, the only nonempty proper open is $\widehat{a} = \{p_1\}$. Thus the frame recovers the space whose characteristic maps classified opens in Section 16.5.

17.3 The opens–points adjunction

Theorem 17.2. *The functor Ω is left adjoint to* Pt*. For a space X and a locale L, the correspondence is*

$$\mathrm{Hom}_{\mathbf{Loc}}(\Omega X, L) \cong \mathrm{Hom}_{\mathbf{Top}}(X, \mathrm{Pt}(L)).$$

Proof. A locale map $u\colon \Omega X \to L$ gives a frame map $u^*\colon L \to \mathcal{O}(X)$. Send $x \in X$ to the composite of u^* with evaluation at x. Writing the resulting map as f_u, we have

$$f_u(x)(a) = 1 \quad \Longleftrightarrow \quad x \in u^*(a).$$

In particular $f_u^{-1}(\widehat{a}) = u^*(a)$, proving continuity. Conversely, from a continuous $f\colon X \to \mathrm{Pt}(L)$ take the frame map $a \mapsto f^{-1}(\widehat{a})$. This preserves joins and finite meets by the identities for $\widehat{a}$ and by the laws of inverse image. Starting with either map and applying the two rules gives it back. The displayed formula recovers the open $u^*(a)$ from f_u, and membership in each $\widehat{a}$ recovers the value of $f(x)$ at every a.

For $g\colon X' \to X$, the rule sends $u \circ \Omega g$ to $f_u \circ g$. For $v\colon L \to M$, it sends $v \circ u$ to $\mathrm{Pt}(v) \circ f_u$. These identities follow by composing inverse-image maps and give naturality in both variables. □

The unit $\eta_X\colon X \to \mathrm{Pt}(\Omega X)$ takes x to evaluation at x. The counit $\varepsilon_L\colon \Omega\,\mathrm{Pt}(L) \to L$ has inverse-image map $\varepsilon_L^*(a) = \widehat{a}$. The correspondence in Theorem 17.2 can be read in either direction as a factorization.

$$\begin{array}{ccc} X & \xrightarrow{\eta_X} & \mathrm{Pt}(\Omega X) \\ & {\scriptstyle f_u}\searrow & \downarrow{\scriptstyle \mathrm{Pt}(u)} \\ & & \mathrm{Pt}(L) \end{array} \qquad \begin{array}{ccc} \Omega X & \xrightarrow{\Omega f} & \Omega\,\mathrm{Pt}(L) \\ & {\scriptstyle u_f}\searrow & \downarrow{\scriptstyle \varepsilon_L} \\ & & L. \end{array}$$

The identities of the adjunction can be checked directly. On the frame of opens of X, the composite $(\varepsilon_{\Omega X} \circ \Omega\eta_X)^*$ sends an open U to $\eta_X^{-1}(\widehat{U}) = U$. On a point p of L, the composite $\mathrm{Pt}(\varepsilon_L) \circ \eta_{\mathrm{Pt}(L)}$ gives the map taking a to 1 exactly when $p \in \widehat{a}$. This map is p.

17.4 Spatial locales and sober spaces

A locale is *spatial* if distinct opens have distinct sets of points. The frame map $a \mapsto \widehat{a}$ is surjective by the definition of the topology on $\mathrm{Pt}(L)$. Thus spatiality means precisely that the counit ε_L is an isomorphism. A frame of the form $\mathcal{O}(X)$ is always spatial. Unequal opens of X differ at some $x \in X$, and evaluation at x separates them. It follows that a locale is spatial exactly when it is isomorphic to the locale of opens of some space.

For spaces we instead test the *unit* η_X. Its inverse-image map on opens is an isomorphism, since all opens of $\mathrm{Pt}(\Omega X)$ are of the form $\widehat{U}$ and $\eta_X^{-1}(\widehat{U}) = U$. In particular η_X is a homeomorphism exactly when it is bijective. Injectivity means that any two distinct points are distinguished by an open, the T_0 separation condition. Surjectivity asks whether every completely prime filter on $\mathcal{O}(X)$ is the filter of neighborhoods of a point. A space X is *sober* when η_X is a homeomorphism.

An equivalent test makes sobriety concrete. A nonempty closed set C is *irreducible* if any two opens meeting C have intersection meeting C. Equivalently C cannot be expressed as the union of two proper closed subsets of C, by taking complements in C. A point x is a *generic point* of C if $\overline{\{x\}} = C$.

Proposition 17.3. *The points of $\mathcal{O}(X)$ correspond bijectively to the nonempty irreducible closed subsets of X. The point $\eta_X(x)$ corresponds to $\overline{\{x\}}$. Thus a space is sober exactly when every nonempty irreducible closed set has a unique generic point.*

Proof. For a completely prime filter P, form

$$B_P = \bigcup \{U \in \mathcal{O}(X) \mid U \notin P\}, \qquad C_P = X \setminus B_P.$$

Complete primeness gives $B_P \notin P$, so C_P is nonempty. An open U belongs to P exactly when it meets C_P. Indeed, an open outside P is included in B_P, and an open in P cannot be included there by upward closure. Since P is closed under meets, C_P is irreducible.

Conversely the opens meeting a nonempty irreducible closed set C form a completely prime filter. Irreducibility gives closure under binary meets, while a union meets C precisely when one of its terms does. The union of the opens disjoint from C is its complement, so this construction inverts $P \mapsto C_P$.

For evaluation at x, the opens outside the filter are those omitting x. Their union is the complement of $\overline{\{x\}}$, which proves the claimed correspondence for η_X. The characterization follows from the bijectivity test for η_X. □

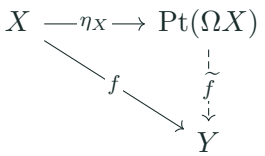
For a Hausdorff space, two distinct points of an irreducible closed set would lie in disjoint opens meeting that set. Consequently each nonempty irreducible closed set is a singleton and has a unique generic point. The next example shows that closed singletons alone do not suffice.

Example 17.4 (Adding a generic point)**.** Let X be an infinite set with the cofinite topology. The closed sets are the finite subsets and X. The nonempty irreducible closed sets are the singletons and X itself. A finite set with two or more elements splits into two proper finite closed sets. Two nonempty cofinite opens always intersect, so X is irreducible. Since the points of X are closed, no point has closure X.

By Proposition 17.3, $\mathrm{Pt}(\Omega X)$ is X together with one point γ corresponding to the filter of all nonempty opens. Its nonempty opens are $U \cup \{\gamma\}$ for nonempty cofinite $U \subseteq X$. Every nonempty open contains γ, so γ has the whole enlarged space as its closure. Restricting the enlarged opens to X recovers precisely the old opens.

For every locale L, the triangular identity $\mathrm{Pt}(\varepsilon_L) \circ \eta_{\mathrm{Pt}(L)} = \mathrm{id}$ shows that $\eta_{\mathrm{Pt}(L)}$ is injective. To see surjectivity, take a point q of $\mathcal{O}(\mathrm{Pt}(L))$ and set $p = q \circ \varepsilon_L^*$. For every $a \in L$, $\eta_{\mathrm{Pt}(L)}(p)(\widehat{a}) = p(a) = q(\widehat{a})$. Since every open has the form $\widehat{a}$, evaluation at p is q. Thus $\mathrm{Pt}(L)$ is sober.

In particular $\eta_X \colon X \to \mathrm{Pt}(\Omega X)$ is a map into a sober space. For any continuous $f \colon X \to Y$ with Y sober, naturality of the unit gives the factor $\widetilde{f} = \eta_Y^{-1} \circ \mathrm{Pt}(\Omega f)$.



For uniqueness, the inverse images of opens along any factor are determined by their inverse images along η_X, which induces an isomorphism on opens. Since Y is T_0, these inverse images determine the factor pointwise. This universal map is the *sobrification* of X. Restricting the opens–points adjunction to sober spaces and spatial locales gives an equivalence, since its unit and counit there are isomorphisms.

17.5 An algebra of opens with no points

Some frames have too few points to recover any of their nontrivial opens. We construct one using the usual topology of the real line, whose underlying metric

is constructed in Chapter 18. For an open $U \subseteq \mathbb{R}$ write $r(U) = \operatorname{int}(\overline{U})$. This operation is monotone, contains U, and is idempotent. Indeed $U \subseteq r(U) \subseteq \overline{U}$, so the two opens $U, r(U)$ have the same closure. The fixed opens are called *regular open sets*.

For open U, V, monotonicity gives $r(U \cap V) \subseteq r(U) \cap r(V)$. Every nonempty open subset of $r(U) \cap r(V)$ meets U, and its intersection with U is open and meets V. It follows that $U \cap V$ is dense in $r(U) \cap r(V)$, giving the reverse inclusion. The regular opens have finite meets given by intersection and joins given by $r(\bigcup_i U_i)$. These joins are least upper bounds by monotonicity and idempotence, and arbitrary meets exist as joins of lower bounds. If A is regular, then $A \cap r(U) = r(A \cap U)$ by the binary intersection formula. Applied to a union, this gives the frame law for the regular opens.

For a regular open A, the open $A' = \operatorname{int}(\mathbb{R} \setminus A)$ is regular. In fact, for any open U its open complement $U' = \operatorname{int}(\mathbb{R} \setminus U)$ satisfies $(U')' = r(U)$ and $(r(U))' = U'$, using the equality of closures of $U, r(U)$. Thus $r(U') = U'$. We have $A \cap A' = \varnothing$, and their union is dense since an open disjoint from A is contained in A'. So $A \vee A' = \mathbb{R}$ in the regular-open frame. Its elements are all complemented.

Suppose this frame had a point p. The join d of all elements taken to 0 by p is itself taken to 0, and its complement $c = d'$ is taken to 1. Every nonempty regular open below c must be taken to 1. Otherwise it would be below d as well, and $c \cap d = \varnothing$. But the nonempty open c contains two disjoint nonempty bounded open intervals. These intervals are regular opens, so p takes both to 1 and their meet to 0, contradicting preservation of meets. The resulting locale has no points even though its bottom and top are distinct. It exhibits the difference between the existence of a frame of opens and its reconstruction by an ordinary space.

CHAPTER 18

Real numbers from completion

Nested rational intervals can determine a number even when their common point cannot be rational. We can construct that number by identifying rational approximations whose discrepancy goes to zero. The resulting space of real numbers admits a map from $\mathbb{Q}$ and satisfies a universal property. Every distance-nonincreasing map from $\mathbb{Q}$ to a complete metric space extends uniquely to the reals. We first build the numbers and their order using rational bounds. This supplies the real-valued distances used in the general completion theorem.

18.1 Cauchy sequences and rational bounds

Chapter 11 obtains $\mathbb{Q}$ by inverting every nonzero integer. Its usual order gives the absolute value used here. A rational sequence $(a_n)_{n\ge 1}$ is *Cauchy* if for each positive rational ϵ there is an N such that $|a_n - a_m| < \epsilon$ whenever $m, n \ge N$. It is *null* if for each positive rational ϵ there is an N such that $|a_n| < \epsilon$ for $n \ge N$. Every null sequence is Cauchy by the triangle inequality. Every Cauchy sequence is bounded. Taking $\epsilon = 1$ bounds the tail by $|a_N| + 1$, and a larger bound handles the finite initial segment.

Let C be the set of rational Cauchy sequences. With coordinatewise addition and multiplication it is a subring of $\mathbb{Q}^{\mathbb{N}_{>0}}$. For sums this follows from the triangle inequality. For products, if $B \ge 1$ bounds the absolute values of the terms of both sequences, then

$$|a_n b_n - a_m b_m| \le B\big(|a_n - a_m| + |b_n - b_m|\big).$$

This estimate follows by adding and subtracting $a_n b_m$. The null sequences form an ideal I of C. They are closed under addition and negation, and the product of a null sequence with a bounded sequence is null.

Define $\mathbb{R} = C/I$, using the quotient of Section 11.1. This is the Cauchy sequence construction of the reals [12]. Write $[a_n]$ for the class of a sequence. Two sequences give the same class exactly when their termwise difference is null. The inclusion $\iota\colon \mathbb{Q} \to \mathbb{R}$ sends q to the class of the constant sequence q. A constant null sequence has value zero, so this ring map is injective. We identify each rational with its image.

Proposition 18.1. *The ring $\mathbb{R}$ is an ordered field. Its positive elements are exactly the classes $[a_n]$ for which there exist a positive rational δ and an N such that $a_n \geq \delta$ for all $n \geq N$. The order extends the order of $\mathbb{Q}$.*

Proof. A change of representative by a null sequence preserves eventual positivity with a uniform lower bound, using $\delta/2$ for the new bound. If $[a_n] \neq 0$, then for some $\delta > 0$ there are arbitrarily late terms with $|a_n| \geq 2\delta$. Pick one such term in a Cauchy tail where all pairwise differences are less than δ. Every term in the tail has the same sign as the chosen term and has absolute value at least δ. Thus for a nonzero class exactly one of $[a_n]$ and $-[a_n]$ is positive. The sum and product of two positive classes have positive rational lower bounds on a common tail, by addition and multiplication of their bounds. The rule $x < y$ when $y - x$ is positive therefore defines a strict total order compatible with addition and multiplication by positive elements. Constant sequences give the rational order.

On a tail where $|a_n| \geq \delta$, the sequence of reciprocals satisfies

$$\left|\frac{1}{a_n} - \frac{1}{a_m}\right| \leq \delta^{-2}|a_n - a_m|.$$

It is Cauchy. Fill its finite initial segment with zero. Its product with (a_n) is eventually 1, so its class is the reciprocal of $[a_n]$. Thus every nonzero class is invertible. □

Absolute value satisfies $|[a_n]| = [|a_n|]$, and $(|a_n|)$ is Cauchy by the reverse triangle inequality. For a nonzero class the formula follows from the eventual sign proved above, and for zero both sequences are null. A class represented by an eventually nonnegative sequence is nonnegative. These facts give a way to compare a rational approximation with its class. Given $x = [a_n]$ and a rational $\epsilon > 0$, choose N such that $|a_n - a_m| < \epsilon/2$ for $n, m \geq N$. For any fixed $m \geq N$ the sequence $\epsilon/2 - |a_n - a_m|$ is eventually nonnegative. Hence

$$|x - a_m| \leq \epsilon/2 < \epsilon.$$

Each positive real u exceeds a positive rational, by taking half its eventual positive bound. Consequently a_m converges to x for the distance $|x - y|$, and $\mathbb{Q}$ is dense in $\mathbb{R}$. Density here means that each point lies within every positive distance of a rational point.

Theorem 18.2. *Every Cauchy sequence for the distance $|x - y|$ on $\mathbb{R}$ converges in $\mathbb{R}$.*

Proof. For a real Cauchy sequence (x_j), choose $q_j \in \mathbb{Q}$ with $|q_j - x_j| < 2^{-j}$. This choice can be made by taking the first suitable term in a fixed enumeration of $\mathbb{Q}$. The estimate

$$|q_j - q_k| \leq 2^{-j} + |x_j - x_k| + 2^{-k}$$

shows that (q_j) is rational Cauchy. Its class $x = [q_j]$ is its limit by the preceding argument. The bound $|x_j - x| \le 2^{-j} + |q_j - x|$ proves that $x_j \to x$. □

Corollary 18.3 (Least upper bounds)**.** *Every nonempty subset of* $\mathbb{R}$ *bounded above has a least upper bound.*

Proof. Let S be such a subset. Start with an upper bound b_0 and a number a_0 that is below some element of S. Bisect at each step. If the midpoint is an upper bound, use it as the new upper endpoint. Otherwise use it as the new lower endpoint. The nested intervals have lengths $2^{-n}(b_0 - a_0)$ tending to zero, so their two endpoint sequences are Cauchy with a common limit b. Every $s \in S$ satisfies $s \le b_n$ for all n, hence $s \le b$. For every $c < b$, some a_n exceeds c. By construction some element of S exceeds that a_n, so c fails to bound S. Thus b is the least upper bound. □

18.2 An irrational number by bisection

Construct rational endpoints $a_0 = 1$ and $b_0 = 2$. At step n, let $m_n = (a_n + b_n)/2$. If $m_n^2 \le 2$, take the next interval to be $[m_n, b_n]$. Otherwise take $[a_n, m_n]$. The intervals are nested, have length 2^{-n} at step n, and satisfy $a_n^2 \le 2 \le b_n^2$. For $k \ge n$, a_k lies between the endpoints at step n. This proves that (a_n) is Cauchy. The upper endpoints give the same class x since $|b_n - a_n| = 2^{-n}$. Furthermore

$$0 \le 2 - a_n^2 \le b_n^2 - a_n^2 = (b_n - a_n)(b_n + a_n) \le 4 \cdot 2^{-n}.$$

The class of (a_n^2) is therefore 2. By the definition of multiplication this is x^2, and $x > 0$ since $a_n \ge 1$. There can be only one positive solution, since $x^2 = y^2$ implies $(x - y)(x + y) = 0$. Write this class as $\sqrt{2}$.

If a rational p/q, written with coprime integers and $q > 0$, had square 2, then $p^2 = 2q^2$ would force p even. Substitution of $p = 2r$ gives $q^2 = 2r^2$ and forces q even as well. This contradicts coprimality. The example exhibits a rational Cauchy sequence with a limit in $\mathbb{R}$ and no limit in $\mathbb{Q}$.

18.3 Completing a metric space

A metric on a set X is a symmetric map $d_X \colon X \times X \to \mathbb{R}_{\ge 0}$ satisfying the triangle inequality and with $d_X(x, y) = 0$ exactly when $x = y$. A sequence (x_n) is Cauchy if for each real $\epsilon > 0$ all sufficiently late terms are pairwise within ϵ. A metric space is *complete* if every Cauchy sequence converges. The empty space is allowed and is complete. A map $f \colon X \to Y$ is *1-Lipschitz* if $d_Y(f(x), f(y)) \le d_X(x, y)$ for every pair x, y. With these maps, metric spaces form a category $\mathbf{Met}_1$. Write $\mathbf{CMet}_1$ for its full subcategory of complete spaces, with inclusion J. Fullness

means that every 1-Lipschitz map between complete spaces remains a morphism in the subcategory.

For Cauchy sequences (x_n) and (y_n) in X, the sequence $d_X(x_n, y_n)$ is real Cauchy by the estimate

$$|d_X(x_n, y_n) - d_X(x_m, y_m)| \leq d_X(x_n, x_m) + d_X(y_n, y_m),$$

which follows by applying the triangle inequality in both directions. Its limit exists by Theorem 18.2. Declare $(x_n) \sim (y_n)$ when this limit is zero. Reflexivity and symmetry follow from the metric laws. Transitivity follows by taking limits in the triangle inequality. The quotient set $\widehat{X}$ of classes has metric

$$d_{\widehat{X}}([x_n], [y_n]) = \lim_{n\to\infty} d_X(x_n, y_n).$$

The same estimate with replacement sequences in place of index m shows independence of representatives. Nonnegativity, symmetry, and the triangle inequality pass to the limit. The relation defining classes ensures that distance zero implies equality.

The map $\eta_X \colon X \to \widehat{X}$ sends a point to its constant sequence. It is an isometry, a map preserving all distances. Its image is dense. Given $[x_n]$ and $\epsilon > 0$, take a Cauchy tail where distances are less than $\epsilon/2$. For each m in that tail,

$$d_{\widehat{X}}([x_n], \eta_X(x_m)) = \lim_{n\to\infty} d_X(x_n, x_m) \leq \epsilon/2.$$

In particular the points $\eta_X(x_m)$ converge to $[x_n]$.

Proposition 18.4. *The space $\widehat{X}$ is complete.*

Proof. Let (z_j) be a Cauchy sequence in $\widehat{X}$. By density choose $x_j \in X$ with $d_{\widehat{X}}(z_j, \eta_X(x_j)) < 2^{-j}$. The isometry property and the triangle inequality give

$$d_X(x_j, x_k) \leq 2^{-j} + d_{\widehat{X}}(z_j, z_k) + 2^{-k}.$$

Thus (x_j) is Cauchy. Put $z = [x_j]$. As just shown, $\eta_X(x_j) \to z$, and their distances to z_j go to zero. The triangle inequality gives $z_j \to z$. □

18.4 The universal extension

Theorem 18.5 (Completion is left adjoint to inclusion)**.** *For every metric space X and complete metric space Y, precomposition with η_X gives a natural bijection*

$$\mathrm{Hom}_{\mathbf{CMet}_1}(\widehat{X}, Y) \cong \mathrm{Hom}_{\mathbf{Met}_1}(X, JY).$$

Proof. Given a 1-Lipschitz map $f\colon X \to JY$, set $\overline{f}([x_n]) = \lim_n f(x_n)$. Since f decreases distances, $(f(x_n))$ is Cauchy and hence converges in Y. Equivalent sequences map to sequences with distances tending to zero, so the limits are equal. For two classes we have

$$d_Y(\overline{f}([x_n]), \overline{f}([y_n])) = \lim_n d_Y(f(x_n), f(y_n)) \le d_{\widehat{X}}([x_n], [y_n]).$$

The equality follows from the triangle inequality applied to the distances from each term to its limit. Thus $\overline{f}$ is 1-Lipschitz. On constant sequences it agrees with f. Any 1-Lipschitz extension is continuous and must send the approximants $\eta_X(x_n)$ to $f(x_n)$. Its value at their limit must therefore be $\overline{f}([x_n])$.

For a 1-Lipschitz map $u\colon X' \to X$, there is an induced map $\widehat{u}\colon \widehat{X}' \to \widehat{X}$ with $\widehat{u}([x_n]) = [u(x_n)]$. This formula respects classes and distances, and it preserves identities and composition. It gives the naturality square for the maps η_X.

$$\begin{array}{ccc} X' & \xrightarrow{\eta_{X'}} & \widehat{X}' \\ \downarrow{\scriptstyle u} & & \downarrow{\scriptstyle \widehat{u}} \\ X & \xrightarrow{\eta_X} & \widehat{X} \end{array}$$

Composing with η is natural in X by this square and natural in Y by associativity of composition. Since each such function on hom sets is bijective, the bijections are natural. □

The dashed arrow in the diagram is the extension of f.

$$\begin{array}{ccc} X & \xrightarrow{\eta_X} & \widehat{X} \\ & {\scriptstyle f}\searrow & \vdots\,{\scriptstyle \overline{f}} \\ & & \downarrow \\ & & JY \end{array}$$

When $X = JY$ is already complete, the extension of the identity sends $[y_n]$ to $\lim_n y_n$ and is inverse to η_Y. The full subcategory is called *reflective* because its inclusion has this left adjoint. Any complete metric space equipped with a map having the extension property is uniquely isometric to $\widehat{X}$ over X, since the two universal arrows extend through each other and uniqueness makes the composites identities. An isometric embedding with dense image into a complete space realizes the same construction: its extension preserves distances by the limit formula and reaches every point by taking a sequence of approximants. In particular ι realizes $\widehat{\mathbb{Q}}$ as $\mathbb{R}$.

18.5 Operations and the choice of distance

The categorical product of metric spaces with 1-Lipschitz maps uses the maximum distance.

$$d_{\max}((x,y),(x',y')) = \max\{d_X(x,x'), d_Y(y,y')\}.$$

The projections decrease distance, and the pairing of two 1-Lipschitz maps into the factors is 1-Lipschitz by the maximum formula. A sequence in the product is Cauchy precisely when both coordinates are Cauchy. This yields an isometric bijection $[(x_n,y_n)] \mapsto ([x_n],[y_n])$ from $\widehat{X \times Y}$ to $\widehat{X} \times \widehat{Y}$. Indeed the distance agrees because the maximum commutes with a limit of pairs of real numbers, and any two classes can be paired termwise. The one-point space is both terminal and complete. This shows that completion preserves finite products.

The completion theorem applies to addition if we equip the pairs with the sum distance $d_1((a,b),(a',b')) = |a-a'| + |b-b'|$. The estimate $|(a+b)-(a'+b')| \le d_1((a,b),(a',b'))$ is exactly the 1-Lipschitz bound. Rational pairs are dense in the complete space of real pairs for this distance, by choosing close rational approximations to each coordinate. Therefore addition on rational pairs extends to addition on real pairs as follows.

$$\begin{array}{ccc} (\mathbb{Q}^2, d_1) & \xrightarrow{\iota\times\iota} & (\mathbb{R}^2, d_1) \\ & \searrow^{\iota\circ+} & \downarrow{\scriptstyle +} \\ & & \mathbb{R} \end{array}$$

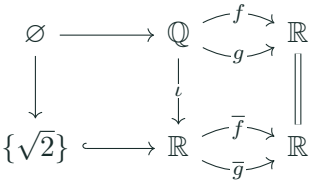
The value on $([a_n],[b_n])$ is $\lim_n (a_n+b_n) = [a_n+b_n]$, as in the quotient construction. For the categorical product distance $d_{\max}$, addition has Lipschitz bound 2, and the pairs $(0,0),(1,1)$ attain that bound. Thus the distance on the domain matters to the hom set in the universal property.

Multiplication has no finite global Lipschitz bound for either of these distances. The inputs $(0,N),(1,N)$ are at distance 1, and their products differ by N. On each box $[-B,B]^2$ it has Lipschitz bound B for the sum distance, by the product estimate in Section 18.1. Boundedness of Cauchy sequences is what makes termwise multiplication work on their classes. It also explains continuity of real multiplication on each bounded box.

18.6 Equalizers and continuous maps

Example 18.6 (Completion can change an equalizer)**.** Let $f, g\colon \mathbb{Q} \to \mathbb{R}$ be the inclusion and the constant map with value $\sqrt{2}$. Both are 1-Lipschitz. Their equalizer in $\mathbf{Met}_1$ is the empty space, since no rational equals $\sqrt{2}$. An equalizer of two 1-Lipschitz maps in this category is the subspace on which the maps agree, with the restricted distance. The corresponding set-theoretic factorization

preserves its Lipschitz bound because the inclusion is isometric. After completion, the parallel maps become the identity and a constant map from $\mathbb{R}$ to $\mathbb{R}$. They have equalizer $\{\sqrt{2}\}$, a complete space. The completion of the original equalizer is still empty. Completion preserves finite products and can fail to preserve equalizers.



The morphisms in the completion theorem matter because continuity by itself may allow a Cauchy sequence to have unbounded image. For instance the continuous map $x \mapsto 1/x$ from $(0, 1)$ to $\mathbb{R}$ sends $1/(n+1)$ to $n+1$. A continuous extension to $[0, 1]$ would require those values to tend to its value at 0, which is impossible. The inclusion $(0, 1) \to [0, 1]$ is a dense isometry into a complete space, so this is the metric completion of the domain.

Uniform continuity, where a single input tolerance works throughout the domain for each output tolerance, suffices for extending maps to a complete target. It sends Cauchy sequences to Cauchy sequences and sends equivalent sequences to equivalent sequences. These two facts establish the same formula $\overline{f}([x_n]) = \lim_n f(x_n)$. For a given output tolerance $\epsilon > 0$, choose an input tolerance $\delta > 0$ on X for $\epsilon/2$. If two points of $\widehat{X}$ are at distance less than δ, the distances between sufficiently late terms of their representative sequences are also less than δ. The image limits are therefore at distance at most $\epsilon/2$. Thus the extension is uniformly continuous. Density again gives uniqueness. The 1-Lipschitz case records the sharper fact that the extension preserves the specific input-to-output distance bound.

CHAPTER 19

Differentiation from the tangent functor

A differentiable map carries a point and a velocity to a new point and a new velocity. The derivative is the linear map on velocities over a chosen point. Keeping the two outputs together gives a map whose ordinary composition law is the chain rule. The same construction can be recovered from evaluations in an algebra with a square-zero element.

Throughout, $U \subseteq E$, $V \subseteq F$, and $W \subseteq G$ are open subsets of finite-dimensional real vector spaces. We use Euclidean norms in chosen coordinates. The coordinate descriptions of the constructions will specify how they behave under a change of variables.

19.1 The linear term of a map

Definition 19.1. A map $f\colon U \to V$ is *differentiable at* $x \in U$ if there is a linear map $L\colon E \to F$ and a remainder $r(h)$ such that, for $x + h \in U$ sufficiently close to x,

$$f(x+h) = f(x) + Lh + r(h), \qquad \lim_{h \to 0,\, h \neq 0} \frac{\|r(h)\|}{\|h\|} = 0.$$

We write $r(h) = o(\|h\|)$. The linear map L is called the *derivative* and is denoted by Df_x.

To see uniqueness, suppose that two linear maps L, L' satisfy the condition. For a nonzero vector v and a sufficiently small nonzero real number t, subtraction at $h = tv$ gives

$$\frac{\|(L - L')(tv)\|}{\|tv\|} = \frac{\|(L - L')v\|}{\|v\|}.$$

The left side tends to zero because it is the quotient associated to the difference of the two remainders. Hence $(L - L')v = 0$. The linear map is therefore determined on all vectors. Its value on v is also the limit of $(f(x + tv) - f(x))/t$ as $t \to 0$. The single remainder condition controls these increments as the direction varies.

A linear map is its own derivative at every point. For the multiplication map $m\colon \mathbb{R}^2 \to \mathbb{R}$,

$$m(a+s, b+t) - m(a,b) = bs + at + st.$$

The bound $|st| \le (s^2+t^2)/2$ makes the last term a remainder of the required order, so $Dm_{(a,b)}(s,t) = bs + at$. Taking h in each coordinate direction in turn gives the columns of the derivative matrix in general.

Proposition 19.2 (Chain rule)**.** *If $f\colon U \to V$ is differentiable at x and $g\colon V \to W$ is differentiable at $y = f(x)$, then*

$$D(g \circ f)_x = Dg_{f(x)} \circ Df_x.$$

Proof. Write $L = Df_x$, $M = Dg_y$, and $k(h) = f(x+h) - f(x) = Lh + r(h)$. Finite-dimensional linear maps are bounded, so some constant C satisfies $\|k(h)\| \le C\|h\|$ for all sufficiently small h. In particular $k(h) \to 0$. If $s(k)$ is the remainder of g at y, substitution gives

$$g(f(x+h)) = g(f(x)) + MLh + Mr(h) + s(k(h)).$$

The term $Mr(h)$ is $o(\|h\|)$ by boundedness. Given $\eta > 0$, for sufficiently small h we have $\|s(k(h))\| \le \eta\|k(h)\| \le \eta C\|h\|$. This also holds when $k(h) = 0$, since $s(0) = 0$. Both terms are therefore part of a remainder of order $o(\|h\|)$. Uniqueness gives the claimed derivative. □

19.2 The tangent functor and its projection

The category used here has open subsets of finite-dimensional real vector spaces as objects and smooth maps as arrows. A map is smooth when it has continuous derivatives of every order, obtained by iterating the derivative construction. In particular $x \mapsto Df_x$ is a smooth map into the vector space of linear maps $E \to F$.

Put $TU = U \times E$, an open subset of $E \times E$, and define

$$Tf\colon TU \longrightarrow TV, \qquad Tf(x,v) = (f(x), Df_x(v)).$$

This is smooth because the evaluation of a linear map on a vector is a polynomial in their coordinates. The derivative of the identity is the identity, so $T(\mathrm{id}_U) = \mathrm{id}_{TU}$, and the chain rule gives $T(g \circ f) = Tg \circ Tf$. These are precisely the two laws for T to be a functor.

The projections $p_U\colon TU \to U$ form a natural transformation $p\colon T \Rightarrow \mathrm{Id}$. The zero sections $z_U(x) = (x, 0)$ form a natural transformation $z\colon \mathrm{Id} \Rightarrow T$. The naturality squares record the point over which a vector lies and the preservation of zero vectors.

$$\begin{array}{ccc} TU & \xrightarrow{Tf} & TV \\ \downarrow{\scriptstyle p_U} & & \downarrow{\scriptstyle p_V} \\ U & \xrightarrow{f} & V \end{array} \qquad \begin{array}{ccc} U & \xrightarrow{f} & V \\ \downarrow{\scriptstyle z_U} & & \downarrow{\scriptstyle z_V} \\ TU & \xrightarrow{Tf} & TV \end{array}$$

The fiber $T_xU = p_U^{-1}(x)$ has the vector space structure of E. Naturality of p makes Tf restrict to a map $T_xU \to T_{f(x)}V$. This restriction, identified with its map of vector coordinates, is Df_x.

A smooth curve $c\colon (-\delta, \delta) \to U$ through x has velocity $c'(0) = Dc_0(1)$. The line $t \mapsto x + tv$ is a curve through x with velocity v after shrinking δ if necessary. The chain rule gives $(f \circ c)'(0) = Df_x(c'(0))$. Thus T_xU can also be described as the curves through x identified when their velocities agree, with Tf acting by postcomposition.

For a smooth change of coordinates φ with smooth inverse, $T\varphi$ has inverse $T(\varphi^{-1})$ by functoriality. In particular a change of chart transports a tangent vector by the linear isomorphism $D\varphi_x$. Successive changes of chart transport it by the composite of their linear maps.

19.3 Products and the pullback of covectors

The usual cartesian product is a product in the category of smooth open sets. For a map into a product, its derivative pairs the derivatives of its two components. Indeed, pair their two first-order expansions. The sum of the norms of the two remainders bounds the norm of the paired remainder. Applying this fact to projections or differentiating them as linear maps gives a product comparison

$$(T\pi_U, T\pi_V)\colon T(U \times V) \xrightarrow{\cong} TU \times TV.$$

On elements it sends $((x, y), (u, v))$ to $((x, u), (y, v))$. Its inverse is the reverse permutation of coordinates. Naturality follows by composing with the projections of the target. The one-point space, open in the zero-dimensional vector space, also has a one-point tangent space. Thus T preserves finite products.

The dual vector space $T_x^*U = \mathrm{Hom}_{\mathbb{R}}(T_xU, \mathbb{R})$ uses the linear duality of Chapter 14. Its elements are called *covectors*. Evaluation $\mathrm{ev}\colon E^* \times E \to \mathbb{R}$ is bilinear. An increment (β, v) at (α, u) changes the evaluation by $\beta(u) + \alpha(v) + \beta(v)$. The bound $|\beta(v)| \le \|\beta\|\|v\|$ for the operator norm gives the derivative

$$D\mathrm{ev}_{(\alpha,u)}(\beta, v) = \beta(u) + \alpha(v).$$

For $\lambda \in T_{f(x)}^*V$, precomposition with Df_x gives the covector

$$(Df_x)^*\lambda = \lambda \circ Df_x \in T_x^*U.$$

Evaluation of the two corresponding pairs gives the same scalar.

$$\begin{array}{ccc} T_{f(x)}^*V \times T_xU & \xrightarrow{(Df_x)^* \times \mathrm{id}} & T_x^*U \times T_xU \\ \downarrow{\scriptstyle \mathrm{id} \times Df_x} & & \downarrow{\scriptstyle \mathrm{ev}} \\ T_{f(x)}^*V \times T_{f(x)}V & \xrightarrow{\qquad \mathrm{ev} \qquad} & \mathbb{R} \end{array}$$

This is a fiberwise construction. Writing $T^*U = U \times E^*$, the corresponding arrow over all x has domain $U \times_V T^*V$. A point of that pullback is a pair of x and a covector over $f(x)$. The arrow sends the pair to the pulled-back covector over x.

$$\begin{array}{ccccc} T^*U & \longleftarrow & U \times_V T^*V & \longrightarrow & T^*V \\ \downarrow & & \downarrow & & \downarrow \\ U & \xleftarrow{\mathrm{id}_U} & U & \xrightarrow{f} & V \end{array}$$

The chain rule and associativity of composition give

$$(D(g \circ f)_x)^* \lambda = (Df_x)^* \big((Dg_{f(x)})^* \lambda \big).$$

Thus covectors travel through a composite in the reverse order of the forward tangent maps.

19.4 Velocities as algebra maps

Let $A = \mathbb{R}[\epsilon]/(\epsilon^2)$, the algebra of *dual numbers*, and let $\rho \colon A \to \mathbb{R}$ send ϵ to zero. Each element is uniquely $a + b\epsilon$. Let $C^\infty_x(U)$ denote the real algebra of germs of smooth real-valued functions at x. A germ is represented by a function on a neighborhood of x, where two representatives agree as germs when they agree on a possibly smaller neighborhood. Write ev_x for evaluation at x.

Proposition 19.3. *The vectors $v \in T_xU$ correspond bijectively to unital real algebra maps $\Phi \colon C^\infty_x(U) \to A$ with $\rho \circ \Phi = \mathrm{ev}_x$. The map for v is*

$$\Phi_v(a) = a(x) + Da_x(v)\epsilon.$$

Proof. Differentiation respects sums and real scalars by adding and scaling the first-order expansions. For products, apply the chain rule to the pair of functions and the real multiplication map. This gives $D(ab)_x(v) = a(x)Db_x(v) + b(x)Da_x(v)$, exactly the coefficient of ϵ in $\Phi_v(a)\Phi_v(b)$. The derivative of a constant is zero, so Φ_v fixes real scalars and has the required reduction to $\mathbb{R}$.

Conversely write $\Phi(a) = a(x) + \delta(a)\epsilon$. Multiplicativity yields $\delta(ab) = a(x)\delta(b) + b(x)\delta(a)$. Let $z_1, \dots, z_n$ be linear coordinates, and let v have components $v_i = \delta(z_i - z_i(x))$. For each germ a, the fundamental theorem of calculus on a sufficiently small ball centered at x gives

$$a(z) - a(x) = \sum_i (z_i - z_i(x)) a_i(z), \qquad a_i(z) = \int_0^1 \partial_i a\big(x + t(z - x)\big)\, dt.$$

Each a_i is smooth. Derivatives in z can be passed under the integral since the derivatives of the integrand are continuous on the product of any closed

smaller ball with the compact interval $[0,1]$. There the difference quotients converge uniformly by the mean value theorem and uniform continuity. At x, $a_i(x) = \partial_i a(x)$. Applying δ to the displayed sum and using that $z_i - z_i(x)$ vanishes at x gives

$$\delta(a) = \sum_i v_i \partial_i a(x) = Da_x(v).$$

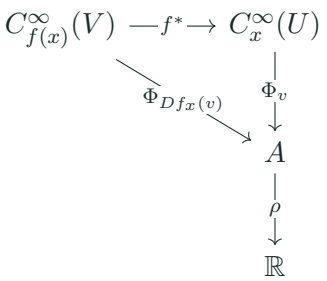
The coordinate germs determine all v_i, proving uniqueness. □

Write I_x for the ideal of germs vanishing at x. On I_x, the map $a \mapsto Da_x$ is linear and surjects onto E^*, since any linear functional α occurs for the germ $z \mapsto \alpha(z - x)$. The product rule shows that I_x^2, consisting of finite sums of products of two elements of I_x, lies in its kernel. If $a(x) = 0$ and $Da_x = 0$, every a_i in the integral formula vanishes at x. The formula then expresses a as an element of I_x^2. Thus

$$I_x/I_x^2 \cong T_x^*U, \qquad C_x^\infty(U)/I_x^2 \cong \mathbb{R} \oplus T_x^*U.$$

In the second isomorphism any product of two covector terms is zero. This algebra records just the value and first derivative of each germ. The derivation description of tangent vectors extends to smooth manifolds [13, Chapter 3].

A smooth map $f\colon U \to V$ induces a homomorphism $f^*\colon C_{f(x)}^\infty(V) \to C_x^\infty(U)$ by composition. A representative of a germ near $f(x)$ can be composed with f near x because f is continuous. Precomposition with f^* sends Φ_v to $\Phi_{Df_x(v)}$.



To verify the upper triangle on the germ b, compare the ϵ-coefficients using $D(b \circ f)_x(v) = Db_{f(x)}(Df_x(v))$. Composition reverses the arrows on algebras, and precomposition reverses them again on the sets of evaluations in A. This recovers the covariant direction of the tangent functor from an algebra of first-order tests.

Pullback by f also takes germs vanishing at $f(x)$ to germs vanishing at x. As an algebra map it takes the square of the first ideal into the square of the second. It therefore induces a linear map $I_{f(x)}/I_{f(x)}^2 \to I_x/I_x^2$. Under the identifications with covectors, this is exactly $(Df_x)^*$, since $D(b \circ f)_x = Db_{f(x)} \circ Df_x$. The same first-order algebra thus gives both the pullback of covectors and the forward map of velocities.

19.5 A two-parameter model

Consider $f(a,b) = (a+b, ab)$, and measure its discrepancy from the target $(3,1)$ by the loss

$$\ell(u,w) = \tfrac{1}{2}(u-3)^2 + \tfrac{1}{2}(w-1)^2.$$

The tangent map first carries a parameter increment (s,t) to $(s+t, bs+at)$. The derivative of the loss is the covector

$$D\ell_{(u,w)}(p,q) = (u-3)p + (w-1)q.$$

At $(a,b) = (1,2)$, the prediction is $(3,2)$ and the loss covector selects the second coordinate. Pulling it back gives

$$D(\ell \circ f)_{(1,2)}(s,t) = D\ell_{(3,2)}(s+t, 2s+t) = 2s + t.$$

With the Euclidean inner product, the gradient is the vector representing this covector by the inner product, namely $(2,1)$. The negative gradient gives the directional derivative -5. The first-order expansion therefore gives a loss decrease for all sufficiently small positive steps in that direction.

For example, taking the step size $1/10$ gives new parameters $(4/5, 19/10)$ and a prediction of $(27/10, 38/25)$. The new loss is $901/5000$, compared with the old loss of $1/2$. The choice of inner product enters the update when the covector is converted to a vector. For the inner product $\langle (s,t),(s',t')\rangle = 2ss' + tt'$, the same covector $(s,t) \mapsto 2s+t$ gives the gradient $(1,1)$.

To compute the update, the point traveled forward from parameters to predictions, and the loss covector traveled backward through the derivative of the model. This composition of covector pullbacks will return when predictions are assembled from multiple stages in machine learning.

CHAPTER 20

Probability from stochastic maps

A function sends each input to one output. A stochastic map sends each input to a distribution of possible outputs. If the input already has a distribution, the output is obtained by mixing the distributions assigned to its possible values. The composition law for stochastic maps makes this rule precise. A joint distribution then admits two factorizations, one for each order in which its coordinates can be sampled. Comparing the factorizations gives Bayes' rule.

20.1 Finite distributions and substitution

For a set X, let $\mathcal{D}X$ be the set of functions $p\colon X \to [0,1]$ with finite support $\operatorname{supp}(p) = \{x \in X \mid p(x) > 0\}$ and total weight $\sum_x p(x) = 1$. We write $p = \sum_x p(x)\delta_x$, where δ_x is the point mass at x. The notation records how to reconstruct p from a finite list of elements with weights. A sum indexed by a whole set will always have only finitely many nonzero terms here. On the empty set there is no distribution of total weight one.

A function $f\colon X \to Y$ sends a distribution on X to one on Y by sending every mass at x to $f(x)$. If several elements of X have the same image, their weights add. Thus

$$(\mathcal{D}f)(p)(y) = \sum_{x:f(x)=y} p(x).$$

The support lies in $f(\operatorname{supp}(p))$, and each input contributes its weight exactly once when we sum over Y. Pushing along an identity leaves every weight at its original point. Pushing along f and then $g\colon Y \to Z$ adds the weights over all x with $g(f(x)) = z$ at each z. This is exactly the pushforward along $g \circ f$, so $\mathcal{D}\colon \mathbf{Set} \to \mathbf{Set}$ is a functor.

A distribution $P \in \mathcal{D}\mathcal{D}X$ assigns weights to distributions on X. Selecting p with weight $P(p)$ and then selecting x with weight $p(x)$ contributes $P(p)p(x)$ to the total weight at x. The resulting mixture defines a function

$$\mu_X\colon \mathcal{D}\mathcal{D}X \longrightarrow \mathcal{D}X, \qquad \mu_X(P)(x) = \sum_p P(p)p(x).$$

Its support is contained in the finite union of the supports of the p in $\operatorname{supp}(P)$. Its total weight is $\sum_p P(p) \sum_x p(x) = 1$. Pushforward and mixture commute

because summing over p and summing over a fiber of f can be done in either order. In particular the following square commutes.

$$\begin{array}{ccc} \mathcal{D}\mathcal{D}X & \xrightarrow{\mu_X} & \mathcal{D}X \\ \downarrow{\scriptstyle \mathcal{D}\mathcal{D}f} & & \downarrow{\scriptstyle \mathcal{D}f} \\ \mathcal{D}\mathcal{D}Y & \xrightarrow{\mu_Y} & \mathcal{D}Y \end{array}$$

The point-mass function $\eta_X \colon X \to \mathcal{D}X$, $x \mapsto \delta_x$, is natural since pushing δ_x along f gives $\delta_{f(x)}$. The maps η and μ satisfy the following identities.

$$\mu_X \circ \eta_{\mathcal{D}X} = \mathrm{id}_{\mathcal{D}X} = \mu_X \circ \mathcal{D}\eta_X, \qquad \mu_X \circ \mu_{\mathcal{D}X} = \mu_X \circ \mathcal{D}\mu_X.$$

The first equality says that a mixture with one term of weight one has that term as its value. The second says that the mixture of point masses with weights $p(x)$ reconstructs p. For the last equality, a distribution $Q \in \mathcal{D}\mathcal{D}\mathcal{D}X$ yields at x the sum $\sum_{P,p} Q(P)P(p)p(x)$ on either route. Evaluating the inner mixture first and evaluating the outer mixture first simply group the same finite contributions in two ways. A functor with natural maps η and μ satisfying these identities is a *monad*. Here η introduces an outcome with certainty and μ composes two stages of sampling into one distribution. Jacobs [14] develops this monad and its general relationship to Kleisli categories.

20.2 Stochastic maps and matrix multiplication

A *stochastic map* from X to Y means a function $k \colon X \to \mathcal{D}Y$. Write $k(y \mid x) = k(x)(y)$. If $\ell \colon Y \to \mathcal{D}Z$ is another stochastic map, their composition is induced by the mixture operation.

$$X \xrightarrow{k} \mathcal{D}Y \xrightarrow{\mathcal{D}\ell} \mathcal{D}\mathcal{D}Z \xrightarrow{\mu_Z} \mathcal{D}Z \qquad (\ell \odot k)(z \mid x) = \sum_y k(y \mid x)\ell(z \mid y).$$

For each x, only the finitely many y in the support of $k(x)$ are used, and each supplies finitely many possible z. The identity on X is η_X, or $\mathrm{id}_X(x' \mid x) = 1$ for $x' = x$ and zero otherwise. Inserting a point mass into either position of the composition formula leaves the other map unchanged. Given a further stochastic map m out of Z, both possible compositions have weight $\sum_{y,z} k(y \mid x)\ell(z \mid y)m(w \mid z)$ at w for input x. This verifies associativity directly. The resulting category is the *Kleisli category* of $\mathcal{D}$. We write $\mathbf{Stoch}_{\mathrm{fin}}$ for its full subcategory on finite sets. In diagrams in this category, an arrow $X \to Y$ means a stochastic map, with its representing function $X \to \mathcal{D}Y$ understood.

When X and Y are finite and their elements have been ordered, put the input x in the row position and the output y in the column position. Then $K_{xy} = k(y \mid x)$ is a nonnegative matrix whose rows each sum to one. Every

such matrix defines a stochastic map. With this convention, executing k and then ℓ is matrix multiplication in the order KL, since $(KL)_{xz} = \sum_y K_{xy}L_{yz}$. A deterministic function $f\colon X \to Y$ gives the stochastic map $x \mapsto \delta_{f(x)}$, whose row has a single unit entry. Composition of those matrices recovers composition of functions.

Write $1 = \{*\}$. A stochastic map $1 \to X$ gives one distribution on X and is called a *state*. Composing a state p with $k\colon X \to Y$ gives the output state $q = k \odot p$ with

$$q(y) = \sum_x p(x)k(y \mid x).$$

This is the law of total probability. In the matrix convention, p is a row vector and $q = pK$. The unique map $X \to 1$ sends every input to δ_*. This is the discard map, expressing that the total weight of each row is one.

20.3 Independent channels and joint outputs

For stochastic maps $k\colon X \to Y$ and $k'\colon X' \to Y'$, define their tensor $k \otimes k'\colon X \times X' \to Y \times Y'$ by sampling the two outputs independently for each pair of inputs.

$$(k \otimes k')((y, y') \mid (x, x')) = k(y \mid x)k'(y' \mid x').$$

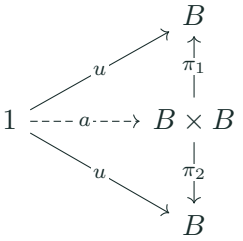
The weights are normalized since the sum of products splits into the product of the two marginal sums, both one. The possible outputs have finite support because they lie in the product of two finite supports. For further maps ℓ and ℓ', summing over the intermediate pair splits into separate sums over the two intermediate coordinates. Thus

$$(\ell \otimes \ell') \odot (k \otimes k') = (\ell \odot k) \otimes (\ell' \odot k').$$

The tensor of two identities is the identity on the pair. The ordinary rebracketing of triples, identification of $1 \times X$ and $X \times 1$ with X, and swap of pairs give the associativity, unit, and symmetry isomorphisms. They act on products of weights by relabeling, so their naturality follows from the formula. Each coherence equation compares two deterministic rearrangements with the same effect on tuples. This is a symmetric monoidal structure on $\mathbf{Stoch}_{\mathrm{fin}}$.

On a state of $X \times Y$, the two deterministic projections give the two marginals. Their values do not determine the joint state. For $B = \{0, 1\}$, let u be the fair state on B, assigning weight $1/2$ to each bit. The distribution $u \otimes u$ gives weight $1/4$ to each pair. The distribution d with weights $d(0, 0) = d(1, 1) = 1/2$ gives weight zero to the other two pairs. Both fill the same diagram as the dashed

map a.



The universal property in Chapter 2 would require a unique such map. In fact B paired with itself has no categorical product in $\mathbf{Stoch}_{\mathrm{fin}}$. To see this, suppose P were a product with stochastic projections. Each of the four pairs of point masses at $b, c \in B$ would then arise from a unique state of P. For any state giving point-mass marginals, all points in its support must give those same marginals, since its weights are positive and mass cannot cancel. Picking one such point therefore gives a point mass $\delta_{p_{bc}}$ with those marginals, and the four points p_{bc} must be distinct. The states $\frac{1}{2}(\delta_{p_{00}} + \delta_{p_{11}})$ and $\frac{1}{2}(\delta_{p_{01}} + \delta_{p_{10}})$ on P would then be distinct states with the same fair marginals, violating uniqueness.

The diagonal function $\Delta_X \colon X \to X \times X$ copies an already sampled value. Applying Δ_Y after sampling via $k \colon X \to Y$ puts all mass on the diagonal, with mass $k(y \mid x)$ at (y, y) for input x. Copying x first and then applying $k \otimes k$ gives a product distribution with mass $k(y \mid x)k(y' \mid x)$ at (y, y'). These maps agree exactly when k is deterministic. If a row assigns positive mass to two different outcomes, the product gives positive mass to their ordered pair off the diagonal. If every row is a point mass, both routes give the same pair with certainty. For the fair state u, copying its sampled output gives d while two independent samples give $u \otimes u$.

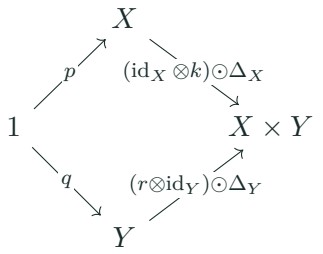
20.4 Two factorizations and Bayes' rule

Suppose $p \colon 1 \to X$ describes a prior distribution and $k \colon X \to Y$ describes the possible observations given x. Both sets are finite in this section. Sample from p, keep that sample, and generate an observation from it using k. Keeping both values gives the joint state $j \colon 1 \to X \times Y$.

$$1 \xrightarrow{p} X \xrightarrow{\Delta_X} X \times X \xrightarrow{\mathrm{id}_X \otimes k} X \times Y \qquad j(x, y) = p(x)k(y \mid x).$$

The X-marginal of j is p, and the Y-marginal is $q = k \odot p$. More generally, a joint state j with marginal q on Y can be generated by first sampling y from q. For a given y, a row $r(- \mid y)$ for sampling x must satisfy $q(y)r(x \mid y) = j(x, y)$. When $q(y) > 0$, division gives the row and $\sum_x j(x, y) = q(y)$ verifies its normalization. When $q(y) = 0$, nonnegativity forces $j(x, y) = 0$ for all x. The row can then be any distribution on X, for example the X-marginal of j. No sample from q uses that row.

This factorization of a joint state into a marginal and a conditional map is a *disintegration*. Interchanging X and Y gives a disintegration in the other direction. For the joint state built from p and k, the two routes agree as stochastic maps to $X \times Y$.



At each (x, y), equality of the two routes says $p(x)k(y \mid x) = q(y)r(x \mid y)$. This gives Bayes' rule on the support of q.

$$r(x \mid y) = \frac{p(x)k(y \mid x)}{\sum_{x'} p(x')k(y \mid x')} \qquad (q(y) > 0).$$

The map r is a *Bayesian inverse* of k relative to p. It depends on the prior and is determined by the joint state only on the observations of positive probability. Projecting the diagram to X gives $r \odot q = p$. Thus averaging the posteriors with the law of the observations recovers the prior.

Example 20.1 (Updating after an observation)**.** Let $X = \{h, \bar{h}\}$ indicate whether a hypothesis holds, with $p(h) = 1/10$. A test reports an outcome from $Y = \{+, -, ?\}$. Conditional on h, it reports $+$ with probability $4/5$. Conditional on $\bar{h}$, that probability is $1/5$. Under either hypothesis the outcome ? has probability zero, so $-$ receives the remaining mass. Multiplying each row of k by the corresponding prior weight gives the joint state.

	$+$	$-$	?	p
h	$4/50$	$1/50$	0	$5/50$
$\bar{h}$	$9/50$	$36/50$	0	$45/50$
q	$13/50$	$37/50$	0	1

The two positive columns can be normalized to give the posterior probabilities $r(h \mid +) = 4/13$ and $r(h \mid -) = 1/37$. The observed positive result raises the prior $1/10$ to $4/13$. Setting the whole row $r(\cdot \mid ?)$ to the prior gives one total stochastic map. This choice is arbitrary because its joint column has zero mass. Averaging the posteriors at h gives

$$\frac{13}{50}\frac{4}{13} + \frac{37}{50}\frac{1}{37} = \frac{1}{10}.$$

Each posterior at h after an observation supported by $k(\cdot \mid h)$ is strictly less than one. Hence $(r \odot k)(h \mid h) < 1$, whereas the identity on X gives weight one there. The inverse property here compares samplings of a joint state with a specified prior.

20.5 From finite sums to kernels

The finite formulas use three operations on distributions. Pushforward relocates weights, composition mixes conditional distributions, and the tensor constructs a joint distribution from independent samples. For measurable spaces the analogue of k is a *Markov kernel*. Each input x specifies a probability measure $k(- \mid x)$, and each measurable output event B specifies a measurable function $x \mapsto k(B \mid x)$. The composition formula then takes the form

$$(\ell \odot k)(C \mid x) = \int_Y \ell(C \mid y)\, k(dy \mid x)$$

for a measurable event C in the final output space. If X and Y are standard Borel spaces, such as Euclidean spaces with their Borel sets, a joint state admits disintegrations. For the disintegration over Y, the finite identity has the following integral form on measurable rectangles $A \times B$.

$$j(A \times B) = \int_B r(A \mid y)\, q(dy).$$

This is an identity of joint distributions even if every individual observation has probability zero. Fritz gives this disintegration formula for kernels [15, Example 11.3]. The same joint-state identity will guide the later treatment of Bayesian learning.

20.6 Combinatorics of walks on a cycle

Start at 0 in $C_m = \mathbb{Z}/m\mathbb{Z}$, for $m \geq 3$, and choose each step independently, going left or right with equal probability. Every word of n signed steps has probability 2^{-n}. Its endpoint is the sum of the steps modulo m. A word with j right steps contributes the endpoint $2j - n$, and there are $\binom{n}{j}$ such words. Thus the number of walks ending at a is

$$N_n(a) = \sum_{\substack{0 \leq j \leq n \\ 2j - n \equiv a \bmod m}} \binom{n}{j}.$$

For a five-cycle and six steps, $2j - 6$ is in $[-6, 6]$. Its only even multiple of five in that range is zero, so there are $\binom{6}{3} = 20$ return walks among 64 words.

Characters turn this congruence count into a sum of m powers. By the orthogonality in Section 8.6,

$$\mathbf{1}_{b=a} = \frac{1}{m} \sum_{k=0}^{m-1} e^{2\pi i k(b-a)/m} \qquad (a, b \in C_m).$$

Apply this indicator to the endpoint of each walk and sum over the choices of each step. Since a character of a sum is the product of its values, the sum over all words factors:

$$N_n(a) = \frac{1}{m} \sum_{k=0}^{m-1} e^{-2\pi i k a/m} \big(2\cos(2\pi k/m)\big)^n.$$

This is the same count in Fourier coordinates. The Markov operator $Pf(x) = (f(x+1) + f(x-1))/2$ has eigenvalue $\cos(2\pi k/m)$ on the k-th character, so P^n accounts for the n-th powers. For odd m the nontrivial eigenvalues have absolute value strictly less than one. Consequently $2^{-n} N_n(a)$ tends to $1/m$ for each a. For even m, the character $k = m/2$ has eigenvalue -1 and records the fact that only endpoints of the same parity as n are reachable. The same formula therefore distinguishes when walk counts approach a uniform distribution and when they retain a parity constraint.

Part V Logic, semantics, and quantification

CHAPTER 21

Propositions from subobjects

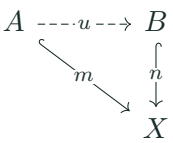
The predicate that an integer is even determines an inclusion $2\mathbb{Z} \hookrightarrow \mathbb{Z}$. To show that a function takes only even values is to factor it through that inclusion. If the function does take only even values, the factor is unique. The same factorization test defines a predicate in any category. Pullback then gives the rules for combining predicates and changing the variables on which they depend.

21.1 Subobjects and their order

Recall from Chapter 3 that a map $m\colon A \to X$ is a *monomorphism*, or *mono*, if $m \circ a = m \circ b$ implies $a = b$ for all pairs of arrows $a, b\colon T \to A$. Two monos $m\colon A \to X$ and $n\colon B \to X$ represent the same *subobject* of X if there is an isomorphism $i\colon A \to B$ with $n \circ i = m$. Thus a subobject specifies both the domain and its position in X, up to isomorphism over X. Write $\mathrm{Sub}(X)$ for the collection of these classes, using categories where it is a set when a set is needed.

In **Set**, monos are precisely the injections. If a function identifies two elements of its domain, the two maps from a singleton picking those elements witness a failure of cancellation. Conversely, an injection cancels pointwise. An injection $A \to X$ is isomorphic over X to the inclusion of its image, and two subset inclusions can be isomorphic over X only if their images coincide. Thus $\mathrm{Sub}(X)$ is the powerset of X in this case.

For subobjects $P = [m]$ and $Q = [n]$ of X, put $P \leq Q$ if m factors through n. The comparison is independent of the representatives because isomorphisms over X transport factors. There is at most one factor u in the following triangle, by cancellation against n.



Identities give reflexivity, and composition gives transitivity. If there are factors $u\colon A \to B$ and $v\colon B \to A$ in both directions, then $m \circ v \circ u = m$ and $n \circ u \circ v = n$. Cancellation gives $v \circ u = \mathrm{id}_A$ and $u \circ v = \mathrm{id}_B$. This proves antisymmetry on subobjects. For subset inclusions the order is ordinary inclusion.

We regard X as a *context*, the space of assignments to the variables in use. A predicate $P = [m]$ in context X holds along a map $x\colon T \to X$ if $x = m \circ \widetilde{x}$ for

some $\widetilde{x}\colon T \to A$. The factor is unique. The map x is a generalized element, with T allowing a family of assignments to be tested together. If $P \leq Q$, then every map satisfies P satisfies Q by composition with the factor $A \to B$. Conversely, if this implication holds for every test object T and every $x\colon T \to X$, apply it to $T = A$ and $x = m$. Thus entailment in a context is precisely the subobject order.

There is an equivalent formulation using the slice category $\mathcal{C}/X$. Its objects are arrows into X, and its maps are the commuting triangles over X. An object of a category is *subterminal* if it receives at most one map from every object. An arrow $m\colon A \to X$ is subterminal in $\mathcal{C}/X$ precisely when it is mono in $\mathcal{C}$. Indeed, a map from $x\colon T \to X$ to m in the slice is a choice of factor of x through m. Uniqueness of such factors is exactly the cancellation criterion. In this sense predicates are the objects over a context whose only data over an assignment are whether a factor exists.

21.2 Truth and conjunction

The identity of X is a mono, and every mono into X factors through it. Its class is therefore the greatest element $\top_X \in \mathrm{Sub}(X)$. Every map into X factors through id_X, so this predicate always holds.

Suppose the category has pullbacks. For $P = [m]$ and $Q = [n]$ as above, form the square

$$\begin{array}{ccc} A \times_X B & \xrightarrow{p_B} & B \\ {\scriptstyle p_A}\downarrow & \lrcorner & \downarrow{\scriptstyle n} \\ A & \xrightarrow{m} & X. \end{array}$$

Its two composites into X agree. Write r for their common value. If $r \circ a = r \circ b$, cancellation against m and n shows that the p_A- and p_B-components of a, b agree. Uniqueness in the pullback gives $a = b$, so r is mono. By construction $[r]$ lies below P, Q. If a mono $s\colon C \to X$ factors through both m, n, the two factors have common composite s into X. The pullback gives a unique map from C over X into $A \times_X B$. Thus $[r]$ is the greatest lower bound, written $P \wedge Q$.

The same argument for an arbitrary $x\colon T \to X$ shows that $P \wedge Q$ holds along x exactly when both P and Q do. The greatest lower bound therefore interprets conjunction. In the poset viewed as a category, the greatest element is terminal and the conjunction is a product. Associativity, commutativity, and idempotence of conjunction follow from the greatest-lower-bound property. In sets, $A \times_X B$ identifies with the intersection of A and B in X, recovering the set formula for conjunction.

21.3 Changing the context

Let $f\colon Y \to X$ specify a change of variables. A predicate $P = [m]$ on X should hold for $y\colon T \to Y$ after substitution exactly when P holds for $f \circ y$. Pullback provides this predicate.

$$\begin{array}{ccc} Y \times_X A & \xrightarrow{p_A} & A \\ {\scriptstyle m_f}\downarrow & \lrcorner & \downarrow{\scriptstyle m} \\ Y & \xrightarrow{f} & X \end{array}$$

In fact m_f is mono. If two maps into this pullback have the same component in Y, the square shows their components in A have equal composite with m and hence agree. The pullback uniqueness then identifies the original maps. This also proves that pulling a mono back along any map gives a mono.

Write $f^*P = [m_f]$. If m factors through n over X, the same factor composed with p_A and the map m_f yield a compatible pair into B, Y. Their induced arrow over Y gives $f^*P \leq f^*Q$. An isomorphism over X yields both inequalities, so reindexing is well defined on subobjects. The pullback of a mono along an identity has the original mono as a representative. For a second change of context $g\colon Z \to Y$, pasting the successive pullbacks gives a pullback along $f \circ g$.

$$\begin{array}{ccccc} Z \times_Y (Y \times_X A) & \longrightarrow & Y \times_X A & \longrightarrow & A \\ \downarrow & & \downarrow & & \downarrow{\scriptstyle m} \\ Z & \xrightarrow{g} & Y & \xrightarrow{f} & X \end{array}$$

The unique comparison isomorphism with $Z \times_X A$ lies over Z. After passage to subobject classes this gives the equations $\mathrm{id}_X^* P = P$ and $(f \circ g)^* P = g^* f^* P$. This is functoriality of substitution.

The substitution f^* takes $\top_X$ to $\top_Y$, as seen by pulling back id_X. It also preserves conjunction. A map $y\colon T \to Y$ satisfies $f^*(P \wedge Q)$ exactly when $f \circ y$ satisfies both P and Q. This says exactly that y satisfies $f^*P \wedge f^*Q$. Predicates that agree on all tests lie below one another by the order test of Section 21.1, so they are equal. In sets, $f^*P = f^{-1}(P)$, and these laws are the inverse-image rules for the whole set and intersections.

21.4 Equality and a calculation in context

When products exist, the map $\Delta_X = \langle \mathrm{id}_X, \mathrm{id}_X \rangle\colon X \to X \times X$ is mono, since either projection is a left inverse. Its class is the equality predicate in context $X \times X$. A pair $\langle x, y \rangle\colon T \to X \times X$ factors through Δ_X precisely when $x = y$. For two maps $s, t\colon Y \to X$, the substituted predicate $\langle s, t \rangle^*[\Delta_X]$ has the following

representative.

$$\begin{array}{ccc} E & \longrightarrow & X \\ \downarrow{\scriptstyle e} & \lrcorner & \downarrow{\scriptstyle \Delta_X} \\ Y & \xrightarrow{\langle s,t\rangle} & X \times X \end{array}$$

A map $z\colon T \to Y$ factors through e precisely when $s \circ z = t \circ z$. Thus e is the equalizer of s and t, as in Section 3.2. Equality also gives a substitution rule for predicates. On E the arrows $s \circ e$ and $t \circ e$ coincide, so $e^*s^*P = e^*t^*P$ for every subobject P of X. The two expressions give identical truth conditions wherever $s = t$.

Example 21.1 (Constraints on pairs)**.** Let $S = \{0, 1, 2, 3, 4, 5\}$ and work in the context $S \times S$. Define $s(x, y) = x + y$ and $c(x, y) = 6$ as maps to $\mathbb{Z}$. The predicate E given by pulling back $[\Delta_{\mathbb{Z}}]$ along $\langle s, c\rangle$ has extent

$$E = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}.$$

Let R be the subset of pairs with $x < y$, and let P be the even elements $\{0, 2, 4\}$ of S. The expression

$$E \wedge R \wedge \pi_1^* P$$

represents the conjunction $x + y = 6$, $x < y$, and x even. Meeting E with R leaves $(1, 5), (2, 4)$, and the last condition leaves $(2, 4)$ alone. Now take $j\colon \{1, 2, 3, 4, 5\} \to S \times S$, $j(t) = (t, 6 - t)$. Pulling the conjunction back along j yields the single allowed parameter $t = 2$. Equivalently, apply j^* to each conjunct. They become the whole parameter set, $\{1, 2\}$, and $\{2, 4\}$, whose intersection is $\{2\}$.

21.5 Boolean operations in sets

In the powerset order, the empty subset is the least element $\bot_X$. Disjunction is the join, given by union because $P \cup Q \subseteq R$ holds exactly when both $P \subseteq R$ and $Q \subseteq R$ do. Conjunction distributes over disjunction since membership in an intersection and a union is determined pointwise. The complement $X \setminus P$ has intersection $\varnothing$ and union X with P. These operations give the Boolean algebra of subobjects of X, with the same laws used in the propositional presentation of Section 5.5.2.

The implication predicate is characterized by its comparison with tests $R \subseteq X$.

$$R \cap P \subseteq Q \quad \Longleftrightarrow \quad R \subseteq (X \setminus P) \cup Q.$$

Indeed, for elements of R lying in P the left side asks for membership in Q. The elements of R outside P already belong to the complement. This shows that meeting with P is left adjoint to implication from P in the powerset order.

Taking $Q = \varnothing$ recovers negation. Inverse image preserves unions, intersections, and complements, so in sets substitution also preserves this implication.

Disjunction and implication require additional structure on the subobject order. Chapter 17 constructs implication in the frame of opens. There the largest open disjoint from an open U is the interior of $X \setminus U$. On the real line the open predicates $x > 0$ and $x < 0$ are negations of one another in this sense, and their union omits 0. The meaning of negation depends on the available predicates.

21.6 Monotone predicates and Boolean antichains

An assignment to n Boolean variables can be written as a subset $S \subseteq [n] = \{1, \dots, n\}$ of its true variables. A predicate is *monotone* if truth at S implies truth at every $T \supseteq S$. Its true assignments form an upper set in the inclusion order. By finiteness, every member of that upper set contains a minimal member. The minimal members form an *antichain*, meaning no two distinct ones are comparable. Conversely any antichain $\mathcal{A}$ determines the upper set of supersets of its members, and the resulting predicate is

$$\bigvee_{A \in \mathcal{A}} \bigwedge_{i \in A} x_i.$$

The empty antichain yields the false predicate, and the singleton $\{\varnothing\}$ yields the true predicate. The correspondence identifies monotone predicates with the choices of irredundant minimal clauses in this form.

Obtain a maximal chain of subsets by starting from $\varnothing$ and adding the elements in a permutation of $[n]$. An antichain meets each chain at most once. A subset A of size j lies on exactly $j!(n-j)!$ chains, because its elements must be the first j entries in the permutation. Counting incidences between chains and $\mathcal{A}$, and dividing by $n!$, gives the Lubell–Yamamoto–Meshalkin inequality

$$\sum_{A \in \mathcal{A}} \frac{1}{\binom{n}{|A|}} \le 1.$$

The ratio $(n-j)/(j+1)$ of consecutive binomial coefficients shows that each denominator is at most $\binom{n}{\lfloor n/2 \rfloor}$. Thus the middle rank gives both an upper bound on $|\mathcal{A}|$ and an antichain attaining that bound. This is Sperner's theorem for the Boolean lattice [69, Section 1.2].

For three variables there are 20 monotone predicates. There is one empty antichain and there are eight singletons. The two intermediate ranks each contain three subsets, giving three pairs in each rank. Between those ranks a singleton subset is incomparable with exactly its complementary pair, giving three more antichains of size two. A mixed pair cannot be extended within either rank, so only the two entire intermediate ranks give antichains of size three. The bound rules out larger antichains, and the count is $1 + 8 + 9 + 2 = 20$.

21.7 The scope of a classifier

An open subset of a space supplies a subobject in **Top** by its inclusion with the subspace topology. Section 16.5 describes those inclusions by maps to the Sierpiński space. Every pullback of $1 \to \mathbb{S}$ selecting the open point is an open inclusion, since its domain is the open inverse image with the subspace topology. This test therefore describes exactly the open predicates.

For a different subobject, take a two-element set with its indiscrete topology as X and the same set with the discrete topology as D. The underlying identity $D \to X$ is continuous and mono, since an injective underlying map cancels continuous maps. It cannot be isomorphic over X to the inclusion of an open subspace. Its image is all of X, and its inverse as a function fails continuity. Every map from a one-point space to X factors through D. Testing with id_X detects that $D \to X$ is a proper subobject.

This example rules out a classifier for all subobjects in **Top**. Suppose such a classifier $\top\colon 1 \to \Omega$ existed, and let $\chi\colon X \to \Omega$ classify $D \to X$. By the pullback property, χ takes every point of X to the point selected by $\top$. Thus χ is the composite $X \to 1 \to \Omega$, whose classified subobject would be the identity of X. This contradicts the properness of $D \to X$. The classifier axiom of Chapter 23 asks for a universal truth map with the pullback property for every mono.

CHAPTER 22

The logic–semantics and structure–semantics adjunctions

A theory determines the models that satisfy it. A category of models also determines operations that are respected by all its homomorphisms. These are two adjunctions. The first uses the relation of satisfaction between sentences and models. The second uses functions on the underlying sets of models and keeps track of their naturality. The first recovers semantic consequences. The second recovers operations, including their substitution laws.

22.1 Satisfaction and semantic closure

Let S be a set of sentences and M a set of interpretations, with a relation $m \models s$ between them. We can take these to be all sentences of a fixed language and a selected collection of structures. A theory here means any subset $T \subseteq S$. For $K \subseteq M$, define

$$\operatorname{Mod}(T) = \{m \in M \mid m \models s \text{ for all } s \in T\},$$
$$\operatorname{Th}(K) = \{s \in S \mid m \models s \text{ for all } m \in K\}.$$

Both functions reverse inclusion. Taking the opposite order on $\mathcal{P}(M)$ makes them monotone and gives the adjunction

$$\operatorname{Mod}\colon \mathcal{P}(S) \rightleftarrows \mathcal{P}(M)^{\mathrm{op}}\colon \operatorname{Th}, \qquad \operatorname{Mod} \dashv \operatorname{Th}.$$

Its hom-set equivalence is

$$\operatorname{Mod}(T) \leq_{\mathrm{op}} K \iff K \subseteq \operatorname{Mod}(T) \iff T \subseteq \operatorname{Th}(K).$$

Both inclusions assert that every pair in $K \times T$ belongs to the satisfaction relation. Speirs [3, Section 6.1] describes this Galois connection for first-order structures. This recovers the two-variable example in Section 5.5.3 for the four classical valuations of p, q.

The unit gives $T \subseteq \operatorname{Th}\operatorname{Mod}(T)$. The counit, read in the ordinary inclusion order on model classes, gives $K \subseteq \operatorname{Mod}\operatorname{Th}(K)$. Applying Mod to the first inclusion reverses it. The second inclusion at $K = \operatorname{Mod}(T)$ goes the other way. Repeating the argument with Th gives

$$\operatorname{Mod}\operatorname{Th}\operatorname{Mod}(T) = \operatorname{Mod}(T), \qquad \operatorname{Th}\operatorname{Mod}\operatorname{Th}(K) = \operatorname{Th}(K).$$

These are the triangle identities written as equalities of sets. Thus both $\operatorname{Th}\operatorname{Mod}$ and $\operatorname{Mod}\operatorname{Th}$ are monotone, extensive, and idempotent in their ordinary inclusion orders. The first closure gives the sentences that follow semantically from T relative to M. The fixed points of the second are exactly the classes of the form $\operatorname{Mod}(T)$. On the two fixed-point posets, Mod and Th are mutually inverse and reverse inclusion. Taking the opposite on the model side therefore makes them an order isomorphism.

Example 22.1 (A class enlarged by semantic closure)**.** Use first-order sentences in the language with equality and no other symbols. Take M to contain one set $\{1, \dots, n\}$ of each positive finite size and the countably infinite set $\mathbb{N}$. Let K consist of all the finite members of M. We claim that $\operatorname{Mod}\operatorname{Th}(K) = M$.

For a formula of quantifier depth r, compare two assignments of its free variables with the same equality pattern. This means that two variables have equal values in one assignment precisely when they do in the other. If both assignments leave at least r elements unused, they give the same truth value to the formula. For an atomic equality this follows from the pattern itself, and the Boolean connectives preserve the claim. For an existential quantifier, a witness can repeat an assigned element or use a new one. In the other structure choose the corresponding assigned element or a fresh one. The new assignments have the same pattern and leave at least $r - 1$ elements unused. Induction through the quantified formula applies in each direction. Universal quantification follows by negating an existential.

For a sentence, the starting assignments are empty. Therefore a sentence of depth r true in every finite member of M is true in $\mathbb{N}$ by comparing it with any finite member of size at least r. The claim follows. By comparison, the singleton $\{\mathbb{N}\}$ is a fixed point. For each $n \geq 1$, use the sentence asserting the existence of n pairwise distinct elements. Only the infinite member satisfies all of these sentences at once.

All these closures use the selected M. For example, restricting M to infinite structures would make the sentence asserting at least two elements a consequence of the empty theory. The adjunction itself is a calculation with a relation. To recover ordinary first-order consequence, one can choose M to contain a representative of every realized set of true sentences. Then soundness and completeness of the first-order proof calculus identify semantic consequence with derivability [73, Sections 1.9 and 10.8].

22.2 Equations as a category of operations

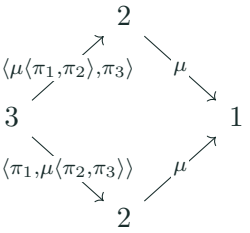
An algebraic term in variables $x_1, \dots, x_n$ describes a way of producing one output from n inputs. If a map is to produce m outputs, it can give a term for each

position. Composing two tuples of terms substitutes the first tuple into the second. The variables themselves form the identity tuple.

A *Lawvere theory* is a category L with objects $0, 1, 2, \ldots$, and with chosen projections $\pi_i^n \colon n \to 1$ that exhibit n as a product of n copies of 1. The projection π_1^1 is the identity. The object 0 is terminal, and $n + m$ is a product of n and m, by grouping projections in two blocks. A map of theories is a functor that fixes the objects and preserves the chosen projections. Write **Law** for this category. Lawvere [4] formulated algebraic semantics in terms of these finite-product categories.

Given algebraic generators and equations, take the arrows $n \to 1$ to be terms in n variables modulo the equations and their consequences under substitution and formation of terms. The equivalence is the least one respected by these operations. For $n \to m$, take m-tuples of these classes. Substitution of well-defined classes is well-defined because of the closure conditions on the equivalence. Substitution is associative and substituting variables is an identity. The evident operation of taking the i-th output has the product property because the whole tuple is determined by those outputs. This builds the theory from the presentation. Given an object and chosen powers in a target category, a functor taking the projections to those powers is determined by its interpretation of the generators. Every interpretation satisfying the equations defines such a functor by interpreting each term recursively.

For example, the monoid theory has generators $\mu \colon 2 \to 1$ and $e \colon 0 \to 1$, for multiplication and the empty word. For associativity the following two maps from 3 to 1 are made equal.



Here juxtaposition with a tuple denotes composition, and the projections shown all start from 3. The two unit equations send $\mu\langle e!, \mathrm{id}_1\rangle$ and $\mu\langle \mathrm{id}_1, e!\rangle$ to id_1, where $! \colon 1 \to 0$ is the unique map. Each term is equivalent to the word obtained by flattening its products and removing occurrences of e, by associativity and the unit laws. Distinct words give distinct term classes, since evaluating at the generators in the free monoid of Chapter 5 distinguishes them. As a result $n \to m$ is an ordered m-tuple of words on n letters.

22.3 Models and natural operations

We fix product powers of sets, with X^0 a chosen singleton, $X^1 = X$, and the other X^n sets of tuples. A *model* of L is a functor $A\colon L \to \mathbf{Set}$ with $A(n) = A(1)^n$ taking the chosen projections of L to these tuple projections. Any product-preserving functor has an isomorphic model in this form, by transporting its maps along the unique product comparisons. A morphism of models is a natural transformation $\alpha\colon A \to B$. Naturality at projections forces $\alpha_n = (\alpha_1)^n$, so α is determined by its underlying function α_1. Naturality at the generators suffices for all the terms they generate, because it is preserved by tupling and composition. For the monoid theory the models and their morphisms are precisely monoids and monoid homomorphisms under this description.

Write $U_L\colon \mathbf{Mod}(L) \to \mathbf{Set}$ for evaluation at 1. A map $j\colon L \to L'$ of theories gives a functor $j^*\colon \mathbf{Mod}(L') \to \mathbf{Mod}(L)$ by precomposition. It commutes strictly with the evaluation functors, since $j(1) = 1$. This defines a functor Sem from $\mathbf{Law}^{\mathrm{op}}$ to the category $\mathbf{Cat}/\mathbf{Set}$. In the latter, an object is a functor $U\colon \mathcal{C} \to \mathbf{Set}$. An arrow to $V\colon \mathcal{E} \to \mathbf{Set}$ is a functor $H\colon \mathcal{C} \to \mathcal{E}$ with $VH = U$. The superscript $^{\mathrm{op}}$ is required because precomposition goes from models of the target theory to models of the source.

For each such U, let U^n be its pointwise power. Define $\mathrm{Str}(U)$ to have objects n and arrows

$$\mathrm{Str}(U)(n, m) = \mathrm{Nat}(U^n, U^m).$$

A map $o\colon U^n \Rightarrow U$ is an n-argument operation on all objects C of $\mathcal{C}$, satisfying

$$U(g)\, o_C(\mathbf{a}) = o_{C'}\big(U(g)^n(\mathbf{a})\big) \qquad (g\colon C \to C').$$

This equation specifies compatibility with every map in $\mathcal{C}$. Composition is composition of natural transformations. The pointwise tuple projections exhibit n as the required power of 1, since a tuple of natural operations is natural coordinate by coordinate. An arrow $H\colon U \to V$ over $\mathbf{Set}$ restricts a natural transformation $V^n \Rightarrow V^m$ along H to give one $U^n \Rightarrow U^m$. These restrictions preserve the projections and compose, defining Str from $\mathbf{Cat}/\mathbf{Set}$ to $\mathbf{Law}^{\mathrm{op}}$.

For size, one can take $\mathbf{Set}$ to consist of sets in a fixed universe and form both categories using categories and theories small in a larger universe. The underlying $\mathbf{Set}$ is small in the larger universe. Specifying a model, a natural transformation, or a natural operation takes a set of choices indexed by a set in that larger universe. Thus the two constructions stay in the indicated size range. In the recovery result below, the hom sets of the theory are assumed to belong to the smaller universe so that representable models take values in $\mathbf{Set}$.

22.4 The structure–semantics adjunction

Lawvere [4, p. 870] relates the natural operations of an underlying-set functor to interpretations of a theory.

Theorem 22.2. *For each $U\colon \mathcal{C} \to \mathbf{Set}$ and each theory L, there is a natural bijection*

$$\mathrm{Hom}_{\mathbf{Cat}/\mathbf{Set}}(U, U_L) \cong \mathrm{Hom}_{\mathbf{Law}}\big(L, \mathrm{Str}(U)\big).$$

Equivalently $\mathrm{Str} \dashv \mathrm{Sem}$ *with* Str *landing in* $\mathbf{Law}^{\mathrm{op}}$.

Proof. Given $H\colon \mathcal{C} \to \mathbf{Mod}(L)$ over $\mathbf{Set}$, interpret each $a\colon n \to m$ of L at every $H(C)$. The functions

$$o^a_C = H(C)(a)\colon U(C)^n \longrightarrow U(C)^m$$

are natural in C, since $H(g)$ is a morphism of models with underlying function $U(g)$. Interpretations preserve identities, composition, and the chosen projections, so $a \mapsto o^a$ defines a map of theories $\widehat{H}\colon L \to \mathrm{Str}(U)$.

Conversely let $k\colon L \to \mathrm{Str}(U)$ be a map of theories. On an object C take the model $A_C(n) = U(C)^n$, with arrow a interpreted as $k(a)_C$. On a morphism $g\colon C \to C'$, take the components $U(g)^n$. These give a model morphism by the naturality of each $k(a)$, as recorded by the square.

$$\begin{array}{ccc} U(C)^n & \xrightarrow{k(a)_C} & U(C)^m \\ \downarrow{\scriptstyle U(g)^n} & & \downarrow{\scriptstyle U(g)^m} \\ U(C')^n & \xrightarrow{k(a)_{C'}} & U(C')^m \end{array}$$

This defines $H_k\colon \mathcal{C} \to \mathbf{Mod}(L)$ by functoriality of the powers of $U(g)$. By construction its composite with U_L is exactly U.

$$\begin{array}{ccc} \mathcal{C} & \xrightarrow{H_k} & \mathbf{Mod}(L) \\ & \searrow^{U} & \downarrow{\scriptstyle U_L} \\ & & \mathbf{Set} \end{array}$$

The constructions recover each other by evaluation on C and a, and the morphism of models over $U(g)$ has prescribed components. Precomposition by a functor H_0 over $\mathbf{Set}$ restricts the family of operations along H_0. Precomposition by a map of theories $j\colon L' \to L$ amounts to replacing the interpretation of each arrow a of L' by that of $j(a)$. These equalities give naturality in both variables. □

A functor H_k over **Set** can also be viewed as a section of the left projection of a pullback of categories.

$$\begin{array}{ccc} \mathcal{C} \times_{\mathbf{Set}} \mathbf{Mod}(L) & \longrightarrow & \mathbf{Mod}(L) \\ \downarrow{\scriptstyle \mathrm{pr}_1} & \lrcorner & \downarrow{\scriptstyle U_L} \\ \mathcal{C} & \xrightarrow{\quad U \quad} & \mathbf{Set} \end{array}$$

Objects of this pullback are pairs (C, A) satisfying $U(C) = A(1)$, and arrows are pairs whose functions agree in **Set**. The section takes C to $(C, H_k(C))$ and g to $(g, H_k(g))$. Any section gives an interpretation by its second projection, so the adjunction describes all sections.

At U, the unit is the functor over **Set** into the models of $\mathrm{Str}(U)$ sending C to evaluation at C. The counit is an arrow from $\mathrm{Str}(U_L)$ to L in $\mathbf{Law}^{\mathrm{op}}$. In the category of theories it is the comparison

$$c_L \colon L \longrightarrow \mathrm{Str}(U_L), \qquad (c_L(a))_A = A(a).$$

For example, if a concrete category U interprets the theory via k, its functor of models factors as evaluation at objects of $\mathcal{C}$, followed by restriction of the resulting $\mathrm{Str}(U)$-models along k. Each arrow of L then acts by precisely the specified operation $k(a)_C$. This is the universal interpretation property of the unit.

22.5 Recovering the theory from its models

Suppose the hom sets of L belong to the smaller universe used for **Set**. Fix n, and take the representable functor $L(n, -)$. The projections identify $L(n, m)$ with $L(n, 1)^m$. Transport along these bijections to make a model F_n with $F_n(1) = L(n, 1)$. The tuple $\mathbf{x} = (\pi_1^n, \ldots, \pi_n^n)$ in $F_n(1)^n$ corresponds to id_n. The covariant form of Yoneda (Chapter 4) identifies

$$\mathrm{Hom}_{\mathbf{Mod}(L)}(F_n, A) \;\cong\; U_L(A)^n$$

naturally in A. Explicitly, a tuple $\mathbf{a}$ in $A(1)^n$ specifies the unique model map $h_{\mathbf{a}}$ carrying the generic tuple $\mathbf{x}$ to $\mathbf{a}$. For $t \colon n \to 1$ the underlying map sends t to $A(t)(\mathbf{a})$. At any other object the same rule uses all the output coordinates. Its naturality follows from the functor law for A.

Now a natural operation $o \colon U_L^n \Rightarrow U_L^m$ has a value $o_{F_n}(\mathbf{x})$ in $L(n, 1)^m$, so it specifies a single arrow $a \colon n \to m$. Naturality gives the following square for every

choice of model A and tuple $\mathbf{a}$.

$$\begin{array}{ccc} F_n(1)^n & \xrightarrow{o_{F_n}} & F_n(1)^m \\ \downarrow{\scriptstyle U_L(h_{\mathbf{a}})^n} & & \downarrow{\scriptstyle U_L(h_{\mathbf{a}})^m} \\ A(1)^n & \xrightarrow{o_A} & A(1)^m \end{array}$$

Following $\mathbf{x}$ around the square gives $o_A(\mathbf{a}) = A(a)(\mathbf{a})$. Hence $o = c_L(a)$. Conversely, for any a the value of $c_L(a)$ at the generic tuple is the tuple of its projections, so it gives back a. This proves that c_L is bijective on each hom set. It fixes all objects and preserves composition and projections, and is therefore an isomorphism of theories.

The adjunction now computes maps of model categories over **Set**. For two theories with hom sets in the smaller universe, a functor $\mathbf{Mod}(L') \to \mathbf{Mod}(L)$ over **Set** comes from restriction along a unique theory map $L \to L'$. Indeed, its transpose has target $\mathrm{Str}(U_{L'})$, and the comparison above identifies that target with L'.

For monoids F_n is the monoid of words on the n generic generators. Every natural function $B^n \to B$ of a monoid B is evaluation of a single word, with the empty word giving the unit. For $n = 1$ this says precisely that the natural unary operations are $b \mapsto b^k$ for integers $k \geq 0$. The word is found by applying the proposed operation to the generic generator. It can then be evaluated in any monoid by the unique homomorphism sending that generator to the chosen element. The underlying-set functor supplies the input positions for this test, and model homomorphisms supply the substitutions.

Example 22.3 (What the unit can add)**.** Let $\mathcal{C}$ be the category of nonempty sets and all functions between them, and U its inclusion in **Set**. For $n \geq 1$, a natural operation $o\colon U^n \Rightarrow U$ takes some value i on the tuple $(1, \ldots, n)$ in the set $\{1, \ldots, n\}$. Every tuple of n elements in a nonempty set is the image of this one under a function. Naturality forces o to select its i-th coordinate. A nullary operation would have to choose a point of every nonempty set compatibly with all functions. In a two-element set the swap fixes neither point, so no such operation exists.

The resulting theory has only projections and tuples of projections. A model is specified by any set X, including the empty set, with the tuple projections interpreted as usual on its powers. Its morphisms are all functions. Under this identification the unit of the structure–semantics adjunction is the inclusion of nonempty sets into all sets. The property of having an element imposes no natural choice of element and no additional natural operations. In the satisfaction adjunction the sentence $\exists x\,(x = x)$ tests this property directly in equality structures when empty domains are admitted.

22.6 Pontryagin duality and linear satisfaction

Fix a finite abelian group A as the space of assignments. An equation is a character $\chi \in \widehat{A}$ with satisfaction defined by $\chi(a) = 0$ in $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. For subgroups $S \leq \widehat{A}$ of equations and $B \leq A$ of assignments, put

$$S^{\perp} = \{a \in A \mid \chi(a) = 0 \text{ for all } \chi \in S\},$$
$$B^{\perp} = \{\chi \in \widehat{A} \mid \chi(b) = 0 \text{ for all } b \in B\}.$$

The adjunction is the equivalence $B \leq S^{\perp}$ if and only if $S \leq B^{\perp}$, with the order on assignment subgroups reversed as in Section 22.1. Taking the subgroup generated by a list of equations preserves all its solutions, since evaluation adds equations pointwise.

Here the two annihilator operations are inverse. If $a \notin B$, the image of a in A/B is nonzero. Finite character separation supplies a character of A/B nonzero on that image. Pulling it back gives $\chi \in B^{\perp}$ with $\chi(a) \neq 0$, so $B^{\perp\perp} = B$. Applying the same argument to $\widehat{A}/S$ and using the double-dual evaluation isomorphism gives $S^{\perp\perp} = S$. Thus every subgroup of assignments is the solution set of some linear equations, and the subgroup generated by a list consists of exactly the character equations entailed by that list.

For $A = (\mathbb{Z}/4\mathbb{Z})^2$, the character with coefficients (u, v) evaluates at (x, y) as $(ux + vy)/4 \bmod \mathbb{Z}$. Imposing $x + 2y = 0$ gives the four assignments $(-2y, y)$. A character (u, v) vanishes on all four exactly when $v = 2u \bmod 4$, as evaluation at $y = 1$ tests this congruence. These are precisely the multiples of $(1, 2)$, including $(2, 0)$, the consequence $2x = 0$. The quotient argument extends to closed subgroups in the locally compact setting of Section 16.6.

CHAPTER 23

Internal logic from elementary toposes

A subobject of X is a predicate in context X. To calculate with predicates, it helps to represent them by maps $X \to \Omega$ into a single object of truth values. In sets, Ω is the two-element set, and the representing maps are characteristic functions. In other categories the same universal property leads to different truth values. The logic comes from the operations imposed on these values by subobject inclusion and substitution.

23.1 A universal predicate

Definition 23.1. A *subobject classifier* in a category with finite limits consists of an object Ω and an arrow $\top\colon 1 \to \Omega$. For every monomorphism $m\colon A \hookrightarrow X$ there is a unique arrow $\chi_m\colon X \to \Omega$ making the square below a pullback.

$$\begin{array}{ccc} A & \longrightarrow & 1 \\ {\scriptstyle m}\downarrow & \lrcorner & \downarrow{\scriptstyle \top} \\ X & \xrightarrow{\chi_m} & \Omega \end{array}$$

An *elementary topos* has finite limits, a subobject classifier, and exponentials. An exponential B^A has an evaluation $\mathrm{ev}\colon B^A \times A \to B$ inducing natural bijections $\mathrm{Hom}(X, B^A) \cong \mathrm{Hom}(X \times A, B)$.

The arrow $\top$ is monic since its domain is terminal. Pulling it back along a predicate $p\colon X \to \Omega$ therefore gives a subobject, denoted $[p]$. The classifier property identifies $\mathrm{Hom}(X, \Omega)$ with $\mathrm{Sub}(X)$. If $f\colon Y \to X$, then $[p \circ f] = f^*[p]$, by pasting pullbacks. This is the substitution of Chapter 21 expressed as composition. In particular, the formula represented by p is true throughout X if $p = \top \circ !_X$, with $!_X\colon X \to 1$ the terminal arrow. A *global truth value* is an arrow $1 \to \Omega$, equivalent to a subobject of the terminal object.

The Sierpiński space of Section 16.5 classifies open inclusions in **Top**. Here the pullback test applies to every monomorphism. Even in a topos, the global truth values alone need not determine the predicates. The representing property uses maps from every context X.

For an object A, the classifier χ_{Δ_A} of the diagonal $\Delta_A \colon A \hookrightarrow A \times A$ represents equality in type A. For $s, t \colon X \to A$, the classifier $\chi_{\Delta_A} \circ \langle s, t \rangle$ represents their equalizer subobject of X. Indeed, a map into X factors through the pullback of the diagonal precisely when it makes s and t agree. This explains both equality and its substitution law using the same classifier.

23.2 Conjunction and implication

Work in an elementary topos $\mathcal{E}$. Write $\wedge_\Omega \colon \Omega \times \Omega \to \Omega$ for the classifier of the monomorphism $\langle \top, \top \rangle \colon 1 \hookrightarrow \Omega \times \Omega$. Given $p, q \colon X \to \Omega$, a map $u \colon U \to X$ makes $\wedge_\Omega \circ \langle p, q \rangle \circ u$ constantly true exactly when it makes both $p \circ u$ and $q \circ u$ constantly true. The subobject it classifies is $[p] \wedge [q]$, their meet.

Let $E \hookrightarrow \Omega \times \Omega$ be the equalizer of $\wedge_\Omega$ and the first projection π_1, and let $\Rightarrow_\Omega \colon \Omega \times \Omega \to \Omega$ be the classifier of E.

$$\begin{array}{ccc} E & \longrightarrow & 1 \\ \downarrow & \lrcorner & \downarrow{\scriptstyle \top} \\ \Omega \times \Omega & \xrightarrow{\Rightarrow_\Omega} & \Omega \end{array} \qquad\qquad E \hookrightarrow \Omega \times \Omega \underset{\pi_1}{\overset{\wedge_\Omega}{\rightrightarrows}} \Omega$$

The two diagrams describe the same subobject in different ways. For subobjects $P = [p]$ and $Q = [q]$ of X, define $P \Rightarrow Q$ as the subobject classified by $\Rightarrow_\Omega \circ \langle p, q \rangle$. A subobject $U \hookrightarrow X$ lies below $P \Rightarrow Q$ if and only if its map to $\Omega \times \Omega$ factors through the equalizer. This happens exactly when

$$(P|_U) \wedge (Q|_U) = P|_U,$$

where $P|_U$ and $Q|_U$ are their restrictions to U. That condition means $P|_U \leq Q|_U$, or equivalently $U \wedge P \leq Q$ in $\mathrm{Sub}(X)$. Consequently,

$$U \wedge P \leq Q \quad \Longleftrightarrow \quad U \leq (P \Rightarrow Q).$$

Implication is thus the right adjoint to meeting with P in the subobject poset. As in the case of opens in Chapter 17, it is determined by the order and conjunction. Moreover, because the same arrow $\Rightarrow_\Omega$ computes it in each context,

$$f^*(P \Rightarrow Q) = f^*P \Rightarrow f^*Q$$

for every $f \colon Y \to X$. The equality follows by composing the predicate with f. The same reasoning applies to conjunction.

23.3 Testing a family of truth values

Exponentials give a predicate for the parameter values at which a whole family of predicates is true. Let $R \hookrightarrow X \times Y$ be classified by $r \colon X \times Y \to \Omega$. Its

exponential transpose is $\widehat{r}\colon X \to \Omega^Y$. Let $t_Y : 1 \to \Omega^Y$ transpose $\top \circ !_Y$, using $1 \times Y \cong Y$. Define $A_Y(R) \hookrightarrow X$ by the pullback

$$\begin{array}{ccc} A_Y(R) & \longrightarrow & 1 \\ \downarrow & \lrcorner & \downarrow{\scriptstyle t_Y} \\ X & \xrightarrow{\widehat{r}} & \Omega^Y. \end{array}$$

For $U \hookrightarrow X$, factoring through this pullback means that $\widehat{r}|_U = t_Y \circ !_U$. Transposing this equality back says that r is constantly true on $U \times Y$, so

$$U \leq A_Y(R) \quad \Longleftrightarrow \quad U \times Y \leq R \text{ in } \mathrm{Sub}(X \times Y).$$

The transpose of $r \circ (f \times \mathrm{id}_Y)$ is $\widehat{r} \circ f$. Hence pulling back the displayed square along $f\colon X' \to X$ shows that

$$f^* A_Y(R) = A_Y\big((f \times \mathrm{id}_Y)^* R\big).$$

This is universal quantification along a projection. We use it here to derive falsity and disjunction. Chapter 24 will treat quantification along arbitrary arrows.

23.4 Falsity and disjunction from the universal test

Write $\pi\colon X \times \Omega \to X$ for the projection, and let $R \hookrightarrow X \times \Omega$ be the predicate classified by the other projection, $r(x, \omega) = \omega$. Define $\bot_X = A_\Omega(R)$. For any predicate $w\colon X \to \Omega$, use the section $s_w = \langle \mathrm{id}_X, w \rangle$ of π. By definition $\pi^* \bot_X \leq R$. Pulling this inequality back along s_w gives $\bot_X \leq [w]$, since $r \circ s_w = w$. Thus $\bot_X$ is the least subobject of X.

For $P, Q \in \mathrm{Sub}(X)$, form the subobject of $X \times \Omega$ given by

$$C = \big((\pi^* P \Rightarrow R) \wedge (\pi^* Q \Rightarrow R)\big) \Rightarrow R.$$

We claim that $J = A_\Omega(C)$ is the least upper bound of P, Q. To see that $P \leq J$, it suffices to show $\pi^* P \leq C$. By the implication adjunction, this amounts to

$$\pi^* P \wedge (\pi^* P \Rightarrow R) \wedge (\pi^* Q \Rightarrow R) \leq R.$$

The inequality holds already upon meeting the first two terms. Exchanging P, Q gives $Q \leq J$.

Conversely, let $W = [w]$ contain both P and Q. Pull back $\pi^* J \leq C$ along s_w. The rules for restricting conjunction and implication give

$$J \leq \big((P \Rightarrow W) \wedge (Q \Rightarrow W)\big) \Rightarrow W = W.$$

The equality uses that each of $P \Rightarrow W$ and $Q \Rightarrow W$ is the greatest subobject. This proves $J = P \vee Q$.

All the constructions commute with substitution by the stability of conjunction, implication, and the universal test. The rule for $\bot_X$ is therefore induced by the classifier of $\bot_1$. For disjunction, take the join of the two projection predicates on $\Omega \times \Omega$ and classify it. Composition with the resulting arrow $\vee_\Omega \colon \Omega \times \Omega \to \Omega$ then computes the join of any pair of predicates.

A bounded lattice in which meeting with each element has a right adjoint is a *Heyting algebra*. It is distributive because each such meet map, being a left adjoint, preserves finite joins. We have proved that every $\mathrm{Sub}(X)$ in a topos is a Heyting algebra and every reindexing map preserves its operations. In particular $\neg P = P \Rightarrow \bot_X$ is the largest subobject disjoint from P. The equality $P \wedge \neg P = \bot_X$ follows from the adjunction. Whether $P \vee \neg P$ is the greatest subobject depends on the topos. In **Set**, the largest subset disjoint from P is its complement, so excluded middle holds in every context. Mac Lane and Moerdijk [16] give a systematic treatment of the internal language of toposes.

23.5 Truth at two stages

Let D be the category with two objects $0, 1$ and one nonidentity arrow $0 \to 1$. A presheaf of sets on D is a contravariant functor into **Set**. It consists of a restriction $\rho_X \colon X_1 \to X_0$, and a map of presheaves is a commuting square of functions. Limits are formed at the two stages, since the induced restriction is determined by the two cones of functions. A map is monic precisely when both component functions are injective. Injectivity implies left cancellation. For the converse, maps from the restrictions $\varnothing \to \{*\}$ and $\{*\} \to \{*\}$ select elements of X_0 and X_1, respectively, so a failure of injectivity would violate left cancellation for one of these test objects. Thus a subobject consists of subsets $P_i \subseteq X_i$ with $\rho_X(P_1) \subseteq P_0$.

For clarity, the other topos axioms can be checked for this category directly. For X, Y, the exponential has

$$(Y^X)_0 = \mathbf{Set}(X_0, Y_0),$$
$$(Y^X)_1 = \{(a_0, a_1) \mid a_i \colon X_i \to Y_i, \rho_Y \circ a_1 = a_0 \circ \rho_X\}.$$

Restriction forgets a_1, and evaluation is evaluation of the specified function at each stage. A map $h \colon Z \times X \to Y$ transposes at stage 0 to $z_0 \mapsto h_0(z_0, -)$. At stage 1 its transpose sends z_1 to $\big(h_0(\rho_Z(z_1), -), h_1(z_1, -)\big)$. The condition on (a_0, a_1) follows from commutation for h. Each component of a transpose is forced by evaluation and restriction, which proves the exponential property.

Define the candidate classifier by the restriction

$$\Omega_1 = \{0, a, 1\} \xrightarrow{\rho_\Omega} \Omega_0 = \{0, 1\} \qquad \rho_\Omega(0) = 0, \quad \rho_\Omega(a) = \rho_\Omega(1) = 1.$$

The terminal object $\mathbf{1}$ is $\{*\} \to \{*\}$. Its map $\top$ selects 1 at both stages. For a subobject $P \hookrightarrow X$, use its ordinary characteristic function at stage 0. At stage 1 send x to

$$\begin{cases} 1 & x \in P_1, \\ a & x \notin P_1 \text{ and } \rho_X(x) \in P_0, \\ 0 & \rho_X(x) \notin P_0. \end{cases}$$

This is a map of presheaves because the image at stage 0 of 1 or a is 1, as required for the restricted elements lying in P_0. The inverse image of truth is P at each stage. In any classifying map, the values at stage 0 are forced by that pullback. So is the value 1 on P_1. Outside P_1, the value is forced by compatibility with the classifier at stage 0. This proves the required uniqueness.

The global truth values are the three subobjects of $\mathbf{1}$, in a chain. From least to greatest their restrictions are $\varnothing \to \varnothing$, $\varnothing \to \{*\}$, and $\{*\} \to \{*\}$. Write M for the middle element. It is represented by a at stage 1, and its restriction is true at stage 0. The largest subobject disjoint from M is the least subobject, since every other subobject meets M in M. Therefore $\neg M = \bot_{\mathbf{1}}$ and $M \vee \neg M = M \neq \mathbf{1}$. This identifies a proposition that becomes true at the earlier stage, while its excluded middle fails in the whole terminal context.

The middle subobject also illustrates why global elements cannot replace maps from arbitrary contexts. As a presheaf M has no global element, since a map $\mathbf{1} \to M$ would need a function from $\{*\}$ to $\varnothing$ at stage 1. Yet the two predicates given by the empty and the full subobject of M are distinct. They fail to agree in context M even though testing them on all global elements of M gives no distinction.

23.6 Implication and negation in context

The two-stage calculation extends to arbitrary predicates $P, Q \hookrightarrow X$ in the same presheaf topos. Let $T_i = \{x_i \in X_i \mid x_i \in P_i \text{ implies } x_i \in Q_i\}$ be the ordinary implication set in stage i. The subpresheaf $H = P \Rightarrow Q$ has components

$$H_0 = T_0, \qquad H_1 = T_1 \cap \rho_X^{-1}(T_0).$$

The formula makes $\rho_X(H_1) \subseteq H_0$, so these are the components of a subpresheaf. Since $H_i \subseteq T_i$ at both stages, $H \wedge P \leq Q$. For any subpresheaf U with $U \wedge P \leq Q$, we have $U_i \subseteq T_i$. Closure under restriction then gives $U_1 \subseteq T_1 \cap \rho_X^{-1}(T_0)$. Thus $U \leq H$, proving the formula by the implication adjunction.

Example 23.2 (Double negation and restriction)**.** Take $X_0 = \{a, b\}$ and $X_1 = \{u, v, w\}$, with $\rho_X(u) = \rho_X(v) = a$ and $\rho_X(w) = b$. Let $P_0 = \{a\}$ and $P_1 = \{u\}$. The empty subsets at both stages give falsity. Using them for Q in the implication formula computes $\neg P$. Applying the same computation again gives $\neg\neg P$.

	P	$\neg P$	$\neg\neg P$	$P \vee \neg P$
0	$\{a\}$	$\{b\}$	$\{a\}$	$\{a, b\}$
1	$\{u\}$	$\{w\}$	$\{u, v\}$	$\{u, w\}$

For the last column, the two stagewise unions form a subpresheaf, since the restriction of any element in either component lies in the corresponding union. Any upper bound must contain both unions, so this subpresheaf is the join.

The element v is outside P_1 and its restriction a belongs to P_0. For $\neg P$ to hold at v, P would have to omit both v and a at their respective stages. The last column therefore omits v. The table also records $v \in (\neg\neg P)_1$, so double negation can enlarge a predicate.

CHAPTER 24

Quantifiers from pullback

A predicate on pairs (x, y) can be turned into a predicate on y by asking for some admissible x or by testing every admissible x. The map that drops x also acts in the other direction. A predicate on y can be pulled back to a predicate on pairs that ignores x. The two ways to remove a variable are the two adjoints of this pullback. Lawvere [17] formulates quantification in these terms. The same order tests apply in an elementary topos.

24.1 The two adjoints of inverse image

For a function $f\colon X \to Y$, write $X_y = f^{-1}(\{y\})$ for its fiber over y. Reindexing a subset $Q \subseteq Y$ along f gives $f^*Q = f^{-1}(Q)$. For a subset $P \subseteq X$, set

$$\exists_f P = \{y \in Y \mid X_y \cap P \neq \varnothing\},$$
$$\forall_f P = \{y \in Y \mid X_y \subseteq P\}.$$

Both operations preserve inclusion. The existential quantifier gives the usual direct image $f(P)$. The universal quantifier collects those fibers every member of which belongs to P.

Proposition 24.1 (Quantification adjunctions in sets)**.** *For subsets ordered by inclusion,* $\exists_f \dashv f^* \dashv \forall_f$.

Proof. The inclusion $\exists_f P \subseteq Q$ says that $f(x)$ belongs to Q for each $x \in P$. This is the same as $P \subseteq f^*Q$. The inclusion $f^*Q \subseteq P$ says that for each $y \in Q$, every x in the fiber X_y belongs to P. This is the same as $Q \subseteq \forall_f P$. These equivalences give the hom-set bijections because an order has one arrow for a true inequality and none otherwise. □

The inequalities $P \subseteq f^*\exists_f P$ and $f^*\forall_f P \subseteq P$ follow by testing identity arrows in the two adjunctions. Thus $\exists_f P$ is the least predicate on Y whose pullback contains P, and $\forall_f P$ is the greatest whose pullback lies inside P. For a projection $Y \times A \to Y$ these are the usual rules for binding a variable of type A.

Example 24.2 (Testing the available runs)**.** Let f assign each program run its input in $Y = \{0, 1, 2, 3\}$. The predicate $P = \{a, b\}$ selects the successful runs.

y	X_y	$y \in \exists_f P$	$y \in \forall_f P$
0	$\varnothing$	no	yes
1	$\{a\}$	yes	yes
2	$\{b, c\}$	yes	no
3	$\{d\}$	no	no

Input 0 has no witness for success and no witness for failure. Its fiber is contained in every subset, so the universal condition holds vacuously. The adjunctions account for the two answers as well. Whatever happens over the other inputs, changing Q by adding or removing 0 makes no change to f^*Q. To satisfy the least and greatest tests, the least choice omits 0 and the greatest includes it.

Composite context changes obey the corresponding laws. For $f\colon X \to Y$ and $g\colon Y \to Z$, the identity $(g \circ f)^* = f^*g^*$ yields

$$\exists_{g\circ f} = \exists_g \exists_f, \qquad \forall_{g\circ f} = \forall_g \forall_f.$$

Indeed a composite of two left adjoints is left adjoint to the composite of their right adjoints in reverse order. The same rule applies with the roles reversed. Adjoints between posets are unique as functions, since two candidates having the same order tests are below each other.

24.2 Substituting parameters under a quantifier

If a context map $g\colon Y' \to Y$ replaces the parameter y, the admissible inputs also change. Their new space is the pullback, with $X' = Y' \times_Y X$ and projections as shown.

$$\begin{array}{ccc} X' & \xrightarrow{\;h\;} & X \\ {\scriptstyle f'}\downarrow & \lrcorner & \downarrow{\scriptstyle f} \\ Y' & \xrightarrow{\;g\;} & Y \end{array}$$

In **Set**, the map h sends (y', x) to x and restricts to a bijection from $X'_{y'}$ to $X_{g(y')}$. A point of the first fiber belongs to h^*P exactly when its corresponding point belongs to P. Thus the following identities hold as predicates on Y'.

$$g^*\exists_f P = \exists_{f'} h^* P, \qquad g^*\forall_f P = \forall_{f'} h^* P.$$

These are the *Beck–Chevalley identities*. They say that substituting a new parameter before quantifying gives the same predicate as substituting it afterward. A merely commutative square may fail these identities. For example, take

$X = Y = Y' = 1$, $f = g = \mathrm{id}_1$, and $X' = \varnothing$. In this commutative square, $P = X$ yields $g^*\exists_f P = 1$ and $\exists_{f'} h^* P = \varnothing$.

A condition $Q \subseteq Y$ involving just the outside parameter can also be moved across an existential quantifier.

$$\exists_f (P \cap f^*Q) = (\exists_f P) \cap Q.$$

This is the *Frobenius identity*. For a point $y \in Y$, the condition $f(x) \in Q$ on a witness $x \in X_y$ is precisely $y \in Q$. Thus a witness in $P \cap f^*Q$ amounts to a witness in P and the one condition $y \in Q$.

The two formulas play different roles in a calculation. Suppose that $R \subseteq Y \times A$ is a relation and $\pi\colon Y \times A \to Y$ is the projection. The query returning parameters in Q with a matching row of R gives $Q \cap \exists_\pi R$. Frobenius allows the filter on Q to be applied to the rows before projecting. If parameters are instead indexed by $g\colon Y' \to Y$, Beck–Chevalley allows the lookup of rows to be reindexed before projecting.

24.3 Universal quantification in a topos

Work now in an elementary topos. The notation $\mathrm{Sub}(X)$ and f^* is as in Chapter 21. Write $\top_X$ for the largest subobject of X. For subobjects of a common context, $\wedge$ is their meet and $\Rightarrow$ satisfies

$$R \leq (P \Rightarrow Q) \quad \Longleftrightarrow \quad R \wedge P \leq Q.$$

Both operations are preserved by pullback, as shown for implication in Section 23.2.

Let $\pi\colon Y \times A \to Y$ be a projection and P a subobject of $Y \times A$. The universal test of Section 23.3 can be written as an equalizer. Its parallel maps are the transpose $\widehat{\chi_P}$ of the classifier and the map τ obtained by transposing the constantly true predicate on $Y \times A$.

$$\forall_\pi P \hookrightarrow Y \underset{\tau}{\overset{\widehat{\chi_P}}{\rightrightarrows}} \Omega^A$$

The order test proved there is $Q \leq \forall_\pi P$ if and only if $\pi^*Q \leq P$. Thus $\forall_\pi$ is right adjoint to π^*. In formulas with variables we also write the result as $\forall_{a:A} P(y, a)$, a subobject of Y. Pulling the equalizer back gives the equalizer of the parallel maps precomposed with the base-change map. This yields Beck–Chevalley for projections, the case used in the next construction.

24.4 Existential quantification in a topos

To obtain a left adjoint along π, consider an arbitrary truth value t as a variable of type Ω. The formula

$$E(y) = \forall_{t:\Omega}((\forall_{a:A}\,(P(y,a) \Rightarrow t)) \Rightarrow t)$$

gives a predicate on Y. It says that every truth value implied by all the instances of P holds at y. It uses quantification over truth values and is called an *impredicative* definition. Its meaning and the adjunction can be checked using subobjects alone.

Put $D = Y \times \Omega$, write $p\colon D \to Y$ for the first projection, and let $T \in \mathrm{Sub}(D)$ be the subobject classified by the second projection. Thus T interprets the variable t. Set $C = D \times A$ and use the following pullback diagram, where α drops the A-coordinate and β drops the Ω-coordinate.

$$\begin{array}{ccc} C & \xrightarrow{\ \alpha\ } & D \\ {\scriptstyle\beta}\downarrow & \lrcorner & \downarrow{\scriptstyle p} \\ Y \times A & \xrightarrow{\ \pi\ } & Y \end{array}$$

Define $U = \forall_\alpha(\beta^* P \Rightarrow \alpha^* T)$ on D. The displayed formula for E then means $\forall_p(U \Rightarrow T)$.

Proposition 24.3 (Existential projection quantification)**.** *For E as defined above and $Q \in \mathrm{Sub}(Y)$,*

$$E \le Q \quad\Longleftrightarrow\quad P \le \pi^* Q.$$

In particular $E = \exists_\pi P$.

Proof. The counit for $\alpha^* \dashv \forall_\alpha$ gives $\alpha^* U \wedge \beta^* P \le \alpha^* T$. Using implication and pullback stability, this says $\beta^* P \le \alpha^*(U \Rightarrow T)$. By the adjunction for β, the latter holds precisely if $P \le \forall_\beta \alpha^*(U \Rightarrow T)$. The square is a base change of the projection p along π, with β a projection after permuting coordinates. Stability of universal projection quantification gives $\forall_\beta \alpha^*(U \Rightarrow T) = \pi^* E$. Thus $P \le \pi^* E$, which implies $P \le \pi^* Q$ whenever $E \le Q$.

Conversely suppose $P \le \pi^* Q$. The map $s = \langle \mathrm{id}_Y, \chi_Q \rangle\colon Y \to D$ substitutes the truth value of Q for t. It is a section of p, and $s^* T = Q$. By stability of $\forall_\alpha$ under pullback along s,

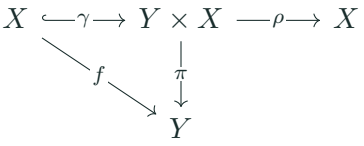
$$s^* U = \forall_\pi(P \Rightarrow \pi^* Q) = \top_Y.$$

Here $P \Rightarrow \pi^* Q = \top_{Y \times A}$ by the assumed inequality, and a right adjoint preserves the largest element. The counit for $p^* \dashv \forall_p$ gives $p^* E \wedge U \le T$. Pull back this inequality along s to obtain $E \wedge \top_Y \le Q$, so $E \le Q$. □

Every ingredient of the formula for E commutes with reindexing along $g\colon Y' \to Y$. In the inner universal quantifier the parameter change is $g \times \mathrm{id}_\Omega$, and in the outer one it is g. The truth-value predicate T also pulls back to the second projection in the new context. Thus the existential projection quantifier is stable under change of parameters.

24.5 Graphs, arbitrary maps, and images

For an arbitrary map $f\colon X \to Y$, let G be the subobject of $Y \times X$ given by the graph monomorphism $\gamma = \langle f, \mathrm{id}_X \rangle$. Write π, ρ for the two projections. The graph is monic because $\rho\gamma = \mathrm{id}_X$.



For $P \in \mathrm{Sub}(X)$, define

$$\exists_f P = \exists_\pi (G \wedge \rho^* P), \qquad \forall_f P = \forall_\pi (G \Rightarrow \rho^* P).$$

In variable notation these test, respectively, for a witness x with $f(x) = y$ and $P(x)$, and for $P(x)$ at every x with $f(x) = y$.

To check the adjunctions, a subobject of $Y \times X$ below G amounts to a subobject of X by factoring through γ. Intersection with G takes each subobject to its pullback along γ, viewed as a subobject below G. Consequently

$$\begin{aligned} \exists_f P \le Q &\iff G \wedge \rho^* P \le \pi^* Q \iff P \le f^* Q, \\ Q \le \forall_f P &\iff \pi^* Q \wedge G \le \rho^* P \iff f^* Q \le P. \end{aligned}$$

For a base change $g\colon Y' \to Y$ as in Section 24.2, pull back G along $g \times \mathrm{id}_X$. Its new domain is $X' = Y' \times_Y X$, and the monomorphism into $Y' \times X$ is $\langle f', h \rangle$. The pullback of $\rho^* P$ restricts along this monomorphism to $h^* P$. The same order tests with that monomorphism and projection to Y' therefore characterize $\exists_{f'} h^* P$ and $\forall_{f'} h^* P$. Stability for projections and preservation of $\wedge, \Rightarrow$ now give both Beck–Chevalley identities for arbitrary maps. Frobenius also follows without reference to points. For any target predicate R, the tests

$$\begin{aligned} \exists_f (P \wedge f^* Q) \le R &\iff P \wedge f^* Q \le f^* R \\ &\iff P \le f^* (Q \Rightarrow R) \\ &\iff \exists_f P \wedge Q \le R \end{aligned}$$

agree by the two adjunctions for $\exists_f$ and implication, and by pullback stability of implication. Hence the two predicates compared against R are equal.

Taking $P = \top_X$ shows that $I = \exists_f \top_X$ is the least subobject of Y through which f factors. This is the *image* of f. The unit gives $\top_X \leq f^* I$, hence the factorization $X \to I \hookrightarrow Y$. The first map is an epimorphism. If two maps on I agree after it, their equalizer in I gives a subobject of Y through which f factors. Minimality forces that equalizer to be all of I, so the maps agree. The existential Beck–Chevalley identity at $\top_X$ says that images pull back to images. Mac Lane and Moerdijk [16] develop the resulting first-order semantics in more detail.

Example 24.4 (Existence and global elements)**.** For the two-element group G, the category of G-sets from Chapter 9 is an elementary topos. Finite limits use underlying set limits with coordinatewise actions. For G-sets A, B, the set of all functions $A \to B$ with action $(g \cdot k)(a) = g \cdot k(g^{-1} \cdot a)$ is an exponential. A function $C \times A \to B$ is equivariant exactly when its curried function on C is equivariant for this action, by substituting $g^{-1} \cdot a$ in the equivariance equation. Subobjects are invariant subsets. In fact a noninjective equivariant map fails to be monic, since distinct elements in a common fiber give distinct equivariant maps from G with its left multiplication action. Characteristic functions of invariant subsets are exactly the equivariant maps to $\{0, 1\}$ with trivial action, giving a classifier.

Let $A = G$ with its left multiplication action. The map $A \to 1$ has full image since its only possible smaller image is empty. Hence $\exists_{A \to 1} \top_A = \top_1$. A global element of A would be a fixed point, and the nonidentity element interchanges the two points of A. This example exhibits a true existential predicate whose map of witnesses to the terminal object has no section.

The predicates of this chapter are subterminal objects in their slice categories. Pullback also reindexes general families, retaining all their elements in each fiber. Part VI uses such families and their dependent sums and products to interpret types. A sum family records its witnesses, whereas the existential subobject here is its image and records whether the witness condition holds.

CHAPTER 25

Extending diagrams by universality

A diagram of sets on a span specifies two maps with a common source. If we enlarge the indexing category by adding a target for the span, a pushout supplies a value at that target and compatible maps into it. A universal extension makes this construction systematic. Its inputs are a diagram and a map of indexing categories. The same method gives colimits, limits, and the quantifiers of Chapter 24.

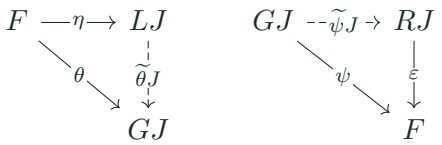
25.1 Restriction and its adjoints

Write $[\mathcal{A}, \mathcal{E}]$ for the category of functors from $\mathcal{A}$ to $\mathcal{E}$, with natural transformations as arrows. For this chapter take $\mathcal{A}, \mathcal{B}$ small and $\mathcal{E}$ locally small, so the natural transformations between two diagrams form a set. A functor $J\colon \mathcal{A} \to \mathcal{B}$ defines restriction by composition,

$$J^*\colon [\mathcal{B}, \mathcal{E}] \longrightarrow [\mathcal{A}, \mathcal{E}], \qquad J^*G = GJ, \qquad (J^*\gamma)_a = \gamma_{Ja}.$$

Restriction can select part of a diagram, repeat values, or replace an arrow by a composite according to the action of J.

A *left Kan extension* of $F\colon \mathcal{A} \to \mathcal{E}$ along J is a functor $L\colon \mathcal{B} \to \mathcal{E}$ equipped with a natural transformation $\eta\colon F \Rightarrow LJ$ such that every $\theta\colon F \Rightarrow GJ$ factors uniquely as $\widetilde{\theta} J \circ \eta$, where $\widetilde{\theta}\colon L \Rightarrow G$. A *right Kan extension* is a functor $R\colon \mathcal{B} \to \mathcal{E}$ equipped with $\varepsilon\colon RJ \Rightarrow F$ such that every $\psi\colon GJ \Rightarrow F$ factors uniquely as $\varepsilon \circ (\widetilde{\psi} J)$ for $\widetilde{\psi}\colon G \Rightarrow R$. The two factorization triangles lie in $[\mathcal{A}, \mathcal{E}]$.



The dashed arrows are restrictions of natural transformations on $\mathcal{B}$ determined by the required factorizations. The maps induced by η and ε give bijections

$$\mathrm{Nat}(L, G) \cong \mathrm{Nat}(F, J^*G), \qquad \mathrm{Nat}(J^*G, F) \cong \mathrm{Nat}(G, R).$$

These bijections are natural in G by associativity of composition. Each extension, with its specified transformation, is unique up to unique isomorphism by comparing two extensions in both directions.

When such extensions exist for all F, the universal properties supply the functors Lan_J and Ran_J. For $\alpha\colon F \Rightarrow F'$, factor $\eta_{F'} \circ \alpha$ through η_F to get $\mathrm{Lan}_J\,\alpha$. For the right extension factor $\alpha \circ \varepsilon_F$ through $\varepsilon_{F'}$ to get $\mathrm{Ran}_J\,\alpha$. The uniqueness of each factorization gives the identity and composition laws and naturality in F. Thus these functors give the adjoint triple

$$\mathrm{Lan}_J \dashv J^* \dashv \mathrm{Ran}_J\,.$$

The maps η_F are the unit for the left adjunction, and ε_F are the counit for the right adjunction. For $G\colon \mathcal{B} \to \mathcal{E}$, the remaining counit and unit are the extensions of the identity transformation of J^*G.

$$\mathrm{Lan}_J(J^*G) \Rightarrow G, \qquad G \Rightarrow \mathrm{Ran}_J(J^*G).$$

For instance, the composite of the restriction of the left counit with η_{J^*G} is the identity by its defining factorization. The other triangle follows by applying the same uniqueness to the identity of $\mathrm{Lan}_J\,F$. The two right triangles follow by the factorization through ε. The unit and counit thus have the roles derived for adjunctions in Chapter 5.

25.2 Which arrows contribute at an object?

Fix $b \in \mathcal{B}$. If $\theta\colon F \Rightarrow GJ$, then an arrow $u\colon Ja \to b$ gives a map $G(u) \circ \theta_a\colon F(a) \to G(b)$. If $h\colon a \to a'$ and $u'\colon Ja' \to b$ satisfy $u' \circ J(h) = u$, the map for u equals the composite of $F(h)$ and the map for u'. The category $(J \downarrow b)$ records these data. Its objects are $(a, u\colon Ja \to b)$, and its arrows $(a, u) \to (a', u')$ are the arrows h satisfying the triangle condition on the left below.



For a transformation $\psi\colon GJ \Rightarrow F$, an arrow $v\colon b \to Ja$ gives a map $\psi_a \circ G(v)\colon G(b) \to F(a)$. The compatible families of such maps are indexed by $(b \downarrow J)$. Its objects are $(a, v\colon b \to Ja)$, and an arrow $(a, v) \to (a', v')$ is an $h\colon a \to a'$ with $J(h) \circ v = v'$, as in the triangle on the right. Composition in both categories comes from $\mathcal{A}$, since substituting the triangle identities shows that composites satisfy the same conditions. The projections $p_b\colon (J \downarrow b) \to \mathcal{A}$ and $q_b\colon (b \downarrow J) \to \mathcal{A}$ forget the arrow involving b. These are instances of *comma categories*.

Theorem 25.1 (Pointwise Kan extensions)**.** *If all the indicated colimits exist in $\mathcal{E}$, the left extension of F along J can be computed by*

$$(\mathrm{Lan}_J\,F)(b) = \operatorname*{colim}_{(J \downarrow b)} F p_b.$$

If all the indicated limits exist, the right extension can be computed by

$$(\mathrm{Ran}_J F)(b) = \lim_{(b \downarrow J)} F q_b.$$

In particular a cocomplete $\mathcal{E}$ supplies the left adjoint to J^*, and a complete $\mathcal{E}$ supplies the right adjoint. Here cocompleteness and completeness refer to all small diagrams.

Proof. For the left extension choose the indicated colimit $L(b)$ with legs $\lambda_{a,u} \colon F(a) \to L(b)$. For $s \colon b \to b'$, postcomposition sends (a, u) to $(a, s \circ u)$ and preserves the conditions on arrows of the comma category. The legs into $L(b')$ therefore form a cocone on Fp_b. The colimit gives the unique $L(s)$ satisfying

$$L(s) \circ \lambda_{a,u} = \lambda_{a,s \circ u}.$$

The equations for identities and composites hold after precomposition with every leg and hence hold on the colimit. Define $\eta_a = \lambda_{a,\mathrm{id}_{Ja}}$. For $h \colon a \to a'$ the arrow h from $(a, J(h))$ to $(a', \mathrm{id}_{Ja'})$ gives

$$L(J(h)) \circ \eta_a = \lambda_{a,J(h)} = \eta_{a'} \circ F(h).$$

This is naturality of η.

For $\theta \colon F \Rightarrow GJ$, the maps $G(u) \circ \theta_a$ are a cocone on Fp_b by naturality of θ and the defining equation for arrows in $(J \downarrow b)$. The induced map $\widetilde{\theta}_b \colon L(b) \to G(b)$ is specified by the equation in the triangle.

$$\begin{array}{ccc} F(a) & \xrightarrow{\lambda_{a,u}} & L(b) \\ & \searrow^{G(u) \circ \theta_a} & \downarrow{\scriptstyle \widetilde{\theta}_b} \\ & & G(b) \end{array}$$

For $s \colon b \to b'$, the composites $G(s) \circ \widetilde{\theta}_b$ and $\widetilde{\theta}_{b'} \circ L(s)$ agree on every leg, since both give $G(s \circ u) \circ \theta_a$. This proves naturality. The leg indexed by id_{Ja} gives the required factorization. If a natural transformation $L \Rightarrow G$ has that factorization, naturality at u forces its value on $\lambda_{a,u} = L(u) \circ \eta_a$. Uniqueness of the map out of each colimit then proves uniqueness of the extension.

For the right extension choose $R(b)$ with legs $\rho_{a,v} \colon R(b) \to F(a)$. Precomposition with $s \colon b \to b'$ sends $(a, v' \colon b' \to Ja)$ to $(a, v' \circ s)$. Reindexing the cone in this way gives $R(s) \colon R(b) \to R(b')$, determined by

$$\rho_{a,v'} \circ R(s) = \rho_{a,v' \circ s}.$$

The identity and composition laws follow by comparing projections. Define $\varepsilon_a = \rho_{a,\mathrm{id}_{Ja}}$. The arrow $h \colon a \to a'$ from (a, id_{Ja}) to $(a', J(h))$ in $(Ja \downarrow J)$ gives

$$\varepsilon_{a'} \circ R(J(h)) = \rho_{a',J(h)} = F(h) \circ \varepsilon_a.$$

A transformation $\psi\colon GJ \Rightarrow F$ gives a cone at $G(b)$ whose legs are $\psi_a \circ G(v)$. The triangle determines $\widetilde{\psi}_b\colon G(b) \to R(b)$.

$$\begin{array}{ccc} G(b) & \dashrightarrow^{\widetilde{\psi}_b} & R(b) \\ & {\scriptstyle \psi_a \circ G(v)} \searrow & \downarrow {\scriptstyle \rho_{a,v}} \\ & & F(a) \end{array}$$

For $s\colon b \to b'$ both naturality composites have the same projection $\psi_a \circ G(v' \circ s)$ at every (a, v'). Factoring at id_{Ja} yields $\psi_a = \varepsilon_a \circ \widetilde{\psi}_{Ja}$. Since $\rho_{a,v} = \varepsilon_a \circ R(v)$, that factorization and naturality force every projection of $\widetilde{\psi}_b$. This proves uniqueness.

The objects in each indexing category form a set because the objects a form a set and each possible arrow involving Ja and b belongs to a hom-set. Its arrows are taken from pairs of objects and arrows of $\mathcal{A}$, so they also form a set. The two existence conclusions therefore follow from the stated size hypotheses. □

The formulas justify the term *pointwise*. They determine each value on $\mathcal{B}$ from a single limit or colimit and then compute each transition map from the corresponding universal legs. Riehl gives further examples and the general theory of pointwise Kan extensions in Sections 6.2–6.3 of *Category Theory in Context* [2].

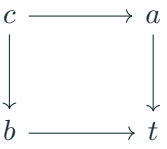
25.3 Collapsing a diagram and filling a missing corner

Let $\mathbf{1}$ be the one-object category with only its identity, and let $!\colon \mathcal{A} \to \mathbf{1}$ be the unique functor. Here restriction takes an object to its constant diagram on $\mathcal{A}$. Each of the two comma categories at the single object of $\mathbf{1}$ is isomorphic to $\mathcal{A}$. Thus, when they exist, the colimit and limit are the left and right Kan extensions along $!$. The same adjunctions say directly that maps out of the colimit correspond to cocones, and maps into the limit correspond to cones.

For a group G, form the one-object category BG whose arrows are the elements of G, with composition given by multiplication. A functor $BG \to \mathbf{Set}$ is a left action on its value X at the sole object. A cocone to Y is a function $h\colon X \to Y$ with $h(gx) = h(x)$ for all g, x. It factors uniquely through the orbit set X/G, which is therefore the colimit. A cone from Y consists of a function $k\colon Y \to X$ with $gk(y) = k(y)$ for all g, y, so it factors uniquely through the subset X^G of points fixed by every element. This gives the limit. Thus the orbit and fixed-point constructions from Chapter 9 are Kan extensions along $BG \to \mathbf{1}$.

Example 25.2 (Completing a span)**.** Let $\mathcal{B}$ be the ordered set with $c < a < t$ and $c < b < t$, with a and b incomparable, regarded as a category. The two paths $c \to t$ in the square are the same arrow. Let $J\colon \mathcal{A} \hookrightarrow \mathcal{B}$ include the full subcategory on

c, a, b, so a functor $F\colon \mathcal{A} \to \mathbf{Set}$ consists of the span $F(a) \leftarrow F(c) \to F(b)$.



At every old object $d \in \{a, b, c\}$, the category $(J \downarrow d)$ has a terminal object (d, id_d), and $(d \downarrow J)$ has the same pair as an initial object. Every cocone in the first category is determined by its leg at that terminal object, and every cone in the second by its leg at that initial object. Hence both extensions agree with F there up to the canonical isomorphisms. This argument works for any fully faithful J at objects of the form Ja, since every arrow between two such objects lifts uniquely along J.

At t, the category $(J \downarrow t)$ is the original span category. Every old object has exactly one arrow into t, and each old arrow satisfies the condition for a comma arrow. Consequently $(\mathrm{Lan}_J F)(t)$ is the pushout of the span. The two maps to the new value are its cocone legs. Meanwhile $(t \downarrow J)$ is empty because t has no arrows to old objects. Its limit in $\mathbf{Set}$ is a singleton, whose unique incoming functions give the arrows to $(\mathrm{Ran}_J F)(t)$.

For example choose $F(c) = \{0, 1\}$, $F(a) = \{x\}$, and $F(b) = \{y, z, w\}$. Send both elements to x on the first leg and send 0 to y and 1 to z on the second. The pushout has two classes in the disjoint union, $[x] = [y] = [z]$ and $[w]$. For any commutative square of sets on $\mathcal{B}$ with compatible maps from the original span, the two maps on $F(a)$ and $F(b)$ therefore induce a unique map from the pushout to its new target. In the reverse comparison the singleton at t receives the unique map from that target. These are the left and right extension comparisons at the new corner.

25.4 Quantifiers and families on discrete categories

For a set X, write $\mathrm{Disc}(X)$ for the category with objects $x \in X$ and only identity arrows. A function $f\colon X \to Y$ induces a functor $J = \mathrm{Disc}(f)$. For $y \in Y$ both comma categories in Theorem 25.1 are discrete on the fiber $f^{-1}(y)$. Indeed, an arrow in either direction between $f(x)$ and y exists exactly when the two objects are equal, and then the arrow is uniquely the identity.

Let $\mathbf{2}$ be the order $0 < 1$ as a category. A functor $\mathrm{Disc}(X) \to \mathbf{2}$ assigns a truth value to every object. It corresponds to the subset of objects assigned 1. A natural transformation exists precisely when the values are ordered pointwise, which means inclusion of the corresponding subsets. Thus the functor category is the poset $\mathcal{P}(X)$, and restriction is inverse image. Colimits in $\mathbf{2}$ are joins and

limits are meets. Hence for a predicate $P\colon X \to \{0,1\}$ the two formulas become

$$(\mathrm{Lan}_J P)(y) = \bigvee_{f(x)=y} P(x), \qquad (\mathrm{Ran}_J P)(y) = \bigwedge_{f(x)=y} P(x).$$

The join is true when some element in the fiber satisfies P. The meet is true when every element in the fiber satisfies P. The empty join is 0, and the empty meet is 1. These are exactly $\exists_f$ and $\forall_f$ from Chapter 24. The pointwise Kan extension proof yields both adjunctions to inverse image.

With target **Set**, a functor on $\mathrm{Disc}(X)$ is a family of sets (E_x). Represent it by the map $\coprod_x E_x \to X$ sending each summand to its index. Conversely any map to X gives the family of its fibers. Maps over X are exactly collections of functions on the individual fibers. This gives an equivalence with the slice category $\mathbf{Set}/X$. Restriction sends (D_y) to $(D_{f(x)})$, the fibers of the pullback along f. The discrete comma categories give

$$(\mathrm{Lan}_J E)_y = \coprod_{f(x)=y} E_x, \qquad (\mathrm{Ran}_J E)_y = \prod_{f(x)=y} E_x.$$

These are the dependent sum $\Sigma_f E$ and dependent product $\Pi_f E$. The sum records a witness x and an element of E_x. The product specifies an element of E_x for each x in the fiber. When each E_x is empty or a singleton, taking the image of the sum in Y retains the existential truth value. The product is itself empty or a singleton and gives the universal truth value.

The action of a pullback on these families can be checked directly. Take a pullback square of sets as below, with $X' = \{(x, y') \mid f(x) = v(y')\}$ and u, f' its projections.

$$\begin{array}{ccc} X' & \xrightarrow{u} & X \\ \downarrow{\scriptstyle f'} & & \downarrow{\scriptstyle f} \\ Y' & \xrightarrow{v} & Y \end{array}$$

Over each y', the points of $(f')^{-1}(y')$ are the pairs (x, y') with $f(x) = v(y')$. Forgetting y' indexes the same summands or factors as $f^{-1}(v(y'))$. Reindexing thus gives natural isomorphisms

$$\Sigma_{f'} u^* \cong v^* \Sigma_f, \qquad \Pi_{f'} u^* \cong v^* \Pi_f.$$

The bijections relabel witnesses in the sum and arguments in the product. Both commute with maps of families, which proves naturality and gives the Beck–Chevalley property for families of sets.

Part VI Type theory and proof

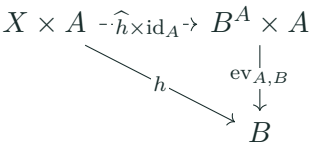
CHAPTER 26

Simply typed syntax and free cartesian closed categories

In the product presentation of Section 5.5.4, terms assemble values and take them apart. To make a function itself a value, we need a type whose generalized elements are families of functions. The exponential has this property. Its universal arrow interprets application, and its universal factorization interprets abstraction. We will use both operations to construct a category whose arrows are terms modulo the equations they satisfy.

26.1 Function types from exponentials

Choose a terminal object and binary products in a category $\mathcal{E}$. An exponential B^A comes with an evaluation map $\mathrm{ev}_{A,B}\colon B^A \times A \to B$. For a map $h\colon X \times A \to B$, its *transpose* $\widehat{h}\colon X \to B^A$ is the unique arrow making the triangle commute.



Equivalently, composing with evaluation gives the bijection

$$\mathrm{Hom}_{\mathcal{E}}(X, B^A) \cong \mathrm{Hom}_{\mathcal{E}}(X \times A, B).$$

For $u\colon Y \to X$, the defining evaluation equation gives $h \circ \widehat{(u \times \mathrm{id}_A)} = \widehat{h} \circ u$. Thus the correspondence is natural in the test object X. A map $v\colon B \to C$ induces $v^A\colon B^A \to C^A$ by transposing $v \circ \mathrm{ev}_{A,B}$. Composing with evaluation verifies the functor laws for $(-)^A$ and the naturality in B, giving $(- \times A) \dashv (-)^A$. A category with these products and an exponential for every pair A, B is *cartesian closed*. Exponentials are also part of the definition of a topos in Chapter 23.

In **Set**, B^A is the set of functions from A to B, and $\widehat{h}(x)$ is the function $a \mapsto h(x, a)$. The parameter x may carry other variables. The same construction in a cartesian closed category explains how a term with an extra variable of type A can become a term of function type with that variable bound.

26.2 Terms in context

Take a set of base type symbols $\mathcal{B}$. Form types from them using 1, $A \times B$, and $A \to B$, where the last denotes the type interpreted by B^A. Types are formal expressions with their parentheses retained. A context Γ is a finite list of declarations $x_i \colon A_i$ with distinct names. A judgment $\Gamma \vdash t \colon B$ gives a term of type B with free variables among those declared in Γ. We use only the following term operations and the variables of the context.

- The term $*$ has type 1 in any context.
- From $s \colon A$ and $t \colon B$ in a common context we form $\langle s, t \rangle \colon A \times B$. The two projections of a term of type $A \times B$ have types A and B.
- If $\Gamma, x \colon A \vdash t \colon B$, with x fresh for Γ, then $\Gamma \vdash \lambda x.t \colon A \to B$. The lambda binds the occurrences of x in t that are free there. We may rename bound variables with fresh names.
- In a common context, $f \colon A \to B$ and $a \colon A$ give an application $f\,a \colon B$.

These rules give the simply typed lambda calculus with products and a unit type.

For $\Gamma = (x_1 \colon A_1, \ldots, x_n \colon A_n)$, a *substitution* σ from Δ to Γ gives one term $\Delta \vdash s_i \colon A_i$ for each i. It acts on a term $\Gamma \vdash t \colon B$ by replacing x_i with s_i simultaneously, yielding $\Delta \vdash t[\sigma] \colon B$. Replacement passes through pairing, projection, and application component by component. Under a lambda, first choose the bound name fresh for the source and target contexts and the substituted terms. Then extend the substitution by the new variable itself and work in the body. Freshness ensures that the binder binds exactly the intended occurrences. For instance, substituting free y for x in $\lambda y.x$ gives $\lambda z.y$, with z fresh.

Substitution of the declared variables acts as the identity. Given ρ from Θ to Δ, define $\sigma\rho$ from Θ to Γ by substituting ρ into every component of σ. Then

$$t[\sigma][\rho] = t[\sigma\rho]$$

up to renaming bound variables. The variable case is the definition of $\sigma\rho$. For pairing, projection, and application the equality follows on the immediate subterms, and it is automatic for $*$. For a lambda, choose a name fresh for both substitutions. Extending each substitution by that variable commutes with their composition, so the same inductive equality holds for the body.

26.3 Equations and their interpretation

We impose the unit and product equations from the earlier presentation, and two equations for function types.

$$t = * \qquad (t\colon 1),$$
$$\pi_1\langle s,t\rangle = s, \qquad \pi_2\langle s,t\rangle = t, \quad \langle \pi_1 u, \pi_2 u\rangle = u,$$
$$(\lambda x.t)\, a = t[x := a], \qquad \lambda x.(f\, x) = f \qquad (x \text{ fresh for } f).$$

The function β-equation runs an abstraction on an argument. The function η-equation reconstructs a function by applying it to a fresh variable and abstracting that variable again. Equality of terms means the least equivalence on well-typed terms in common contexts containing these equations and bound renaming, and preserved by the term operations and substitution. Write $\equiv_{\beta\eta}$ for this equality, including the displayed unit and product equations.

In a cartesian closed category with chosen objects, projections, and evaluation maps, assign one object to each base type. Interpret 1, $\times$, and $\to$ by the chosen terminal object, product, and exponential. Write $\overline{A}$ for the interpretation of A. Interpret the empty context by the terminal object and an extended context $\Gamma, x\colon A$ by $\overline{\Gamma} \times \overline{A}$. This fixes a parenthesization even when Γ is empty. A variable is interpreted by its context projection. The interpretations of $*$, pairing, and projection use the corresponding product operations. The map for $\lambda x.t$ is the transpose of the map for t. The map for application is the composite through evaluation.

$$\overline{\Gamma} \xrightarrow{\langle \overline{f}, \overline{a}\rangle} \overline{B}^{\overline{A}} \times \overline{A} \xrightarrow{\mathrm{ev}_{\overline{A},\overline{B}}} \overline{B}$$

The interpretation $\overline{\sigma}\colon \overline{\Delta} \to \overline{\Gamma}$ of a substitution is the tuple of its component maps into the context product. The substitution equation is $\overline{t[\sigma]} = \overline{t} \circ \overline{\sigma}$. It holds for variables by the projection equations, and for $*$ by terminality. It passes through pairing and projection by the product property, and through application by composing with evaluation. Under a lambda, the extended substitution interprets as $\overline{\sigma} \times \mathrm{id}_{\overline{A}}$. Naturality of transposition gives the equation for the abstraction.

The β-equation for functions follows by precomposing $\mathrm{ev} \circ (\widehat{h} \times \mathrm{id}) = h$ with $\langle \mathrm{id}, \overline{a}\rangle$. This map interprets the substitution that inserts a for the last variable. For η, transposing $\mathrm{ev} \circ (f \times \mathrm{id})$ returns f by uniqueness. Product and unit equations hold by their universal properties. This proves soundness of the term equations for every cartesian closed interpretation.

26.4 The syntactic cartesian closed category

Let $\mathcal{L}_{\mathcal{B}}$ be the category whose objects are types and whose arrows $A \to B$ are terms $x\colon A \vdash t\colon B$ modulo $\equiv_{\beta\eta}$, with the input variable understood up to

renaming. The identity is $[x]$. If s represents $A \to B$ and t represents $B \to C$ in input variable y, their composite is $[t[y := s]]$. Congruence makes this well-defined, and the substitution equations give associativity and the identity laws.

The type 1 is terminal by its equation. The type $A \times B$ with the two projection terms is a product. Pairing gives the map of hom sets

$$\mathrm{Hom}_{\mathcal{L}_\mathcal{B}}(X, A) \times \mathrm{Hom}_{\mathcal{L}_\mathcal{B}}(X, B) \longrightarrow \mathrm{Hom}_{\mathcal{L}_\mathcal{B}}(X, A \times B),$$

and projecting gives its inverse by the three product equations.

The exponential for A, B is the type $A \to B$, with evaluation represented by $(\pi_1 w)(\pi_2 w)$ in input $w\colon (A \to B) \times A$. An arrow $X \times A \to B$ represented by $z\colon X \times A \vdash h\colon B$ has the transpose represented by $\lambda a.h[z := \langle x, a\rangle]$ in input $x\colon X$. In the other direction a term $x\colon X \vdash k\colon A \to B$ gives an arrow $X \times A \to B$ represented by $(k[x := \pi_1 z])(\pi_2 z)$. Going from h to its transpose and back uses function β, then product η to replace $\langle \pi_1 z, \pi_2 z\rangle$ by z. Going from k to evaluation and back uses product β, then function η. These inverse constructions give the exponential property.

Theorem 26.1 (Free cartesian closed category)**.** *For any cartesian closed category $\mathcal{E}$ with specified terminal object, products, and exponentials, each assignment of objects of $\mathcal{E}$ to symbols of $\mathcal{B}$ extends to a unique functor $\mathcal{L}_\mathcal{B} \to \mathcal{E}$ preserving these choices, including the structure maps, on the nose.*

Proof. Use the interpretation above. To interpret $[x\colon A \vdash t\colon B]$ as a map $\overline{A} \to \overline{B}$, precompose its context map with $\langle !, \mathrm{id}_{\overline{A}}\rangle\colon \overline{A} \to 1 \times \overline{A}$. Projection onto $\overline{A}$ is inverse to this canonical isomorphism. The substitution equation yields functoriality. The product and evaluation terms give exactly the chosen structure maps by their interpretation.

For uniqueness, let P_Γ be the type obtained from 1 by appending the types of Γ as binary products, in declaration order. Package a judgment $\Gamma \vdash t\colon B$ as an arrow $P_\Gamma \to B$ by replacing the variables by projections from one variable of type P_Γ. This arrow is built from projections, the terminal map, pairing, composition, evaluation, and transposition, by induction on the term. For abstraction, $P_{\Gamma, x\colon A} = P_\Gamma \times A$, so the induction uses exactly the transpose just constructed. A functor preserving the specified maps also preserves transposes, since these are the unique solutions of the evaluation equations. Its value on every packaged judgment is therefore forced. In a one-variable context the original arrow is obtained from its package by precomposition with $\langle !, \mathrm{id}_A\rangle$. This forces the functor on all arrows. □

The choice convention in the theorem keeps the uniqueness assertion literal. It includes specified projections and evaluation maps. For general theory and its connection with typed syntax, see Lambek and Scott [18].

There is also a completeness consequence for this specific equality of terms. If two terms in a common context have the same interpretation in all cartesian closed categories under all assignments, take the canonical model $\mathcal{L}_{\mathcal{B}}$ with each generator assigned to itself. The interpretations there are the packaged judgment arrows just constructed. Equality of these arrows is equality of their representative terms modulo $\equiv_{\beta\eta}$. Substitute the tuple of the original context variables, with leading component $*$, for the packaging variable. Each projection computes its original variable, so the original terms are $\equiv_{\beta\eta}$-equal.

26.5 A calculation with a function returning a pair

Example 26.2 (Two components of a function)**.** The same calculation can be made in syntax and in any cartesian closed category. The terms Φ and Ψ have types $(A \to (B \times C)) \to ((A \to B) \times (A \to C))$ and $((A \to B) \times (A \to C)) \to (A \to (B \times C))$, respectively.

$$\Phi = \lambda f.\langle \lambda a.\pi_1(f\,a), \lambda a.\pi_2(f\,a)\rangle,$$
$$\Psi = \lambda p.\lambda a.\langle (\pi_1 p)\,a, (\pi_2 p)\,a\rangle.$$

On the input f, applying Ψ after Φ gives $\lambda a.\langle \pi_1(f\,a), \pi_2(f\,a)\rangle$ by the β-equations. Product η gives $\lambda a.f\,a$, and function η gives f. For the other composite, its first component at p reduces to $\lambda a.(\pi_1 p)\,a$, and the second to $\lambda a.(\pi_2 p)\,a$. Function η gives the two projections of p, and product η gives p itself.

In any cartesian closed category, Φ is the pairing of the two maps induced by the projections of $B \times C$. For example, write π_B^A for the transpose of evaluation followed by π_B. Its effect on evaluation is the commuting square

$$\begin{array}{ccc} (B \times C)^A \times A & \xrightarrow{\pi_B^A \times \mathrm{id}_A} & B^A \times A \\ \downarrow{\scriptstyle \mathrm{ev}_{A,B\times C}} & & \downarrow{\scriptstyle \mathrm{ev}_{A,B}} \\ B \times C & \xrightarrow{\quad\pi_B\quad} & B. \end{array}$$

The analogous square for C supplies the other component. The syntax computation proves they give an isomorphism $(B \times C)^A \cong B^A \times C^A$. Equivalently, the exponential is right adjoint to $- \times A$, and this is its product comparison from the right-adjoint product property.

26.6 Pontryagin duality and a finite character type

Interpret types as finite sets and fix $n \geq 1$. Let C_n be the cyclic group $\mathbb{Z}/n\mathbb{Z}$, and let $\mathbb{T}[n]$ be the finite subgroup of $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ killed by n. The function type $C_n \to \mathbb{T}[n]$ contains the subtype Char_n of functions satisfying

$$h(0) = 0, \qquad h(x + y) = h(x) + h(y) \quad (x, y \in C_n).$$

These conditions can be interpreted by equalizers in finite sets. Any character $C_n \to \mathbb{T}$ automatically takes values in $\mathbb{T}[n]$, since $nh(x) = h(nx) = h(0)$. Thus the subtype models the full character group of C_n.

Two functions give an equivalence of types $\mathsf{Char}_n \cong \mathbb{T}[n]$:

$$\mathsf{eval}_1(h) = h(\overline{1}),$$
$$\mathsf{decode}(t)(\overline{k}) = kt.$$

Changing the representative k by a multiple of n preserves kt, and addition of representatives proves that $\mathsf{decode}(t)$ is a character. The first composite sends t to $1t = t$. For the other composite, $\mathsf{decode}(\mathsf{eval}_1(h))(\overline{k}) = kh(\overline{1}) = h(\overline{k})$, so the functions are equal. The subtype uses equality conditions in sets, which carry no additional choice of proof. Pointwise addition of functions makes both maps homomorphisms.

The type calculation is the universal property of the presentation of C_n with one generator g and the relation $ng = 0$. Finite Pontryagin duality in Section 8.6 recovers C_n intrinsically by evaluating on its characters. The explicit equivalence here uses the generator to compute each character from one value.

Here a context lists types fixed before any variable is chosen. The next chapter allows a variable's type to depend on the preceding values. Its context extension then uses a family over a context, and substitution reindexes that family.

CHAPTER 27

Dependent types from families

A list of variables may include a natural number n and a vector of length n. The choice of the first value determines the set of available second values. A single type for the vector variable would forget this condition. A family of sets indexed by n records it, and the category of families supplies the operations for adding, substituting, and binding variables.

27.1 Contexts and their extensions

Interpret a context by a set Γ of environments. The empty context has one environment, so its interpretation is a singleton 1. A type $\Gamma \vdash A$ type assigns a set $A(\gamma)$ to each $\gamma \in \Gamma$. A term $\Gamma \vdash a : A$ assigns an element $a(\gamma) \in A(\gamma)$ to each environment. An empty fiber prevents the existence of a term when that fiber lies over an environment of Γ.

Write $\mathrm{Fam}(\Gamma)$ for the category of families over Γ. A map $u : A \to B$ is a function $u_\gamma : A(\gamma) \to B(\gamma)$ for each γ, and composition is fiberwise. A term of A is equivalently a map from the family of singletons to A. To add a variable $x : A$ to Γ, form

$$\Gamma.A = \{(\gamma, a) \mid \gamma \in \Gamma,\ a \in A(\gamma)\}, \qquad p_A(\gamma, a) = \gamma.$$

The extension is the total set of A, with projection $p_A : \Gamma.A \to \Gamma$. A term is also a section of p_A, by sending a to $\gamma \mapsto (\gamma, a(\gamma))$. Every function $e : E \to \Gamma$ conversely determines the family of fibers $\gamma \mapsto e^{-1}(\gamma)$, whose total set is isomorphic to E over Γ. The two constructions give an equivalence of $\mathrm{Fam}(\Gamma)$ with the slice $\mathbf{Set}/\Gamma$, as in Section 25.4.

A context map $\sigma : \Delta \to \Gamma$ sends environments for Δ to environments for Γ. It substitutes into types and terms by the formulas

$$A[\sigma](\delta) = A(\sigma(\delta)), \qquad a[\sigma](\delta) = a(\sigma(\delta)).$$

The expression on the right in the term formula has type $A[\sigma](\delta)$. This yields the substitution functor $\sigma^* : \mathrm{Fam}(\Gamma) \to \mathrm{Fam}(\Delta)$. Reindexing along an identity leaves every fiber unchanged. For $\tau : \Theta \to \Delta$, reindexing along σ and then τ is literally reindexing along $\sigma \circ \tau$. The same equations hold for terms and maps of families because their components are selected by the same indices.

There is a term q_A of $A[p_A]$ in context $\Gamma.A$, given by $q_A(\gamma, a) = a$. It interprets the variable just adjoined. For a term b of $A[\sigma]$ in context Δ, pairing gives $\langle \sigma, b\rangle(\delta) = (\sigma(\delta), b(\delta))$. It is the unique map displayed in the triangle with the indicated term property.

$$\begin{array}{ccc} \Delta & \dashrightarrow^{\langle\sigma,b\rangle} & \Gamma.A \\ & {\scriptstyle\sigma}\searrow & \downarrow{\scriptstyle p_A} \\ & & \Gamma \end{array} \qquad q_A[\langle \sigma, b\rangle] = b.$$

The triangle fixes the first component of the map, and the term equation fixes the second. Conversely each map into $\Gamma.A$ gives its substitution and term by these two projections. This representing property is called *context comprehension*. In particular, $\langle p_A, q_A\rangle = \mathrm{id}_{\Gamma.A}$.

Taking the generic term of $A[\sigma]$ gives the lifted map $\sigma^+ : \Delta.A[\sigma] \to \Gamma.A$, which sends (δ, a) to $(\sigma(\delta), a)$. Its square is a pullback.

$$\begin{array}{ccc} \Delta.A[\sigma] & \longrightarrow^{\sigma^+} & \Gamma.A \\ \downarrow{\scriptstyle p_{A[\sigma]}} & \lrcorner & \downarrow{\scriptstyle p_A} \\ \Delta & \longrightarrow_{\sigma} & \Gamma \end{array}$$

Indeed, a point in a pullback is a pair $\delta, (\gamma, a)$ with $\gamma = \sigma(\delta)$, which is specified by (δ, a). In particular, reindexing families is pullback of total sets. Equipping a category of contexts with terminal context, strictly reindexed types and terms, and comprehension by these equations gives a *category with families*. This is one way to formulate the structure required to interpret context-dependent syntax [19].

27.2 Dependent pairs

Take A over Γ and B over $\Gamma.A$. The type $\Sigma_{x:A}B$ over Γ contains pairs (a, b) with $a \in A(\gamma)$ and $b \in B(\gamma, a)$ in fiber γ. Use $\Sigma_A B$ for the same family, and use $p_A^* C$ for the family $(\gamma, a) \mapsto C(\gamma)$ whenever C is over Γ. A family map $u : \Sigma_A B \to C$ gives a map $v : B \to p_A^* C$ by restriction to the summands. The reverse map combines the functions on the summands. Explicitly,

$$v_{\gamma,a}(b) = u_\gamma(a, b), \qquad u_\gamma(a, b) = v_{\gamma,a}(b).$$

The two rules are inverse and commute with composition in B and C, giving the natural bijection

$$\mathrm{Hom}_{\mathrm{Fam}(\Gamma)}(\Sigma_A B, C) \cong \mathrm{Hom}_{\mathrm{Fam}(\Gamma.A)}(B, p_A^* C).$$

Under the slice description the total set of $\Sigma_A B$ is identified with $(\Gamma.A).B$ over Γ. Its projection to Γ is thus the composite of the two display projections.

$$\begin{array}{ccc} (\Gamma.A).B & \xrightarrow{\cong} & \Gamma.(\Sigma_A B) \\ \downarrow{\scriptstyle p_B} & & \\ \Gamma.A & & \downarrow{\scriptstyle p_{\Sigma_A B}} \\ \downarrow{\scriptstyle p_A} & & \\ \Gamma & = & \Gamma \end{array}$$

The top map takes $((\gamma, a), b)$ to $(\gamma, (a, b))$. This also gives dependent elimination. If D is a family over $\Gamma.(\Sigma_A B)$, a term of D is specified by giving a term after reindexing along that map. On the pair (a, b) the resulting term evaluates to the prescribed value at $((\gamma, a), b)$. The result type can vary with both entries of the pair.

In particular a term a of A and a term b of $B[\langle \mathrm{id}_\Gamma, a\rangle]$ pair to form a term of $\Sigma_A B$. Its two projections recover a and b, the two β-equations. For an arbitrary term s of the sum type, the first projection a is a term of A, and the second is a term of $B[\langle \mathrm{id}_\Gamma, a\rangle]$. Pairing these projections yields s again, the η-equation. Both equations follow from the description of the elements as pairs.

27.3 Dependent functions

The type $\Pi_{x:A} B$, written $\Pi_A B$, has in fiber γ all functions h on $A(\gamma)$ such that $h(a) \in B(\gamma, a)$. If b is a term of B in context $\Gamma.A$, its abstraction λb sends γ to $a \mapsto b(\gamma, a)$. A term f of $\Pi_A B$ in context Γ gives a term $\mathrm{app}(f)$ of B in context $\Gamma.A$ with $\mathrm{app}(f)(\gamma, a) = f(\gamma)(a)$. To apply f to a term a_0 of A in context Γ, substitute $\langle \mathrm{id}_\Gamma, a_0\rangle$ into this term. Its result type is $B[\langle \mathrm{id}_\Gamma, a_0\rangle]$.

Abstraction and application are induced by an adjunction. For a family map $v : p_A^* C \to B$, there is a unique $\widetilde{v} : C \to \Pi_A B$ with $\widetilde{v}_\gamma(c)(a) = v_{\gamma,a}(c)$. It factors v as shown over $\Gamma.A$.

$$\begin{array}{ccc} p_A^* C & \xrightarrow{p_A^* \widetilde{v}} & p_A^* \Pi_A B \\ & \searrow^{v} & \downarrow{\scriptstyle \mathrm{ev}} \\ & & B \end{array}$$

Here $\mathrm{ev}_{\gamma,a}(h) = h(a)$. Evaluation gives the inverse correspondence because a function is determined by its values. Precomposition applies a map of C-fibers before filling the function. Postcomposition applies a map of B-fibers to each value. Both actions commute with the correspondence, so it is natural. Together with dependent pairs, this gives the adjoint triple

$$\Sigma_A \dashv p_A^* \dashv \Pi_A.$$

Specializing the second bijection to the family of singletons gives the equations $\mathrm{app}(\lambda b) = b$ and $\lambda(\mathrm{app}(f)) = f$. The first, β, evaluates the function just defined. The second, η, reconstructs a function by its values. If $A(\gamma) = \varnothing$, there is one such function in the Π-fiber, the empty function, and no element in the Σ-fiber. As in Chapter 24, binding over an empty fiber distinguishes the two adjoints.

27.4 Binding after substitution

Suppose that $\sigma : \Delta \to \Gamma$ changes the parameters. A type B in the context $\Gamma.A$ must now be reindexed along the lifted map σ^+. Its fiber at (δ, a) becomes $B(\sigma(\delta), a)$, where $a \in A[\sigma](\delta)$. This gives the equations of families

$$(\Sigma_A B)[\sigma] = \Sigma_{A[\sigma]}(B[\sigma^+]),$$
$$(\Pi_A B)[\sigma] = \Pi_{A[\sigma]}(B[\sigma^+]).$$

Both sum fibers consist of the same pairs (a, b), and both product fibers are functions with the same domain and permitted values at each argument. Evaluation and pairing also agree under substitution since those operations act on exactly these pairs and functions. These are the Beck–Chevalley comparisons for the comprehension square. In an arbitrary pullback square of sets the adjoints for families use the corresponding fibers, and the comparisons are canonical isomorphisms as in Section 25.4. The chosen fiber formulas here make the displayed comparisons equalities.

Example 27.1 (Choosing a coordinate or choosing in every coordinate)**.** Let $[n] = \{i \in \mathbb{N} \mid i < n\}$ and take the context $\Gamma = \mathbb{N}$. Set $A(n) = [n]$ and $B(n, i) = [i + 1]$. An element of $(\Sigma_A B)(n)$ is a pair (i, j) with $0 \le j \le i < n$. A member of $(\Pi_A B)(n)$ is a function assigning a choice j_i in $[i + 1]$ to every $i < n$. Consequently the two fibers have respective sizes

$$\sum_{i=0}^{n-1}(i+1) = \frac{n(n+1)}{2}, \qquad \prod_{i=0}^{n-1}(i+1) = n!.$$

At $n = 2$, the pairs are $(0, 0), (1, 0), (1, 1)$. The functions are given by the value lists $(0, 0)$ and $(0, 1)$. At $n = 0$, there are no pairs and there is one function, given by the empty list. The section of the product family given by $j_i = i$ for all $i < n$ abstracts the term $b(n, i) = i$. Its application at i recovers precisely that term. Substitution along $m \mapsto m + 1$ reindexes the sizes to $(m + 1)(m + 2)/2$ and $(m + 1)!$. Computing the two families after reindexing A and B gives the same choices with $n = m + 1$.

A subobject $R \hookrightarrow \Gamma.A$ from Chapter 21 corresponds to a family whose fibers are empty or singletons. Its dependent sum remembers each eligible a together

with the unique proof of membership in R. Taking the image of that sum over Γ yields the existential predicate of Chapter 24. Its dependent product has at most one element in each fiber, so it directly represents the universal predicate. For general families the terms retain the actual choices and the right adjoint organizes one choice for each argument.

27.5 Families in other categories

Let $\mathcal{E}$ have finite limits. A categorical version of a type in context Y is an object $f : X \to Y$ in $\mathcal{E}/Y$. Terms are sections, and substitution along a map into Y is pullback. For $B \xrightarrow{\beta} X$ the dependent sum along f is $f \circ \beta : B \to Y$. A map from this composite to $C \to Y$ over Y corresponds to a map from $B \to X$ to the pullback $X \times_Y C \to X$, by the defining pullback square. Thus composition is left adjoint to pullback.

Suppose also that each slice is cartesian closed, so $\mathcal{E}$ is *locally cartesian closed*. Regard $\beta : B \to X$ as a map in the slice over Y, using $f \circ \beta$ on B. Write B^X and X^X for exponentials in this slice and $\ulcorner \mathrm{id}_X \urcorner : 1_Y \to X^X$ for the transpose of id_X. Here 1_Y is $\mathrm{id}_Y : Y \to Y$. Define $\Pi_f B$ by the equalizer in $\mathcal{E}/Y$

$$\Pi_f B \hookrightarrow B^X \underset{-\circ !\to \ulcorner \mathrm{id}_X \urcorner}{\overset{-\circ \beta \to}{\rightrightarrows}} X^X.$$

The arrow β^X is induced by postcomposition with β. A map $C \to B^X$ over Y transposes to $u : C \times_Y X \to B$ over Y. Equality of its two composites means that $\beta \circ u$ is the projection to X. This is precisely the condition for u to be a map from f^*C to B over X. The equalizer and transposition bijections are natural, so this gives $f^* \dashv \Pi_f$.

27.6 Base change in slices

The adjoints for slices also commute with pullback in the following sense. Take a pullback square in $\mathcal{E}$.

$$\begin{array}{ccc} X' & \xrightarrow{h} & X \\ {\scriptstyle f'}\downarrow & \lrcorner & \downarrow{\scriptstyle f} \\ Y' & \xrightarrow{g} & Y \end{array}$$

For $B \to X$, both $g^*\Sigma_f B$ and $\Sigma_{f'} h^* B$ are represented by the pullback of $B \to Y$ along g. For the second expression, this follows by pasting the square defining h^*B with the displayed square. Taking the same argument with the square read in the other direction gives $f^*\Sigma_g C \cong \Sigma_h (f')^* C$ for $C \to Y'$.

Write Hom_Y for hom sets in $\mathcal{E}/Y$, with the same convention for the other bases. In the locally cartesian closed case, apply the adjunctions and the last

isomorphism to get

$$\begin{aligned}\mathrm{Hom}_{Y'}(C, g^*\Pi_f B) &\cong \mathrm{Hom}_Y(\Sigma_g C, \Pi_f B)\\ &\cong \mathrm{Hom}_X(f^*\Sigma_g C, B)\\ &\cong \mathrm{Hom}_X(\Sigma_h (f')^* C, B)\\ &\cong \mathrm{Hom}_{X'}((f')^* C, h^* B)\\ &\cong \mathrm{Hom}_{Y'}(C, \Pi_{f'} h^* B).\end{aligned}$$

Naturality in C shows by Yoneda that $g^*\Pi_f B \cong \Pi_{f'} h^* B$. This is the product Beck–Chevalley isomorphism. It expresses compatibility of binding with substitution using only pullbacks and adjunctions.

For general chosen pullback squares, the two routes for successive substitution give canonically isomorphic objects over their common base. Pullback uniqueness supplies the comparison and makes the comparisons coherent for three successive substitutions, since both ways of comparing preserve the limiting projections. Literal syntactic substitution uses a split presentation with strictly functorial reindexing. Its chosen type formers must also satisfy their prescribed substitution equations as equalities. The indexed-set construction above has both properties for the type formers just considered. Chapter 29 uses a further kind of family. It starts by interpreting identity over two values of a type by the fibers of the diagonal.

CHAPTER 28

Inductive types from initial algebras

To define a function on lists, give its value at the empty list and say how a new entry changes the result on the tail. This specifies a map out of the list type by requiring it to preserve two operations. The natural numbers have the same pattern, with zero and successor. Initial algebras express that pattern for a general functor of constructors. A family over the carrier can also carry constructors, and applying initiality to its total space gives a dependent eliminator.

28.1 Constructors and recursion

Let $F\colon \mathcal{C} \to \mathcal{C}$ be an endofunctor. An *F-algebra* is an object X with a structure map $\alpha\colon FX \to X$. A morphism $h\colon (X,\alpha) \to (Y,\beta)$ is a map $h\colon X \to Y$ making the following square commute.

$$\begin{array}{ccc} FX & \xrightarrow{\alpha} & X \\ {\scriptstyle Fh}\downarrow & & \downarrow{\scriptstyle h} \\ FY & \xrightarrow{\beta} & Y \end{array}$$

Identities and composition preserve this equation by functoriality. Write $\mathrm{Alg}(F)$ for the resulting category. If (W,ω) is initial in this category, each structure $\beta\colon FY \to Y$ determines a unique morphism

$$\mathrm{fold}_\beta\colon W \to Y, \qquad \mathrm{fold}_\beta\ \omega = \beta\, F(\mathrm{fold}_\beta).$$

The displayed equation computes the result at a constructor by first applying the fold to the constructor's recursive inputs. Initiality supplies both the map and the uniqueness test for it. The choice $F(X) = 1 \amalg X$ recovers the zero-and-successor recursion of Chapter 6.

Lemma 28.1 (Lambek)**.** *The structure map of an initial F-algebra is an isomorphism in $\mathcal{C}$.*

Proof. The object FW has the structure $F\omega$, and $\omega\colon (FW, F\omega) \to (W,\omega)$ is a morphism of algebras by its defining equation. Let j be the unique morphism in the other direction. The endomorphism ωj of the initial algebra is id_W. The

square for j gives the other inverse equation.

$$\begin{array}{ccc} FW & \xrightarrow{\omega} & W \\ {\scriptstyle Fj}\downarrow & & \downarrow{\scriptstyle j} \\ FFW & \xrightarrow{F\omega} & FW \end{array} \qquad j\omega = F\omega\, Fj = F(\omega j) = \mathrm{id}_{FW}\,.$$

□

For the natural numbers this is the inverse decomposition of Chapter 6. If ω is a coproduct of constructors in **Set**, an element of W has one constructor tag and a uniquely determined tuple of inputs. The isomorphism alone supplies that decomposition, while the universal property supplies the fold. For instance the identity endofunctor of **Set** has $(\varnothing, \mathrm{id}_\varnothing)$ as its initial algebra. The algebra $(1, \mathrm{id}_1)$ also has invertible structure map, and it has no map to the empty algebra.

28.2 Folds on lists

Fix a set A and put $F_A(X) = 1 \amalg (A \times X)$. Its action on a function h is the identity on the first summand and $\mathrm{id}_A \times h$ on the second. An F_A-algebra on X consists of an element $e \in X$ and a function $c\colon A \times X \to X$. An algebra on the finite words of Chapter 5 uses the empty word $[]$ and the map $(a, w) \mapsto a :: w$ that puts a at the front of w. Write $\mathrm{List}(A)$ for this carrier.

For any (X, e, c) assign $f([]) = e$ and $f(a :: w) = c(a, f(w))$. The length of the tail decreases at every step, so induction on length constructs exactly one function with these equations. These are the algebra morphism equations on the two summands. Thus $\mathrm{List}(A)$ is initial for F_A. When $A = 1$, taking lengths identifies the list algebra with the natural numbers algebra. A different universal property of the same lists gives the free monoid. In that property the map into a monoid must preserve concatenation, and is determined by its values on the one-entry lists.

Suppose $k\colon (X, \alpha) \to (Y, \beta)$ is an algebra morphism and W is initial. The composite $k \,\mathrm{fold}_\alpha$ is an algebra morphism from W. Its structure equation follows from the two commuting squares.

$$\begin{array}{ccc} FW & \xrightarrow{\omega} & W \\ {\scriptstyle F(\mathrm{fold}_\alpha)}\downarrow & & \downarrow{\scriptstyle \mathrm{fold}_\alpha} \\ FX & \xrightarrow{\alpha} & X \\ {\scriptstyle Fk}\downarrow & & \downarrow{\scriptstyle k} \\ FY & \xrightarrow{\beta} & Y \end{array}$$

By uniqueness, $k \operatorname{fold}_\alpha = \operatorname{fold}_\beta$. This is the *fusion law*. It allows an operation following a fold to be absorbed into the target algebra when the two structure maps satisfy the morphism equation.

Example 28.2 (Folding directly into residues)**.** Use $A = \mathbb{Z}$, set the value of [] to zero, and take the list step $c(a, n) = a^2 + n$ on $\mathbb{Z}$. The resulting fold computes the sum of the squares of the entries. For modulus five there is also an algebra on $\mathbb{Z}/5\mathbb{Z}$, with zero and the step $\overline{c}(a, r) = \overline{a^2} + r$. The quotient $k(n) = \overline{n}$ preserves both the zero and the step, by the ring quotient equations of Chapter 11. Consequently the sum in $\mathbb{Z}$ followed by reduction equals the fold in the residue ring. For $[2, -1, 3]$, the integer fold gives $4 + 1 + 9 = 14$. The residue fold works from the tail and gives 0, then 4, then 0, then 4 in $\mathbb{Z}/5\mathbb{Z}$. The morphism equation verifies the replacement for every list.

28.3 Dependent induction as a section

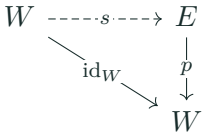
Recursion into a single set uses a fixed target for every constructor. In dependent induction the target at w is a fiber $P(w)$. Write $p\colon E \to W$ for the total space of this family, so $E = \coprod_{w \in W} P(w)$ and $p(w, z) = w$. A dependent term amounts to a section of p, as in Chapter 27. Constructors on the fibers can be packaged into an algebra $e\colon FE \to E$ such that

$$pe = \omega F p.$$

The equation says that the result of a constructor on E lies over the corresponding constructor on W. More generally we can ask for such an algebra on any map p in $\mathcal{C}$.

Proposition 28.3. *For an initial algebra (W, ω), every algebra morphism $p\colon (E, e) \to (W, \omega)$ has a unique section which is an algebra morphism. If $\mathcal{C}$ has finite limits, the existence of such a section for every p already implies initiality.*

Proof. Initiality gives a unique algebra morphism $s\colon W \to E$. The composite ps is an algebra endomorphism of W, so it equals id_W. This proves the section property.



Conversely, for any algebra (X, a), give $W \times X$ the structure $\langle \omega F \pi_W, a F \pi_X \rangle$. Both projections are algebra morphisms, as verified by composing with them. A section of π_W yields a morphism to X by composing with π_X. If $f, g\colon W \to X$

are algebra morphisms, let $i\colon I \to W$ be their equalizer in $\mathcal{C}$. The map $\omega F i$ equalizes f, g, since $f\omega F i = aF(fi) = aF(gi) = g\omega F i$. Thus $\omega F i$ factors through i as a map $FI \to I$ that makes i an algebra morphism. A section r of i gives $f = fir = gir = g$. □

For lists, the fiber constructors are data

$$z_0 \in P([]), \qquad d_{a,w}\colon P(w) \to P(a :: w) \quad (a \in A,\ w \in \mathrm{List}(A)).$$

They give e by the empty list value $([], z_0)$ and the step $(a, (w, z)) \mapsto (a :: w, d_{a,w}(z))$. The projection p preserves the constructors. Writing the section of Proposition 28.3 as $w \mapsto (w, t_w)$, we obtain

$$t_{[]} = z_0, \qquad t_{a::w} = d_{a,w}(t_w).$$

The equations follow by projecting the algebra morphism equation onto the corresponding fibers. If each $P(w)$ is a singleton when a stated property holds and empty otherwise, this is proof by induction. General families can carry additional data.

Example 28.4 (Numbering a list)**.** Write $[n] = \{0, \dots, n-1\}$ and let $\ell(w)$ denote the length of w. Take $P(w)$ to be the set of functions $[\ell(w)] \to \mathbb{N} \times A$. The base element is the empty function. For $v \in P(w)$, define $d_{a,w}(v)$ to have value $(0, a)$ at 0, and value $(i+1, \pi_2 v(i))$ at $i+1$ for $i \in [\ell(w)]$. Since $\ell(a :: w) = 1 + \ell(w)$, this specifies a value at every position of the new list. Dependent induction gives a function on the positions of each input list. At $[a, b, c]$ the values are $(0, a), (1, b), (2, c)$. The domain of the output is indexed by the length of the input.

28.4 Well-founded trees

Let A be a set of constructor labels and let B_a be a set of input positions for each $a \in A$. Consider the *polynomial functor*

$$F(X) = \coprod_{a \in A} X^{B_a}, \qquad F(h)(a, u) = (a, h \circ u).$$

The label chooses a constructor and the function $u\colon B_a \to X$ supplies its recursive inputs. The variable X occurs in the values of these functions, and postcomposition gives the covariant action. Natural numbers use a nullary and a unary label. Lists over a set L use a nullary label and one unary label for each element of L.

An initial algebra for this functor in **Set** can be built as a set of rooted labelled trees. Put $S = \coprod_{a \in A} B_a$. An address is a finite word on S, with the empty word as root. Use a prefix-closed subset of the addresses containing the root, and

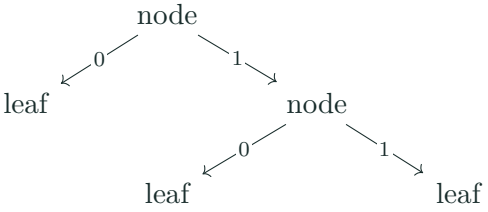
label each node by an element of A. At a node with label a, require exactly the children obtained by appending (a, b) for $b \in B_a$. Require the child relation to be *well-founded*. Every nonempty subset of nodes must contain a node with no child in that subset. Such labelled trees form a set, since the addresses form a set and the labels can be specified by the graph of a function on a subset of the addresses. Call this set W.

There is a constructor $\sup(a, u)$ for each label a and tuple $u\colon B_a \to W$. Form a root with label a, and attach the tree $u(b)$ below the address (a, b). The result is well-founded. For any subset of nodes meeting a branch, choose a node with no child in the part of the subset in that branch. It has no child in the rest of the subset either. If the subset contains only the root, that root meets the condition. Conversely, the root label and the subtrees immediately below it recover the constructor input.

To prove initiality, take an algebra $\alpha\colon FX \to X$. On the nodes of each tree, require a value at a node with label a to be $\alpha(a, v)$, where v assigns each child position the value already given to that child. There is exactly one assignment on the subtree at each node. If some subtrees lacked a unique assignment, choose a node in that set with no child in it. Each child subtree then has a unique assignment. Their domains are disjoint, so they combine and the equation fixes the value at the root of the chosen subtree. Every assignment there must use those same child assignments and that same root value. This contradicts the choice of node. Taking the value at the root of each tree defines the fold. The constructor equation holds by the node rule. Any morphism to (X, α) assigns values throughout a tree by applying that morphism to its subtrees. These values satisfy the same rule, so uniqueness of the assignment forces the morphism to be the fold.

Example 28.5 (Counting leaves)**.** Use a nullary label for a leaf and a binary label for a node. The resulting set consists of finite full binary trees. Indeed each leaf subtree is finite, and a node with two finite subtrees is finite, so well-founded induction proves finiteness. Let I count internal nodes and L count leaves. The fold into $\mathbb{N} \times \mathbb{N}$ takes a leaf to $(0, 1)$ and combines $(i_1, l_1), (i_2, l_2)$ as $(1 + i_1 + i_2, l_1 + l_2)$. The pairs with $l = i + 1$ form a subalgebra. The base pair lies there, and if both inputs lie there then $l_1 + l_2 = i_1 + i_2 + 2$. Folding into the subalgebra and including the result gives the original count by fusion. Hence $L = I + 1$.

In the tree below the labels on the arrows name the two positions at a binary node. Its right subtree has count $(1, 2)$, so the root combines $(0, 1)$ and $(1, 2)$ to give $(2, 3)$.



28.5 Counting binary trees by their constructors

For the full binary trees above, the two child positions are ordered. Let W_n be the set of trees with n internal nodes, and write $c_n = |W_n|$. The leaf is the only element of W_0. The two subtrees extracted at the root determine the inverse of the graded constructor bijection

$$\coprod_{i+j=n} W_i \times W_j \;\cong\; W_{n+1}.$$

Taking cardinalities gives $c_0 = 1$ and $c_{n+1} = \sum_{i+j=n} c_i c_j$. In the formal power series $C(z) = \sum_{n \geq 0} c_n z^n$, the same recursion is $C(z) = 1 + zC(z)^2$. The disjoint choice of a constructor becomes a sum, the two fields of a node become a product, and the additional internal node accounts for the factor z.

A direct bijection gives the closed count. Encode the leaf by the empty path and a node with children u, v by an up-step, the path for u, a down-step, and the path for v. A tree in W_n gives a path of $2n$ steps that starts and ends at height zero and stays at nonnegative height. Every nonempty such path starts with an up-step. Its first return to zero identifies the closing down-step and separates the two child paths. Repeating the decomposition reconstructs a unique tree. In the path below, the teal steps mark the two steps contributed by the root. Each child contributes a two-step path.



There are $\binom{2n}{n}$ unrestricted paths of $2n$ steps from zero back to zero. For a path that falls below zero, exchange up- and down-steps through its first visit to height -1. This produces a path ending at height 2, with $n + 1$ up-steps. Conversely, such a path first reaches 1 somewhere. Reflecting through that first visit recovers the original path. Thus $\binom{2n}{n+1}$ paths must be removed, and the

count is the n-th Catalan number

$$c_n = \binom{2n}{n} - \binom{2n}{n+1} = \frac{1}{n+1}\binom{2n}{n}.$$

The first counts are $1, 1, 2, 5, 14$. For $n = 3$, the ordered sizes of the two child trees are $(0, 2), (1, 1), (2, 0)$. They give $2 + 1 + 2 = 5$ trees, agreeing with the path count. The reflection count also appears in Postnikov [68, Lecture 2].

28.6 Induction for trees and change of parameters

For a family $P(w)$ over W, constructor data for dependent induction consist of functions

$$d_{a,u} \colon \prod_{b \in B_a} P(u(b)) \longrightarrow P(\sup(a, u)).$$

To construct the algebra on the total space, an input chooses a and a map $B_a \to E$. Write this map as $b \mapsto (u(b), v_b)$. Its output is $(\sup(a, u), d_{a,u}((v_b)_b))$. The projection p is an algebra morphism. Proposition 28.3 gives a section t, and its algebra equation reads

$$t_{\sup(a,u)} = d_{a,u}\big((t_{u(b)})_{b \in B_a}\big).$$

This is the induction rule for the W-type in sets.

The constructor data can vary in a set of parameters Γ. For every $\gamma \in \Gamma$, take a set A_γ of labels and sets $B_{\gamma,a}$ of positions. Their polynomial functors act on families over Γ by applying the indicated functor in each fiber. The trees W_γ give an initial algebra of this family functor, because a map of families is a collection of maps of its fibers and each required algebra map is unique there. For $\sigma \colon \Delta \to \Gamma$, the substituted data at δ are $A_{\sigma(\delta)}$ and $B_{\sigma(\delta),a}$. They give exactly the same tree and constructor equations as the fiber at $\sigma(\delta)$. Uniqueness of the fold and of the algebra section shows that both reindex by the same rule. For the explicit families of Chapter 27, this is evaluation at $\sigma(\delta)$. Their total spaces compare with chosen pullbacks by the canonical bijections.

In general categories the fold assigns a map to each target algebra and gives equations of morphisms. To interpret rules of dependent type theory requires a semantics of families and constructor data over contexts, with stability under substitution as in the preceding construction. Judgmental computation rules also require a choice of semantic structure in which the specified equalities hold. Homotopy type theory uses a corresponding homotopy-initial property in which the space of algebra morphisms is contractible. The distinction between equality of maps and terms of an identity type is developed in the next chapter. See *Homotopy Type Theory* [21, Sections 5.3–5.5] for the homotopy-initial formulation.

CHAPTER 29

Identity types and universes

The diagonal of a set describes pairs whose entries are equal. An identity type also asks for a witness of that equality. In sets there is at most one such witness for each pair. In a groupoid the witnesses are arrows, and two arrows between the same objects can act differently on a dependent type. The identity rules describe how to use a witness in either interpretation. When the objects are types themselves, the same rules let a path act on the elements of the types it connects.

29.1 Identity and its elimination rule

Let A be a type in context Γ. In the extended context $\Gamma, x\colon A, y\colon A$, the *identity type* $\mathrm{Id}_A(x, y)$ is a type. The notation $x =_A y$ abbreviates it. The introduction rule gives

$$\Gamma, x\colon A \;\vdash\; \mathrm{refl}_x\colon \mathrm{Id}_A(x, x).$$

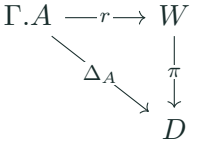
We call a term of $\mathrm{Id}_A(x, y)$ a *path* from x to y. The judgment $t \equiv u$ denotes judgmental equality, the equality used by the typing rules to convert terms and types. An identity witness $p\colon \mathrm{Id}_A(x, y)$ is a term to which the following elimination rule can be applied.

Suppose $C(x, y, p)$ is a type in context $\Gamma, x\colon A, y\colon A, p\colon \mathrm{Id}_A(x, y)$. This type is called the *motive*. Given a term $d(x)$ of $C(x, x, \mathrm{refl}_x)$ in context $\Gamma, x\colon A$, the eliminator J provides

$$J_C(d, x, y, p)\colon C(x, y, p),$$
$$J_C(d, x, x, \mathrm{refl}_x) \equiv d(x).$$

The second line is the β-rule for J. It computes the eliminator at the constructor. Both rules allow parameters in Γ. In the set interpretation of Chapter 27, a parameter assignment γ fixes the fiber $A(\gamma)$ in which x and y live.

In the language of context extensions, write D for $\Gamma, x\colon A, y\colon A$ and W for D extended by the identity witness. The map r sends x to (x, x, refl_x) over its parameter. Its composite with the projection to D is the diagonal over Γ.



The type C has a display map $E \to W$, where E is its total context. The term d is a section of the pullback of this map along r. Equivalently it is a map $d\colon \Gamma.A \to E$ over r. The interpretation of J supplies the dashed section in the next square.

$$\begin{array}{ccc} \Gamma.A & \xrightarrow{d} & E \\ \downarrow{\scriptstyle r} & \overset{J_C(d)}{\nearrow} & \downarrow \\ W & \xrightarrow{\mathrm{id}_W} & W \end{array}$$

The upper triangle states the computation equation. This is a dependent extension property. It must apply to motives that vary with the witness p, as well as to motives depending on its endpoints. In particular the rule for identity is stronger than the assertion that the display map has a reflexivity term.

29.2 Transport and path operations

For a family B over A, take the motive $C(x, y, p) = B(x) \to B(y)$. The reflexivity term can be the identity function on $B(x)$. The associated J gives

$$\mathrm{tr}^B_p\colon B(x) \longrightarrow B(y), \qquad \mathrm{tr}^B_{\mathrm{refl}_x} \equiv \mathrm{id}_{B(x)}\,.$$

This is *transport*. It gives a typed substitution of equals inside a family. A map $f\colon A \to A'$ acts on paths by taking the motive $f(x) =_{A'} f(y)$ and the reflexivity term $\mathrm{refl}_{f(x)}$. Its action is denoted ap_f.

The motive $y =_A x$ gives an inverse p^{-1} from y to x, by taking reflexivity as the starting term. To compose $p\colon x =_A y$ with $q\colon y =_A z$, eliminate q with motive $(x =_A y) \to (x =_A z)$ and start with the identity on $x =_A y$. Applying the result to p gives $q \circ p$. The computation rule gives $\mathrm{refl}_y \circ p \equiv p$. For the other unit, eliminate p with motive

$$(p \circ \mathrm{refl}_x) =_{\mathrm{Id}_A(x,y)} p.$$

At reflexivity both sides compute to refl_x, so reflexivity at that term supplies the starting case. The inverse laws follow by eliminating p in the same way. For associativity of $s \circ (q \circ p)$ and $(s \circ q) \circ p$, eliminate s. Both sides at $s = \mathrm{refl}_z$ compute to $q \circ p$. Thus the groupoid laws hold by paths between paths, with the stated computation rule giving some of the laws judgmentally.

Transport along p^{-1} is an inverse to transport along p up to paths. The two inverse laws for functions applied to an argument reduce by eliminating p to reflexivity. For p, q as above, eliminating q gives a path from $\mathrm{tr}^B_{q\circ p}(b)$ to $\mathrm{tr}^B_q(\mathrm{tr}^B_p(b))$. Both expressions compute to $\mathrm{tr}^B_p(b)$ in the reflexivity case. These constructions use identity types again at the types of the computed terms.

29.3 Equality in the set model

In sets, interpret $\mathrm{Id}_A(x, y)$ as $\{*\}$ if $x = y$ in the fiber and as $\varnothing$ otherwise. The map r above sends (γ, x) to $(\gamma, x, x, *)$. It is a bijection onto W, with inverse forgetting the repeated entry and the witness. Given d, define $J_C(d, \gamma, x, x, *) = d(\gamma, x)$. There is exactly one value prescribed at each element of W, and the values belong to the correct fibers of C. This calculation supplies the eliminator and its computation rule. Reindexing the context along $f\colon \Gamma' \to \Gamma$ replaces γ by $f(\gamma')$ on both sides of the calculation, which verifies substitution.

In this interpretation $W \to D$ is a monomorphism representing the diagonal subobject of Chapter 21. A path exists just when the two entries agree, so the model validates uniqueness of identity proofs. In particular the set model interprets each identity type as a set with at most one element. The groupoid interpretation will use the same formal rules with different witness sets.

29.4 Identity witnesses in a groupoid

A *groupoid* is a category in which every arrow has an inverse. In the groupoid interpretation a context Γ is a groupoid. A family over it is a functor A assigning a groupoid $A(\gamma)$ to each object and a functor $A(u)$ to each arrow $u\colon \gamma \to \gamma'$. The total groupoid $\Gamma.A$ has objects (γ, x), $x \in A(\gamma)$, and arrows (u, g), where $g\colon A(u)(x) \to x'$ in $A(\gamma')$. For a further arrow (u', g') the fiber part of the composite is $g' \circ A(u')(g)$, and its base part is $u' \circ u$. A section of the projection to Γ assigns an object in each fiber and compatible arrows for each u. This is the groupoid version of a term of a family. Reindexing a family precomposes its functor with the map of contexts.

For (γ, x, y) the witnesses of $x =_A y$ form the *discrete* groupoid on the set $A(\gamma)(x, y)$. A discrete groupoid has identity arrows only. The action on witnesses along an arrow $(u, g, h)\colon (\gamma, x, y) \to (\gamma', x', y')$ of D is

$$p \longmapsto h \circ A(u)(p) \circ g^{-1}.$$

Here $g\colon A(u)(x) \to x'$ and $h\colon A(u)(y) \to y'$. The action preserves identity arrows and composites by functoriality of A and cancellation with inverses. In the total groupoid W, the arrow from (γ, x, y, p) to (γ', x', y', p') over (u, g, h) therefore exists precisely when $h \circ A(u)(p) = p' \circ g$. The equality is the commutation of the following square in $A(\gamma')$.

$$\begin{array}{ccc} A(u)(x) & \xrightarrow{A(u)(p)} & A(u)(y) \\ \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle h} \\ x' & \xrightarrow{\quad p' \quad} & y' \end{array}$$

Reflexivity r sends (γ, x) to $(\gamma, x, x, \mathrm{id}_x)$ and (u, g) to (u, g, g). Let $s\colon W \to \Gamma.A$ forget the second endpoint and the witness. Then $sr = \mathrm{id}_{\Gamma.A}$. At $w = (\gamma, x, y, p)$ the arrow

$$c_w = (\mathrm{id}_\gamma, \mathrm{id}_x, p)\colon rs(w) \longrightarrow w$$

is an isomorphism in W. For $k = (u, g, h)\colon w \to w'$ the square above gives $k \circ c_w = c_{w'} \circ rs(k)$. The c_w are thus a natural isomorphism $rs \Rightarrow \mathrm{id}_W$ which is the identity at reflexivity.

Now interpret C as a family of groupoids over W. A section d over r assigns an object $d(z)$ of $C(r(z))$ to each $z \in \Gamma.A$. For an arrow $v\colon z \to z'$ it has a compatible arrow $d_v\colon C(r(v))(d(z)) \to d(z')$. For $J_C(d)$ assign the object $C(c_w)(d(s(w)))$ at w. Its arrow over $k\colon w \to w'$ is obtained by applying $C(c_{w'})$ to $d_{s(k)}$. The source of this arrow is $C(k)C(c_w)(d(s(w)))$ by naturality of c, and its target is $C(c_{w'})(d(s(w')))$. The section axioms follow from the section axioms for d, using $kc_w = c_{w'}rs(k)$ to move each transition of C past the indicated component of c. At $r(z)$, the arrow $c_{r(z)}$ is an identity, and sr is the identity on objects and arrows. Hence the new section restricts to d exactly. The formulas depend only on the family at the substituted parameters and its action on arrows. They commute with precomposition by a context map. This gives the identity-type rules for these groupoid families. Hofmann and Streicher [20] develop the groupoid model of intensional type theory.

Example 29.1 (Two loops and a dependent term)**.** Let A be the one-object groupoid BG for the group $G = \{e, \sigma\}$ with $\sigma^2 = e$. Its object x has two loops e, σ, so the identity type $x =_A x$ has two objects. They are distinct in a discrete groupoid, and the identity type between them is empty.

Take the family B sending x to the discrete two-element groupoid $\{0, 1\}$, with $B(e)$ the identity and $B(\sigma)$ the swap. This defines a functor because the swap squared is the identity. Transport along an arrow p here is $B(p)$. Indeed the formula for J moves the identity function in the reflexivity motive along c_w, which fixes the source fiber and acts by $B(p)$ on the target fiber. In particular $\mathrm{tr}^B_e(0) = 0$ and $\mathrm{tr}^B_\sigma(0) = 1$.

$$\begin{array}{ccc} (x, 0) & \overset{(\sigma, \mathrm{id}_1)}{\longrightarrow} & (x, 1) \\ \downarrow & & \downarrow \\ x & \xrightarrow{\quad\sigma\quad} & x \end{array}$$

The top arrow is in $A.B$, and the vertical arrows denote its projection to A. Its fiber part is the identity at 1, since $B(\sigma)(0) = 1$.

29.5 Universes as families of types

A *Tarski-style universe* $\mathcal{U}_i$ is a type of codes, with a decoding family $\mathrm{El}_i(a)$ for $a : \mathcal{U}_i$. A universe supplies constructors for the type formers under which its decoded types are closed. For example, for $a : \mathcal{U}_i$ and $b : \prod_{x:\mathrm{El}_i(a)} \mathcal{U}_i$, the code $\ulcorner\Pi\urcorner_i(a, b)$ satisfies the decoding rule

$$\mathrm{El}_i(\ulcorner\Pi\urcorner_i(a, b)) \equiv \prod_{x:\mathrm{El}_i(a)} \mathrm{El}_i(b(x)).$$

There are corresponding rules for dependent sums, and an identity code at a, x, y decodes to $\mathrm{Id}_{\mathrm{El}_i(a)}(x, y)$. A code for $\mathcal{U}_i$ itself may be provided in $\mathcal{U}_{i+1}$. The index records the level at which the universe can be used as a decoded type.

For one set model, assume in the metatheory a sequence of increasing strongly inaccessible cardinals κ_i. Use $\mathcal{U}_i = V_{\kappa_i}$, the set of all sets of rank less than κ_i, and decode each set by itself. The family is represented by the projection $\{(a, x) \mid a \in \mathcal{U}_i,\ x \in a\} \to \mathcal{U}_i$. A family of sets over Γ with values in $\mathcal{U}_i$ determines a code map $a : \Gamma \to \mathcal{U}_i$. Its total set is the pullback of the decoding projection.

$$\begin{array}{ccc} \{(\gamma, x) \mid x \in a(\gamma)\} & \longrightarrow & \{(u, x) \mid x \in u \in \mathcal{U}_i\} \\ \downarrow & \lrcorner & \downarrow \\ \Gamma & \xrightarrow{\qquad a \qquad} & \mathcal{U}_i \end{array}$$

Under substitution by $f : \Gamma' \to \Gamma$, the new code map is $a \circ f$. This is the same reindexing as pulling back the family. Decoding rules are interpreted by choosing the codes to be the actual sets built by the indicated type formers.

The size hypothesis ensures the required closure. For any $\alpha < \kappa_i$, the set V_α has cardinality less than κ_i, by the strong limit property at successor stages and regularity at limit stages. Thus a set $a \in V_{\kappa_i}$ has cardinality less than κ_i. The ranks of the sets $b(x)$, $x \in a$, have a common bound below κ_i by regularity. The dependent sum is a set of ordered pairs and the dependent product is a set of functions whose graphs are sets of such pairs. Their ranks therefore lie below κ_i as well. The empty and singleton sets used for identities belong to V_{κ_i}. So does $\mathbb{N}$, whose rank in this model is ω. Finally V_{κ_i} has rank κ_i and belongs to $V_{\kappa_{i+1}}$. This accounts for the successor level in the decoding rule for a smaller universe.

29.6 The comparison with equivalences

Transport in El_i sends a path $p : a =_{\mathcal{U}_i} b$ to a function $\mathrm{El}_i(a) \to \mathrm{El}_i(b)$. The inverse transport and its inverse laws exhibit this function as an equivalence. We write the resulting canonical map as

$$\mathrm{idtoequiv} : (a =_{\mathcal{U}_i} b) \longrightarrow (\mathrm{El}_i(a) \simeq \mathrm{El}_i(b)).$$

On reflexivity it gives the identity equivalence. The *univalence* condition for a universe says that this map is an equivalence for each a, b. Type equivalences and their higher coherence are developed further in the later chapters on homotopy type theory. In the examples here, the decoded types are sets, for which equivalences are precisely bijections.

For the set model above take $a = b$ the code for $\{0, 1\}$. The source of idtoequiv is a singleton, and the target contains both the identity and the swap. The single identity witness maps to the identity bijection. This model validates the identity and decoding rules but fails the univalence condition.

There is another interpretation of the same code objects in the groupoid setting. Form the groupoid with objects the small sets and arrows their bijections, using the levels above for size. Its decoding family sends a set to the discrete groupoid on its elements and sends a bijection to its action on those elements. A groupoid-indexed family of small discrete types, with its chosen transition maps, is a functor into this groupoid of codes. Each transition is a bijection because each arrow in the indexing groupoid is invertible. Pullback of the decoding family along the code functor recovers the original family.

At a code pair (a, b) the identity type in this groupoid is the discrete type of bijections $a \to b$. Transport in the decoding family applies exactly the given bijection. The canonical comparison above therefore acts as the identity on the set of bijections. Its compatibility with changes of endpoints follows from the same action on both sides, sending $p\colon a \to b$ to $v \circ p \circ u^{-1}$ along bijections $u\colon a \to a'$ and $v\colon b \to b'$. The swap has its own loop at the two-element code. The bijection $0 \mapsto 2,\ 1 \mapsto 3$ is a path between the distinct code objects $\{0, 1\}$ and $\{2, 3\}$. This calculation concerns the universe of small discrete types with its specified decoding family. Allowing codes for types with higher identity structure requires correspondingly richer paths between codes. The treatment of univalence in such universes returns in Part XII [21, Sections 2.3 and 2.10].

CHAPTER 30

Proofs as terms

Suppose that an assumption $a\colon A$ allows us to derive a conclusion $t\colon B$. Discharging the assumption produces a proof of $A \Rightarrow B$. Given a proof of A, we can apply that implication and recover a proof of B. The same pattern defines a function by abstraction and runs it by application. The *Curry–Howard correspondence* identifies these rules and their computations. It gives a way to read a derivation as a term and the term as a map with a universal property. Lambek and Scott [18] develop the relationship between proofs, types, and cartesian closed categories.

30.1 Assumptions and implication

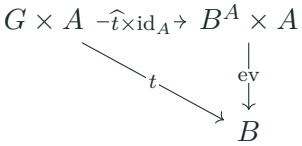
Use the judgments of Chapter 26 with the types read as propositions. A context Γ lists named assumptions. A derivation of $\Gamma \vdash t\colon A$ records a proof of A from those assumptions. An assumption can be used by naming it, and a proof can be substituted for any assumption of its type. Weakening, exchange, and contraction are implemented by changing the context along projections, symmetry, and diagonals.

In an intuitionistic natural deduction calculus, implication introduction discharges a named assumption. Implication elimination uses a proof of the antecedent to obtain the consequent. Writing $A \to B$ for the function type, the rules with their terms are

$$\frac{\Gamma, x\colon A \vdash t\colon B}{\Gamma \vdash \lambda x.\, t\colon A \to B} \qquad \frac{\Gamma \vdash f\colon A \to B \qquad \Gamma \vdash a\colon A}{\Gamma \vdash f\, a\colon B}.$$

A bound name records which assumption is discharged. If there are other assumptions of type A, they remain available.

In a cartesian closed category write G for the product interpreting Γ. The premiss of introduction is a map $t\colon G \times A \to B$. Its transpose $\widehat{t}\colon G \to B^A$ makes the top and right arrows of the following triangle compose to t.



If $a\colon G \to A$ is a proof of the antecedent, the proof $f\, a$ is $\mathrm{ev} \circ \langle f, a \rangle$. Substituting $\langle \mathrm{id}_G, a \rangle$ into the triangle proves $(\lambda x.\, t)\, a = t[a/x]$. The exponential uniqueness

gives the converse calculation $\lambda x.\, f\, x = f$ when x is fresh for f. These are the β- and η-equations, respectively. The former computes an elimination after an introduction.

Example 30.1 (Composition as an intuitionistic proof)**.** From proofs $f\colon A \to B$ and $g\colon B \to C$, assume $a\colon A$, obtain $f\, a\colon B$, and then obtain $g(f\, a)\colon C$. Discharging a, then g, then f gives the closed term

$$\lambda f.\, \lambda g.\, \lambda a.\, g(f\, a) \quad : \quad (A \to B) \to (B \to C) \to (A \to C),$$

with arrows associated to the right. Applying this term to f, g, and a computes $g(f\, a)$ through three β-steps. As an arrow, this proof composes the two evaluations $B^A \times A \to B$ and $C^B \times B \to C$ and then transposes the resulting map in the required arguments.

30.2 Conjunction, disjunction, and their cases

The type $A \times B$ reads as conjunction. Introduction pairs two proofs, and the two projections give elimination. The product property relates the proof of a conjunction to its two components.

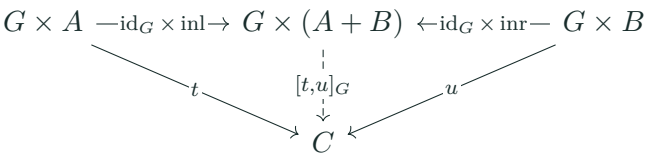
$$\mathrm{Hom}(G, A \times B) \cong \mathrm{Hom}(G, A) \times \mathrm{Hom}(G, B).$$

The unit type 1 reads as truth. Its canonical proof $*$ in context G is the unique arrow $G \to 1$.

For disjunction work in a cartesian closed category with an initial object 0 and binary coproducts. The two disjunction introduction rules take $a\colon A$ to $\mathrm{inl}(a)\colon A + B$ and $b\colon B$ to $\mathrm{inr}(b)\colon A + B$. To eliminate a term $s\colon A + B$, give a branch $t\colon C$ in context $\Gamma, x\colon A$ and a branch $u\colon C$ in context $\Gamma, y\colon B$. Then in context Γ the term

$$\text{case } \; s \; \text{ of } \; \big(\mathrm{inl}(x) \mapsto t,\, \mathrm{inr}(y) \mapsto u\big)\colon C$$

uses the branch for the given injection. The categorical proof of this rule must pass the common context to both branches. For fixed G, the functor $G \times -$ is left adjoint to $(-)^G$. The exponential and coproduct tests identify maps from $G \times (A + B)$ with pairs of maps from $G \times A$ and $G \times B$. Thus the diagram below gives a coproduct in that context and uniquely specifies the map $[t, u]_G$.



Compose the dashed arrow with $\langle \mathrm{id}_G, s\rangle$ to interpret the case term. The two commuting triangles give the β-equations on injections. Coproduct uniqueness gives the corresponding η-equation for a map defined on both injections. The same exponential test shows that $G \times 0$ is initial. A term of type 0 therefore eliminates to a term of any type by composing with the unique map $0 \to C$. Read 0 as falsity, and read $\neg A$ as $A \to 0$.

Example 30.2 (Distributing a proof across cases)**.** The map

$$d\colon (A \times B) + (A \times C) \longrightarrow A \times (B + C)$$

sends $\mathrm{inl}\langle a, b\rangle$ to $\langle a, \mathrm{inl}(b)\rangle$ and $\mathrm{inr}\langle a, c\rangle$ to $\langle a, \mathrm{inr}(c)\rangle$. For the reverse map e, split the input z into $\pi_1 z$ and $\pi_2 z$, and match on the second component. Its two branches are

$$\mathrm{inl}(b) \longmapsto \mathrm{inl}\langle \pi_1 z, b\rangle,$$
$$\mathrm{inr}(c) \longmapsto \mathrm{inr}\langle \pi_1 z, c\rangle.$$

On the first injection, ed takes a pair $\langle a, b\rangle$ back to $\mathrm{inl}\langle a, b\rangle$ by the product and case β-equations. It takes the second injection back to itself by the other case equation. Coproduct uniqueness gives $ed = \mathrm{id}$. On inputs $\langle a, \mathrm{inl}(b)\rangle$ and $\langle a, \mathrm{inr}(c)\rangle$, de yields the same inputs. These are the two coproduct injections in context A. Their universal property gives $de = \mathrm{id}$.

These equations correspond to reducing a proof immediately followed by use of its conclusion. For instance, choosing the left injection and then matching on it leaves just the selected branch with its assumption replaced by the provided proof. Such local reductions preserve types by the substitution rule. For example, the branch $\Gamma, x\colon A \vdash t\colon C$ and input $\Gamma \vdash a\colon A$ give the reduct $\Gamma \vdash t[a/x]\colon C$, the same type as the case term. A normalization theorem concerns termination of reductions for an entire calculus. Its hypotheses depend on the term formers and computation rules.

30.3 When truth forgets a proof

Different terms can substantiate the same assertion. In **Set**, the type $1 + 1$ has two closed terms $\mathrm{inl}(*)$ and $\mathrm{inr}(*)$. Their values in the coproduct are different points. If the assertion is only that this type is inhabited, both give the same answer. The unique map $1 + 1 \to 1$ sends either witness to $*$.

This loss of information is built into predicates as subobjects. Fix a context X in an elementary topos. A morphism of predicates over X is a factorization of one subobject through another and is unique if it exists, because the target is monic. As Chapter 23 shows, $\mathrm{Sub}(X)$ is a Heyting algebra. In its order viewed as a category, the exponential law is precisely

$$(R \wedge P) \leq Q \quad \Longleftrightarrow \quad R \leq (P \Rightarrow Q).$$

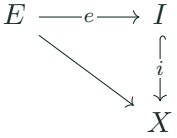
Thus this category also interprets the implication rules. The meet and join interpret conjunction and disjunction by the product and coproduct properties. In this reading, a proof from one predicate to another is the unique order arrow whenever the entailment holds. Arbitrary objects and maps can record multiple ways of witnessing a type. In **Set**, the map $A \to 1$ forgets which element of A was supplied.

Example 30.3 (Excluded middle and the proof rules)**.** In the two-stage presheaf topos of Section 23.5, take the middle subobject M of the terminal object. Its negation is bottom, so $M \vee \neg M = M < \top$. In the corresponding Heyting order the arrow $\top \to M \vee \neg M$ does not exist. Every closed derivation with the rules above would interpret as such an arrow under this assignment to its proposition symbol. Consequently there is no derivation of the general formula $A \vee \neg A$ in these intuitionistic rules.

30.4 Quantifiers and the information in witnesses

Dependent types change what can be recorded in a proof. A family B on A gives the type $\sum_{a:A} B(a)$, whose term is a pair $\langle a, b\rangle$ with $b\colon B(a)$. Its first projection returns the witness a, and its second projection has type $B(a)$. A term of $\prod_{a:A} B(a)$ is a rule that gives a term of $B(a)$ for each input a. The dependent introduction and application are the abstraction and evaluation forms of Chapter 27.

In **Set**, let $E \to X$ be the total space of a family. Factor the map through its image $I \hookrightarrow X$.



The fiber I_x is a singleton precisely when E_x has an element, and otherwise is empty. Write $\|E\| \to X$ for this subterminal family, the *propositional truncation* of the family in sets. Every map from E to a subterminal family T over X factors uniquely through e. Indeed, such a map can exist only when T_x has an element for each $x \in I$, and its value there has a unique choice. This is the adjunction from families to subterminal families given by image and inclusion. Chapter 3 of the *Homotopy Type Theory* book [21] treats propositional truncation in type theory.

For $f\colon X \to Y$ and a predicate $P \hookrightarrow X$, regard P as a family with empty or singleton fibers. The fiber of $\Sigma_f P$ at y records the points $x \in P$ with $f(x) = y$. The truncation $\|\Sigma_f P\|$ retains just whether that fiber is inhabited. It is $\exists_f P$ from Chapter 24. The family $\Pi_f P$ is already subterminal. Its fiber at y has the

unique tuple of singleton elements exactly when every $x \in f^{-1}(y)$ belongs to P. Thus $\Pi_f P$ represents $\forall_f P$. The uniqueness of the entries gives this tuple without requiring a choice of representatives.

30.5 A proof that constructs a function

Let $R(a, b)$ be a family of types for $a\colon A$ and $b\colon B$. Suppose

$$h\colon \prod_{a:A} \sum_{b:B} R(a, b).$$

The following terms build a function and its certificate from h.

$$f_h = \lambda a.\, \pi_1(h\, a), \qquad r_h = \lambda a.\, \pi_2(h\, a).$$

Here $f_h\colon A \to B$, and substitution into the type of the second projection gives $r_h\colon \prod_{a:A} R(a, f_h(a))$. The term $\lambda h.\, \langle f_h, r_h \rangle$ therefore has type

$$\left(\prod_{a:A} \sum_{b:B} R(a, b) \right) \longrightarrow \sum_{f:A \to B} \prod_{a:A} R(a, f(a)).$$

From a pair $\langle f, r \rangle$ on the right the function $\lambda a.\, \langle f(a), r(a) \rangle$ recovers a term on the left. The two functions are inverse in the family interpretation in **Set**. In the first composite, evaluating at a rebuilds $h(a)$ from its two projections. In the other composite the separate functions are recovered because set functions are determined by their values. Every witness used by this construction was supplied by evaluating h.

Truncating each inner sum gives the input type

$$\prod_{a:A} \left\| \sum_{b:B} R(a, b) \right\|.$$

The projected values $f_h(a)$ are then unavailable. The group-action example of Chapter 24 illustrates this loss of witness data. The regular two-element action has full image in the terminal object, and a section would have to supply a fixed point. The object of witnesses has no such point.

Together with the inductive types of Chapter 28, proof terms supply programs that build and analyze structured data. The next part studies how substitutions, reductions, and interpretations of these terms become operations on programs.

Part VII Computation and programming languages

CHAPTER 31

Program syntax and compiler correctness

A compiler translates a program into another language. To state what the translation preserves, we need to specify how each language describes a computation. An arithmetic expression denotes an integer once its variables have values. A stack program describes a function on stacks. The compiler below connects these descriptions by putting the integer denoted by an expression on the stack. Its correctness law specifies both the new entry and the fate of every entry already there.

31.1 The initial algebra of expressions

Fix a set V of variables. The constants are the integers, and the set of binary operation symbols is $O = \{\mathsf{add}, \mathsf{sub}, \mathsf{mul}\}$. Each symbol o has an interpretation $\overline{o}\colon \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$. These are addition, subtraction, and multiplication, in that order. In particular $\overline{\mathsf{sub}}(a, b) = a - b$, so the positions of the arguments matter.

An expression is a finite rooted tree. Its leaves are labelled by constants or variables, and each internal vertex is labelled by an operation and has two ordered children. Use the notation $\underline{z}$ for the leaf of constant z, v for the leaf of variable v, and $o(a, b)$ for the tree with ordered subtrees a, b. These constructors give the structure map ι on the set E_V of expressions for the polynomial functor

$$F_V(X) = \mathbb{Z} \amalg V \amalg (O \times X \times X).$$

On a function f, this functor keeps the two leaf summands unchanged and sends (o, x, y) to $(o, f(x), f(y))$.

An algebra $\alpha\colon F_V(X) \to X$ specifies a value in X for each constant and each variable, and one binary operation on X for each o. Define $\mathrm{fold}_\alpha\colon E_V \to X$ by using these prescribed values at leaves and applying the prescribed operation to the results at the children. Induction on the height of the tree proves that this defines a function and that any function preserving the constructors agrees with it. Thus (E_V, ι) is initial by the tree construction of Chapter 28. The diagram

gives its algebra morphism equation.

$$\begin{array}{ccc} F_V(E_V) & \xrightarrow{\iota} & E_V \\ \downarrow{\scriptstyle F_V(\mathrm{fold}_\alpha)} & & \downarrow{\scriptstyle \mathrm{fold}_\alpha} \\ F_V(X) & \xrightarrow{\alpha} & X \end{array}$$

31.2 Evaluation and substitution

Let $\mathrm{Env}_V = \mathbb{Z}^V$ be the set of environments. An environment ρ supplies an integer $\rho(v)$ for every variable. Interpret expressions in the algebra of functions $D_V = \mathrm{Hom}_{\mathbf{Set}}(\mathrm{Env}_V, \mathbb{Z})$. A constant is the constant function, and a variable is the corresponding projection. The operation o combines $d, d' \in D_V$ pointwise by $\rho \mapsto \overline{o}(d(\rho), d'(\rho))$. The unique fold into this algebra is the evaluation map $\mathrm{eval}_V \colon E_V \to D_V$. Writing $[\![e]\!]_\rho = \mathrm{eval}_V(e)(\rho)$ gives the familiar recursion equations

$$[\![\underline{z}]\!]_\rho = z, \qquad [\![v]\!]_\rho = \rho(v), \qquad [\![o(a,b)]\!]_\rho = \overline{o}([\![a]\!]_\rho, [\![b]\!]_\rho).$$

This map is compositional because the value at each constructor uses only the values of its immediate inputs.

A substitution $\sigma \colon V \to E_W$ assigns an expression in a new set W of variables to each variable of V. Give E_W an F_V-algebra structure by interpreting a variable leaf v as $\sigma(v)$, keeping the constants and operation constructors of E_W. The induced fold $\widehat{\sigma} \colon E_V \to E_W$ substitutes all the assigned expressions at once. On environments the corresponding map goes in the other direction.

$$h_\sigma \colon \mathrm{Env}_W \longrightarrow \mathrm{Env}_V, \qquad h_\sigma(\rho)(v) = [\![\sigma(v)]\!]_\rho.$$

Use h_σ^* for precomposition with h_σ.

$$\begin{array}{ccc} E_V & \xrightarrow{\widehat{\sigma}} & E_W \\ \downarrow{\scriptstyle \mathrm{eval}_V} & & \downarrow{\scriptstyle \mathrm{eval}_W} \\ D_V & \xrightarrow{h_\sigma^*} & D_W \end{array}$$

To see why the square commutes, interpret an F_V-variable leaf in D_W as $\mathrm{eval}_W(\sigma(v))$. Both routes now preserve every constructor. For the bottom route, the value at that leaf is $\rho \mapsto h_\sigma(\rho)(v)$ by precomposition, and the operations are pointwise. Initiality makes the two routes equal. The equation says that evaluating a substituted expression amounts to evaluating the original expression in the environment obtained by evaluating its replacement expressions.

31.3 Stack programs as composable arrows

The objects of a category $\mathbf{Code}_V$ are stack heights $n \in \mathbb{N}$. Its arrows are paths of primitive instructions, whose input and output heights are given below. Composition is concatenation of paths in execution order, and an empty path is an identity. Write $p \cdot q$ for executing p and then q when their heights match.

instruction	height change	action
$\mathsf{push}(z)$	$n \to n+1$	append the integer z
$\mathsf{load}(v)$	$n \to n+1$	append the integer $\rho(v)$
$\mathsf{bin}(o)$	$n+2 \to n+1$	replace the top pair (a, b) by $\overline{o}(a, b)$

Here the top of a stack is its right end. Thus b in the last instruction is popped first, and a is the left operand. Each primitive with the indicated heights is defined for every stack of its input height.

Fixing ρ, write $S_n = \mathbb{Z}^n$ for the set of stacks of height n. Denote the result of appending entries to s by $(s, a_1, \ldots, a_j)$. The last primitive sends (s, a, b) to $(s, \overline{o}(a, b))$. Empty paths and concatenations are interpreted by identity functions and composition. Associativity of function composition makes this an interpretation functor $\mathrm{run}_\rho \colon \mathbf{Code}_V \to \mathbf{Set}$, with $\mathrm{run}_\rho(n) = S_n$. In particular, every arrow $p \colon n \to m$ interprets as a total function $S_n \to S_m$. An instruction that consumes two entries can occur only at a height with those entries available.

The compiler produces code for any starting height. To express this by a fold, take the algebra carrier

$$C_V = \prod_{n \geq 0} \mathrm{Hom}_{\mathbf{Code}_V}(n, n+1).$$

Interpret constants by families of pushes and variables by families of loads. Given $P, Q \in C_V$, the family for a binary operation has component $P_n \cdot Q_{n+1} \cdot \mathsf{bin}(o)$ at height n. Its types fit as follows.

$$n \xrightarrow{P_n} n+1 \xrightarrow{Q_{n+1}} n+2 \xrightarrow{\mathsf{bin}(o)} n+1$$

The unique fold $E_V \to C_V$ is the compiler. Let $c_n(e)$ denote its n-th component at e. In a binary expression the compiler first emits code for the left subexpression, then code for the right subexpression over the resulting larger stack, and then the binary instruction. The compiler definition uses the same constructor decomposition as evaluation but assigns code to each constructor.

31.4 Correctness with a continuation

A fragment of code often runs as one part of a larger program. The rest of that program is called the continuation. Its input height must match the output

height of the fragment. A theorem quantified over continuations lets us replace a compiled fragment by its semantic action while leaving the rest of the program in place.

Theorem 31.1 (Compiler correctness)**.** *For every expression e, environment ρ, stack $s \in S_n$, and code $k\colon n+1 \to m$, one has*

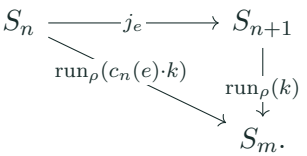
$$\mathrm{run}_\rho\,(c_n(e) \cdot k)(s) = \mathrm{run}_\rho(k)(s, [\![e]\!]_\rho).$$

In particular $c_n(e)$ appends one entry and leaves the original stack intact.

Proof. We prove the equality by induction on e, with input height and continuation arbitrary at every step. For a constant, its push appends the value of that constant. For a variable, its load appends the value from ρ. In either case the continuation then receives exactly the stack on the right of the formula.

Suppose $e = o(a, b)$. Apply the induction hypothesis to a with continuation $c_{n+1}(b) \cdot \mathsf{bin}(o) \cdot k$. It remains to execute that continuation from $(s, [\![a]\!]_\rho)$. Apply the induction hypothesis to b at height $n+1$ with continuation $\mathsf{bin}(o) \cdot k$. The stack presented to $\mathsf{bin}(o)$ is now $(s, [\![a]\!]_\rho, [\![b]\!]_\rho)$. That instruction replaces the last two entries by $\overline{o}([\![a]\!]_\rho, [\![b]\!]_\rho)$, which is $[\![e]\!]_\rho$. The remaining code is k, giving the required equality. Taking k to be the identity yields the assertion about the original stack. □

For fixed ρ and e, let $j_e(s) = (s, [\![e]\!]_\rho)$. The theorem first identifies j_e with $\mathrm{run}_\rho(c_n(e))$. The effect of the continuation is then the commuting triangle



Two expressions have equal denotations exactly when their compiled fragments give equal results for all environments, input stacks, and well-typed continuations. One direction follows from the theorem. For the converse take the empty stack and the identity continuation, then read the single output entry. This comparison uses the final stack as the observable result in the machine just defined.

The proof also admits a fold formulation. Consider the carrier of families of functions

$$M_V = \prod_{n \geq 0} \mathrm{Hom}_{\mathbf{Set}}(\mathrm{Env}_V \times S_n, S_{n+1}).$$

Give it the algebra of pushes, loads, and sequential composition followed by the binary instruction. Each component of a binary operation sends (ρ, s) through its first family at n, then its second at $n+1$, then the primitive binary function.

The value of ρ stays fixed throughout. Interpreting the code defines an algebra map $C_V \to M_V$ by the functor law for execution. The map $D_V \to M_V$ sending d to the family $(\rho, s) \mapsto (s, d(\rho))$ is an algebra morphism as well. The leaf cases use the definitions of push and load. For a binary operation the two appended values are precisely the two arguments of the primitive at the end. Initiality therefore identifies the two composites from E_V.

$$\begin{array}{ccc} E_V & \xrightarrow{\text{compile}} & C_V \\ \downarrow{\scriptstyle \text{eval}_V} & & \downarrow{\scriptstyle \text{run}} \\ D_V & \xrightarrow{\text{append}} & M_V \end{array}$$

The square identifies the execution of the compiled fragment with its semantic action. Composition with the remaining code yields the full theorem.

31.5 An execution trace

Example 31.2 (An order-sensitive trace)**.** Compile $e = \mathsf{mul}(\mathsf{sub}(x, \underline{3}), \mathsf{sub}(\underline{2}, y))$. Take $\rho(x) = 10$ and $\rho(y) = 6$. The left subtraction gives 7 and the right gives -4, so the expected value is -28. Begin with the stack $(8, 13)$. As continuation take one subtraction instruction that acts on the result and the entry beneath it.

instruction just executed	current stack
begin	$(8, 13)$
$\mathsf{load}(x)$	$(8, 13, 10)$
$\mathsf{push}(3)$	$(8, 13, 10, 3)$
$\mathsf{bin}(\mathsf{sub})$	$(8, 13, 7)$
$\mathsf{push}(2)$	$(8, 13, 7, 2)$
$\mathsf{load}(y)$	$(8, 13, 7, 2, 6)$
$\mathsf{bin}(\mathsf{sub})$	$(8, 13, 7, -4)$
$\mathsf{bin}(\mathsf{mul})$	$(8, 13, -28)$
$\mathsf{bin}(\mathsf{sub})$	$(8, 41)$

The horizontal rule marks where the compiled fragment ends. The new entry -28 sits over the original stack as required. The continuation computes $13 - (-28) = 41$, using an entry the compiled fragment preserved.

31.6 Stack height and semantic equality

Let $H(e)$ be the greatest number of entries above the starting height during execution of $c_n(e)$, counting the initial and final stacks. The instruction height rules give

$$H(\underline{z}) = H(v) = 1, \qquad H(o(a, b)) = \max\{H(a),\, 1 + H(b)\}.$$

For a binary expression, executing the first subexpression costs $H(a)$ extra slots. Its result stays on the stack while the second subexpression uses $H(b)$ more slots. The last primitive reduces the height, so it cannot increase this maximum. This proves the formula inductively and shows that it is independent of n and the environment. In the trace each subtraction requires two extra slots relative to its own starting height. The multiplication needs three extra slots in total, as witnessed by the stack of height five.

Example 31.3 (Reassociation and stack use)**.** The expressions

$$e_\ell = \mathsf{add}(\mathsf{add}(x, y), z), \qquad e_r = \mathsf{add}(x, \mathsf{add}(y, z))$$

have the same denotation by associativity of integer addition. The compiler theorem makes them interchangeable in any continuation when the observation is the final stack. Their peak stack heights are different. The inner addition in each expression takes two extra slots. In e_ℓ that addition runs before z is loaded, giving $H(e_\ell) = 2$. In e_r it runs with the value of x still on the stack, giving $H(e_r) = 3$. Thus reassociating this sum to the left reduces the number of stack slots needed by the compiled fragment by one.

The compiler theorem concerns finite expression trees, exact integers, and the given instruction semantics. Winskel [22] develops operational and denotational semantics for larger languages.

31.7 Pontryagin duality and fast Fourier programs

For a positive integer N, the characters of $G = \mathbb{Z}/N\mathbb{Z}$ are indexed by $j \in \mathbb{Z}/N\mathbb{Z}$ and take the complex values $x \mapsto \zeta^{jx}$, where $\zeta = e^{2\pi i/N}$. Use the exact field $K = \mathbb{Q}(\zeta)$ for this example. For $f \in K^G$, the discrete Fourier transform is $\mathcal{F}f(j) = \sum_{x \in G} f(x)\zeta^{-jx}$. The character sum $\sum_{j} \zeta^{j(x-y)}$ is N when $x = y$ and zero otherwise, by the geometric series. Consequently $f(x) = N^{-1} \sum_j \mathcal{F}f(j)\zeta^{jx}$. Define convolution by $(f * g)(x) = \sum_y f(y)g(x - y)$. The following commuting square expresses an interchangeable implementation of convolution:

$$\begin{array}{ccc} K^G \times K^G & \xrightarrow{\ *\ } & K^G \\ \downarrow{\scriptstyle \mathcal{F}\times\mathcal{F}} & & \downarrow{\scriptstyle \mathcal{F}} \\ K^{\widehat{G}} \times K^{\widehat{G}} & \xrightarrow{\ \cdot\ } & K^{\widehat{G}}. \end{array}$$

Indeed, substituting $z = x - y$ into $\mathcal{F}(f * g)(j)$ splits $\zeta^{-j(y+z)}$ into two factors and gives $\mathcal{F}f(j)\,\mathcal{F}g(j)$. An implementation can transform both inputs, multiply pointwise, and apply the inverse.

For $N = 2m$, write E, O for the length-m transforms, using ζ^2, of $r \mapsto f(2r)$ and $r \mapsto f(2r+1)$. Splitting the defining sum by parity gives, for $0 \le j < m$,

$$\mathcal{F}f(j) = E(j) + \zeta^{-j} O(j),$$
$$\mathcal{F}f(j+m) = E(j) - \zeta^{-j} O(j).$$

The second equation uses $\zeta^{-m} = -1$. With the required roots of unity available as constants, for N a power of two the arithmetic operation count satisfies $T(N) = 2T(N/2) + O(N)$ and $T(1) = O(1)$, giving $O(N \log N)$. The inverse uses the opposite sign in the exponent and division by N, so the full convolution implementation has the same arithmetic bound. This count treats operations in K as single steps. The field calculation proves exact semantic equality of the two finite programs. Tao [64] develops the Fourier transform on finite abelian groups. Section 16.6 gives the setting for Fourier analysis beyond finite groups.

Termination of the expression compiler follows from recursion on the syntax tree. Allowing a program to call itself introduces computations for which evaluation can keep unfolding. The next chapter studies how to assign a meaning to such recursion.

CHAPTER 32

Recursion from least fixed points

A loop checks a condition and either returns its current state or continues from a new state. One can approximate this process by allowing a bounded number of steps. A larger bound can supply an answer where a smaller bound supplied none. Once an answer is available, increasing the bound preserves it. We can therefore interpret the loop by an increasing sequence of partial functions. Its limit solves the equation obtained by unfolding the loop once.

32.1 Ordering partial answers

For a set X, put $X_\perp = \{\perp\} \amalg X$, and write $\uparrow x$ for the copy of x in this set. Give $X_\perp$ the order in which $\perp$ lies below every element and the elements $\uparrow x$ are comparable only to themselves. This is the *flat lift* of X. An element of $X_\perp$ says either that no answer has been supplied or that the answer is x. A partial function from a set S to X is encoded by a total function $S \to X_\perp$, using $\perp$ at inputs outside its domain. Order these functions pointwise. In this order, $f \leq g$ exactly when g gives the same answer everywhere f gives one.

A chain $d_0 \leq d_1 \leq \cdots$ in a flat lift either consists entirely of $\perp$ or becomes constant at the first answer it supplies. It has a least upper bound in either case. The least upper bound of a chain of partial functions is obtained by taking these bounds at each input. Thus input s has an answer at the limit exactly when some finite stage answers there.

Definition 32.1. An *ω-complete partial order*, or *ω-CPO*, is a partially ordered set in which every ascending sequence has a least upper bound, written $\bigvee_n d_n$. It is *pointed* if it also has a least element $\perp$. A map of ω-CPOs is *continuous* if it is monotone and preserves these sequence bounds.

The bounds in this definition concern nonempty sequences, so continuity permits a map to send $\perp$ to a defined answer. A map between pointed orders is *strict* if it sends the least element to the least element. A set with the discrete order is an ω-CPO since its ascending sequences are constant. In particular every map from a discrete order to an ω-CPO is continuous.

Composition of continuous maps preserves sequence bounds and the identities do as well. Products in this category are ordered coordinatewise, and their sequence bounds are computed coordinatewise. The projections are continuous,

and the pairing of two continuous maps is continuous by the same rule for bounds. The singleton is terminal. In particular, a product of pointed orders is pointed with the tuple of least elements at its bottom.

32.2 A domain of functions

For ω-CPOs A, D, let $[A, D]$ denote the set of continuous maps $A \to D$ with its pointwise order. If $f_0 \leq f_1 \leq \cdots$, put $f(a) = \bigvee_n f_n(a)$. This map is monotone. To prove its continuity, use the following observation about a family $b_{n,m}$ increasing in each index. Every $b_{n,m}$ is bounded by $b_{k,k}$ for $k = \max(n, m)$. Thus both iterated sequence bounds equal the bound of the diagonal sequence $b_{k,k}$. Applying this to $b_{n,m} = f_n(a_m)$ for an ascending sequence a_m gives

$$\begin{aligned} f\left(\bigvee_m a_m\right) &= \bigvee_n \bigvee_m f_n(a_m) \\ &= \bigvee_m \bigvee_n f_n(a_m) = \bigvee_m f(a_m). \end{aligned}$$

So f belongs to $[A, D]$ and is the least upper bound of the original sequence. If D is pointed, the constant map with value $\bot$ is bottom in this function space.

Evaluation $\mathrm{ev} \colon [A, D] \times A \to D$ is continuous jointly in both arguments. It is monotone by the pointwise order and monotonicity of each map. For a sequence (f_n, a_n) in its domain, the bound of the evaluations $f_n(a_n)$ equals the iterated bound of $f_n(a_m)$ by the diagonal argument. The iterated bound is the evaluation of the two coordinate bounds.

For a continuous map $g \colon P \times A \to D$, the function $\widehat{g}(p)(a) = g(p, a)$ has continuous values in $[A, D]$ because the map $a \mapsto (p, a)$ is continuous for fixed p. The map $p \mapsto \widehat{g}(p)$ is continuous by computing its bounds at every a. Conversely, evaluation recovers g from $\widehat{g}$, and this factorization forces every value $\widehat{g}(p)(a)$. We obtain the exponential universal property, now in a category where functions can return partial answers.

$$\begin{array}{ccc} P \times A & \xrightarrow{\widehat{g} \times \mathrm{id}_A} & [A, D] \times A \\ & \searrow^{g} & \downarrow{\scriptstyle \mathrm{ev}} \\ & & D \end{array}$$

When $A = S$ has the discrete order and $D = X_\bot$, this function space is exactly the order of partial functions described above.

32.3 Taking the least solution

A recursive specification can describe how to use a candidate denotation to compute another candidate. This is an endomap $F \colon D \to D$, where D contains

the possible partial denotations. A solution is a fixed point of F. Iteration from the least element lets us build the solution by taking the bound of the finite stages. In a function space with flat results, each defined answer of this solution must appear at some finite stage.

Theorem 32.2 (Kleene fixed-point theorem)**.** *Let D be a pointed ω-CPO and $F\colon D \to D$ a continuous map. With $d_0 = \bot$ and $d_{n+1} = F(d_n)$, the element $\mathrm{fix}(F) = \bigvee_n d_n$ is a fixed point. Moreover $\mathrm{fix}(F) \leq e$ whenever $F(e) \leq e$.*

Proof. The inequality $d_0 \leq d_1$ holds by leastness of $\bot$. Monotonicity then gives $d_n \leq d_{n+1}$ at each step. For the bound $d = \bigvee_n d_n$, continuity gives

$$F(d) = \bigvee_n F(d_n) = \bigvee_n d_{n+1} = \bigvee_n d_n = d.$$

Deleting the first stage leaves the same upper bounds because $d_0 \leq d_1$. Now suppose $F(e) \leq e$. Since $d_0 \leq e$ and $d_n \leq e$ implies $F(d_n) \leq F(e) \leq e$, induction gives $d_n \leq e$ for all n. Taking the bound gives $d \leq e$. □

In particular the fixed point constructed in the theorem is the least fixed point. Regard D as a category with an arrow for each inequality. An F-algebra is then an arrow $F(e) \to e$. Its morphisms are the inequalities between such e, since any parallel composites in an order agree. The leastness conclusion therefore makes $\mathrm{fix}(F)$ the initial F-algebra in this order, in the sense of Chapter 28. The sequence bound in the construction is the colimit of its ascending chain.

For example, the equation $f = f$ holds for every partial function, and its least solution gives no answer at any input. Thus a fixed-point equation can hold for functions that supply more information than iteration justifies.

Example 32.3 (A discontinuous step)**.** Consider the chain $0 < 1 < 2 < \cdots < \omega < \omega + 1$. It is a pointed ω-CPO. An ascending sequence of naturals which is unbounded in the naturals has bound ω, and a sequence reaching either of the two extra elements stabilizes there or at the greater one. Take $F(n) = n + 1$ on naturals, $F(\omega) = \omega + 1$ and $F(\omega + 1) = \omega + 1$. This step is monotone. Its least fixed point is $\omega + 1$, while the bound of its finite iterates at 0 is ω. At that bound the step gives $\omega + 1$, even though every value on the finite stages lies below ω. Continuity is the hypothesis that rules out this extra step at a sequence bound.

Winskel [22] develops this construction for the denotational semantics of programming languages.

32.4 Recursion in a context

Free variables in a recursive definition give parameters. Suppose $F\colon P \times D \to D$ is continuous, with D pointed. Write $F_p(d) = F(p, d)$. Each F_p is continuous by

fixing the first argument. The parameterized fixed point is

$$F^{\dagger}(p) = \mathrm{fix}(F_p).$$

To prove $F^{\dagger}$ is continuous, perform the approximation inside $[P, D]$. The map u_0 is constantly $\bot$, and $u_{n+1} = F \circ \langle \mathrm{id}_P, u_n \rangle$. Each stage is continuous by pairing and composition. These maps are increasing pointwise, and their pointwise bound is $F^{\dagger}$. The function space construction shows that this bound is continuous. Applying the result to the evaluation map $[D, D] \times D \to D$ also makes $F \mapsto \mathrm{fix}(F)$ a continuous map $[D, D] \to D$.

For a change of context $\sigma \colon Q \to P$, the new step is $F_\sigma = F \circ (\sigma \times \mathrm{id}_D)$. Its approximations are $u_n \circ \sigma$ by induction at each stage, so its fixed point is $F^{\dagger} \circ \sigma$. In the diagram the dashed arrow is obtained by solving the substituted equation.

$$\begin{array}{ccc} Q & \xrightarrow{\sigma} & P \\ & \searrow^{F^{\dagger}_{\sigma}} & \downarrow{\scriptstyle F^{\dagger}} \\ & & D \end{array}$$

This is the substitution law for a least fixed point in a context, using the same direction of substitution as in Chapter 27.

Proposition 32.4 (Eliminating a simultaneous recursion)**.** *Let S, T be pointed ω-CPOs and let $f \colon S \times T \to S$, $g \colon S \times T \to T$ be continuous. Define $h(s) = \mathrm{fix}(g(s, -))$ and $m(s) = f(s, h(s))$. The least fixed point of $\langle f, g \rangle$ is $(s_*, h(s_*))$, where $s_* = \mathrm{fix}(m)$.*

Proof. The parameter construction makes h continuous, and m is continuous by pairing. The two displayed coordinates satisfy their respective fixed-point equations. If (s, t) is any pair with $f(s, t) \leq s$ and $g(s, t) \leq t$, then $h(s) \leq t$ by leastness for $g(s, -)$. Monotonicity gives $m(s) \leq f(s, t) \leq s$, so $s_* \leq s$. Also $h(s_*) \leq h(s) \leq t$. This proves leastness in the product order. □

32.5 Moving a map through recursion

Let $F \colon D \to D$ and $G \colon E \to E$ be continuous endomaps of pointed orders. Suppose $h \colon D \to E$ is continuous and strict and the square commutes.

$$\begin{array}{ccc} D & \xrightarrow{F} & D \\ {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle h} \\ E & \xrightarrow{G} & E \end{array}$$

Strictness gives $h(\bot_D) = \bot_E$. From this equality and the square, $h(F^n(\bot_D)) = G^n(\bot_E)$ for each n. Continuity then gives the *fixed-point fusion law*

$$h(\mathrm{fix}(F)) = \mathrm{fix}(G).$$

This resembles the fold fusion of Section 28.2. Here the preservation of the starting stage is essential. For instance, take the two-element chain $\bot < \top$, let both endomaps be identities, and let h be constantly $\top$. The square still commutes, while h takes the least fixed point to $\top$.

32.6 A loop and its finite runs

Let S be a set of states, $b\colon S \to \{0,1\}$ a total guard and $t\colon S \to S$ a total one-step update. While $b(s) = 1$, the loop moves to $t(s)$. When $b(s) = 0$, it stops with state s. Its unfolding step on the partial function domain $D = [S, S_\bot]$, with S discrete, is

$$F(u)(s) = \begin{cases} \uparrow s & b(s) = 0, \\ u(t(s)) & b(s) = 1. \end{cases}$$

The step is monotone because the first branch has a fixed result and the second branch evaluates at a fixed state. For an ascending sequence of candidates, the bound is computed at each state. On the first branch the bound of the answers is again $\uparrow s$. On the second it is the bound of the candidates at $t(s)$. So F is continuous.

Put $u_n = F^n(\bot)$, where $\bot$ is the nowhere-defined partial function. A run starting at s has state $t^k(s)$ after k updates, as long as the guard has stayed true before each update. For each n, $u_n(s)$ has a defined result exactly when the first false guard occurs after $k < n$ updates. In that case it returns $\uparrow t^k(s)$. For $n = 0$ the bound allows no guard checks and the function is nowhere defined. For the passage to $n + 1$, if the first guard is false we have $k = 0$ and $F(u_n)(s) = \uparrow s$. If the guard is true, there must be one update followed by a run from $t(s)$. Its first false guard occurs within the bound for $u_n(t(s))$ precisely when the original time k is less than $n + 1$. The returned state is unchanged by this reindexing of the run. This proves the claim by induction.

Flatness makes each defined limit value occur at a finite stage. The least fixed

point therefore returns exactly the final states of terminating runs.

Example 32.5 (Decrementing an integer counter)**.** Take $S = \mathbb{Z} \times \mathbb{Z}$, with counter k and accumulator a. The guard asks whether $k \neq 0$, and the step is $t(k, a) = (k - 1, a + k)$. From the state $(3, 0)$ the run is

$$(3,0) \xrightarrow{t} (2,3) \xrightarrow{t} (1,5) \xrightarrow{t} (0,6)$$

The final state is first recorded by u_4, since the stop test follows the third update. More generally a nonnegative counter k takes k steps, adding $k, k-1, \ldots, 1$ to the accumulator. Its returned state is $(0, a + k(k+1)/2)$, first recorded by u_{k+1}. If $k < 0$, every successive counter is still negative. Each u_n consequently gives $\perp$ at such an input, and the least solution does too. The fixed-point equation would also admit assigning the same defined answer, such as $\uparrow (0, 0)$, to every negative-counter state while keeping the usual answers elsewhere. The least solution excludes that assignment because no finite run from a negative counter reaches a stop test with result 0.

The order on partial functions allows the least fixed point to capture termination as part of a program's meaning. The following chapter treats partiality as an effect and studies how partial answers and other effects compose.

CHAPTER 33

Computational effects from monads

A computation can return a value, report an exception, read and update state, or sample an outcome. Each of these possibilities changes the type of the result. For a particular choice of effect, write TA for the type of computations whose ordinary results have type A. Returning a value requires a map $A \to TA$. Sequencing requires a way to run a computation which produces another computation. These maps and the equations they satisfy give a category of effectful programs.

33.1 Unit and multiplication from an adjunction

Let $F\colon \mathcal{C} \to \mathcal{D}$ be left adjoint to $U\colon \mathcal{D} \to \mathcal{C}$, with unit η and counit ε as in Chapter 5. Set $T = UF$ and use the same unit $\eta\colon \mathrm{id}_{\mathcal{C}} \Rightarrow T$. The counit reduces two successive applications of T to one by

$$\mu_A = U(\varepsilon_{FA})\colon T^2A = UFUFA \longrightarrow UFA = TA.$$

These maps are natural by the naturality of the unit and counit. The triangle identities give $\mu_A \eta_{TA} = \mathrm{id}_{TA}$ and $\mu_A T\eta_A = \mathrm{id}_{TA}$. The first is the U-triangle at FA, and the second comes from applying U to the F-triangle at A.

$$\begin{array}{ccccc} TA & \xrightarrow{\eta_{TA}} & T^2A & \xleftarrow{T\eta_A} & TA \\ & \searrow^{\mathrm{id}_{TA}} & \downarrow{\scriptstyle \mu_A} & \swarrow_{\mathrm{id}_{TA}} & \\ & & TA & & \end{array}$$

Naturality of the counit at $\varepsilon_{FA}\colon FUFA \to FA$ gives

$$\varepsilon_{FA}\varepsilon_{FUFA} = \varepsilon_{FA}\, FU(\varepsilon_{FA}).$$

Applying U yields $\mu_A\mu_{TA} = \mu_A T\mu_A$. This identity is the associativity of flattening three layers, as in the commuting square below.

$$\begin{array}{ccc} T^3A & \xrightarrow{\mu_{TA}} & T^2A \\ \downarrow{\scriptstyle T\mu_A} & & \downarrow{\scriptstyle \mu_A} \\ T^2A & \xrightarrow{\mu_A} & TA \end{array}$$

A *monad* is an endofunctor T with natural transformations $\eta\colon \mathrm{id}_{\mathcal{C}} \Rightarrow T$ and $\mu\colon T^2 \Rightarrow T$ satisfying these two unit laws and the associativity law. The adjunction gives one construction of such data. For the free monoid adjunction, TA

is the set of lists of elements of A, η forms a one-entry list, and μ concatenates a list of lists. Chapter 20 uses the same laws for finite distributions, where η puts mass one at a value and μ mixes distributions. Moggi [23] develops the interpretation of programs with effects in monadic terms.

33.2 Sequencing computations

A program with an input of type A and effectful output of type B is represented by $f\colon A \to TB$. Its *Kleisli extension* $f^\sharp\colon TA \to TB$ is $\mu_B \circ Tf$. Given $g\colon B \to TC$, sequencing first f and then g gives $g \odot f = g^\sharp \circ f$.

$$A \xrightarrow{f} TB \xrightarrow{Tg} T^2C \xrightarrow{\mu_C} TC$$

In sets we can write $m \gg\!= f = f^\sharp(m)$ for $m \in TA$, and call this operation *bind*. The value of $\eta_A(a)$ is often written $\mathrm{return}(a)$. The equations of the monad imply three laws for extensions.

$$f^\sharp \eta_A = f, \qquad (\eta_A)^\sharp = \mathrm{id}_{TA}, \qquad (g^\sharp f)^\sharp = g^\sharp f^\sharp.$$

For the first equality use $Tf\,\eta_A = \eta_{TB} f$ and the first unit law. The second is the other unit law. For the third, use associativity of μ followed by naturality at $g\colon B \to TC$.

$$\begin{aligned}(g^\sharp f)^\sharp &= \mu_C\, T\mu_C\, T^2 g\, Tf \\ &= \mu_C \mu_{TC}\, T^2 g\, Tf \\ &= \mu_C\, Tg\, \mu_B\, Tf = g^\sharp f^\sharp.\end{aligned}$$

The identity arrow at A is represented by η_A. The first two laws give the identity equations for $\odot$, and the last law gives associativity. This is the *Kleisli category* $\mathrm{Kl}(T)$, with the same objects as $\mathcal{C}$.

An ordinary arrow $h\colon A \to B$ enters the Kleisli category as $J(h) = \eta_B h$. For ordinary maps $u\colon A' \to A$ and $v\colon B \to B'$, the left unit law and functoriality of T give

$$f \odot J(u) = fu, \qquad J(v) \odot f = Tv\, f.$$

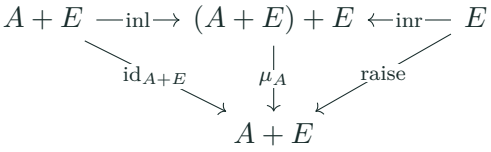
Indeed $(\eta_{B'} v)^\sharp = \mu_{B'} T \eta_{B'} Tv = Tv$. Thus ordinary preprocessing and postprocessing can be inserted on either side of an effectful computation. Taking f itself to be pure shows that J preserves composition, and its identity law is the definition of the Kleisli identity. Naturality of η gives the equivalent formula $\eta_B h = Th\, \eta_A$.

The extension laws also make $V\colon \mathrm{Kl}(T) \to \mathcal{C}$ a functor with $V(A) = TA$ and $V(f) = f^\sharp$. There is an adjunction $J \dashv V$, since its hom-set bijection is the identity on the underlying arrows $A \to TB$. Naturality in A follows from $f \odot J(u) = fu$, and naturality along a Kleisli arrow g out of B is the equation

$g \odot f = V(g)f$. The counit at B is represented by $\mathrm{id}_{TB} \colon TB \to TB$, whose Kleisli extension is μ_B. In this way the Kleisli adjunction gives back the original monad. Jacobs [14, Section 5.2] treats the construction in general.

33.3 Exceptions as extra results

Fix a set E of exception labels. Let $T_E A = A + E$, using disjoint summands for returned values and raised exceptions. Write $\mathrm{ok}(a) = \mathrm{inl}(a)$ and $\mathrm{raise}(e) = \mathrm{inr}(e)$. The category under E has objects maps $i \colon E \to X$ and arrows functions commuting with those maps. The forgetful functor to sets has a left adjoint taking A to $\mathrm{inr} \colon E \to A + E$. Indeed any function $A \to X$ extends to a unique map under E, by using the given i on the other summand. The hom-set bijections are natural by restriction to the A summand. The induced monad is T_E, with unit ok. Its multiplication is the map $(A + E) + E \to A + E$ which is the identity on the first summand and raises the exception in the second summand.



Here $T_E f$ for an ordinary function acts by f on returned values and preserves the labels of exceptions. For $k \colon A \to B + E$ its extension satisfies

$$\mathrm{ok}(a) \gg\!= k = k(a), \qquad \mathrm{raise}(e) \gg\!= k = \mathrm{raise}(e).$$

The second equation is propagation of the first exception in a sequence. For example, a checked subtraction of 7 on $\mathbb{N}$ returns $\mathrm{ok}(n-7)$ when $n \geq 7$ and raises insufficient otherwise. Follow it by checked division of 12 by the result, returning a rational when the denominator is positive and raising division-by-zero at zero. An input 6 raises insufficient, an input 7 raises division-by-zero, and an input 10 returns 4. The first exception takes the raised branch of bind, so the second computation is reached exactly on the successful branch.

33.4 Reading and updating state

Fix a set S of states. A computation returning an A-value takes an incoming state and produces a pair of the value and outgoing state. The result type is $T_S A = (A \times S)^S$. This monad comes from the adjunction $- \times S \dashv (-)^S$ in sets. The unit is the transpose of the identity on $A \times S$, so $\eta_A(a)(s) = (a, s)$. The counit at X is evaluation $X^S \times S \to X$. Consequently μ_A sends $M \colon S \to (T_S A) \times S$ to the computation which at s runs m at s', where $M(s) = (m, s')$.

$$S \xrightarrow{M} (A \times S)^S \times S \xrightarrow{\mathrm{ev}} A \times S$$

For an ordinary map $h\colon A \to B$, $T_S h$ changes the returned value by h and preserves the outgoing state. Applying $T_S k$ for an arrow $k\colon A \to T_S B$ and then multiplying has the following evaluation rule.

$$(m \gg\!= k)(s) = k(a)(s') \quad \text{where } m(s) = (a, s').$$

This rule also checks the bind laws directly. Substituting $m = \eta_A(a)$ makes the right-hand side $k(a)(s)$. Taking $k = \eta_A$ makes it (a, s'), the pair from m. For associativity take $\ell\colon B \to T_S C$ and write $k(a)(s') = (b, s'')$. Evaluating either $(m \gg\!= k) \gg\!= \ell$ or $m \gg\!= (a \mapsto k(a) \gg\!= \ell)$ at s gives $\ell(b)(s'')$. Each stage receives the outgoing state of its predecessor.

The computation $\mathrm{get} \in T_S S$ sends s to (s, s). For $t \in S$, $\mathrm{put}(t) \in T_S 1$ sends s to $(*, t)$. In particular $\mathrm{get} \gg\!= \mathrm{put} = \eta_1(*)$, since the update puts back the state just read. In the other order, $\mathrm{put}(t) \gg\!= (* \mapsto \mathrm{get})$ returns (t, t) for every input state.

Example 33.1 (Reading a counter)**.** Let $S = \mathbb{N}$, with $\mathrm{tick} \in T_S 1$ sending n to $(*, n+1)$. Define two computations of type $T_S \mathbb{N}$ by the following expressions.

$$r = \mathrm{get} \gg\!= (a \mapsto \mathrm{tick} \gg\!= (* \mapsto \eta_{\mathbb{N}}(a))),$$
$$t = \mathrm{tick} \gg\!= (* \mapsto \mathrm{get}).$$

At input n, the first reads n, increments the stored state, and returns $(n, n+1)$. The second increments the state and then reads, returning $(n+1, n+1)$. The two computations have the same outgoing state and different returned values. The monad law permits changing parentheses in the sequence, while these calculations show that interchanging the read and increment can change its meaning.

33.5 Which state is visible on failure?

To model a computation that uses state and can raise an exception, we must specify the result of a failed run. Two choices of result type are

$$T_0 A = \big((A \times S) + E\big)^S, \qquad T_1 A = \big((A + E) \times S\big)^S.$$

In $T_0 A$ the successful branch supplies a value and an outgoing state, and the exception branch supplies an exception label. In $T_1 A$ both branches supply an outgoing state. Each type comes from an adjunction. For a set K, use the category under K, let U_K forget to sets, and let $F_K(X)$ adjoin the summand K as in Section 33.3. Composing the adjunctions gives

$$F_K(- \times S) \;\dashv\; (U_K(-))^S.$$

Indeed a map from $(A \times S) + K$ under K amounts to a function from $A \times S$, then to a function from A by currying. For $K = E$ the induced monad is T_0.

For $K = E \times S$ the induced monad is T_1, written using the natural partition of $(A + E) \times S$ into $A \times S$ and $E \times S$.

In the composite adjunction, the counit evaluates on a pair of a state-dependent computation and a state, and uses the given map on the additional K summand. Thus bind in T_0 runs the continuation at s' on a result $\mathrm{inl}(a, s')$, and propagates the label on $\mathrm{inr}(e)$. In T_1 it runs the continuation at s' on $(\mathrm{ok}(a), s')$, and propagates the whole pair $(\mathrm{raise}(e), s')$ on failure.

For $m \in T_i A$ and $k \colon E \to T_i A$, a handler is a new computation $H_i(m, k) \in T_i A$. We can let it return $m(s)$ on a successful run of m and use these rules on a failure.

$$H_0(m, k)(s) = k(e)(s) \quad \text{if } m(s) = \mathrm{inr}(e),$$
$$H_1(m, k)(s) = k(e)(s') \quad \text{if } m(s) = (\mathrm{raise}(e), s').$$

These handler equations specify an operation in addition to bind.

Example 33.2 (Handling a cancelled increment)**.** Use a counter again, and let $E = \{\mathrm{cancelled}\}$. Run an increment, then discard its unit result and raise cancelled. In T_0 this composite run gives $\mathrm{inr}(\mathrm{cancelled})$ at input n. In T_1 the same sequence gives $(\mathrm{raise}(\mathrm{cancelled}), n + 1)$. Choose $k(\mathrm{cancelled})$ to read the current state. In the first semantics the handled run returns n and leaves state n. In the second it returns $n + 1$ and leaves state $n + 1$. Thus the same source operations give different observations under the two sequencing rules.

CHAPTER 34

Automata and streams from coalgebras

A state may output a number and move to a new state, where the same process starts again. Its behavior is a stream. A different machine may accept or reject at each state and let the next input letter select its successor. Its behavior is a language. Coalgebras describe these systems by their observation and transition maps. A final coalgebra collects their behaviors and gives a unique interpretation of each state.

34.1 States and behavior maps

Let $F\colon \mathcal{C} \to \mathcal{C}$ be an endofunctor. An F-coalgebra is an object X with a map $c\colon X \to FX$. A morphism from (X, c) to (Y, d) is a map $h\colon X \to Y$ satisfying $d \circ h = Fh \circ c$.

$$\begin{array}{ccc} X & \xrightarrow{c} & FX \\ {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle Fh} \\ Y & \xrightarrow{d} & FY \end{array}$$

Functoriality makes the identity and composition of the underlying maps into identity and composition of coalgebra morphisms. Write $\mathrm{Coalg}(F)$ for the resulting category. Passing to the opposite category makes these algebras for F^{op} and reverses their morphisms. In particular, the initial algebras of Chapter 28 specify maps *out of* a carrier by recursion. A *final* F-coalgebra (Z, z) specifies a unique morphism $\mathrm{beh}_c\colon X \to Z$ *into* its carrier for each (X, c). Defining a map by this property is called *coiteration*.

The morphisms into Z are compatible with changes of state space. If $h\colon (X, c) \to (Y, d)$ is a coalgebra morphism, then both beh_c and $\mathrm{beh}_d \circ h$ are coalgebra morphisms into Z, so they coincide. This is the coiteration counterpart of the fold fusion law in Section 28.2.

$$\begin{array}{ccc} (X, c) & \xrightarrow{h} & (Y, d) \\ & {\scriptstyle \mathrm{beh}_c}\searrow \quad \downarrow{\scriptstyle \mathrm{beh}_d} & \\ & & (Z, z) \end{array}$$

The dual of Lambek's lemma in Chapter 28 also applies. The structure map $z\colon Z \to FZ$ of a final coalgebra is an isomorphism. Indeed it is a coalgebra morphism into (FZ, Fz). If $j\colon FZ \to Z$ is the unique morphism back, then $j \circ z = \mathrm{id}_Z$ by finality. The morphism equation for j gives $z \circ j = Fj \circ Fz = F(j \circ z) = \mathrm{id}_{FZ}$.

34.2 Streams as a final coalgebra

Fix a set O of possible outputs. A coalgebra for $F(X) = O \times X$ consists of a function $o\colon X \to O$ and a transition $t\colon X \to X$. On a map h the functor acts by $\mathrm{id}_O \times h$, so a morphism preserves the output and commutes with the transition.

Put $Z = O^{\mathbb{N}}$, the set of streams $s\colon \mathbb{N} \to O$. Its map to $O \times Z$ is

$$z(s) = (\mathrm{hd}(s), \mathrm{tl}(s)), \qquad \mathrm{hd}(s) = s(0), \quad \mathrm{tl}(s)(n) = s(n+1).$$

Given $a \in O$ and $s \in Z$, the stream with first entry a and remaining entries s is the unique inverse image of (a, s) under z.

Proposition 34.1. *The stream coalgebra is final for $O \times (-)$.*

Proof. For a state x of an arbitrary coalgebra, define $\mathrm{beh}(x)(n) = o(t^n(x))$, where $t^0 = \mathrm{id}_X$ and $t^{n+1} = t^n \circ t$. At $n = 0$ this gives the required head. The remaining entries give $\mathrm{beh}(t(x))$, so the transition square commutes.

$$\begin{array}{ccc} X & \xrightarrow{\langle o, t\rangle} & O \times X \\ {\scriptstyle \mathrm{beh}}\downarrow & & \downarrow{\scriptstyle \mathrm{id}_O \times \mathrm{beh}} \\ O^{\mathbb{N}} & \xrightarrow{\langle \mathrm{hd}, \mathrm{tl}\rangle} & O \times O^{\mathbb{N}} \end{array}$$

Any morphism into the stream coalgebra must have these entries, because commuting with the tail forces its n-th entry to be the head of the behavior of $t^n(x)$. Thus the morphism is unique. □

Example 34.2 (Generating the Fibonacci stream)**.** Use states $(u, v) \in \mathbb{N}^2$, output u, and transition $(u, v) \mapsto (v, u + v)$. Starting from $(0, 1)$ the first five states are

$$(0,1), \quad (1,1), \quad (1,2), \quad (2,3), \quad (3,5).$$

If $x_n = (u_n, v_n)$ is the state reached after n transitions and $s(n) = u_n$, then $v_n = s(n+1)$. The next state formula gives $s(n+2) = s(n) + s(n+1)$, with $s(0) = 0$ and $s(1) = 1$. The coalgebra determines the infinite stream from a rule on a pair of numbers. Evaluating any requested entry uses only finitely many applications of the rule.

34.3 Languages as a final coalgebra

Let A be an alphabet and put $2 = \{0, 1\}$. A coalgebra for $F(X) = 2 \times X^A$ specifies an output $o\colon X \to 2$ and a transition $\delta_a\colon X \to X$ for each $a \in A$. The functor acts on maps by $F(h)(i, \tau) = (i, h \circ \tau)$, where $\tau\colon A \to X$. With a designated start state this is a deterministic automaton. An output of 1 means acceptance. We allow arbitrary sets of states, and a transition is defined for every state and letter.

Write A^* for the finite words and ϵ for the empty word. Read a word from left to right. The extended transition maps are $\delta_\epsilon = \mathrm{id}_X$ and $\delta_{aw} = \delta_w \circ \delta_a$ for $a \in A$ and $w \in A^*$. Induction on the length of u gives $\delta_{uv} = \delta_v \circ \delta_u$. For each state x, its language is the set of words accepted when read from x.

$$L_x = \{w \in A^* \mid o(\delta_w(x)) = 1\}.$$

The language after an initial letter a is the derivative $D_a L = \{w \in A^* \mid aw \in L\}$. This gives the powerset $\mathcal{P}(A^*)$ a coalgebra structure with output $e(L) = 1$ if $\epsilon \in L$, and $e(L) = 0$ otherwise. The transition at a is D_a.

Proposition 34.3. *The language coalgebra is final for* $2 \times (-)^A$.

Proof. Since $\delta_\epsilon = \mathrm{id}_X$, we have $e(L_x) = o(x)$. The transition satisfies $D_a L_x = L_{\delta_a(x)}$, since reading aw amounts to reading w from $\delta_a(x)$. Together these equations give the behavior morphism.

$$\begin{array}{ccc} X & \xrightarrow{\quad c \quad} & 2 \times X^A \\ \downarrow{\scriptstyle \mathrm{beh}} & & \downarrow{\scriptstyle F(\mathrm{beh})} \\ \mathcal{P}(A^*) & \xrightarrow{\langle e, (D_a)_{a \in A} \rangle} & 2 \times \mathcal{P}(A^*)^A \end{array}$$

For uniqueness, let H be any coalgebra morphism into $\mathcal{P}(A^*)$. It must agree with $x \mapsto L_x$ on membership of ϵ. Membership of aw in $H(x)$ is equivalent to membership of w in $D_a H(x) = H(\delta_a(x))$. Induction on word length shows this is equivalent to membership of w in $L_{\delta_a(x)}$, or of aw in L_x. Testing every word proves equality of the languages. □

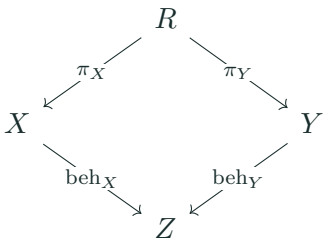
For any word u, put $D_u L = \{w \mid uw \in L\}$. Then $D_v(D_u L) = D_{uv} L$, since $vw \in D_u L$ means $uvw \in L$. In particular, running the language automaton on u from L yields $D_u L$. This identifies the states reached from L as the languages of continuations after finite prefixes of input.

34.4 Coinduction by a relation

Streams are the one-transition case of a coalgebra with an output set O and an indexed family of transitions $(t_a)_{a \in A}$. In this section use $F(X) = O \times X^A$ and suppose it has a final coalgebra. For two systems X, Y of this form, a *bisimulation* is a relation $R \subseteq X \times Y$ such that for each $(x, y) \in R$ the outputs agree, and every successor pair $(t_a^X(x), t_a^Y(y))$ is again in R.

Proposition 34.4. *Two states have the same behavior precisely when some bisimulation relates them.*

Proof. Give R the common output of each pair and transitions given by the successor pairs. The conditions on R ensure that these define a coalgebra. Its two projections are coalgebra morphisms.



The two composites into the final coalgebra Z are equal by finality. Thus every related pair has equal behavior. Conversely, take R to consist of all pairs whose behaviors agree. Their outputs are equal because the behavior maps preserve output. Their successor pairs have equal behavior because the behavior maps preserve each transition. This makes R a bisimulation. □

To prove equality of two given streams by *coinduction*, it suffices to exhibit a relation containing them in which related streams have equal heads and related tails. Their behaviors are the streams themselves, since the sole endomorphism of a final coalgebra is the identity. For the Fibonacci system, another generator uses states $(u, d) \in \mathbb{N}^2$, output u, and transition $(u, d) \mapsto (u + d, u)$. Relate $(u, u + d)$ in the first generator to (u, d) in the second. Their outputs agree. Their successors are $(u + d, 2u + d)$ and $(u + d, u)$, which again have the required form. The pair of start states $(0, 1)$ is related, so their streams are equal. The second state coordinate of one generator tracks the difference between the two coordinates of the other.

34.5 The reachable quotient

Let x_0 be the start state of a deterministic automaton, and put $L = L_{x_0}$. Its reachable states form the set

$$X_{\mathrm{r}} = \{\delta_u(x_0) \mid u \in A^*\}.$$

This is closed under transitions by the concatenation law. The restricted behavior map has image

$$\mathrm{Der}(L) = \{D_u L \mid u \in A^*\}.$$

Indeed $L_{\delta_u(x_0)} = D_u L_{x_0}$. The derivative set contains L and is closed under the letter derivatives, so it is a pointed subcoalgebra of the language coalgebra.

On X_{r}, declare $x \sim y$ if $L_x = L_y$. By the relation proof above, equivalent states have the same output and their successors are equivalent. There are therefore induced outputs and transitions on $X_{\mathrm{r}}/{\sim}$. Factoring the behavior map through this quotient gives a bijection

$$\begin{array}{ccc} X_{\mathrm{r}} & \xrightarrow{\;q\;} & X_{\mathrm{r}}/{\sim} \\ & \searrow^{\mathrm{beh}} & \downarrow \cong \\ & & \mathrm{Der}(L). \end{array}$$

It preserves the outputs, transitions, and start state. Its inverse preserves them by applying the bijection to the defining equations. For any other reachable automaton accepting L, the same factorization yields a surjective morphism to $\mathrm{Der}(L)$. Every state of $\mathrm{Der}(L)$ is reachable from L. Thus if L is accepted by a finite automaton, $\mathrm{Der}(L)$ attains the least possible number of states. A finite automaton attaining that number is reachable, since restricting to the reachable states still accepts L. Its surjective behavior map onto $\mathrm{Der}(L)$ is then a bijection. This proves uniqueness of the minimal finite automaton up to pointed isomorphism. It also shows that a language admits a finite deterministic automaton exactly when its derivative set is finite.

Example 34.5 (Testing a two-letter suffix)**.** Take $A = \{a, b\}$, and let L consist of all words ending in ab. Here are three candidate states and their transition and output table. The start state is q_0.

	a	b	output
q_0	q_1	q_0	0
q_1	q_1	q_2	0
q_2	q_1	q_0	1

After each word, the state is q_1 if the word ends in a, q_2 if it ends in ab, and q_0 in all other cases. This holds at ϵ, and the table preserves the rule when a letter

is appended. The state languages are

$$L_{q_0} = L, \qquad L_{q_1} = L \cup \{b\}, \qquad L_{q_2} = L \cup \{\epsilon\}.$$

For example q_1 is reached after a, and a one-letter continuation completes the suffix precisely when it is b. For continuations of at least two letters the suffix depends only on those letters. At q_2 the empty continuation is the only new accepting continuation, since ab followed by one letter ends in ba or bb. Thus the three states are distinguished by their acceptance of b and ϵ.

Enlarge the machine to record the parity $p \in \{0, 1\}$ of the number of letters read. Its states are (q_i, p), it starts at $(q_0, 0)$, and every transition uses the table and flips parity. The output is still the table output at q_i. All six states occur, as shown by the following prefixes.

	$p = 0$	$p = 1$
q_0	ϵ	b
q_1	aa	a
q_2	ab	aab

The relation that pairs states over the same q_i preserves output and successor pairs. Coinduction identifies the languages across each such pair. The distinguishing continuations above keep the three resulting classes separate, so the quotient recovers the three-state table.

On a finite reachable state set, begin with the pairs that agree on output. To pass from R_n to R_{n+1}, use

$$R_{n+1} = \{(x, y) \in R_n \mid (\delta_a(x), \delta_a(y)) \in R_n \text{ for every } a \in A\}.$$

Induction shows R_n agrees exactly on words of length at most n. A word of length $n+1$ is a first letter followed by a word of length n, which is the additional transition test. For a finite state set the descending chain of relations stabilizes, and the stationary relation is a bisimulation. Every pair with equal language remains at every stage. Thus the stationary relation computes the quotient above. When the alphabet is finite, each step uses finitely many output and transition comparisons.

In the suffix example, the initial partition has one block of four nonaccepting states and one of two accepting states. The letter b splits the first block into the states over q_0 and those over q_1. The resulting three blocks are stable under transitions. Jacobs [14, Chapters 2–3] develops final coalgebras and bisimulation in a general setting.

34.6 Combinatorics of automata and transfer matrices

For finite state and alphabet sets, let P_n be the set of length-n runs, each specified by a start state and a word. The two boundary maps $s, t\colon P_n \to X$ give its start and end states. Splitting a run after m letters gives a pullback, since two runs concatenate uniquely when their adjacent boundary states agree.

$$\begin{array}{ccc} P_{m+n} & \xrightarrow{\text{suffix}_n} & P_n \\ \downarrow{\scriptstyle \text{prefix}_m} & & \downarrow{\scriptstyle s} \\ P_m & \xrightarrow{\quad t \quad} & X \end{array}$$

A span of finite sets $X \xleftarrow{s} P \xrightarrow{t} X$ has a matrix whose entry at (x, y) is the size of the joint fiber over (x, y). Pullback composition becomes matrix multiplication. For fixed outer boundaries x, z, the fiber of a composite is the disjoint union over intermediate states y of the two boundary fibers' products. The zero-length span gives the identity matrix. Thus the matrix of P_n is T^n, where

$$T_{xy} = |\{a \in A \mid \delta_a(x) = y\}|.$$

If e_{x_0} is the row supported at the start state and o is the column of acceptance indicators, then $e_{x_0} T^n o$ counts accepted words. Each word has a unique run from x_0. The ordinary generating function is $e_{x_0}(I - uT)^{-1} o$, because the inverse in formal power series is $\sum_{n \geq 0} u^n T^n$. This is the transfer-matrix method [70, Section 4.7].

Example 34.6 (Binary words avoiding a forbidden block)**.** For words on $\{a, b\}$ with no occurrence of aba, track the longest suffix that is one of ϵ, a, ab. Reading a from ab reaches a rejecting sink, which cannot contribute to an accepted run. The three other states are accepting. In the displayed order their transition matrix and counting series are

$$T = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \qquad e_0(I - uT)^{-1}\mathbf{1} = \frac{1 + u^2}{1 - 2u + u^2 - u^3}.$$

To obtain the series, the three components of $g = (I - uT)^{-1}\mathbf{1}$ satisfy $g = \mathbf{1} + uTg$. Substitute $g_2 = 1 + ug_0$ in $(1 - u)g_1 = 1 + ug_2$, then use $(1 - u)g_0 = 1 + ug_1$. The coefficients start $1, 2, 4, 7, 12, 21$. If c_n denotes the coefficient of u^n, multiplication by the denominator gives $c_n = 2c_{n-1} - c_{n-2} + c_{n-3}$ for $n \geq 3$.

CHAPTER 35

Computability and undecidability

A program can take another program's description as input. Once descriptions are encoded by natural numbers, the same evaluator can run them all. Some runs produce an answer and others run forever. This partiality matters when the evaluator is applied to a program's own description. We derive the halting theorem from that diagonal application, and then use specialization to transfer the obstruction to properties of computed functions.

35.1 Partial maps as computations

A *partial function* $f\colon X \rightharpoonup Y$ consists of a subset $D_f \subseteq X$ and a function $f_D\colon D_f \to Y$. We write $f(x)\downarrow$ for $x \in D_f$ and $f(x)\uparrow$ otherwise. Equality of partial functions requires equality of their domains and agreement there. Equivalently, f is a span

$$X \xleftarrow{i_f} D_f \xrightarrow{f_D} Y$$

with its left arrow an inclusion. This presentation connects definedness to the subobjects of Chapter 21.

To compose f with $g\colon Y \rightharpoonup Z$, restrict D_f to the inputs whose f-values lie in D_g. The following pullback gives the restriction.

$$\begin{array}{ccc} D_{gf} & \xrightarrow{\bar{f}} & D_g \\ \downarrow & \lrcorner & \downarrow{\scriptstyle i_g} \\ D_f & \xrightarrow{f_D} & Y. \end{array}$$

The composite is represented by the inclusion $D_{gf} \subseteq X$ and the map $g_D\bar{f}$ to Z. For three maps, both parenthesizations impose the same successive definedness conditions and give the same value. The identity on a set has full domain, so these operations give a category of partial maps. Another description sends f to the total set map $X \to Y_\perp = Y \amalg \{\perp\}$ with value $\perp$ off D_f. Composition keeps $\perp$ whenever the first stage gives it and otherwise applies the second stage to the ordinary value. This is the Kleisli composition for the monad that adjoins one possible failure value, as in Chapter 33.

A partial function on natural numbers is *computable* if a program returns its value for each input in its domain and never returns for inputs outside it. The

same convention applies to finite tuples of inputs. These maps are closed under composition. Run a program for the first map and pass its output to the second if the first returns. The resulting domain and output are exactly those of the pullback construction. A computable partial map with full domain is called *total computable*. The encoding of $\perp$ as an element of $Y_\perp$ is a set-theoretic description of a partial map. It does not equip the running program with a test that detects its own divergence.

35.2 A universal evaluator

Fix deterministic Turing machines as our model of programs, with finite transition tables and a common convention for representing natural numbers on the input and output tapes. A halting configuration is decoded as a number, using 0 as the default for a malformed output. A finite table can be written as a finite string on a fixed alphabet. For example, one may list all strings by length and then lexicographically within a length. A number coding a string which fails the table's format checks denotes a program that loops on every input. We write $\varphi_e \colon \mathbb{N} \rightharpoonup \mathbb{N}$ for the partial function computed by code e.

The finite simulation of these machines is effective uniformly in their codes. At each stage a configuration is represented by the state, head positions, and finite lists recording the nonblank tape cells. A transition is obtained by reading the cells at the head positions, looking up the corresponding entry in the finite table, and updating the lists, heads, and state. Each operation consists of finite scans, copies, or comparisons of finite strings. These operations are Turing computable and can be repeated for any prescribed number of steps. We treat a halting configuration as absorbing in this simulation so that its output persists at later times.

For $t \in \mathbb{N}$, define a bounded evaluator with a tagged output by

$$E_t(e, x) = \begin{cases} b & \text{if } e \text{ halts on } x \text{ by step } t \text{ with output } b, \\ \perp & \text{otherwise.} \end{cases}$$

The operation $(t, e, x) \mapsto E_t(e, x)$ is total computable with the tags encoded as numbers. Running the same simulation and returning only when it detects a halt gives a single partial computable map

$$U \colon \mathbb{N} \times \mathbb{N} \rightharpoonup \mathbb{N}, \qquad U(e, x) = \varphi_e(x).$$

The equality is an equality of partial functions in x for every e, including malformed codes. This is the universality of U. The machine stops its own simulation exactly when the simulated machine has stopped. Turing's universal machine supplies the historical construction behind this enumeration [24].

In the flat domain $\mathbb{N}_\perp$ of Chapter 32, $\perp$ is below every number and different numbers are incomparable. Each sequence $(E_t(e,x))_{t\in\mathbb{N}}$ is increasing. It either stays at $\perp$ or reaches its final numerical value. Its supremum is therefore the partial value of $U(e,x)$, encoded in $\mathbb{N}_\perp$. A finite simulation supplies a stage of this sequence. For a fixed t, the result $\perp$ says only that the simulation has yet to return by that stage.

Example 35.1 (A partial computation with a known domain)**.** Consider a machine with an input counter n. At $n = 1$ return 7. At $n = 0$ keep the counter at 0 and repeat the same instruction. For $n \geq 2$, subtract 2 and repeat. This counter description can be implemented on a tape storing a unary numeral. Counting counter instructions, input $2k + 1$ returns 7 on instruction $k + 1$, while input $2k$ eventually reaches 0 and continues there. The partial function represented by the machine thus has domain the odd numbers and constant value 7 on that domain. The exact instruction bound refers to the counter description. A tape implementation can spend several moves on one instruction.

35.3 Specializing a program

We use an effective coding $\langle a, x\rangle$ of pairs with effective projections. For example,

$$\langle a, x\rangle = \frac{(a + x)(a + x + 1)}{2} + x$$

enumerates the diagonals $a + x = n$ consecutively. To recover a pair from a code, find the unique diagonal whose consecutive block contains it and read the offset x within that block. Searching for the diagonal terminates because the block endpoints are unbounded and increasing.

Passing a fixed parameter into a program is a construction on its code. Given e, a, form the program which on input x writes the literals e, a, computes $\langle a, x\rangle$, and runs the fixed universal evaluator with that code and input. Its description can be assembled by a total computable operation. The instruction sequence which writes a numeral is generated from its digits, and the pairing and simulation routines are fixed finite tables. Rename states by distinct tags as the pieces are attached, then encode the resulting table. Assembly terminates for every e, a because it operates on finite descriptions. Any divergence arises when the resulting table is run.

Write $s(e, a)$ for the new code. With $w(e, a, x) = (e, \langle a, x\rangle)$ and $v(e, a, x) = (s(e, a), x)$, the following diagram expresses the identity of partial evaluations.

$$\begin{array}{ccc} \mathbb{N}^3 & \xrightarrow{\;v\;} & \mathbb{N}^2 \\ \downarrow{\scriptstyle w} & & \downarrow{\scriptstyle U} \\ \mathbb{N}^2 & \xrightarrow{\;U\;} & \mathbb{N}. \end{array}$$

The dashed arrows are partial. Their equality after the indicated precompositions gives

$$\varphi_{s(e,a)}(x) = \varphi_e(\langle a, x\rangle)$$

with both sides defined together. The operation s is the one-parameter instance of the *s-m-n* theorem. For tuples coded by nested pairing, iterating s specializes the successive parameters. This form of compiler correctness uses the same distinction as Chapter 31. The compiler constructs a description, and a separate equation describes what that description does when run.

There is a relation with currying in Chapter 26. As a total map of sets into $\mathbb{N}_\perp$, U has a transpose $\mathbb{N} \to (\mathbb{N}_\perp)^{\mathbb{N}}$ sending each code to its partial function. The transpose lands in the computable partial functions, and all such functions occur. The specialization operation adds effective control of codes to this account. It produces one index of the residual function after fixing an argument.

35.4 The diagonal obstruction

Theorem 35.2 (Halting is undecidable)**.** *There is no total computable function* $H\colon \mathbb{N}^2 \to \{0,1\}$ *with* $H(e,x) = 1$ *exactly when* $U(e,x)\downarrow$.

Proof. Suppose H were such a function. Build a program d which on input e computes $H(e,e)$. If the answer is 0, it returns 0. If the answer is 1, it loops. The program exists by closure under composition and branching on a computed bit. Let k be any code for it. Its construction gives

$$\varphi_k(k)\downarrow \quad\Longleftrightarrow\quad H(k,k) = 0.$$

The assumed specification of H gives $\varphi_k(k)\downarrow$ if and only if $H(k,k) = 1$. Since $H(k,k)$ is one of these bits, its two requirements are incompatible. □

The categorical form of the construction is a composite through the diagonal followed by a change in behavior. Let $b\colon \{0,1\} \rightharpoonup \mathbb{N}$ be defined only at 0, with $b(0) = 0$. Then $bH\Delta$ is the program's partial function. Writing $k\colon 1 \to \mathbb{N}$ for the point given by its code, we get the diagonal instance in the two paths below.

$$\begin{array}{ccccc} 1 & \xrightarrow{k} & \mathbb{N} & & \\ & \searrow^{(k,k)} & \downarrow{\scriptstyle \Delta} & & \\ & & \mathbb{N}^2 & \xrightarrow{H} \{0,1\} & \overset{b}{\dashrightarrow} \mathbb{N}. \end{array}$$

The composite along the row from the pair has the opposite definedness to what the bit $H(k,k)$ asserts for the same computation. The use of universality is essential. The diagonal composite produces a computation that itself occurs among the indexed programs.

The same diagonal arrangement gives a version of Lawvere's fixed-point argument [74] for a category with finite products. Suppose $u\colon A \times A \to B$ has the property that every arrow $f\colon A \to B$ is represented by some global point $a\colon 1 \to A$, meaning $f = u\langle a!, \mathrm{id}_A\rangle$, with $!\colon A \to 1$. For any arrow $g\colon B \to B$, take $f = gu\Delta_A$ and choose its representative a. Evaluating at a gives

$$g(u\langle a, a\rangle) = fa = u\langle a, a\rangle.$$

Thus g has a fixed global point. Apply this in the category of natural-number powers and total computable maps with $A = B = \mathbb{N}$ and $g(n) = n+1$. There can be no total computable evaluator whose rows represent every total computable function. In our enumeration the universal evaluator is partial, which permits it to represent those functions together with all the partial computable ones.

A set $S \subseteq \mathbb{N}^m$ is *semidecidable* when it is the domain of a partial computable map. The halting set is semidecidable because it is the domain of U. Its complement cannot be semidecidable. If programs semidecide a set and its complement, alternate one step of each on the same input. One of them must eventually return. Which one returns determines membership, giving a total decision procedure. This explains the asymmetry of bounded simulation. A finite return witnesses halting, while a general method providing finite evidence of nonhalting as well would decide the halting set.

35.5 Extensional properties of programs

An *extensional* property of codes depends only on their partial functions. If $\varphi_e = \varphi_{e'}$, the property must give the same answer on e and e'. One such property asks whether the function's range contains 1.

Example 35.3 (Does the computation ever produce 1?)**.** Fix e, x and construct a program which ignores its own input, simulates $U(e, x)$, and returns 1 if that simulation halts. Call its code $r(e, x)$. The specialization procedure gives a total computable map $(e, x) \mapsto r(e, x)$. If $U(e, x)$ returns, the new partial function is the constant 1 function. If $U(e, x)$ diverges, the new function has empty domain. Thus

$$1 \in \operatorname{rng} \varphi_{r(e,x)} \quad \Longleftrightarrow \quad U(e, x) \downarrow .$$

A decision procedure for range membership of 1 in an arbitrary program would decide halting by first forming this code. The property is nevertheless semidecidable. For each t, simulate on all inputs $y \leq t$ for t steps. If any run ever returns 1, one of these finite stages will record it.

Theorem 35.4 (Rice's theorem for partial functions)**.** *Let P be a class of partial computable functions $\mathbb{N} \rightharpoonup \mathbb{N}$. If P contains at least one such function and omits at least one, then $\{e \mid \varphi_e \in P\}$ is undecidable.*

Proof. Let z denote the function with empty domain. If $z \in P$, replace P by its complement among the partial computable functions. A total decision procedure for one class would give a total decision procedure for the other by exchanging the output bits. Thus assume $z \notin P$. Choose an index j with $\varphi_j \in P$, and use j as a fixed constant in the following construction.

Use a fixed program p receiving $\langle\langle e, x\rangle, y\rangle$. It first runs $U(e, x)$ and discards the result when it returns. It then runs $U(j, y)$ and returns that value if defined. Set $q(e, x) = s(p, \langle e, x\rangle)$, using the specialization of Section 35.3. This is a total computable code map. Its behavior is

$$\varphi_{q(e,x)} = \begin{cases} \varphi_j & \text{if } U(e, x)\downarrow, \\ z & \text{if } U(e, x)\uparrow. \end{cases}$$

The first case preserves both the outputs and the domain of φ_j. The second case never reaches the second simulation and therefore returns on no input. Thus $q(e, x)$ has the desired property exactly when $U(e, x)$ halts. If D were a decider for the index set of P, the following triangle would give a computable halting decision.

$$\begin{array}{ccc} \mathbb{N}^2 & \xrightarrow{\;q\;} & \mathbb{N} \\ & \searrow^{D \circ q} & \downarrow{\scriptstyle D} \\ & & \{0, 1\}. \end{array}$$

This contradicts Theorem 35.2. □

Rice's original theorem concerns classes of recursively enumerable sets [25]. Totality is a nontrivial extensional property, as the constant functions and the empty partial function show. A finite table can be checked for a noninitial state with no incoming transitions. Adding such a state preserves the partial function.

35.6 Programs as witnesses

An index e witnesses that a map is executable on specified inputs when $U(e, x)$ returns the required value at each one. Given e, j, compile the program that runs $U(e, x)$ and passes its result to $U(j, -)$. Specialization makes this composition uniform in the codes. Chapter 36 uses these operations on witnesses to interpret objects and predicates with computational evidence.

CHAPTER 36

Realizability and the effective topos

A proof of existence can specify a witness. If the witness depends on a parameter, a program can produce it from evidence for the parameter. Realizability builds this requirement into logic. It assigns a set of numerical realizers to a proposition and requires one program to turn realizers of a premise into realizers of a conclusion. Assemblies give sets an analogous kind of evidence. Their morphisms are functions that admit uniform algorithms for transforming that evidence.

36.1 One program for every instance

Use the numbering of partial computable functions from Chapter 35. Write $e \cdot n \downarrow = m$ if the program with number e, run on input n, halts with output m. This partial application on $\mathbb{N}$ is *Kleene's first partial combinatory algebra*, denoted $\mathcal{K}_1$. The universal evaluator lets one program call another from its code. By parameterization, a computable procedure can also return code containing a supplied numerical parameter. These two properties make the numbering *acceptable*. Use the same computable pairing $\langle -, - \rangle$ with computable projections.

A *realizability predicate* on a set X is an assignment $P : X \to \mathcal{P}(\mathbb{N})$. Members of $P(x)$ are the available realizers at x. Write $P \vdash_X Q$ if there is a single program number e such that for all $x \in X$ and all $a \in P(x)$, the computation $e \cdot a$ halts with an output in $Q(x)$. The same number must work for every x and every permitted input. The identity program proves reflexivity. To compose two entailments, run the first program and pass its output to the second. Their guarantees ensure both runs halt. This gives a preorder, and mutual entailment gives its associated partial order.

For example, let $X = \mathbb{N}$ and $P(n) = \{n\}$. Each $P(n)$ is inhabited, and for each fixed n a constant program takes any input to its member. In the context X, however, the predicate $\top(n) = \mathbb{N}$ does not entail P. The run on input 0 would have to return an element of every $P(n)$. The program has the same input in each instance.

Conjunction uses pairs. Implication uses codes that act on all permitted inputs. Define both operations at every x by

$$(P \wedge Q)(x) = \{\langle a, b \rangle \mid a \in P(x),\ b \in Q(x)\},$$
$$(P \Rightarrow Q)(x) = \{e \mid \text{for all } a \in P(x),\ e \cdot a \downarrow \in Q(x)\}.$$

The notation $e \cdot a \downarrow \in Q(x)$ requires termination and membership of the returned value. Pair projections track the two entailments from $P \wedge Q$. If programs send R to P and Q, their outputs on one input can be paired, so conjunction is the meet. For implication the order comparison is

$$R \wedge P \vdash_X Q \quad \Longleftrightarrow \quad R \vdash_X (P \Rightarrow Q).$$

Given a program k for the first entailment, parameterization computably sends $r \in R(x)$ to the code of $a \mapsto k \cdot \langle r, a \rangle$. For the other direction, given a program for the second entailment, apply it to the first component of a pair and run the resulting code on the second. These are the two constructions of the implication adjunction of Chapter 23 at the level of realizers. Falsity is the empty set of realizers. Disjunction uses a tag 0 or 1 together with a realizer of the selected disjunct. Projections from the tag and the evidence allow a single case program to combine two entailments out of the disjuncts.

36.2 Substitution and quantification

For a function $f : X \to Y$, reindexing sends Q to $f^*Q(x) = Q(f(x))$. A program tracking an entailment over Y still works after substitution. At the level of these set-indexed predicates, quantification along f has formulas

$$(\exists_f P)(y) = \bigcup_{f(x)=y} P(x),$$
$$(\forall_f P)(y) = \{e \mid \text{for all } x \text{ with } f(x) = y \text{ and all } n \in \mathbb{N},\ e \cdot n \downarrow \in P(x)\}.$$

The universal predicate permits every code if the fiber is empty. For a nonempty fiber a realizer is a code whose output on any input realizes every instance in the fiber.

Proposition 36.1. *In the entailment order, $\exists_f \dashv f^* \dashv \forall_f$.*

Proof. A program sends every member of the union into $Q(y)$ precisely when it sends each $P(x)$ into $Q(f(x))$. For the other adjunction, suppose k tracks $f^*Q \vdash_X P$. From an input $q \in Q(y)$, return the code of the program $n \mapsto k \cdot q$. Parameterization produces the code without running k. If there is an x over y, the guarantee on k ensures that this code halts with output in $P(x)$ on every n. If there is no such x, the returned code already meets the vacuous requirement. This tracks $Q \vdash_Y \forall_f P$. Conversely, for any x, a program tracking $Q \vdash_Y \forall_f P$ gives a code from $q \in Q(f(x))$. Running that code on 0 gives a realizer of $P(x)$. □

In a pullback of sets the fibers before and after reindexing have a canonical bijection. Both quantifier formulas depend on the same family of sets of realizers

along that bijection, so they satisfy Beck–Chevalley as in Chapter 24. Pairing distributes over the union defining $\exists_f$, which gives the Frobenius identity with conjunction. This indexed logic is the effective *tripos*. Its generic predicate on $\mathcal{P}(\mathbb{N})$ sends a set of realizers to itself. Every set-indexed predicate is obtained by reindexing it along the function specifying that predicate.

Example 36.2 (Why the universal realizer is a code)**.** Let $K = \{n \mid n \cdot n \text{ halts}\}$ with inclusion $j : K \hookrightarrow \mathbb{N}$. Put $Q(n) = \{n\}$. For $n \in K$, let $P(n)$ be the set of numbers of steps by which $n \cdot n$ has halted. The program that simulates $n \cdot n$ and returns its runtime tracks $j^*Q \vdash_K P$.

Suppose universal quantification used the unmodified intersection of the $P(n)$ over each fiber, with $\mathbb{N}$ over empty fibers. Any tracker from Q to that predicate would halt on every n, and for $n \in K$ would return a bound on the runtime of $n \cdot n$. Simulating for that many steps would decide K, contradicting Chapter 35. In the actual universal quantifier the tracking program returns the code for a simulation. It need not complete that simulation to return its code.

36.3 Assemblies and tracked functions

An *assembly* is a set X with a nonempty set of realizers $E_X(x) \subseteq \mathbb{N}$ for each element x. A morphism $(X, E_X) \to (Y, E_Y)$ is a function $f : X \to Y$ with a *tracker*. This is a program number e such that

$$a \in E_X(x) \quad \Longrightarrow \quad e \cdot a \downarrow \in E_Y(f(x)).$$

One tracker works on every permitted representation of every input. The arrow is the function, with existence of a tracker as a condition. Different trackers for the same function give the same arrow. Identity and composition use the identity program and composition of programs. Write **Asm** for this category.

The singleton with realizer set $\{0\}$ is terminal, since a constant program tracks every map to it. From it there is one arrow to an assembly for each underlying element. To track the arrow, choose a realizer for that element and return it constantly. In particular, a monomorphism is injective on underlying sets by testing the arrows from the terminal object. The converse follows from cancellation of injective functions.

For a set Y, let ∇Y give every element of Y all of $\mathbb{N}$ as its realizers. Every ordinary function $X \to Y$ defines an assembly map into ∇Y, tracked by the constant program with output 0. Thus the underlying-set functor Γ is left adjoint to ∇. The uniformity test depends on the realizer assignments on both sides of a map.

A realizability predicate P on the underlying set of an assembly gives a subassembly on the x for which $P(x)$ is nonempty. Its realizers at x are the pairs

in $E_X(x) \wedge P(x)$. The first projection tracks the inclusion. Conversely, any mono into the assembly is represented by some subset with a realizer assignment on the subset. Extend that assignment by empty sets outside the subset to obtain a predicate P. The original inclusion has a tracker. Pairing its output with its input gives an isomorphism from that subobject to the subassembly with realizers $E_X(x) \wedge P(x)$, with inverse tracked by the second projection. This accounts for the extra evidence of the parameter in comparisons of predicates on an assembly. For ∇X the extra factor $\mathbb{N}$ can be added by pairing with 0 or removed by projection.

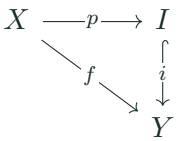
36.4 Images and quotient maps of assemblies

Finite limits have the underlying limits of sets. For a binary product use pairs of realizers, and for an equalizer use the realizers of the input assembly restricted to those elements where the two functions agree. Pairing trackers proves the product property. A map into the input assembly whose image lies in the equalizer already has a tracker for the equalizer with its restricted realizers. Together with the terminal assembly these constructions give finite limits.

For an assembly map $f : X \to Y$, put $I = f(X)$. Give it the following realizers.

$$E_I(y) = \bigcup_{f(x)=y} E_X(x).$$

The first arrow $p : X \to I$ in the image factorization is tracked by the identity program, and the inclusion $i : I \to Y$ is tracked by a tracker for f.



Suppose a map $u : X \to Z$ agrees on pairs with the same p-image. It induces a unique function $I \to Z$. Any tracker for u tracks that function. Indeed each $a \in E_I(y)$ belongs to at least one $E_X(x)$ with $f(x) = y$, and the required output belongs to the same target realizer set for every such x. Thus p is the coequalizer of its kernel pair and in particular a regular epimorphism.

For clarity, an assembly map $f : X \to Y$ is a regular epimorphism exactly when it is surjective and has a single program d satisfying

$$b \in E_Y(y) \quad \Longrightarrow \quad d \cdot b \downarrow \in \bigcup_{f(x)=y} E_X(x).$$

For sufficiency the program tracks an inverse to i. For necessity, if f is the coequalizer of a pair, the first factor p also coequalizes that pair because i

is monic. The induced map $Y \to I$ is inverse to i by cancellation with the epimorphism f and the monomorphism i. This gives the desired surjectivity and program.

The test is stable under pullback along any assembly map $g : Z \to Y$ with tracker k. In the pullback, a pair (x, z) with $f(x) = g(z)$ has the paired realizers from the product. Given a realizer c of z, run k on c and then d on the result. The output a realizes some x over $g(z)$, so $\langle a, c\rangle$ realizes a point over z.

$$\begin{array}{ccc} X \times_Y Z & \longrightarrow & X \\ \downarrow{\scriptstyle f'} & \lrcorner & \downarrow{\scriptstyle f} \\ Z & \xrightarrow{\;g\;} & Y \end{array}$$

Thus f' is also a regular epimorphism. Finite limits and these pullback-stable regular-epimorphism–monomorphism factorizations make **Asm** a regular category.

36.5 Adjoining quotients

An internal equivalence relation on X is a subobject $R \hookrightarrow X \times X$ equipped with the reflexivity, symmetry, and transitivity maps. An equivalence relation is *effective* when it is the kernel pair of its quotient. In a regular category every kernel pair has a quotient given by the regular epimorphism in the image factorization. The requirement that all internal equivalence relations be effective is the additional property of an *exact category*.

Example 36.3 (An equivalence relation with informative evidence)**.** Let X have two elements $0, 1$, each realized by 0. Take R on the full set of underlying pairs of elements of X, with realizers

$$E_R(i, j) = \{i \oplus j\},$$

where $\oplus$ is addition modulo two. The inclusion $R \hookrightarrow X \times X$ is tracked by the constant program returning $\langle 0, 0\rangle$. Reflexivity $i \mapsto (i, i)$ is tracked by the constant 0. Symmetry keeps the same evidence. For transitivity, the input for a composable pair $((i, j), (j, k))$ contains $(i \oplus j)$ and $(j \oplus k)$ as its two components. Their sum modulo two realizes (i, k). This gives all three maps of an internal equivalence relation.

Any map $q : X \to Y$ equating the projections of R is constant on the underlying set. Its kernel pair is therefore $X \times X$ with the product realizers. Every underlying pair there has the common realizer $\langle 0, 0\rangle$. An inverse to the inclusion of R over $X \times X$ would have to act as the identity on pairs and transform this one input into 0 for equal pairs and 1 for unequal pairs. Such a tracker cannot exist. The relation is therefore not the kernel pair of any assembly map.

The *exact completion* of a regular category adjoins effective quotients of its internal equivalence relations. In its relational presentation an object is represented by (X, R). A map to (Y, S) is a relation F from X to Y saturated with respect to R, S, total on R-classes, and single-valued up to S. In relation notation these conditions are

$$S \circ F \circ R = F, \qquad R \leq F^{\circ} \circ F, \qquad F \circ F^{\circ} \leq S,$$

where F° is the reversed relation. The composite of relations is formed by pulling back over their common object and taking the image in the product of the outer objects. The identity of (X, R) is R. An ordinary object X enters by its diagonal relation. Between diagonal objects the displayed conditions say that $F \to X$ is regular epic and monic, so F is the graph of a unique map $X \to Y$. This is how the completion keeps the original maps between embedded objects.

For **Asm** on $\mathcal{K}_1$, the exact completion as a regular category is the *effective topos* **Eff**. The realizability theorem identifies it with the topos obtained from the effective tripos in Section 36.2 [26, 27]. In particular it is an elementary topos, and the diagonal embedding of assemblies is fully faithful, preserves finite limits, and preserves regular epimorphisms. These properties allow the two descriptions to be used together. The example gives a new object represented by (X, R). In **Eff** the embedded relation has the following quotient.

$$R \underset{r_2}{\overset{r_1}{\rightrightarrows}} X \xrightarrow{q} (X, R).$$

If this quotient were isomorphic to an embedded assembly, full faithfulness would make its quotient map an assembly arrow. Preservation of finite limits would then make R its kernel pair in **Asm**, contrary to the example.

36.6 The computable maps on natural numbers

The assembly N has underlying set $\mathbb{N}$ and $E_N(n) = \{n\}$. Zero and successor are assembly maps. For an assembly X with a point $z : 1 \to X$ and an endomorphism $s : X \to X$, the number n must be sent to $s^n(z(*))$ by the recursion equations. Choose a realizer of $z(*)$ and a tracker of s. On input n, begin with the chosen realizer and run the tracker n times, passing the latest output to the next run. Induction proves that every run halts and that the final output realizes $s^n(z(*))$. This yields the unique recursive map out of N in **Asm**. In **Eff** the embedded N, with these same two maps, is again a natural numbers object by the realizability theorem [27, Proposition 3.1.1].

An assembly morphism $N \to N$ has a tracker that on the sole realizer n must halt with the sole output realizer $f(n)$. Conversely, an index of a total computable function tracks that function as an assembly map. Full faithfulness

now gives

$$\mathrm{Hom}_{\mathbf{Eff}}(N, N) \cong \{f : \mathbb{N} \to \mathbb{N} \mid f \text{ is total computable}\}.$$

Each global element $1 \to N$ is a numeral, because the embedding preserves the terminal object and all assembly maps from it. The assertion about endomorphisms thus describes the effect of those arrows on their numerical inputs.

For a concrete contrast, let $h(n)$ be 1 if $n \cdot n$ halts and 0 otherwise. Each individual $h(n)$ gives a map $1 \to N$. A map inducing this assignment on points would decide the diagonal halting set. Equivalently, in context N the graph predicate with witness set $\{h(n)\}$ at n admits a witness for each external number. A program taking the available parameter realizer n to its witness would compute h. Even after furnishing the realizer of the parameter, uniform realizability imposes a stronger test than pointwise existence in **Set**.

This also tests whether countable families of points can be assembled freely. If the coproduct $C = \coprod_{n \in \mathbb{N}} 1$ existed, assign zero to its zeroth injection and let successor shift the injections by one. For any z and s as in the recursion test, the maps $s^n z$ would define the required arrow from C. The coproduct injections verify its equations and uniqueness, so C would be a natural numbers object. Its isomorphism with N would take the injections to the numerals. Applying the coproduct property to the points $h(n)$ would then produce the impossible endomorphism above. Thus **Eff** has a natural numbers object even though it lacks that countable coproduct.

Part VIII Machine learning and statistics

CHAPTER 37

Data, predictors, and empirical risk

A supervised sample pairs inputs with observed labels. A predictor uses the input to produce an action, and a loss compares the action with the label. Averaging these losses gives a risk. The composition of these maps explains how the risk depends on the sample.

37.1 A sample and its empirical law

Let X be the set of inputs, Y the set of labels, and $Z = X \times Y$ the set of records. A sample is a map $z\colon I \to Z$ from a finite nonempty index set I. Write $z_i = (x_i, y_i)$ for its value at i. The product property describes z as the pair of input and label families. Since the map need not be injective, identical records can occur at several indices.

Giving I its uniform distribution u_I, the sample defines a state on Z by pushforward. In the notation of Chapter 20, it is

$$\widehat{p}_z = \mathcal{D}z(u_I) = \frac{1}{|I|}\sum_{i \in I} \delta_{z_i}.$$

This is the *empirical law*. Its mass at a record a is $|z^{-1}(a)|/|I|$. Sampling from this law amounts to choosing a uniformly distributed index and reading the record there. For example, the sample with three records a, a, b gives mass $2/3$ to a. A state uniform on the two-element set of observed records would give a mass $1/2$.

For fixed I, empirical laws are natural in the record set. Pushing along a recoding $g\colon Z \to W$ gives the empirical law of $g \circ z$.

$$\begin{array}{ccc} \mathrm{Hom}_{\mathbf{Set}}(I, Z) & \xrightarrow{\mathrm{Emp}_I} & \mathcal{D}Z \\ \downarrow{\scriptstyle g\circ(-)} & & \downarrow{\scriptstyle \mathcal{D}g} \\ \mathrm{Hom}_{\mathbf{Set}}(I, W) & \xrightarrow{\mathrm{Emp}_I} & \mathcal{D}W \end{array}$$

The two routes agree by functoriality of $\mathcal{D}$. A recoding can identify records and thereby sum their masses.

A reindexing $r\colon J \to I$ from a finite nonempty J gives $\widehat{p}_{z\circ r} = \mathcal{D}z(\mathcal{D}r(u_J))$. The weights before applying z are $|r^{-1}(i)|/|J|$. Equal positive fiber sizes give the original uniform weights, including when r is a bijection.

37.2 Multisets and pooling

Two indexed samples are isomorphic over Z when their index sets have a bijection that commutes with the record maps. Such an isomorphism restricts to a bijection on each fiber over Z, so the two samples have the same counts. Conversely, matching counts allow each nonempty fiber to be matched separately. The union of the finitely many fiber bijections is then an isomorphism of samples. The isomorphism class is therefore given by the finite-support multiplicity function $m_z(a) = |z^{-1}(a)|$.

Allow an empty index set when working with counts. Disjoint union adds the functions m_z, with the empty sample giving the zero function. The resulting commutative monoid is free on Z. For any function $v\colon Z \to M$ to the elements of a commutative monoid M, send a multiplicity function m to $\sum_a m(a)v(a)$, where $m(a)v(a)$ is repeated addition in M. The sum is finite, addition of multiplicities makes the map a monoid homomorphism, and $v(a)$ is its value on the single record a. These single records generate the monoid, which proves uniqueness. This is the commutative version of the free monoid property of Chapter 5.

Normalization to total mass one uses the sample size $n = |I|$. For two nonempty samples $z\colon I \to Z$ and $w\colon J \to Z$, write $[z, w]\colon I \amalg J \to Z$ for their pooled sample. Every occurrence contributes one unit to its count, so

$$\widehat{p}_{[z,w]} = \frac{|I|}{|I| + |J|}\,\widehat{p}_z + \frac{|J|}{|I| + |J|}\,\widehat{p}_w.$$

The same formula follows by choosing a batch in proportion to its size and sampling uniformly within that batch. This procedure assigns probability $1/(|I| + |J|)$ to every pooled index. Pooling therefore combines normalized laws with weights determined by their original sample sizes. Multiplying every count by the same positive integer preserves its normalized law and changes the size available for weighting future batches.

37.3 Averaging the loss

Let A be a set of actions. A predictor $h\colon X \to A$ and a loss $\ell\colon A \times Y \to \mathbb{R}_{\geq 0}$ give a record-wise loss map $\ell_h = \ell \circ (h \times \mathrm{id}_Y)\colon Z \to \mathbb{R}_{\geq 0}$. When $A = Y$, the zero–one loss is 0 if the action equals the label and 1 otherwise. For real labels and real actions, a common choice is $\ell(a, y) = (a - y)^2$.

A family of candidate predictors may be presented by a single map $H\colon \Theta \times X \to A$. Each parameter $\theta \in \Theta$ defines $h_\theta(x) = H(\theta, x)$. The equivalent map $\Theta \to A^X$ sends θ to h_θ, by transposition from Chapter 26. Its image is the hypothesis class of functions described by H. Different parameters can give the same predictor and therefore the same losses.

Write $L = \mathbb{R}_{\geq 0}$. The map $b\colon \mathcal{D}L \to L$ takes the mean $b(p) = \sum_a p(a)a$. For a distribution p on records, its risk and the empirical risk on z are

$$R_p(h) = b\big(\mathcal{D}\ell_h(p)\big) = \sum_{(x,y)\in Z} p(x,y)\ell(h(x),y),$$
$$\widehat{R}_z(h) = R_{\widehat{p_z}}(h) = \frac{1}{|I|}\sum_{i\in I}\ell(h(x_i),y_i).$$

The second formula is the first applied after the uniform index has been read. Equivalently, in the following triangle both routes from $\mathcal{D}I$ give the distribution of the losses.

$$\begin{array}{ccc} \mathcal{D}I & \xrightarrow{\mathcal{D}z} & \mathcal{D}Z \\ & {\scriptstyle \mathcal{D}(\ell_h\circ z)}\searrow & \downarrow{\scriptstyle \mathcal{D}\ell_h} \\ & & \mathcal{D}L \end{array}$$

The average of a mixture $P \in \mathcal{D}\mathcal{D}L$ is the average of its individual means.

$$b\big(\mu_L(P)\big) = \sum_{p\in\mathcal{D}L} P(p)\sum_{a\in L} p(a)a = b\big(\mathcal{D}b(P)\big).$$

Together with $b(\delta_a) = a$, this makes b an algebra for the distribution monad of Chapter 20. The empirical risk of a fixed predictor on pooled data is therefore the batch-size-weighted mean of its two empirical risks.

A feature map $f\colon X' \to X$ takes a predictor on X to the predictor $h \circ f$ on X'. It also sends a distribution on $X' \times Y$ forward along $f \times \mathrm{id}_Y$. The two changes are related by the equality

$$R_p(h \circ f) = R_{\mathcal{D}(f\times\mathrm{id}_Y)(p)}(h) \qquad \big(p \in \mathcal{D}(X' \times Y)\big).$$

Indeed $\ell_{h\circ f} = \ell_h \circ (f \times \mathrm{id}_Y)$, as the following square expresses before applying $\mathcal{D}$.

$$\begin{array}{ccc} X' \times Y & \xrightarrow{f\times\mathrm{id}_Y} & X \times Y \\ \downarrow{\scriptstyle (h\circ f)\times\mathrm{id}_Y} & & \downarrow{\scriptstyle h\times\mathrm{id}_Y} \\ A \times Y & = & A \times Y \end{array}$$

Thus predictor substitution is contravariant in the feature map, and changing the law on the record set is covariant. The risk calculation pairs these two operations. When f identifies two inputs, a predictor of the form $h \circ f$ must give the same action at both. Pushing data along f therefore evaluates a restricted class of predictors on the original inputs.

37.4 Minimizing empirical risk

For a hypothesis class $\mathcal{H} \subseteq A^X$, empirical risk minimization selects an element of

$$\arg\min_{h \in \mathcal{H}} \widehat{R}_z(h).$$

A nonempty finite class always has a minimizer. Ties can give several minimizers, and a selection rule must say which to return. An infinite class can have a finite infimum that is unattained. For example, let X have one input, all labels be zero, and let $\mathcal{H}$ consist of the constant real predictors with value $a > 0$. The squared risks are a^2, with infimum 0 and no minimum. Shalev-Shwartz and Ben-David [28, Chapter 2] develop the statistical learning framework around this minimization principle.

Proposition 37.1. *The set of empirical risk minimizers is invariant under bijections of sample indices. An empirical risk minimization rule with a fixed tie-breaker on $\mathcal{H}$ has the same invariance.*

Proof. Reindexing by a bijection gives the same empirical law and hence the same objective on $\mathcal{H}$. A fixed tie-breaker selects the same element from the common minimizer set. □

Example 37.2 (Repeated records and the selected predictor)**.** Let $X = \{a, b\}$ and $Y = A = \{0, 1\}$. Use zero–one loss and compare h_0, which always gives 0, with h_+, which gives 0 at a and 1 at b. The data have these counts.

	$y = 0$	$y = 1$
$x = a$	4	1
$x = b$	1	3

The risk of h_0 is $4/9$, counting the four labels 1. The risk of h_+ is $2/9$, counting the single error at each input. Thus h_+ uniquely minimizes empirical risk in the two-element class. Add three further copies of the existing record $(b, 0)$. The counts now give risks $4/12$ and $5/12$ respectively. The unique minimizer becomes h_0. Both tables have the same set of observed records. The change comes from their relative multiplicities.

The two batch minima are $2/9$ and 0, whose weighted average is $2/12$. The pooled minimum is $4/12$. In general the weighted average of the separate minima bounds the pooled minimum below, since each predictor incurs at least the minimum on each batch. With positive weights, a predictor attains the lower bound precisely when it minimizes on both batches.

37.5 Population risk and conditional decisions

Let $p \in \mathcal{D}(X \times Y)$ be the population law. The population risk $R_p(h)$ is the mean loss on one record drawn from p. If each random sample record has marginal law p, linearity of expectation gives

$$\mathbb{E}[\widehat{R}_z(h)] = \frac{1}{|I|} \sum_{i \in I} \mathbb{E}[\ell(h(x_i), y_i)] = R_p(h)$$

for each fixed h, using only the marginal laws. For a single input with a fair binary label, suppose all n sample labels copy one fair bit. Each fixed constant predictor has expected empirical risk $1/2$. The rule selecting the sampled bit has empirical risk zero and population risk $1/2$. Controlling the chosen predictor requires more than the fixed-predictor identity.

To calculate the best decision for a known finite population law, let p_X be its input marginal. For x with $p_X(x) > 0$, let $q(y \mid x) = p(x, y)/p_X(x)$, using the factorization of Chapter 20. Then

$$R_p(h) = \sum_{x : p_X(x) > 0} p_X(x) \sum_{y \in Y} q(y \mid x) \ell(h(x), y).$$

If A is finite and nonempty and all maps $X \to A$ are allowed, each term can be minimized at its own input. Choose an action of smallest conditional mean loss at every positive-mass input, and choose any action elsewhere. The resulting predictor is a *Bayes decision rule*. A stochastic decision at an input has conditional risk a weighted average of the losses for the available actions, and so cannot lower the minimum. The pointwise choices need not all be realizable in a restricted class $\mathcal{H}$.

For zero–one loss the conditional risk of predicting a is $1 - q(a \mid x)$, so a conditional most probable label is optimal. With binary labels, suppose the loss is 5 for predicting 0 on label 1, 1 for predicting 1 on label 0, and 0 for correct predictions. If the conditional probability of 1 is u, the conditional loss of action 0 is $5u$ and that of action 1 is $1 - u$. Action 1 is optimal when $u \geq 1/6$, with a tie at equality. Using the original table as a population law gives $u = 1/5$ at a and $u = 3/4$ at b. The Bayes decision under this loss predicts 1 at both inputs, with risk $5/9$. Under zero–one loss the optimal rule instead predicts 0 at a and 1 at b.

Loss minimization becomes a concrete calculation once the class and loss have been specified. The next chapter makes that calculation explicit for regression and kernel predictors.

CHAPTER 38

Regression and kernel methods

Squared loss turns a table of observations into a point in a Euclidean space. A linear model predicts points in a subspace of that space. Fitting the model amounts to projecting the observations onto the subspace. When the parameters themselves carry an inner product, a penalty on their length gives a unique fit and leads to a second description. Its linear system is indexed by observations, and its entries are inner products of features.

38.1 The adjoint of an evaluation map

Let $(x_i, y_i)_{i=1}^n$ be an indexed sample with real labels, where $n \geq 1$, and write $y = (y_i)_{i=1}^n \in \mathbb{R}^n$. Suppose that the predictors are parametrized by a finite-dimensional real inner product space E and that each prediction is linear in the parameter. Thus the evaluation map

$$A \colon E \longrightarrow \mathbb{R}^n, qquad (Aw)_i = f_w(x_i)$$

is linear. The squared empirical risk of Chapter 37 is $n^{-1}\|Aw - y\|^2$. Multiplication by n leaves its minimizers the same, so we use the sum of squares.

The dual of A acts on linear functionals by precomposition, as in Chapter 14. An inner product identifies a vector v with the functional $z \mapsto \langle z, v\rangle$. Transferring the dual map across these identifications gives the *adjoint* A^*. It is characterized by

$$\langle Aw, z\rangle = \langle w, A^* z\rangle.$$

Write R for the inner product identification with the dual.

$$\begin{array}{ccc} \mathbb{R}^n & \xrightarrow{\;A^*\;} & E \\ \downarrow{\scriptstyle \cong} & & \downarrow{\scriptstyle \cong} \\ R_{\mathbb{R}^n} & & R_E \\ \downarrow & & \downarrow \\ (\mathbb{R}^n)^* & \xrightarrow{\;A^\vee\;} & E^* \end{array}$$

In orthonormal bases the adjoint is represented by the transpose matrix. The adjoint reverses composition since $\langle BAw, z\rangle = \langle Aw, B^* z\rangle = \langle w, A^* B^* z\rangle$. The identity has itself as adjoint, and $(A^*)^* = A$ by symmetry of the inner products.

38.2 Least squares as an orthogonal projection

Let $M = \operatorname{im} A$ and write P_M for the orthogonal projection onto M. To construct it, choose an orthonormal basis of M and sum the components of a vector along the basis. The result lies in M, and the remaining component is orthogonal to every basis vector. These two conditions determine the result uniquely, since a vector in $M \cap M^\perp$ is zero. For the inclusion $i\colon M \hookrightarrow \mathbb{R}^n$, this says $P_M = i^*$, since $\langle i(v), y\rangle = \langle v, P_M y\rangle$ for every $v \in M$. In particular $i^* i = \mathrm{id}_M$, and ii^* is the projection written as an endomorphism of $\mathbb{R}^n$. For every $w \in E$ the two summands on the right of

$$\|Aw - y\|^2 = \|Aw - P_M y\|^2 + \|y - P_M y\|^2$$

come from perpendicular vectors. Thus the minimum is attained exactly when $Aw = P_M y$. The residual $Aw - y$ is then orthogonal to M. The adjoint gives the normal equation

$$A^* A w = A^* y.$$

Indeed $A^*(Aw - y) = 0$ holds exactly when $\langle Av, Aw - y\rangle = 0$ for every $v \in E$, which is precisely the orthogonality condition.

All minimizers differ by vectors in $\ker A$. The restriction $A_\perp$ of A to $(\ker A)^\perp$ has trivial kernel and image M, since the component of any parameter in $\ker A$ contributes zero to its prediction. The unique solution in $(\ker A)^\perp$ has least norm by the perpendicular decomposition of the other solutions. It is the map $w_{\min}$ applied to y.

$$\begin{array}{ccc} \mathbb{R}^n & \xrightarrow{\quad P_M \quad} & M \\ & \searrow^{w_{\min}} & \downarrow{\scriptstyle A_\perp^{-1}} \\ & & (\ker A)^\perp \end{array}$$

In particular $A^* A$ is invertible if $\ker A = 0$, because $\langle w, A^* A w\rangle = \|Aw\|^2$. When the kernel is nonzero, the fitted vector is still unique.

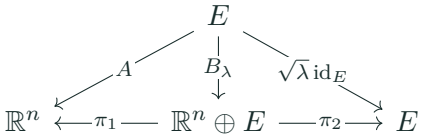
Example 38.1 (Two parameters with the same effect)**.** Let $A(w_1, w_2) = (w_1 + w_2, w_1 + w_2)$ and $y = (3, 1)$. Then $P_M y = (2, 2)$, so the least-squares solutions are exactly $w_1 + w_2 = 2$. The minimum-norm solution is $(1, 1)$. A change by $(t, -t)$ adds $2t^2$ to the squared parameter norm and leaves the fitted values at $(2, 2)$.

38.3 Ridge regression

For $\lambda > 0$, the ridge objective on E is

$$Q_\lambda(w) = \|Aw - y\|^2 + \lambda \|w\|^2.$$

This is a sum of squares convention. Adding $\rho\|w\|^2$ to the empirical mean corresponds to taking $\lambda = n\rho$. Let $B_\lambda \colon E \to \mathbb{R}^n \oplus E$ send w to $(Aw, \sqrt{\lambda}w)$. The objective is the squared distance from $B_\lambda w$ to $(y, 0)$, so this is another projection problem.



The adjoint sends (z, v) to $A^*z + \sqrt{\lambda}v$. The least-squares condition for the enlarged map is therefore

$$(A^*A + \lambda\,\mathrm{id}_E)w = A^*y.$$

The left-hand operator has trivial kernel since its quadratic form is $\|Aw\|^2 + \lambda\|w\|^2$. It is invertible in finite dimensions, and the solution is unique even when A has a kernel. In the two-parameter example, choosing $\lambda = 2$ gives $w_1 = w_2 = 2/3$. The penalty shrinks each fitted value from 2 to $4/3$.

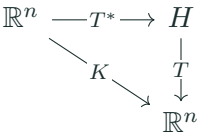
38.4 Feature vectors and the representer calculation

Let X be the input set, H a real Hilbert space, and $\phi \colon X \to H$ a feature map. Allowing H to be infinite dimensional does not change the number of evaluations in the sample. A parameter $w \in H$ predicts $f_w(x) = \langle w, \phi(x)\rangle$. Write T for evaluation at the training inputs. Its adjoint and the resulting Gram matrix are given by

$$(Tw)_i = \langle w, \phi(x_i)\rangle, \qquad T^*c = \sum_{i=1}^{n} c_i \phi(x_i),$$

$$K_{ij} = k(x_i, x_j), \qquad k(x, z) = \langle \phi(x), \phi(z)\rangle.$$

The formula for T^* follows by pairing Tw with c. Thus $TT^* = K$. The triangle distinguishes synthesis of a feature-space vector from its evaluations on the sample.



The matrix is symmetric and positive semidefinite, because for $c \in \mathbb{R}^n$,

$$c^{\mathsf{T}} K c = \|T^*c\|^2 \geq 0.$$

These properties hold for every finite choice of inputs, with repeated inputs allowed. A symmetric real function on $X \times X$ with this property is called a positive semidefinite kernel.

Proposition 38.2 (Finite-sample representer property)**.** *Let an objective on H have the form $\ell(Tw) + g(\|w\|)$, where $\ell\colon \mathbb{R}^n \to \mathbb{R}$ depends only on the sample evaluations and the real-valued function g is strictly increasing. If a minimizer exists, it belongs to $H_D = \operatorname{span}\{\phi(x_i)\}_{i=1}^n$. For $\ell(v) = \|v - y\|^2$ and $g(r) = \lambda r^2$ with $\lambda > 0$, the minimizer exists, is unique, and is given by*

$$w = T^*\alpha, qquad\alpha = (K + \lambda \operatorname{id}_{\mathbb{R}^n})^{-1} y.$$

Proof. The subspace H_D is finite dimensional. Projection onto an orthonormal basis of it decomposes any w as $u + v$ with $u \in H_D$ and $v \perp H_D$. Then $Tv = 0$, while $\|u + v\|^2 = \|u\|^2 + \|v\|^2$. If $v \neq 0$, replacing w by u strictly decreases the objective. This proves the assertion about every minimizer.

For the ridge objective, evaluation restricted to H_D is a map between finite-dimensional inner product spaces. The ridge calculation there gives a unique solution to the normal equation, and the perpendicular decomposition gives uniqueness on all of H. Since K is positive semidefinite, $K + \lambda \operatorname{id}_{\mathbb{R}^n}$ is invertible. For the proposed $w = T^*\alpha$ the normal equation holds by the identity

$$T^*TT^*\alpha + \lambda T^*\alpha = T^*(K + \lambda \operatorname{id}_{\mathbb{R}^n})\alpha = T^*y.$$

Since this w lies in H_D, it is the unique solution there. □

This proof uses only the finite span of the training features. Schölkopf, Herbrich, and Smola give a general representer theorem with this projection argument [29]. After calculating α, the prediction at a new input is

$$f_w(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x).$$

Thus the fitting and prediction calculations require only evaluations of k. This is the kernel trick for ridge regression.

38.5 Kernels as inner-product data

A kernel given directly also supplies a suitable feature space. Start with the vector space of finite formal sums of symbols $[x]$. Define a symmetric bilinear form by $\langle [x], [z] \rangle = k(x, z)$, extended linearly. Kernel positivity says that its quadratic form is nonnegative. For a vector v of squared length zero, the squared length of $u + tv$ is nonnegative for all $t \in \mathbb{R}$. Its linear coefficient must vanish. Hence $\langle v, u \rangle = 0$ for every u. Quotienting by these zero-length vectors gives an inner product space. Complete it in the induced norm, as in Chapter 18, to obtain H. The inner product extends to the completion by the Cauchy–Schwarz inequality, and the class of $[x]$ gives a feature with the prescribed inner products.

In any other Hilbert realization, send each $[x]$ to its feature vector. The linear map on formal sums preserves squared lengths and kills the zero-length subspace. It therefore extends to an isometry on the completion. Its image is the closed linear span of the other feature vectors, and those vectors determine it uniquely by density. Thus the kernel determines the feature space generated by the inputs up to isometry.

Example 38.3 (A quadratic feature fit)**.** On $X = \mathbb{R}$, take $k(x, z) = (1 + xz)^2$. The feature vector $(1, \sqrt{2}x, x^2)$ proves its positivity. Use inputs $(-1, 0, 1)$, labels $(1, 0, 1)$, and $\lambda = 1$. Then

$$K = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 4 \end{pmatrix}.$$

By symmetry the first and last coordinates of α agree. Writing them as a and the middle coordinate as b, the system is $5a + b = 1$ and $2a + 2b = 0$. It follows that $a = 1/4$ and $b = -1/4$. Substituting in the kernel expansion gives

$$f_w(x) = \tfrac{1}{4}\big((1 - x)^2 - 1 + (1 + x)^2\big) = \tfrac{1}{4} + \tfrac{1}{2}x^2.$$

The predicted values at the sample inputs are $(3/4, 1/4, 3/4)$. Their difference from y is $-\alpha$, as predicted by $K\alpha - y = -\lambda\alpha$.

38.6 Tangent features of a parametrized model

Consider scalar predictions $f_\theta(x)$ with a finite-dimensional Euclidean parameter θ, differentiable at θ_0 for each input x. The derivative in the parameter is a covector. The inner product identifies it with a feature $\phi_0(x)$ satisfying

$$D_\theta f_{\theta_0}(x)[v] = \langle v, \phi_0(x) \rangle.$$

The first-order expansion of Chapter 19 gives the affine model $f_{\theta_0}(x) + \langle v, \phi_0(x) \rangle$ for the displacement v. For a fixed finite sample, write $F(\theta) = (f_\theta(x_i))_{i=1}^n$ and $J_0 = DF_{\theta_0}$. The empirical neural tangent kernel matrix at initialization is $K_0 = J_0 J_0^*$, with entries $\langle \phi_0(x_i), \phi_0(x_j) \rangle$ [30].

Training the affine model with a squared penalty on v is exactly kernel ridge regression with the centered labels $y - F(\theta_0)$, followed by adding $f_{\theta_0}(x)$ to its prediction. There is also a prediction-space formula for gradient flow on the unpenalized affine model. Let $z(t) = F(\theta_0) + J_0 v(t)$, with the objective $\frac{1}{2}\|z - y\|^2$. Its parameter gradient is $J_0^*(z - y)$, so the differential equation for predictions is

$$\dot{v} = -J_0^*(z - y) \qquad \Longrightarrow \qquad \dot{z} = -K_0(z - y).$$

For a unit eigenvector u of K_0 with eigenvalue κ, the residual component $r_u(t) = \langle z(t) - y, u \rangle$ satisfies $\dot{r}_u = -\kappa r_u$. Thus $r_u(t) = e^{-\kappa t} r_u(0)$. This component decays

for $\kappa > 0$. When $\kappa = 0$, the equation $K_0 u = 0$ implies $J_0^* u = 0$, so changing the parameter in the affine model leaves this component fixed.

Along a gradient-flow solution of the original smooth model, the chain rule gives

$$\frac{d}{dt} F(\theta(t)) = -DF_{\theta(t)}\, DF_{\theta(t)}^*(F(\theta(t)) - y).$$

Here the kernel matrix is evaluated at the current parameter and can vary over time. The fixed-kernel calculation describes the affine model with its derivative held at initialization. The original model has this same prediction equation when its gradient Gram matrix stays equal to K_0 along the trajectory.

Example 38.4 (Fitting a frozen tangent model)**.** Take $f_{(a,b)}(x) = abx$ and initialize at $(a, b) = (1, 2)$. The tangent feature is $\phi_0(x) = (2x, x)$, so that $k_0(x, z) = 5xz$. Use the single observation $(1, 5)$ and $\lambda = 1$. The centered label is $5-2 = 3$, hence $\alpha = 3/(5+1) = 1/2$. The displacement inferred in the affine model is $v = \alpha\phi_0(1) = (1, 1/2)$, yielding the fitted prediction $2x + \langle v, \phi_0(x)\rangle = 9x/2$. Evaluating the original parametrization at the displaced parameter $(2, 5/2)$ gives $5x$. The difference is exactly the second-order term $v_a v_b x = x/2$ of the original product.

38.7 Pontryagin duality and stationary kernels

Let G be a finite abelian group of order N, written additively. A kernel invariant under simultaneous translation of its two inputs has the form $k(x, z) = b(x - z)$, with $b(t) = k(t, 0)$. The characters in the Pontryagin dual $\widehat{G}$ give coordinates for the full Gram operator

$$(Kf)(x) = \sum_{z \in G} b(x - z) f(z), \qquad f \in \mathbb{C}^G.$$

Finite Pontryagin duality (Section 8.6) gives $|\widehat{G}| = N$. The vectors $N^{-1/2}\chi$, for $\chi \in \widehat{G}$, are an orthonormal basis for the counting inner product. Indeed, the sum of any nontrivial character over G is zero, since translating the index multiplies the sum by a character value different from one. Applying this to $\chi\overline{\psi}$ gives orthogonality.

Write $a_\chi = N^{-1} \sum_{t \in G} b(t)\overline{\chi(t)}$, so that $b(t) = \sum_\chi a_\chi \chi(t)$. The substitution $t = x - z$ and the character law give

$$(K\chi)(x) = \chi(x) \sum_{t \in G} b(t)\overline{\chi(t)} = N a_\chi \chi(x).$$

Thus the unitary Fourier coordinate map $\mathcal{F}$, whose coordinates are taken in the normalized character basis, takes K to the diagonal operator D with entries

Na_χ.

$$\begin{array}{ccc} \mathbb{C}^G & \xrightarrow{\;K\;} & \mathbb{C}^G \\ \downarrow{\scriptstyle \mathcal{F}} & & \downarrow{\scriptstyle \mathcal{F}} \\ \mathbb{C}^{\widehat{G}} & \xrightarrow{\;D\;} & \mathbb{C}^{\widehat{G}} \end{array}$$

Every sample Gram matrix of the kernel is Hermitian positive semidefinite exactly when every a_χ is real and nonnegative. Necessity follows by testing K on each character. For sufficiency, the coordinates $\phi(x)_\chi = \sqrt{a_\chi}\,\chi(x)$ give

$$k(x,z) = \sum_{\chi \in \widehat{G}} \phi(x)_\chi \, \overline{\phi(z)_\chi}.$$

Every sample quadratic form is therefore a sum of squared moduli, including for samples with repeated inputs. When k is real, conjugating the coefficient formula gives $a_{\chi^{-1}} = \overline{a_\chi} = a_\chi$. The real part of the feature inner product then gives a real Hilbert realization. This is the finite Bochner theorem. Section 16.6 gives the locally compact setting for the character group.

Example 38.5 (A kernel on the four-point circle)**.** On $G = \mathbb{Z}/4\mathbb{Z}$, let $b(t) = \frac{1}{2} + \frac{1}{2}\cos(\pi t/2)$. For $\chi_j(t) = e^{2\pi i j t/4}$ the weights in order $j = 0,1,2,3$ are $(1/2, 1/4, 0, 1/4)$. Thus the full Gram matrix has eigenvalues $2, 1, 0, 1$, and real features are $2^{-1/2}(1, \cos(\pi x/2), \sin(\pi x/2))$. On the full group, the ridge prediction $K(K + \lambda\,\mathrm{id})^{-1} y$ multiplies the Fourier coordinate of the labels by $Na_\chi/(Na_\chi + \lambda)$. The alternating mode $\chi_2(x) = (-1)^x$ contributes zero to the fitted values for every $\lambda > 0$.

CHAPTER 39

Neural networks from parameterized maps

A layer takes an input and some parameters and returns a new value. When two layers are composed, their parameters can be chosen independently or constrained to agree. The underlying composition formula is the same in either case. First assemble a parameterized map with slots for both choices, then specify the allowed choices by a map into those slots. This separates evaluation from the relations on the parameters.

39.1 A category of parameterized maps

Let $\mathcal{C}$ have finite products, with terminal object 1. For example, we may use sets and functions or finite-dimensional real vector spaces and smooth maps. A *parameterized map* from X to Y is a pair (P, f) with an object P and a map $f\colon P \times X \to Y$. A point $p\colon 1 \to P$ gives an ordinary map $f_p\colon X \to Y$ by restriction.

$$\begin{array}{ccc} X & \xrightarrow{\langle p!, \mathrm{id}_X \rangle} & P \times X \\ & \searrow^{f_p} & \downarrow f \\ & & Y \end{array}$$

Here $!\colon X \to 1$ is the terminal map, so the first component of the pairing is constant at p. In sets this says $f_p(x) = f(p, x)$. If the function object Y^X exists, transposing f gives a map $P \to Y^X$ that classifies the family.

For a second parameterized map (Q, g) from Y to Z, use $P \times Q$ for the parameter object of the composite, in order of execution. On $(P \times Q) \times X$ write p, q, x for its three projections. The composite $(P \times Q, h)$ is obtained by following the top and right arrows in this diagram.

$$\begin{array}{ccc} (P \times Q) \times X & \xrightarrow{\langle q, f \circ \langle p, x \rangle \rangle} & Q \times Y \\ & \searrow^{h} & \downarrow g \\ & & Z \end{array}$$

In set notation $h((p, q), x) = g(q, f(p, x))$, and fixing a pair (p, q) gives $h_{(p,q)} = g_q \circ f_p$. The universal property of products constructs this composite in any $\mathcal{C}$ with finite products.

We identify (P, f) and (P', f') when there is an isomorphism $r\colon P' \to P$ satisfying $f' = f \circ (r \times \mathrm{id}_X)$. This is an equivalence relation by the identity, composition, and inverse laws. Composing with a third map respects it. Indeed, renaming the P, Q slots by two isomorphisms renames their product by the product of those isomorphisms and gives the same composite formula. All sets or categories of parameters can be taken in a fixed universe.

Composition of three representatives with parameters P, Q, R gives the nested expression $k(r, g(q, f(p, x)))$ on the appropriate projections. The parameter objects $(P\times Q)\times R$ and $P\times(Q\times R)$ differ by the product rebracketing isomorphism. After transport, each of the three projections agrees, and so do the maps defined by the expression. This gives associativity on classes. The identity has parameter 1 and map $\pi_X\colon 1 \times X \to X$. Composing with it inserts a terminal slot, which the product unit isomorphism removes. Thus the classes form a category, denoted $\mathrm{Para}(\mathcal{C})$. Fong, Spivak, and Tuyéras use this construction for Euclidean spaces in their compositional treatment of learning [31].

Equality of the attainable functions is a coarser test. In sets, take the ordinary identity on X as a family with parameter 1. Take it again with parameter $\{0, 1\}$ by projecting onto X at each parameter. Both families have image consisting of the same single function $X \to X$, and their parameter objects have different cardinalities. The parameterized maps therefore give distinct arrows. The distinction keeps information about how each function is indexed.

39.2 Parallel maps and shared parameters

Given (P, f) from X to Y and (Q, g) from X' to Y', their tensor runs the maps in parallel. Its source is $X \times X'$, its target is $Y \times Y'$, and its parameter object is $P \times Q$. The defining map is

$$((p, q), (x, x')) \longmapsto (f(p, x), g(q, x')).$$

In $\mathcal{C}$ this is the composite of the product permutation that groups (p, x) and (q, x') with $f \times g$. If each parallel branch is followed by another parameterized map, composing before tensoring and tensoring before composing only changes the order of the parameter slots. For example, with second parameters R, S, the orders are (P, R, Q, S) and (P, Q, R, S), and the permutation matches the slots used by each branch. The product associator, unit maps, and symmetry of $\mathcal{C}$ are parameter-free arrows in this construction. All the needed coherence equations compare product maps with the same projections. This yields a symmetric monoidal structure.

The product in the underlying category thus supplies the tensor of parameterized maps. Its unit 1 can have many parameterized endomorphisms, since $(P, !)\colon 1 \to 1$ depends on the isomorphism class of P. For the categories of sets

and smooth vector spaces there are many such classes, so this unit is not terminal in their parameterized categories.

To send the same input to two branches, precompose their tensor with the parameter-free diagonal $\Delta_X \colon X \to X \times X$. For $(P, f) \colon X \to Y$ and $(Q, g) \colon X \to Z$, this gives the family $((p, q), x) \mapsto (f(p, x), g(q, x))$. An ordinary map out of $Y \times Z$ may then combine the branch values. With a vector space as output and with one branch equal to the identity, combining by addition gives a skip map $x \mapsto x + f_p(x)$.

Parameters can also be shared. For any $r \colon T \to P$, the new representative is obtained by the upper route below.

$$\begin{array}{ccc} T \times X & \xrightarrow{r \times \mathrm{id}_X} & P \times X \\ & \searrow^{f \circ (r \times \mathrm{id}_X)} & \downarrow{\scriptstyle f} \\ & & Y \end{array}$$

Its family at t is $f_{r(t)}$. Reindexing first by r and then by $s \colon U \to T$ gives the same result as reindexing by $r \circ s$, since $(r \times \mathrm{id}_X) \circ (s \times \mathrm{id}_X) = (r \circ s) \times \mathrm{id}_X$. In particular, if two successive maps use parameters in the same object P, form the composite with its two independent slots, then reindex along $\Delta_P \colon P \to P \times P$. This gives the family in which one parameter is used twice. The same restriction can share the parameters of parallel branches. In sets the restriction evaluates both branches on pairs of the form (p, p).

39.3 Affine layers and an exact network

An affine layer from $\mathbb{R}^n$ to $\mathbb{R}^m$ has parameter space $P = \mathbb{R}^{m \times n} \times \mathbb{R}^m$. Its evaluation is the smooth map

$$((W, b), x) \longmapsto Wx + b.$$

The parameters W and b have ordinary vector coordinates, and evaluation uses matrix multiplication and addition. A fixed smooth scalar function σ gives an activation $(y_1, \ldots, y_m) \mapsto (\sigma(y_1), \ldots, \sigma(y_m))$. This map has parameter 1. Composing the affine map and activation gives a layer $x \mapsto \sigma(Wx + b)$, with the activation applied componentwise. For two layers the composite is

$$x \longmapsto \sigma_2(W_2\, \sigma_1(W_1 x + b_1) + b_2).$$

The two affine parameter slots enter by the product law. An architecture supplies the maps, their order of composition, and any parameter restrictions. Each choice of the remaining parameters then gives one predictor in the sense of Chapter 37.

Example 39.1 (A shared offset)**.** Use a first scalar affine layer $x \mapsto x + a$, the activation $u \mapsto u^2$, and a last affine layer $v \mapsto v + b$. The two independent offsets give

$$H((a, b), x) = (x + a)^2 + b.$$

The free and shared computations can be read in the square

$$\begin{array}{ccc} \mathbb{R} \times \mathbb{R} & \xrightarrow{\Delta_{\mathbb{R}} \times \mathrm{id}_{\mathbb{R}}} & (\mathbb{R} \times \mathbb{R}) \times \mathbb{R} \\ \downarrow{\scriptstyle H_{\mathrm{sh}}} & & \downarrow{\scriptstyle H} \\ \mathbb{R} & = & \mathbb{R}. \end{array}$$

At $x = 3$ the independent choice $(a, b) = (2, -1)$ gives $5^2 - 1 = 24$. The shared choice $a = 2$ gives $5^2 + 2 = 27$. The independent choice $(0, 1)$ gives the whole function $x^2 + 1$. No shared choice gives that function. Comparing the coefficients of x in $(x + a)^2 + a$ and $x^2 + 1$ forces $2a = 0$. At $a = 0$ the constant coefficient is 0, contradicting the required constant coefficient 1.

Reindexing also specifies how derivatives depend on the shared coordinates. For fixed x, the derivative of H with respect to (a, b) at that point acts on a pair of velocities by $(u, v) \mapsto 2(x + a)u + v$. The derivative of the parameter diagonal sends $t \mapsto (t, t)$. Thus the chain rule gives

$$D_a H_{\mathrm{sh}}(a, x)(t) = \big(2(x + a) + 1\big)t.$$

Both parameter uses contribute to differentiation of the shared coordinate. More generally the dual of the diagonal derivative sends a pair of covectors (λ, μ) to their sum, since evaluating on t gives $\lambda(t) + \mu(t)$. This accumulation of contributions from both uses of a weight will reappear in Chapter 40.

39.4 Attention as a weighted evaluation

Some layers choose coefficients by comparing inputs with a family of reference vectors. Consider a finite set I of query positions, a nonempty finite set J of key and value positions, query vectors $q_i \in \mathbb{R}^d$, key vectors $k_j \in \mathbb{R}^d$, and value vectors $v_j \in \mathbb{R}^m$. For $d \geq 1$, the scaled dot-product rule assigns scores and weights

$$s_{ij} = \frac{\langle q_i, k_j \rangle}{\sqrt{d}}, \qquad A_{ij} = \frac{\exp(s_{ij})}{\sum_{j' \in J} \exp(s_{ij'})}.$$

This is the scaled dot-product attention rule of Vaswani et al. [32]. The denominator is positive, and summing the weights over j yields 1. Adding the same constant to each score in one row cancels the same exponential factor in

numerator and denominator. Each row therefore gives a probability distribution on key positions that depends only on score differences within that row.

Holding the queries and keys fixed gives a stochastic map from I to J in the sense of Chapter 20. Write $\mathcal{D}$ for the finitely supported distribution functor and let $v\colon J \to \mathbb{R}^m$ take j to v_j. The layer takes the distribution on J, pushes it forward along v, and uses its mean as the output vector.

$$I \xrightarrow{A} \mathcal{D}J \xrightarrow{\mathcal{D}v} \mathcal{D}(\mathbb{R}^m) \xrightarrow{\text{mean}} \mathbb{R}^m$$

The output at i is $\sum_j A_{ij} v_j$, which is computed deterministically. If several key positions have the same value, their weights add under pushforward and contribute the same total to the mean. For fixed weights, mixing two rows before taking their mean gives the same vector as mixing their means. This follows by distributing the two finite sums. Thus if a fixed stochastic map B feeds positions into I, the combined weights for the corresponding means are the matrix product BA.

Example 39.2 (Two key positions)**.** Take $d = 1$, a query $q = 1$, and keys 0 and $\log 3$. Their scores are 0 and $\log 3$, so the row of weights is $(1/4, 3/4)$. For values $v_1 = (2, 0)$, $v_2 = (0, 4)$, the resulting vector is

$$\tfrac{1}{4}(2, 0) + \tfrac{3}{4}(0, 4) = (\tfrac{1}{2}, 3).$$

Adding 5 to both scores gives the same output, since the common factor $\exp(5)$ cancels.

In a network, the queries, keys, and values can themselves be produced by learned affine maps. At fixed lengths the operations above give a smooth map of their real coordinates, since the only division is by positive denominators. Composing it with the parameterized maps for the queries, keys, and values puts all their parameters into the new slots. A parameter restriction can share those slots across positions. The stochastic rows in this construction vary with the input and parameter values. Chapter 40 computes covectors through such compositions, and Chapter 41 uses them in parameter updates.

CHAPTER 40

Backpropagation from reverse differentiation

A network sends parameters and inputs forward to a prediction. A loss assigns a scalar to that prediction. Its differential measures the first-order change of that scalar, so it is a covector. To find the sensitivity to an earlier quantity, pass that covector back through the derivative of the map that used the quantity. When several maps use the same value, the resulting covectors add. These two rules give backpropagation on a graph of smooth maps.

40.1 The reverse arrow of a smooth map

Let $U \subseteq V$, $X \subseteq W$, and $Y \subseteq Z$ be open subsets of finite-dimensional real vector spaces. Write V^* for the space of real-valued linear maps on V. A smooth map $f\colon U \to X$ gives the linear map $Df_x\colon V \to W$ at $x \in U$. Taking the dual of this linear map gives its *reverse derivative* at x.

$$R[f]_x\colon W^* \longrightarrow V^*, \qquad R[f]_x(\lambda)(v) = \lambda(Df_x(v)).$$

This is the pullback of covectors from Chapter 19. Evaluation satisfies the following equality of pairings, with all arrows at the indicated base point.

$$\begin{array}{ccc} W^* \times V & \xrightarrow{R[f]_x \times \mathrm{id}} & V^* \times V \\ \downarrow{\scriptstyle \mathrm{id} \times Df_x} & & \downarrow{\scriptstyle \mathrm{ev}} \\ W^* \times W & \xrightarrow{\qquad \mathrm{ev} \qquad} & \mathbb{R} \end{array}$$

Given λ, the square specifies a unique covector by its value on every tangent vector. In chosen bases, if Df_x is an m-by-n matrix and λ is an m-component column of coefficients in the dual basis, $R[f]_x(\lambda) = (Df_x)^{\mathsf{T}}\lambda$ is an n-component column. This transpose describes the dual map using the two bases. The underlying covectors require no inner products.

For $g\colon X \to Y$, the chain rule gives

$$\begin{array}{ccc} Z^* & \xrightarrow{R[g]_{f(x)}} & W^* \\ & {\scriptstyle R[g\circ f]_x} \searrow & \downarrow{\scriptstyle R[f]_x} \\ & & V^* \end{array} \qquad R[g \circ f]_x = R[f]_x \circ R[g]_{f(x)}.$$

Indeed, evaluating the right-hand side on v yields $\lambda(Dg_{f(x)}(Df_x(v)))$, the defining evaluation of the left-hand side. For identity maps the reverse derivative is the identity on each fiber.

The reverse linear maps define a bundle map $f^*(T^*X) \to T^*U$ over U. Its fiber at x carries a covector at $f(x)$ to one at x. More generally, objects (U, E) consist of a smooth base U and a fixed finite-dimensional fiber E. An arrow $(U, E) \to (X, F)$ consists of a smooth map $f\colon U \to X$ and a smoothly varying family of linear maps $r_x\colon F \to E$. Composing (f, r) and (g, s) gives $(g \circ f, t)$ with $t_x = r_x \circ s_{f(x)}$. Associativity follows from associativity of both ordinary and linear composition. Identities use constant identity fiber maps. Sending an open subset $U \subseteq V$ to (U, V^*) and a smooth map f to $(f, R[f])$ is a functor by the chain rule. The base map in an arrow records the point at which to evaluate the next reverse map.

40.2 Products and the cotangent of copying

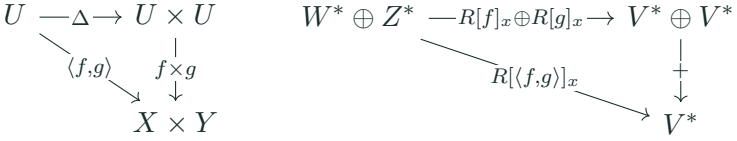
Write $V \oplus W$ for the product of the two vector spaces, with its coordinatewise addition. Every covector on the product is uniquely $(v, w) \mapsto \alpha(v) + \beta(w)$ for a pair $(\alpha, \beta) \in V^* \oplus W^*$. This follows from $(v, w) = (v, 0) + (0, w)$. Thus the dual of the first projection $(v, w) \mapsto v$ sends α to $(\alpha, 0)$, and the dual of the second sends β to $(0, \beta)$.

The diagonal $\Delta\colon V \to V \oplus V$ sends v to (v, v). Its derivative is itself. For every v and (α, β),

$$(\alpha, \beta)(\Delta(v)) = \alpha(v) + \beta(v), \qquad R[\Delta]_x(\alpha, \beta) = \alpha + \beta.$$

A forward copy of a value therefore becomes addition of the two sensitivities in the reverse direction. The unique map from V to the zero-dimensional space has zero reverse derivative. It models dropping a value, which contributes the zero covector in a reverse computation.

The same reasoning gives a rule for paired maps. If $f\colon U \to X$ and $g\colon U \to Y$ both use x, the forward pairing factors through Δ and the product map $f \times g$.



The derivative of $f \times g$ is block diagonal, as each component uses only its own input. Its dual therefore takes the two covectors back separately before the final addition. This also follows directly by differentiating $\langle f, g\rangle$ and evaluating at a pair of covectors.

40.3 The reverse sweep on a finite graph

Consider a finite directed acyclic graph of computations. Choose an order of its nodes in which all input nodes come first and every input to an operation precedes that operation. Such an order exists because a finite acyclic graph always has a node with no incoming edge, and removing one repeatedly gives an order. A node v holds a value z_v in an open subset of a real vector space E_v. For an operation node v, let $u_{v,1}, \dots, u_{v,k_v}$ be its input nodes in argument order. Repeated entries are allowed, since the same value can be used in two argument positions. Let F_v be the smooth map computing z_v from this tuple. Assume the final node L computes the scalar loss as a smooth function of earlier values.

At a fixed input, evaluate the graph in the chosen order. At the resulting argument tuple of F_v, the derivative in its j-th argument is a linear map

$$A_{v,j} \colon E_{u_{v,j}} \longrightarrow E_v.$$

It is the derivative of F_v precomposed with the inclusion of the j-th summand of the product of the input vector spaces. If the input nodes vary by tangent vectors δz_u, the chain rule gives

$$\delta z_v = \sum_{j=1}^{k_v} A_{v,j}(\delta z_{u_{v,j}}).$$

This uses linearity on the product, so the sum runs over occurrences of arguments, even when some come from the same input node.

Proposition 40.1 (Backpropagation)**.** *Assign to each node a covector $\bar{z}_v \in E_v^*$, initially zero except for $\bar{z}_L$ equal to the functional $t \mapsto t$ on $\mathbb{R}$. Visit the operation nodes in decreasing order. At v, for every argument occurrence j, add the contribution*

$$\bar{z}_{u_{v,j}} \longleftarrow \bar{z}_{u_{v,j}} + A_{v,j}^*(\bar{z}_v).$$

The resulting covectors at the input nodes together give the differential of the scalar loss with respect to those inputs.

Proof. For any input variation, the differential initially equals $\bar{z}_L(\delta z_L)$. Keep an expression for the same differential as a sum $\sum_v \bar{z}_v(\delta z_v)$ over nodes still to be visited and input nodes. At v, substitute the chain-rule expression for δz_v. Its term becomes

$$\bar{z}_v(\delta z_v) = \sum_{j=1}^{k_v} (A_{v,j}^* \bar{z}_v)(\delta z_{u_{v,j}}).$$

Each predecessor is still present in the sum because it is earlier in the forward order. Adding to its existing coefficient and removing the term at v preserves the expression. At the end only input nodes remain. Their variations are independent, so the resulting coefficients are precisely the components of the differential. □

At the time a node is visited in reverse, each of its uses has already contributed to its covector. This allows the computation to add these contributions first and apply the node's reverse map once. A variable that has no path to the loss receives zero. If a primitive uses the same node twice, both terms in the update still contribute, as for $z \mapsto z \cdot z$, whose two argument contributions sum to $2z\bar{z}_v$. The rule also applies with a chosen output covector in place of the scalar seed. It then computes the reverse derivative of the whole graph at that covector.

The forward evaluation stores the values at which the local reverse maps must be applied, or makes them available by recomputation. One reverse visit per operation then suffices. This count concerns visits to graph nodes. The cost of each visit depends on the primitive and the dimensions of its arguments. The local rule can apply the transposed derivative to a covector directly, without forming the matrix of the full network. Baydin et al. [33] discuss these forward and reverse calculations for computer programs.

40.4 A worked network with shared parameters

Use the same parameters $\theta = (w, b)$ to predict at two fixed inputs $x_1 = 1$ and $x_2 = 2$. The predictor is the smooth network

$$z_i = wx_i + b, \qquad u_i = z_i^2 \qquad (i = 1, 2).$$

The two branches share w, b, as in the parameter reuse of Chapter 39. Take target labels $y_1 = 3$ and $y_2 = 6$ and use the mean of the two half-squared losses. Write $r_i = u_i - y_i$ and set $L = (r_1^2 + r_2^2)/4$. With $P = \mathbb{R}^2$ for the parameter space and $f_i(\theta) = u_i$, the forward computation factors as

$$P \longrightarrow P \times P \xrightarrow{f_1 \times f_2} \mathbb{R}^2 \longrightarrow \mathbb{R}.$$

The loss ℓ takes the two predictions and returns their mean half-squared error. The diagonal represents one parameter node with two outgoing edges. At $w = b = 1$, the forward values are $(z_1, z_2) = (2, 3)$, $(u_1, u_2) = (4, 9)$, and $(r_1, r_2) = (1, 3)$, with $L = 5/2$.

Starting with seed 1 at L, the loss contributes $r_i/2$ to $\bar{u}_i$. The square pulls back $\bar{u}_i$ to $2z_i\bar{u}_i$ at z_i. The affine map then sends $\bar{z}_i$ to the parameter covector with coefficients $(x_i\bar{z}_i, \bar{z}_i)$. The exact values are as follows.

i	r_i	$\bar{u}_i$	$\bar{z}_i$	$(x_i\bar{z}_i, \bar{z}_i)$
1	1	1/2	2	$(2, 2)$
2	3	3/2	9	$(18, 9)$

Adding the two contributions gives $(20, 11)$. If the branches had separate parameters, these would have been covectors on distinct parameter factors, with

components $(2, 2)$ and $(18, 9)$. Pullback along the parameter diagonal gives their sum, the covector for the shared parameters. To verify directly, let (h, k) be a parameter variation. The two affine outputs vary by $h + k$ and $2h + k$. Differentiating the scalar loss gives

$$DL_{(1,1)}(h, k) = 2(h + k) + 9(2h + k) = 20h + 11k.$$

40.5 Differentials and gradient updates

The calculation gives a linear map $DL_\theta \colon E \to \mathbb{R}$ on parameter variations. A gradient vector requires an inner product $\langle -, - \rangle_G$ on E. It is the unique vector $\nabla_G L(\theta)$ with

$$DL_\theta(v) = \langle \nabla_G L(\theta), v \rangle_G \qquad (v \in E).$$

Existence and uniqueness follow because a positive-definite inner product identifies E with E^*, by the linear map sending a vector to its pairing with arguments. In coordinates with Gram matrix G, if c is the column of coefficients of the covector, the column of the gradient is $G^{-1}c$. For the worked network the Euclidean gradient is $(20, 11)$, while for $G = \mathrm{diag}(4, 1)$ the gradient is $(5, 11)$. Both gradients represent the same covector with their respective inner products.

For a nonzero gradient, its negative direction has directional derivative $-\langle \nabla_G L, \nabla_G L \rangle_G < 0$. The first-order expansion then gives a strict decrease for sufficiently small positive steps along that direction, while staying inside the open parameter domain. The reverse sweep computes the differential at the current parameters. Chapter 41 uses it to define a state update and studies how such updates compose over multiple steps.

CHAPTER 41

Optimization as a state update

A loss assigns a value to each parameter. Changing the parameter requires another map. The gradient method builds this map from a first-order description of the loss and a choice of distance. An optimizer may also keep a record of previous gradients or draw new data at each step. These choices determine the state of the optimization and how its successive updates compose.

41.1 Turning a differential into a step

Let $P = \mathbb{R}^d$, and let $L\colon P \to \mathbb{R}$ be differentiable. Its differential DL_w at w is a linear functional on parameter increments. The Euclidean gradient $\nabla L(w)$ represents that functional by the dot product: $DL_w(v) = \langle \nabla L(w), v\rangle$. The reverse sweep of Chapter 40 supplies the functional. An inner product identifies it with a vector.

For a stepsize $\eta > 0$, consider the following function of a candidate increment v.

$$m_w(v) = L(w) + DL_w(v) + \frac{\|v\|^2}{2\eta}.$$

The first two terms describe the first-order approximation of L around w. The last term charges for moving. Completing the square gives

$$m_w(v) = L(w) - \frac{\eta}{2}\|\nabla L(w)\|^2 + \frac{1}{2\eta}\|v + \eta\nabla L(w)\|^2.$$

Its unique minimizer is $v = -\eta\nabla L(w)$. This derives the update $G_\eta(w) = w - \eta\nabla L(w)$ from the differential and the quadratic cost.

The construction depends on the chosen inner product. To see the effect on a change of coordinates, write $w = Mq$ for an invertible real matrix M. The chain rule gives $\nabla_q(L \circ M)(q) = M^{\mathsf{T}}\nabla L(Mq)$ using the Euclidean inner product on the q-coordinates. Taking that Euclidean step and mapping it to the w-coordinates gives the update $\widetilde{G}_\eta$ in the following square.

$$\begin{array}{ccc} \mathbb{R}^d_q & \xrightarrow{G_\eta^{L\circ M}} & \mathbb{R}^d_q \\ \downarrow{\scriptstyle M} & & \downarrow{\scriptstyle M} \\ \mathbb{R}^d_w & \xrightarrow{\widetilde{G}_\eta} & \mathbb{R}^d_w \end{array} \qquad \widetilde{G}_\eta(w) = w - \eta M M^{\mathsf{T}}\nabla L(w).$$

For an orthogonal M, the transformed step is G_η. For arbitrary invertible M, the same agreement holds if the quadratic cost in the q-coordinates is $\|Mv\|^2/(2\eta)$ for an increment v. Indeed, minimizing in v then gives the original quadratic problem after substituting Mv for the increment. The differential obeys the chain rule in either description.

41.2 From a model decrease to a loss decrease

Suppose that L is continuously differentiable and its gradient satisfies $\|\nabla L(w') - \nabla L(w)\| \le K\|w' - w\|$ for some $K > 0$ and all w, w'. Integrating along the segment from w to $w + v$ gives

$$\begin{aligned} L(w+v) - L(w) - DL_w(v) &= \int_0^1 \langle \nabla L(w+tv) - \nabla L(w), v\rangle \, dt \\ &\le \tfrac{K}{2}\|v\|^2. \end{aligned}$$

The bound follows because the integrand is at most $Kt\|v\|^2$. Inserting $v = -\eta\nabla L(w)$ yields

$$L(G_\eta(w)) \le L(w) - \eta(1 - K\eta/2)\|\nabla L(w)\|^2.$$

Thus each step with $0 < \eta < 2/K$ lowers the loss at a point with nonzero gradient. If L is bounded below by L_*, summing these inequalities over T steps gives

$$\min_{0 \le t < T} \|\nabla L(w_t)\|^2 \le \frac{L(w_0) - L_*}{T\eta(1 - K\eta/2)}.$$

Indeed the sum of the decreases telescopes and is at most $L(w_0) - L_*$. This is a bound on a stationarity measure. For convergence of parameters to a minimizer we can prove more when the loss has additional structure.

Example 41.1 (A quadratic with a known rate)**.** Let A be a real symmetric positive definite $d \times d$ matrix, and set $L(w) = \frac{1}{2}\langle w - w_*, A(w - w_*)\rangle$. The differential gives $\nabla L(w) = A(w - w_*)$, so

$$w_{t+1} - w_* = (I - \eta A)(w_t - w_*).$$

In an orthonormal eigenbasis of A, each component of the error is multiplied at each step by $1 - \eta\lambda_i$, where $\lambda_i > 0$ is the corresponding eigenvalue. If $0 < \eta < 2/\max_i \lambda_i$ and $r = \max_i |1 - \eta\lambda_i|$, then $\|w_t - w_*\| \le r^t\|w_0 - w_*\|$ with $r < 1$. For $A = \mathrm{diag}(1, 4)$, $w_* = 0$, and $\eta = 1/4$, the first coordinate shrinks by $3/4$ per step and the second vanishes on the first step.

41.3 Composing updates with memory

Let H be a set of auxiliary states, such as previous gradient estimates. Set $W = P \times H$, and let Z be a set of possible data inputs to one update. A deterministic step consuming an input has the form $U : W \times Z \to W$. Each $z \in Z$ specifies an endomorphism U_z of W. A sequence $(z_1, \dots, z_t)$ specifies $U_{z_t} \circ \cdots \circ U_{z_1}$. There is an identity update for the empty sequence and concatenation composes updates in execution order. The sequence-to-update map is the homomorphism from the free monoid of data words to endomorphisms with execution-order multiplication, by the universal property in Chapter 5.

For example, for $0 \le \beta < 1$, one version of momentum uses a gradient estimate $g(w, z)$ and stores $m' = \beta m + (1 - \beta) g(w, z)$. First V takes (w, m, z) to (w, m'). Then G takes (w, m') to $(w - \eta m', m')$.

$$\begin{array}{ccc} (P \times H) \times Z & \xrightarrow{V} & P \times H \\ & \searrow^{U} & \downarrow{\scriptstyle G} \\ & & P \times H \end{array}$$

Here $H = \mathbb{R}^d$. For the one-dimensional gradient $g(w) = w$, take $\eta = \beta = 1/2$ and $w = 2$. With $m = 0$ the next parameter is $3/2$. With $m = 2$ the next parameter is 1. The stored estimate is therefore part of the input to a deterministic prediction of the next parameter.

These updates have the same state-passing form as the computations of Chapter 33. A step with a returned diagnostic $a \in A$ would have output in $A \times W$. Taking $A = 1$ recovers an update of W alone. To incorporate random data, for each $w \in W$ we can push the law of z forward by $z \mapsto U(w, z)$. For a finite dataset this gives a row of a stochastic map $K : W \to \mathcal{D}W$ as in Chapter 20. With fresh independent draws at each step and a fixed update rule, the state law evolves by repeated Kleisli composition of K. For an evolving stepsize or other schedule, one can use kernels indexed by step, or include the step number in W.

41.4 Stochastic steps and their variance

Consider the empirical risk $L(w) = n^{-1} \sum_{i=1}^{n} \ell_i(w)$ of a fixed dataset with $n > 0$, as in Chapter 37. Write $g_i(w) = \nabla \ell_i(w)$ and $\bar{g}(w) = \nabla L(w)$. For a uniform random index I, the vector $g_I(w)$ has mean $\bar{g}(w)$. If the current parameter arose from earlier draws, this statement holds conditional on the current state when the next index is drawn independently of those draws.

For a batch of $b > 0$ independent uniform indices I_j, use $g_B(w) = b^{-1} \sum_{j=1}^{b} g_{I_j}(w)$. In the following calculation the parameter is fixed, or everything is conditional on its value before the new batch. Put $\sigma^2(w) =$

$n^{-1}\sum_i \|g_i(w)-\bar g(w)\|^2$. Independence and the zero means of $g_{I_j}(w)-\bar g(w)$ make every cross term in the squared norm of their sum have mean zero. Therefore

$$\mathbb{E}[g_B(w)] = \bar g(w), \qquad \mathbb{E}\|g_B(w)-\bar g(w)\|^2 = \sigma^2(w)/b.$$

The update with a batch and its distribution can be written as the composite below, using the sampled batch at the current parameter w.

$$\{1,\dots,n\}^b \xrightarrow{B\mapsto g_B(w)} \mathbb{R}^d \xrightarrow{v\mapsto w-\eta v} P$$

The first arrow sends each tuple of indices to its averaged gradient. Pushing the uniform product law along this composite gives the transition law of the parameter.

If L satisfies the gradient bound of Section 41.2, the same segment estimate applies to each batch. Taking its conditional mean gives

$$\mathbb{E}[L(w-\eta g_B(w))] \le L(w) - \eta(1-K\eta/2)\|\bar g(w)\|^2 + \frac{K\eta^2}{2b}\sigma^2(w).$$

Here the linear term uses the mean gradient, and $\mathbb{E}\|g_B(w)\|^2 = \|\bar g(w)\|^2+\sigma^2(w)/b$. The final term records the contribution of batch variance to the bound. The stochastic gradient method and refinements of its convergence estimates are treated by Bottou, Curtis, and Nocedal [34].

Example 41.2 (A mean step and a mean loss)**.** In one dimension take the two losses $\ell_y(w) = (w-y)^2/2$ for labels $y = -1, 3$. Their average is $L(w) = 2 + (w-1)^2/2$, so $K = 1$. At $w = 0$, the full gradient is -1 and a step with $\eta = 1/2$ gives $w' = 1/2$, with loss $17/8$. The initial loss is $5/2$.

For a batch of size one, the two possible steps give $w' = -1/2$ and $w' = 3/2$ with equal probabilities. Their mean is the full-gradient parameter $1/2$. Their losses are $25/8$ and $17/8$, whose mean is $21/8 > 5/2$. Here the variance term in the bound is exact, since the loss is quadratic. With a batch of size b, the next expected loss is $17/8 + 1/(2b)$, which is below $5/2$ for $b \ge 2$.

The draws used here are from the fixed dataset. A step's expected loss is therefore an average over the optimizer's randomness at fixed training data. Chapter 43 takes an average over the sampling of the dataset itself and relates empirical risk to risk on new observations.

CHAPTER 42

Bayesian learning from stochastic kernels

A Bayesian model starts with a state on a space of hypotheses and a kernel that assigns an observation law to each hypothesis. Conditioning turns these two pieces into a kernel from observations back to hypotheses. That kernel summarizes what the model learns from each observation. Composing it with the law of a future observation gives a prediction. These constructions use the same disintegration identity as Chapter 20. The choice of joint law determines what counts as new evidence and which data can be compressed.

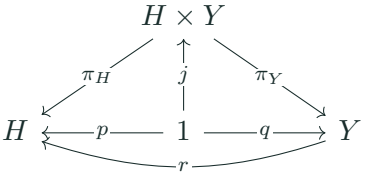
42.1 An experiment and its posterior

Let H be a finite nonempty hypothesis set and Y a finite observation set. Write $p\colon 1 \to H$ for a prior state and $k\colon H \to Y$ for an observation kernel. Following Section 20.4, the marginal $q = k \odot p$ on Y admits a Bayesian inverse $r\colon Y \to H$ relative to p. We call its row $r(- \mid y) = p_y$ the *posterior* at y. At any y with $q(y) > 0$, the defining identity and its solution are

$$j(h, y) = p(h)k(y \mid h) = q(y)p_y(h), \qquad p_y(h) = \frac{p(h)k(y \mid h)}{q(y)}.$$

The function of h given by $k(y \mid h)$ at the observed y is called the *likelihood*. Its sum over hypotheses need not be one. Normalization takes place after multiplying by the prior. In particular, the posterior gives positive weight exactly to hypotheses with positive prior weight and positive likelihood at y.

The joint state keeps both the hypothesis and the observation. Let π_H, π_Y also denote the deterministic kernels that project onto the two coordinates. Marginalization and disintegration give the equal routes from 1 to the two outputs below.



In particular, $r \odot q = p$. Thus a fixed prior is the average of the possible posterior states, with weights given by the model's observation distribution. If $q(y) = 0$, the

entire column $j(-, y)$ is zero. The factorization leaves the row p_y unconstrained there. Assigning a posterior to such an event requires information beyond this disintegration identity. Cho and Jacobs [35] develop disintegration and Bayesian inversion in categorical form.

The effect of changing the prior can be measured explicitly when there are two hypotheses h_0, h_1 with positive likelihoods at y. For a prior weight ρ on h_1, set $\lambda = k(y \mid h_1)/k(y \mid h_0)$. Bayes' rule gives

$$p_y(h_1) = \frac{\lambda\rho}{1 + (\lambda - 1)\rho}, \qquad \frac{p_y(h_1)}{p_y(h_0)} = \lambda \, \frac{\rho}{1 - \rho} \qquad (0 < \rho < 1).$$

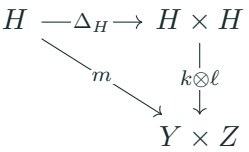
Observing y multiplies the prior odds by the likelihood ratio. The first formula makes the sensitivity to ρ explicit. Its derivative is $\lambda/[1+(\lambda-1)\rho]^2 > 0$. Different interior prior weights therefore give different posterior weights on the same data. At the endpoints, a zero prior weight remains zero.

42.2 Learning from a sequence

Suppose there are two experiments $k\colon H \to Y$ and $\ell\colon H \to Z$. Conditional independence means we sample their outcomes separately given the same h. More precisely, the prescribed kernel $m\colon H \to Y \times Z$ is

$$m(y, z \mid h) = k(y \mid h)\ell(z \mid h).$$

Using the tensor of stochastic maps from Chapter 20, we can either sample the pair together or copy the latent hypothesis and run the two experiments. The two paths are equal.



Copying h shares the hidden cause. After averaging over h, the two observations can be dependent even though the factorization holds at every fixed h.

Proposition 42.1 (Batch and sequential updates)**.** *For the product likelihood above, if (y, z) has positive joint probability, updating p by k at y and then updating by ℓ at z gives the same posterior as updating p once by m at (y, z).*

Proof. Let $q(y) = \sum_h p(h)k(y \mid h)$. The probability $q_m(y, z)$ of the pair is at most $q(y)$, so the first update is defined. With p_y as prior for the second experiment, its evidence is

$$\sum_{h \in H} p_y(h)\ell(z \mid h) = \frac{q_m(y, z)}{q(y)} > 0.$$

Substituting the formula for p_y into the second update and canceling $q(y)$ gives

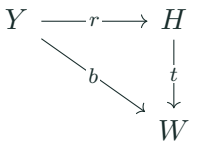
$$(p_y)_z(h) = \frac{p(h)k(y \mid h)\ell(z \mid h)}{q_m(y,z)}.$$

This is the disintegration of the original batch joint law at (y,z). □

The result extends by induction to finitely many experiments conditionally independent given h. At every history of positive probability, multiplying the successive likelihood factors and then normalizing yields the posterior. Permuting the factors together with their observations preserves that product. If the next experiment instead depends on the earlier outcome, use a kernel with row $\ell(- \mid h, y)$. The joint likelihood then uses $k(y \mid h)\ell(z \mid h, y)$, and the second update uses the kernel at the observed y. For instance, if the second report copies the first, observing the same report twice contributes just one original likelihood factor. Treating the reports as independent given h would square that factor and describe a different experiment.

42.3 Posterior predictive laws and decisions

Let $t\colon H \to W$ specify the future observation. Suppose the future outcome and the past data are conditionally independent given h. The predictive kernel $b\colon Y \to W$ is the composite in the upper route of the triangle.



$$b(w \mid y) = \sum_{h \in H} p_y(h)t(w \mid h).$$

Indeed the joint mass of y, w is $\sum_h p(h)k(y \mid h)t(w \mid h)$. Dividing by $q(y) > 0$ shows that $b(- \mid y)$ is its conditional law, as the triangle asserts. Averaging it over possible past data gives $b \odot q = t \odot p$, by associativity and $r \odot q = p$.

Suppose we choose an action a from a finite nonempty set of actions and incur loss $L(w,a)$ at outcome w. Conditional on the observed y, an action minimizing $\sum_w b(w \mid y)L(w,a)$ is a Bayes decision. The unconditional expected loss of a rule d with $d(y)$ the chosen action is

$$\sum_y q(y) \sum_w b(w \mid y)L(w, d(y)).$$

Thus choosing a separate minimizer for each row minimizes the total expected loss. This calculation uses the model distribution just as expected risk does in Chapter 37.

Example 42.2 (Predicting another success)**.** There are two possible coins. Under F the success probability is $1/2$, and under B it is $3/4$. Give each hypothesis prior weight $1/2$, and assume all tosses are independent given the coin. After one success the posterior weight of B is $3/5$. After two successes the likelihoods are $1/4$ and $9/16$, so the weight of B is $9/13$. The two predictions for the next success are therefore

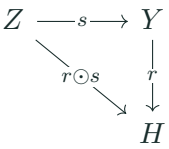
$$\frac{2}{5}\frac{1}{2} + \frac{3}{5}\frac{3}{4} = \frac{13}{20}, \qquad \frac{4}{13}\frac{1}{2} + \frac{9}{13}\frac{3}{4} = \frac{35}{52}.$$

The initial one-toss success probability is $5/8$. The joint probability of two successes is $13/32$, which exceeds $(5/8)^2 = 25/64$. This difference comes from sharing the coin across tosses.

Suppose predicting success incurs loss 2 when the next toss fails, while predicting failure incurs loss 1 when it succeeds. A correct prediction costs zero. At predictive success probability v, the two expected losses are $2(1 - v)$ and v, respectively. Predicting success is preferred exactly when $v > 2/3$. The numbers $13/20$ and $35/52$ lie on opposite sides of $2/3$, so the optimal decision changes at the second success. If the prior weight on B had been $1/4$, its posterior weight after two successes would be $3/7$ and the next-success probability would be $17/28$. That prior gives a different decision on the same observed pair.

42.4 Learning after coarsening the data

Suppose that Y is available in the model and a kernel $c\colon Y \to Z$ describes a report generated using only Y. We observe the report Z. Let $r\colon Y \to H$ be the inverse of k relative to p, and $s\colon Z \to Y$ the inverse of c relative to $q = k \odot p$. Put $q_Z = c \odot q$. The displayed composite is a Bayesian inverse of $c \odot k$ relative to p. Its value at positive $q_Z(z)$ is uniquely determined.



To prove this, use the two disintegration identities inside a sum.

$$\begin{aligned} q_Z(z) \sum_y s(y \mid z) r(h \mid y) &= \sum_y q(y) c(z \mid y) r(h \mid y) \\ &= \sum_y p(h) k(y \mid h) c(z \mid y) \\ &= p(h)(c \odot k)(z \mid h). \end{aligned}$$

This proof also works when some $q(y)$ are zero, since the corresponding summands vanish. The posterior given the report is a mixture of the posteriors for the

underlying data. Fritz [15] treats conditional independence and sufficient statistics in Markov categories.

For a deterministic report $S\colon Y \to Z$, each row of s at positive probability is supported on the fiber of S. If the posterior is constant on the positive-probability members of each fiber, then mixing does not change it. A sufficient condition for this constancy, for any prior, is a likelihood factorization

$$k(y \mid h) = a(y)b(S(y), h)$$

with $a(y)$ independent of h. At positive $q(y)$ the factor $a(y)$ cancels from the posterior. Its remaining expression depends on y only through $S(y)$. The same cancellation holds for densities with respect to a fixed reference measure on the hypothesis space. The next example uses this with $S(y)$ the number of successes.

42.5 A conjugate update for a continuum of coins

Let the hypothesis θ range over $[0, 1]$ and put a prior density on θ with respect to Lebesgue measure. For positive integers a, b, set

$$I(a, b) = \int_0^1 u^{a-1}(1-u)^{b-1}\,du, \qquad g_{a,b}(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{I(a, b)}.$$

This density is positive on the open interval and integrates to one. It is the beta prior with parameters a, b. Conditional on θ, a Bernoulli trial produces success with probability θ and failure with probability $1 - \theta$. We assume finitely many trials conditionally independent given θ.

For an ordered sequence y with s successes and f failures, the likelihood is $\theta^s(1-\theta)^f$. To compute its posterior density, require that multiplying the density by the probability of y recovers $g_{a,b}(\theta)\theta^s(1-\theta)^f$. Integrating over any measurable set of hypotheses recovers the joint probability for that set and y, so this is the disintegration identity with a density. Since the likelihood is positive on the open interval, the event y has positive probability. Division by its probability gives the posterior density

$$g_{a+s,b+f}(\theta).$$

If only the number of successes in a fixed number $n = s + f$ of trials is observed, its likelihood has the extra factor $\binom{n}{s}$ from the orders of those trials. That factor cancels. The posterior is therefore the same for the ordered sequence and for the count. Because updating stays in the same family of densities, this prior family is called *conjugate* to the Bernoulli likelihood.

The predictive probability of success in one further trial is the integral of θ against the posterior. For any positive integers A, B, inserting $\theta + (1 - \theta) = 1$

under the integral gives $I(A, B) = I(A + 1, B) + I(A, B + 1)$. Integrating the derivative of $\theta^A(1 - \theta)^B$ gives

$$AI(A, B + 1) - BI(A + 1, B) = 0.$$

The endpoints contribute zero. Combining the two equalities gives $I(A + 1, B)/I(A, B) = A/(A + B)$. Consequently the posterior predictive probability is

$$\int_0^1 \theta g_{a+s,b+f}(\theta)\, d\theta = \frac{a + s}{a + b + s + f}.$$

The categorical operation is the same as before. Compose the posterior state with the one-trial kernel and integrate over the latent hypothesis.

Example 42.3 (Updating a uniform density)**.** Take $a = b = 1$, so the prior density is uniform. The data $(1, 0, 1)$ give $s = 2, f = 1$. The new density is $g_{3,2}$ and the next-success probability is $3/5$. The probability of two future successes together is

$$\frac{I(5, 2)}{I(3, 2)} = \frac{I(5, 2)}{I(4, 2)} \frac{I(4, 2)}{I(3, 2)} = \frac{4}{6} \frac{3}{5} = \frac{2}{5}.$$

This exceeds $(3/5)^2 = 9/25$. The same number comes from predicting a first success with probability $3/5$, updating to $g_{4,2}$ if it occurs, and predicting the second with probability $4/6$. Both routes use the same latent coin for the two future trials.

Here every finite observation sequence has positive probability, although each individual θ has prior mass zero. In the standard Borel setting of Chapter 20, the disintegration identity determines the conditional kernel up to equality almost everywhere with respect to the observation law. For continuously distributed observations, the choice of version can affect its value at a specified observation of probability zero. In either setting, the learning calculation refers to the joint law supplied by the model.

CHAPTER 43

Generalization from sampling assumptions

An empirical risk compares predictors on observed data. A population risk compares them on a fresh draw from the law of the data. To relate the two, one must specify how observations were drawn. Independent sampling supplies a product distribution, and that product factors certain expectations. Bounded loss then converts the factorization into an error bound. Uniformity over a class of predictors allows the same bound to apply after a predictor has been selected from that class.

43.1 Sampling and the law of a score

Let Z be a finite set of possible examples, such as a finite set of input and label pairs. Let p be its population law and let $\mathcal{H}$ be a nonempty finite set of predictors. For each $h \in \mathcal{H}$, use the loss function $\ell_h \colon Z \to [0,1]$. The population risk and the risk on an observed tuple $S = (z_1, \dots, z_n)$ are

$$R_p(h) = \sum_{z \in Z} p(z)\ell_h(z), \qquad \widehat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell_h(z_i), \qquad n \geq 1.$$

These are expectations under p and the empirical law $\widehat{p}_S = \frac{1}{n}\sum_i \delta_{z_i}$, respectively, as in Chapter 37. The following diagram has two ways to obtain the same empirical score. Let avg_n average the coordinates and let $\mathbb{E}_{(-)}\ell_h$ take expectation against a distribution.

$$\begin{array}{ccc} Z^n & \xrightarrow{\ell_h^{\times n}} & [0,1]^n \\ \downarrow{\scriptstyle S \mapsto \widehat{p}_S} & & \downarrow{\scriptstyle \mathrm{avg}_n} \\ \mathcal{D}Z & \xrightarrow{\mathbb{E}_{(-)}\ell_h} & [0,1]. \end{array}$$

The upper path evaluates the loss separately on each observation. On the lower path repeated observations first contribute to the same empirical mass. Distributing the sum of losses over these masses gives the upper path.

For independent draws from p, the sampling state is $p^{\otimes n} \colon 1 \to Z^n$ in the stochastic category of Chapter 20. Its mass at $(z_1, \dots, z_n)$ is $\prod_i p(z_i)$. If each

coordinate is scored by the same loss map, the resulting score state is the tensor of n copies of the one-draw score law $(\mathcal{D}\ell_h)(p)$. In particular,

$$\mathbb{E}_{p^{\otimes n}} \prod_{i=1}^{n} v_i(Z_i) = \prod_{i=1}^{n} \mathbb{E}_p v_i(Z),$$

for any real functions v_i on Z. Expanding the expectation as a sum over Z^n proves this by distributivity. We will apply the identity to exponential functions of the losses. Linearity alone gives $\mathbb{E}\widehat{R}_S(h) = R_p(h)$ for any joint sample law whose marginals are all p. The product law gives the extra factorization used in the deviation estimate.

43.2 A bound for one predictor

Consider first a single random variable W in $[0,1]$, with mean u. Convexity of the exponential gives the chord inequality $e^{tw} \le 1 - w + we^t$ for $0 \le w \le 1$ and $t \ge 0$. Taking expectations and centering yields

$$\log \mathbb{E}e^{t(W-u)} \le b(t), \qquad b(t) = \log(1 - u + ue^t) - ut.$$

The derivative of the logarithm is $q(t) = ue^t/(1-u+ue^t)$. Hence $b(0) = b'(0) = 0$ and $b''(t) = q(t)(1-q(t)) \le 1/4$, because $0 \le q(t) \le 1$. Integrating this bound twice from zero gives $b(t) \le t^2/8$. This proof also covers $u = 0$ and $u = 1$, where $b(t) = 0$.

Theorem 43.1 (Bounded independent losses)**.** *If S has law $p^{\otimes n}$, then for a fixed predictor h and $\epsilon > 0$,*

$$\Pr\{|\widehat{R}_S(h) - R_p(h)| \ge \epsilon\} \le 2e^{-2n\epsilon^2}.$$

Proof. Put $W_i = \ell_h(Z_i)$ and $u = R_p(h)$. The variables W_i are independent since each is a function of its own draw. For the event that $\widehat{R}_S(h) - u \ge \epsilon$, the random variable $e^{t\sum_i (W_i - u)}$ is at least $e^{tn\epsilon}$ for each $t > 0$. A nonnegative random variable V satisfies $\Pr\{V \ge a\} \le \mathbb{E}V/a$ for $a > 0$, since the contribution of that event to its expectation is at least $a \Pr\{V \ge a\}$. Here this gives

$$\begin{aligned}
\Pr\{\widehat{R}_S(h) - u \ge \epsilon\} &\le e^{-tn\epsilon} \mathbb{E} e^{t\sum_i (W_i - u)} \\
&= e^{-tn\epsilon} \prod_i \mathbb{E} e^{t(W_i - u)} \\
&\le \exp\bigl(n(t^2/8 - t\epsilon)\bigr).
\end{aligned}$$

Take $t = 4\epsilon$ to obtain $e^{-2n\epsilon^2}$. Applying the same argument to the losses $1 - W_i$ bounds the lower tail by the same number. The union of the two events gives the stated result. □

This is the two-sided Hoeffding bound for losses in a unit interval [36]. If all losses instead lie in $[a, b]$ with $a < b$, applying the estimate to $(\ell_h - a)/(b - a)$ replaces ϵ in the exponent by $\epsilon/(b - a)$. For squared loss on unbounded inputs, this particular bound requires an additional boundedness condition.

43.3 Selecting from a finite class

The fixed-predictor theorem assumes that the same h is used throughout its probability calculation. To use a predictor selected on S, first bound the event on which any member of the class has a large deviation. The probability of a union is at most the sum of the probabilities because the sum counts each point of the union at least once. Hence

$$\Pr\left\{\max_{h\in\mathcal{H}} |\widehat{R}_S(h) - R_p(h)| \geq \epsilon\right\} \leq 2|\mathcal{H}|e^{-2n\epsilon^2}.$$

The loss estimates for two different predictors on the same data can be dependent. The union bound requires only a separate bound for each predictor. For $0 < \delta < 1$, it follows that with probability at least $1 - \delta$ the maximum error is at most

$$c_n = \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2n}}.$$

Let $\widehat{h}(S)$ be an empirical risk minimizer, chosen by a fixed tie rule, and let h_* minimize R_p over $\mathcal{H}$. On the event just identified, the two deviations and the defining inequality of empirical risk minimization give

$$\begin{aligned} R_p(\widehat{h}(S)) &\leq \widehat{R}_S(\widehat{h}(S)) + c_n \\ &\leq \widehat{R}_S(h_*) + c_n \\ &\leq R_p(h_*) + 2c_n. \end{aligned}$$

Both comparisons to population risk matter. One involves the chosen predictor and one involves the population minimizer. The argument also applies if the algorithm attains empirical risk within τ of the minimum. Then the middle comparison contributes τ, and the excess population risk is at most $2c_n + \tau$. Thus, for exact empirical minimization, it suffices to take

$$n \geq \frac{2}{\varepsilon^2}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

to bound excess population risk by ε with probability at least $1-\delta$. This compares with the best member of the given class. Approximation to a predictor outside the class is a separate question about the class and the population law.

Example 43.2 (Four classifiers on two inputs)**.** Let the inputs be a, b and labels be $0, 1$. The class contains the four possible functions. Use 0-1 loss. Use the following population masses and counts from one sample of size ten.

z	$(a,0)$	$(a,1)$	$(b,0)$	$(b,1)$
$p(z)$	$9/20$	$1/20$	$3/20$	$7/20$
observed count	4	1	3	2

The possible risks, indexed by the pair of predictions, are

$(h(a), h(b))$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$R_p(h)$	$2/5$	$1/5$	$4/5$	$3/5$
$\widehat{R}_S(h)$	$3/10$	$4/10$	$6/10$	$7/10$

For example $(0,0)$ errs on the two positive-label outcomes, of population mass $1/20 + 7/20$ and sample count $1 + 2$. The sample selects $(0,0)$. The population minimizer is $(0,1)$, so the excess on this sample is $1/5$. For a guarantee of excess at most $1/10$ with probability at least 0.95 under independent sampling, the bound gives the sufficient sample size $\lceil 200 \log 160 \rceil = 1016$. It makes a claim across repeated draws of the full sample.

43.4 Identical marginals and shared draws

Compare $p^{\otimes n}$ with the sample obtained by drawing once from p and copying the result. Write $\Delta_n(z) = (z, \ldots, z)$. The copied law is $\mathcal{D}\Delta_n(p)$. Every coordinate projection π_i takes this law back to p by the following diagram in sets.

$$\begin{array}{ccc} Z & \xrightarrow{\Delta_n} & Z^n \\ & \searrow^{\mathrm{id}_Z} & \downarrow{\scriptstyle \pi_i} \\ & & Z. \end{array}$$

One draw with loss 0 or 1 each of probability $1/2$ already shows the difference. Under copying, the empirical risk is always that single loss, so its error from $1/2$ is $1/2$ with probability one for every n. For two such scores, $\mathbb{E}(W_1 W_2) = 1/2$ under copying, while $\mathbb{E}W_1 \mathbb{E}W_2 = 1/4$. A large tuple can give the same evidence as one observation.

There is a related distinction for a conditional sampling kernel $k \colon U \to \mathcal{D}Z$. Drawing a shared parameter u and then sampling n times from $k(u)$ gives a mixture of products, with mass $\sum_u r(u) \prod_i k(u)(z_i)$ for parameter law r. Drawing parameters independently for each observation gives a product of mixtures, with mass $\prod_i \sum_u r(u) k(u)(z_i)$. Conditional independence suffices for the preceding bound about the conditional risk $R_{k(u)}$ at fixed u. To compare that risk with a population averaged over r, one needs a bound on the variation of the conditional risks, or a different sampling design.

43.5 Validation after training

Suppose a rule A selects a predictor from a training sample $T \in Z^n$. Draw a validation sample $V \in Z^m$ independently of T, with law $p^{\otimes m}$. The evaluation first applies the rule and then scores the resulting predictor on V. The right vertical map below sends (h, V) to $\widehat{R}_V(h)$.

$$\begin{array}{ccc} Z^n \times Z^m & \xrightarrow{A\times \mathrm{id}} & \mathcal{H} \times Z^m \\ & \searrow^{\text{validate}} & \downarrow{\scriptstyle \text{score}} \\ & & [0,1]. \end{array}$$

Condition on a possible training tuple $T = t$ of positive probability. Then $A(t)$ is fixed and the validation sample still has law $p^{\otimes m}$. The fixed-predictor bound applies to this conditional experiment. Averaging the bound over t gives

$$\Pr\{|\widehat{R}_V(A(T)) - R_p(A(T))| \geq \epsilon\} \leq 2e^{-2m\epsilon^2}.$$

This calculation uses independence of the split. A training rule with an independent random seed can be treated by conditioning on the seed as well.

A validation set can also be used for selection. If M candidates were fixed by the training data before validation, condition on training and use the uniform bound with class size at most M. The fixed-predictor bound alone does not apply to that selected candidate. For example, take a fair binary label and two predictors that always predict 0 and 1, respectively. With a single validation observation, choosing the predictor that matches its label gives validation risk zero every time. Each candidate has population risk $1/2$. A fresh independent test of the final selection again admits the conditional fixed-predictor argument.

43.6 Changing the population law

The comparison so far uses one distribution for both sampling and population risk. For a different deployment law q on Z, put

$$d_{\mathrm{TV}}(p, q) = \frac{1}{2} \sum_z |p(z) - q(z)|.$$

Because the masses of p and q each sum to one, the positive and negative portions of $p - q$ have equal mass $d_{\mathrm{TV}}(p, q)$. A loss in $[0, 1]$ weights each portion by an amount between zero and its full mass. Thus $|R_p(h) - R_q(h)| \leq d_{\mathrm{TV}}(p, q)$ for every predictor. On the uniform event of Section 43.3,

$$|\widehat{R}_S(h) - R_q(h)| \leq c_n + d_{\mathrm{TV}}(p, q) \quad (h \in \mathcal{H}).$$

Replacing the population minimum by its counterpart under q in the empirical minimization argument gives $R_q(\widehat{h}) - \min_{h\in\mathcal{H}} R_q(h) \le 2c_n + 2d_{\mathrm{TV}}(p,q)$.

The distance is compatible with stochastic composition. If the same kernel k acts on both laws, the two resulting laws are $k \odot p$ and $k \odot q$. The composition formula of Chapter 20 and the triangle inequality give

$$\begin{aligned} 2d_{\mathrm{TV}}(k\odot p, k\odot q) &= \sum_w \left| \sum_z k(w \mid z)(p(z) - q(z)) \right| \\ &\le \sum_z |p(z) - q(z)| \sum_w k(w \mid z) = 2d_{\mathrm{TV}}(p,q). \end{aligned}$$

One instance is the deterministic map ℓ_h into its finite set of score values. Passing from examples to scores cannot increase the distance between the laws. The risk comparison then applies to the identity function on the score values.

Example 43.3 (The optimal classifier changes with the law)**.** Return to the four classifiers of Example 43.2. Keep the two probability masses at a and exchange the masses at $(b,0)$ and $(b,1)$. The change at each of the two b-outcomes has magnitude $4/20$, so $d_{\mathrm{TV}}(p,q) = 1/5$. Under q, the classifier $(0,0)$ has risk $1/5$ and is optimal. The classifier $(0,1)$, optimal under p, has risk $2/5$ under q. Its risk change attains the total variation bound. Even knowledge of the exact training law p would leave that deployment excess of $1/5$ for its optimal classifier.

43.7 Combinatorics of shattering and learning

A binary predictor is determined by its subset of positive inputs. Identify a nonempty hypothesis class $\mathcal{H}$ on X with a family of subsets of X. On a finite set $B \subseteq X$, the trace $\mathcal{H}|_B = \{H \cap B \mid H \in \mathcal{H}\}$ records the labelings the class realizes there. Trace restriction along inclusions composes, so $B \mapsto \mathcal{H}|_B$ is a presheaf on the poset of finite subsets of X. We say $\mathcal{H}$ *shatters* B when its trace is the whole powerset of B. Its Vapnik–Chervonenkis dimension is the supremum of the sizes of shattered finite subsets.

Theorem 43.4 (Sauer–Shelah bound)**.** *If $\mathcal{H}$ has Vapnik–Chervonenkis dimension at most d, then for every m-element set $B \subseteq X$,*

$$|\mathcal{H}|_B| \le \sum_{j=0}^{\min(d,m)} \binom{m}{j}.$$

Proof. For any family $\mathcal{F} \subseteq \mathcal{P}(B)$, we prove that the number of its shattered subsets is at least $|\mathcal{F}|$. This holds for $B = \varnothing$, where the family is empty or contains just $\varnothing$. Otherwise choose $b \in B$ and delete it from each member of $\mathcal{F}$. Let U be the resulting image family, and let $D \subseteq U$ consist of traces with

both extensions in $\mathcal{F}$, one with b and one without. Counting the fibers gives $|\mathcal{F}| = |U| + |D|$.

Any set shattered by U is shattered by $\mathcal{F}$ and omits b. If C is shattered by D, then $\mathcal{F}$ shatters $C \cup \{b\}$, since each labeling of C can be extended with either label at b. These two contributions are disjoint, so induction on $B \setminus \{b\}$ gives at least $|U| + |D|$ shattered subsets. Apply this to $\mathcal{F} = \mathcal{H}|_B$. It can shatter only subsets of B of size at most d, which gives the bound. □

The bound is attained on B by taking all subsets of size at most d. Sauer and, independently, Perles and Shelah obtained this bound [71, 75].

For a learning application, fix inputs $x_1, \dots, x_n$ and draw the labels Y_i independently. Let $\widehat{R}(h)$ be the mean 0-1 loss and $R(h)$ its expectation over these labels. Predictors with the same trace on $\{x_1, \dots, x_n\}$ give the same losses and risks. The proof of Theorem 43.1 applies with each loss centered at its own mean, since the exponential estimate is uniform in that mean. Applying the union bound over traces yields

$$\Pr\{\sup_{h \in \mathcal{H}} |\widehat{R}(h) - R(h)| \ge \epsilon\} \le 2 e^{-2n\epsilon^2} \sum_{j=0}^{\min(d,n)} \binom{n}{j}.$$

For a fixed design, the combinatorial count replaces the cardinality of the full hypothesis class in the risk estimate.

Part IX Sheaves, descent, schemes, and classifying toposes

CHAPTER 44

Sheaves from compatible local data

A continuous function can be restricted to smaller open sets. It can also be recovered from its restrictions on a cover, provided that those restrictions agree where their domains overlap. A presheaf specifies the restriction maps. A sheaf specifies the additional test that identifies a section on the covered open with a compatible family of local sections. Taking germs gives another description of the same test.

44.1 Restriction and gluing

Let X be a topological space. Use the order category $\mathcal{O}(X)$ of Chapter 17, with one arrow $V \to U$ when $V \subseteq U$. A presheaf of sets is a functor $F\colon \mathcal{O}(X)^{\mathrm{op}} \to \mathbf{Set}$. Its elements on U are called sections over U. Write $s|_V$ for the restriction of $s \in F(U)$ to V. Functoriality gives $s|_U = s$ and $(s|_V)|_W = s|_W$ for $W \subseteq V \subseteq U$. A map of presheaves is a natural transformation, so a map on sections commutes with every restriction.

Suppose $U = \bigcup_{i \in I} U_i$. The restrictions of any section on U agree when both are restricted to $U_i \cap U_j$. The converse asks whether this is the only condition on a family of local sections. Form the parallel maps in the diagram by restricting the i-th and j-th components, respectively.

$$F(U) \longrightarrow \prod_i F(U_i) \rightrightarrows \prod_{i,j} F(U_i \cap U_j)$$

Here $\rho(t) = (t|_{U_i})_i$. The equality $r\rho = s\rho$ follows from functoriality. The equalizer of r, s consists precisely of all compatible families, since equality in the target is componentwise. As in Chapter 3, a map into this equalizer amounts to a family of such tests parametrized by the source of the map. The sheaf condition is that ρ exhibit $F(U)$ as this equalizer for every open cover.

Equivalently, each compatible family has exactly one section restricting to it. Injectivity of ρ says that agreement on a cover implies equality. A presheaf satisfying this injectivity test for every cover is called *separated*. The empty cover of $\varnothing$ requires a sheaf to have exactly one section there, because both empty products in the diagram are singletons.

Example 44.1 (Functions and open inclusions)**.** Fix a space Y, and let $F(U)$ consist of the continuous maps $U \to Y$. Given compatible functions $f_i \colon U_i \to Y$, define $f(x) = f_i(x)$ for any i with $x \in U_i$. Compatibility makes the value independent of the choice of i. For an open $B \subseteq Y$, the inverse image $f^{-1}(B)$ is the union of the opens $f_i^{-1}(B)$ in U_i, which are also open in U. This proves continuity. Equality on all U_i determines the function, so this presheaf is a sheaf.

For a fixed open V, the representable presheaf $h_V(U) = \mathrm{Hom}_{\mathcal{O}(X)}(U, V)$ is also a sheaf. Each section set has at most one element. It has one just when $U \subseteq V$. Since a union lies in V if and only if each member lies in V, the equalizer condition holds. This description will reappear when covers are specified on categories other than $\mathcal{O}(X)$.

Example 44.2 (From constant data to locally constant sections)**.** Let $A = \{0, 1\}$. Give each nonempty open the set A, with identity restrictions between nonempty opens, and give the empty open one section. This defines the normalized constant presheaf P_A. The ordinary constant presheaf has value A on the empty open as well. For the space $\{-1, 1\}$ with the discrete topology, cover by the two singleton opens. The intersection is empty, so the prescription 0 on $\{-1\}$ and 1 on $\{1\}$ is compatible. Every element of $P_A(\{-1, 1\})$ restricts to the same value on both singletons. Thus P_A is separated in this example and fails the existence part of gluing.

Continuous maps to A with its discrete topology are exactly locally constant functions. They form a sheaf by the preceding example and admit the constant sections as a presheaf map from P_A. In Section 44.4 we identify this map by its universal property, for any space X and any set A.

44.2 The stalk at a point

The local behavior of a section at $x \in X$ is measured by its germ. The neighborhoods of x form a category of restrictions, with an arrow $U \to V$ when $x \in V \subseteq U$ and both are open. This category is directed toward smaller neighborhoods. Any two have a common successor given by their intersection. The stalk F_x is the colimit of the sets $F(U)$ over this category. Explicitly, a germ $[s]_x$ is represented by $s \in F(U)$ for an open U containing x. Two representatives define the same germ if they agree on some open neighborhood contained in both domains. This is an equivalence relation since neighborhoods can be intersected. Compatible maps from all $F(U)$ to a fixed set assign the same value to equivalent representatives and hence factor uniquely through F_x. This verifies the colimit property.

A presheaf map $\varphi \colon F \to G$ induces $\varphi_x \colon F_x \to G_x$ by sending $[s]_x$ to $[\varphi_U(s)]_x$. Naturality guarantees that equivalent representatives have equivalent images. For sheaves, the stalk maps detect isomorphisms.

Proposition 44.3. *A map of sheaves is an isomorphism if and only if every stalk map is a bijection.*

Proof. An inverse on sheaves induces inverses on all stalks. Conversely, suppose the stalk maps are bijections. If two sections s, t over U have the same image, then $[s]_x = [t]_x$ at every $x \in U$. Choose open neighborhoods on which the representatives agree. These cover U, so sheaf uniqueness gives $s = t$. Thus φ_U is injective on each open.

For $y \in G(U)$ and $x \in U$, use surjectivity on stalks to lift $[y]_x$ to the germ of some local F-section. Shrink its domain to a neighborhood in U on which its image is the restriction of y. On the overlap of any two domains, the two lifts have the same image and are equal by the injectivity just proved. The lifts glue to a section of $F(U)$, and the image is y by sheaf uniqueness. The component bijections give an inverse natural transformation. □

For the sheaf of continuous real-valued functions, a germ records more than a value. Evaluation at a point gives a well-defined map $F_x \to \mathbb{R}$, since equality on a neighborhood implies equality at x. At $x = 0$ in $\mathbb{R}$, the constant zero function and $t \mapsto t$ have value zero and distinct germs.

44.3 A space of germs

For a presheaf F, form the set $E_F = \coprod_{x \in X} F_x$ with projection $\pi_F \colon E_F \to X$. Use as basic opens the sets

$$B(U, s) = \{[s]_x \in E_F \mid x \in U\}, \qquad s \in F(U).$$

Every germ lies in such a set. If a germ lies in $B(U, s) \cap B(V, t)$, equality of germs supplies a neighborhood $W \subseteq U \cap V$ of its base point with $s|_W = t|_W$. Thus $B(W, s|_W)$ is contained in the intersection. These sets define a topology. For V open in X, the intersection of $\pi_F^{-1}(V)$ with $B(U, s)$ is $B(U \cap V, s|_{U \cap V})$, so π_F is continuous. It is open because it takes each basic open onto its domain. Consequently its restriction to $B(U, s)$ is a homeomorphism onto U, with inverse $x \mapsto [s]_x$.

A continuous map that gives a homeomorphism onto an open of X on a neighborhood of every point of its domain is called a local homeomorphism, or an *étale space* over X. For such a map $p \colon E \to X$, let $\Gamma_p(U)$ consist of continuous sections of p over U. Each section u makes the triangle commute.

$$\begin{array}{ccc} U & \xrightarrow{\;u\;} & E \\ & \searrow^{i_U} & \downarrow{\scriptstyle p} \\ & & X \end{array}$$

These sections form a sheaf. Gluing the compatible continuous maps $U_i \to E$ gives a unique continuous map $U \to E$ as in the example of functions. The equation with p can be checked on the cover, so the glued map is again a section.

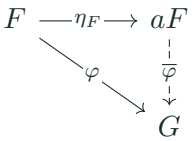
Conversely, recover E from the germs of Γ_p. The map $E_{\Gamma_p} \to E$ sends the germ of a local section at x to its value over x. Every point e of E is such a value, by inverting p on a neighborhood of e. If two sections have value e at x, continuity lets us restrict both to a neighborhood of x where their values land in the same local chart of p. They are then equal there because p is injective on that chart. Thus the map is bijective. On every basic open $B(U, u)$ it is identified with u by projection to U. A local section u maps small neighborhoods homeomorphically onto opens in E. Indeed, choose a chart at the value of u and restrict to the open preimage of that chart. There u is the inverse of the chart projection. This proves that evaluation of germs is a homeomorphism over X.

44.4 Sheafification and its universal property

Define $aF = \Gamma_{\pi_F}$. Every section s of F on U gives a section of aF on U by $x \mapsto [s]_x$. These maps define a natural transformation $\eta_F \colon F \to aF$. In the other direction, every $q \in aF(U)$ is locally of this form. Choose a basic open $B(V, s)$ containing $q(x)$. The preimage of this basic open along q is an open neighborhood W of x contained in $U \cap V$. Over W, the value $q(y)$ lies in $B(V, s)$ and projects to y. There is only one such germ, so $q|_W = \eta_F(s|_W)$.

Theorem 44.4. *For every sheaf G, a map $F \to G$ of presheaves extends uniquely along η_F to a map $aF \to G$.*

Proof. Given $\varphi \colon F \to G$ and $q \in aF(U)$, choose a cover of U on which q is given by F-sections s_i as above. For any point in an overlap, the two representatives have the same germ in F. On a neighborhood of that point the representatives agree, so their images under φ agree there. Sheaf uniqueness in G makes the images agree on the entire overlap. Thus the images $\varphi(s_i)$ glue to a unique section of $G(U)$. The same argument on intersections of two covers shows the result is independent of the choices. Restricting a cover to a smaller open shows that the construction commutes with restriction. The map $\overline{\varphi}$ obtained this way extends φ, by using a single representative when $q = \eta_F(s)$. Any extension must give the sections $\varphi(s_i)$ on the local representatives, so uniqueness of gluing forces its value.



□

This identifies a as left adjoint to the inclusion of sheaves into presheaves. The extension of $\eta_{F'} \circ \varphi$ defines $a\varphi\colon aF \to aF'$ for a presheaf map $\varphi\colon F \to F'$. Uniqueness gives the identity and composition rules. The stalk map $(\eta_F)_x$ is bijective for any presheaf F. Its inverse takes the germ of a local section of π_F at x to the value of that section in the fiber F_x. Local representability and equality of germs give the two inverse equations.

If F is already a sheaf, Proposition 44.3 therefore makes η_F an isomorphism. Together with the germ evaluation map for Γ_p, this gives an equivalence between sheaves and local homeomorphisms over X. A presheaf map acts on germ spaces by $[s]_x \mapsto [\varphi(s)]_x$. This map is continuous because on every basic open it is the continuous section determined by $\varphi(s)$ after projection to U. In the other direction a continuous map over X acts by composition on sections. The germ and section comparisons commute with these maps. Mac Lane and Moerdijk [16, Chapter II] develop both descriptions.

For P_A, each germ at x is just an element of A. Its basic opens have the form $U \times \{a\}$. Thus E_{P_A} is $X \times A$ with the product topology for discrete A, and aP_A is the sheaf of locally constant functions to A. The ordinary constant presheaf has the same germs and hence the same sheafification.

44.5 Local square roots on a circle

Let $X = S^1 = \{z \in \mathbb{C} : |z| = 1\}$ and $p\colon S^1 \to S^1$ be $p(z) = z^2$. On the open $U = S^1 \setminus \{1\}$, choose the argument $0 < \theta < 2\pi$. The function $s_U(e^{i\theta}) = e^{i\theta/2}$ is a continuous section. On $V = S^1 \setminus \{-1\}$ use $-\pi < \theta < \pi$ to define s_V by the same formula. On each arc the two inverse branches $s_U, -s_U$ or $s_V, -s_V$ are homeomorphisms to their open images. This shows that p is a local homeomorphism.

The two chosen local sections agree on the upper component of $U \cap V$. On the lower component their arguments differ by 2π, so $s_U = -s_V$ there. Each of the two arcs is connected. On either arc a second root divided by the chosen root is a continuous function to $\{1, -1\}$, hence a constant sign. Therefore every local section on U is ϵs_U for some $\epsilon \in \{1, -1\}$, and every local section on V is δs_V for some $\delta \in \{1, -1\}$. Use s_U as the reference root on both overlap components. The restrictions of ϵs_U and δs_V then have sign pairs $r_U(\epsilon, \delta) = (\epsilon, \epsilon)$ and $r_V(\epsilon, \delta) = (\delta, -\delta)$, respectively.

$$\Gamma_p(X) \longrightarrow \{1, -1\}^2 \underset{r_V}{\overset{r_U}{\rightrightarrows}} \{1, -1\}^2$$

Equality requires both $\epsilon = \delta$ and $\epsilon = -\delta$, so the equalizer is empty. The sheaf Γ_p has two germs over each point and no global section.

The test used here involves the cover and the two maps assigning restrictions to each overlap. The next chapter describes how to specify these tests when the objects replacing opens form a general category.

CHAPTER 45

Grothendieck toposes from sites

An open cover specifies the domains on which data can be given before they are glued. The same gluing test can use arrows other than inclusions. A map $V \to U$ restricts data on U to data on V, and all further restrictions are controlled by composition. A site records which collections of such tests are sufficient to recover a section.

45.1 Sieves and their pullbacks

Fix a small category $\mathcal{C}$. Write yX for the presheaf $\mathrm{Hom}_{\mathcal{C}}(-, X)$. A *sieve* R on X is a subpresheaf of yX. In arrows, this means that if $f\colon V \to X$ belongs to R, then fg belongs to R for every $g\colon W \to V$. We write $f \in R$ for $f \in R(V)$. A family $(f_i\colon U_i \to X)_i$ generates the sieve of all arrows factoring through at least one f_i.

For $u\colon Y \to X$, the pullback u^*R consists of the arrows $g\colon W \to Y$ for which $ug \in R$. This is a sieve by closure under precomposition. It is the objectwise pullback in the presheaf category, since $y(u)$ takes an arrow to its composite with u.

$$\begin{array}{ccc} u^*R & \longrightarrow & R \\ \downarrow & \lrcorner & \downarrow \\ yY & \xrightarrow{y(u)} & yX \end{array}$$

This description gives $\mathrm{id}_X^* R = R$ and $(uv)^*R = v^*(u^*R)$ for $v\colon Z \to Y$. The construction uses pullbacks of presheaves even when $\mathcal{C}$ itself has no pullbacks.

Definition 45.1. A *Grothendieck topology* J on $\mathcal{C}$ assigns to each X a collection $J(X)$ of *covering sieves*. It satisfies the following conditions.

1. The maximal sieve yX covers X.
2. If R covers X and $u\colon Y \to X$, then u^*R covers Y.
3. If R covers X and S is a sieve on X such that u^*S covers the domain of each $u \in R$, then S covers X.

The pair $(\mathcal{C}, J)$ is a *site*.

The second condition makes coverage stable on restriction. The third allows a cover to be checked on a cover. An enlargement $S \supseteq R$ of a covering sieve covers, since u^*S is maximal for each $u \in R$. If R, S cover X, their intersection covers as well. For each $u \in R$, $u^*(R \cap S) = u^*S$, which covers by stability. Transitivity then applies to $R \cap S$.

Example 45.2 (Open covers as covering sieves)**.** A sieve R on an open set U in the category of opens consists of a collection of opens in U closed under taking smaller opens. Declare it covering when $\bigcup_{V \in R} V = U$. Pullback to $W \subseteq U$ intersects each open of R with W, so their intersections cover W. For transitivity, if R covers and S covers after restricting to each $V \in R$, each such V is contained in the union of the opens in S. Thus their union is U. The sieve generated by an ordinary open cover satisfies this rule, including the empty cover of the empty open.

Example 45.3 (Jointly surjective families of finite sets)**.** Use a small skeleton of **FinSet**, with an object 1 having one element. A sieve on X covers if the images of its arrows have union X. Equivalently it contains each point $1 \to X$. Indeed, a point in the image of $f \in R$ factors through f by choosing one element in its domain. If R covers, every point of Y has its composite under $u\colon Y \to X$ in R, giving stability. For transitivity, test the assumed cover u^*S at each point $u\colon 1 \to X$ in R. It must contain id_1, so $u \in S$. This proves that S covers. The empty sieve covers the empty set, since there are no points to test.

45.2 The sheaf test is an extension property

For a presheaf F, a natural transformation $a\colon R \to F$ is a *matching family*. It gives $a_f \in F(V)$ for each $f\colon V \to X$ in R, with $F(g)(a_f) = a_{fg}$ for every $g\colon W \to V$. A section $s \in F(X)$ determines such a family by $a_f = F(f)(s)$. By Yoneda, these sections are the maps $yX \to F$. Thus the condition that a matching family have a unique glued section says that restriction gives a bijection

$$F(X) \cong \mathrm{Nat}(yX, F) \longrightarrow \mathrm{Nat}(R, F) \qquad (R \in J(X)).$$

A *sheaf* satisfies this test for every covering sieve. A presheaf is *separated* when restriction is injective. Write $\mathrm{PSh}(\mathcal{C})$ for the presheaf category and $\mathrm{Sh}(\mathcal{C}, J)$ for the full subcategory of sheaves.

When $(f_i\colon U_i \to X)_i$ generates R and the pullbacks $U_i \times_X U_j$ exist, a matching family is determined by the sections $s_i = a_{f_i}$. If $f_i g_i = f_j g_j$, the induced map to $U_i \times_X U_j$ tests agreement between the two restrictions. Conversely, agreement on all those pullbacks makes $a_f = F(g_i)(s_i)$ independent of a chosen factorization $f = f_i g_i$. These values commute with further restriction. In categorical terms

the presheaf diagram

$$\coprod_{i,j} y(U_i \times_X U_j) \rightrightarrows \coprod_i yU_i \longrightarrow R$$

is a coequalizer. At each test object T, the last map is surjective by the definition of R, and two choices of arrow $T \to U_i$ with the same composite are paired by the corresponding pullback. Applying $\mathrm{Nat}(-, F)$ recovers the gluing equalizer of Chapter 44 when the site is a category of opens.

Example 45.4 (Recovering a sheaf from its value at 1)**.** On the finite-set site above, a sheaf F has $F(\varnothing)$ a singleton by the empty-sieve test. For any X, take the covering sieve of maps factoring through a point of X. A matching family on this sieve is determined by the sections at the points. Conversely, any such pointwise assignment extends to a matching family. A map from a nonempty set in this sieve has a unique constant value x, and the unique map of its domain to 1 pulls back the assigned section at x. A map with empty domain has its unique section in $F(\varnothing)$. This yields a natural identification

$$F(X) \cong \mathbf{Set}(X, F(1)).$$

Naturality follows by tracing which section each point of a new domain selects after composition. For any set A, the presheaf $X \mapsto \mathbf{Set}(X, A)$ passes every covering-sieve test. A matching family defines a function $X \to A$ by its values on points, and its function on each arrow is then determined by agreement at all points of that arrow's domain. Natural transformations between these sheaves are determined by their maps at 1. Evaluation at 1 is consequently an equivalence $\mathrm{Sh}(\mathbf{FinSet}, J) \simeq \mathbf{Set}$.

A topology is *subcanonical* if every yX is a sheaf. In that case the Yoneda embedding lands in the sheaf category and is fully faithful there. The finite-set topology is subcanonical by the example, since $yX = \mathbf{Set}(-, X)$ on the chosen skeleton. The topology with only maximal sieves covering is also subcanonical, and every presheaf is then a sheaf.

45.3 Sheafification by refining covers

The presheaf F^+ is formed by taking local data modulo agreement after refinement. Its elements at X are pairs (R, a), with $R \in J(X)$ and $a\colon R \to F$, modulo the relation of agreeing on a common covering subsieve. Intersections of covers make this relation transitive. Restriction along $u\colon Y \to X$ pulls back the sieve and takes the family with value a_{ug} on $g \in u^*R$. Stability makes this well-defined. Using sections on the maximal sieve gives a natural map $F \to F^+$. Two sections

have the same class precisely when they agree on a cover, so this map is injective at every object if F is separated.

Proposition 45.5. *For every presheaf F, the presheaf F^+ is separated. If F is separated, F^+ is a sheaf.*

Proof. Take representatives $(R, a), (S, b)$ on X that restrict to equal classes along each $f \in T$, where T covers X. Let D be the subsieve of $R \cap S$ on which a and b agree. For $f \in T$, equality of the restricted classes means that f^*D contains a cover. Thus f^*D covers. Transitivity makes D a cover, which proves separatedness.

Now assume F is separated. Take a matching family of classes $b_f \in F^+(V)$ on a cover T of X, where $f\colon V \to X$ lies in T. Represent b_f by a family a^f on a cover R_f of V. The composites fg with $f \in T$ and $g \in R_f$ form a covering sieve D. Closure follows because each R_f is a sieve. For $f \in T$, f^*D contains R_f, so transitivity gives the covering condition. Assign a^f_g to fg. Its class in F^+ equals the given b_{fg}, by matching and by the description of restriction of a representative. If $fg = f'g'$, these classes coincide and injectivity of $F \to F^+$ gives $a^f_g = a^{f'}_{g'}$. The assignment is therefore well-defined, and its restrictions agree because each a^f is matching. Its class on D glues the original data, since its restriction on R_f represents b_f for each f. Separatedness of F^+ gives uniqueness. □

It follows that $aF = (F^+)^+$ is a sheaf. To extend a map $F \to H$ into a sheaf, send a class in $F^+(X)$ to the gluing in $H(X)$ of the image of its matching family. Agreement after refinement gives the same gluing, and restriction commutes with gluing by its uniqueness. The extension is unique because a class restricts along its chosen covering sieve to classes of sections in F. Applying this extension twice gives the adjunction

$$a\colon \mathrm{PSh}(\mathcal{C}) \rightleftarrows \mathrm{Sh}(\mathcal{C}, J)\colon i, \qquad a \dashv i.$$

Here i is inclusion. Since the restriction maps in the sheaf test are bijections for H, $a(iH) \cong H$.

For each X, the operation $F \mapsto F^+(X)$ is the colimit of the sets $\mathrm{Nat}(R, F)$ indexed by covering sieves ordered by reverse inclusion. This index is nonempty and filtered by the intersection property. Each $\mathrm{Nat}(R, -)$ preserves limits, and filtered colimits of sets commute with finite limits. In the latter assertion, finitely many representatives can be moved to one common stage, and finitely many equalities can be imposed at a further common stage. These two steps prove the assertion for finite products and equalizers, which form all finite limits. Thus a preserves finite limits, using this calculation twice and the fact that sheaf limits are formed on the underlying presheaves. Indeed a pointwise limit of sheaves passes each sheaf test, since $\mathrm{Nat}(R, -)$ and $\mathrm{Nat}(yX, -)$ both preserve

limits. Colimits of sheaves are obtained by sheafifying presheaf colimits, by the adjunction. Sheafification thus gives the universal way to impose gluing, and its preservation of finite limits gives compatibility with pullback.

45.4 Why sheaves form a topos

By definition, a *Grothendieck topos* is a category equivalent to $\mathrm{Sh}(\mathcal{C}, J)$ for a small site. The description also gives the structure of an elementary topos from Chapter 23. We already have finite limits. For sheaves F, G the presheaf exponential has value at X given by $\mathrm{Nat}(yX \times F, G)$, using F, G also as presheaves. This formula represents maps from products. To see this, for a presheaf H and $h \in H(X)$, Yoneda gives $yX \to H$ and hence a map $yX \times F \to H \times F$. A map $H \times F \to G$ supplies the exponential value at h by composing with this map. Conversely one evaluates the corresponding map $yX \times F \to G$ at (id_X, s) for $s \in F(X)$. These assignments are inverse by naturality with respect to every arrow into X.

For a covering sieve $R \hookrightarrow yX$, its sheafification $aR \to ayX$ is an isomorphism. Indeed, the sheaf test and the adjunction give a bijection on maps into each sheaf. The finite-product property of a makes $a(R \times F) \to a(yX \times F)$ an isomorphism as well. Since G is a sheaf, maps from these two presheaves to G therefore correspond bijectively. By the exponential property this is the sheaf test for the presheaf exponential. It supplies the exponential G^F inside the sheaf category.

To get a truth object, call a sieve S on X *closed* if f^*S covering implies $f \in S$. Closure assigns the sieve $\overline{S}$ of all f for which f^*S covers. Stability makes it a sieve, transitivity makes it closed, and the composition formula for pullbacks shows that closure commutes with pullback. Let $\Omega_J(X)$ be the set of closed sieves on X, with restriction given by pullback. To glue closed sieves S_f matching along a covering sieve R, take the sieve S' of composites fg for $f \in R$ and $g \in S_f$. For each $f \in R$, $f^*S' = S_f$. In fact if $fg = f'g'$ and $g' \in S_{f'}$, matching gives $g^*S_f = (g')^*S_{f'}$, which contains the identity. Thus $g \in S_f$, proving the only needed inclusion. The closure $\overline{S'}$ has these same restrictions because each S_f is already closed. For uniqueness, if two closed sieves S, T have the same restrictions to R and $u \in S$, then u^*T contains u^*R. This holds because every $ug \in R$ with $u \in S$ belongs to S, and membership along R agrees. Hence $u \in T$ by stability and closedness. The reverse inclusion follows by exchanging the sieves. This proves that Ω_J is a sheaf.

For a subobject $A \hookrightarrow F$ in sheaves, represent it by a subpresheaf. This is possible since inclusion preserves finite limits and therefore monomorphisms. For $s \in F(X)$ put

$$\chi_A(s) = \{f \colon V \to X \mid F(f)(s) \in A(V)\}.$$

It is closed because restrictions in A along a cover glue in A, and their gluing in F is unique. Pullback of the sieve gives restriction of $\chi_A(s)$, so this is a map $F \to \Omega_J$. The truth arrow $1 \to \Omega_J$ selects maximal sieves.

$$\begin{array}{ccc} A & \longrightarrow & 1 \\ \downarrow & \lrcorner & \downarrow{\scriptstyle \top} \\ F & \xrightarrow{\chi_A} & \Omega_J \end{array}$$

The square is a pullback because $\chi_A(s)$ is maximal exactly when its identity arrow lies in it, or equivalently $s \in A(X)$. Any classifying map must give the same sieve, since f lies in a sieve precisely when its pullback along f is maximal. Thus this is the required subobject classifier. For the general theory of sites and these constructions, see Mac Lane and Moerdijk [16, Chapter III].

45.5 Maps between toposes

The adjunction $a \dashv i$ illustrates the direction of a *geometric morphism*. A geometric morphism $f : \mathcal{E} \to \mathcal{F}$ has an adjunction $f^* \dashv f_*$ in which $f^* : \mathcal{F} \to \mathcal{E}$ preserves finite limits. The inverse image points from the target topos to the source, as restriction of data does for spaces. Thus $a \dashv i$ describes a geometric morphism from $\mathrm{Sh}(\mathcal{C}, J)$ to $\mathrm{PSh}(\mathcal{C})$. The inverse image imposes gluing along the covers.

For any sheaf topos $\mathcal{E}$, global sections $\Gamma(F) = \mathrm{Hom}_{\mathcal{E}}(1, F)$ give the direct image of a geometric morphism $\mathcal{E} \to \mathbf{Set}$. Its inverse image takes a set A to the sheafification of the constant presheaf with that value. A map from the constant presheaf to F amounts to a compatible global section of F for each element of A. The sheafification adjunction transfers this test to sheaves. Constant presheaves preserve finite limits of sets objectwise, and a preserves them as just proved. Geometric morphisms with source $\mathbf{Set}$ and with an arbitrary sheaf topos as source will give two tests for the models specified by a classifying topos later in this part.

CHAPTER 46

Bundles from descent

A sheaf glues sections of an object already defined over the base. For a bundle we can ask how the object itself is constructed. Take a bundle on each open set in a cover, and give isomorphisms identifying the bundles over their overlaps. These isomorphisms must compose consistently wherever three charts meet. The resulting object and its universal property are described by descent.

46.1 Reindexing a family

A family of spaces over X is a continuous map $p\colon E \to X$. For $f\colon Y \to X$, the pullback has total space

$$f^*E = Y \times_X E = \{(y, e) \mid f(y) = p(e)\},$$

with its subspace topology in $Y \times E$ and projection to Y. Over y its fiber is identified with the original fiber over $f(y)$. Pullback takes a map of families over X to a map over Y. This is the topological version of the reindexing of sets in Chapter 27.

$$\begin{array}{ccc} Y \times_X E & \longrightarrow & E \\ \downarrow & \lrcorner & \downarrow{\scriptstyle p} \\ Y & \xrightarrow{\;f\;} & X \end{array}$$

For an open inclusion $U \hookrightarrow X$, the projection to E identifies the pullback with $p^{-1}(U)$ and its inherited topology. We write $E|_U$ for this restriction.

A real vector bundle of rank r is a family whose fibers are r-dimensional real vector spaces with local product descriptions. Each $x \in X$ has an open neighborhood U on which $E|_U$ is homeomorphic over U to $U \times \mathbb{R}^r$ by a map linear on fibers. We call this a trivialization. Bundle maps over X are continuous maps over X linear on fibers. Write $\mathrm{Vect}_r(X)$ for this category. Pullback preserves bundles because a trivialization over U pulls back to one over $f^{-1}(U)$, using the same vector coordinates.

46.2 Descent along a surjection of sets

The underlying identification can be seen in **Set**. Let $f\colon Y \to X$ be surjective. A family $(A_y)_{y \in Y}$ descends to X if its variation between points over the same x

is given by compatible identifications. A *descent datum* consists of the family and bijections $\theta_{yz} \colon A_z \to A_y$ whenever $f(y) = f(z)$. We require $\theta_{yy} = \mathrm{id}$ and $\theta_{yz} \circ \theta_{zw} = \theta_{yw}$ whenever $f(y) = f(z) = f(w)$. In particular $\theta_{zy} = \theta_{yz}^{-1}$. A morphism of such data is a family of functions that commutes with every θ_{yz}. The two points indexing an identification belong to $Y \times_X Y$, and the three points in the composition law belong to $Y \times_X Y \times_X Y$.

$$\begin{array}{ccc} A_w & \xrightarrow{\theta_{zw}} & A_z \\ & \searrow^{\theta_{yw}} & \downarrow{\scriptstyle \theta_{yz}} \\ & & A_y \end{array}$$

A family $(B_x)_{x \in X}$ reindexes to $A_y = B_{f(y)}$ with identity identifications. Passing from families over X to these data is a functor.

Proposition 46.1. *Reindexing along a surjection gives an equivalence between families of sets over X and descent data over Y.*

Proof. Given a descent datum, define B_x as the quotient of the set of pairs (y, a) with $f(y) = x$ and $a \in A_y$. Identify (y, a) with $(z, \theta_{zy}(a))$ for each z over x. The identity, inverse, and composition laws make this relation an equivalence. For a fixed y over x, the map $A_y \to B_x$ sends a to $[y, a]$. Every class has a representative indexed by y, since an element of A_z can be transported to A_y by θ_{yz}. Each class has at most one such representative by the identity and composition laws. Thus $A_y \to B_x$ is a bijection that commutes with the given identifications.

A morphism with components $u_y \colon A_y \to A'_y$ sends $[y, a]$ to $[y, u_y(a)]$. Commutation with the identifications is exactly the condition for this rule to be well-defined. For a family already defined over X, the quotient of its reindexed family identifies copies of each element of B_x. The map from that quotient to B_x is bijective because each x has at least one preimage. These bijections are natural in the families and the data, so the two constructions are inverse up to natural isomorphism. □

A datum is *effective* if it comes from a global family up to isomorphism. Surjections of sets give effective descent for all families, with the quotient supplying the inverse construction. The same cocycle pattern appears for schemes in the Stacks Project [37, Section 35.34].

46.3 Bundle data on an open cover

Let $(U_i)_{i \in I}$ cover X, and put $U_{ij} = U_i \cap U_j$ and $U_{ijk} = U_i \cap U_j \cap U_k$. The map $\coprod_i U_i \to X$ records the cover as a single map. Its double and triple pullbacks are

respectively the sums of the spaces U_{ij} and U_{ijk}, with each intersection indexed by the corresponding ordered tuple.

A bundle datum consists of bundles E_i of rank r on U_i and bundle isomorphisms $\theta_{ij}\colon E_j|_{U_{ij}} \to E_i|_{U_{ij}}$. The map θ_{ii} is the identity. After restriction to each U_{ijk}, we require $\theta_{ij} \circ \theta_{jk} = \theta_{ik}$. For a second datum (E_i', θ_{ij}'), a morphism consists of maps $u_i\colon E_i \to E_i'$ over U_i satisfying the following square on each U_{ij}.

$$\begin{array}{ccc} E_j|_{U_{ij}} & \xrightarrow{\theta_{ij}} & E_i|_{U_{ij}} \\ \downarrow{\scriptstyle u_j} & & \downarrow{\scriptstyle u_i} \\ E_j'|_{U_{ij}} & \xrightarrow{\theta_{ij}'} & E_i'|_{U_{ij}} \end{array}$$

Write $\mathrm{Desc}_r((U_i))$ for the resulting category. A bundle over X yields these data by restriction. Its identifications come from taking both restrictions inside the same total space.

Theorem 46.2 (Effective descent for bundles on an open cover)**.** *Restriction gives an equivalence*

$$\mathrm{Vect}_r(X) \longrightarrow \mathrm{Desc}_r((U_i)).$$

Proof. Form the disjoint union of the total spaces of the E_i and identify $e \in E_j|_{U_{ij}}$ with $\theta_{ij}(e) \in E_i$. The cocycle makes the displayed identifications transitive, since each leg preserves the same base point and a composite lies over the triple overlap. Inverses and identities give symmetry and reflexivity. Let E be the quotient with the quotient topology. The projections $p_i\colon E_i \to U_i \subseteq X$ agree along the identifications and so give a continuous projection p.

$$\begin{array}{ccc} \coprod_i E_i & \xrightarrow{q} & E \\ & \searrow^{(i,e)\mapsto p_i(e)} & \downarrow{\scriptstyle p} \\ & & X \end{array}$$

The quotient map on E_i gives a bijection $j_i\colon E_i \to p^{-1}(U_i)$. Every class over U_i has a representative from E_i by transport, and transport back to i recovers the same representative. To see that j_i is open, take $O \subseteq E_i$ open. Its image in the quotient has preimage in E_k equal to

$$\theta_{ki}(O \cap E_i|_{U_{ik}}).$$

The right-hand set is open in $E_k|_{U_{ik}}$ by the homeomorphism property, and $E_k|_{U_{ik}}$ is open in E_k. The quotient topology now makes $j_i(O)$ open in E. Thus each j_i identifies E_i with an open subspace of E. The fiber vector operations agree in these charts because every θ_{ij} is linear. The original local trivializations therefore

make p a bundle of the prescribed rank. Its restriction data are isomorphic to the given ones.

A compatible family of maps u_i gives a single map on the quotient by $j_i(e) \mapsto j_i'(u_i(e))$. The quotient topology ensures continuity, and each component is linear on fibers. Conversely any map of bundles restricts to such a family, and the restrictions determine it because the U_i cover X. Applying the same construction to local maps between restrictions of global bundles proves full faithfulness. Thus restriction is an equivalence. □

The proof builds the topology and the vector-space operations from the descent datum. This goes beyond gluing sections. Once the bundle is constructed, a section can be described by maps $s_i \colon U_i \to E_i$ over U_i that satisfy $s_i|_{U_{ij}} = \theta_{ij} \circ s_j|_{U_{ij}}$. The bijections j_i turn those maps into continuous local sections of one projection. Since the U_i are open, continuous local sections agreeing on overlaps define a unique continuous global section. In the language of Chapter 44, the sections of p form a sheaf of sets.

46.4 Transition functions and a twisted line

When $E_i = U_i \times \mathbb{R}^r$, the isomorphism θ_{ij} has the form $(x, v) \mapsto (x, g_{ij}(x)v)$. The matrix $g_{ij}(x)$ is invertible, and its entries depend continuously on x, as seen by applying θ_{ij} to each constant basis section. Conversely continuous maps $g_{ij} \colon U_{ij} \to \mathrm{GL}_r(\mathbb{R})$ give such isomorphisms by matrix multiplication and continuous inversion. The cocycle becomes $g_{ij}g_{jk} = g_{ik}$ on triple overlaps. Changing the coordinate on each chart by $h_i(x) \in \mathrm{GL}_r(\mathbb{R})$ gives the new transition $g'_{ij} = h_i g_{ij} h_j^{-1}$. This formula follows by converting a new input coordinate to the old one, applying the transition, and converting the output to the new coordinate.

Example 46.3 (The Möbius line bundle)**.** Let $X = \mathbb{R}/\mathbb{Z}$ with quotient map π, and set

$$U_0 = \pi((-1/4, 3/4)), \qquad U_1 = \pi((1/4, 5/4)).$$

Each is open and the quotient map restricts to a homeomorphism from the given interval. In fact π is open since the saturation of an open set is the union of its integer translates. The two arcs cover X. Their intersection has two components

$$V_+ = \pi((1/4, 3/4)), \qquad V_- = \pi((-1/4, 1/4)).$$

Use a trivial real line on each arc. Take $g_{01} = +1$ on V_+ and $g_{01} = -1$ on V_-. Take $g_{10} = g_{01}^{-1}$ and $g_{ii} = 1$. The cocycle holds because with two chart indices every triple relation follows from the identities and inverses.

The total space can also be written as the quotient of $\mathbb{R}\times\mathbb{R}$ by $(t, v) \sim (t+1, -v)$ and its iterates. For $x \in U_i$, let $t_i(x)$ be its unique lift to the interval defining

U_i. The chart map from $U_i \times \mathbb{R}$ sends (x, v) to $[t_i(x), v]$. It is a homeomorphism onto the corresponding part of the quotient, since translating any open piece of the strip by $(t, v) \mapsto (t + 1, -v)$ preserves openness. On V_+ the two lifts agree, and on V_- they differ by $t_1 = t_0 + 1$. The chart transitions are therefore exactly g_{01} and its inverse.

If the bundle were trivial, the constant section 1 of the trivial line would give a section everywhere nonzero. Its two coordinate functions s_i would each have constant sign because U_i is connected. On V_+ the signs would have to agree by the section equation $s_0 = g_{01} s_1$, whereas on V_- they would have to be opposite. This is impossible. The two coordinate descriptions therefore glue to a nontrivial bundle.

The same transition functions give a two-sheeted cover if we replace the line by its two signs, $+1$ and -1. Repeating the interval description gives the quotient of $\mathbb{R} \times \{+1, -1\}$ under $(t, \epsilon) \sim (t + 1, -\epsilon)$. This is isomorphic to the cover $\mathbb{R}/2\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$. Indeed send $(t, +1)$ to t and $(t, -1)$ to $t + 1$, modulo $2\mathbb{Z}$. The formula respects the relation and is bijective on classes, with inverse $t \mapsto [(t, +1)]$. Both maps are continuous by the quotient descriptions. Following a point in the fiber once around the base switches its sign. This change is visible in a discrete fiber because the original transition was an identification of whole fibers.

46.5 Transporting descent data

For a map $f\colon Y \to X$, the $f^{-1}(U_i)$ cover Y. Pulling back an identification θ_{ij} gives a bundle isomorphism on $f^{-1}(U_{ij})$. The cocycle holds after pullback because pullback preserves composites with the canonical identifications of iterated restrictions. Let E be the bundle obtained from the datum on X. Then f^*E restricts on $f^{-1}(U_i)$ to the pullback of E_i along $f|_{f^{-1}(U_i)}$.

$$\begin{array}{ccc} f^{-1}(U_i) \times_{U_i} E_i & \hookrightarrow & Y \times_X E \\ \downarrow & \lrcorner & \downarrow \\ f^{-1}(U_i) & \hookrightarrow & Y \end{array}$$

The square is a pullback by the description of its points and subspace topologies. Its local identifications on Y are the pulled-back ones, so Theorem 46.2 says that gluing these data gives f^*E up to a unique isomorphism compatible with all local identifications. This compatibility with base change will also govern gluing in algebraic geometry.

46.6 Counting colorings by descent

Fix an integer $q \geq 2$ and write $\mathrm{Col}_q(H)$ for the set of proper q-colorings of a finite graph H. A proper coloring assigns a color to each vertex and assigns different colors to the endpoints of each edge. On the poset of subgraphs of H, these sets form a sheaf for covers whose union contains every vertex and edge. Indeed, agreement on overlaps gives one well-defined color at each vertex. Every edge belongs to some covering subgraph, so its endpoints have different colors in the glued coloring.

For $H = H_1 \cup H_2$, put $K = H_1 \cap H_2$. Restriction gives the pullback

$$\begin{array}{ccc} \mathrm{Col}_q(H) & \longrightarrow & \mathrm{Col}_q(H_1) \\ \downarrow & & \downarrow \\ \mathrm{Col}_q(H_2) & \longrightarrow & \mathrm{Col}_q(K). \end{array}$$

For $c \in \mathrm{Col}_q(K)$, let $e_i(c)$ be the number of extensions to H_i. The fiber over c of the pullback is a product of the two extension sets. This yields

$$|\,\mathrm{Col}_q(H)| = \sum_{c \in \mathrm{Col}_q(K)} e_1(c) e_2(c).$$

Example 46.4 (Two triangles joined along an edge)**.** Let H_1 have vertices a, b, c and H_2 have vertices a, b, d, with all three edges in each triangle. They share the edge ab. Its endpoints have $q(q-1)$ possible assignments. Each such assignment leaves $q-2$ choices at c and $q-2$ choices at d, independently. The number of global colorings is $q(q-1)(q-2)^2$. With three colors the shared edge determines both remaining colors, giving six global colorings.

CHAPTER 47

Schemes from functors of points

The equation $t^2 = 0$ has only the solution 0 in a field. In the ring $k[\epsilon]/(\epsilon^2)$ it also has the solution ϵ. Changing the ring of coefficients makes the equation detectable at a scale that fields alone miss. A scheme can be tested by maps from other schemes. For affine schemes these tests are ring maps in the reverse direction. The topology and sheaf of an affine scheme let us apply the same approach locally when one ring cannot describe the whole space.

47.1 Testing equations in a ring

Write **CRing** for the category of commutative unital rings and unit-preserving homomorphisms, including the zero ring. For a ring A, define $h_A \colon \mathbf{CRing} \to \mathbf{Set}$ by

$$h_A(R) = \mathrm{Hom}_{\mathbf{CRing}}(A, R).$$

A map $R \to S$ acts by postcomposition, sending solutions over R to solutions over S. For a presentation $A = \mathbb{Z}[x_1, \ldots, x_n]/I$, a map $A \to R$ amounts to a tuple in R^n that sends every polynomial of I to zero. This follows from the universal properties of polynomials and quotients in Chapters 10 and 11.

The covariant Yoneda calculation of Chapter 4 gives

$$\mathrm{Nat}(h_A, h_B) \cong \mathrm{Hom}_{\mathbf{CRing}}(B, A).$$

In fact a transformation $\eta \colon h_A \to h_B$ applied to $\mathrm{id}_A \in h_A(A)$ gives a map $v \colon B \to A$. Naturality at $a \colon A \to R$ gives $\eta_R(a) = a \circ v$. Conversely each v defines the transformation by this formula. Composition is reversed on the representing rings. We will write an affine object as $\operatorname{Spec} A$, so that the transformation induced by v has direction $\operatorname{Spec} A \to \operatorname{Spec} B$.

47.2 Prime ideals and distinguished opens

A *prime ideal* $\mathfrak{p}$ of A is a proper ideal such that $A/\mathfrak{p}$ is an integral domain. Equivalently, $ab \in \mathfrak{p}$ entails $a \in \mathfrak{p}$ or $b \in \mathfrak{p}$. The underlying set of $\operatorname{Spec} A$ is the set of prime ideals. A map $A \to K$ into a field has prime kernel $\mathfrak{p}$. It factors uniquely through the field of fractions $\kappa(\mathfrak{p})$ of $A/\mathfrak{p}$, since it sends every nonzero element of this domain to a unit. Each prime is obtained in this way from its

canonical map $A \to \kappa(\mathfrak{p})$. The prime ideal records where a field-valued test lands. The map of residue fields records the rest of the test.

For $f \in A$, put

$$D_A(f) = \{\mathfrak{p} \mid f \notin \mathfrak{p}\}.$$

These sets form a basis for the *Zariski topology*. They cover the space since $D_A(1) = \operatorname{Spec} A$, and primality gives $D_A(f) \cap D_A(g) = D_A(fg)$. A ring map $v\colon A \to B$ gives the continuous map $\mathfrak{q} \mapsto v^{-1}(\mathfrak{q})$, since inverse image takes $D_A(f)$ to $D_B(v(f))$. This is the point map of $\operatorname{Spec} v$.

Write $A_f = A[f^{-1}]$ for the localization from Chapter 11. Its ring maps into R correspond to maps $A \to R$ sending f to a unit. In particular prime ideals of A_f contract to primes in $D_A(f)$. For $\mathfrak{p} \in D_A(f)$, the kernel of $A_f \to \kappa(\mathfrak{p})$ consists of the fractions with numerator in $\mathfrak{p}$ and contracts back to it. This kernel is the only prime of A_f over $\mathfrak{p}$, since denominators are units. The resulting bijection $\operatorname{Spec} A_f \to D_A(f)$ is a homeomorphism. The distinguished open of a fraction a/f^n corresponds to $D_A(a) \cap D_A(f)$, and these sets are a basis on either side. The localization map and the corresponding open inclusion go in opposite directions.

$$A \longrightarrow A_f$$

$$\operatorname{Spec} A \xleftarrow{\operatorname{Spec}} \operatorname{Spec} A_f$$

The spectrum is quasi-compact, meaning that every open cover has a finite subcover. For a cover by $D_A(f_i)$, the ideal generated by the f_i is the whole ring. A proper ideal would be contained in a maximal ideal, which is prime and would be outside every $D_A(f_i)$. An expression $1 = \sum_{i \in I_0} c_i f_i$ uses a finite set I_0 and gives a finite subcover. Refining arbitrary opens by the basis proves the claim. The argument applies also to each $D_A(f)$ using A_f. See the Stacks Project [9] for further algebraic properties of the spectrum.

47.3 The sheaf of local fractions

On $D_A(f)$ prescribe the ring A_f. If $D_A(g) \subseteq D_A(f)$, the image of f in A_g lies in no maximal ideal by the prime correspondence. It is therefore a unit. The restriction $A_f \to A_g$ is the unique map extending $A \to A_g$ by localization. Uniqueness ensures that restrictions compose. This also gives the canonical identification when two choices of element specify the same open.

We verify the sheaf equalizer on the basis. First suppose $D_A(f_1), \dots, D_A(f_m)$ cover the whole spectrum. The maps in the equalizer test of Chapter 44 are the

localization restrictions

$$A \longrightarrow \prod_i A_{f_i} \rightrightarrows \prod_{i,j} A_{f_i f_j}.$$

The two maps to the last ring compare the i-th and j-th sections on the overlap. Suppose the sections agree there. Choose a common exponent N and write them as a_i/f_i^N. For each pair i, j, agreement says that a power of $f_i f_j$ annihilates $f_j^N a_i - f_i^N a_j$. There are finitely many pairs, so one power M works for all of them. Set $K = N + M$ and $b_i = f_i^M a_i$. We then have $f_j^K b_i = f_i^K b_j$ in A. The f_i^K generate the unit ideal because their distinguished opens still cover. Choose coefficients r_i with $\sum_i r_i f_i^K = 1$. The element $a = \sum_i r_i b_i$ satisfies $f_j^K a = b_j$ for each j, so a restricts to a_j/f_j^N. For uniqueness, if an element restricts to zero everywhere, some power of each f_i annihilates it. A common power suffices, and a linear combination of those powers is 1. The element is zero.

The same proof on a distinguished open $D_A(f)$ uses A_f. Any cover of this open by distinguished opens has a finite subcover. After gluing there, the proposed section agrees on any extra member by uniqueness on its intersections with the finite family. For arbitrary opens U, define $\mathcal{O}_A(U)$ to be the ring of families of sections on the distinguished opens contained in U, compatible with all restrictions. To glue over a cover of U, cover each distinguished open in U by smaller distinguished opens lying in members of the given cover. The basis equalizer produces a unique section on the first open. Two choices of refinement give the same section by comparing on their intersections. The resulting sections are compatible under restriction and give the required unique family. This defines the structure sheaf $\mathcal{O}_A$. In particular, $\mathcal{O}_A(\operatorname{Spec} A) \cong A$. The construction agrees with the affine structure sheaf in the Stacks Project [38, Section 26.5].

The open $D_A(f)$, with this sheaf restricted to it, is the affine scheme of A_f. A basic open inside $\operatorname{Spec} A_f$ is specified by a fraction g/f^n and corresponds to $D_A(fg)$. On the two sides the section rings are $(A_f)_{g/1}$ and A_{fg}. They agree by localization since making fg a unit in a commutative ring is equivalent to making both f and g units. The same identification commutes with restrictions because all the maps extend the maps from A.

The stalk at $\mathfrak{p}$ is the colimit of the A_f for $f \notin \mathfrak{p}$, ordered by refinement of neighborhoods. This follows from the germ construction of Chapter 44 because the distinguished opens are a basis. A compatible family of maps from these rings into a ring R is the same as one map $A \to R$ making every $f \notin \mathfrak{p}$ invertible. Thus the stalk is $A_{\mathfrak{p}}$, the localization at $A \setminus \mathfrak{p}$. It is a local ring. Its unique maximal ideal is $\mathfrak{p}A_{\mathfrak{p}}$, since the fractions whose numerators are outside $\mathfrak{p}$ are units, and its quotient by that ideal is $\kappa(\mathfrak{p})$.

47.4 Affine maps and base change

A *locally ringed space* is a space with a sheaf of rings whose stalks are local rings. A morphism $(X, \mathcal{O}_X) \to (Y, \mathcal{O}_Y)$ is a continuous map $\phi\colon X \to Y$ with ring maps $\mathcal{O}_Y(U) \to \mathcal{O}_X(\phi^{-1}U)$ compatible with restriction, inducing local maps on the stalks. A ring homomorphism of local rings is local if the inverse image of the maximal ideal of the target is the maximal ideal of the source. For $v\colon A \to B$ and $\mathfrak{q} \in \operatorname{Spec} B$, put $\mathfrak{p} = v^{-1}(\mathfrak{q})$. The maps on sections over distinguished opens are $A_f \to B_{v(f)}$. They come from localization and induce the local map $A_{\mathfrak{p}} \to B_{\mathfrak{q}}$. This constructs $\operatorname{Spec} v$ on spaces and sheaves.

Conversely a morphism $\operatorname{Spec} B \to \operatorname{Spec} A$ induces a map $v\colon A \to B$ on global sections. Suppose it sends $\mathfrak{q}$ to $\mathfrak{p}$. The global-to-stalk maps commute with the induced stalk map. Since the stalk map is local, an element $a \in A$ lies in $\mathfrak{p}$ if and only if $v(a) \in \mathfrak{q}$. This recovers the point map by contraction. The section map on $D_A(f)$ extends $A \to B \to B_{v(f)}$ and is forced by localization. Restricting to the basis then determines all the section maps. Hence the affine locally ringed spaces $(\operatorname{Spec} A, \mathcal{O}_A)$ give a category equivalent to $\mathbf{CRing}^{\mathrm{op}}$. Its maps from $\operatorname{Spec} R$ are exactly $h_A(R)$. Under this bijection, factoring through $D_A(f)$ is the condition that the image of f is invertible in R.

Example 47.1 (Two tests for the equation $t^2 = 0$)**.** Work over the field $k = \mathbb{F}_3$, so all ring maps on coordinates preserve the given maps from k. The affine line $\mathbb{A}^1_k = \operatorname{Spec} k[t]$ has one R-point for each $r \in R$, given by $t \mapsto r$. Its subscheme with coordinate ring $k[t]/(t^2)$ consists in this test of the r with $r^2 = 0$. Over $R = k$, the line has three points and this subscheme has the single point 0. Over $R = k[\epsilon]/(\epsilon^2)$ the line has nine points. The subscheme has the three points $0, \epsilon, 2\epsilon$. Indeed, $(a + b\epsilon)^2$ has constant term a^2, which vanishes exactly when $a = 0$, and the square of $b\epsilon$ is zero. In every field extension of k the subscheme and $\operatorname{Spec} k$ have the same singleton set of points. The dual-number test distinguishes them.

For ring maps from R to A and B, the tensor product $A \otimes_R B$ carries the multiplication $(a \otimes b)(a' \otimes b') = aa' \otimes bb'$ and unit $1 \otimes 1$. The tensor relations of Chapter 14 make the multiplication well-defined. A map from it to a commutative ring C amounts to maps $i\colon A \to C$ and $j\colon B \to C$ agreeing on R. The extension sends $a \otimes b$ to $i(a)j(b)$, which is balanced and multiplicative. The tensor product is thus a pushout of rings. Reversing maps gives the affine pullback

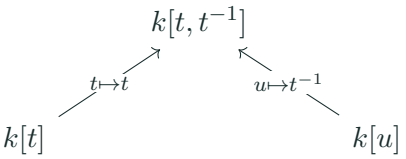
$$\begin{array}{ccc} \operatorname{Spec}(A \otimes_R B) & \longrightarrow & \operatorname{Spec} B \\ \downarrow & \lrcorner & \downarrow \\ \operatorname{Spec} A & \longrightarrow & \operatorname{Spec} R. \end{array}$$

For example, given $v\colon A \to B$, one has $B \otimes_A A_f \cong B_{v(f)}$. Both have maps to C parametrized by maps from B to C inverting $v(f)$. This is the algebraic form of pulling a distinguished open back along $\operatorname{Spec} v$.

47.5 Gluing affine charts

A *scheme* is a locally ringed space with an open cover by affine schemes. An affine open in this definition has the restricted structure sheaf. Isomorphisms between open parts of charts can be used as gluing data if they have the cocycle property on triple overlaps, as in Chapter 46. The topological gluing identifies points according to the given isomorphisms, with each chart embedded as an open. A section on an open of the glued space is a family of chart sections agreeing under the identifications. These families form a sheaf by gluing in each chart. Its restrictions recover the chart sheaves, so the stalks are local rings. Maps from the result are determined by chart maps that agree on the identified overlaps, since both continuous maps and the compatible maps on sections glue.

Example 47.2 (The projective line)**.** For a field k, glue $U = \operatorname{Spec} k[t]$ and $V = \operatorname{Spec} k[u]$ by identifying $D_U(t)$ and $D_V(u)$ via $u = t^{-1}$. This means that the two open inclusions come from the ring maps in the diagram.



The resulting scheme is $\mathbb{P}^1_k$. A global section is a pair $p(t), q(u)$ with $p(t) = q(t^{-1})$ as Laurent polynomials. The equality allows only the constant term. Thus the global ring is k, and the scheme has at least two points, the origins of the two charts. If it were affine, its global ring would identify it with $\operatorname{Spec} k$, whose only point is the zero ideal. The line is therefore assembled from affines without being affine itself.

Example 47.3 (A test that meets both charts)**.** Take $R = k \times k$. Its spectrum is the disjoint union of two copies of $\operatorname{Spec} k$. The two basic opens are specified by the complementary idempotents $(1, 0)$ and $(0, 1)$. Their product is zero and their sum is one, so every prime lies in exactly one of these opens. Localizing at either idempotent gives k by killing the other idempotent and keeping its own factor. Map the two components to the two chart origins of $\mathbb{P}^1_k$. They give a single R-point. It does not lie in the image of either chart on R-points, since each chart omits the other origin. The open cover provides local factorizations of this test.

47.6 Affine tests determine maps

For an arbitrary scheme X, set $X(R) = \mathrm{Hom}_{\mathbf{Sch}}(\operatorname{Spec} R, X)$, with test-ring maps acting by precomposition on the affine domains. This functor of points still recovers all scheme morphisms. To see this, let $\eta\colon X(-) \to Y(-)$ be natural. For every affine open $U = \operatorname{Spec} A \subseteq X$, apply η_A to the inclusion $U \to X$. The resulting maps $U \to Y$ agree over an overlap because their restrictions to any affine open in the overlap are equal by naturality. Those affine opens cover the overlap. Thus the maps glue to a morphism $\phi\colon X \to Y$. For a test map $a\colon \operatorname{Spec} R \to X$, cover its domain by affine opens mapping into charts of X. On each such open W, the restricted test factors through the chosen affine chart. Naturality then gives $\eta_R(a)|_W = \phi \circ (a|_W)$. The two maps from $\operatorname{Spec} R$ are therefore equal by gluing. The chart inclusions also show that ϕ is unique. This is a Yoneda test using only affine domains, with the reduction to that smaller collection justified by the open covers.

Maps from any scheme into an affine also have a formula in terms of sections. A map $X \to \operatorname{Spec} A$ induces a ring map $A \to \Gamma(X, \mathcal{O}_X)$, where Γ denotes the global section ring. Given such a ring map, restrict its values to each affine chart of X and take spectra. The chart maps agree on all affine opens of the overlaps, since the restricted ring maps agree there. They therefore glue to a map from X. The two constructions are inverse by the affine calculation and the sheaf uniqueness of global sections. In particular every map from $\mathbb{P}^1_k$ to $\mathbb{A}^1_k$ over k is constant, since its coordinate is an element of $\Gamma(\mathbb{P}^1_k, \mathcal{O}) = k$.

The affine pullback in Section 47.4 also has the universal property among all schemes. Its affine-domain comparison is bijective by the tensor product calculation. For any two compatible maps from a scheme, the unique test factorizations over its affine charts agree on affines in the overlaps and glue to the unique global factorization. In the next chapter, tests over a field will organize finite étale schemes by their Galois actions.

CHAPTER 48

Galois theory from finite étale schemes

A finite separable field extension has one geometric point for each embedding into a separable closure. Choosing a closure lets its automorphisms permute these points. The resulting finite set with an action remembers the extension, and equivariant maps remember maps of schemes. The same description applies to finite étale covers of a connected scheme, with the group recovered from the fiber functor.

48.1 Finite algebras that split over a closure

Let k be a field and k^{sep} a separable closure. Every scheme finite over $\operatorname{Spec} k$ is affine. Its coordinate algebra is a commutative finite-dimensional k-algebra. For such a scheme, the étale condition over k says that the algebra is a finite product of finite separable field extensions of k. We use this field case of the general characterization of étale schemes over a field [39, Lemma 58.2.2]. The empty scheme corresponds to the empty product, the zero k-algebra.

This characterization is equivalent to a test after extension of scalars. Namely, A is finite étale over k precisely when $A \otimes_k k^{\mathrm{sep}}$ is a finite product of copies of k^{sep}. To see one direction, write a finite separable field extension L/k as $k[t]/(f)$ by the primitive element theorem. The minimal polynomial f has distinct roots in k^{sep}, and evaluation at these roots gives

$$L \otimes_k k^{\mathrm{sep}} \cong k^{\mathrm{sep}}[t]/(f) \cong \prod_{\sigma \in \mathrm{Hom}_{k\text{-alg}}(L, k^{\mathrm{sep}})} k^{\mathrm{sep}}.$$

The second isomorphism is the Chinese remainder construction of Chapter 11. Taking finite products gives the result for A. The isomorphism sends $a \otimes c$ in any summand to the tuple of evaluations $c\,\sigma(a)$, so the description does not depend on the choice of primitive element.

Conversely, if the scalar extension splits, A is reduced because $A \to A \otimes_k k^{\mathrm{sep}}$ is injective and the target has no nonzero nilpotents. Every prime quotient of a finite-dimensional k-algebra is a field. Multiplication by any nonzero element of the quotient is an injective linear endomorphism and hence surjective. There are only finitely many maximal ideals. Otherwise, the Chinese remainder maps to

products of arbitrarily many nonzero field quotients would be surjective, contrary to the dimension bound. The intersection of the maximal ideals is the nilradical, which is zero here. The Chinese remainder theorem therefore expresses A as their product of field quotients. For each field factor L, its identity idempotent is nonzero after scalar extension. Choose a coordinate of the split algebra at which this idempotent is 1. The coordinate map restricts to a k-embedding $L \to k^{\mathrm{sep}}$, so L/k is separable.

The split algebra describes the geometry of the test. For $X = \operatorname{Spec} A$, put $X_{k^{\mathrm{sep}}} = X \times_{\operatorname{Spec} k} \operatorname{Spec}(k^{\mathrm{sep}})$. By affine base change from Chapter 47, the square

$$\begin{array}{ccc} X_{k^{\mathrm{sep}}} & \longrightarrow & X \\ \downarrow & \lrcorner & \downarrow \\ \operatorname{Spec}(k^{\mathrm{sep}}) & \longrightarrow & \operatorname{Spec} k \end{array}$$

expresses $X_{k^{\mathrm{sep}}}$ as the spectrum of the split algebra. Its components are copies of $\operatorname{Spec}(k^{\mathrm{sep}})$. In contrast, the dual numbers have one point over k^{sep}, and the square-zero generator survives after base change. The splitting test concerns the coordinate algebra as well as its points.

48.2 The fiber functor and the Galois action

Write $\mathrm{FEt}(k)$ for the category of schemes finite étale over $\operatorname{Spec} k$, with morphisms over that base. For X in this category, the set of points over the chosen closure is

$$F(X) = \operatorname{Hom}_{\operatorname{Spec} k}\bigl(\operatorname{Spec}(k^{\mathrm{sep}}), X\bigr).$$

When $X = \operatorname{Spec} A$, this is $\operatorname{Hom}_{k\text{-alg}}(A, k^{\mathrm{sep}})$. An object $\operatorname{Spec} L$ for a finite separable extension L/k thus has $[L : k]$ points in the geometric fiber. A product of field factors in A gives a disjoint union of their embedding sets. A map to the field k^{sep} sends precisely one of the factor idempotents to 1, so it chooses one factor.

The *absolute Galois group* $G_k = \operatorname{Gal}(k^{\mathrm{sep}}/k)$ acts from the left by $g \cdot \varphi = g \circ \varphi$ on the algebra maps. For a scheme map $f\colon X \to Y$, with coordinate homomorphism $f^{*}\colon B \to A$, the map $F(f)$ sends φ to $\varphi \circ f^{*}$. This sends $F(X)$ to $F(Y)$, in the direction of the scheme arrow. The diagram shows this composition on coordinate rings.

$$\begin{array}{ccc} B & \xrightarrow{f^{*}} & A \\ & \searrow^{\varphi \circ f^{*}} & \downarrow{\scriptstyle \varphi} \\ & & k^{\mathrm{sep}} \\ & & \downarrow{\scriptstyle g} \\ & & k^{\mathrm{sep}} \end{array}$$

Postcomposition with g makes this triangle the corresponding triangle for $g \cdot \varphi$. Thus $F(f)$ is equivariant.

The topology on G_k comes from the restriction maps to the groups $\mathrm{Gal}(E/k)$, with E running over finite Galois extensions in k^{sep}. Those extensions have union k^{sep}, so a compatible system of automorphisms of the finite extensions gives one automorphism of the closure. This identifies G_k with the inverse limit of the finite groups. A finite set is given the discrete topology. Its action is continuous when each point has open stabilizer, or equivalently when the action factors through a finite continuous quotient. For a fiber element φ arising from a field factor L, an automorphism which fixes the finite image $\varphi(L)$ fixes φ. That image lies in a finite Galois extension, so its pointwise stabilizer is open. We obtain a functor

$$F\colon \mathrm{FEt}(k) \longrightarrow \mathbf{FinSet}^{G_k}_{\mathrm{cont}}.$$

48.3 Recovering fields and their maps

An embedding $L \hookrightarrow k^{\mathrm{sep}}$ identifies L with a subfield. Its stabilizer is then $H_L = \mathrm{Gal}(k^{\mathrm{sep}}/L)$. The classical Galois correspondence identifies the fixed field $(k^{\mathrm{sep}})^{H_L}$ with L, and for any open subgroup $H \leq G_k$ it gives a finite separable fixed field L_H with $H = H_{L_H}$. These are the finite layers of the correspondence. For instance, choose a finite Galois extension E/k whose restriction kernel lies in H. Then L_H is the fixed field in E of the image of H in $\mathrm{Gal}(E/k)$.

Every k-embedding $L \to k^{\mathrm{sep}}$ extends to an automorphism of k^{sep} over k. One can first extend through a finite normal closure of L and then to the separable closure. Consequently,

$$G_k/H_L \xrightarrow{\sim} \mathrm{Hom}_{k\text{-alg}}(L, k^{\mathrm{sep}}), \qquad gH_L \longmapsto g|_L.$$

This is equivariant, with an equality of restrictions precisely when the automorphisms lie in the same left coset of H_L. Conversely, for a transitive finite continuous G_k-set S, choose $s \in S$. The map $g \mapsto g \cdot s$ identifies S with G_k/H, where H is its stabilizer. The algebra L_H therefore realizes this orbit. Replacing s by $g \cdot s$ replaces H by gHg^{-1} and L_H by the k-isomorphic field $g(L_H)$.

The map test has the same form on both sides. Let H, K be open subgroups of G_k. An equivariant map $G_k/H \to G_k/K$ is determined by the image gK of H, because the source is transitive. It exists with that image exactly when gK is fixed by H, which says $H \subseteq gKg^{-1}$. This condition is equivalent to $g(L_K) \subseteq L_H$. In that case the map on fields is $L_K \to L_H,\ a \mapsto g(a)$. Changing the representative g of gK leaves its restriction to L_K the same. Every k-map $L_K \to L_H$ is obtained this way by extending its composite into k^{sep}. The direction on schemes is $\operatorname{Spec} L_H \to \operatorname{Spec} L_K$, in agreement with the equivariant map.

An automorphism of L_H over k extends to an element of G_k preserving L_H. By the fixed-field correspondence, these elements form the normalizer $N_{G_k}(H)$, and restriction has kernel H. Thus $\mathrm{Aut}_k(L_H) \cong N_{G_k}(H)/H$. The corresponding scheme automorphisms have reversed composition. If H is normal, L_H/k is Galois and the orbit is the left regular action of the finite quotient group G_k/H.

Theorem 48.1 (Finite étale schemes over a field)**.** *The geometric fiber functor F is an equivalence between* $\mathrm{FEt}(k)$ *and finite continuous G_k-sets.*

Proof. Disjoint union of connected components in $\mathrm{FEt}(k)$ matches disjoint union of orbits. Every map from $\operatorname{Spec} L$, a one-point scheme, into a finite disjoint union factors uniquely through the component containing its image. An equivariant map from one orbit has the same property with respect to target orbits. The map test on orbits therefore gives a bijection on all hom sets, including maps with empty source or target. Realizing each orbit by its fixed field gives every finite continuous G_k-set up to isomorphism. □

There is also a reconstruction of the algebra using the whole set of points at once. For a finite continuous G_k-set S, form

$$A(S) = \{b\colon S \to k^{\mathrm{sep}} \mid b(g \cdot s) = g(b(s)) \text{ for all } g, s\}.$$

It is a k-algebra under pointwise operations. On an orbit G_k/H, evaluation at H identifies it with L_H. An element $a \in L_H$ extends uniquely by $gH \mapsto g(a)$, a formula well-defined because H fixes a. On a union of orbits this gives a product of the fixed fields. An equivariant map $v\colon S \to T$ gives a ring map $v^*\colon A(T) \to A(S)$ by precomposition. For each $s \in S$ evaluation at s is one geometric point of $\operatorname{Spec} A(S)$, and the following diagram relates it to evaluation at $v(s)$.

$$\begin{array}{ccc} A(T) & \xrightarrow{v^*} & A(S) \\ & \searrow^{\mathrm{ev}_{v(s)}} & \downarrow{\scriptstyle \mathrm{ev}_s} \\ & & k^{\mathrm{sep}} \end{array}$$

The orbit calculation proves that every geometric point arises uniquely in this way. Thus evaluation exhibits $\operatorname{Spec} A(S)$ as the reconstruction functor, with S recovered naturally. For $S = F(\operatorname{Spec} A)$, the map $a \mapsto (\varphi \mapsto \varphi(a))$ identifies A with $A(S)$. Evaluating at one embedding per field factor identifies the corresponding component with the image of that field factor.

48.4 Restriction and factorization over finite fields

Let $k \subseteq L \subseteq k^{\mathrm{sep}}$ be a finite separable extension of k. The same separable closure can be used for L. If $X = \operatorname{Spec} A$, the bijection

$$\operatorname{Hom}_{L\text{-alg}}(A \otimes_k L, k^{\mathrm{sep}}) \cong \operatorname{Hom}_{k\text{-alg}}(A, k^{\mathrm{sep}})$$

sends an L-map to its restriction to A. An algebra map φ on the right extends by $a \otimes \lambda \mapsto \varphi(a)\lambda$ with the chosen inclusion of L. This bijection is equivariant for $G_L \leq G_k$, since G_L fixes λ. Thus scalar extension restricts the group action. Finite products over k become products of geometric fibers with the diagonal action, by the universal property of the product.

Let $k = \mathbb{F}_q$, where q is a prime power. Its separable closure is the union of the fields $\mathbb{F}_{q^r}$, and the arithmetic Frobenius $a \mapsto a^q$ generates their Galois groups over $\mathbb{F}_q$. In degree r its order is r. Its r-th power fixes the field, and for $0 < j < r$ the polynomial $t^{q^j} - t$ has too few roots to vanish on the field. Since the degree is r, these powers give all r automorphisms. Taking inverse limits over divisibility of r gives $G_{\mathbb{F}_q} \cong \varprojlim_r \mathbb{Z}/r$, with Frobenius corresponding to the topological generator 1. A finite continuous action is therefore specified by a permutation, and each cycle of length d corresponds to one factor $\mathbb{F}_{q^d}$.

Example 48.2 (A quartic splits into two quadratic factors)**.** Over $\mathbb{F}_2$, the polynomial $t^4 + t + 1$ has no roots. The only monic irreducible quadratic is $t^2 + t + 1$, and reduction modulo it gives $t^4 + t + 1 \equiv 1$. Hence the quartic is irreducible. Its derivative is 1, so the spectrum of its quotient algebra has a geometric fiber of four roots permuted transitively by Frobenius. Extension to $\mathbb{F}_4$ restricts the action to the group generated topologically by the square of Frobenius. Number the roots $0, 1, 2, 3$ in Frobenius order. The square acts as the two cycles (0 2) and (1 3), giving two quadratic factors over $\mathbb{F}_4$.

To identify the factors, choose $\omega \in \mathbb{F}_4$ with $\omega^2 + \omega + 1 = 0$. Then

$$t^4 + t + 1 = (t^2 + t + \omega)(t^2 + t + \omega^2).$$

Expanding gives the quartic because $\omega + \omega^2 = 1$ and $\omega\omega^2 = 1$. The map $a \mapsto a^2 + a$ on $\mathbb{F}_4$ has image $\{0, 1\}$, as evaluation at $0, 1, \omega, \omega^2$ shows. Both factors are therefore irreducible. The scalar-extended quotient algebra is $\mathbb{F}_{16} \times \mathbb{F}_{16}$, with one field for each of the two cycles.

Products offer a further calculation with the same permutation. On the geometric fiber of $\operatorname{Spec} \mathbb{F}_{q^m} \times_{\operatorname{Spec} \mathbb{F}_q} \operatorname{Spec} \mathbb{F}_{q^n}$, Frobenius sends (i, j) to $(i+1, j+1)$ in $\mathbb{Z}/m \times \mathbb{Z}/n$. Every cycle has length $\operatorname{lcm}(m, n)$, and there are $\gcd(m, n)$ cycles. Thus

$$\mathbb{F}_{q^m} \otimes_{\mathbb{F}_q} \mathbb{F}_{q^n} \cong \prod_{i=1}^{\gcd(m,n)} \mathbb{F}_{q^{\operatorname{lcm}(m,n)}}.$$

A tensor product calculation is thereby reduced to finding orbits of a diagonal action, the operation studied in Chapter 9.

48.5 Reconstructing the group from the functor

The group itself can be recovered once the functor to finite sets is specified. For a profinite group G, let U forget the actions of finite continuous G-sets. Each $g \in G$ gives a natural automorphism of U by $s \mapsto g \cdot s$. Naturality uses precisely that the maps in the source of U are equivariant. The components are bijections of the underlying sets.

For the converse, let α be a natural automorphism of U. For every open normal subgroup N, consider the object G/N with its left regular action and put $g_N = \alpha_{G/N}(N)$. Naturality for right translations, which are equivariant, implies $\alpha_{G/N}(x) = g_N x$ on this finite quotient. Naturality for quotient maps makes the g_N compatible. They therefore give an element $g \in G$. On any finite continuous G-set S the action factors through one such quotient. For $s \in S$, the equivariant orbit map $x \mapsto x \cdot s$ gives the square

$$\begin{array}{ccc} G/N & \xrightarrow{\alpha_{G/N}} & G/N \\ {\scriptstyle x \mapsto x \cdot s}\downarrow & & \downarrow{\scriptstyle x \mapsto x \cdot s} \\ S & \xrightarrow{\alpha_S} & S. \end{array}$$

At the identity coset, it gives $\alpha_S(s) = g \cdot s$. This proves $\mathrm{Aut}(U) \cong G$. The isomorphism respects the inverse-limit topology when $\mathrm{Aut}(U)$ is given the topology of its finite components. The regular objects G/N already detect the restrictions to all finite quotients.

For a connected scheme X, choose a geometric point $\bar{x}\colon \operatorname{Spec}\Omega \to X$ with Ω algebraically closed. A finite étale morphism $Y \to X$ has finite fiber

$$F_{\bar{x}}(Y) = \mathrm{Hom}_X(\operatorname{Spec}\Omega, Y).$$

The connected-scheme Galois category theorem states that $\pi_1^{\text{ét}}(X, \bar{x}) = \mathrm{Aut}(F_{\bar{x}})$ is profinite and that this functor, equipped with the natural action, gives an equivalence

$$\mathrm{FEt}(X) \simeq \mathbf{FinSet}_{\mathrm{cont}}^{\pi_1^{\text{ét}}(X,\bar{x})}.$$

Here $\mathrm{FEt}(X)$ contains all finite étale schemes over X, with arrows over X. This is a theorem for connected schemes with the specified geometric point [39, Theorem 58.6.2]. In the field case, geometric points valued in an algebraic closure give the same fiber sets used above, because the images of finite separable fields lie in the separable closure inside it. Theorem 48.1 and the reconstruction of G identify the field's fundamental group with G_k.

For a map $f\colon X' \to X$ of connected schemes with geometric points related by $f \circ \bar{x}' = \bar{x}$, pulling back a finite étale cover over X gives a cover over X'. The two fibers are canonically the same set by the universal property of pullback. Restricting a natural automorphism of $F_{\bar{x}'}$ to these pullback covers therefore gives a continuous homomorphism $\pi_1^{\text{ét}}(X', \bar{x}') \to \pi_1^{\text{ét}}(X, \bar{x})$. The pullback functor corresponds to restricting finite actions along that homomorphism. The two descriptions of restriction follow from the same operation of testing a cover at a geometric point.

Example 48.3 (A cyclic cover over an algebraically closed field)**.** Let k be algebraically closed and choose $n \geq 2$ invertible in k. Over $X = \operatorname{Spec} k[t, t^{-1}]$, take $Y = \operatorname{Spec} k[u, u^{-1}]$ with $t \mapsto u^n$. The algebra has the presentation $k[t, t^{-1}][u]/(u^n - t)$, since $u^n = t$ makes u invertible. The monic relation gives a free module with basis $1, u, \ldots, u^{n-1}$. Its derivative nu^{n-1} is invertible, which is the standard étale criterion for this presentation [9, Definition 10.144.1]. Thus the morphism is finite étale.

At the point $t = 1$, the pullback square is

$$\begin{array}{ccc} \coprod_{\zeta^n=1} \operatorname{Spec} k & \longrightarrow & Y \\ \downarrow & \lrcorner & \downarrow \\ \operatorname{Spec} k & \xrightarrow{t=1} & X. \end{array}$$

The upper left corner splits because $u^n - 1$ has n distinct roots. The cover itself is connected because the Laurent polynomial algebra of Y is an integral domain. Each $\zeta^n = 1$ gives a deck transformation by $u \mapsto \zeta u$, and these transformations act freely and transitively on the fiber. They exhaust the deck transformations. Indeed a $k[t, t^{-1}]$-automorphism sends u to an element v with $(v/u)^n = 1$. Factoring this equation over k in the integral domain forces v/u to be one of the roots.

In the pointed Galois category, the transitive fiber is G/H for $G = \pi_1^{\text{ét}}(X, t = 1)$. Its equivariant automorphisms have cardinality $[N_G(H) : H]$ by the same coset map test as over a field. There are $n = [G : H]$ of them here, so $N_G(H) = G$. Thus this cover exhibits a quotient of G isomorphic to the cyclic group of n-th roots of unity in k.

48.6 Pontryagin duality for profinite abelian groups

The abelian quotients of a profinite Galois group can be probed by characters with finite image. Let P be a profinite abelian group, and write $\widehat{P}$ for its continuous characters into the circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, with the compact-open topology from Section 16.6. A sufficiently small open arc about 0 contains no nonzero

subgroup of $\mathbb{T}$. The inverse image of that arc under any character contains an open subgroup $U \subseteq P$, so the character kills U and comes from the finite group P/U. The compact-open condition that the image of all of P lie in the arc isolates the zero character, so $\widehat{P}$ is discrete. Consequently, with transitions given by pullback along quotient maps,

$$\widehat{P} \cong \operatorname*{colim}_{U} \widehat{P/U}.$$

Conversely, a discrete torsion abelian group A is the filtered union of its finite subgroups B, since any finitely generated torsion abelian group is finite. A character on A is a compatible family of characters on the B. Each compact subset of A is finite, so the compact-open topology on $\widehat{A}$ gives precisely the inverse limit topology on $\lim_B \widehat{B}$. Together with bidual evaluation this gives the contravariant equivalence between profinite abelian groups and discrete torsion abelian groups [65, Theorem 1.1.11].

Example 48.4 (Characters of Frobenius)**.** The arithmetic Frobenius identifies $G_{\mathbb{F}_q}$ with the profinite completion $\varprojlim_n \mathbb{Z}/n$ of $\mathbb{Z}$. At a finite quotient $\mathbb{Z}/n$, a character assigns $a/n + \mathbb{Z}$ to 1 mod n, as in Section 8.6. For $n \mid m$, the transition between character groups comes from reduction:

$$\mathbb{Z}/m \xrightarrow{\text{mod } n} \mathbb{Z}/n$$

$$\widehat{\mathbb{Z}/m} \xleftarrow{\text{pullback}} \widehat{\mathbb{Z}/n}.$$

Under the identification by the value on Frobenius, this pullback is inclusion of the n-torsion in the m-torsion of the circle. Taking the union gives $\widehat{G_{\mathbb{F}_q}} \cong \mathbb{Q}/\mathbb{Z}$ as a discrete group.

CHAPTER 49

Classifying toposes from geometric theories

A sheaf can assemble local solutions of a system of equations, with different solutions on different opens. This makes a topos a setting in which to interpret a theory. For certain theories there is a topos with a universal solution, such that every solution in every other Grothendieck topos is its inverse image. We first identify the logical operations that travel along inverse images and then build the classifier for a single proposition.

49.1 Formulas that travel by inverse image

Take a set-sized signature of sorts, function symbols, and relation symbols of finite arity. A *geometric formula* is built from atomic formulas, including equality of terms, by finite conjunctions, set-indexed disjunctions, and existential quantification. The empty conjunction is $\top$, and the empty disjunction is $\bot$. A geometric theory $\mathbb{T}$ gives axioms in the form of *sequents*

$$\phi \;\vdash_{\vec{x}}\; \psi.$$

Both formulas have their free variables among the finite context $\vec{x}$. For example, that an endomorphism s of a sort A is surjective is the sequent $\top \vdash_{y:A} \exists x\,(s(x) = y)$. Its context specifies an inclusion for every value of y.

A model in a Grothendieck topos $\mathcal{E}$ assigns objects to sorts, arrows to function symbols, and subobjects to relation symbols. The product $A_{\vec{x}} = \prod_i A_i$ of the interpreted sorts interprets the context. Equality pulls back a diagonal and conjunction intersects subobjects, as in Chapter 21. Disjunction takes the join of the subobjects in the context. For $\exists y\,\phi(\vec{x}, y)$, take the image of the predicate for ϕ under the projection that drops y. This last operation is the existential adjoint of Chapter 24. For example, the surjectivity sequent says that the image of s is the whole of A. An axiom holds in M when the interpreted subobjects satisfy $\phi_M \le \psi_M$. A homomorphism of models is a family of maps on the sorts commuting with the functions and taking each relation into the corresponding relation.

A geometric morphism $f\colon \mathcal{E} \to \mathcal{F}$ has an adjunction $f^* \dashv f_*$ with left exact inverse image $f^*\colon \mathcal{F} \to \mathcal{E}$. Thus f^* preserves finite limits and all small colimits.

It preserves a mono by the pullback test for its diagonal. It also preserves images. Indeed, an image factorization is obtained by coequalizing the kernel pair of the map, then including the quotient in the target. For sheaves of sets, an epimorphism in such a factorization has local lifts of every section. A map out of its source that agrees on the kernel pair gives the same result on different local lifts. Those results therefore glue to a map out of the target. This gives the coequalizer description. Inverse image preserves the kernel pair by finite limits and its coequalizer by colimits. For the image of $B \to A$ this gives the following comparison, where both rows are image factorizations.

$$\begin{array}{ccccc} f^*B & \longrightarrow & f^*\,\mathrm{im}(B \to A) & \hookrightarrow & f^*A \\ \| & & \downarrow\cong & & \| \\ f^*B & \longrightarrow & \mathrm{im}(f^*B \to f^*A) & \hookrightarrow & f^*A. \end{array}$$

A join of subobjects is the image of the map from their coproduct. This description follows from the universal property of the image: the coproduct map factors through a subobject precisely when all its summand maps do. Hence inverse image preserves the joins as well. It preserves the operations that interpret geometric formulas and thus takes models of $\mathbb{T}$ in $\mathcal{F}$ to models of $\mathbb{T}$ in $\mathcal{E}$.

49.2 The universal model and its direction

A *classifying topos* for $\mathbb{T}$ is a Grothendieck topos $\mathcal{B}_\mathbb{T}$ with a model $U_\mathbb{T}$ such that the assignment

$$f \longmapsto f^*U_\mathbb{T} \qquad \text{gives} \qquad \mathrm{Geom}(\mathcal{E}, \mathcal{B}_\mathbb{T}) \simeq \mathrm{Mod}_\mathbb{T}(\mathcal{E})$$

for every Grothendieck topos $\mathcal{E}$, naturally with respect to inverse images. In this formula a map between geometric morphisms f, g is taken to be a natural transformation $f^* \Rightarrow g^*$. Such a transformation gives maps on the sorts and commutes with the structure of $U_\mathbb{T}$ by naturality. For a model M in $\mathcal{E}$, its classifying morphism therefore points from $\mathcal{E}$ to $\mathcal{B}_\mathbb{T}$. The model moves in the opposite direction, by the inverse-image functor.

For $g\colon \mathcal{D} \to \mathcal{E}$, the composite $f \circ g$ has inverse image g^*f^*. The model classified by $f \circ g$ is consequently g^*M.

$$\begin{array}{ccc} \mathcal{D} & \xrightarrow{g} & \mathcal{E} \\ & \searrow^{f\circ g} & \downarrow f \\ & & \mathcal{B}_\mathbb{T} \end{array} \qquad\qquad \begin{array}{ccc} & & U_\mathbb{T} \\ & & \top \\ & & \downarrow f^* \\ g^*M & \xleftarrow{g^*} & M \end{array}$$

In particular, a model in sets is obtained from a point $\mathbf{Set} \to \mathcal{B}_\mathbb{T}$. The logic–semantics adjunction in Chapter 22 compares sentences and their satisfying

models in a fixed setting. The classifying property concerns a whole family of categories of models, with inverse image relating their interpretations across topoi.

The classifying property determines the pair up to equivalence. Given a second classifying pair, use the universal model in each topos to find a geometric morphism into the other. Each composite pulls the universal model back to an isomorphic copy of itself. Full faithfulness of the defining equivalence then makes the composite isomorphic to the identity morphism.

49.3 The site of definable objects

For a set-sized geometric theory, the classifying topos can be constructed from the *geometric syntactic site* $(\mathcal{C}_\mathbb{T}, J_\mathbb{T})$. An object $\{\vec{x} \mid \phi\}$ specifies values of the context satisfying ϕ. An arrow to $\{\vec{y} \mid \psi\}$ is given by a formula $\theta(\vec{x}, \vec{y})$ that the theory proves to be the graph of a total single-valued relation between these objects. Formulas defining the same relation by provable equivalence specify the same arrow. For two such arrows, the composite graph is obtained by conjoining their graph formulas and existentially binding the shared intermediate context. For example,

$$(\eta \circ \theta)(\vec{x}, \vec{z}) \quad \text{is represented by} \quad \exists \vec{y}\, (\theta(\vec{x}, \vec{y}) \wedge \eta(\vec{y}, \vec{z})).$$

The empty true context is terminal. Products use disjoint concatenations of contexts and conjunction, and equalizers impose equality of the two outputs. These give finite limits. For the size of the site, geometric formulas can be put in the form of disjunctions of formulas built using only atoms, finite conjunction, and existential quantification. This reduces size to a set of choices of subsets of a set of finite expressions. Finite meets distribute over joins and existential quantification commutes with joins, so this reduction follows by recursion on formulas.

A family of arrows into $\{\vec{y} \mid \psi\}$ with graph formulas $\theta_i(\vec{x}_i, \vec{y})$ covers if the following sequent is provable:

$$\psi \;\vdash_{\vec{y}}\; \bigvee_i \exists \vec{x}_i\, \theta_i(\vec{x}_i, \vec{y}).$$

Each graph already lands in ψ, so this says that the image formulas jointly exhaust the target. Pulling back the family substitutes the context and pulls back these image predicates. Existential base change gives the pullback condition on covering sieves. For the transitivity condition, coverage of each member of a cover implies coverage after composing the two families by distribution and repeated existential quantification. The identity exhausts its target, giving the maximal-sieve axiom. The covering families generate $J_\mathbb{T}$ as in Chapter 45.

Theorem 49.1 (Geometric syntactic site)**.** *The topos* $\mathrm{Sh}(\mathcal{C}_{\mathbb{T}}, J_{\mathbb{T}})$ *classifies* $\mathbb{T}$*. The universal model sends each sort to the sheafified representable of its one-variable true context. Its functions and relations are induced by their defining graphs and subobjects in the syntax.*

Here is the route from the site to the universal property. Interpretation in a model in $\mathcal{E}$ gives a functor $\mathcal{C}_{\mathbb{T}} \to \mathcal{E}$. Functional relations interpret as maps. Contexts and the constructions of finite limits show that this functor is left exact. The coverage condition becomes joint epimorphy. Conversely, a left exact functor taking covers to jointly epimorphic families interprets the sorts, maps, and atomic relations. It interprets conjunction by finite limits. The inclusions of the disjuncts cover the formula for their disjunction, so their images join to that formula. The projection to the formula for an existential quantification is a cover. It follows that its interpretation is the image under the projected map. These interpretations respect the axioms, since an axiom gives a syntactic inclusion between the two predicates.

The passage from sites to morphisms uses the finite-limit-site form of Diaconescu's theorem. If a small site has finite limits, inverse image on sheafified representables identifies geometric morphisms into its sheaf topos with left exact functors on the site that send covers to jointly epimorphic families. For a functor F of this kind, the resulting direct image sends an object Y of $\mathcal{E}$ to the sheaf $\mathrm{Hom}_{\mathcal{E}}(F(-), Y)$. Matching sections glue because the image of a covering family is jointly epic and its pairwise pullbacks test agreement. The inverse image is obtained by extending F by colimits. The theorem supplies left exactness of this extension and identifies its natural transformations with those of the functors on the site. Applying it to the interpretation above gives the classifying equivalence. Mac Lane and Moerdijk give the full construction in Chapters VIII and X [16]. Caramello develops how choices of sites encode the same classifying topos [40, Chapter 2].

49.4 A topos classifying a single proposition

Consider the theory with no sorts, one nullary relation symbol p, and no axioms. A model in $\mathcal{E}$ is a subobject $P \hookrightarrow 1$ of the terminal object. A map between such models exists exactly when $P \leq Q$ in $\mathrm{Sub}(1)$. We can identify its classifying topos directly.

Let D have the arrow $0 \to 1$ and its two identities. Its presheaf topos $\widehat{D}$ is the category of restrictions $\rho_X \colon X_1 \to X_0$ studied in Section 23.5. Write y_0, y_1 for the two representables. The presheaf y_1 is terminal, while y_0 is $\varnothing \to \{*\}$, with a unique inclusion into y_1. On the Sierpiński space $\mathbb{S}$ of Chapter 17, a sheaf consists of its sets on $\mathbb{S}$ and its one nonempty proper open, with a restriction map between them. The empty open is assigned a singleton, and each cover

of a nonempty open contains the open itself. Thus $\widehat{D} \simeq \mathrm{Sh}(\mathbb{S})$ and y_0 is the subterminal of the nonempty proper open.

The syntax gives the same two-stage description. Every formula reduces by the meet and join laws to $\bot$, p, or $\top$. The arrows are entailments, forming the chain $\bot \leq p \leq \top$. A family covers its target when its join is the target, so every cover of p or $\top$ includes the target itself, and the empty family covers $\bot$. A sheaf therefore assigns a singleton to $\bot$ and has an arbitrary restriction from $\top$ to p. These are the two stages of $\widehat{D}$.

Proposition 49.2. *Pulling back $y_0 \hookrightarrow y_1$ gives an equivalence*

$$\mathrm{Geom}(\mathcal{E}, \widehat{D}) \simeq \mathrm{Sub}_{\mathcal{E}}(1)$$

for each Grothendieck topos $\mathcal{E}$.

Proof. For a set I and an object B, write $I \cdot B$ for the coproduct of I copies of B. Any presheaf X fits into the pushout

$$\begin{array}{ccc} X_1 \cdot y_0 & \longrightarrow & X_1 \cdot y_1 \\ {\scriptstyle \rho_X \cdot y_0}\downarrow & & \downarrow \\ X_0 \cdot y_0 & \longrightarrow & X. \end{array}$$

The top arrow uses the inclusion of representables. At stage 1 the pushout has set X_1, since both terms using y_0 are empty there. At stage 0 the top arrow is the identity of X_1, and the pushout has set X_0. Its restriction is ρ_X. In particular a colimit-preserving functor that takes y_0 to P and y_1 to 1 must take this pushout to

$$L_P(X) = (X_1 \cdot 1) \sqcup_{X_1 \cdot P} (X_0 \cdot P).$$

The maps again use $P \hookrightarrow 1$ and ρ_X.

We show that this formula defines an inverse image. Use a presentation of $\mathcal{E}$ as sheaves on a small site $(\mathcal{A}, J)$. Regard P as a subterminal sheaf. Before sheafifying, the same pushout in the presheaf category has values

$$M_P(X)(a) = \begin{cases} X_1 & P(a) = \varnothing, \\ X_0 & P(a) \text{ is a singleton.} \end{cases}$$

Across an arrow $a' \to a$, the restriction is the identity when the same stage is used at both objects. When the stage changes from 1 to 0, it is ρ_X. This is the only possible change because a section of P restricts to a section. Thus M_P preserves finite limits, since at each site object it evaluates X at one of the two stages and limits of presheaves are computed stagewise. Sheafification takes

this presheaf pushout to the displayed $L_P(X)$ and preserves finite limits by left exactness. So L_P preserves finite limits.

For an object Y of $\mathcal{E}$, maps from $L_P(X)$ to Y are pairs of functions from X_1 to $\mathrm{Hom}(1, Y)$ and from X_0 to $\mathrm{Hom}(P, Y)$ agreeing over $X_1 \cdot P$. This is precisely the condition for a map from X to the presheaf $R_P(Y)$ with restriction $\mathrm{Hom}(1, Y) \longrightarrow \mathrm{Hom}(P, Y)$ given by precomposition. The bijection is natural and gives $L_P \dashv R_P$. Therefore L_P is the inverse image of a geometric morphism, and $L_P(y_0) \hookrightarrow L_P(y_1)$ is $P \hookrightarrow 1$.

Conversely, any geometric inverse image must send the inclusion of representables to some $P \hookrightarrow 1$. Since inverse image preserves colimits, the pushout gives it the formula for L_P, naturally in X. A natural transformation $L_P \Rightarrow L_Q$ is determined by its components on y_0, y_1, again by the pushout. At y_1 the component is the identity of 1; at y_0 it must be the map over 1 witnessing $P \leq Q$. Conversely any such map induces the transformation by the pushout. These correspondences give the claimed equivalence. □

49.5 What the classifier records

Example 49.3 (One proposition on a space)**.** In $\mathrm{Sh}(X)$, a subterminal corresponds to an open $V \subseteq X$. There is a single local section on $W \subseteq V$, and no section on any other open W. Conversely the opens on which a subterminal has a section are closed under restriction and under unions by the sheaf axiom. Their union is the required V. Proposition 49.2 therefore identifies the models on X with maps $\mathrm{Sh}(X) \to \mathrm{Sh}(\mathbb{S})$ up to isomorphism. Let $h \colon X \to \mathbb{S}$ take V to the open point and its complement to the other point. Its direct image sends a sheaf Y to the arrow $Y(X) \to Y(V)$, by taking sections over the inverse images of the two nonempty opens. This is $R_P(Y)$. Indeed a map $P \to Y$ is determined by a section of Y on V, with all other components given by restriction. Thus the geometric morphism induced by h is the one classified by V.

Example 49.4 (Why negation needs a separate test)**.** Take $\mathcal{E} = \mathbf{Set}$ and $P = \varnothing$. The formula for L_P gives the geometric inverse image $X \mapsto X_1$. In the two-stage topos $\widehat{D}$, the negation of y_0 is the bottom subterminal. Hence both y_0 and its negation go to the empty set under L_P. In sets the negation of the empty subterminal is the terminal object, which L_P obtains from y_1. Thus the geometric preservation statement applies to the listed operations while this inverse image fails to preserve negation.

The syntactic site keeps the operations of formulas visible. The classifying topos keeps the same interpretations visible after changing the site. Instances such as the single proposition let us calculate both descriptions with diagrams of sets.

Part X Homology, cohomology, and derived categories

CHAPTER 50

Exactness from kernels and cokernels

A kernel describes the inputs a map kills. An image describes the outputs that another map can supply. Exactness says that at a given object these descriptions agree. In modules, one can test the agreement by using elements. In an abelian category one can make the same comparison by universal properties. The snake lemma will also show how a comparison of two exact sequences produces a new map between a kernel and a cokernel.

50.1 An abelian category

In an *additive category*, each hom set is an abelian group, composition is additive in each variable, and there are a zero object and finite biproducts. The zero maps are the additive zeros of the hom groups. The binary biproduct $A \oplus B$ has injections j_A, j_B and projections p_A, p_B. They satisfy

$$p_A j_A = \mathrm{id}_A, \quad p_B j_B = \mathrm{id}_B, \quad p_A j_B = p_B j_A = 0, \qquad j_A p_A + j_B p_B = \mathrm{id}_{A \oplus B}\,.$$

The projections exhibit a product and the injections exhibit a coproduct. Maps into and out of the biproduct can therefore be specified by columns and rows, whose composites use addition and composition of the entries.

Suppose kernels and cokernels exist. As in Chapter 13, take them to be the equalizer and coequalizer with the zero map. For $f\colon A \to B$, define $\operatorname{coim} f = \operatorname{coker}(\ker f \to A)$ and $\operatorname{im} f = \ker(B \to \operatorname{coker} f)$. The map f factors through $\operatorname{im} f$ because the cokernel projection kills it. The induced map from A to $\operatorname{im} f$ kills $\ker f$, since composing it with the image inclusion gives f and that inclusion is monic. It therefore factors through $\operatorname{coim} f$.

$$A \xrightarrow{\quad f \quad} B \qquad A \twoheadrightarrow \operatorname{coim} f \xrightarrow{\overline{f}} \operatorname{im} f \hookrightarrow B$$

An *abelian category* is an additive category with all kernels and cokernels in which this canonical $\overline{f}$ is an isomorphism for every f. The category of left modules over any ring satisfies the definition by the image–coimage theorem of Chapter 13. The category **Ab** of abelian groups is the case of $\mathbb{Z}$-modules.

A mono has zero kernel by cancellation, so its coimage is its source. The factorization above identifies the mono with the kernel of its cokernel, up to the

unique isomorphism over its target. Dually an epi is the cokernel of its kernel. A map with zero kernel is monic, since any difference of maps that it sends to zero factors through the zero object. A map with zero cokernel is epic by the dual argument. Equalizers are kernels of differences, so biproducts and kernels also give finite limits. For example, the pullback of $p\colon B \to C$ and $k\colon K \to C$ is the kernel of $(p, -k)\colon B \oplus K \to C$. Maps into this kernel correspond to pairs b, v with $pb = kv$.

50.2 Short exact sequences and splitting

Let $A \xrightarrow{i} B \xrightarrow{p} C$ have $pi = 0$. The induced inclusion $\operatorname{im} i \to \ker p$ is an isomorphism exactly when the sequence is *exact at* B. Exactness at all three objects of

$$0 \longrightarrow A \xrightarrow{i} B \xrightarrow{p} C \longrightarrow 0$$

says in addition that i is monic and p epic. Thus a short exact sequence exhibits i as a kernel of p and p as a cokernel of i. Indeed $\operatorname{im} i = i$ as a subobject, and p is the cokernel of its kernel. Conversely, having these two universal properties makes the sequence short exact.

A section $s\colon C \to B$ of p makes the sequence split. Since $p(\mathrm{id}_B - sp) = 0$, there is a unique $r\colon B \to A$ such that $ir = \mathrm{id}_B - sp$. This gives $ri = \mathrm{id}_A$ and $rs = 0$ by composing with i and cancelling it. The row and column below are inverse by the biproduct equations.

$$A \oplus C \underset{(r,p)}{\overset{(i,s)}{\rightleftarrows}} B$$

In these coordinates i is the injection of A and p the projection to C. A retraction r of i also suffices to split the sequence. In fact $\mathrm{id}_B - ir$ kills i, hence equals sp for a unique s. Then $psp = p$, so $ps = \mathrm{id}_C$ by cancellation of p. A splitting is therefore an additional choice of section or retraction. For example, the sequence

$$0 \longrightarrow \mathbb{Z}/2 \xrightarrow{i} \mathbb{Z}/4 \xrightarrow{p} \mathbb{Z}/2 \longrightarrow 0, \qquad i(\overline{1}) = \overline{2}, \quad p(\overline{1}) = \overline{1},$$

is exact. Every lift of $\overline{1}$ along p has order four, so no lift can be its image under a section of groups.

50.3 The connecting map

Consider a commutative diagram whose two rows are short exact.

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A & \xrightarrow{i} & B & \xrightarrow{p} & C & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle a} & & \downarrow{\scriptstyle b} & & \downarrow{\scriptstyle c} & & \\ 0 & \longrightarrow & A' & \xrightarrow{i'} & B' & \xrightarrow{p'} & C' & \longrightarrow & 0. \end{array}$$

Write K_a, K_b, K_c for the kernels and Q_a, Q_b, Q_c for the cokernels of the three vertical maps. The squares induce maps $K_a \to K_b \to K_c$ and $Q_a \to Q_b \to Q_c$ by the kernel and cokernel properties. The new map has source K_c and target Q_a.

For modules, start with $z \in K_c$ and lift it along p to $y \in B$. Since $p'b(y) = c(z) = 0$, there is a unique $x' \in A'$ with $i'(x') = b(y)$. A different lift is $y + i(x)$ for some $x \in A$. Its corresponding element of A' is $x' + a(x)$, so the class of x' in Q_a depends only on z. This is the connecting map ∂. The same procedure on a sum uses the sum of the chosen lifts, and scalar multiplication uses the corresponding multiple. Thus the map is linear.

To carry out the construction in an abelian category, let $k\colon K_c \hookrightarrow C$ and form the pullback shown below.

$$\begin{array}{ccc} \widetilde{B} & \xrightarrow{j} & B \\ \downarrow{\scriptstyle r} & \lrcorner & \downarrow{\scriptstyle p} \\ K_c & \xrightarrow{k} & C. \end{array}$$

The projection r is epic. To see the pullback property for epis in general, take p epic and k arbitrary. If $h\colon K_c \to T$ satisfies $hr = 0$, the map $(0, h)\colon B \oplus K_c \to T$ kills the kernel $\widetilde{B}$ of $(p, -k)$. The latter row is epic because its composite with the first injection is p. It is consequently the cokernel of its kernel, so $(0, h) = g(p, -k)$ for a map $g\colon C \to T$. The first component gives $gp = 0$, hence $g = 0$ and $h = 0$. Thus r has zero cokernel and is epic. The kernel of r comes from $i\colon A \to B$ paired with the zero map $A \to K_c$, by the pullback property and $i = \ker p$.

Now $p'bj = ckr = 0$, so $bj = i'd$ for a unique $d\colon \widetilde{B} \to A'$. On the kernel of r, the map d restricts to a by the left square of the original diagram. Thus the composite of d with the quotient $q_a\colon A' \to Q_a$ kills $\ker r$. Since r is its cokernel, there is a unique $\partial\colon K_c \to Q_a$ with $\partial r = q_a d$. In modules this equation recovers the formula on lifts. The equation also proves naturality in a map of the original two-row diagrams. The induced map of pullbacks respects d by the kernel property of i'. Hence the two relevant maps into the target cokernel agree after precomposition with r, and r is epic.

50.4 The snake lemma

Theorem 50.1. *The induced maps and the connecting map give an exact sequence*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & K_a & \longrightarrow & K_b & \longrightarrow & K_c & & \\ & & & & & & \downarrow{\scriptstyle \partial} & & \\ 0 & \longleftarrow & Q_c & \longleftarrow & Q_b & \longleftarrow & Q_a. & & \end{array}$$

Here the top row is read from left to right, then the bottom row from right to left.

Proof. We can chase maps $T \to X$ in place of individual elements. An epi need not lift a map $T \to Y$ directly. Pulling the epi back to T gives a lift after precomposing with an epi $T' \to T$, by the argument in the previous section. A map into X factors through $\operatorname{im} f \subseteq X$ whenever it does so after such precomposition. Its composite with $\operatorname{coker} f$ is then zero by cancellation. Conversely, a map into the image lifts through f after an epi by the image factorization and pullback. To establish the inclusion of a kernel in an image, apply the test to the kernel arrow itself. In modules the same chase can be read on individual elements, where epis are surjections.

The map $K_a \to K_b$ is monic by monicity of i. If y in K_b maps to zero under p, write $y = i(x)$. Then $i'a(x) = b(y) = 0$, hence x is in K_a. Conversely $i(x)$ for x in K_a maps to zero in K_c.

At K_c, an image of y in K_b has zero connecting class, since y gives a lift with $b(y) = 0$. For the converse, start with z in K_c with zero connecting class. Lift to y along p, so $b(y) = i'(x')$. Since $q_a(x') = 0$, lift x' through a to x. Then $y - i(x)$ lies in K_b and maps to z.

At Q_a, a connecting class represented by x' has zero image in Q_b since $i'(x') = b(y)$. Conversely, lift a class with zero image to $x' \in A'$. Lift $i'(x')$ through b to y, which is possible after an epi because $q_b i'(x') = 0$. Then $z = p(y)$ lies in K_c and connects to the given class.

At Q_b, the composite out of Q_a through Q_b is zero because $p'i' = 0$. Conversely, lift a class with zero image in Q_c to y'. Since $q_c p'(y') = 0$, lift $p'(y')$ through c to z. Lift z along p to y. Now $y' - b(y)$ maps to zero under p', so it is $i'(x')$ for some x', whose class gives the original one. Finally $Q_b \to Q_c$ is epic, since precomposing it with q_b gives $q_c p'$, a composite of epis. □

The Stacks Project gives the snake lemma for exact rows without requiring a monomorphism at the start of the upper row or an epimorphism at the end of the lower row [42, Lemma 12.5.17].

Example 50.2 (A nonzero connecting map)**.** Use the sequence for $\mathbb{Z}/4$ above for both rows, with vertical maps multiplication by 2. The outside vertical maps are zero. The class $\overline{1} \in K_c = \mathbb{Z}/2$ lifts to $\overline{1} \in \mathbb{Z}/4$, whose double is $i(\overline{1})$. Thus ∂ is the identity from $\mathbb{Z}/2$ to $\mathbb{Z}/2$. In the snake sequence $K_a \to K_b$ is an isomorphism because the kernel of doubling in $\mathbb{Z}/4$ is the even subgroup. The map $Q_b \to Q_c$ is also an isomorphism since p quotients by the even subgroup. The maps on either side of ∂ are zero.

50.5 Exactness of functors

A functor between abelian categories is additive if it preserves addition of arrows. It preserves zero maps and zero objects. An additive functor is *left exact* if it

preserves kernels, *right exact* if it preserves cokernels, and *exact* if it preserves both. An exact functor preserves short exact sequences by their two universal properties. Conversely, preserving short exact sequences suffices for an additive functor to be exact. Factor f into an epi onto its image and a mono from there to its target. From the two short exact sequences of the factorization, the functor preserves the kernel of the first arrow and the cokernel of the second. After applying the functor the second arrow is monic and the first epic. Thus these are the kernel and cokernel of the functor's value on f. A left exact functor preserves exactness in $0 \to A \to B \to C$, since the map $A \to B$ is the kernel of the one following it. A right exact functor preserves the dual pattern $A \to B \to C \to 0$.

For example, $\mathrm{Hom}_R(P, -)$ takes values in abelian groups, with addition of maps defined pointwise. It preserves kernels, since a map $P \to B$ killed by $B \to C$ factors uniquely through the kernel. With $R = \mathbb{Z}$ and $P = \mathbb{Z}/2$, evaluation at $\overline{1}$ identifies this functor with $M \mapsto \ker(2\colon M \to M)$. Applying it to the sequence for $\mathbb{Z}/4$ gives

$$0 \longrightarrow \mathbb{Z}/2 \xrightarrow{\cong} \{0, \overline{2}\} \xrightarrow{0} \mathbb{Z}/2.$$

The resulting sequence is exact at its middle term and the last arrow fails to be surjective. In fact the snake connecting map in the example identifies its cokernel with Q_a. This is how the connecting map measures the failure to lift a map from P to the quotient into the middle module.

For a commutative ring map $R \to S$, extension of scalars $M \mapsto S \otimes_R M$ is additive and preserves cokernels, because it is a left adjoint by Chapter 15. For $R = \mathbb{Z}$, $S = \mathbb{Z}/2$, it sends the injection in the same short exact sequence to the zero map between two copies of $\mathbb{Z}/2$, since the image of the generator in $\mathbb{Z}/4$ is twice a generator. Thus this functor is right exact and fails to preserve that mono. The next chapters use complexes to record failures of exactness in successive degrees.

CHAPTER 51

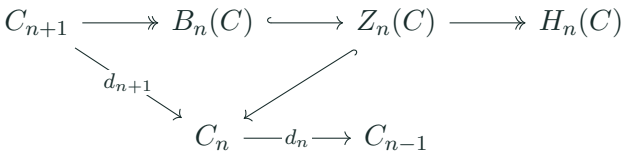
Homology from chain complexes

Three oriented edges around a triangle give a closed chain. If the triangle is filled, that chain is the boundary of the face. The same boundary equations on the vertices and edges therefore give different homology according to the available faces. The construction makes this distinction by taking the quotient of a kernel by an image. A short exact sequence compares such quotients through a map that changes the degree.

51.1 Cycles, boundaries, and maps

Let $\mathcal{A}$ be an abelian category as in Chapter 50. A *chain complex* C consists of objects C_n for integers n and differentials $d^C_n \colon C_n \to C_{n-1}$ with $d^C_{n-1} d^C_n = 0$. For a complex concentrated in a range of degrees we use zero objects in all the other degrees. We omit the superscript on a differential when the complex is determined by its domain.

The *cycles* $Z_n(C) \hookrightarrow C_n$ are the kernel of d_n. The *boundaries* $B_n(C) \hookrightarrow C_n$ are the image of d_{n+1}. The composite with d_n kills that image because the map from C_{n+1} onto it is epic. Consequently the boundary subobject factors through the cycles. Define $H_n(C)$ by the cokernel of this inclusion.



Thus maps from $H_n(C)$ are precisely maps from $Z_n(C)$ that kill every boundary. For modules this is the quotient $\ker(d_n)/\operatorname{im}(d_{n+1})$. The equality $H_n(C) = 0$ holds precisely when the complex is exact at C_n, since in an abelian category a mono with zero cokernel is an isomorphism.

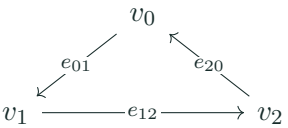
A *chain map* $f \colon C \to D$ is a family of maps f_n commuting with the differentials. It restricts to the cycles by the kernel property and takes boundaries to boundaries by the image factorization. Indeed the map from C_{n+1} to the boundaries is epic, and $f_n d^C_{n+1} = d^D_{n+1} f_{n+1}$ factors through $B_n(D)$. Composing with the cokernel of

the latter inclusion shows that the restriction to $B_n(C)$ factors there as well.

$$\begin{array}{ccccc} B_n(C) & \hookrightarrow & Z_n(C) & \twoheadrightarrow & H_n(C) \\ \downarrow & & \downarrow & & \downarrow{\scriptstyle H_n(f)} \\ B_n(D) & \hookrightarrow & Z_n(D) & \twoheadrightarrow & H_n(D) \end{array}$$

This gives the unique map on the quotients. The identity and composition laws for it follow by precomposing with the epimorphism from the cycles. We have obtained the functor H_n. If two maps are added degree by degree, the induced map on homology is the sum of their induced maps, by the same epi comparison. In particular homology is additive. The zero object and the direct sum of two complexes are computed degree by degree, and H_n preserves these biproducts.

Example 51.1 (The boundary of a triangle)**.** Take free abelian groups on the vertices v_0, v_1, v_2 in degree zero and the oriented edges e_{01}, e_{12}, e_{20} in degree one. Let $d_1(e_{ij}) = v_j - v_i$, and put zero groups in higher degrees. The diagram records oriented edges, with the subscripts specifying the signs in their boundaries.



The differential of $ae_{01} + be_{12} + ce_{20}$ has coefficients $(c - a, a - b, b - c)$ on the three vertices. It vanishes exactly when $a = b = c$. Thus $z = e_{01} + e_{12} + e_{20}$ freely generates H_1. In degree zero the image of d_1 identifies the three vertex classes. Their common class freely generates H_0, since summing the vertex coefficients kills d_1 and sends that class to 1.

Now add a generator a_2 in degree two, with $d_2(a_2) = z$. The new complex is that of the filled triangle. The differential condition follows from the cycle calculation, and d_2 is injective with image $\mathbb{Z}z$. The homology in positive degrees vanishes, and H_0 is still $\mathbb{Z}$.

51.2 What a chain homotopy preserves

For chain maps $f, g\colon C \to D$, a *chain homotopy* from g to f is a family $s_n\colon C_n \to D_{n+1}$ such that

$$f_n - g_n = d^D_{n+1} s_n + s_{n-1} d^C_n.$$

Restriction to the cycles kills the second term, and the first factors through $B_n(D)$. Quotienting therefore gives $H_n(f) = H_n(g)$. The relation is an equivalence by zero homotopies, negation, and addition. Precomposing or postcomposing with a chain map preserves the homotopy equations by commutation with the differentials.

Thus composition passes to the equivalence classes, giving the *homotopy category* of complexes, through which each H_n factors.

A chain homotopy equivalence consists of two chain maps whose composites are homotopic to the respective identities. It induces inverse isomorphisms on homology. In particular the complex with M in degrees one and zero and differential id_M is homotopy equivalent to zero. In the homotopy equation for its identity map and zero, take $s_0 = \mathrm{id}_M$ and all other components zero. A complex whose identity is homotopic to zero is called *contractible*.

Example 51.2 (Contracting the filled triangle)**.** Let D be the filled triangle from Example 51.1. The augmentation $\varepsilon\colon D \to \mathbb{Z}[0]$, where $\mathbb{Z}[0]$ is concentrated in degree zero, sums the vertex coefficients. Let $j\colon \mathbb{Z}[0] \to D$ take 1 to v_0. Then $\varepsilon j = \mathrm{id}$. A homotopy for $\mathrm{id}_D - j\varepsilon$ is given by the following values, extended linearly.

x	v_0	v_1	v_2
$s_0(x)$	0	e_{01}	$e_{01} + e_{12}$

x	e_{01}	e_{12}	e_{20}
$s_1(x)$	0	0	a_2

On vertices, $d_1 s_0(v_i) = v_i - v_0$. On e_{01} and e_{12}, the term $s_0 d_1$ returns the edge itself. On e_{20} it returns $-e_{01} - e_{12}$, while $d_2 s_1$ contributes z, yielding e_{20}. In degree two, $s_1 d_2(a_2) = s_1(z) = a_2$. These equations exhibit a homotopy equivalence to $\mathbb{Z}[0]$.

51.3 The connecting map

A sequence of complexes $0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$ is *short exact* when it is short exact at each degree. A cycle in C_n lifts to B_n in the category of modules. Its lift may have a nonzero differential, which lies in A_{n-1}. That differential determines a homology class one degree lower. We first give the categorical meaning of this lift.

For an epimorphism $e\colon X \to Y$ and a map $t\colon T \to Y$, let T' be their pullback. The map $u\colon T' \to T$ is epic. To prove this, identify the pullback with the kernel of $[e, -t]\colon X \oplus T \to Y$. The latter is epic because its composite with the injection of X is e. Suppose a map $h\colon T \to W$ kills u. Then $[0, h]$ kills the kernel of $[e, -t]$, so it factors through that epimorphism, which is the cokernel of its kernel. This factor kills e, hence is zero. Restricting to the summand T gives $h = 0$. Applying this to differences of maps proves that u is epic.

Thus when t need not lift over T, it lifts over the replacement $T' \to T$. We use such replacements freely in a chase. The rules for descending its conclusions are simple. An equation tested after an epimorphism holds before it, by right cancellation. If a map factors through a subobject after precomposition by an epimorphism, its composite with the subobject's cokernel is zero by cancellation. It then factors through the subobject itself, which is the kernel of that cokernel.

These rules allow us to reason with maps from a test object using representative notation such as $db = ia$.

For the short exact sequence above, form the following pullback, where the bottom inclusion is the kernel of d_n^C.

$$\begin{array}{ccc} P & \xrightarrow{\;b\;} & B_n \\ \downarrow{\scriptstyle u} & \lrcorner & \downarrow{\scriptstyle p_n} \\ Z_n(C) & \hookrightarrow & C_n \end{array}$$

The map $d_n^B b$ is killed by p_{n-1}, so it uniquely factors through i_{n-1} as a map to A_{n-1}. The factor lies in $Z_{n-1}(A)$, since its differential followed by the monomorphism i_{n-2} is zero. Its composite to $H_{n-1}(A)$ kills $\ker u$. Indeed $\ker u$ is identified with A_n by exactness of the degree n row, and there the factor is d_n^A, a boundary. Since u is the cokernel of its kernel, this defines a map $Z_n(C) \to H_{n-1}(A)$.

To see that this map kills boundaries, test it on the map $C_{n+1} \to Z_n(C)$ induced by d_{n+1}^C. After precomposing with the epimorphism $p_{n+1}\colon B_{n+1} \to C_{n+1}$, take d_{n+1}^B as the lift to B_n. Its differential is zero. Cancellation gives the required vanishing, since $C_{n+1} \to B_n(C)$ is epic. Therefore there is a unique map

$$\partial_n \colon H_n(C) \longrightarrow H_{n-1}(A).$$

In representative notation choose a cycle c and a lift b with $pb = c$, write $db = ia$, and set $\partial_n[c] = [a]$. The construction above proves the choice independence and defines the map in any abelian category. A map of short exact sequences sends a lift and its differential to a lift and its differential in the second sequence. Uniqueness after the epimorphisms used in the construction gives naturality of ∂_n.

51.4 Exactness across the degrees

Theorem 51.3 (Long exact sequence of homology)**.** *A short exact sequence $0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$ of complexes induces an exact sequence with successive terms*

$$H_n(A) \xrightarrow{H_n(i)} H_n(B) \xrightarrow{H_n(p)} H_n(C) \xrightarrow{\partial_n} H_{n-1}(A) \xrightarrow{H_{n-1}(i)} H_{n-1}(B) \xrightarrow{H_{n-1}(p)} H_{n-1}(C).$$

The connecting map is the one defined above. In each degree the maps commute with the maps induced by morphisms of short exact sequences.

Proof. We use the representative notation of Section 51.3. Each representative of a homology class is lifted along the quotient from the cycles, and each representative of a boundary is lifted along the epi onto the image of the differential. In modules these are ordinary lifts. In an abelian category each lift may require replacing the current test object by one mapping epically onto it. To check a kernel lies in an image it suffices to start with its inclusion as a test map. The subobject descent rule then gives the factorization globally once it has been found after the lifts.

At $H_n(B)$, a cycle b maps to zero in $H_n(C)$ when $pb = dc'$. Lift c' to b'. Then $b - db'$ is a cycle representing the same class as b and is killed by p. It comes from a cycle of A, since i is monic and commutes with the differential. Conversely a cycle from A maps to zero because $pi = 0$.

At $H_n(C)$, lift c to b with $db = ia$. If $[a] = 0$, write $a = da'$. The element $b - ia'$ is then a cycle lifting c. Conversely if a class comes from a cycle of B, that cycle can be used as the lift so the connecting class is zero.

At $H_{n-1}(A)$, a cycle a killed in $H_{n-1}(B)$ satisfies $ia = db$. Now pb is a cycle and $\partial_n[pb] = [a]$. Conversely any class from ∂_n has representative a with $ia = db$, so it is killed in $H_{n-1}(B)$. Varying n proves exactness everywhere. Naturality follows from the construction of the maps. □

Example 51.4 (The incidence of the filled face)**.** Let $C \hookrightarrow D$ be the inclusion from the boundary triangle into the filled triangle of Example 51.1. Their quotient has $\mathbb{Z}$ in degree two, generated by the class $\overline{a}_2$ of the face, and zero objects in all the other degrees. In the connecting map for this short exact sequence, lift $\overline{a}_2$ to a_2. Its differential in D is the cycle z in C. Thus $\partial_2 \colon H_2(D/C) \to H_1(C)$ sends the chosen generator to the chosen generator. Exactness also forces this map to be an isomorphism, since the adjacent terms $H_2(D)$ and $H_1(D)$ both vanish. The connecting map records the attaching boundary of the added face.

51.5 Counting graph cycles with an incidence complex

Orient each edge of a finite graph H and form the chain complex $\mathbb{Z}^{E(H)} \xrightarrow{\partial} \mathbb{Z}^{V(H)}$, where $\partial(e) = [v] - [u]$ for an edge e from u to v. The cokernel is freely generated by the connected components, since a path identifies its endpoint generators and the boundary respects the component of each vertex. The kernel has a basis built from a spanning forest T, chosen by taking a spanning tree in each component. Each edge e outside T gives a cycle γ_e formed by adding to e the path in T from its target back to its source, with each path edge given the sign of its traversal relative to the chosen orientation.

The coefficient of a nontree edge e in γ_f is 1 if $e = f$ and 0 otherwise. From any cycle subtract the combination of the γ_f with its nontree edge coefficients. The difference is a cycle supported on T. At a leaf its only incident coefficient must be zero, and repeating after deleting the leaf removes every edge. This proves the basis assertion. If H has v vertices, e edges, and c components, the forest has $v - c$ edges. Thus

$$\operatorname{rank} H_0(H) = c, \qquad \operatorname{rank} H_1(H) = e - v + c.$$

These ranks give the Euler identity $\operatorname{rank} H_0(H) - \operatorname{rank} H_1(H) = v - e$.

Example 51.5 (Two cycles and two face relations)**.** Take the graph of two triangles sharing an edge from Section 46.6. The three edges ab, ac, ad form a spanning tree, so the oriented boundaries of abc and abd form a basis of H_1, after orienting each triangle to agree with its nontree edge. Filling the triangle abc adds one two-cell whose boundary is the first basis vector. The first homology then has rank one. Filling both triangles makes the boundary map in degree two an isomorphism onto the cycle group. At that stage both H_1 and H_2 vanish.

51.6 Changing coefficients

Tensoring a complex of abelian groups with a group S gives a new complex with differentials $d_n \otimes \mathrm{id}_S$, by functoriality and preservation of zero maps. An additive exact functor preserves the kernels, images, and cokernels used to define homology. Tensoring can change these constructions. The two-term complex

$$C_1 = \mathbb{Z} \xrightarrow{m} \mathbb{Z} = C_0 \qquad (m \geq 2)$$

has $H_1(C) = 0$ and $H_0(C) = \mathbb{Z}/m$. With $S = \mathbb{Z}/m$, its two terms tensor to S and the resulting differential is zero. Thus $H_1(C \otimes S) = \mathbb{Z}/m$ and $H_0(C \otimes S) = \mathbb{Z}/m$. One can locate this new class in the long exact sequence. Multiplication by m is injective on each original term, giving

$$0 \longrightarrow C \xrightarrow{m} C \longrightarrow C \otimes (\mathbb{Z}/m) \longrightarrow 0.$$

The connecting map sends the degree-one class of 1 modulo m to the class of 1 in $H_0(C)$. Indeed, the lift 1 has differential m in the middle copy, and its preimage under the first arrow is 1 in degree zero.

The same two-term complex gives a map $C \to (\mathbb{Z}/m)[0]$ by reduction modulo m in degree zero. It induces an isomorphism in every homology degree. Such a map is called a *quasi-isomorphism*. This one has no chain homotopy inverse, since any chain map in the other direction has a zero degree-zero component. Indeed any homomorphism $\mathbb{Z}/m \to \mathbb{Z}$ is zero. A homotopy inverse would have to induce an inverse on H_0. Chapter 52 studies what happens when all quasi-isomorphisms are made invertible.

51.7 Cochain indexing

A *cochain complex* K has arrows $d^n\colon K^n \to K^{n+1}$ with $d^{n+1}d^n = 0$. It can be read as a chain complex by setting $C_n = K^{-n}$. Then $H^n(K) = \ker(d^n)/\operatorname{im}(d^{n-1})$ is $H_{-n}(C)$, and the connecting map for a short exact sequence $0 \to K \to L \to M \to 0$ has direction $H^n(M) \to H^{n+1}(K)$. This convention will be used for resolutions and sheaf cohomology.

For example, applying $\operatorname{Hom}_{\mathbb{Z}}(-,\mathbb{Z})$ to the two-term chain example gives a cochain complex with two copies of $\mathbb{Z}$ in degrees zero and one and differential multiplication by m. Its degree-one cohomology is $\mathbb{Z}/m$. The calculation uses the maps of the original complex, and it can retain information beyond the duals of its homology groups. Weibel develops both indexing conventions and the exact sequence construction in his treatment of chain complexes [41, Chapter 1].

CHAPTER 52

Derived categories from localization

A module can be specified by generators and relations, and the relations can themselves have relations. A complex records the successive stages. Many complexes can resolve the same module. The derived category makes the resolution and the original module isomorphic by formally inverting quasi-isomorphisms. We construct this category for bounded-above complexes of modules and use it to calculate what tensoring does to a resolution.

52.1 Complexes, homotopies, and quasi-isomorphisms

Fix a commutative unital ring R. We use cochain notation $X = (X^i, d_X^i)$, with $d_X^i \colon X^i \to X^{i+1}$ and $d_X^{i+1} d_X^i = 0$. The cohomology $H^i(X)$ is the quotient of the cycles in degree i by the boundaries, as in Chapter 51. A chain complex indexed by lower indices can be put in this convention by $X^i = X_{-i}$. A module by itself is regarded as a complex in degree zero. The category $\mathbf{Ch}^-(R)$ has complexes that are zero in all sufficiently high degrees and cochain maps as arrows. The bound may depend on the complex.

Two maps $f, g \colon X \to Y$ are homotopic if there are maps $h^i \colon X^i \to Y^{i-1}$ with $f - g = d_Y h + h d_X$. Composing on either side with cochain maps preserves the homotopy relation. Passing to its classes gives the *homotopy category* $\mathbf{K}^-(R)$, with the same objects and arrows denoted $[f]$. An isomorphism in this category is a homotopy equivalence. Cohomology is defined on homotopy classes because homotopic maps take each cycle to the same class. A *quasi-isomorphism* is a map or homotopy class inducing isomorphisms on all cohomology modules. Two quasi-isomorphisms compose to a quasi-isomorphism. If two of f, g, gf are quasi-isomorphisms, the third is as well, by cancellation of isomorphisms on each cohomology module. A complex with zero cohomology is called *acyclic*.

Example 52.1 (An acyclic complex without a contraction)**.** Let $m \geq 2$, and use the following complex in degrees $-2, -1, 0$.

$$0 \longrightarrow \mathbb{Z} \xrightarrow{m} \mathbb{Z} \xrightarrow{q} \mathbb{Z}/m \longrightarrow 0$$

Here q is reduction modulo m. The sequence is exact, so the complex has zero cohomology. A complex is *contractible* if its identity is homotopic to zero. A contracting homotopy here would give $qh^0 = \mathrm{id}_{\mathbb{Z}/m}$ in degree 0. Any map

from $\mathbb{Z}/m$ to $\mathbb{Z}$ is zero, since its image is killed by m. The required section therefore cannot exist. The map from this complex to the zero complex is a quasi-isomorphism. It cannot be inverted in the homotopy category.

52.2 Shifts and cones

For an integer r, set

$$X[r]^i = X^{i+r}, \qquad d^i_{X[r]} = (-1)^r d^{i+r}_X.$$

For a cochain map $f\colon X \to Y$, its *mapping cone* $C(f)$ has $C(f)^i = Y^i \oplus X^{i+1}$ and differential

$$d^i_{C(f)}(y, x) = (d^i_Y y + f^{i+1}x,\ -d^{i+1}_X x).$$

The sign gives $d^2_{C(f)} = 0$, since the two mixed terms in the first component cancel by the cochain map equation. Inclusion into the first summand and projection onto the second give a degreewise split short exact sequence of complexes.

$$0 \longrightarrow Y \xhookrightarrow{i} C(f) \xrightarrow{p} X[1] \longrightarrow 0$$

The connecting map from $H^i(X[1])$ to $H^{i+1}(Y)$ is $H^{i+1}(f)$. In fact a cycle x in the shifted complex lifts to $(0, x)$, with differential $(f^{i+1}x, 0)$. The long exact sequence of Chapter 51 now gives the exact portion

$$H^i(X) \xrightarrow{H^i(f)} H^i(Y) \xrightarrow{H^i(i)} H^i(C(f)) \xrightarrow{H^i(p)} H^{i+1}(X) \xrightarrow{-H^{i+1}(f)} H^{i+1}(Y).$$

It follows that $C(f)$ has zero cohomology exactly when f is a quasi-isomorphism. This argument uses no bound on the nonzero degrees. The cone of id_X is contractible. On $X^i \oplus X^{i+1}$, the degree -1 map $h(y, x) = (0, y)$ satisfies $dh + hd = \mathrm{id}$.

52.3 Projective complexes compute maps

A module P is projective if every map $P \to B$ lifts along every surjective module map $A \to B$. Free modules are projective by lifting the images of their basis elements. We need two consequences of this lifting property. Every bounded-above complex admits a quasi-isomorphism from a bounded-above complex of free modules. Maps from such projective complexes, taken up to homotopy, are insensitive to quasi-isomorphisms of the target.

Lemma 52.2 (A free replacement)**.** *For every X in $\mathbf{Ch}^-(R)$ there is a quasi-isomorphism $u\colon P \to X$ in $\mathbf{Ch}^-(R)$ with each P^i free.*

Proof. Choose b with $X^i = 0$ for $i > b$ and put $P^i = 0$ in those degrees. Construct the remaining terms downward. Given the terms above i, let

$$T^i = \{(x, p) \in X^i \oplus P^{i+1} \mid d_X^i x = u^{i+1} p, \quad d_P^{i+1} p = 0\}.$$

Take P^i to be free on the set T^i, and define u^i and d_P^i by the first and second projections on its generators. This set is the pullback of d_X^i along the restriction of u^{i+1} to $\ker d_P^{i+1}$. The two conditions on T^i give the map equation and $d_P^{i+1} d_P^i = 0$. A cycle $x \in X^i$ lifts to the generator indexed by $(x, 0)$, which is a cycle of P^i. If a cycle $p \in P^i$ maps to a boundary $u^i p = d_X^{i-1} x$, then (x, p) indexes a generator of P^{i-1} whose differential is p. These observations give surjectivity and injectivity on cohomology. □

Lemma 52.3 (Invariance with projective source)**.** *Let P be bounded above with projective terms, and let $s\colon Z \to Y$ be a quasi-isomorphism of bounded-above complexes. Postcomposition with s gives a bijection*

$$\operatorname{Hom}_{\mathbf{K}^-(R)}(P, Z) \xrightarrow{\sim} \operatorname{Hom}_{\mathbf{K}^-(R)}(P, Y).$$

Proof. First suppose A is acyclic and $f\colon P \to A$. We construct a homotopy from f to zero in descending degree, starting with zero above the bound of P. With h^{i+1} already chosen, the map $f^i - h^{i+1} d_P^i$ has image in $\ker d_A^i$. Indeed $d_A h^{i+1} = f^{i+1} - h^{i+2} d_P^{i+1}$ and $d_P^{i+1} d_P^i = 0$. Acyclicity gives a surjection $A^{i-1} \to \ker d_A^i$ via the differential. Projectivity gives h^i lifting the indicated map. This completes the homotopy equation in degree i.

For any complex B, consider the complex

$$\mathcal{H}^j(P, B) = \prod_i \operatorname{Hom}_R(P^i, B^{i+j}), \qquad D(v) = d_B v - (-1)^j v d_P.$$

Its square-zero equation follows by canceling the two mixed terms. Its degree-zero cohomology is the set of homotopy classes $P \to B$, and in degree j it gives homotopy classes $P \to B[j]$, by the sign in the shift. The cone formula gives $\mathcal{H}(P, C(s)) \cong C(\mathcal{H}(P, s))$. The two component families are precisely the families into Y and into Z shifted by one, and substituting the formula for D gives the cone differential on them. By Section 52.2, $C(s)$ is acyclic, as are all its shifts. The homotopy construction just proved then makes $\mathcal{H}(P, C(s))$ acyclic. Applying the cone criterion again proves the desired isomorphism on degree-zero cohomology. □

In particular, a quasi-isomorphism $s\colon P \to Q$ between bounded-above projective complexes is a homotopy equivalence. Taking Q as source in the lemma lifts id_Q to $[r]\colon Q \to P$ with $[s][r] = \mathrm{id}_Q$. Taking P as source and using injectivity gives $[r][s] = \mathrm{id}_P$.

52.4 Localization and roofs

The *localization* of a category at a class of arrows comes with a functor that inverts those arrows. Every functor that inverts them factors through this functor, and the factorization is unique up to natural isomorphism. This is the same test on maps out that characterized localizing rings in Chapter 11. For $\mathbf{K}^-(R)$ we localize at the quasi-isomorphisms. Write $\mathbf{D}^-(R)$ for the resulting derived category.

Here is a construction with hom sets. Choose a replacement $p_X \colon P_X \to X$ as in Lemma 52.2 for each complex. One can use the greatest index in which X is nonzero, or the bound zero for the zero complex, and take the free module on the specified set at each stage. Keep the objects of $\mathbf{K}^-(R)$. In the new category the arrows $X \to Y$ are $\operatorname{Hom}_{\mathbf{K}^-(R)}(P_X, P_Y)$. Composition and identities are those of projective complexes. The groups of maps and the sets of homotopy classes are sets, so this construction gives a locally small category.

For a map $f \colon X \to Y$ in $\mathbf{K}^-(R)$ the comparison lemma gives a unique homotopy class $\widetilde{f} \colon P_X \to P_Y$ filling the square.

$$\begin{array}{ccc} P_X & \overset{\widetilde{f}}{\dashrightarrow} & P_Y \\ \downarrow{\scriptstyle p_X} & & \downarrow{\scriptstyle p_Y} \\ X & \xrightarrow{f} & Y \end{array}$$

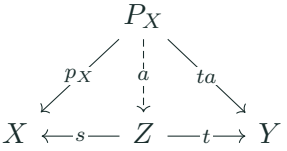
The square commutes in the homotopy category. Uniqueness shows that $f \mapsto \widetilde{f}$ respects identities and composition, giving a functor L. If f is a quasi-isomorphism then so is its lift, by applying cohomology to the square. The lift is invertible in the homotopy category of projective complexes, as just proved. Thus $L(f)$ is invertible.

Every arrow $X \to Y$ in this construction has a *roof*. If a class $h \colon P_X \to P_Y$ represents the arrow, its roof is $X \xleftarrow{p_X} P_X \xrightarrow{p_Y h} Y$. This expression means $L(p_Y h)L(p_X)^{-1}$. To verify it, the defining square for p_X gives $\widetilde{p}_X = p_{P_X} \colon P_{P_X} \to P_X$ by uniqueness. The defining square for $p_Y h$ similarly gives $h p_{P_X}$. Cancelling proves the formula. If F is any functor from $\mathbf{K}^-(R)$ sending quasi-isomorphisms to isomorphisms, the rule

$$h \longmapsto F(p_Y)F(h)F(p_X)^{-1}$$

defines the factorization on arrows, with object X sent to $F(X)$. It respects composition by cancellation and agrees with F on $L(f)$ by the defining square. Any such factorization has these values because it must respect the roof formulas. Natural transformations are also determined through L, since naturality for the inverted arrows follows by composing with the inverses. This proves the localization property.

More generally take a roof $X \xleftarrow{s} Z \xrightarrow{t} Y$ in $\mathbf{K}^-(R)$ with s a quasi-isomorphism. There is a unique class $a\colon P_X \to Z$ with $sa = p_X$, and a is a quasi-isomorphism. Thus the roof is represented by the map $ta\colon P_X \to Y$, or equivalently by its lift to P_Y.



Two roofs give the same map exactly when they agree after a common refinement by quasi-isomorphisms in the homotopy category. Indeed all roofs can be refined by P_X, and their refined numerators agree precisely when their lifts to P_Y agree. Conversely, any common refinement yields the same expression $L(t)L(s)^{-1}$ after cancellation. Bounded-above projective complexes therefore provide an equivalent model of $\mathbf{D}^-(R)$. Their own arrows are homotopy classes, since their replacements are already homotopy equivalent to them. General derived categories also use localization at quasi-isomorphisms, with the available replacements depending on the category and the degrees allowed [43, Sections 13.11 and 13.19].

The composite $\mathbf{Ch}^-(R) \to \mathbf{K}^-(R) \to \mathbf{D}^-(R)$ also localizes at the quasi-isomorphisms. To see why these inverses already identify homotopic maps, use the cone of $(-\operatorname{id}_X, \operatorname{id}_X)\colon X \to X \oplus X$. Its elements in degree i are (x_0, x_1, t) with $t \in X^{i+1}$. The map q to X that adds the first two coordinates has kernel $C(\operatorname{id}_X)$, identified by $(x, t) \mapsto (-x, x, t)$. Thus q is a quasi-isomorphism. The two inclusions $i_0, i_1\colon X \to C((-\operatorname{id}_X, \operatorname{id}_X))$ into the first and second copies both satisfy $qi_j = \operatorname{id}_X$. If $f - g = d_Y h + h d_X$, the cochain map

$$(x_0, x_1, t) \longmapsto g(x_0) + f(x_1) + h(t)$$

restricts along i_0, i_1 to g, f. The map equation follows from the homotopy identity, since the cone differential adds $(-t, t)$ to the first two coordinates. Every functor inverting q identifies i_0, i_1 by cancellation, hence also f, g. Such a functor factors through $\mathbf{K}^-(R)$ and then through its localization.

52.5 Exact triangles

The cone determines the triangle

$$X \xrightarrow{f} Y \xrightarrow{i} C(f) \xrightarrow{p} X[1]$$

in the derived category. The *distinguished triangles* are the triangles isomorphic to ones of this form. They give the triangulated structure on the derived category, with the shift above as translation [43, Sections 13.9 and 13.12]. The cone exact

sequence shows that applying cohomology to such a triangle and its shifts gives a long exact sequence.

For a short exact sequence $0 \to X \xrightarrow{f} Y \xrightarrow{q} Z \to 0$ of complexes, the cone has a map to Z given by $(y, x) \mapsto qy$. It is a cochain map, since $qf = 0$, and is surjective in every degree. Its kernel identifies with $C(\mathrm{id}_X)$ via $(a, x) \mapsto (fa, x)$. This identification preserves the differentials by $d_Y f = f d_X$. The kernel is contractible, so the map $C(f) \to Z$ is a quasi-isomorphism by the long exact sequence. In the localized category it is invertible and gives the dashed connecting arrow below.

$$\begin{array}{ccccccc} X & \xrightarrow{f} & Y & \xrightarrow{i} & C(f) & \xrightarrow{p} & X[1] \\ \| & & \| & & \downarrow \wr & & \| \\ X & \xrightarrow{f} & Y & \xrightarrow{q} & Z & \xrightarrow{\delta} & X[1] \end{array}$$

This recovers the long exact sequence of a short exact sequence from one triangle, with the sign determined by the cone convention. For a cycle z in Z^i, lift z to y in Y^i and write $d_Y y = fa$. Then $(y, -a)$ is a cycle in $C(f)$ over z, so $H^i(\delta)[z] = -[a]$. Negating this map gives the lift-then-differentiate convention of Chapter 51. The triangle records an arrow before taking cohomology.

52.6 What derived tensor remembers

Fix an R-module N. Tensoring every term with N preserves homotopies, because it is additive and preserves composition. It therefore takes homotopy equivalences to homotopy equivalences. Replacing X by a bounded-above projective complex gives the left derived tensor

$$X \otimes^{\mathbf{L}}_R N = P_X \otimes_R N \qquad \text{in } \mathbf{D}^-(R).$$

A map is calculated by lifting to the replacements, then tensoring. Composition of lifts is unique up to homotopy, so this gives a functor. A different projective replacement has a comparison homotopy equivalence by Lemma 52.3. Tensoring preserves that equivalence, giving the same result up to isomorphism. For modules M, N in degree zero, $H^{-i}(M \otimes^{\mathbf{L}}_R N)$ is denoted $\operatorname{Tor}^R_i(M, N)$.

Example 52.4 (Tensoring a relation)**.** Over $\mathbb{Z}$, let $M = \mathbb{Z}/m$ and $N = \mathbb{Z}/n$ for integers $m, n \geq 2$. Use the two-term complex $P^{-1} = \mathbb{Z} \xrightarrow{m} P^0 = \mathbb{Z}$. Reduction modulo m gives $P \to M$ and an isomorphism on every cohomology module. This is a projective replacement. Its termwise tensor with N is

$$\mathbb{Z}/n \xrightarrow{\cdot m} \mathbb{Z}/n \qquad \text{in degrees } -1,\ 0.$$

Write $g = \gcd(m, n)$. The image of multiplication by m consists of the multiples of g modulo n, by the Bézout identity, so its cokernel is $\mathbb{Z}/g$. The kernel consists

of the multiples of n/g modulo n because m/g and n/g are relatively prime. The kernel is cyclic of order g. Consequently

$$\operatorname{Tor}_0^{\mathbb{Z}}(M,N) \cong \mathbb{Z}/g, \qquad \operatorname{Tor}_1^{\mathbb{Z}}(M,N) \cong \mathbb{Z}/g,$$

and all higher groups vanish for this pair. For $m = 6$, $n = 4$, the tensor differential is multiplication by 2 modulo 4. Both the cokernel and kernel have order 2. The ordinary tensor $M \otimes_{\mathbb{Z}} N$ is the degree-zero cokernel by the right exactness shown in Chapter 15. In degree -1, the derived tensor also records the elements annihilated by the relation. In particular $P \otimes_{\mathbb{Z}} N \to M \otimes_{\mathbb{Z}} N$ fails to be a quasi-isomorphism when $g > 1$.

Example 52.5 (Changing scalars in both degrees)**.** Return to the presentation of $Q = \operatorname{coker} A$ in Chapters 13 and 15, where

$$A = \begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix}.$$

The determinant is -12, so the map is injective over $\mathbb{Z}$. Thus $A\colon \mathbb{Z}^2 \to \mathbb{Z}^2$ in degrees $-1, 0$ is a projective replacement of Q. The integral changes of bases from the presentation give the diagonal entries $2, 6$. After tensoring with $\mathbb{F}_2$ both entries vanish, so the cohomology in each of the two degrees is $\mathbb{F}_2^2$. After tensoring with $\mathbb{F}_3$ one entry is invertible and one vanishes, so both cohomology modules are $\mathbb{F}_3$. In degree 0 these are the cokernels already computed by extension of scalars. In degree -1 they account for its new kernels.

52.7 Pontryagin duality on complexes with finite cohomology

For a complex of abelian groups, take the character coefficient $J = \mathbb{Q}/\mathbb{Z}$. It is injective because it is divisible: by Baer's criterion it suffices to extend a map from $n\mathbb{Z} \subseteq \mathbb{Z}$, which amounts to finding an n-th root of the image of n in J. For a bounded cochain complex C, put $DC = \mathcal{H}(C, J)$, with J in degree zero and the internal Hom differential. In degree i this means

$$(DC)^i = \operatorname{Hom}_{\mathbb{Z}}(C^{-i}, J), \qquad d_{DC}^i(\phi) = (-1)^{i+1}\phi \circ d_C^{-i-1}.$$

A cocycle in $(DC)^i$ kills boundaries in C^{-i}. Restricting it to cycles gives a character on $H^{-i}(C)$. Every such character extends to C^{-i} by injectivity of J, and a character vanishing on the cycles factors through $\operatorname{im} d_C^{-i}$ and extends to C^{-i+1}. Thus the latter characters are precisely the coboundaries, and

$$H^i(DC) \cong \operatorname{Hom}_{\mathbb{Z}}(H^{-i}(C), J).$$

This natural isomorphism shows directly that D takes quasi-isomorphisms to quasi-isomorphisms. The double evaluation $C^i \to (DDC)^i$, sending x to $\phi \mapsto (-1)^i \phi(x)$, is a cochain map with these signs. For complexes with finite cohomology it is a quasi-isomorphism by finite character duality in Section 8.6: all circle characters of a finite group land in J. This is a contravariant involution on the bounded derived subcategory with finite cohomology.

Example 52.6 (Dualizing a cyclic resolution)**.** The complex $\mathbb{Z} \xrightarrow{n} \mathbb{Z}$ in degrees $-1, 0$, for $n > 0$, represents $\mathbb{Z}/n$. Its dual is $J \xrightarrow{-n} J$ in degrees $0, 1$. The map $-n$ is surjective and its kernel is $\frac{1}{n}\mathbb{Z}/\mathbb{Z}$. This kernel is the character group, now obtained from a dual complex even though the original resolution used infinite groups.

For the derived tensor calculation, projective terms make comparisons computable by homotopy. The next chapter uses injective resolutions to derive global sections of sheaves.

CHAPTER 53

Sheaf cohomology from global sections

A section of a quotient sheaf can lift in a neighborhood of each point, while the local lifts can differ on the overlaps. Those differences have values in the kernel of the quotient. Sheaf cohomology puts this obstruction into a group. Its construction starts with the global sections functor and applies that functor to resolutions in which the terms admit extensions.

53.1 Global sections and local lifts

Fix a topological space X, and write $\mathrm{Sh}(X, \mathbf{Ab})$ for its sheaves of abelian groups. The group operations commute with restriction. Kernels are computed on sections, since a compatible family of sections in the kernel glues in the source and its image is locally zero. Cokernels are formed by sheafifying the presheaf of quotient groups. The construction by germs in Chapter 44 also works for abelian groups, using the group operations on germs. In particular sheafification preserves stalks.

Taking the stalk at x is exact. A relation among germs is already a relation on a common smaller neighborhood, by the filtered colimit description of stalks. This proves that stalks commute with kernels, and the stalk of a sheaf cokernel is the cokernel of the stalk map. The abelian category tests of Chapter 50 now apply. In particular, the map from coimage to image of a sheaf morphism induces an isomorphism on every stalk, and so is an isomorphism by the stalk test of Chapter 44. Exactness can be checked on stalks for the same reason. For an epimorphism $p\colon \mathcal{B} \to \mathcal{C}$, the stalk condition means every section of $\mathcal{C}$ lifts near every point of its domain. Given a section and a germ lifting it, choose a representative of the germ and shrink until its image equals the section.

The global sections functor is

$$\Gamma(X, -)\colon \mathrm{Sh}(X, \mathbf{Ab}) \longrightarrow \mathbf{Ab}, \qquad \mathcal{F} \longmapsto \mathcal{F}(X).$$

Applied to a short exact sequence, it yields the exact sequence

$$0 \longrightarrow \mathcal{A}(X) \longrightarrow \mathcal{B}(X) \longrightarrow \mathcal{C}(X).$$

Indeed it preserves kernels by their description on sections. This is left exactness. To have surjectivity at the last term would require every locally liftable section

to have lifts which agree. If b_i, b_j are lifts of the same section on two opens, their difference $b_j - b_i$ lies in $\mathcal{A}$ on the intersection. Changing b_i by $a_i \in \mathcal{A}(U_i)$ changes that difference by $a_j - a_i$ on the overlap. Thus a global lift amounts to eliminating all the differences by such changes.

53.2 Injectives from stalks

For a point $x \in X$ and a group A, define the *skyscraper sheaf* $J_x(A)$ by

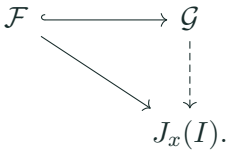
$$J_x(A)(U) = \begin{cases} A & x \in U, \\ 0 & x \notin U. \end{cases}$$

Restrictions between opens containing x are identities. The other restrictions have target zero. A matching family on a cover containing x uses the same value of A in all members that contain x, and those values determine its gluing. The construction applies also when x is not closed.

The stalk and the skyscraper construction are adjoint.

$$\mathrm{Hom}_{\mathrm{Sh}(X,\mathbf{Ab})}(\mathcal{F}, J_x(A)) \cong \mathrm{Hom}_{\mathbf{Ab}}(\mathcal{F}_x, A).$$

The maps $\mathcal{F}(U) \to A$ for U containing x are compatible exactly when they factor through the colimit $\mathcal{F}_x$. On other opens the component map is zero. If I is an injective abelian group, $J_x(I)$ is injective as a sheaf. For a mono $\mathcal{F} \to \mathcal{G}$, the stalk map is a mono, so a map $\mathcal{F}_x \to I$ extends to $\mathcal{G}_x$. Its transpose supplies the dashed arrow below.



Products of these sheaves are injective, since an extension into the product can be supplied in every component. Products of sheaves are computed on sections because the gluing condition holds separately in each component.

For completeness, every abelian group admits an injection into an injective group. The group $\mathbb{Q}/\mathbb{Z}$ is divisible, so the equation $ny = a$ has a solution for each a and each positive n. A homomorphism from a subgroup of an abelian group into a divisible group has a maximal extension by Zorn's lemma. If its domain B omits c, extend to $B + \mathbb{Z}c$. When $nc \in B$ for a least positive n, choose the image of c to be a solution of $ny = f(nc)$. If no such n exists, choose the image to be zero. These choices define an extension, proving injectivity. On a cyclic subgroup generated by any nonzero element there is a map into $\mathbb{Q}/\mathbb{Z}$ nonzero on that generator. Extend it to the whole group by injectivity. Evaluation on all

maps to $\mathbb{Q}/\mathbb{Z}$ therefore embeds the group in a product of injective groups, which is again injective.

Choose injections $\mathcal{F}_x \hookrightarrow I_x$ for all points x of X. By the stalk adjunction these define an arrow

$$\mathcal{F} \longrightarrow \prod_{x \in X} J_x(I_x).$$

At each x the composite with projection to the factor $J_x(I_x)$ induces the chosen injection on the stalk at x. Hence this arrow is a monomorphism. The target is injective, so $\mathrm{Sh}(X, \mathbf{Ab})$ has enough injectives.

53.3 Deriving global sections

Embed $\mathcal{F}$ into an injective $\mathcal{I}^0$, embed its cokernel into an injective $\mathcal{I}^1$, and continue. The resulting exact augmented complex is an *injective resolution*. Its differential first takes a cokernel and then includes it in the next injective. Thus consecutive differentials compose to zero. For each index $q \geq 0$ set

$$H^q(X, \mathcal{F}) = H^q(\Gamma(X, \mathcal{I}^\bullet)).$$

The differential after applying Γ can have a smaller image than the kernel of the next differential. Their quotient is this cohomology group. For $q = 0$, left exactness identifies $\ker \Gamma(d^0)$ with $\mathcal{F}(X)$.

These groups depend functorially on $\mathcal{F}$. A sheaf map between the augmented terms extends to a comparison of injective resolutions. Start by using injectivity of the target in degree zero. At each next step define the comparison on the image of the preceding differential using the already defined map. Exactness and the preceding square ensure that this prescription descends from the preceding term, and injectivity extends it to the entire current term.

$$\begin{array}{ccccccccc}
\mathcal{F} & \hookrightarrow & \mathcal{I}^0 & \xrightarrow{d^0} & \mathcal{I}^1 & \xrightarrow{d^1} & \cdots \\
\downarrow & & \downarrow{\scriptstyle f^0} & & \downarrow{\scriptstyle f^1} & & \\
\mathcal{G} & \hookrightarrow & \mathcal{J}^0 & \longrightarrow & \mathcal{J}^1 & \longrightarrow & \cdots .
\end{array}$$

Two comparisons extending the same map differ by a homotopy. Their degree-zero difference vanishes on $\mathcal{F}$, so it factors through $\operatorname{im} d^0$. Extend that factorization to $h^1 \colon \mathcal{I}^1 \to \mathcal{J}^0$ by injectivity. If the difference is already written as $dh + hd$ below degree q, subtract dh^q from its degree-q component. The remaining map vanishes on $\operatorname{im} d^{q-1}$, hence factors through $\operatorname{im} d^q$. Extend that factorization to the next homotopy component h^{q+1}. This proves the claim inductively. Taking sections preserves the homotopy equations, so the induced cohomology map is independent of the comparison by Chapter 51. Comparisons over the identity give inverse maps on cohomology for any two resolutions.

In the formalism of Chapter 52, $\mathcal{F}$ in degree zero is quasi-isomorphic to $\mathcal{I}^\bullet$. The right derived functor $R\Gamma(X, -)$ on the bounded-below derived category is obtained by taking sections of bounded-below injective representatives. Here we use the general existence theorem for such representatives in an abelian category with enough injectives, as in the construction of right derived functors [43, Section 13.18]. For the sheaf cohomology construction, see the Stacks Project [44, Section 20.2]. The bounded-below condition has a useful consequence. An acyclic bounded-below complex of injectives is contractible. At the lowest degree its first differential is monic, so it splits by injectivity of that term. The image is then injective, and splitting its inclusion in the next term makes the next image a direct summand of an injective. Repeating this argument splits the complex into adjacent pairs with identity differential after the splitting identifications. Thus a quasi-isomorphism between bounded-below complexes of injectives has a contractible cone and is a homotopy equivalence. Taking sections on such representatives respects this equivalence. In particular the H^q above are the cohomology of $R\Gamma(X, \mathcal{F})$.

53.4 The connecting homomorphism

A short exact sequence $0 \to \mathcal{A} \to \mathcal{B} \to \mathcal{C} \to 0$ admits a degreewise split short exact sequence of injective resolutions. In the first degree embed $\mathcal{A}, \mathcal{C}$ into $\mathcal{I}^0_A, \mathcal{I}^0_C$. Extend the first embedding to a map from $\mathcal{B}$ by injectivity. Together with the map through $\mathcal{C}$ it defines an injection into $\mathcal{I}^0_A \oplus \mathcal{I}^0_C$. The map to the second factor detects any element outside $\mathcal{A}$ on stalks, and the first is injective on $\mathcal{A}$. The diagram is

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathcal{A} & \longrightarrow & \mathcal{B} & \longrightarrow & \mathcal{C} & \longrightarrow & 0 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 0 & \longrightarrow & \mathcal{I}^0_A & \longrightarrow & \mathcal{I}^0_A \oplus \mathcal{I}^0_C & \longrightarrow & \mathcal{I}^0_C & \longrightarrow & 0. \end{array}$$

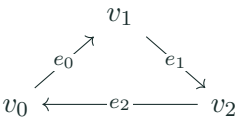
Its cokernels form a short exact sequence by the snake lemma. Apply the same construction to that sequence to obtain the next degree, and repeat. The differentials formed through the cokernels respect the inclusions and projections, giving the short exact sequence of complexes. Each row remains exact on taking sections because it is split. The long exact cohomology sequence of Chapter 51 now begins as follows.

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \Gamma(X, \mathcal{A}) & \longrightarrow & \Gamma(X, \mathcal{B}) & \longrightarrow & \Gamma(X, \mathcal{C}) & & \\ & & & & & & \downarrow{\scriptstyle \delta} & & \\ \cdots & \longleftarrow & H^1(X, \mathcal{C}) & \longleftarrow & H^1(X, \mathcal{B}) & \longleftarrow & H^1(X, \mathcal{A}). & & \end{array}$$

The connecting map δ lifts a cycle in the quotient complex to the middle complex and applies the differential. Since its image downstairs is zero, the result represents a class in the first complex. By exactness $\delta(c) = 0$ precisely when a global section c lifts to $\mathcal{B}(X)$. The class thus records the obstruction described in Section 53.1.

53.5 Computing on a finite cycle

Let X be the poset with vertices v_0, v_1, v_2 and edges e_0, e_1, e_2, with e_i above v_i and v_{i+1} and indices taken modulo three. Give X the topology whose opens are the upward closed subsets. The minimal open U_x containing x is $\{y : y \geq x\}$. For the calculation, orient the underlying graph as follows.



Consequently a stalk at x is the section group on U_x. For any abelian group A, let $\underline{A}$ be its sheaf of locally constant functions. Its section group on U_x is A, since a neighborhood of x inside U_x is the whole U_x.

Use the two sheaves

$$\mathcal{E} = \bigoplus_{x \in X} J_x(A), \qquad \mathcal{B} = \bigoplus_{v < e} J_v(A).$$

Here the second sum has one summand per vertex-edge incidence, and both sums are finite. A section of $\mathcal{E}$ on U consists of values a_x for $x \in U$. Where $v \in U$ one also has $e \in U$ for every incident e. Define $d \colon \mathcal{E} \to \mathcal{B}$ by $d(a)_{v,e} = a_e - a_v$ on those incidences. A locally constant function maps into $\mathcal{E}$ by its values, giving a complex

$$0 \longrightarrow \underline{A} \longrightarrow \mathcal{E} \xrightarrow{d} \mathcal{B} \longrightarrow 0.$$

At an edge stalk there is only the edge coordinate and no incidence coordinate, so the inclusion is an isomorphism there. At a vertex stalk the sequence is $0 \to A \to A^3 \to A^2 \to 0$, with the diagonal inclusion and the two differences against the vertex coordinate. Their kernel is the diagonal, and the differences can be prescribed independently. Hence the sheaf sequence is exact.

Each $J_x(A)$ has zero positive cohomology. The functor J_x sends an injective resolution of A to an injective resolution of $J_x(A)$, since it is exact on sections and preserves injectives by the stalk adjunction. Taking global sections of that resolution gives the original group resolution. The same vanishing holds for its finite sums, because finite sums of injectives are injective and taking sections commutes with finite direct sums. The connecting sequence therefore identifies $H^0(X, \underline{A})$ with the kernel of $d(X)$, $H^1(X, \underline{A})$ with its cokernel, and $H^q(X, \underline{A}) = 0$ for $q \geq 2$.

The kernel of $d(X)$ consists of constant assignments by connectedness of the incidence cycle, so it is A. For the cokernel define

$$\tau(b) = \sum_{i=0}^{2} (b_{v_i,e_i} - b_{v_{i+1},e_i}) \quad \text{for } b \in \mathcal{B}(X).$$

This sum is zero on $d(X)a$ by telescoping. Conversely if $\tau(b) = 0$, start at $a_{v_0} = 0$ and set $a_{v_{i+1}} = a_{v_i} + b_{v_i,e_i} - b_{v_{i+1},e_i}$ around the cycle. The zero-sum condition returns to the starting value. Then $a_{e_i} = a_{v_i} + b_{v_i,e_i}$ gives $d(X)a = b$. Finally τ is onto by using a single incidence coordinate. It gives $H^1(X, \underline{A}) \cong A$. For $A = \mathbb{Z}$, the vector b with $b_{v_0,e_0} = 1$ and all other coordinates zero has local lifts by stalkwise surjectivity. Since $\tau(b) = 1$, it has no global lift. This also gives an explicit failure of right exactness of $\Gamma(X, -)$.

53.6 The Mayer–Vietoris sequence

Suppose $X = U \cup V$ with U, V open and put $W = U \cap V$. For a resolution of $\mathcal{F}$ we may choose products of injective skyscrapers in every degree, as in Section 53.2. Their restrictions to an open retain the factors at points of that open, so they are again injective. The restrictions of the resolution are exact by the stalk test. In each degree the sequence of sections

$$0 \to \mathcal{I}^j(X) \to \mathcal{I}^j(U) \oplus \mathcal{I}^j(V) \xrightarrow{r_U - r_V} \mathcal{I}^j(W) \to 0$$

is exact by gluing and extension by zero in the point product. For surjectivity, place the desired coordinates into the U-component and use zero at its other points and on V. The associated long exact sequence is the two-open *Mayer–Vietoris sequence*. In particular, if the restrictions of $\mathcal{F}$ to U, V, W have zero positive cohomology, then the kernel and cokernel of

$$\mathcal{F}(U) \oplus \mathcal{F}(V) \xrightarrow{r_U - r_V} \mathcal{F}(W)$$

give H^0, H^1 and all higher groups vanish. This proves the required cover comparison under the stated vanishing hypothesis.

53.7 Twisting sheaves on the projective line

Apply the calculation to the projective line of Chapter 47 over a field k, with affine charts U, V of coordinates t, u and overlap $u = t^{-1}$. Form the twisting sheaf $\mathcal{O}(m)$ by gluing free rank-one modules, with local generators s_U, s_V satisfying $s_V = t^m s_U$ on the overlap. A sheaf of $\mathcal{O}$-modules is quasi-coherent if on each affine chart it comes from a module by localization. Here the two modules are free of rank one and their identification on the overlap is multiplication by a unit.

On U, V, W the restrictions are thus quasi-coherent. We use affine vanishing for quasi-coherent sheaves to deduce that on those three opens their higher cohomology as abelian sheaves is zero [45].

Use s_U to read sections on the overlap. The difference map d_m sends (f, g) to $f(t) - t^m g(t^{-1})$.

$$k[t] \oplus k[u] \xrightarrow{-d_m} k[t, t^{-1}].$$

Its kernel corresponds to Laurent polynomials whose powers are simultaneously ≥ 0 and $\leq m$. Its cokernel has a basis represented by the powers with exponents $m < i < 0$, since $k[t]$ spans exponents ≥ 0 and $t^m k[t^{-1}]$ spans exponents $\leq m$. Explicitly,

$$\dim_k H^0(\mathbb{P}^1_k, \mathcal{O}(m)) = \max(0, m+1), \qquad \dim_k H^1(\mathbb{P}^1_k, \mathcal{O}(m)) = \max(0, -m-1).$$

For $m = -2$ the missing power t^{-1} is a nonzero class. It is a section on the overlap that cannot be expressed as a difference of the permitted sections on the two charts. In degrees $q \geq 2$ the groups vanish by the same two-open calculation.

CHAPTER 54

Fundamental groupoids from paths

A path from x to y can be followed by a path from y to z. The result depends on how much time we give each part of the journey. Identifying paths that deform into one another with fixed endpoints makes composition associative and makes reversal an inverse. The resulting groupoid can retain several points of a space at once. This matters when the two open arcs covering a circle intersect in two components.

54.1 From continuous paths to arrows

Put $I = [0,1]$. A path $p\colon I \to X$ from x to y is continuous with $p(0) = x$ and $p(1) = y$. A homotopy *relative to the endpoints* from p to p' is a continuous map $H\colon I \times I \to X$ satisfying

$$H(t,0) = p(t), \quad H(t,1) = p'(t), \qquad H(0,s) = x, \quad H(1,s) = y.$$

This is an equivalence relation on paths with given endpoints. Use the constant homotopy for reflexivity, reverse its second coordinate for symmetry, and paste two homotopies along their common edge for transitivity. Write $[p]$ for the class of p.

For $p\colon x \to y$ and $q\colon y \to z$, let $q * p$ mean traversal of p followed by q, parametrized by

$$(q * p)(t) = \begin{cases} p(2t) & 0 \le t \le \tfrac{1}{2}, \\ q(2t-1) & \tfrac{1}{2} \le t \le 1. \end{cases}$$

The two formulas agree at their join. We compose homotopies by the same formula in the first coordinate. The endpoint condition then makes the two homotopies agree at the join for every value of the second coordinate. Hence $[q * p]$ depends only on $[p]$ and $[q]$.

Reparametrizing a path by a continuous $r\colon I \to I$ with $r(0) = 0$ and $r(1) = 1$ leaves its class unchanged. The homotopy is $p((1-s)t + sr(t))$. For three consecutive paths, the two parenthesized concatenations each traverse the path with three equal time intervals after a piecewise linear change of parameter. This proves associativity on classes. Inserting a constant path adds a pause, and the same argument with a piecewise linear parameter constant on the pause proves the unit laws.

Let $\bar{p}(t) = p(1-t)$. The path $\bar{p} * p$ equals $p(\min(2t, 2-2t))$. Replacing the argument of p by $(1-s)\min(2t, 2-2t)$ contracts it to the constant path at x, with both endpoints fixed. Applying the formula to $\bar{p}$ contracts $p * \bar{p}$ at y. Thus $[\bar{p}]$ is inverse to $[p]$.

Definition 54.1. The *fundamental groupoid* $\Pi_1(X)$ has the points of X for objects and endpoint-fixed homotopy classes for arrows. Composition is $[q] \circ [p] = [q * p]$. The *fundamental group* $\pi_1(X, x)$ is the automorphism group of x in $\Pi_1(X)$.

This construction has a universal property. An assignment of points to objects of a category $\mathcal{D}$, and paths to arrows between the corresponding objects, induces a unique functor $\Pi_1(X) \to \mathcal{D}$ if it preserves constant paths and concatenation and sends homotopic paths with fixed endpoints to equal arrows. Existence follows by passing to classes, and uniqueness follows because every arrow is a class. Any such assignment lands in isomorphisms on arrows, since it must send $[\bar{p}]$ to the inverse of the image of $[p]$.

54.2 Maps, homotopies, and basepoints

A continuous map $f\colon X \to Y$ sends x to $f(x)$ and $[p]$ to $[f \circ p]$. Postcomposition takes endpoint-fixed homotopies to endpoint-fixed homotopies and preserves the formulas for constants and concatenation. Composition and identity for continuous maps thus give those for the induced functors. We have a functor from spaces to groupoids.

A homotopy $H\colon X \times I \to Y$ from f to g gives an isomorphism of functors $\eta\colon \Pi_1(f) \Rightarrow \Pi_1(g)$. Its component η_x is the class of $s \mapsto H(x, s)$. For $p\colon x \to y$, the map $(t, s) \mapsto H(p(t), s)$ fills a square. The two routes around that square from $(0, 0)$ to $(1, 1)$ are homotopic with fixed endpoints inside I^2, by linear interpolation between the routes. Their images under the filling give the equality of composites in

$$\begin{array}{ccc} f(x) & \xrightarrow{[f\circ p]} & f(y) \\ \downarrow{\scriptstyle \eta_x} & & \downarrow{\scriptstyle \eta_y} \\ g(x) & \xrightarrow{[g\circ p]} & g(y). \end{array}$$

This is the naturality square. Each component is invertible by path reversal. In particular a homotopy equivalence of spaces induces an equivalence of their fundamental groupoids.

The isomorphism classes of objects of $\Pi_1(X)$ are the path components of X. In each component choose a point x. The full subgroupoid on the chosen points is equivalent to $\Pi_1(X)$, since its inclusion is fully faithful and essentially surjective. It is a disjoint union of one-object groupoids, with arrow groups $\pi_1(X, x)$. For a

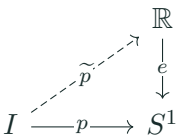
path $p\colon x \to y$, the corresponding change-of-basepoint isomorphism is

$$\pi_1(X,x) \longrightarrow \pi_1(X,y), \qquad a \longmapsto [p] \circ a \circ [p]^{-1}.$$

If p' is another choice, its isomorphism differs by conjugation by the loop $[p'] \circ [p]^{-1}$ at y. The whole groupoid records the paths that implement these changes. A full subgroupoid on a set $A \subseteq X$ of marked points, denoted $\Pi_1(X,A)$, keeps exactly the arrows whose two endpoints lie in A. It can contain several objects in the same component, which lets the same set of marks serve simultaneously in different open subsets.

54.3 Winding around a circle

Write $S^1 \subset \mathbb{C}$ for the unit circle and $e\colon \mathbb{R} \to S^1$ for $e(u) = \exp(2\pi i u)$. Every small open arc admits a continuous argument θ with $e\theta = \mathrm{id}$ there. A path p in S^1, with a chosen lift of its initial point, has a unique lift $\widetilde{p}\colon I \to \mathbb{R}$. For existence, subdivide I so the image of each closed subinterval lies in a single argument arc. Such a finite subdivision exists by the Lebesgue number lemma applied to the inverse images of the arcs. On each subinterval use $\theta \circ p$ translated by the integer needed to match the prescribed value at its left endpoint. Uniqueness follows because the difference of two lifts is an integer-valued continuous function on the interval, hence constant.



$$\widetilde{p}(0) = 0 \quad \text{if } p(0) = 1.$$

For a loop p at 1 use the lift with initial value 0. The final value $w(p) = \widetilde{p}(1)$ is an integer called the *winding number*. To see that it is invariant under an endpoint-fixed homotopy H, fix a parameter s_0 and choose a subdivision and argument arcs for $H(-, s_0)$. Compactness of each closed subinterval and openness of its chosen arc give a neighborhood of s_0 on which the same choices work. On this neighborhood, with subdivision points $0 = t_0 < \cdots < t_m = 1$ and corresponding arguments θ_j, the winding number is the continuous function

$$\sum_{j=1}^{m} \bigl(\theta_j(H(t_j, s)) - \theta_j(H(t_{j-1}, s))\bigr).$$

Indeed the integer shifts cancel in each difference, and the differences of the lift telescope. This locally continuous integer-valued function is constant on the parameter interval.

If two loops have winding number n, their lifts have common endpoints 0 and n. Linear interpolation between their lifts, followed by e, is a homotopy relative

to the endpoints. Every integer n occurs, through the loop $t \mapsto e(nt)$. Finally the lift of $q * p$ follows $\widetilde{p}$ and then $w(p) + \widetilde{q}$, with the same concatenation of time intervals. This proves

$$w\colon \pi_1(S^1, 1) \xrightarrow{\sim} \mathbb{Z}, \qquad w([q] \circ [p]) = w(q) + w(p).$$

The lifting argument is a special case of the use of covering spaces to compute fundamental groups [46, Sections 1.1 and 1.3].

54.3.1 The action on a covering fiber

Fix a positive integer n and let $c_n(z) = z^n$ on S^1. For a path p starting at x and a root $e(v)$ with $e(nv) = x$, take the real lift of p starting at nv. Then $\widehat{p}(t) = e(\widetilde{p}(t)/n)$ is a lift through c_n starting at that root. The lift is unique because the ratio of two lifts is continuous with values in the discrete set of n-th roots of unity, and their ratio starts at 1. In particular the construction depends only on the root $e(v)$.

The finite-sum argument for winding also applies to paths with any pair of fixed endpoints. The change in the real lift is continuous locally in the homotopy parameter, and the final value belongs to the discrete fiber of e over the common endpoint. The final value, and hence the final root of $\widehat{p}$, is constant along the homotopy. Taking the final root therefore gives a function $c_n^{-1}(x) \to c_n^{-1}(y)$ for each arrow $[p]\colon x \to y$. Uniqueness of lifts gives the identity and composition laws. This is a functor $\Pi_1(S^1) \to \mathbf{Set}$.

At $x = 1$, write $\zeta = e(1/n)$. For a loop of winding k, the real lift starting at j ends at $j + k$. The corresponding lift through c_n gives the fiber permutation in the right column.

$$\begin{array}{ccc} \mathbb{Z}/n & \xrightarrow{[j] \mapsto \zeta^j} & c_n^{-1}(1) \\ \Big\downarrow{\scriptstyle +k} & & \Big\downarrow{\scriptstyle \zeta^j \mapsto \zeta^{j+k}} \\ \mathbb{Z}/n & \xrightarrow{[j] \mapsto \zeta^j} & c_n^{-1}(1). \end{array}$$

Thus the winding-one loop acts by a cyclic shift, and a loop returns a given root to itself exactly when its winding is in $n\mathbb{Z}$. This action by lifting loops, called *monodromy*, is the topological counterpart of the action on fibers in Chapter 48.

54.4 Gluing across a disconnected overlap

Theorem 54.2 (Groupoid van Kampen theorem for two open sets)**.** *Let $X = U \cup V$ with U, V open, and put $W = U \cap V$. Suppose $A \subseteq X$ meets every path component of U, V, and W. When Z is one of these subspaces, $\Pi_1(Z, A \cap Z)$ uses*

paths and homotopies in Z. The inclusion functors form a pushout of groupoids

$$\begin{array}{ccc} \Pi_1(W, A \cap W) & \longrightarrow & \Pi_1(U, A \cap U) \\ \downarrow & & \downarrow \\ \Pi_1(V, A \cap V) & \longrightarrow & \Pi_1(X, A). \end{array}$$

Proof. First keep every point of X. Let $F_U \colon \Pi_1(U) \to G$ and $F_V \colon \Pi_1(V) \to G$ agree on $\Pi_1(W)$. The value on each point is determined. Subdivide each path into pieces contained in U or V, and compose their images. Refinement preserves the composite by reparametrization and the functor laws. On pieces in W either functor gives the same arrow. Comparing with a common refinement proves independence of the subdivision and open-set assignment.

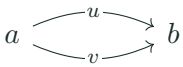
For a homotopy relative to endpoints, the Lebesgue number lemma gives a rectangular grid with each closed cell mapping into U or V. An edge assigned to both sets lies in W, so each edge has one image in G. Each cell gives a homotopy of the two routes around its boundary in the assigned set, so the image square commutes. Pasting gives the boundary equality for the whole grid. The vertical sides are constant, so the images of the paths agree. This defines the required functor $\Pi_1(X) \to G$, uniquely by subdivision.

Now choose, for each x, a path $\eta_x \colon r_x \to x$ from a mark $r_x \in A$ in the same component as x. Make the choice inside W for $x \in W$, inside U for $x \in U \setminus V$, and inside V for $x \in V \setminus U$. At a mark take $r_x = x$ and the constant path. Conjugating arrows by these paths defines a retraction onto the full subgroupoid on marks, on objects by $x \mapsto r_x$ and on arrows by

$$[p] \colon x \to y \;\longmapsto\; [\eta_y]^{-1} \circ [p] \circ [\eta_x] \colon r_x \to r_y.$$

These retractions commute with the inclusion functors. Precompose the given marked functors with the retractions for U, V to obtain agreeing full functors. Extend by the full pushout and restrict to the marks. Any two extensions on the marks give full extensions by precomposition with the retraction for X. These coincide, and each restricts to its original extension. □

This is the open two-set form of Brown's groupoid theorem [47]. For an explicit computation, take $X = S^1$, $U = S^1 \setminus \{-i\}$, and $V = S^1 \setminus \{i\}$. The overlap is the disjoint union of the open left and right semicircles. Put $a = 1$, $b = -1$, and take $A = \{a, b\}$. Both U and V are homeomorphic to intervals, so there is one path class from a to b in each. Call them u and v, represented by the upper and lower closed semicircles, both traversed from a to b. The marked overlap groupoid has two objects and only identity arrows.



The pushout is the free groupoid on these two arrows. By its universal property, a functor out of it assigns objects to a, b and chooses two invertible arrows between them. Loops at a are words in u, v and their inverses. Once immediate inverse pairs are cancelled, a nonempty loop repeats either $v^{-1}u$ or $u^{-1}v$, so every loop is a power of $v^{-1}u$. To see that no nonzero power is the identity, send both marks to the unique object of the one-object groupoid of $\mathbb{Z}$, with $u \mapsto 1$ and $v \mapsto 0$. The generator $v^{-1}u$ is also the positive loop in the lifting computation. Applying the automorphism group functor at a to the three local groupoids gives only trivial groups. The new loop uses a path between the marks from each side of the overlap.

54.5 What passes to the groupoid

In the fundamental groupoid, an endpoint-fixed homotopy equates the two arrows it connects. To see what this can discard, cover S^2 by the complements of its north and south poles. The two open sets are contractible and their intersection is path connected. With one marked point in the intersection, Theorem 54.2 shows that $\pi_1(S^2)$ at that point is trivial. As S^2 is path connected, its fundamental groupoid is equivalent to that of a point.

At the next dimension a based map $S^2 \to X$ can be specified by a square $I^2 \to X$ sending the whole boundary to the basepoint, since $I^2/\partial I^2 \cong S^2$. It gives a homotopy from the constant loop to itself with fixed endpoints. Taking based homotopy classes of these maps defines $\pi_2(X)$ at the chosen point. The identity of S^2 gives a nontrivial class. It acts as the identity on $H_2(S^2; \mathbb{Z}) \cong \mathbb{Z}$, while a constant map factors through a point, which has zero homology in that degree [46, Chapter 2]. Thus two spaces can have equivalent fundamental groupoids while their homotopies between paths behave differently. Retaining those homotopies requires a way to compose them and to account for the homotopies between them.

CHAPTER 55

Coherence in two dimensions

The category of categories has categories as objects and functors as arrows. Between parallel functors there can also be natural transformations. To retain them, we need a composition within each category of functors, and a composition that follows the composition of functors. Their compatibility gives the first law of two-dimensional composition. Pullbacks then explain why associativity itself sometimes needs an isomorphism.

55.1 Two compositions of transformations

Fix a universe of small categories. For each pair $\mathcal{A}, \mathcal{B}$, the functor category $[\mathcal{A}, \mathcal{B}]$ has functors as objects and natural transformations as arrows. We now call these *1-cells* and *2-cells*, respectively, in the two-dimensional category of small categories. Given $\alpha \colon F \Rightarrow F'$ and $\alpha' \colon F' \Rightarrow F''$, their *vertical composition* $\alpha' \cdot \alpha$ is the usual composition in $[\mathcal{A}, \mathcal{B}]$. Its component at x is $\alpha'_x \circ \alpha_x$. The identity 2-cell 1_F has components id_{Fx}.

Suppose $F, F' \colon \mathcal{A} \to \mathcal{B}$ and $G, G' \colon \mathcal{B} \to \mathcal{C}$. Given $\alpha \colon F \Rightarrow F'$ and $\beta \colon G \Rightarrow G'$, the two sides of the naturality square of β along α_x give candidate components for a 2-cell from GF to $G'F'$.

$$\begin{array}{ccc} GFx & \xrightarrow{G(\alpha_x)} & GF'x \\ \downarrow{\scriptstyle \beta_{Fx}} & & \downarrow{\scriptstyle \beta_{F'x}} \\ G'Fx & \xrightarrow{G'(\alpha_x)} & G'F'x \end{array}$$

They agree by naturality. In particular, the formula

$$(\beta * \alpha)_x = \beta_{F'x} \circ G(\alpha_x) = G'(\alpha_x) \circ \beta_{Fx}$$

defines the *horizontal composition* $\beta * \alpha$. Indeed, the components are a vertical composite of the natural transformations $G\alpha$ and $\beta F'$. Their naturality follows from the naturality of α after applying G, and the naturality of β along arrows in the image of F'. Composing horizontally with an identity 2-cell is called *whiskering*. It is precisely applying a functor to each component, or evaluating each component at the value of another functor, according to which 2-cell is the identity.

Now add $\alpha' \colon F' \Rightarrow F''$ and $\beta' \colon G' \Rightarrow G''$. The following diagram is in the hom category $[\mathcal{A}, \mathcal{C}]$. The direct route composes vertically in each input before composing horizontally. The other route composes horizontally in each row before composing vertically.

$$\begin{array}{ccc} GF & \xrightarrow{\beta * \alpha} & G'F' \\ & {\scriptstyle (\beta' \cdot \beta) * (\alpha' \cdot \alpha)} \searrow & \downarrow {\scriptstyle \beta' * \alpha'} \\ & & G''F'' \end{array}$$

The direct route has component

$$\beta'_{F''x} \circ \beta_{F''x} \circ G(\alpha'_x) \circ G(\alpha_x).$$

Naturality of β along α'_x replaces the middle two factors by $G'(\alpha'_x) \circ \beta_{F'x}$. The resulting four factors are the component of the other route. This proves the *interchange law*

$$(\beta' \cdot \beta) * (\alpha' \cdot \alpha) = (\beta' * \alpha') \cdot (\beta * \alpha).$$

Also $1_G * 1_F = 1_{GF}$. These equations express the functoriality of composition in both variables together.

Horizontal composition is associative in this example. For a third transformation $\gamma \colon H \Rightarrow H'$, both parenthesizations give the component

$$\gamma_{G'F'x} \circ H(\beta_{F'x}) \circ HG(\alpha_x).$$

The identity functor acts as a horizontal unit by the component formula. In general, a *strict 2-category* has an ordinary hom category $\mathcal{K}(A, B)$ for every pair of objects, composition functors

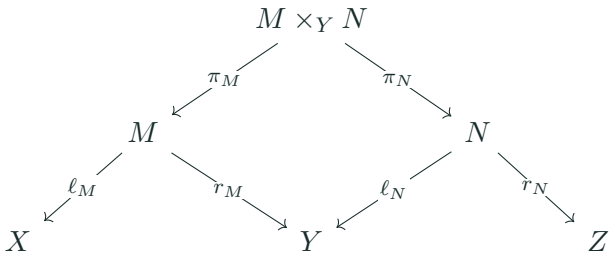
$$\mathcal{K}(B, C) \times \mathcal{K}(A, B) \longrightarrow \mathcal{K}(A, C),$$

and identity 1-cells. These functors satisfy the associative and unit laws as equalities of functors. Interchange inside any strict 2-category is preservation of composition by such a functor. The construction above makes **Cat** a strict 2-category [49].

55.2 Composing spans by pullback

Let $\mathcal{E}$ be a category with a choice of pullback for each cospan, as in Chapter 3. A *span* from X to Y consists of a diagram $X \xleftarrow{\ell_M} M \xrightarrow{r_M} Y$. The object M is its *apex*. In sets the elements of M can record witnesses that pair an element of X with an element of Y. Several witnesses may give the same pair. A map of spans over X, Y is a map $u \colon M \to M'$ that preserves both legs. These are the 2-cells. The usual composition and identities of maps of apices give the hom category $\mathrm{Span}_{\mathcal{E}}(X, Y)$.

For a span F from X to Y with apex M and a span G from Y to Z with apex N, the composite $G \odot F$ uses witnesses that agree at Y. The universal object for this agreement is the pullback $M \times_Y N$.



The left and right legs of the composite are $\ell_M \pi_M$ and $r_N \pi_N$. The identity span 1_X has apex X and both legs id_X. To compose 2-cells $u\colon M \to M'$ and $v\colon N \to N'$ horizontally, project the source pullback to M, N and apply u, v. The results agree over Y since u, v preserve the span legs. They therefore give a unique map to $M' \times_Y N'$, and that map preserves the two outer legs. Identity and composition follow by projecting to M' and N' at the target of the proposed map. Those projections determine a map into the pullback. Thus span composition is itself a functor, so horizontal and vertical composition again satisfy interchange.

55.3 Comparing parenthesizations and units

Take a third span H from Z to W with apex P. The two ways of iterating the pullback give

$$(M \times_Y N) \times_Z P \qquad \text{and} \qquad M \times_Y (N \times_Z P).$$

For each object T, a map from T to either apex is equivalent to a triple of arrows $m\colon T \to M$, $n\colon T \to N$, and $p\colon T \to P$ satisfying

$$r_M m = \ell_N n, \qquad r_N n = \ell_P p.$$

For the first apex, use the first equation to map into the inner pullback and the second for the outer pullback. For the second apex, apply the two equations in the reverse order. The correspondences respect precomposition in T. Each apex carries its own compatible triple of projections, which specifies a map to the other apex. Composing the two maps preserves the original projections, so the composite is the identity on each apex. The maps preserve the two outer legs as well. In the order used here the resulting associator is

$$a_{H,G,F}\colon (H \odot G) \odot F \xRightarrow{\cong} H \odot (G \odot F).$$

For pullbacks in sets defined by pairs its map on witnesses is $(m, (n, p)) \mapsto ((m, n), p)$. Given maps of all three input spans, both orders of applying these

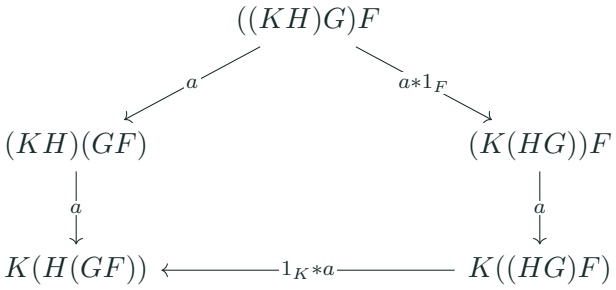
maps and reparenthesizing have the same three projections. This proves the naturality of a.

The left unitor $\lambda_F \colon 1_Y \odot F \Rightarrow F$ is projection from $M \times_Y Y$ to M. The map in the reverse direction is specified by (id_M, r_M). The right unitor $\rho_F \colon F \odot 1_X \Rightarrow F$ is projection from $X \times_X M$ to M, with inverse specified by (ℓ_M, id_M). In each case the two maps are inverse by the uniqueness part of the pullback property, and they preserve the legs. Naturality follows because projection to M after a map of pullbacks is the same as projection followed by the map of apices.

The common universal property extends to any nonempty chain of composable spans. A map into the apex for a given parenthesization is a list of maps into all the original apices. Each pair of successive maps agrees over the intermediate endpoint. To prove this, split the parenthesization into its two last blocks. The projections to those two block apices account for the equations inside the blocks by induction. The defining pullback adds exactly the one equation between the blocks. An identity span has both legs equal to the identity. Projections to inserted identity spans are determined along each chain of identities by the adjacent leg of an original span. For a chain consisting only of identity spans at X, all projections agree and give a single map to X.

55.4 The pentagon and the triangle

In a *bicategory*, the hom categories and composition functors have the same roles as in a strict 2-category. Associativity and the unit laws are supplied by specified natural isomorphisms a, λ, ρ of the forms just constructed. To make these comparisons compatible, the definition requires two equations [48]. The first compares two ways of reparenthesizing four 1-cells. Suppress $\odot$ in the display. Every arrow is the indicated associator on the applicable factors, or its whiskering with an identity.



For spans, all five apices represent the same lists of four compatible arrows. Each edge preserves the projections to the four original apices. The routes from $((KH)G)F$ to $K(H(GF))$ therefore give equal maps by uniqueness. This is the pentagon equation.

The other equation involves two 1-cells $F\colon X \to Y$ and $G\colon Y \to Z$ and the identity of their shared endpoint.

$$\begin{array}{ccc} (G \odot 1_Y) \odot F & \xrightarrow{a_{G,1_Y,F}} & G \odot (1_Y \odot F) \\ & {\scriptstyle \rho_G * 1_F} \searrow \quad \downarrow {\scriptstyle 1_G * \lambda_F} & \\ & G \odot F & \end{array}$$

For spans with apices M, N, the source represents triples (m, y, n) with $r_M m = y = \ell_N n$. The route through the associator then deletes the projection to Y. The other route deletes it directly. Both maps preserve the projections to M, N, which determine the target map to $M \times_Y N$. This proves the triangle equation and completes the construction of the bicategory of spans. The same projection argument proves that every sequence of these comparisons and their inverses between two fixed span composites of a given word gives the same map. For a general bicategory, the coherence theorem equates parallel formal composites of associators, unitors, their inverses, and identity 2-cells, using vertical and horizontal composition on a fixed string of generating 1-cells. The strictification theorem says that every bicategory is biequivalent to a strict 2-category [48].

55.5 Counting finite span witnesses

For finite sets, let $M_{x,y}$ be the set of witnesses in M above the pair (x, y). Given the same spans F, G as before, the witness set above (x, z) in the composite is the following disjoint union.

$$(M \times_Y N)_{x,z} \cong \coprod_{y \in Y} M_{x,y} \times N_{y,z}.$$

The bijection assigns a pair of agreeing witnesses to the summand indexed by their common intermediate value. Consequently the matrix of witness counts for $G \odot F$, with source indices as rows, is the matrix product of the count matrix of F with that of G. The identity span gives the identity matrix. Two spans over fixed finite endpoints are isomorphic precisely when their count matrices agree. An isomorphism restricts to a bijection on each fiber, and a bijection on every fiber assembles to a map of spans with inverse. Every matrix with nonnegative integer entries occurs by taking a disjoint set of the prescribed size for every pair.

For example, take the outer endpoints to be singletons and the two intermediate sets to be $\{0, 1\}$. In the first span choose two witnesses above 0, called u, v, and one above 1, called w. In the middle span choose one witness each for the transitions $0 \to 1$, $1 \to 0$, and $1 \to 1$. In the final span choose one witness above 1. A compatible triple can start with u or v and take $0 \to 1$, or start with w

and take $1 \to 1$. The two parenthesizations group these same triples differently. Counting either of them gives the scalar

$$\begin{pmatrix} 2 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 3.$$

The associator preserves each of the three witnesses. Passing to isomorphism classes of spans thus gives associativity as equality. On finite sets the resulting category is equivalently described by these matrices, with multiplication in the source-as-row convention.

The full bicategory retains maps between witness sets. For instance, over singleton endpoints the two-element span has a 2-cell to the one-element span that identifies the two witnesses. It also has an invertible 2-cell that exchanges them. The count matrices record the underlying span classes. The 2-cells track what happens to individual witnesses.

55.6 From equations to comparison cells

The fundamental groupoid of Chapter 54 identifies paths when there is an endpoint-fixed homotopy between them. Reparametrizations then give associative composition on path classes. A bicategory retains comparison cells between 1-cells. Its pentagon and triangle are equations between composites of 2-cells. For spans, those equations follow from uniqueness of maps preserving projections. When a comparison between 2-cells is itself part of the data, it requires a next dimension. The identity types of Chapter 29 allow witnesses of equality between path witnesses. The next chapter develops a model of spaces in which composites can be chosen and compared through simplices of successively higher dimensions.

CHAPTER 56

The infinity-category of spaces

The fundamental groupoid of a circle has one isomorphism class of objects and a copy of $\mathbb{Z}$ as the automorphism group at each object. Passing to path components preserves the first statement and forgets the second. There is an analogous loss when we pass from a space of maps to the set of homotopy classes of maps. To retain the paths between maps and the homotopies between those paths, we will equip maps with a simplicial set of higher data.

56.1 The maps to invert

For a topological space X, the set $\pi_0(X)$ consists of its path components. For $x \in X$ and $n \geq 1$, the group $\pi_n(X, x)$ consists of based homotopy classes of maps $(S^n, *) \to (X, x)$. The group law can be given by viewing a map as a map of the cube I^n that sends its boundary to x, then concatenating in the first coordinate. In degree one this is the automorphism group of x in the fundamental groupoid of Chapter 54.

A continuous map $f\colon X \to Y$ is a *weak homotopy equivalence* if $\pi_0(f)$ is a bijection and

$$\pi_n(f)\colon \pi_n(X, x) \xrightarrow{\sim} \pi_n(Y, f(x))$$

is an isomorphism for every $x \in X$ and $n \geq 1$. The tests include every path component. When f has a homotopy inverse, the homotopies supply the changes of basepoint needed to see that f passes these tests. For spaces of the homotopy type of CW complexes, Whitehead's theorem gives the converse [46, §4.1].

Example 56.1 (Two topologies and one weak homotopy type)**.** Let $\mathbb{Q}_{\text{disc}}$ carry the discrete topology and let $\mathbb{Q}$ carry its subspace topology from $\mathbb{R}$. The map $\mathbb{Q}_{\text{disc}} \to \mathbb{Q}$ given by the identity on points is a weak homotopy equivalence. Each connected subset of $\mathbb{Q}$ is a singleton, since two different rationals are separated by an irrational cut. Paths and maps from S^n, $n \geq 1$, are therefore constant in both spaces. Every homotopy into $\mathbb{Q}$ is pointwise stationary, since each point traverses a path in $\mathbb{Q}$. A homotopy inverse would have to be the identity on points from $\mathbb{Q}$ to $\mathbb{Q}_{\text{disc}}$. This function is discontinuous, as no singleton is open in $\mathbb{Q}$.

We seek a construction in which every weak homotopy equivalence becomes an equivalence. This is a version of localization, as in Chapter 52. Here the

universal property will control a category of functors carrying their own higher data. We first need a language in which an individual map, a homotopy between maps, and a higher homotopy are all part of one object.

56.2 Simplices and horn filling

Write $\mathbf{\Delta}$ for the category of finite nonempty ordinals $[n] = \{0, \dots, n\}$ and order-preserving maps. A *simplicial set* K is a functor $\mathbf{\Delta}^{\mathrm{op}} \to \mathbf{Set}$. Its n-simplices are the elements of $K_n = K([n])$. Injections that omit one vertex induce the face maps. Surjections that repeat one vertex induce the degeneracy maps. The functor laws express that the operations agree whenever the ordinal maps defining them agree. The representable $\Delta^n = \mathrm{Hom}_{\mathbf{\Delta}}(-, [n])$ is the standard simplicial n-simplex. By the Yoneda lemma, giving a map $\Delta^n \to K$ is giving an element of K_n.

The geometric n-simplex is

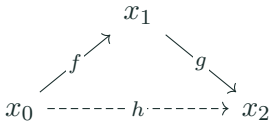
$$|\Delta^n| = \left\{ (t_0, \dots, t_n) \in [0,1]^{n+1} \;\middle|\; \sum_{j=0}^{n} t_j = 1 \right\}.$$

The ordinal maps act on these simplices by sending vertices to vertices and extending affinely. Define the *singular simplicial set* by

$$(\mathrm{Sing}\, X)_n = \mathrm{Hom}_{\mathbf{Top}}(|\Delta^n|, X).$$

A vertex is a point of X, an edge is a parameterized path, and a 2-simplex is a continuous triangle with three boundary paths. The edge $0 \to 2$ of that triangle and the two-edge path $0 \to 1 \to 2$ are homotopic with endpoints fixed. Within $|\Delta^2|$, use straight-line interpolation between the parameterized paths with these endpoints and compose with the map of the triangle to X.

For $n > 0$, the *horn* $\Lambda^n_i \subseteq \Delta^n$ is the union of all the codimension-one faces except the face opposite i. An arrow $\Lambda^n_i \to K$ specifies these faces with their common subfaces identified. An extension to Δ^n is a *filler*. For the inner horn Λ^2_1, a filler includes a choice of edge h and the whole triangle. A dashed arrow in this diagram indicates the edge supplied by the filler.



A *Kan complex* has fillers for all horns. An ∞-category in the simplicial-set model, also called a *quasicategory*, has fillers for the inner horns $0 < i < n$. The ordinary nerve of a category describes composable chains of arrows. In its inner

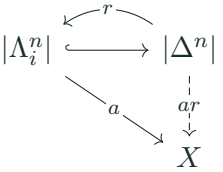
two-dimensional horn, $h = gf$ and the filler is unique. The outer horns impose a different condition. For example Λ^2_0 specifies the arrows $x_0 \to x_1$ and $x_0 \to x_2$. Taking $x_2 = x_0$ and the latter edge to be an identity asks for a left inverse up to a triangle. The other outer horn asks for a right inverse up to a triangle. The two inverses agree after passing to homotopy classes. This is the sense in which a Kan complex models an ∞-groupoid.

Proposition 56.2. *The singular simplicial set of any topological space is a Kan complex.*

Proof. A singular horn gives continuous maps from its closed faces that agree on intersections. Gluing them gives a map $|\Lambda^n_i| \to X$. In barycentric coordinates, put $m = \min_{j \neq i} t_j$ and define $r(t) = u$ by

$$u_i = t_i + nm, \qquad u_j = t_j - m \quad (j \neq i).$$

The coordinates are nonnegative and sum to one. Some $u_j = 0$ with $j \neq i$, so u belongs to $|\Lambda^n_i|$. On that horn $m = 0$, so r fixes it. The formula gives a continuous retraction, and composing with r provides the extension to the simplex.



□

The horn condition leaves room for higher homotopy. For instance a singular 2-simplex in S^2 can have its whole boundary constant and still give the generator of $\pi_2(S^2)$. Indeed the quotient $|\Delta^2|/|\partial\Delta^2|$ is a two-sphere and the quotient map gives the required simplex after identifying that quotient with S^2.

56.3 Mapping spaces and coherent composition

Given simplicial sets K and L, the simplicial mapping object L^K is defined by

$$(L^K)_n = \mathrm{Hom}_{\mathbf{sSet}}(K \times \Delta^n, L).$$

Its faces and degeneracies are obtained by precomposition in the second factor. Vertices are maps $K \to L$. An edge from f to g is a map $K \times \Delta^1 \to L$ restricting to f and g at the endpoints. Higher simplices are families of maps with the corresponding parameter simplices.

If L is Kan, then L^K is Kan for every K. The horn problem for L^K becomes the extension problem

$$\begin{array}{ccc} K \times \Lambda^n_i & \longrightarrow & L \\ \downarrow & \nearrow & \\ K \times \Delta^n & & \end{array}$$

by the product-exponential adjunction. The required lifting fact for simplicial sets says that $K \times \Lambda^n_i \to K \times \Delta^n$ is anodyne. This means it lies in the class generated by the horn inclusions under pushout, transfinite composition, and retract. The claim follows from the pushout-product property for a monomorphism and a horn, applied to $\varnothing \to K$ [50, §3.1.3]. Since each construction preserves the lifting property against $L \to *$, the filler exists.

The Kan complexes form a category enriched in simplicial sets, denoted $\mathbf{Kan}_\Delta$, with mapping objects L^K. Enrichment means that composition is a map of simplicial sets

$$M^L \times L^K \longrightarrow M^K.$$

In degree n it composes families with the same parameter. If $v\colon K \times \Delta^n \to L$ and $u\colon L \times \Delta^n \to M$, their composite sends (x, t) to $u(v(x, t), t)$. Reusing the same parameter makes composition commute with all simplicial operators. Associativity and the identity law follow from those for functions.

The *homotopy coherent nerve* of this enriched category is the ∞-category $\mathcal{S}$ of spaces [51, §1.2.16]. Its simplices can be specified by testing against simplicially enriched versions of $[n]$. The version of $[n]$ has the same objects as $[n]$. For $i \le j$ its mapping simplicial set is the nerve of the poset of subsets of $\{i, \dots, j\}$ containing i, j, ordered by inclusion. For $i > j$ the mapping simplicial set is empty. Composition is union of subsets. An n-simplex of the coherent nerve is an enriched functor from this enriched $[n]$ into $\mathbf{Kan}_\Delta$. For $n = 2$, the mapping object from 0 to 2 contains the inclusion $\{0, 2\} \subseteq \{0, 1, 2\}$. Its image gives an edge in the mapping space from the chosen direct arrow to the composite through vertex 1.

The coherent nerve of a simplicial category whose mapping objects are Kan is an ∞-category. The mapping spaces of the nerve are equivalent to the original mapping objects [50, §§2.4.5 and 4.6.7]. Consequently

$$\begin{aligned} \mathrm{Map}_{\mathcal{S}}(K, L) &\simeq L^K, \\ \mathrm{Hom}_{h\mathcal{S}}(K, L) &= \pi_0(L^K). \end{aligned}$$

Here $h\mathcal{S}$ denotes the ordinary homotopy category. For example, $\mathrm{Map}_{\mathcal{S}}(*, K) \simeq K$, since $* = \Delta^0$. For $K = \mathrm{Sing}\, S^1$ the hom set $\mathrm{Hom}_{h\mathcal{S}}(*, K)$ has one element. The mapping space itself has fundamental group $\mathbb{Z}$ by Chapter 54. Taking components of mapping spaces is what removes that group from this hom set.

56.4 Localization and comparison with topology

Geometric realization $K \mapsto |K|$ glues the topological simplices indexed by elements of K_n, identifying them according to the simplicial operators. A compatible family of continuous maps from those simplices to X is the data of a map $K \to \operatorname{Sing} X$. This gives the adjunction $|-| \dashv \operatorname{Sing}$. Call a map of simplicial sets a weak homotopy equivalence if its realization is one.

The Kan–Quillen model structure on simplicial sets takes these as its weak equivalences, takes the monomorphisms as cofibrations, and takes the maps with the right lifting property against all horns as fibrations [51, §A.2.7]. All simplicial sets are cofibrant, since $\varnothing \to K$ is a monomorphism. A simplicial set is fibrant precisely when it is Kan. This is a simplicial model structure for the mapping objects above. The Dwyer–Kan localization theorem identifies the coherent nerve of the cofibrant-fibrant objects of a simplicial model category with the ∞-categorical localization of its ordinary category of cofibrant objects at the weak equivalences [52, Theorem 1.3.4.20]. In the present case, if W denotes those equivalences, the identification gives

$$\mathrm{N}(\mathbf{sSet})[W^{-1}] \simeq \mathcal{S}.$$

Here N is the ordinary nerve. The universal property is that precomposing with the localization map identifies $\operatorname{Fun}(\mathcal{S}, \mathcal{D})$ with the full subcategory of $\operatorname{Fun}(\mathrm{N}(\mathbf{sSet}), \mathcal{D})$ on functors sending W to equivalences. This identification controls transformations and higher simplices as well as the objects of the functor category.

For every simplicial set K and every topological space X, the unit and counit of the realization–singular adjunction are weak homotopy equivalences in their respective categories. This is the comparison theorem of Milnor and its counit corollary [50, §3.6.4]. Naturality of the counit gives the square

$$\begin{array}{ccc} |\operatorname{Sing} X| & \xrightarrow{|\operatorname{Sing} f|} & |\operatorname{Sing} Y| \\ \downarrow{\scriptstyle \varepsilon_X} & & \downarrow{\scriptstyle \varepsilon_Y} \\ X & \xrightarrow{\qquad f \qquad} & Y. \end{array}$$

The vertical maps are weak homotopy equivalences, so f is one if and only if $|\operatorname{Sing} f|$ is one. Realization preserves weak equivalences by their definition on simplicial sets. The two functors therefore descend to the localizations, where the unit and counit give inverse equivalences. In particular the homotopy theory of all topological spaces with weak homotopy equivalences inverted is presented by $\mathcal{S}$.

Between Kan complexes a map is an equivalence in $\mathcal{S}$ when its class in $h\mathcal{S}$ is invertible, which means it has a simplicial homotopy inverse. The simplicial

Whitehead theorem detects this by a bijection on components and isomorphisms on all based homotopy groups [50, Theorem 3.2.7.1]. For $K = \operatorname{Sing} X$ the groups agree with the topological groups of X by the comparison theorem.

56.5 The Möbius function as an Euler characteristic

For a finite poset P, give integer-valued functions on intervals $[x, y]$ the convolution product

$$(f * g)(x, y) = \sum_{x \le z \le y} f(x, z) g(z, y).$$

Associativity comes from summing over the same chains with either grouping. The unit δ is 1 on degenerate intervals and 0 elsewhere. The *Möbius function* μ is the convolution inverse of the function $\zeta(x, y) = 1$. Write $u = \zeta - \delta$. The power $u^{*k}(x, y)$ counts strict chains $x = x_0 < \cdots < x_k = y$, so powers of sufficiently large degree vanish. The finite geometric series gives

$$\mu = \delta - u + u^{*2} - u^{*3} + \cdots .$$

For $x < y$, a strict chain with k arrows from x to y gives a simplex of dimension $k - 2$ in the order complex of the open interval (x, y). This order complex has the same realization as the ordinary nerve of the interval, whose nondegenerate simplices are the strict chains. The $k = 1$ chain corresponds to the empty simplex and contributes -1. Since the geometric series signs are $(-1)^k = (-1)^{k-2}$, its value is the reduced Euler characteristic:

$$\mu(x, y) = \widetilde{\chi}(|\mathrm{N}(x, y)|).$$

This is Philip Hall's theorem [70, Propositions 3.8.5–3.8.6].

Example 56.3 (The proper faces of a triangle)**.** In the poset of subsets of a three-element set, the open interval between the empty set and the whole set consists of three vertices and three edges of a triangle, ordered by incidence. Its nerve has six vertices and six edges with no higher nondegenerate simplices. It realizes the subdivided boundary circle, so $\mu(\varnothing, \{1, 2, 3\}) = 6 - 6 - 1 = -1$. The calculation packages the alternating counts into the Euler characteristic of a space.

56.6 The fundamental groupoid as truncation

For a simplicial set A, let hA be the category presented by its vertices and edges, with a degenerate edge acting as an identity and each triangle imposing $[e_{02}] = [e_{12}][e_{01}]$. When K is Kan this category is its fundamental groupoid $\Pi_1 K$. The relation on edges with specified endpoints is relative homotopy [50, §1.4.3].

In $\mathrm{Sing}\, X$ this agrees with ordinary endpoint-fixed path homotopy, since $\Delta^1 \times \Delta^1$ realizes as the square triangulated along a diagonal. The inner horn fillers give composition of edge classes, and the outer horns give inverses. Thus $\Pi_1 \,\mathrm{Sing}\, X$ recovers the path groupoid.

The ordinary nerve $\mathrm{N}G$ of a groupoid G is Kan. Its inner horns compose edges and the outer horns use their inverses. It is a 1-*type*, meaning that its homotopy groups above degree one vanish. Indeed, in a pointed Kan complex a class in π_n can be represented by an n-simplex whose faces are constant at the basepoint [50, §3.2.2]. For $n \geq 2$ all the edges of such a simplex are constant. A nerve simplex is determined by its successive edges, so this simplex is itself constant.

There is a canonical map $K \to \mathrm{N}\Pi_1 K$ sending edges to their classes. It preserves π_0 and π_1, since these are the component set and the groups of automorphisms of $\Pi_1 K$. If K is already a 1-type, the map is an equivalence by the simplicial Whitehead theorem. Thus every 1-type is represented up to equivalence by a groupoid nerve.

The same map has a universal property with values in all 1-types. To see it, first note that giving a map $A \to \mathrm{N}G$ is equivalent to giving a functor $hA \to G$. In each higher dimension, the image of a simplex in the nerve is determined by its successive edges, and its compatibility with each longer edge is exactly the relation imposed by the triangles. Also $h(A \times B) \cong hA \times hB$. To verify this, an edge in the product decomposes as an edge in A at a fixed B-vertex followed by an edge in B at a fixed A-vertex. The two orders of these moves agree, by combining the degeneracies of the two edges into triangles in the product. The triangle relations for each factor apply with the other factor held constant. These relations give the inverse to the functor induced by the two projections.

Applying these facts in each degree yields

$$\begin{aligned} ((\mathrm{N}G)^K)_n &\cong \mathrm{Fun}(h(K \times \Delta^n), G)_0 \\ &\cong \mathrm{Fun}(\Pi_1 K \times [n], G)_0 \\ &\cong \big(\mathrm{N}\,\mathrm{Fun}(\Pi_1 K, G)\big)_n. \end{aligned}$$

Here $\mathrm{Fun}(-,-)_0$ denotes the set of objects of the ordinary functor category. The identifications are natural in the ordinal $[n]$. Because natural transformations into a groupoid are invertible, the right-hand side describes the nerve of a groupoid. The calculation with K replaced by $\mathrm{N}\Pi_1 K$ has the same result, and precomposition with the canonical map gives the identification. Thus

$$\mathrm{Map}_{\mathcal{S}}(\mathrm{N}\Pi_1 K, T) \xrightarrow{\sim} \mathrm{Map}_{\mathcal{S}}(K, T) \qquad (T \text{ a 1-type}).$$

In the terminology of Chapter 5, this is the mapping-space form of a left adjoint to the inclusion of 1-types. Its value at K is called the 1-truncation of K.

Example 56.4 (Maps out of a circle)**.** For a group G let BG be its one-object groupoid. The mapping space from $\operatorname{Sing} S^1$ to $\mathrm{N}BG$ is equivalent to the nerve of $\operatorname{Fun}(B\mathbb{Z}, BG)$, since the fundamental groupoid of S^1 is equivalent to $B\mathbb{Z}$. A functor is specified by the image $g \in G$ of the integer 1. A natural transformation from the functor for g to that for g' is an element $a \in G$ satisfying $ag = g'a$. It suffices to impose naturality on the generator, as every other integer is a power of that generator. Thus the components of the mapping space are conjugacy classes and the fundamental group at g is its centralizer $C_G(g) = \{a \in G \mid ag = ga\}$. For $G = S_3$, the identity, a transposition, and a 3-cycle represent the three classes. Their centralizers are S_3, the order-two subgroup generated by the transposition, and the order-three subgroup generated by the cycle, respectively. The hom set in $h\mathcal{S}$ records three maps up to homotopy. The mapping space assigns a different group of loops to each.

Example 56.5 (The sphere and the point)**.** The van Kampen calculation in Chapter 54 shows that S^2 is simply connected. Its cellular chain complex has one generator in degree 0 and one in degree 2, giving $H_2(S^2; \mathbb{Z}) \cong \mathbb{Z}$. By the Hurewicz theorem in degree 2, $\pi_2(S^2) \cong \mathbb{Z}$ [46, Theorem 4.32]. Consequently the map $S^2 \to *$ is an equivalence after 1-truncation and fails to be an equivalence in $\mathcal{S}$.

The next constructions apply universal properties to entire mapping spaces. The distinction will already matter for pullbacks, where compatibility between two arrows is itself specified by a path.

CHAPTER 57

Homotopy limits and colimits

Two maps from a point into the circle can name the same point. Their ordinary pullback then consists of one point. A homotopy pullback asks how the two images are identified. In this example an identification is a loop in the circle, and the possible winding numbers are already enough to distinguish infinitely many components of the answer. The universal property sees these components because its test objects probe an entire mapping space.

57.1 Universality in mapping spaces

Let K be a small indexing shape and $D\colon K \to \mathcal{C}$ a diagram in an infinity-category. Write ΔT for the constant diagram at T. A *coherent cone* with vertex T is a point of $\mathrm{Map}_{\mathrm{Fun}(K,\mathcal{C})}(\Delta T, D)$. Its maps to the objects of D come with homotopies along the arrows of D and compatibility data along higher simplices of K. Paths in the cone space are homotopies of these entire families. Reversing the direction of the transformation gives the space of coherent cocones from D to ΔT.

A limit L has a specified cone $\lambda\colon \Delta L \to D$. For every T, composing with this cone must give an equivalence of spaces. Dually, a colimit Q has a cocone $\mu\colon D \to \Delta Q$ with the corresponding property.

$$\mathrm{Map}_{\mathcal{C}}(T, L) \xrightarrow{\simeq} \mathrm{Map}_{\mathrm{Fun}(K,\mathcal{C})}(\Delta T, D),$$
$$\mathrm{Map}_{\mathcal{C}}(Q, T) \xrightarrow{\simeq} \mathrm{Map}_{\mathrm{Fun}(K,\mathcal{C})}(D, \Delta T).$$

Thus for a given coherent cone to D, the space of maps $T \to L$ together with a homotopy identifying the resulting cone with the given one is contractible. This space is a homotopy fiber of the first comparison. In particular the universal properties determine the limit and colimit up to equivalence. When the limits and colimits exist for every D, the natural mapping-space equivalences exhibit the adjunctions

$$\mathrm{colim}_K \dashv \Delta \dashv \lim_K .$$

This is the same variance as in Chapter 25. For spaces $\mathcal{S}$, all small limits and colimits exist. We use this general existence result and compute two shapes explicitly [50, §7.4].

57.2 Pullbacks from paths with prescribed endpoints

Take two maps of spaces $f\colon X \to Z$ and $g\colon Y \to Z$. Use Kan complexes to present the three spaces and simplicial maps to present the arrows. Below Z^{Δ^1} denotes the simplicial function complex. Its endpoint map is a Kan fibration. This is the simplicial mapping-space property that restriction along a monomorphism into a Kan target is a Kan fibration, applied to $\partial\Delta^1 \hookrightarrow \Delta^1$. The pullback in simplicial sets in the following display defines P.

$$\begin{array}{ccc} P & \longrightarrow & Z^{\Delta^1} \\ \downarrow & & \downarrow{\scriptstyle (e_0,e_1)} \\ X \times Y & \xrightarrow{f\times g} & Z \times Z. \end{array}$$

A point of P is (x, y, γ) with $\gamma\colon f(x) \rightsquigarrow g(y)$. More generally an n-simplex consists of maps $x\colon \Delta^n \to X$ and $y\colon \Delta^n \to Y$, and an interval of maps $\gamma\colon \Delta^n \times \Delta^1 \to Z$ with the same endpoint formulas. The fibration property implies that P is Kan, since $X \times Y$ is Kan.

Exponentiating by a test simplicial set T preserves the ordinary pullback, since it is right adjoint to product with T. Consequently P^T consists of maps $T \to X$ and $T \to Y$ with a path between their composites to Z, and all families of these data. This is the mapping-space test for a pullback cone on the cospan. In this test one can record a map $T \to Z$ and two paths to it as well. The two triangles at the cone vertex correspond to these two paths, and the index has no nondegenerate simplices of higher dimension. The two paths give the same homotopy type of data as a single path between the composites. In the topological path description, join the two legs at their common end, reversing the second, and recover the common end at the midpoint. In the simplicial description the two legs form the horn Λ^2_2 of edges 02 and 12. Both that horn and the edge 01 include into Δ^2 by monomorphisms which are weak equivalences. The two function complex restrictions from Z^{Δ^2} are therefore trivial Kan fibrations over the pair of endpoint values at $0, 1$. This gives an equivalence of the two descriptions together with their families, by pulling back along the pair of prescribed endpoint maps [50, §3.1.3 and §3.4].

There is an equivalent computation by replacing one leg. Define $\widetilde{X} = X \times_{Z,e_0} Z^{\Delta^1}$ and factor f as

$$\begin{array}{ccc} X & \xrightarrow{c} & \widetilde{X} \\ & \searrow^{f} & \downarrow{\scriptstyle e_1} \\ & & Z. \end{array}$$

Here c uses constant paths. The map $\widetilde{X} \to X$ is a trivial Kan fibration, by restriction to one endpoint of Δ^1 and pullback. Hence c is a weak equivalence.

The endpoint map $\widetilde{X} \to Z$ is a Kan fibration, since it factors through $X \times Z$ by base change of the two-endpoint map, and X is Kan. The ordinary pullback $\widetilde{X} \times_Z Y$ is exactly P. In the Kan model an ordinary pullback along a Kan fibration between Kan complexes computes the homotopy pullback [50, §3.4.1]. If one original leg is already such a fibration, this criterion applies without a replacement. The two descriptions give the same object, denoted $X \times^{\mathrm{h}}_Z Y$.

57.3 Fibers, loops, and base change

For a point $b\colon * \to B$ and a map $p\colon E \to B$, the *homotopy fiber* $F_b = E \times^{\mathrm{h}}_B *$ consists of points of E together with paths from their image to b. When $p = b\colon * \to B$, it gives the based loop space, with basepoint the constant loop. In the continuous path model this is

$$\Omega_b B = \{\gamma\colon [0,1] \to B \mid \gamma(0) = \gamma(1) = b\}.$$

For $B = S^1$ and $b = 1$, a path lifts uniquely to the covering $q\colon \mathbb{R} \to S^1$, $q(t) = \exp(2\pi i t)$, after fixing its starting point at zero. A based loop then ends in the lift at an integer. Chapter 54 computes this integer as the winding number. A homotopy through based loops is exactly a path in $\Omega_1 S^1$, so

$$\pi_0(* \times^{\mathrm{h}}_{S^1} *) = \pi_0(\Omega_1 S^1) \cong \mathbb{Z}.$$

The ordinary pullback of the two chosen inclusions has a single point. At the same time, replacing one copy of $*$ over S^1 by $q\colon \mathbb{R} \to S^1$ gives an ordinary pullback with one point for every integer. Indeed $* \to \mathbb{R}$, picking zero, is a weak equivalence over S^1, and the covering map induces a Kan fibration on singular complexes by the homotopy lifting property. Taking singular complexes preserves ordinary pullbacks. The one-leg replacement criterion therefore identifies $\Omega_1 S^1$ with the discrete space on $\mathbb{Z}$ up to equivalence. This also shows that taking components of spaces need not preserve pullbacks. The three terms of the original cospan are all connected.

For base change, take maps $E \to B$, $B' \to B$, and a point $b'\colon * \to B'$. Write b for the image of b' and $E' = E \times^{\mathrm{h}}_B B'$. The two adjacent squares are homotopy pullback squares.

$$\begin{array}{ccccc} F_{b'} & \longrightarrow & E' & \longrightarrow & E \\ \downarrow & & \downarrow & & \downarrow{\scriptstyle p} \\ * & \xrightarrow{b'} & B' & \longrightarrow & B. \end{array}$$

Their composite exhibits $F_{b'} \simeq F_b$. One can verify this pasting rule by replacing p by a Kan fibration over B and using Kan representatives for the other spaces. Its pullback to B' is again a Kan fibration, so both pullbacks along these fibrations

can be computed strictly. The two-step and one-step ordinary pullbacks are isomorphic by the ordinary universal property. This form of base change will recur when interpreting families over a context and when comparing moduli objects after changing the base.

57.4 Pushouts from cylinders on the overlap

For a span $B \xleftarrow{i} A \xrightarrow{j} C$, a coherent cocone with target T specifies maps $u\colon B \to T$, $v\colon C \to T$, and a path from ui to vj in the mapping space from A to T. The comparison with two paths from a common source uses the corresponding horn Λ^2_0. The path is a map $A \times \Delta^1 \to T$ with those end values. This suggests inserting a cylinder on A between B and C. In the simplicial construction, let D be the ordinary pushout

$$\begin{array}{ccc} A \times \partial\Delta^1 & \xrightarrow{i \amalg j} & B \amalg C \\ \downarrow & & \downarrow \\ A \times \Delta^1 & \longrightarrow & D. \end{array}$$

The endpoint copies of A are attached using i and j, respectively. This is the *double mapping cylinder*. For any Kan complex T, mapping into T gives a strict pullback with n-simplices described by the same cocone data parameterized by Δ^n. In formulas,

$$\begin{aligned} T^D &\cong (T^B \times T^C) \times_{T^A \times T^A} (T^A)^{\Delta^1} \\ &\simeq T^B \times^{\mathrm{h}}_{T^A} T^C. \end{aligned}$$

Here the two maps to T^A are restrictions. The pullback formula proved above gives the second line. This is precisely the space of cocones in the universal test. Thus D presents $B \amalg^{\mathrm{h}}_A C$. To present D by a Kan complex one takes a Kan replacement. Mapping out into a Kan target already computes its mapping space, as every simplicial set is cofibrant.

If i is a monomorphism of simplicial sets, then for every Kan T the restriction $T^B \to T^A$ is a Kan fibration. The ordinary pushout $B \amalg_A C$ then has the same mapping property, since its maps to T are the strict pullback of the two restrictions, which computes the homotopy pullback. The same holds if j is a monomorphism. This criterion for using a strict pushout is dual in purpose to the fibration criterion for pullbacks [50, §3.4.2].

57.5 Suspension and the loop adjunction

In a span $* \leftarrow A \rightarrow *$, both endpoint copies of A collapse, each to its own point. The double mapping cylinder presents the *unreduced suspension* of A. For

nonempty A, a concrete topological model is the quotient of $A \times I$, $I = [0,1]$, with a separate collapse at each end. The empty case gives the two endpoints alone. When $A = S^0$ consists of two points, the quotient has two arcs with the same pair of endpoints, so is a circle. When $A = S^1$, the quotient is a 2-sphere. In the latter case the formula

$$(z,t) \longmapsto \big(\sin(\pi t)\,\mathrm{Re}\,z,\ \sin(\pi t)\,\mathrm{Im}\,z,\ \cos(\pi t)\big)$$

collapses exactly the two end circles and gives a continuous bijection from a compact space to the Hausdorff sphere. These models are also suspensions of simplicial presentations by realization, which preserves colimits and the product with Δ^1 in compactly generated spaces. Hatcher describes the geometric suspension construction in Chapter 0 [46].

In pointed spaces $\mathcal{S}_*$, the pushout of the same shape $* \leftarrow X \rightarrow *$ uses pointed maps from a pointed X, so the path along the basepoint of X is kept constant in any pointed cocone. The corresponding *reduced suspension* is denoted ΣX. For a based CW complex with a basepoint x_0 that is a 0-cell, it is presented by the quotient

$$\Sigma X = (X \times I)/(X \times \partial I \ \cup\ \{x_0\} \times I).$$

A pointed map from this quotient to Y is a map $X \times I \to Y$ constant at the basepoint of Y on the collapsed subspace. With the compactly generated function-space topology, the exponential adjunction takes it to a pointed map $X \to \Omega Y$ and gives the same correspondence on families. Equivalently, in $\mathcal{S}_*$ the pushout universal property turns maps from ΣX into loops at the constant map in $\mathrm{Map}_*(X,Y)$. Representable mapping-space functors preserve limits, so the loop pullback for Y turns maps to ΩY into that same loop space. Thus

$$\mathrm{Map}_*(\Sigma X, Y) \simeq \mathrm{Map}_*(X, \Omega Y), \qquad \Sigma \dashv \Omega.$$

Taking $X = S^0$ with one point chosen gives $\Sigma X = S^1$. The adjunction sends the identity of the circle to the pointed map $S^0 \to \Omega S^1$ whose second point is the loop of winding number one, using the increasing orientation of $I/\partial I$. The next chapter uses this adjunction to join iterated suspension and looping into a single category.

CHAPTER 58

Spectra from stabilization

Suspending a pointed space raises its dimension and carries the sphere S^n to S^{n+1}. A map of spheres can therefore be suspended repeatedly. Some maps become nullhomotopic after suspension, while others remain nonzero at every later stage. We construct the category in which suspension can be reversed, compute the first stable groups of the sphere, and use its squares to recover the exact sequence of a mapping cone.

58.1 Suspension and loops

Write $\mathcal{S}_*$ for the ∞-category of pointed spaces. An object has a map $* \to X$, and maps preserve the pointing with the specified homotopy. The one-point space $*$ is both initial and terminal. Let $\mathrm{Map}_*(X, Y)$ be the space of pointed maps. It has a basepoint given by the constant map. Section 57.5 obtained $\Sigma \dashv \Omega$ from two universal properties in pointed spaces. The pushout on the left defines the reduced suspension and the pullback on the right defines the based loop space.

$$\begin{array}{ccc} X & \longrightarrow & * \\ \downarrow & & \downarrow \\ * & \longrightarrow & \Sigma X \end{array} \qquad \begin{array}{ccc} \Omega Y & \longrightarrow & * \\ \downarrow & & \downarrow \\ * & \longrightarrow & Y \end{array}$$

The unit of this adjunction need not be an equivalence. Give S^0 one of its two points as basepoint. Then $\Sigma S^0 \simeq S^1$, since the interval over the other point has its endpoints identified. The unit $S^0 \to \Omega S^1$ sends the basepoint to the constant loop and the other point to a loop of winding number one after a choice of orientation. Chapter 54 gives $\pi_0(\Omega S^1) = \pi_1(S^1) \cong \mathbb{Z}$. Thus the unit hits two of infinitely many components.

58.2 Spectra as compatible deloopings

A *delooping* of X is a pointed space Y together with an equivalence $X \simeq \Omega Y$. A *spectrum* E consists of pointed spaces E_n, for $n \geq 0$, with structure equivalences

$s_n \colon E_n \xrightarrow{\sim} \Omega E_{n+1}$. A map has component maps f_n together with homotopies making the structure squares commute, compatibly at every stage.

One structure square is shown below.

$$\begin{array}{ccc} E_n & \xrightarrow{s_n} & \Omega E_{n+1} \\ \downarrow{\scriptstyle f_n} & & \downarrow{\scriptstyle \Omega f_{n+1}} \\ F_n & \xrightarrow{t_n} & \Omega F_{n+1}. \end{array}$$

The precise construction, including the higher homotopies of maps, is the limit of the tower in the ∞-category of ∞-categories.

$$\mathrm{Sp} = \lim(\mathcal{S}_* \xleftarrow{\Omega} \mathcal{S}_* \xleftarrow{\Omega} \mathcal{S}_* \xleftarrow{\Omega} \cdots).$$

The projection onto the first term is denoted $\Omega^\infty \colon \mathrm{Sp} \to \mathcal{S}_*$. The superscript ∞ names the operation on spectra, which takes E to E_0 [52, Proposition 1.4.2.24]. The limit also describes mapping spaces. For two spectra the transition map from level $n+1$ to level n sends f to $t_n^{-1}\Omega(f)s_n$. The limit of this tower of mapping spaces retains both the component maps and their compatibility data.

The zero spectrum has every level a point. Limits of spectra are computed at each level. In fact Ω is a right adjoint, so applying it to a limit of the spaces at level $n+1$ gives the limit at level n. The mapping-space test at each level then gives the test for the whole tower. In particular the loop functor Ω_{Sp} computed using the zero spectrum replaces each E_n by ΩE_n. Let T be the shift taking the tail of a spectrum, $(TE)_n = E_{n+1}$. The structure equivalences give natural equivalences

$$E \simeq \Omega_{\mathrm{Sp}} TE, \qquad T\Omega_{\mathrm{Sp}} E \simeq E.$$

In the second comparison, the levelwise map is the inverse to s_n. These equivalences show that taking loops has become invertible.

58.3 From a space to its suspension spectrum

A *prespectrum* P has maps $\sigma_n \colon \Sigma P_n \to P_{n+1}$, or equivalently adjoints $u_n \colon P_n \to \Omega P_{n+1}$, with no equivalence condition. The diagrams of a map of prespectra again commute coherently. Colimits are computed at each level using the suspension description, since Σ preserves colimits. There is a left adjoint L to the inclusion of spectra into prespectra. It is called *spectrification* and is computed by

$$(LP)_n \simeq \mathrm{colim}(P_n \longrightarrow \Omega P_{n+1} \longrightarrow \Omega^2 P_{n+2} \longrightarrow \cdots),$$

where the transition out of $\Omega^j P_{n+j}$ is $\Omega^j u_{n+j}$.

To see why the formula works, write QP for the prespectrum with $(QP)_n = \Omega P_{n+1}$ and structure maps given by looping those at the next level. The maps

u_n assemble into $u\colon P \to QP$, with Qu also giving the structure maps of QP. Looping commutes with sequential colimits because filtered colimits in spaces commute with finite limits and the loop space is a pullback of two basepoints [50, Section 9.1.5]. Consequently the structure map $LP \to QLP$ comes from comparing the chain $P \to QP \to Q^2P \to \cdots$ with its tail. The two chains have the same colimit, since every stage maps to a stage of the tail. Thus LP is a spectrum.

If E is a spectrum, a map $f\colon P \to E$ extends to $QP \to E$ by applying Q and using the inverse to $E \xrightarrow{\sim} QE$. Naturality of u proves this extension is inverse on mapping spaces to restriction along $P \to QP$. Applying the same argument at each stage and using the colimit test gives

$$\mathrm{Map}_{\mathrm{Sp}}(LP, E) \simeq \mathrm{Map}_{\mathrm{PSp}}(P, E),$$

where PSp denotes the ∞-category of prespectra [53, Proposition 4.3.19]. In particular, a colimit in Sp is obtained by taking the levelwise colimit of the underlying prespectra and then applying L.

For a pointed space X, the prespectrum $(X, \Sigma X, \Sigma^2 X, \ldots)$ has the identity as each suspension structure map. A map out of it to any prespectrum E is freely generated by $X \to E_0$. At the next level the map must be the composite $\Sigma X \to \Sigma E_0 \to E_1$, up to the given comparison. For a fixed composite, the choices of a map together with such a comparison form a contractible space. The same argument at each level, including the comparisons on maps, gives the mapping space of the free prespectrum. Applying L defines the *suspension spectrum* $\Sigma^\infty X$. The two mapping tests yield

$$\mathrm{Map}_{\mathrm{Sp}}(\Sigma^\infty X, E) \simeq \mathrm{Map}_*(X, \Omega^\infty E).$$

Thus $\Sigma^\infty \dashv \Omega^\infty$ [53, Lemma 4.3.22 and Remark 4.3.27].

58.4 Why exactness takes the form of a square

An ∞-category is *stable* if it is pointed, has finite limits and colimits, and its pushout squares are exactly its pullback squares. The following criterion applies here. In a pointed ∞-category with finite limits and colimits, full faithfulness of suspension makes every pushout a pullback. Dually, full faithfulness of looping makes every pullback a pushout. In particular, invertibility of the loop functor suffices for stability [52, Proposition 1.4.2.11]. The limit and colimit constructions above supply the hypotheses for Sp. The shift T is therefore the intrinsic suspension Σ_{Sp}, the left adjoint to Ω_{Sp}. The functor Σ^∞ preserves the pushout that defines ΣX and sends the initial point to the zero spectrum. Thus $\Sigma^\infty \Sigma X \simeq \Sigma_{\mathrm{Sp}} \Sigma^\infty X$.

Given a map of spectra $f\colon A \to B$, form $C = B \sqcup_A 0$, its *cofiber*. The left square below is a pushout, hence also a pullback. It follows that A is the fiber of

$B \to C$. The square on the right computes the cofiber of $B \to C$.

$$\begin{array}{ccccc} A & \xrightarrow{f} & B & \longrightarrow & 0 \\ \downarrow & & \downarrow & & \downarrow \\ 0 & \longrightarrow & C & \xrightarrow{\partial} & \Sigma_{\mathrm{Sp}} A \end{array}$$

Indeed $C \sqcup_B 0 = (B \sqcup_A 0) \sqcup_B 0 \simeq 0 \sqcup_A 0$ by the pushout test. Thus the cofiber sequence continues to $A \to B \to C \to \Sigma_{\mathrm{Sp}} A$. The last two maps determine another fiber sequence by stability.

These constructions also give a universal property of stabilization. Let $\mathcal{D}$ be a stable ∞-category and let $F \colon \mathcal{D} \to \mathcal{S}_*$ preserve finite limits. It lifts to a finite-limit-preserving functor $\widetilde{F} \colon \mathcal{D} \to \mathrm{Sp}$ by setting $\widetilde{F}(D)_n = F(\Sigma_{\mathcal{D}}^n D)$. Indeed F preserves loop objects, so $\Omega F(\Sigma_{\mathcal{D}}^{n+1} D) \simeq F(\Omega_{\mathcal{D}} \Sigma_{\mathcal{D}}^{n+1} D) \simeq F(\Sigma_{\mathcal{D}}^n D)$. Any lift preserving finite limits also preserves loop objects and therefore their inverses. Its level n is then forced by evaluation at level zero on $\Sigma_{\mathcal{D}}^n D$. This argument works on natural transformations as well, by the limit description of spectra.

58.5 The stable homotopy groups of a sphere

For any integer k and any spectrum E, set $\pi_k(E) = \pi_{k+n}(E_n)$ for any n with $k + n \geq 2$. The structure equivalences identify the choices by $\pi_{k+n}(E_n) \cong \pi_{k+n+1}(E_{n+1})$, and the bound makes all the groups abelian. The shift gives $\pi_k(\Sigma_{\mathrm{Sp}} E) \cong \pi_{k-1}(E)$. For a cofiber sequence $A \to B \to C$ the square is a pullback, so for fixed n the spaces $A_n \to B_n \to C_n$ form a homotopy fiber sequence. Its long exact sequence in homotopy gives the displayed terms on choosing n large enough to read them all as stable groups.

$$\pi_k(A) \longrightarrow \pi_k(B) \longrightarrow \pi_k(C) \longrightarrow \pi_{k-1}(A) \longrightarrow \pi_{k-1}(B).$$

The connecting map comes from the fiber inclusion $\Omega_{\mathrm{Sp}} C \to A$. To express it using the displayed cofiber squares, pull back $A \simeq B \times_C 0$ along $0 \to B$, giving $0 \times_C 0$. The outer square identifies A with $0 \times_{\Sigma_{\mathrm{Sp}} A} 0$. Using the same order of the two zero objects in both pullbacks, the fiber inclusion is induced by $\partial \colon C \to \Sigma_{\mathrm{Sp}} A$. Since n can increase without bound, the stable sequence continues in both directions.

The *sphere spectrum* is $\mathbb{S} = \Sigma^\infty S^0$. Since $\Sigma^j S^0 \simeq S^j$, its groups are

$$\pi_k(\mathbb{S}) \cong \operatorname*{colim}_{j,\, j+k \geq 2} \pi_{j+k}(S^j),$$

with suspension as the transition map. To obtain the formula, apply the levelwise spectrification formula and test with a sphere at a sufficiently high level. Homotopy groups in each fixed degree commute with filtered colimits of pointed spaces.

For a Kan model, the simplices used to specify a map from a finite triangulated sphere and a homotopy of such maps involve only finitely many stages. This also identifies the transition maps with suspension by adjunction.

Example 58.1 (The first stable sphere groups)**.** If $k < 0$, then $j + k < j$ for all sufficiently large j. The sphere S^j is $(j-1)$-connected, so these groups are zero. For $k = 0$, $\pi_j(S^j) \cong \mathbb{Z}$ for $j \geq 2$, measured by degree by the first Hurewicz theorem. With compatible orientations, suspension sends the degree-one identity to the degree-one identity in the next dimension, so the colimit is $\mathbb{Z}$ [46, Theorem 4.32].

For $k = 1$, the Freudenthal suspension theorem states that $\pi_i(S^j) \to \pi_{i+1}(S^{j+1})$ is an isomorphism for $i < 2j - 1$ and a surjection for $i = 2j - 1$ when $j \geq 2$. For $i = j + 1$ the isomorphism range begins at $j = 3$. At $j = 2$ the suspension map is the surjection $\pi_3(S^2) \to \pi_4(S^3)$. The Hopf map $h\colon S^3 \to S^2$ generates $\pi_3(S^2)$. The classical computation identifies the kernel of suspension as the subgroup generated by $2[h]$. Hence $\pi_4(S^3) \cong \mathbb{Z}/2$, generated by Σh, and the identifications in the stable range preserve the generator [46, Corollaries 4.24 and 4J.4]. The result is

$$\pi_k(\mathbb{S}) = \begin{cases} 0 & k < 0, \\ \mathbb{Z} & k = 0, \\ \mathbb{Z}/2 & k = 1. \end{cases}$$

The stable class of h survives even though each suspended map $S^{j+1} \to S^j$ induces zero on reduced integral homology. Its source and target have their only nonzero reduced homology groups in distinct degrees.

Example 58.2 (A cofiber of degree two)**.** Stabilize a degree-two map $S^j \to S^j$ for $j \geq 2$ and desuspend j times. This gives an endomorphism of $\mathbb{S}$ acting by multiplication by two on $\pi_0(\mathbb{S})$, since degrees multiply under composition. Write M_2 for its cofiber, a *Moore spectrum*. The long exact sequence has the portion

$$\mathbb{Z} \xrightarrow{2} \mathbb{Z} \longrightarrow \pi_0(M_2) \longrightarrow 0.$$

The last term is $\pi_{-1}(\mathbb{S})$. Exactness makes $\pi_0(M_2) \cong \mathbb{Z}/2$, so a quotient by a relation of degree two has appeared as a cofiber computation.

58.6 Mapping cones as stable cofibers

The exactness of spectra has a parallel in Chapter 52. Fix a commutative unital ring R. The homotopy category of bounded-above projective cochain complexes is equivalent to $\mathbf{D}^-(R)$, by projective replacement. There is a stable ∞-category enhancing this model, formed by keeping the differential graded complexes of

maps between projective complexes. Its mapping spaces use the good nonpositive truncation of the cochain complex $\mathcal{H}(P,Q)$ of Chapter 52. Its term in degree zero is the group of cycles, that is, the cochain maps. The degree -1 differential records the differences that are chain homotopies, so π_0 recovers chain maps modulo homotopy. The higher homotopy groups of the mapping space at the zero map are $H^{-i}(\mathcal{H}(P,Q))$ in degree $i>0$. This is the differential graded nerve construction described in [52, Sections 1.3.1–1.3.2].

The suspension in this model is the cochain shift $[1]$, whose inverse is $[-1]$, and the cofiber of a cochain map is the mapping cone [52, Remark 1.3.2.17 and Corollary 1.3.2.18]. For the cone $C(f)$ the relevant diagram is

$$\begin{array}{ccccc} P & \xrightarrow{f} & Q & \longrightarrow & 0 \\ \downarrow & & \downarrow & & \downarrow \\ 0 & \longrightarrow & C(f) & \longrightarrow & P[1]. \end{array}$$

The cone differential supplies the homotopy in the left square. The map $C(f)\to P[1]$ is the projection onto the shifted summand. Mapping out of R in degree zero tests these squares by homotopy pullbacks of spaces. The cochain complex of maps from R to P is P itself by evaluation at 1. Thus the homotopy groups of the mapping space into $P[n]$ give $H^{n-k}(P)$ in homotopy degree $k>0$. Varying n reaches every cohomological degree. Under this identification the boundary from $H^i(C(f))$ lands in $H^i(P[1])=H^{i+1}(P)$, recovering the shift in the long exact sequence of the cone.

58.7 Pontryagin duality for finite homotopy groups

Write $[X,Y]$ for the group of homotopy classes of maps of spectra. The functor $X\mapsto \mathrm{Hom}_{\mathbb{Z}}(\pi_0 X,\mathbb{Q}/\mathbb{Z})$ extends to a cohomology theory. Applying $\mathrm{Hom}_{\mathbb{Z}}(-,\mathbb{Q}/\mathbb{Z})$ to the homotopy exact sequence preserves exactness by injectivity, and the theory sends wedges to products since π_0 preserves wedges. Brown representability gives the *Brown–Comenetz spectrum* I, characterized by $[X,I]\cong \mathrm{Hom}_{\mathbb{Z}}(\pi_0 X,\mathbb{Q}/\mathbb{Z})$ naturally in X [66, Section 2.8]. Define the Brown–Comenetz dual $D_{\mathrm{BC}}X=F(X,I)$, where the function spectrum $F(X,I)$ represents maps out of the smash product with X. The smash–function adjunction therefore computes its homotopy groups:

$$\pi_k D_{\mathrm{BC}}X\cong [\Sigma^k X,I]\cong \mathrm{Hom}_{\mathbb{Z}}(\pi_{-k}X,\mathbb{Q}/\mathbb{Z}).$$

If every $\pi_k X$ is finite, these are its Pontryagin duals in reversed degrees. The map $X\to D_{\mathrm{BC}}D_{\mathrm{BC}}X$ obtained by currying evaluation induces character biduality on homotopy groups, with the corresponding suspension identifications. It is an

equivalence since each finite group is character reflexive and homotopy groups detect equivalences of spectra.

Example 58.3 (A single homotopy group in a shifted degree)**.** For finite abelian A, let HA denote its Eilenberg–Mac Lane spectrum. The only nonzero homotopy group of $\Sigma^r HA$ is A in degree r. The formula gives

$$D_{\mathrm{BC}}(\Sigma^r HA) \simeq \Sigma^{-r} H\widehat{A}.$$

This recovers the degree reversal for complexes in Section 52.7 at a spectrum with one homotopy group.

Part XII Higher toposes, homotopy type theory, and derived geometry

CHAPTER 59

Higher toposes from sheaves of spaces

A section of a sheaf of sets gives a family of local sections whose restrictions agree. A section of a sheaf of spaces can also specify paths witnessing those agreements. The paths themselves must agree on triple overlaps, with higher comparisons on further overlaps. A single mapping-space test records all these dimensions of descent. We derive the test from covering sieves and use it to compute a sheaf whose global sections have nontrivial homotopy.

59.1 Testing descent by a covering sieve

Let $\mathcal{C}$ be a small ordinary category equipped with a Grothendieck topology as in Chapter 45. Fix a universe of small spaces and write $\mathcal{P}(\mathcal{C}) = \mathrm{Fun}(\mathcal{C}^{\mathrm{op}}, \mathcal{S})$. We view an ordinary set as a discrete space, so the representable $yU = \mathrm{Hom}_{\mathcal{C}}(-, U)$ and every sieve $R \hookrightarrow yU$ are objects of $\mathcal{P}(\mathcal{C})$. The higher Yoneda lemma gives $\mathrm{Map}_{\mathcal{P}(\mathcal{C})}(yU, F) \simeq F(U)$. The two restrictions from the source in the following square therefore agree under this equivalence.

$$\begin{array}{ccc} \mathrm{Map}_{\mathcal{P}(\mathcal{C})}(yU, F) & \xrightarrow{\sim} & F(U) \\ \downarrow & & \downarrow \\ \mathrm{Map}_{\mathcal{P}(\mathcal{C})}(R, F) & \xrightarrow{\sim} & \lim_{(V \to U) \in R^{\mathrm{op}}} F(V). \end{array}$$

Here the index category for the bottom limit is the opposite of the full subcategory of $\mathcal{C}/U$ on the arrows in R. To see the bottom equivalence, take the colimit of the representables yV over these arrows. Evaluate at T. For a composite $f\colon T \to V \to U$ in $R(T)$, the category of its factorizations through arrows in R has an initial object given by $T \xrightarrow{\mathrm{id}} T \xrightarrow{f} U$. The space of factorizations in the colimit is thus contractible. The resulting colimit is R. Mapping this colimit to F gives the bottom limit by the universal property of colimits from Chapter 57.

We call F a *sheaf of spaces* if the vertical maps are equivalences for every covering sieve [51, Definition 6.2.2.6]. Write $\mathrm{Shv}_{\mathcal{S}}(\mathcal{C})$ for the full ∞-subcategory of these presheaves. The right side of the test is a space of coherent families. For example, compatibility of two restrictions can be supplied by a path. A map

from a test space T into that limit specifies a family of such compatible data parametrized by T, including all its homotopies.

59.2 Why a cover produces a Čech diagram

Suppose $\{U_i \to U\}_{i \in I}$ generates a covering sieve and its iterated fiber products exist in $\mathcal{C}$. Form $E = \coprod_i yU_i$ and its map to yU in presheaves. The *Čech nerve* of this map has level

$$E_n = \underbrace{E \times_{yU} \cdots \times_{yU} E}_{n+1} \simeq \coprod_{i_0,\dots,i_n} y(U_{i_0} \times_U \cdots \times_U U_{i_n}).$$

Its face maps forget a factor and its degeneracies repeat one. The formula follows on evaluation at each T, using the representable test for a fiber product and the distributivity of fiber products over disjoint unions of sets. The coproduct is in presheaves, so no coproduct in the site is required. The first two levels, together with the augmentation, have the form

$$E_1 \underset{d_1}{\overset{d_0}{\rightrightarrows}} E_0 \longrightarrow yU.$$

The geometric realization $|E_\bullet|$ here means the colimit of the entire simplicial diagram in presheaves of spaces. It is equivalent to R. To check this, evaluate at an object T and fix an arrow $f \in yU(T)$. In degree n the simplices over f are $(n+1)$-tuples of tagged lifts of f to one of the U_i. If there are no lifts there are no simplices. Otherwise this is the nerve of the groupoid on the set of lifts with a unique arrow between every ordered pair. The groupoid has an initial object and its nerve is contractible. The realization therefore has one contractible component for each f in the sieve. The realization map to R gives a pointwise equivalence and hence an equivalence of presheaves.

By mapping into F we obtain the descent test

$$F(U) \longrightarrow \lim_{[n] \in \mathbf{\Delta}} \prod_{i_0,\dots,i_n} F(U_{i_0} \times_U \cdots \times_U U_{i_n}).$$

The limit of this cosimplicial diagram is often called its *totalization*. Each of its points specifies local sections in degree zero and compatibility paths in degree one. The degree-two simplex compares the direct path on the first and last index with the composite across the middle index. Degeneracies account for identity comparisons and the higher levels supply further coherence. For the open subsets of a topological space the iterated fiber products in the formula are intersections. When the topology is specified by covering families generating its covering sieves, these tests are equivalent to the sieve condition.

59.3 Gluing a space of sections on a finite site

Consider four points a, b, u, v, ordered by $a < u,\ a < v,\ b < u,\ b < v$, with no other strict comparisons. Take the downward-closed subsets as opens. Write T for the whole space, $U = \{a, b, u\}$, $V = \{a, b, v\}$, and $W = U \cap V$. The opens are exactly

$$\varnothing, \quad A = \{a\}, \quad B = \{b\}, \quad W = \{a, b\}, \quad U, \quad V, \quad T.$$

The undirected graph of the nonidentity arrows is a cycle, with two maximal elements u, v in the order.

$$\begin{array}{ccc} a & \longrightarrow & u \\ \downarrow & & \uparrow \\ v & \longleftarrow & b. \end{array}$$

Let K_D be the space presented by the nerve of an open D with its induced order. Fix a space K and set $F_K(D) = \mathrm{Map}_{\mathcal{S}}(K_D, K)$ with restrictions by precomposition.

To verify the sheaf condition, there are two nonmaximal nonempty covering sieves to check, on W and T. The maximal sieve on any open gives the identity test and the empty covering sieve on $\varnothing$ requires a terminal value, as holds for $\mathrm{Map}(\varnothing, K)$. The other sieves are the ones generated by A, B on W and by U, V on T. Indeed a cover of U must include an open containing u, hence U itself, and the parallel statement holds for V. A cover of T omitting T must therefore include both U and V. For a pair of opens D_1, D_2, the sieve they generate is the homotopy pushout of $yD_1 \leftarrow y(D_1 \cap D_2) \to yD_2$ in presheaves. Evaluation at an open is either empty or a point at each term, and the pushout is a point exactly when the open is contained in D_1 or D_2. Thus the sieve test for each of these pair covers is a homotopy pullback test.

The nerve of W is the disjoint union of the two one-point nerves of A, B. Every chain in T lies in one of U, V, and their nerves intersect in the nerve of W. The nerve of T is therefore the simplicial pushout of these two subsets. Both inclusions of the nerve of W are monomorphisms, so this presents $K_U \amalg^{\mathrm{h}}_{K_W} K_V$ by the criterion in Section 57.4. Mapping both of these pushouts into K gives the needed homotopy pullbacks. This proves that F_K is a sheaf.

The nerves K_U, K_V are contractible, since u, v are terminal in their respective posets. Under the equivalences of their mapping spaces with K, the two restrictions to K_W are both equivalent to the diagonal of K. The comparison can be made by sending a point of K to the constant map on each contractible nerve.

$$\begin{array}{ccc} F_K(T) & \longrightarrow & K \\ \downarrow & & \downarrow{\scriptstyle \Delta} \\ K & \xrightarrow{\Delta} & K \times K. \end{array}$$

The square expresses equivalence with the space of maps from the homotopy pushout $* \amalg^{\mathrm{h}}_{S^0} *$ into K. That pushout is the circle, as in Section 57.5. Here it can also be seen directly in the nerve of T, which has four vertices and four nondegenerate edges arranged in a cycle.

$$F_K(T) \simeq \mathrm{Map}_{\mathcal{S}}(S^1, K).$$

When $K = S^1$ the maps $f \mapsto (f(1), f(1)^{-1} f)$ and $(z, g) \mapsto zg$ separate evaluation from a based loop. These formulas act on families and are inverse, so $\mathrm{Map}(S^1, S^1) \simeq S^1 \times \Omega S^1$. The computation in Section 57.3 identifies ΩS^1 with the discrete space of winding numbers $\mathbb{Z}$. Each of $F_K(U), F_K(V), F_K(W)$ is connected, so each of the two opens has a unique homotopy class of sections. The ways of identifying their restrictions account for the global winding numbers. Taking $K = S^2$ gives an example beyond 1-types. Evaluation at a point of the circle retracts the inclusion of constant maps, so $\pi_2(F_{S^2}(T))$ at the constant map contains a copy of $\pi_2(S^2) \cong \mathbb{Z}$, as computed in Chapter 56.

59.4 Sheafification and the higher topos axioms

An ∞-*topos* is an accessible left exact localization of a presheaf category of spaces on a small ∞-category $\mathcal{C}$ [51, Definition 6.1.0.4]. A localization here means an adjunction $L \dashv i$ with i fully faithful. Left exact means that L preserves finite limits, and accessibility of the localization means that the associated endofunctor iL preserves κ-filtered colimits for some regular cardinal κ.

$$\mathcal{P}(\mathcal{C}) \underset{i}{\overset{L}{\rightleftarrows}} \mathcal{X}$$

The ∞-category of spaces is $\mathcal{P}(*)$, where $*$ denotes the category with one object and only its identity. The sheafification theorem says that for a small site the inclusion of sheaves of spaces has an accessible left exact left adjoint a, and that this localization is generated by the inclusions of covering sieves [51, Proposition 6.2.2.7]. This recovers the form of the sheafification adjunction in Chapter 45. Limits of sheaves are computed in presheaves, since a limit of objects local for the sieve inclusions passes every test. For colimits, first take the presheaf colimit and then apply a, by the universal property of the adjunction.

Among the intrinsic properties of an ∞-topos are universal colimits and effective groupoid objects [51, Theorem 6.1.0.6]. Universal colimits are colimits preserved by base change in slices. In the presheaf presentation this follows from pointwise colimits in spaces, where base change preserves colimits, and from the left exact localization. For a diagram of sheaves over Y with presheaf colimit $Z \to iY$ and a sheaf map $X \to Y$, left exactness gives

$$a(iX \times_{iY} Z) \simeq X \times_Y aZ$$

Computing the presheaf colimit inside the left side after pulling back proves the base change assertion for the sheaf colimit.

For a map $f\colon X \to Y$, the functor f^* pulls a family over Y back to one over X. Its left adjoint Σ_f composes maps with f, by the universal property of the pullback. Since f^* preserves colimits and the slices are presentable, the adjoint functor theorem supplies a right adjoint Π_f [51, Corollary 5.5.2.9 and Proposition 5.5.3.10].

$$\mathcal{X}/X \underset{f^*}{\overset{\Sigma_f}{\rightleftarrows}} \mathcal{X}/Y, \qquad \Sigma_f \dashv f^* \dashv \Pi_f.$$

These are the dependent sum, substitution, and dependent product from Chapter 27. On subobjects, Π_f gives universal quantification, since right adjoints preserve the limits that characterize monomorphisms. The (-1)-truncation of Σ_f in the slice gives existential quantification. It takes the image of the composite, and every map over Y from the composite to a subobject of Y factors through that image with a contractible space of choices. This gives the adjoints on predicates from Chapter 24.

An internal groupoid is given by a simplicial object whose levels encode strings of composable arrows via pullbacks, with the groupoid identities expressing unital composition and invertibility in the ∞-categorical sense [51, §6.1.2]. For such a groupoid $G_\bullet$, effectiveness says that the natural map to the Čech nerve of $G_0 \to |G_\bullet|$ is an equivalence. In particular $G_1 \simeq G_0 \times_{|G_\bullet|} G_0$. For a discrete group G acting on a set X, regard its action groupoid as a simplicial object of spaces. Writing $X//G$ for its realization, effectiveness gives

$$\begin{array}{ccc} X \times G & \xrightarrow{\;t\;} & X \\ \downarrow{\scriptstyle s} & & \downarrow \\ X & \longrightarrow & X//G, \end{array}$$

as a homotopy pullback square, with source and target of an action arrow as s, t. Thus the homotopy fiber over (x, x) of the two-endpoint map is the stabilizer of x as a discrete space. Equivalently, that stabilizer is the loop space of $X//G$ at x. These loops give a categorical reason to retain automorphisms in a moduli problem.

59.5 What truncation recovers

An object is n-truncated if every mapping space into it has vanishing homotopy groups in degrees above n, for $n \geq 0$ and at every basepoint. A sheaf F is n-truncated precisely if every value $F(U)$ is. In one direction use $\mathrm{Map}(a(yU), F) \simeq F(U)$ by Yoneda and sheafification. In the other direction, write a presheaf as a

colimit of representables by the higher Yoneda density theorem [51, §5.1.5]. The mapping space from it into F is a limit of n-types, hence an n-type. This applies in particular when the source is a sheaf.

For $n = 0$, the values are equivalent to discrete sets. Limits of discrete spaces agree with the corresponding limits of sets, so the sieve test agrees with ordinary set-valued descent. Hence the full subcategory of 0-truncated sheaves is the ordinary category of set-valued sheaves on the site, regarded as an ∞-category with discrete mapping spaces. Internally truncating a sheaf F in degree zero requires sheafifying the presheaf $U \mapsto \pi_0(F(U))$. Indeed pointwise truncation is left adjoint to including presheaves of sets, and sheafification is left adjoint to including sheaves. Their composite has the claimed test against set-valued sheaves. Sheafification preserves 0-truncated objects since the condition can be tested by the diagonal and its diagonal, both finite-limit constructions [51, §5.5.6].

In Section 59.3, the restriction from $\pi_0(F_{S^1}(T))$ to the set of compatible local components sends all winding numbers to one local family. The presheaf of components therefore fails the sheaf test. The sheaf of components identifies those winding numbers because sections become equivalent locally. In degree 1, the truncation statement instead recovers sheaves valued in 1-types. Chapter 56 lets us study them using nerves of groupoids, where paths are isomorphisms and descent retains the isomorphisms between local objects.

59.6 Pontryagin duality as an internal character sheaf

Choose an uncountable strong limit cardinal κ. Take the site of profinite sets with fewer than κ points, with finite jointly surjective families as covers. Its set-valued sheaves are κ-condensed sets [67, Definition 1.4]. A locally compact abelian group G gives an abelian group sheaf $\underline{G}$ by the rule $\underline{G}(S) = \mathrm{Cont}(S, G)$ with pointwise addition. These maps glue because a surjection from a compact space onto a Hausdorff space is a quotient map. Thus among the 0-truncated sheaves of spaces on this site one can study continuous characters in families of parameters.

Put $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. Take κ large enough that G, $\mathbb{T}$, and their duals have fewer than κ points. Let $\underline{\mathrm{Hom}}(\underline{G}, \underline{\mathbb{T}})$ be the internal Hom of abelian group objects in the sheaf topos. A section over a profinite S is a map of group objects over the representable sheaf of S. The locally compact Hausdorff spaces $S \times G$ and $\mathbb{T}$ are κ-compactly generated. Each has fewer than κ points, and its topology is detected on its compact neighborhoods. Maps from the sheaf of $S \times G$ to the sheaf of $\mathbb{T}$ are the continuous maps between these spaces [67, Proposition 1.7]. The group-object condition says that for each $s \in S$ the map on G is a homomorphism. By the compact-open exponential law, these continuous maps $S \times G \to \mathbb{T}$ are exactly

the continuous maps $S \to \widehat{G}$. This identification commutes with restriction along maps of profinite sets, giving the following case of Scholze's internal Hom formula [67, Proposition 4.2].

$$\underline{\mathrm{Hom}}(\underline{G}, \underline{\mathbb{T}}) \cong \underline{\widehat{G}}.$$

The internal double evaluation on such sheaves therefore recovers the topological bidual isomorphism of Section 16.6.

For $G = \mathbb{Z}$ with the discrete topology, a family of characters is determined by the continuous function $t\colon S \to \mathbb{T}$ obtained by evaluating at 1. Conversely t gives the continuous character family $(s, n) \mapsto n\,t(s)$. The dual sheaf is $\underline{\mathbb{T}}$, so the internal Hom retains the topology of the character group through its sections over S.

59.7 Hyperdescent and hypercompletion

A morphism $E \to B$ in an ∞-topos $\mathcal{X}$ is *∞-connective* if every finite truncation of the object $E \to B$ in the slice $\mathcal{X}/B$ is equivalent to the terminal object $B \to B$. For spaces this means that every homotopy fiber is weakly contractible. By the equivalence test in Chapter 56, spaces are already hypercomplete in the sense that every ∞-connective map is an equivalence. The analogous Whitehead property is an additional condition on a general ∞-topos.

An object F is *hypercomplete* if $\mathrm{Map}(B, F) \to \mathrm{Map}(E, F)$ is an equivalence for every ∞-connective $E \to B$. The full subcategory $\mathcal{X}^{\wedge}$ of such objects is an accessible left exact localization of $\mathcal{X}$ [51, after Proposition 6.5.2.8]. Every finitely truncated object is hypercomplete. For an n-truncated F, the product $B \times F \to B$ is n-truncated in the slice. The maps over B into that product are precisely maps into F. The slice truncation adjunction gives the equivalence of mapping spaces because $\tau_{\le n}^{\mathcal{X}/B} E \simeq B$.

A *hypercover* of B is an augmented simplicial object $U_\bullet \to B$ such that each map to the matching object for the lower-dimensional boundary is an effective epimorphism. An effective epimorphism is a map whose Čech nerve realizes to its target. For a hypercover the first condition is therefore that $U_0 \to B$ is an effective epimorphism. The next asks that $U_1 \to U_0 \times_B U_0$ be an effective epimorphism. In higher degrees one requires a cover of the object of compatible faces, formed by a limit in the slice over B. For an ordinary Čech nerve of an effective epimorphism, the first map is the given cover and the higher matching maps are equivalences. A hypercover allows newly chosen covers for the successive matching objects.

The augmentation $|U_\bullet| \to B$ of a hypercover is ∞-connective [51, Lemma 6.5.3.11]. Conversely, an ∞-connective map $E \to B$ defines a hypercover constant at E in the slice, whose realization is E [51, Lemma 6.5.3.5 and Corol-

lary 6.5.3.13]. Consequently F is hypercomplete exactly when it satisfies the following test for all hypercovers in $\mathcal{X}$.

$$\mathrm{Map}(B, F) \xrightarrow{\sim} \lim_{[n]\in\mathbf{\Delta}} \mathrm{Map}(U_n, F).$$

This condition is *hyperdescent*. Passing to the hypercomplete subcategory imposes this condition in the sheaf ∞-topos in addition to the covering-sieve tests. On any small ordinary site the sheaves valued in 1-types already pass both tests. This is the level at which groupoid-valued descent describes stacks in the next chapter.

CHAPTER 60

Stacks and moduli from groupoid-valued descent

A bundle can have nontrivial automorphisms over a fixed base. Its isomorphism class forgets those arrows. It also forgets which isomorphism identifies one restriction with another on an overlap. Both kinds of information occur in the gluing proof in Chapter 46. Keeping the objects and their isomorphisms gives a groupoid over each test base. Requiring effective descent for these groupoids leads to stacks.

60.1 Descent data and their isomorphisms

Let $\mathcal{F}(U)$ be a groupoid of objects over each open subset U of a space X. For an inclusion $f\colon V \hookrightarrow U$, let $f^{*}\colon \mathcal{F}(U) \to \mathcal{F}(V)$ be restriction. In many examples, an iterated restriction is canonically isomorphic to restriction along the composite. A *pseudofunctor* specifies these natural isomorphisms, along with the comparison for identity arrows. For three successive inclusions, the two composites of comparison isomorphisms must agree. This is an instance of the coherence of Chapter 55. On opens we write $x|_{V}$ for $f^{*}x$ and include the comparison isomorphisms when identifying iterated restrictions.

For a cover $U = \bigcup_{i} U_{i}$, write $U_{ij} = U_{i} \cap U_{j}$ and $U_{ijk} = U_{i} \cap U_{j} \cap U_{k}$. Its descent groupoid $\mathrm{Desc}_{\mathcal{F}}((U_{i}))$ has objects (x_{i}, α_{ij}) with $x_{i} \in \mathcal{F}(U_{i})$ and

$$\alpha_{ij}\colon x_{j}|_{U_{ij}} \xrightarrow{\ \sim\ } x_{i}|_{U_{ij}}.$$

The identity condition is $\alpha_{ii} = \mathrm{id}$, and the following triangle commutes after restricting to U_{ijk}.

$$\begin{array}{ccc} x_{k}|_{U_{ijk}} & \xrightarrow{\ \alpha_{jk}\ } & x_{j}|_{U_{ijk}} \\ & \searrow^{\alpha_{ik}} & \downarrow{\scriptstyle \alpha_{ij}} \\ & & x_{i}|_{U_{ijk}}. \end{array}$$

The triangle says $\alpha_{ij}\alpha_{jk} = \alpha_{ik}$ and in particular gives $\alpha_{ji} = \alpha_{ij}^{-1}$.

An isomorphism from (x_i, α_{ij}) to (y_i, β_{ij}) is a family $a_i \colon x_i \to y_i$ commuting with the transitions.

$$\begin{array}{ccc} x_j|_{U_{ij}} & \xrightarrow{\alpha_{ij}} & x_i|_{U_{ij}} \\ \downarrow{\scriptstyle a_j} & & \downarrow{\scriptstyle a_i} \\ y_j|_{U_{ij}} & \xrightarrow{\beta_{ij}} & y_i|_{U_{ij}}. \end{array}$$

For fixed data the family is an element of an equalizer of sets. The two maps out of $\prod_i \operatorname{Isom}(x_i, y_i)$ take a to $(a_i \alpha_{ij})_{i,j}$ and $(\beta_{ij} a_j)_{i,j}$, respectively, in the product of isomorphism sets from $x_j|_{U_{ij}}$ to $y_i|_{U_{ij}}$.

Restriction gives a functor

$$\mathcal{F}(U) \longrightarrow \operatorname{Desc}_{\mathcal{F}}((U_i)),$$

using the canonical comparisons between the restrictions of a single object. It is fully faithful for every cover precisely when, for each pair $x, y \in \mathcal{F}(U)$, the presheaf of sets

$$V \longmapsto \operatorname{Isom}_{\mathcal{F}(V)}(x|_V, y|_V)$$

is a sheaf on the opens of U. In fact, for two global objects the displayed equalizer is the matching-family equalizer for this presheaf. To obtain the sheaf condition on every $V \subseteq U$, apply full faithfulness to the restrictions $x|_V, y|_V$ and a cover of V.

The functor is essentially surjective if every datum is *effective*, meaning that some global object restricts to an isomorphic datum. A *stack in groupoids* is a pseudofunctor for which the restriction functor is an equivalence for every cover. The definition extends to a site with the required fiber products by replacing intersections with iterated fiber products over U. For a general site it can be expressed using covering sieves. The formulation by categories fibered in groupoids gives the same conditions on morphisms and objects [54, Section 8.5].

60.2 A descent condition for 1-types

For a cover by two opens the descent groupoid has objects (x_U, x_V, α), where $\alpha \colon x_V|_{U \cap V} \to x_U|_{U \cap V}$ is an isomorphism. Its nerve is the homotopy pullback of the two restriction maps of nerves. Indeed the nerve of the groupoid of isomorphisms in $\mathcal{F}(U \cap V)$ is a path object. Evaluation at its two endpoints is a Kan fibration. Lifting endpoint isomorphisms a, b from the source and target of α amounts to choosing the new comparison $b \alpha a^{-1}$. The endpoint functor is thus an isofibration of groupoids, whose nerve is a Kan fibration. The constant-path inclusion is an equivalence, since every arrow is isomorphic to the identity on its source. A square between isomorphisms with one endpoint arrow on each side

is exactly an arrow between the triples. This gives the path description from Chapter 57, with the path oriented from the second restriction to the first.

$$\begin{array}{ccc} \mathrm{N}\,\mathrm{Desc}_{\mathcal{F}}((U,V)) & \longrightarrow & \mathrm{N}\mathcal{F}(V) \\ \downarrow & & \downarrow \\ \mathrm{N}\mathcal{F}(U) & \longrightarrow & \mathrm{N}\mathcal{F}(U \cap V). \end{array}$$

For a general open cover, apply the Čech test of Chapter 59 to the nerves of the groupoids. The degree-zero component of a coherent section chooses the local objects. In degree one it chooses the comparison arrows. Compatibility in degree two requires the triangle above. For higher degrees a simplex in a groupoid nerve is determined by its successive arrows. A triangle already fixes each longer edge as their composite. Thus the comparison arrows and the triangle law determine the remaining simplices. A path between two coherent sections gives the family of isomorphisms a_i with its compatibility square, and the same nerve description determines the higher simplices of the space of sections. It follows that the Čech limit is the nerve of the descent groupoid up to equivalence. In the other direction, a sheaf of spaces with 1-truncated values gives a stack in groupoids by taking fundamental groupoids. The nerve gives back those values up to equivalence by the 1-truncation calculation of Chapter 56.

60.3 The stack of torsors

Let G be a discrete group. Over a topological space U, a *right G-torsor* is a map $p\colon P \to U$ with a continuous right G-action over U, locally isomorphic to $U \times G$ with action by right multiplication on G. Thus G acts freely and transitively on each fiber. A map over U commuting with the actions is an isomorphism. On local product charts it has the form $(x, g) \mapsto (x, h(x)g)$ for a continuous h into G, so the inverse uses $h(x)^{-1}$. These torsors and maps form a groupoid $\mathrm{Tors}_G(U)$.

Proposition 60.1. *The pseudofunctor* Tors_G *is a stack for open covers.*

Proof. Take a descent datum (P_i, α_{ij}). Glue the total spaces by identifying $p \in P_j|_{U_{ij}}$ with $\alpha_{ij}(p)$. Each equivalence class over a point of U_i meets P_i in one representative. This follows from the identity and cocycle laws, which reduce any chain of identifications over that point to the single transition between its first and last indices. The maps $P_i \to P = (\coprod_i P_i)/\!\sim$ identify the charts with open subspaces. For an open set in P_i, its saturated preimage in P_j is the image of its restriction under α_{ji}, which is open. This proves the chart claim by the quotient topology.

Equivariance gives an action on P by $[p] \cdot g = [p \cdot g]$. It is continuous on every open chart, and the original local trivializations make P a torsor. Its datum

is isomorphic to the original one. For morphisms, a compatible family a_i gives the function $[p] \mapsto [a_i(p)]$, which is well-defined by the compatibility square and continuous on each chart. The rule gives the unique global arrow with those restrictions, since every point is represented in a chart. The resulting descent functor is fully faithful and essentially surjective. □

In a trivial chart choose the section corresponding to the identity in G. An equivariant transition between two such charts is left multiplication by a continuous $g_{ij} \colon U_{ij} \to G$. The triangle gives $g_{ij}g_{jk} = g_{ik}$, in the order shown there. The square for an arrow h_i between two trivial-chart data gives

$$g'_{ij} = h_i g_{ij} h_j^{-1}.$$

A change of charts can therefore alter the written transitions while preserving the glued torsor. The arrows themselves also keep track of the ways that such a change is possible.

60.4 Classifying covers of the circle

Use the cover $U = S^1 \setminus \{-i\}$, $V = S^1 \setminus \{i\}$ of Chapter 54. Each open is homeomorphic to an interval. A torsor on it is trivial. Indeed its projection is a covering map, and lifting paths from a chosen point and chosen fiber element gives a continuous section over the interval. The choice of path from the given point is irrelevant by homotopy lifting, and continuity can be checked in each sheet of an evenly covered neighborhood. A section trivializes the torsor using the action. See Hatcher for the covering-space lifting facts [46, Section 1.3].

The overlap has two components W_0, W_1 containing $1, -1$, respectively. A transition from the V-chart to the U-chart is a pair $(r, s) \in G \times G$, its constants on the two components. Changes of chart coordinates a on U and b on V send (r, s) to (arb^{-1}, asb^{-1}). Taking $a = 1$ and $b = r$ makes the first coordinate 1. Between normalized pairs $(1, g)$ and $(1, g')$, the first equation forces $a = b$ and the second becomes $g' = aga^{-1}$. Thus

$$\mathrm{Tors}_G(S^1) \simeq \mathrm{Fun}(B\mathbb{Z}, BG),$$

using BG for the one-object groupoid of G. The right-hand groupoid has functors determined by the image g of $1 \in \mathbb{Z}$. Its arrows satisfy the same conjugation equation. With the positive loop going from 1 to -1 through U and back through V, the transport based in the U-chart changes the fiber coordinate by left multiplication by rs^{-1}. Choosing the inverse loop as generator identifies the normalized comparison $g = sr^{-1}$ with the generator's action. This agrees with the groupoid calculation of maps from the circle into NBG in Chapter 56.

In particular $C_2 = \{1, \epsilon\}$ has two conjugacy classes. The transition $(1, 1)$ gives the product covering. The map $z \mapsto z^2$ on S^1 gives the other torsor by the action $z \cdot \epsilon = -z$. Its winding-one monodromy interchanges the two roots. Each torsor has automorphism group C_2, since both elements commute with its normalized transition. These two distinct isomorphism classes restrict to the same class on U and on V. The presheaf of isomorphism classes therefore fails even the uniqueness clause of the sheaf condition. The descent groupoid distinguishes the two global objects through their comparison maps on the overlap.

60.5 Quotient stacks and stabilizers

Use the fppf site of schemes over an algebraically closed field k of characteristic different from 2. Its covers are jointly surjective families of flat morphisms locally of finite presentation. Let $G = \mu_2$ be the group scheme of square roots of unity, and let G act on $X = \mathbb{A}^1_k$ by sign. This is also a right action because G is abelian. An object of $[X/G](T)$ is a right G-torsor $P \to T$ and an equivariant map $u\colon P \to X$. An arrow $(P, u) \to (P', u')$ is a torsor isomorphism a with $u = u'a$. Here a torsor is trivial on an fppf cover, with the usual right action on the product. As in the topological example it has local transition maps. Descent for torsors and maps gives the quotient stack [55, Proposition 95.15.3].

In a trivialization of P there is a chosen section. The value $x\colon T \to X$ of u on the section determines u by equivariance. An arrow $x \to x'$ between such presentations is given by $g \in G(T)$ with $x = gx'$, or $x' = gx$ since $g^2 = 1$. Thus on trivial torsors we see the action groupoid. Its presenting groupoid scheme has arrow space $X \times G$ and object space X.

$$X \times G \underset{(x,g)\mapsto gx}{\overset{(x,g)\mapsto x}{\rightrightarrows}} X.$$

The action groupoid embeds fully faithfully on each test scheme as the groupoid of trivial torsors with equivariant maps. Every object admits such a presentation on a cover. A map from the action groupoid to a stack extends by descent of these presentations, with common refinements comparing the choices. This is the universal property of *stackification*. Taking the trivial action on a point gives the *classifying stack* $BG = [*/G]$ of torsors.

To see a torsor introduced by gluing, take $T = \operatorname{Spec} k[t, t^{-1}]$ and

$$P = \operatorname{Spec} k[z, z^{-1}], \qquad t = z^2.$$

The algebra is free with basis $1, z$, so $P \to T$ is an fppf cover. It is a torsor under sign, since the ratio of the two invertible coordinates identifies $P \times_T P$ with $P \times G$, giving a trivialization over P. A section over T would give a Laurent unit at^m squaring to t, which requires $2m = 1$. Thus $(P, u = 0)$ lies outside the essential image of the action groupoid.

60.6 Counting objects with higher symmetries

A space X is *π-finite* if it has finitely many components, its homotopy groups are finite, and they vanish above some degree. Its *homotopy cardinality* sums a weight over components [72, Section 3].

$$|X|_{\mathrm{h}} = \sum_{[x]\in\pi_0 X} w(X,x), \qquad w(X,x) = \prod_{i\geq 1} |\pi_i(X,x)|^{(-1)^i}.$$

It extends ordinary cardinality on finite sets. For a finite groupoid each weight is the reciprocal of an automorphism group order. In higher types, degree-two symmetries contribute a factor in the numerator.

Let $F \to E \to B$ be a homotopy fiber sequence of π-finite spaces with B connected. Fix $b \in B$ and identify F with the fiber over b. Transport along loops acts on $\pi_0 F$ and the components of E correspond to its orbits. For a fiber component $[x]$ with stabilizer $S_x \subseteq \pi_1(B,b)$, the long exact sequence extends from the higher groups down to $\pi_1(F,x) \to \pi_1(E,x) \to S_x \to 1$. Multiplying alternating orders in this finite exact sequence, using $|G| = |\ker f|\,|\operatorname{im} f|$ at each arrow, gives

$$w(E,x) = w(F,x)\,\frac{\prod_{i\geq 2} |\pi_i(B,b)|^{(-1)^i}}{|S_x|}.$$

All fiber components in one orbit O have the same weight by transport, and $1/|S_x| = |O|/|\pi_1(B,b)|$. Summing the displayed formula over orbits therefore gives $|E|_{\mathrm{h}} = |F|_{\mathrm{h}}|B|_{\mathrm{h}}$.

Example 60.2 (Finite actions and a second delooping)**.** For a finite group G acting on a finite set X, the map $X/\!/G \to BG$ has fiber X. Since $|BG|_{\mathrm{h}} = 1/|G|$, the action groupoid has cardinality $|X|/|G|$, equivalently the sum of inverse stabilizer orders over the orbits. Burnside's lemma (Theorem 9.4) gives the number of orbits. For a finite abelian group A, the space $B^2A = K(A,2)$ has $\pi_2 = A$ and all other higher homotopy groups trivial. Its homotopy cardinality is $|A|$. This agrees with $BA \to * \to B^2A$, since $|BA|_{\mathrm{h}} = 1/|A|$.

60.7 What a scheme quotient remembers

Every μ_2-torsor over $\operatorname{Spec} k$ is trivial. For example it is finite étale of degree two by descent of those properties from a trivializing cover, and any such scheme over k splits. Thus over k the quotient stack has isomorphism classes represented by the orbits $\{x, -x\}$, and the automorphism group of such an object is the stabilizer of x. At $x = 0$ it is all of $\mu_2(k)$, while at $x \neq 0$ it consists only of the identity. At the level of invariant polynomials,

$$k[x]^{\mu_2} = k[x^2].$$

An invariant polynomial has zero coefficients in odd degrees because $2 \neq 0$ in k. The map $q\colon X \to \operatorname{Spec} k[x^2]$ is universal for invariant maps to affine schemes. Such a map corresponds to an algebra homomorphism with image in the invariant ring and therefore factors uniquely through that ring.

On a torsor with equivariant map u, the local square of its coordinate is invariant under all changes of trivialization. These squares descend to a map $T \to \operatorname{Spec} k[x^2]$, giving a functor

$$[X/G] \longrightarrow \operatorname{Spec} k[x^2].$$

This functor gives a bijection on k-point classes since each element has a square root and two roots differ by sign. It sends both automorphisms above 0 to the identity. The two torsors over $\operatorname{Spec} k[t, t^{-1}]$ equipped with $u = 0$ also give the same map to the scheme quotient.

Over the locus where the invariant coordinate is invertible, the scheme classifies the families. Write $X^\times = \operatorname{Spec} k[x, x^{-1}]$ and let $q^\times\colon X^\times \to \mathbb{G}_m$ be the square map. For a test map $b\colon T \to \mathbb{G}_m$ set $P_b = T \times_{\mathbb{G}_m} X^\times$. The pullback inherits a torsor from the square map, and its projection to $X^\times$ is equivariant.

$$\begin{array}{ccc} P_b & \longrightarrow & X^\times \\ \big\downarrow & & \big\downarrow{\scriptstyle q^\times} \\ T & \xrightarrow{\;b\;} & \mathbb{G}_m. \end{array}$$

If a torsor P with map $u\colon P \to X^\times$ has invariant coordinate b, its two maps to T and $X^\times$ define an equivariant map $P \to P_b$. Any such map of torsors is an isomorphism by checking on trivializing covers. The pullback property makes it unique among maps commuting with u. Thus each b determines a contractible groupoid of families, and the restriction $[X^\times/G] \to \mathbb{G}_m$ is an equivalence.

Moduli use a space of maps to retain isomorphisms of families. The next chapter uses such spaces to interpret identity types.

CHAPTER 61

Homotopy type theory from spaces and fibrations

A type over a context can have different fibers at different points of that context. In spaces, a path in the context also induces an equivalence between those fibers. Higher homotopies relate the equivalences obtained from different paths. We can interpret dependent type theory by asking what operations on such families recover substitution, dependent functions, and identity elimination. A universe of families then gives a way to regard equivalences of types themselves as paths.

61.1 Families over a context

Write $\mathcal{S}$ for the infinity-category of spaces from Chapter 56. A context Γ is a space, and a type in context Γ is a map $p\colon A \to \Gamma$, an object of the slice $\mathcal{S}_{/\Gamma}$. Its fiber over $\gamma\colon * \to \Gamma$ is the homotopy pullback $A_\gamma = * \times_\Gamma A$. Throughout this chapter, pullbacks and maps in slices are interpreted in the infinity-categorical sense. For display maps we can use Kan fibrations $A \to \Gamma$ over Kan complexes. Every map of spaces to a fixed Kan complex can be replaced by such a fibration over that complex. The strict fiber of a Kan fibration already models its homotopy fiber, by the pullback criterion of Chapter 57.

A map from $q\colon T \to \Gamma$ to p in the slice specifies an arrow $t\colon T \to A$ and a homotopy $pt \simeq q$, along with the higher data when it varies in a family. More precisely the mapping space in the slice is the fiber over q of

$$\mathrm{Map}(T, A) \longrightarrow \mathrm{Map}(T, \Gamma).$$

Thus a term of A in Γ is a section of p in the slice. In a Kan fibration model, the space of strict sections computes this fiber. The projection $\Gamma.A \to \Gamma$ of context extension in Chapter 27 is the map p itself, with total space A.

A substitution $u\colon \Delta \to \Gamma$ acts by pulling back. In the pullback square for reindexing, the fiber of u^*A at δ is $A_{u(\delta)}$.

$$\begin{array}{ccc} u^*A & \longrightarrow & A \\ \downarrow{\scriptstyle u^*p} & & \downarrow{\scriptstyle p} \\ \Delta & \xrightarrow{u} & \Gamma \end{array}$$

A section pulls back to a section, so substitution acts on terms as well. For $v\colon \Theta \to \Delta$, the comparison

$$v^*(u^*A) \simeq (uv)^*A$$

follows from the universal property of the outer pullback. Every further composition compares pullbacks of the same diagram, with their projections specified. The spaces of such comparisons are contractible when the presentations are fixed. This accounts for the coherent substitution laws. In a syntactic model, substitution may be required to satisfy judgmental equalities. That requirement calls for chosen data that are strictly stable under reindexing.

61.2 Dependent pairs and dependent functions

Consider a family $p\colon A \to \Gamma$ and a further family $q\colon E \to A$. Composition $pq\colon E \to \Gamma$ is the dependent sum, denoted $\Sigma_p E$. For a test object $T \to \Gamma$, a map $E \to T$ over Γ is equivalent to a map $E \to A \times_\Gamma T$ over A. The pullback property assembles q and the map to T with their compatibility over Γ, in all dimensions. Consequently Σ_p is left adjoint to p^*. The fiber of $\Sigma_p E$ over γ is the total space of E restricted to A_γ. A point of this fiber consists of a point a of A_γ and a point of E_a. A path between two such pairs specifies a path in the first coordinate and compares the second coordinates after transport along it.

The right adjoint Π_p to p^* represents the dependent function space. It exists because slices of an infinity-topos are presentable and pullback preserves their colimits. The latter is the universality of colimits discussed in Chapter 59. The adjoint functor theorem then supplies Π_p [51, Corollary 5.5.2.9 and Theorem 6.1.0.6]. Applying the adjunction to $T \to \Gamma$ yields

$$\mathrm{Map}_{/\Gamma}(T, \Pi_p E) \simeq \mathrm{Map}_{/A}(p^*T, E).$$

At $T = (* \xrightarrow{\gamma} \Gamma)$, the left side is the fiber $(\Pi_p E)_\gamma$. The right side is the space of sections of E restricted to A_γ. More generally the counit $p^*\Pi_p E \to E$ evaluates a dependent function at a point of A. Transposing a family of terms over A gives dependent abstraction. The triangular identities express evaluation of the abstracted term and reconstruction by abstraction after evaluation.

Example 61.1 (Sections remember loops)**.** Take $\Gamma = *$, $A = S^1$, and q the constant family with value X, given by $S^1 \times X \to S^1$. Then $\Sigma_p E \simeq S^1 \times X$, while $\Pi_p E \simeq \mathrm{Map}(S^1, X)$. A section really does vary over the whole circle, so its values along paths enter the mapping space. For $X = S^1$, maps of different winding numbers lie in different components. A free homotopy could only conjugate their loop classes, and the fundamental group is $\mathbb{Z}$ by Chapter 54. The circle construction at the end of the chapter will identify an arbitrary $\mathrm{Map}(S^1, X)$ by testing at a point and a loop.

61.3 Dependent binding commutes with substitution

For the pullback square

$$\begin{array}{ccc} A' & \xrightarrow{v} & A \\ \downarrow{\scriptstyle p'} & & \downarrow{\scriptstyle p} \\ \Gamma' & \xrightarrow{u} & \Gamma, \end{array}$$

both dependent binding operations commute with base change. The comparison for Σ is $u^*\Sigma_p E \simeq \Sigma_{p'} v^* E$. The total spaces in this equivalence are $\Gamma' \times_\Gamma E$ and $A' \times_A E$, which agree by the same pullback square. The comparison for Π can be deduced from the comparison for Σ. For any $T \to \Gamma'$,

$$\begin{aligned} \mathrm{Map}_{/\Gamma'}(T, u^*\Pi_p E) &\simeq \mathrm{Map}_{/\Gamma}(\Sigma_u T, \Pi_p E) \\ &\simeq \mathrm{Map}_{/A}(p^*\Sigma_u T, E) \\ &\simeq \mathrm{Map}_{/A}(\Sigma_v (p')^* T, E) \\ &\simeq \mathrm{Map}_{/A'}((p')^* T, v^* E) \\ &\simeq \mathrm{Map}_{/\Gamma'}(T, \Pi_{p'} v^* E). \end{aligned}$$

The middle line applies the sum comparison with the square read in the other direction. Since these equivalences are natural in T, the mapping-space Yoneda property gives $u^*\Pi_p E \simeq \Pi_{p'} v^* E$. These Beck–Chevalley equivalences are the homotopy versions of the comparisons in Section 27.6. For a type family and a term involving a new variable, they let us substitute in the context outside the binding and obtain the family computed by substituting under the binding.

61.4 Identity by a path object

In the Kan fibration presentation of $p\colon A \to \Gamma$, put $I = \Delta^1$. The relative path object $P_\Gamma A$ has as simplices the paths in A whose image in Γ is constant in the path coordinate. Equivalently it is $A^I \times_{\Gamma^I} \Gamma$, with $\Gamma \to \Gamma^I$ the constant-path map and exponentials taken in simplicial sets. Evaluation gives a factorization of the diagonal.

$$\begin{array}{ccccc} A & & \xrightarrow{r} & & P_\Gamma A \\ & \searrow^{\Delta_A} & & & \downarrow{\scriptstyle (e_0, e_1)} \\ & & & & A \times_\Gamma A \end{array}$$

Here r chooses a constant path. Restriction along $\partial I \hookrightarrow I$ for the Kan fibration p gives the Kan fibration of paths with specified boundary and specified image path in Γ. Pulling back along the subspace of constant image paths gives (e_0, e_1). Restriction along the trivial cofibration $\{0\} \hookrightarrow I$ by the same simplicial mapping-space property gives $e_0\colon P_\Gamma A \to A$ as a trivial Kan fibration.

Since $e_0 r = \mathrm{id}_A$ and r is a monomorphism, r is a trivial cofibration. These facts exhibit $P_\Gamma A$ as a fibrant presentation of the diagonal over $A \times_\Gamma A$. The identity type at (a, b) is its fiber, the space of paths $a \rightsquigarrow b$ in the indicated fiber of A. The interpretation of reflexivity is r.

The elimination rule of Chapter 29 can be read directly off the same factorization. Let $D \to P_\Gamma A$ be the Kan fibration for the type to be eliminated into. A section of its restriction along r is the diagonal case of the eliminator. The lifting problem is

$$\begin{array}{ccc} A & \xrightarrow{\;d\;} & D \\ {\scriptstyle r}\downarrow & \overset{J(d)}{\nearrow} & \downarrow \\ P_\Gamma A & \xrightarrow{\;\mathrm{id}\;} & P_\Gamma A. \end{array}$$

Its filler exists by the trivial cofibration–fibration lifting property. It is the requested section, and its restriction to r agrees with d. At the level of spaces of sections the restriction is an equivalence, since r is a weak equivalence and D is fibrant over the path object. This gives the corresponding uniqueness up to coherent homotopy.

Transport in a family $p\colon A \to \Gamma$ sends a point of A_γ along a path $\gamma \rightsquigarrow \gamma'$ to a point of $A_{\gamma'}$. In the fibration model the map from paths in A to their initial points and paths in Γ is a trivial Kan fibration, again by restriction along $\{0\} \hookrightarrow I$. Taking the endpoint of a lift defines transport, with a contractible space of choices of lifts. A path in the reverse direction gives inverse transport up to homotopy. The inverse laws follow by lifting the homotopies that contract the two concatenations to constant paths. This implements transport as equivalence of fibers, including its action when the original path varies.

Individual lifting problems give J on a chosen presentation. A model of syntax also needs stable choices for constructing types and their eliminators in every context. Using two inaccessible cardinals, Kapulkin and Lumsdaine's construction from universes supplies this coherence in simplicial sets and proves univalence for the resulting universe [56, Theorem 3.4.2]. The mapping properties above describe the homotopy semantics underlying that construction.

61.4.1 Paths between dependent functions

Let $q\colon E \to A$ be a type family and let f, g be sections. They are points of the dependent function space $F = \mathrm{Map}_{/A}(A, E)$. Put

$$D = A \times_{E \times_A E} P_A E,$$

using (f, g) on the left of the pullback. The fiber of D at a is the space of paths from $f(a)$ to $g(a)$ in E_a.

Taking the space of sections preserves limits, since it is the representable functor $\mathrm{Map}_{/A}(A, -)$. In the slice, $P_A E \to E \times_A E$ represents the diagonal of E, so on section spaces it represents the diagonal of F. Taking sections of D therefore gives

$$\mathrm{Id}_F(f, g) \simeq \mathrm{Map}_{/A}(A, D) \simeq \prod_{a \in A} \mathrm{Id}_{E_a}(f(a), g(a)).$$

The last expression denotes the dependent product of the path family. An element includes coherent variation in a. The comparison sends a path of sections to its values at the points of A. Its being an equivalence is function extensionality in the spaces model.

61.5 Univalence and the universal family

Fix a size bound for small types and work in a larger universe when forming their collection. In spaces the space of small types can be taken as $\mathcal{U} = (\mathcal{S}_{\mathrm{small}})^{\simeq}$, the maximal infinity-groupoid of the category of small spaces. Its points are small spaces and its path spaces are the equivalences between them, together with homotopies of those equivalences. The universal family $\widetilde{\mathcal{U}} \to \mathcal{U}$ has fiber X at X, with transport given by the equivalences $X \to Y$ occurring in the core. The classifying property identifies

$$\mathrm{Map}(\Gamma, \mathcal{U}) \simeq \left(\mathcal{S}^{\mathrm{small}}_{/\Gamma}\right)^{\simeq}.$$

Here the slice on the right is restricted to maps with small fibers. Straightening over the space Γ sends a family to a functor $\Gamma \to \mathcal{S}$ of fibers and coherent transport. Restricting its values to small spaces makes the functor land in $\mathcal{U}$, since every arrow in Γ is invertible. The universal family corresponds to the inclusion $\mathcal{U} \to \mathcal{S}$. In a general infinity-topos, for every sufficiently large regular cardinal κ, the corresponding result classifies relatively κ-compact maps [51, Section 3.2 and Theorem 6.1.6.8].

Transport in $\widetilde{\mathcal{U}}$ gives the map

$$\mathrm{idtoequiv}_{X,Y} \colon \mathrm{Id}_{\mathcal{U}}(X, Y) \longrightarrow \mathrm{Equiv}(X, Y).$$

Here $\mathrm{Equiv}(X, Y)$ consists of the components of $\mathrm{Map}(X, Y)$ whose maps are equivalences. Under the core description, the source is already the space of equivalences from X to Y and the map sends each path to that very equivalence. Thus $\mathrm{idtoequiv}_{X,Y}$ is an equivalence. This gives the semantic content of univalence for the universe of spaces. The type-theoretic inverse sends an equivalence e to its universe path, usually denoted $\mathsf{ua}(e)$. Transport along $\mathsf{ua}(e)$ acts by e. Univalence here concerns paths in a universe of types. The paths between elements of one type use the identity type of that type.

Example 61.2 (The universe of n-element types)**.** Let $n \geq 1$ and put $F_n = \{1, \ldots, n\}$. The component of $\mathcal{U}$ containing F_n contains precisely the spaces equivalent to this discrete space. Its loop space at F_n is the discrete group S_n of permutations. The component is therefore $\mathrm{N}BS_n$ up to equivalence, by the fundamental groupoid description of 1-types in Chapter 56. In particular the swap of two points gives a loop at F_2 distinct from reflexivity. It transports the first point to the second.

The total space of the universal family restricted to this component is the nerve of the action groupoid $F_n /\!\!/ S_n$. Its objects are points $i \in F_n$, and an arrow from i to j is a permutation carrying i to j. The projection sends that arrow to the permutation. Transitivity identifies this groupoid with the one-object groupoid of the stabilizer S_{n-1}. Its projection to BS_n is modeled, under this equivalence, by the inclusion of the stabilizer.

$$\begin{array}{ccc} \mathrm{N}BS_{n-1} & \xrightarrow{\simeq} & \mathrm{N}(F_n /\!\!/ S_n) \\ & \searrow & \downarrow \\ & & \mathrm{N}BS_n \end{array}$$

The upper arrow includes the object n, whose stabilizer is S_{n-1}. A section would pick a point of the n-element fiber fixed by every permutation. This follows by transporting the section around each loop in BS_n, since the fiber is discrete. For $n \geq 2$ there is no such point. Every fiber has elements, but choosing an element of every n-element type by a single term would have to respect all the equivalences in the universe.

61.6 The circle as a higher inductive type

An inductive type can specify what points its eliminator must receive, as in Chapter 28. A higher inductive type can also specify paths and give boundary data for their images. The circle $\mathbb{S}^1$ has one point constructor b and one path constructor $\lambda \colon b = b$. Its homotopy semantics is obtained by gluing the boundary of a single interval to a point.

$$\begin{array}{ccc} \partial\Delta^1 & \hookrightarrow & \Delta^1 \\ \downarrow & & \downarrow{\scriptstyle \lambda} \\ * & \xrightarrow{b} & \mathbb{S}^1 \end{array}$$

The simplicial pushout in this diagram is a model for the homotopy pushout, because the boundary inclusion is a monomorphism. Its realization is the ordinary circle. Here b and λ denote the induced vertex and edge before Kan replacement.

For any Kan complex X, testing the pushout gives

$$\mathrm{Map}(\mathbb{S}^1, X) \simeq X \times_{X \times X} X^{\Delta^1}.$$

The two maps to $X \times X$ are the diagonal and the endpoint fibration. The expression on the right is an ordinary pullback of Kan complexes along a fibration, so it computes the homotopy pullback. At a given $x \in X$ it has fiber $\Omega_x X$. Thus a map out of the circle amounts to a point x and a loop at x, with all higher comparisons also described by the mapping property. This is the recursion principle for the two constructors.

For the dependent principle, let a family $P \to \mathbb{S}^1$ be presented by a Kan fibration over the simplicial pushout model. A section is a lift of its identity map. The pushout says that such a lift consists of a point x over b and an edge above λ whose endpoints are both x. The latter consists, up to equivalence, of a path in P_b from the transport of x along λ back to x. Indeed, lifting λ with its first point specified has a contractible space of choices, and prescribing its last point takes the homotopy fiber of the endpoint map on this lift space. The same pushout argument applies with an additional simplicial parameter, so it describes the space of sections. These two pieces of data are the point and path cases of dependent circle induction [21, Section 6.4].

In this model the identity type $\mathrm{Id}_{\mathbb{S}^1}(b, b)$ is $\Omega_b \mathbb{S}^1$. The circle computation in Chapter 57 identifies it with the discrete space $\mathbb{Z}$, with λ corresponding to 1 after choosing its orientation. The synthetic calculation uses univalence to make a family with fiber $\mathbb{Z}$ and with successor as transport along λ. A loop is encoded by transporting 0 in that family. The inverse uses the powers of λ and circle induction [21, Section 8.1]. Here inductive presentation, categorical universality, and winding number describe the same circle at three levels of its construction.

CHAPTER 62

Derived geometry from homotopy pullbacks

Two copies of the origin in the affine line have the same ordinary intersection as the two coordinate axes in the affine plane, when each intersection is viewed as a scheme over the ground field. In both cases the fiber product has coordinate ring k. The homotopy pullback of the two origins has an additional group in degree one. We will derive it from the requirement that a test point carry a path identifying its two images.

62.1 Rings with higher homotopy

Fix an ordinary commutative ring k. A *simplicial commutative k-algebra* is a functor $\mathbf{\Delta}^{\mathrm{op}} \to \mathbf{CAlg}_k$. In each degree it has addition and multiplication, and the faces and degeneracies preserve them. Its underlying simplicial set is a Kan complex because it is a simplicial abelian group under addition. Write $U(A)$ for this underlying space. Its components $\pi_0(A)$ form a commutative k-algebra. The groups $\pi_i(A)$ for $i > 0$ are based at the additive zero and have a natural $\pi_0(A)$-module structure.

We use the ∞-category sCAlg_k obtained by inverting maps that induce isomorphisms on these homotopy groups. Additive translation identifies the homotopy groups at other basepoints in the same way, so these are exactly the weak homotopy equivalences of underlying spaces. This construction is available in every characteristic. The simplicial model category of these algebras presents the ∞-category of derived commutative k-algebras. Mapping spaces can be computed by replacing the source with a cofibrant simplicial algebra and taking the simplicial mapping object into the target [58, §2.2.1]. We write $\mathrm{Map}_{\mathrm{alg}}$ for these mapping spaces.

The inclusion of ordinary rings as constant simplicial algebras has left adjoint π_0. In fact a simplicial map to a constant target B sends the two endpoints of every edge to the same element, so factors through the components. For maps of simplicial algebras that factor is a map of rings. This argument works in every simplicial degree of the mapping object, since a map into a discrete target is

constant in its parameter simplex. Applying it after cofibrant replacement gives

$$\mathrm{Map}_{\mathrm{alg}}(A, B) \simeq \mathrm{Hom}_{\mathbf{CAlg}_k}(\pi_0 A, B)_{\mathrm{disc}}.$$

The polynomial algebra $k[t]$ is free on one generator in degree zero. The simplicial free-algebra adjunction makes it cofibrant, and evaluating at t gives

$$\mathrm{Map}_{\mathrm{alg}}(k[t], C) \simeq U(C).$$

For an ordinary target this says that a homomorphism is given by the image of t. For a simplicial target the same adjunction supplies families of homomorphisms, with the simplicial operators acting on the element assigned to t.

62.2 Testing an intersection by maps from affines

The ∞-category of derived affines over k is the opposite of sCAlg_k. Write $\mathrm{Spec}^{\mathrm{der}} A$ for the affine of A. In the language of Chapter 47, its points with test algebra C form the space $\mathrm{Map}_{\mathrm{alg}}(A, C)$. A map of algebras $R \to A$ consequently gives a map of affines in the reverse direction.

Take two maps $R \to A$, $R \to B$ of simplicial k-algebras. Their pushout in sCAlg_k is denoted $D = A \otimes^{\mathbf{L}}_R B$. Testing the pushout by maps to C gives an equivalence of spaces

$$\mathrm{Map}_{\mathrm{alg}}(D, C) \simeq \mathrm{Map}_{\mathrm{alg}}(A, C) \times^{\mathrm{h}}_{\mathrm{Map}_{\mathrm{alg}}(R, C)} \mathrm{Map}_{\mathrm{alg}}(B, C).$$

Thus a test point consists of maps out of A and B and a path between their restrictions to R. Chapter 57 describes the higher simplices of this path space. The mapping-space test proves that the opposite square

$$\begin{array}{ccc} \mathrm{Spec}^{\mathrm{der}} D & \longrightarrow & \mathrm{Spec}^{\mathrm{der}} B \\ \downarrow & & \downarrow \\ \mathrm{Spec}^{\mathrm{der}} A & \longrightarrow & \mathrm{Spec}^{\mathrm{der}} R \end{array}$$

is a pullback. This is the *derived fiber product* of the two affines over $\mathrm{Spec}^{\mathrm{der}} R$ [57, §3.1].

Restricting the tests to ordinary C turns this formula into the ordinary pullback of sets of ring homomorphisms. Equivalently, the left adjoint π_0 carries the pushout of derived algebras to the pushout of ordinary algebras. Hence

$$\pi_0(A \otimes^{\mathbf{L}}_R B) \cong \pi_0(A) \otimes_{\pi_0(R)} \pi_0(B).$$

The affine scheme $\mathrm{Spec}\, \pi_0 A$ is called the *classical truncation* of a derived affine. It has a natural map into the derived affine from $A \to \pi_0 A$.

62.3 Free resolutions and higher intersection groups

For this section let R, A, B be ordinary k-algebras. Choose free-cell cofibrant replacements $P_\bullet \to A$ and $Q_\bullet \to B$ as simplicial R-algebras, built by adjoining polynomial generators in simplicial degrees. Each degree is polynomial over R, including the freely adjoined degeneracies of generators from lower degrees. Their degreewise tensor $P_\bullet \otimes_R Q_\bullet$ computes the algebra pushout $A \otimes^{\mathbf{L}}_R B$ [57, §3.1].

Let $C_n(Q) = Q_n$ as an R-module and give $C_\bullet(Q)$ the alternating face differential $\partial_n = \sum_{j=0}^{n} (-1)^j d_j$ for $n > 0$. For a simplicial module its homology computes the homotopy groups of the underlying simplicial abelian group. One way to see this is to pass to the normalized complex, since the subcomplex generated by degeneracies is acyclic [52, §1.2.3]. The augmentation $Q_\bullet \to B$ induces an isomorphism on π_0, and $\pi_i(Q) = 0$ for $i > 0$. Consequently the augmented chain complex $C_\bullet(Q) \to B$ is a free resolution, with polynomial monomials as bases.

The map $P_\bullet \otimes_R Q_n \to A \otimes_R Q_n$ is a weak equivalence for each n, since Q_n is flat and tensoring the alternating face complex with it preserves homology. The diagonal of the resulting map of bisimplicial modules is a weak equivalence as well. Indeed the Eilenberg–Zilber comparison identifies the homology of the diagonal with that of the first-quadrant total complex [41, Theorem 8.5.1]. Taking homology first in the $P_\bullet$ direction proves the comparison for those total complexes. Hence $A \otimes_R Q_\bullet$ still models the pushout. Taking C_n of $A \otimes_R Q_\bullet$ and forming $A \otimes_R C_n(Q)$ gives the same chain complex in every degree. We obtain

$$\pi_i(A \otimes^{\mathbf{L}}_R B) \cong \operatorname{Tor}^R_i(A, B), \qquad i \geq 0.$$

The module computation of Chapter 52 now applies. The index i is a homotopy degree here and corresponds to cohomological degree $-i$ for the derived tensor of modules. The algebra pushout additionally specifies a multiplication and a mapping-space universal property.

In particular, if $A = R/I$ and $B = R/J$, the exact sequence $0 \to I \to R \to R/I \to 0$ and the long exact sequence for $\operatorname{Tor}^R(-, R/J)$ give

$$0 \longrightarrow \operatorname{Tor}^R_1(R/I, R/J) \longrightarrow I/IJ \longrightarrow R/J.$$

The term on the left injects since R is a free R-module. The last map sends the class of $a \in I$ to its class modulo J, so its kernel is $(I \cap J)/IJ$. We have recovered two groups of the intersection directly in terms of the two ideals.

$$\pi_0(D) = R/(I + J), \qquad \pi_1(D) \cong (I \cap J)/IJ.$$

For a repeated immersion, $I = J$, the second formula is the conormal module I/I^2 over R/I. If at least one of the maps is flat as a map of ordinary rings, the positive Tor groups vanish and $D \to \pi_0 D$ is an equivalence.

62.4 Three intersections over a field

Let k now be a field. The self-intersection of the origin in $\mathbb{A}^1_k$ has both incoming maps given by the evaluation $k[t] \to k$ at $t = 0$. The $k[t]$-module k has the resolution

$$0 \longrightarrow k[t] \xrightarrow{t} k[t] \longrightarrow k \longrightarrow 0.$$

Multiplication by t is injective and its cokernel is k. Tensoring with the other copy of k gives the two-term complex $k \xrightarrow{0} k$ in homological degrees $1, 0$. Writing $D_0 = k \otimes^{\mathbf{L}}_{k[t]} k$, we get

$$\pi_0(D_0) = k, \qquad \pi_1(D_0) \cong k, \qquad \pi_i(D_0) = 0 \quad (i > 1).$$

In the ideal description the class of t gives the generator of $(t)/(t^2)$. The product of two degree-one classes in the graded homotopy ring lands in degree two and is therefore zero.

For the two coordinate axes, put $R = k[x, y]$, $A = R/(x)$, and $B = R/(y)$. Resolve A by multiplication by x on R. Tensoring with $B = k[x]$ now gives $k[x] \xrightarrow{x} k[x]$. Its kernel is zero and its cokernel is k. Thus the derived intersection is equivalent to the ordinary point. In terms of ideals, $(x) \cap (y)$ is (xy), so the first positive homotopy group vanishes.

For a different example, intersect the line $y = 0$ and the parabola $y = x^2$ in the same plane. Resolve $R/(y)$ by multiplication by y, then tensor with $R/(y - x^2)$. The complex is $k[x] \xrightarrow{x^2} k[x]$. All positive homotopy groups still vanish, and the ordinary intersection is $\operatorname{Spec}(k[x]/(x^2))$. This last computation exhibits a nonreduced classical intersection in degree zero. The self-intersection of the origin has a reduced classical point and a nonzero group in degree one.

62.5 Infinitesimal tests and the cotangent complex

The initial algebra k has a contractible space of maps into any test C. The two restrictions to $k[t]$ send t to zero, so the universal property gives

$$\operatorname{Map}_{\mathrm{alg}}(D_0, C) \simeq * \times^{\mathrm{h}}_{U(C)} * \simeq \Omega_0 U(C).$$

This is the loop pullback of Chapter 57 with the space of possible values of t as the base. For ordinary C, the pullback is contractible.

For a simplicial k-module M, the split square-zero extension $k \oplus M$ has multiplication $(a, m)(b, n) = (ab, an + bm)$ degreewise. Its additive fiber over $0 \in k$ is M. The nerve of the additive group of k gives such a module, with n-simplices k^n and linear face and degeneracy maps. Its underlying space is the 1-type $K(k, 1)$, so this test gives

$$\operatorname{Map}_{\mathrm{alg}}(D_0, k \oplus K(k, 1)) \simeq \Omega K(k, 1) \simeq k_{\mathrm{disc}}.$$

Here k_{disc} is the discrete space on k. The ordinary point $\operatorname{Spec} k$ has a contractible space of points on the same test.

For an augmentation $a\colon D \to k$, the space of derived derivations at a is the homotopy pullback

$$\begin{array}{ccc} \mathrm{Der}_{k,a}(D,M) & \longrightarrow & \mathrm{Map}_{\mathrm{alg}}(D, k \oplus M) \\ \downarrow & & \downarrow \\ * & \xrightarrow{\ a\ } & \mathrm{Map}_{\mathrm{alg}}(D,k). \end{array}$$

The cotangent complex at a is connective and represents this functor of M by maps in the derived category of modules enhanced to an ∞-category [57, §3.1]. For D_0, the target of that map of spaces is contractible, so the representing problem reduces to $\Omega_0 U(M)$.

Under normalization, write $\mathcal{D}(k)$ for the stable ∞-category of complexes of k-modules and regard M as a complex in cohomological degrees ≤ 0. Mapping out of the free module k gives $U(M)$. The cochain shift $[1]$ is suspension, so

$$\mathrm{Map}_{\mathcal{D}(k)}(k[1], M) \simeq \Omega_0 \,\mathrm{Map}_{\mathcal{D}(k)}(k, M) \simeq \Omega_0 U(M).$$

Both representing objects are connective, so Yoneda identifies the cotangent complex at the point with $k[1]$. Testing by k in degree zero gives a singleton tangent set. With k in degree -1, the components form the additive group of k.

62.6 Localizing and gluing derived affines

A derived affine can be restricted to a principal open of its classical truncation. For $f \in \pi_0(A)$, choose a representative giving $k[t] \to A$ with t sent to f. The required algebra is

$$A[f^{-1}] = A \otimes^{\mathbf{L}}_{k[t]} k[t, t^{-1}].$$

In each homotopy degree, $\pi_i(A[f^{-1}]) \cong (\pi_i A)[f^{-1}]$. Indeed $k[t,t^{-1}]$ is flat over $k[t]$. After a cofibrant replacement of A over $k[t]$, the same diagonal comparison used above allows the second leg $k[t,t^{-1}]$ to remain constant. Tensoring the alternating face complex with this flat algebra commutes with its homology. The action of t on those groups is multiplication by f. This calculation also identifies the classical open with $\operatorname{Spec}((\pi_0 A)[f^{-1}])$.

To see the mapping-space property, write $R = k[t]$ and $S = k[t,t^{-1}]$. Multiplication $S \otimes^{\mathbf{L}}_R S \to S$ is an equivalence. Flatness kills the positive groups, and ordinary localization gives $S \otimes_R S \cong S$. Thus the diagonal of $\mathrm{Map}_{\mathrm{alg}}(S, C) \to \mathrm{Map}_{\mathrm{alg}}(R, C)$ is an equivalence. Its fibers are empty or contractible. They are nonempty at a map carrying t to a unit on $\pi_0(C)$, since in that case $C \to C[t^{-1}]$ is an equivalence

by the homotopy group calculation, and $S \to C[t^{-1}]$ supplies a lift. Every lift sends t to a unit on π_0. The pushout universal property for $A[f^{-1}]$ now identifies its maps to C with the subspace of maps $A \to C$ that take f to a unit on $\pi_0(C)$. This refines the ordinary localization property of Chapter 11. These localizations give the principal opens of derived affines [58, §2.2.2].

A family of derived Zariski open immersions covers its target precisely when the classical opens cover the classical truncation. One glues derived affine charts by identifying the restrictions on overlaps with coherent compatibilities. A derived scheme is a sheaf of spaces on derived affines that admits such Zariski affine charts. Derived affine functors of points satisfy Zariski descent, a special case of derived étale descent [58, §2.2.2]. The homotopy groups of the local algebras glue to quasi-coherent modules on the classical scheme, since on each principal open their restrictions are the ordinary localizations just computed.

The local tensor computation and the gluing step use the same test by spaces of maps. Maps into an intersection agree along its base by a specified path. Local descriptions of that intersection then agree on overlaps with the coherences of descent. This combines the space-valued sheaves of Chapter 59 with the exactness calculation of Chapter 52.

Appendices

APPENDIX A

Diagram notation and duality

A diagram records objects, arrows, and relations between composites. A universal diagram adds a statement about maps from or to every test object. This appendix gives a way to read those statements and to reverse them when passing to the opposite category.

A.1 Reading the routes

An arrow $f\colon X \to Y$ has source X and target Y. In a chain $X \xrightarrow{f} Y \xrightarrow{g} Z$, the composite $g \circ f$ applies f first. Its source and target are X and Z. In the square

$$\begin{array}{ccc} X & \xrightarrow{f} & Y \\ \downarrow{\scriptstyle u} & & \downarrow{\scriptstyle v} \\ X' & \xrightarrow{f'} & Y' \end{array}$$

the two routes from X to Y' are $v \circ f$ and $f' \circ u$. To assert that the square commutes is to assert $v \circ f = f' \circ u$. For instance, when X, Y, X', Y' are sets and f, f' are the functions $x \mapsto x^2$ on $\mathbb{Z}$ and $\mathbb{Z}/5\mathbb{Z}$, respectively, both u and v can be reduction modulo 5. The square commutes because the quotient map preserves multiplication.

A diagram of shape $\mathcal{J}$ in $\mathcal{C}$ is formally a functor $D\colon \mathcal{J} \to \mathcal{C}$. The shape may be specified by generating arrows and relations among paths. For a commutative square the shape has a relation between its two length-two paths. The shape used for an equalizer has two parallel arrows $f, g\colon X \rightrightarrows Y$, with no relation between them. The relation appears only after precomposing with the equalizer arrow $e\colon E \to X$, when $f \circ e = g \circ e$. Keeping track of where a relation holds is part of a diagram chase [2, §§1.4, 1.6].

Remark A.1 (Arrow marks). We use a hook arrow $X \hookrightarrow Y$ to mark an inclusion or monomorphism, and a double-headed arrow $X \twoheadrightarrow Y$ for a surjection or epimorphism. A dashed arrow often marks a factorization or a filler. Its existence and any uniqueness come from the accompanying property. A corner mark signals the universal property of a pullback square.

A.2 Reversing a category

The *opposite category* $\mathcal{C}^{\mathrm{op}}$ has the same objects as $\mathcal{C}$ and an arrow $f^{\mathrm{op}}\colon Y \to X$ for each $f\colon X \to Y$ of $\mathcal{C}$. Its composition is

$$(g \circ f)^{\mathrm{op}} = f^{\mathrm{op}} \circ g^{\mathrm{op}}.$$

Associativity and the identity laws in $\mathcal{C}$ give the same laws in $\mathcal{C}^{\mathrm{op}}$. Reversing twice recovers $\mathcal{C}$. Likewise D gives a functor $D^{\mathrm{op}}\colon \mathcal{J}^{\mathrm{op}} \to \mathcal{C}^{\mathrm{op}}$. A relation among composites in $\mathcal{J}$ becomes the reversed relation in $\mathcal{J}^{\mathrm{op}}$.

For example, $m\colon A \to B$ is monic if $m \circ r = m \circ s$ implies $r = s$ for all pairs $r, s\colon T \to A$. In the opposite category the same test reads

$$r^{\mathrm{op}} \circ m^{\mathrm{op}} = s^{\mathrm{op}} \circ m^{\mathrm{op}} \quad \Longrightarrow \quad r^{\mathrm{op}} = s^{\mathrm{op}}.$$

This is the cancellation property of an epimorphism. Thus m is monic in $\mathcal{C}$ exactly when m^{op} is epic in $\mathcal{C}^{\mathrm{op}}$. For sets, monicity gives injectivity by choosing $T = *$ and testing two elements. Injectivity proves the cancellation property by evaluation at each point of any test set. For epimorphisms, surjectivity gives cancellation by checking at elements in the image. If the image is proper, the indicator of its complement and the constant-zero function to $\{0, 1\}$ become equal after precomposition, though they differ on the complement. In commutative unital rings, the inclusion $\mathbb{Z} \to \mathbb{Q}$ is epic. Any two maps out of $\mathbb{Q}$ agreeing on $\mathbb{Z}$ send a/b to the image of a times the inverse of the image of b.

An assertion valid in every category can be applied to $\mathcal{C}^{\mathrm{op}}$ and translated back to $\mathcal{C}$. This gives the duality principle [2, §1.2]. For example, the proof in Section 3.1 that every equalizer is monic simultaneously gives that every coequalizer is epic. An existence hypothesis also reverses. In the category of nonempty sets a singleton is terminal. An initial object would have a unique map to a two-element set, but from any nonempty set there are at least two constant maps. The terminal object of this category instead gives an initial object of its opposite.

A.3 Dual universal properties

For a small shape $\mathcal{J}$, let $\Delta T\colon \mathcal{J} \to \mathcal{C}$ be constant at T. A cone $\Delta T \Rightarrow D$ consists of maps $\lambda_j\colon T \to D(j)$ with $D(a) \circ \lambda_i = \lambda_j$ for $a\colon i \to j$. A cocone is a transformation $D \Rightarrow \Delta T$. The limit L and colimit Q give the tests

$$\mathrm{Hom}_{\mathcal{C}}(T, L) \cong \mathrm{Nat}(\Delta T, D),$$
$$\mathrm{Hom}_{\mathcal{C}}(Q, T) \cong \mathrm{Nat}(D, \Delta T).$$

For the first bijection, composing with the universal cone sends a map $T \to L$ to its induced family. For the second, composing with the universal cocone sends a

map $Q \to T$ to a family with target T. Reversing all arrows exchanges the two bijections.

The pullback gives a concrete instance. The left square below comes with a projection to each end of a cospan. Its opposite is the pushout square on the right, with arrows labeled in the directions of $\mathcal{C}^{\mathrm{op}}$.

$$\begin{array}{ccc} P & \xrightarrow{p_X} & X \\ {\scriptstyle p_Y}\downarrow & \lrcorner & \downarrow{\scriptstyle f} \\ Y & \xrightarrow{g} & Z \end{array} \qquad \begin{array}{ccc} P & \xleftarrow{p_X^{\mathrm{op}}} & X \\ \uparrow{\scriptstyle p_Y^{\mathrm{op}}} & & \uparrow{\scriptstyle f^{\mathrm{op}}} \\ Y & \xleftarrow{g^{\mathrm{op}}} & Z \end{array}$$

In the left square, maps $x\colon T \to X$ and $y\colon T \to Y$ with $f \circ x = g \circ y$ factor through a unique $t\colon T \to P$. The reversed assertion says that maps $x^{\mathrm{op}}\colon X \to T$ and $y^{\mathrm{op}}\colon Y \to T$ agreeing on Z factor through a unique $t^{\mathrm{op}}\colon P \to T$. These are the two tests of Chapter 3. The dashed arrows show the respective factorizations.



A diagram on two objects and their identities alone gives products and coproducts. For the empty diagram the compatible family consists of no maps and there is exactly one such family. Consequently a limit of the empty diagram is terminal and a colimit is initial.

The direction of that test also gives the usual uniqueness statement. Given two limiting cones on the same diagram with vertices L, L', apply each universal property to the other cone to obtain $L \to L'$ and $L' \to L$. Their composites preserve the universal cones. Since an identity arrow has the same property, the uniqueness clause identifies each composite with the identity. Thus the comparison is the unique isomorphism preserving the cones. Reversal gives the corresponding statement for colimits and their cocones. This argument uses the structure maps as well as the objects named in a diagram [2, Chapter 3].

A.4 Variance and two directions of reversal

The hom functor $\mathrm{Hom}_{\mathcal{C}}(-,-)$ has source $\mathcal{C}^{\mathrm{op}} \times \mathcal{C}$. For arrows $u\colon A' \to A$, $v\colon B \to B'$, an element $h\colon A \to B$ is sent to $v \circ h \circ u$. This gives the square of sets

$$\begin{array}{ccc} \mathrm{Hom}(A,B) & \xrightarrow{v\circ(-)} & \mathrm{Hom}(A,B') \\ {\scriptstyle (-)\circ u}\downarrow & & \downarrow{\scriptstyle (-)\circ u} \\ \mathrm{Hom}(A',B) & \xrightarrow{v\circ(-)} & \mathrm{Hom}(A',B') \end{array}$$

which commutes by associativity. The first variable runs opposite to the original arrows because precomposition changes the source. The second runs in their direction because postcomposition changes the target. This is the variance used in Chapter 4.

Reversal exchanges the sides of an adjunction. If $F\colon \mathcal{C} \to \mathcal{D}$ is left adjoint to $G\colon \mathcal{D} \to \mathcal{C}$, then the adjunction bijection becomes

$$\mathrm{Hom}_{\mathcal{C}^{\mathrm{op}}}(G^{\mathrm{op}}Y, X) \cong \mathrm{Hom}_{\mathcal{D}^{\mathrm{op}}}(Y, F^{\mathrm{op}}X).$$

Naturality follows by reversing the two variables of the original bijection. Thus $G^{\mathrm{op}} \dashv F^{\mathrm{op}}$.

A strict 2-category has two kinds of arrows and therefore two separate reversals. For a strict 2-category $\mathcal{K}$, the hom categories of $\mathcal{K}^{\mathrm{op}}$ are $\mathcal{K}^{\mathrm{op}}(A, B) = \mathcal{K}(B, A)$, with the order of horizontal composition reversed. This reverses 1-cells and retains the direction of the 2-cells between them. The hom categories of $\mathcal{K}^{\mathrm{co}}$ are $\mathcal{K}^{\mathrm{co}}(A, B) = \mathcal{K}(A, B)^{\mathrm{op}}$. Here the objects and 1-cells keep their direction, and the 2-cells reverse. The operations commute, giving $\mathcal{K}^{\mathrm{coop}}$ when both are applied.

Taking opposites of the *individual categories* in an example has a separate effect. For functors $F, G\colon \mathcal{A} \to \mathcal{B}$, a natural transformation $\eta\colon F \Rightarrow G$ induces $\eta^{\mathrm{op}}\colon G^{\mathrm{op}} \Rightarrow F^{\mathrm{op}}$. Its component is the reversed arrow $\eta^{\mathrm{op}}_a\colon G(a) \to F(a)$ in $\mathcal{B}^{\mathrm{op}}$. Reversing the naturality square for $a \to b$ proves naturality for $b \to a$ in $\mathcal{A}^{\mathrm{op}}$. This gives a 2-functor $\mathbf{Cat}^{\mathrm{co}} \to \mathbf{Cat}$ sending a category to its opposite. The distinction between vertical and horizontal composition is developed in Chapter 55 [49, §2].

A.5 Commuting up to specified homotopy

For maps of spaces, one can specify a homotopy between the two routes in a square. In a homotopical interpretation of the square in Section A.1, this is a path in $\mathrm{Map}(X, Y')$ between the two composites.

$$X \;\overset{v\circ f}{\underset{f'\circ u}{\Downarrow H}}\; Y'$$

The equality of homotopy classes $[v \circ f] = [f' \circ u]$ says that such a path exists. Keeping the path itself makes it possible to compare different choices by homotopies between paths. In Chapter 56 this distinction is expressed by simplices of the mapping spaces. For diagrams with several compositions, coherence includes compatibility with the specified comparisons on their faces.

For a homotopy pullback of $X \xrightarrow{f} Z \xleftarrow{g} Y$, a test object T supplies maps $x\colon T \to X$, $y\colon T \to Y$, and a path $\gamma\colon f \circ x \rightsquigarrow g \circ y$ in $\mathrm{Map}(T, Z)$. The universal

property is an equivalence of spaces

$$\mathrm{Map}(T, X \times^{\mathrm{h}}_{Z} Y) \simeq \mathrm{Map}(T, X) \times^{\mathrm{h}}_{\mathrm{Map}(T,Z)} \mathrm{Map}(T, Y).$$

The right side includes paths between triples and all their higher homotopies. Once a comparison with the given triple is included, the space of factorizations is contractible, as a homotopy fiber of the displayed equivalence.

Take $T = X = Y = *$, $Z = S^1$, and let both maps choose the same base point. The strict square test using equality of maps has one pair of legs. For the homotopy test the paths γ are based loops. Their components are indexed by the winding number, giving $* \times^{\mathrm{h}}_{S^1} * \simeq \Omega S^1 \simeq \mathbb{Z}_{\mathrm{disc}}$, as computed in Section 57.3. Here all the different components come from the comparison along the base of the cospan. Reversal exchanges coherent cones and coherent cocones, with a mapping-space universal property in each direction.

APPENDIX B

Size, universes, and smallness assumptions

A universal property can quantify over every test object of a category. A product has a separate requirement on the family being assembled. To construct a product of sets as a set of tuples, the coordinates must themselves form a set of the allowed size. This appendix tracks the sizes in these two roles and in constructions involving whole categories.

B.1 A universe for the data

Fix a Grothendieck universe $\mathcal{U}$. It is a transitive set containing $\mathbb{N}$, closed under pairs and power sets, and closed under unions of families of its members indexed by a member of $\mathcal{U}$. Transitivity means that if $x \in y$ and $y \in \mathcal{U}$, then $x \in \mathcal{U}$. One standard choice is the rank segment V_κ for an uncountable strongly inaccessible cardinal κ. Its regularity bounds the supremum of fewer than κ ranks less than κ, and the strong limit condition bounds the cardinalities of the power sets at each stage. When collecting the objects of $\mathcal{U}$, work in a larger ambient universe $\mathcal{V}$ with $\mathcal{U} \in \mathcal{V}$. This is the kind of size convention used in higher category theory [51, Section 1.2.15].

Asking for such a supply of universes is an additional set-theoretic assumption. Ordinary ZFC already supplies bounds for specified set-indexed constructions. For example, Replacement and Union collect the data of a given family of sets in a single set, and Power Set can then be used to form their product as a set of graphs. The universe convention gives one common bound that is closed under these operations. In a development using ZFC and definable classes, one instead specifies the set-indexed constructions as they arise and treats large collections by formulas in the ambient theory.

A set is $\mathcal{U}$-small if it is in bijection with a member of $\mathcal{U}$. We may choose such a representative when using the closure properties. For a family $(X_i)_{i \in I}$ of members with $I \in \mathcal{U}$, closure under indexed unions puts $\bigcup_i X_i$ in $\mathcal{U}$. A product of the X_i is a set of graphs in $I \times \bigcup_i X_i$. Pairs, power sets, and transitivity show that this set of graphs belongs to $\mathcal{U}$. A disjoint union is a subset of that same cartesian product. Subsets and quotients of a member of $\mathcal{U}$ are again members, since a quotient is a set of subsets and the universe is closed under power sets.

B.2 Small categories and small hom sets

A category is $\mathcal{U}$-small if its object and arrow sets are $\mathcal{U}$-small. It is *locally* $\mathcal{U}$-small if each $\mathrm{Hom}(A, B)$ is $\mathcal{U}$-small. It is *essentially* $\mathcal{U}$-small if it is equivalent to a $\mathcal{U}$-small category. A small full subcategory meeting every isomorphism class exhibits the last property. For example, the full subcategory of finite sets on $[n] = \{0, \dots, n-1\}$, $n \in \mathbb{N}$, is small and equivalent to the category of finite sets.

Let $\mathbf{Set}_{\mathcal{U}}$ have all members of $\mathcal{U}$ as objects and all functions between them as arrows. Its hom sets are small, since functions form a subset of a power set. It is a category in the ambient universe $\mathcal{V}$. To test essential smallness, suppose a family of objects $(A_i)_{i \in I}$ with I small meets every isomorphism class. Let $B = \coprod_i A_i$, an object of $\mathbf{Set}_{\mathcal{U}}$. The power set of B is another object of $\mathbf{Set}_{\mathcal{U}}$. Every A_i injects into B, so Cantor's theorem makes $\mathcal{P}(B)$ too large to be isomorphic to any A_i. Thus even a category with small hom sets may require a larger universe for its collection of isomorphism classes.

B.3 Functor categories and presheaves

Suppose $\mathcal{C}$ is $\mathcal{U}$-small and $\mathcal{D}$ is locally $\mathcal{U}$-small. For fixed functors $F, G \colon \mathcal{C} \to \mathcal{D}$, a natural transformation is a point of a product of small hom sets that satisfies every naturality square.

$$\begin{array}{ccc} F(c) & \xrightarrow{\eta_c} & G(c) \\ {\scriptstyle F(f)}\downarrow & & \downarrow{\scriptstyle G(f)} \\ F(d) & \xrightarrow{\eta_d} & G(d) \end{array} \qquad (f \colon c \to d)$$

Equivalently, the following two products and parallel maps express natural transformations as an equalizer.

$$\mathrm{Nat}(F, G) \hookrightarrow \prod_{c} \mathrm{Hom}(Fc, Gc) \rightrightarrows \prod_{f : c \to d} \mathrm{Hom}(Fc, Gd)$$

At an arrow f, the two outputs are $G(f)\eta_c$ and $\eta_d F(f)$. Smallness of the indexing sets and local smallness of $\mathcal{D}$ put the products and the equalizer in the same universe. For an essentially small source, restriction along an equivalence with a small category induces an equivalence of functor categories, so the same bound on their hom sets applies.

The argument can fail for a large source. Let $\mathcal{U} = V_\kappa$ and take $\mathcal{C}$ discrete on the ordinals $i < \kappa$. In the ambient universe the natural transformations from the constant singleton presheaf to the constant two-element presheaf are precisely functions $\kappa \to \{0, 1\}$. The only arrows of $\mathcal{C}$ are identities, so all functions are

allowed. There are 2^κ such families, greater than the cardinality of any member of V_κ. Pointwise small values therefore allow transformations of larger size.

The Yoneda test of Chapter 4 has a separate reason for giving a small result. Suppose only that $\mathcal{C}$ is locally $\mathcal{U}$-small, with its data in $\mathcal{V}$, and P takes values in $\mathbf{Set}_\mathcal{U}$. Use small representatives for the hom sets when necessary. For $c \in \mathcal{C}$, evaluation at id_c identifies $\mathrm{Nat}(\mathrm{Hom}(-, c), P)$ with $P(c)$. An element $x \in P(c)$ determines every component by $f \mapsto P(f)(x)$. Thus natural transformations with this representable source have a small parameter set even when the product over all objects is large. A small site in the presheaf and sheaf constructions gives local smallness uniformly for all pairs of presheaves.

B.4 Small diagrams in limits and Kan extension

Given a small diagram of sets $D\colon I \to \mathbf{Set}_\mathcal{U}$, its limit is a subset of $\prod_{i \in I} D(i)$, obtained by requiring agreement along every arrow. Its colimit is a quotient of $\coprod_{i \in I} D(i)$, obtained by identifying an element with each of its images. Both stay in $\mathbf{Set}_\mathcal{U}$ by the closure argument. This explains which diagrams the completeness and cocompleteness statements for sets cover.

For the pointwise left Kan extension along $J\colon \mathcal{A} \to \mathcal{B}$ in Chapter 25, the indexing category $(J \downarrow b)$ has objects $(a, u : Ja \to b)$. Its arrows are the $f\colon a \to a'$ satisfying the triangle below.

$$\begin{array}{ccc} Ja & \xrightarrow{Jf} & Ja' \\ & {\scriptstyle u}\searrow & \downarrow{\scriptstyle u'} \\ & & b \end{array}$$

If $\mathcal{A}$ is $\mathcal{U}$-small and $\mathcal{B}$ locally $\mathcal{U}$-small, the possible objects lie in the small disjoint union of the sets $\mathrm{Hom}(Ja, b)$. The arrows can be counted by taking the small set of pairs of objects and taking, for each pair, the subset of arrows of $\mathcal{A}$ making the triangle commute. The dual computation bounds $(b \downarrow J)$. Consequently a target with all small colimits and limits supplies the pointwise values

$$(\mathrm{Lan}_J F)(b) = \operatorname*{colim}_{(a,u) \in (J \downarrow b)} F(a), \qquad (\mathrm{Ran}_J F)(b) = \lim_{(a,u) \in (b \downarrow J)} F(a).$$

The pointwise formulas apply whenever the particular colimit and limit exist. The size hypotheses are one way to supply those objects for every b.

Theorem B.1 (Freyd)**.** *A $\mathcal{U}$-small category with all $\mathcal{U}$-small products has at most one arrow between each ordered pair of objects.*

Proof. Write I for the set of all arrows, and suppose $f, g\colon A \to B$ are different. The power B^I exists. Each subset $S \subseteq I$ determines the map $A \to B^I$ whose i-th

component is f for $i \in S$ and g otherwise. If two subsets differ at i, composing their maps with the i-th projection gives different arrows. This injects $\mathcal{P}(I)$ into $\mathrm{Hom}(A, B^I) \subseteq I$, contradicting Cantor's theorem. □

This cardinality argument also applies to coproducts by duality [2, Proposition 3.7.3]. For instance, the small category of representatives for finite sets has all finite products, whereas the countable power of a two-element set would have infinitely many global elements. The theorem concerns products indexed by sets up to the size of the entire arrow collection.

B.5 A solution set for an adjoint

For a functor $U\colon \mathcal{A} \to \mathcal{B}$, a left adjoint is obtained by representing $\mathrm{Hom}_{\mathcal{B}}(b, U(-))$ at each b, as in Chapter 5. To find a representing object at b is to find an initial object of $(b \downarrow U)$. A solution set is a family of pairs $(a_i, \eta_i : b \to Ua_i)$ indexed by a small set, such that every $\eta : b \to Ua$ factors as $U(f)\eta_i$ for some i and $f : a_i \to a$. This is a weakly initial family in the comma category. If U preserves small limits and $\mathcal{A}$ is locally small and has those limits, Freyd's general adjoint functor theorem uses the solution set to obtain an initial object [2, Theorem 4.7.3].

For a concrete failure, order the ordinals $< \kappa$ by reverse comparison and view this order as a category $\mathcal{A}$. Here κ is the cardinal for $\mathcal{U} = V_\kappa$. Every small diagram has a limit. Its value is the supremum of its objects in the usual ordinal order, which is still below κ by regularity. Indeed a cone has a bound above each object in the usual order, and any such bound maps to the least one in the reverse order. The empty diagram has limit 0. The functor $U\colon \mathcal{A} \to \mathbf{1}$ to the one-object category preserves all these limits. An initial object of $\mathcal{A}$ would be a greatest ordinal below κ, which cannot exist. The same obstruction defeats a small solution set. The supremum of any small candidate family is below κ, and its successor receives no arrow from a member of that family. The large diagram given by the identity on $\mathcal{A}$ likewise has no limit.

B.6 Spaces, universes of types, and localization

In Chapter 59, a small site gives a presheaf category valued in small spaces. In a simplicial presentation, the simplices of a small space in all degrees form small data, since the simplicial degrees are indexed by $\mathbb{N}$. The resulting category of small spaces can be assembled in a larger universe, as for $\mathbf{Set}_{\mathcal{U}}$. The accessible left exact localization used to define an infinity-topos has a small indexing infinity-category for its presheaf presentation [51, Definition 6.1.0.4].

The universe of types in Section 61.5 lies at the larger size. For example, even the discrete small types have isomorphism classes unbounded among small

cardinalities. Any small family of candidate representatives is bounded by its disjoint union, and the power set gives a larger small type. A universe space whose points classify all small types up to equivalence must therefore have a larger component set. Once it is formed in the larger universe, the universal family has small fibers. Its pullback to a context gives a family of the prescribed size.

$$\begin{array}{ccc} E & \longrightarrow & \widetilde{\mathcal{U}}_{\mathrm{type}} \\ \downarrow & & \downarrow \\ \Gamma & \longrightarrow & \mathcal{U}_{\mathrm{type}} \end{array}$$

The simplicial univalent model cited there uses two inaccessible cardinal bounds for the stated construction [56, Theorem 3.4.2]. The bounds specify both the allowable fibers and an ambient setting for their classifier.

Localization gives a different size check. In Section 52.4, roofs between two fixed bounded-above complexes can involve any complex of the chosen size as an intermediate object. To obtain a small hom set, fix the ring in $\mathcal{U}$ and use the bounded-above free replacements of that section. The construction uses small sets at each degree, and there are only countably many degrees. Maps between the chosen replacements form a subset of a small product of module hom sets, and dividing by homotopy gives a small quotient. The replacement calculation bounds the arrows in the localization even though the initial collection of roofs may range over the entire category.

APPENDIX C

Worked proofs

The proofs in this appendix compare constructions by following their maps. An adjunction transports a comparison between functors. A norm compares two constructions applied to the same group action. A free resolution compares the ordinary and derived intersections of the same subschemes. In each case a proposed map is specified before its properties are proved.

C.1 Taking mates across adjunctions

Chapter 5 transposes a map across an adjunction by using its unit or counit. This procedure applies to a natural transformation when each of its components has the indicated shape. Suppose $L \dashv R$ and $L' \dashv R'$ are adjunctions, with the left adjoints shown in the following square.

$$\begin{array}{ccc} \mathcal{C} & \xrightarrow{L} & \mathcal{D} \\ \downarrow{\scriptstyle T} & & \downarrow{\scriptstyle S} \\ \mathcal{C}' & \xrightarrow{L'} & \mathcal{D}' \end{array}$$

A natural transformation $\alpha\colon L'T \Rightarrow SL$ determines a *mate* $\beta\colon TR \Rightarrow R'S$. Write η, ε for the unit and counit of $L \dashv R$ and η', ε' for those of $L' \dashv R'$. The formula for β at $D \in \mathcal{D}$ is

$$TRD \xrightarrow{\eta'_{TRD}} R'L'TRD \xrightarrow{R'\alpha_{RD}} R'SLRD \xrightarrow{R'S\varepsilon_D} R'SD.$$

The inverse correspondence takes β at $C \in \mathcal{C}$ to the composite

$$L'TC \xrightarrow{L'T\eta_C} L'TRLC \xrightarrow{L'\beta_{LC}} L'R'SLC \xrightarrow{\varepsilon'_{SLC}} SLC.$$

To prove these formulas inverse, first use $L' \dashv R'$ componentwise to transpose α to $\delta\colon T \Rightarrow R'SL$. Naturality of the hom-set bijection makes the transposed components natural. Put $F = R'S$. The remaining correspondence sends $\delta_C\colon TC \to FLC$ to $F(\varepsilon_D)\delta_{RD}\colon TRD \to FD$. In the other direction it sends β to $\beta_{LC}T(\eta_C)$. If we start with δ, naturality at η_C gives the recovered component

$$F(\varepsilon_{LC})\delta_{RLC}T(\eta_C) = F(\varepsilon_{LC}\, L\eta_C)\delta_C = \delta_C.$$

If we start with β, its naturality at ε_D gives

$$F(\varepsilon_D)\beta_{LRD}T(\eta_{RD}) = \beta_D T(R\varepsilon_D\,\eta_{RD}) = \beta_D.$$

The last equality in each display is a triangle identity. This gives the mates bijection $\mathrm{Nat}(L'T, SL) \cong \mathrm{Nat}(TR, R'S)$ [2, Proposition 4.3.8].

C.2 Identifying the base-change map

Take a locally cartesian closed category $\mathcal{E}$, so pullback along p has both adjoints $\Sigma_p \dashv p^* \dashv \Pi_p$. As in Chapter 27, Σ_p composes the structure map of a slice object with p. Suppose that

$$\begin{array}{ccc} A' & \xrightarrow{v} & A \\ \downarrow{\scriptstyle p'} & & \downarrow{\scriptstyle p} \\ \Gamma' & \xrightarrow{u} & \Gamma \end{array}$$

is a pullback. The equality of the two composites gives an isomorphism $\alpha\colon (p')^*u^* \Rightarrow v^*p^*$ by uniqueness of pullbacks. Take its mate with respect to the right adjoints $\Pi_p, \Pi_{p'}$. At $E \to A$ this is

$$\beta_E\colon u^*\Pi_p E \longrightarrow \Pi_{p'}v^*E.$$

Its transpose under $(p')^* \dashv \Pi_{p'}$ is the composite

$$(p')^*u^*\Pi_p E \xrightarrow{\alpha_{\Pi_p E}} v^*p^*\Pi_p E \xrightarrow{v^*\varepsilon_E} v^*E.$$

The pulled-back evaluation map therefore specifies the direction and the value of β_E.

To prove it invertible, use a test object $W \to \Gamma'$. The pullback condition gives a canonical isomorphism

$$\chi_W\colon \Sigma_v(p')^*W \xrightarrow{\sim} p^*\Sigma_u W.$$

Indeed, over A the source has total object $A' \times_{\Gamma'} W$. Its projections to A and W identify it with $A \times_\Gamma W$, the total object of the target. Writing $\mathrm{Hom}_{/\Gamma'}$ and $\mathrm{Hom}_{/A}$ for hom-sets in the corresponding slices, the test of β_E factors into the following bijections.

$$\begin{aligned} \mathrm{Hom}_{/\Gamma'}(W, u^*\Pi_p E) &\cong \mathrm{Hom}_{/\Gamma}(\Sigma_u W, \Pi_p E) \\ &\cong \mathrm{Hom}_{/A}(p^*\Sigma_u W, E) \\ &\cong \mathrm{Hom}_{/A}(\Sigma_v(p')^*W, E) \\ &\cong \mathrm{Hom}_{/A'}((p')^*W, v^*E) \\ &\cong \mathrm{Hom}_{/\Gamma'}(W, \Pi_{p'}v^*E). \end{aligned}$$

Here the middle step is precomposition with χ_W. For the identification with the mate, let $f\colon W \to u^*\Pi_p E$ and let $\bar{f}\colon \Sigma_u W \to \Pi_p E$ be its transpose. The second line sends f to $\varepsilon_E p^* \bar{f}$. Under $\Sigma_v \dashv v^*$, the transpose of χ_W is $\alpha_{\Sigma_u W}$ following the pullback of the unit $W \to u^*\Sigma_u W$. This follows by taking the two projections of the total objects just described. Naturality of α at $\bar{f}$ therefore gives the fourth-line arrow

$$v^*\varepsilon_E\, \alpha_{\Pi_p E}\, (p')^* f.$$

Its transpose in the last line is $\beta_E f$, as claimed. Postcomposition with β_E is bijective on every test hom-set, so β_E is an isomorphism by Chapter 4. This is the dependent-product Beck–Chevalley comparison. The analogous mapping-space proof is in Section 61.3.

The pullback hypothesis controls this conclusion. For a commuting square in sets with $\Gamma = \Gamma' = *$, let $A = \{0, 1\}$ and let $v\colon A' = \{0\} \hookrightarrow A$ be inclusion. Pullback along the equal composites still gives α invertible. For a family $E \to A$, its mate is the projection

$$E_0 \times E_1 \longrightarrow E_0.$$

Choosing $E_0 = *$ and $E_1 = \varnothing$ makes this the map $\varnothing \to *$. The two fiber products used to define χ_W now represent different families.

C.3 The norm on invariants and coinvariants

Let G be a finite group, let k be a commutative ring, and let M be a $k[G]$-module. The inclusion of G-invariants and the quotient to G-coinvariants are the linear versions of the fixed-point and orbit constructions in Section 9.4. Write

$$M^G = \{m \in M : gm = m \text{ for each } g \in G\}, \qquad M_G = M/\langle gm - m\rangle_k.$$

Let triv give a k-module the trivial G-action. An equivariant map $M \to \mathrm{triv}(Q)$ kills every $gm - m$, so factors uniquely through M_G. An equivariant map $\mathrm{triv}(Q) \to M$ takes values in M^G, so factors uniquely through the inclusion. The factorizations identify the two adjunctions $(-)_G \dashv \mathrm{triv} \dashv (-)^G$. They also give the comparison $j\colon M^G \to M_G$ by composing the inclusion with the quotient.

The map $N\colon M_G \to M^G$ given by $N([m]) = \sum_{g\in G} gm$ uses both factorizations. Multiplying the sum on the left by an element of G permutes its terms, so it is invariant. Replacing m by hm has the same effect by right multiplication on the summation index. Thus N is well-defined. The following square commutes.

$$\begin{array}{ccc} M & \xrightarrow{\sum_g g} & M \\ \downarrow{\scriptstyle q} & & \uparrow{\scriptstyle i} \\ M_G & \xrightarrow{N} & M^G \end{array}$$

For invariant m, every summand is m, giving $Nj(m) = |G|m$. For any m, every gm represents $[m]$, giving $jN([m]) = |G|[m]$. Both j and N are natural because equivariant linear maps commute with the action of each summand. When $|G|$ is a unit in k, j is an isomorphism with inverse $|G|^{-1}N$. The averaging projection of Section 9.8.4 is $i\,|G|^{-1}Nq$. On a permutation module it first takes the orbit class and then recovers the invariant vector with that orbit class.

There is an exactness consequence. The invariants functor is additive and preserves kernels as a right adjoint. The coinvariants functor is additive and preserves cokernels as a left adjoint. When $|G|$ is a unit the natural isomorphism therefore makes both functors exact. In particular, for a surjective equivariant map $M \to P$, an invariant element of P can be lifted and then averaged to give an invariant lift.

For a modular example let $k = \mathbb{F}_p$, $G = C_p$, and $M = k[G]$ with its left regular action. The translates of a basis element run through the whole basis, so M_G has basis the class $\overline{1}$ of the identity element. The invariant vectors are constant on the same basis orbit, so M^G has basis $s = \sum_{g \in G}[g]$, where $[g]$ denotes the group-algebra basis element. The norm sends the coinvariant class of 1 to s, while $j(s) = p\overline{1} = 0$ in M_G. For the trivial module k, j is an isomorphism and N is zero. In particular neither direction is an isomorphism uniformly in the module.

The augmentation $k[G] \to k$ sends each basis element to 1. It is a surjective equivariant map for the trivial action on k. On invariants it sends s to $p = 0$, so its image is zero in $k^G = k$. This exhibits failure of right exactness of invariants when the group order vanishes in the coefficient field.

C.4 A derived intersection in several coordinates

Chapter 62 computes the self-intersection of the origin in a line from a two-term resolution. For the origin in affine d-space over a field k, put $R = k[t_1, \dots, t_d]$ and use the augmentation $R \to k$ sending each t_i to zero. The derived affine pullback is represented by $D = k \otimes^{\mathbf{L}}_R k$. We compute $\pi_i(D)$ from a resolution of one copy of k, using the comparison with Tor in Section 62.3.

Use homological degrees. For t_i, let $K(t_i)$ be the complex $Re_i \xrightarrow{t_i} R$ in degrees $1, 0$, where the label means $e_i \mapsto t_i$. Set $K^{(r)} = K(t_1) \otimes_R \cdots \otimes_R K(t_r)$, with $K^{(0)} = R$ in degree zero. The tensor differential sends $x \otimes y$ to $dx \otimes y + (-1)^{\deg x} x \otimes dy$ for homogeneous x. In degree i the basis of $K^{(r)}$ is indexed by subsets $j_1 < \cdots < j_i$ of $\{1, \dots, r\}$. Writing the associated basis element as $e_{j_1} \wedge \cdots \wedge e_{j_i}$, its differential is

$$d(e_{j_1} \wedge \cdots \wedge e_{j_i}) = \sum_{a=1}^{i} (-1)^{a-1} t_{j_a} e_{j_1} \wedge \cdots \wedge \widehat{e_{j_a}} \wedge \cdots \wedge e_{j_i}.$$

The hat means omission. This is the Koszul complex on the first r coordinate functions.

Here is the exactness check. The first summand of $K_n^{(r)}$ is $K_n^{(r-1)}$ and comes from tensoring with the degree-zero copy of R in $K(t_r)$. The second summand, tensored with e_r, is $K_{n-1}^{(r-1)}$. In the quotient complex by the first summand, the second summand keeps its internal differential. Its homology in degree n is $H_{n-1}(K^{(r-1)})$. The boundary in the long exact sequence from this quotient to the subcomplex is multiplication by $(-1)^{n-1}t_r$. Indeed a cycle x in the second summand lifts as $x \otimes e_r$, and its differential in the first summand is $(-1)^{n-1}t_r x$. If $K^{(r-1)}$ has homology only in degree zero, put $Q = R/(t_1, \ldots, t_{r-1})$ for that homology. All degrees above one for $K^{(r)}$ then vanish by exactness, and the remaining part of the sequence is

$$0 \longrightarrow H_1(K^{(r)}) \longrightarrow Q \xrightarrow{-t_r} Q \longrightarrow H_0(K^{(r)}) \longrightarrow 0.$$

Multiplication by t_r is injective on $Q = k[t_r, \ldots, t_d]$ by independence of the monomials. The cokernel is $R/(t_1, \ldots, t_r)$. Starting at $r = 0$, induction proves that $K^{(d)}$ resolves k by free R-modules. This is the regular-sequence exactness argument in this particular polynomial ring [59, Lemma 15.31.2].

After tensoring the resolution with k, each t_i acts by zero. The differential vanishes in every degree. Thus $\pi_i(D) \cong \operatorname{Tor}_i^R(k, k)$ is the k-vector space with basis the i-element subsets of the d coordinates. Its dimension is $\binom{d}{i}$ for $0 \le i \le d$, and it vanishes for $i > d$. In degree one, the classes of the e_i correspond to the classes of t_i in $\mathfrak{m}/\mathfrak{m}^2$, where $\mathfrak{m} = (t_1, \ldots, t_d)$. This agrees with the conormal description of π_1 in Section 62.3. The basis transforms under a linear coordinate change $t_i' = \sum_j a_{ij} t_j$ by the same matrix on the generators e_i. The indicated map on the generators extends to a chain isomorphism on the exterior powers because it commutes with the differential. In top degree the comparison acts by the determinant.

For $d = 2$, the resolution is

$$0 \longrightarrow R \xrightarrow{\binom{-t_2}{t_1}} R^2 \xrightarrow{(t_1 \;\; t_2)} R \longrightarrow k \longrightarrow 0.$$

The ordinary self-intersection has coordinate ring k. The derived one has $\pi_0 = k$, $\pi_1 = k^2$, and $\pi_2 = k$ in the given coordinates. Its degree-two class comes from the generator $e_1 \wedge e_2$, whose differential records the relation between the two coordinate equations before tensoring.

APPENDIX D

Index of universal properties and their diagrams

The entries are grouped by construction. Each gives the mapping test, adjunction, or factorization used in the text. Select the page reference to open its treatment and any accompanying diagram.

D.1 Maps, limits, and colimits

Product § 2.1, p. 6

A map $T \to A \times B$ is uniquely determined by its two projections, and every pair of maps out of T occurs. The empty product is terminal.

Coproduct Def. 2.4, p. 7

A map $A \amalg B \to T$ is uniquely determined by its two restrictions, and every pair of maps into T occurs. The empty coproduct is initial.

Equalizer § 3.1, p. 12

Maps $T \to X$ on which $f, g\colon X \rightrightarrows Y$ agree factor uniquely through the equalizer $E \to X$.

Coequalizer § 3.4, p. 15

Maps $h\colon Y \to T$ with $hf = hg$ for $f, g\colon X \rightrightarrows Y$ factor uniquely through $Y \to Q$.

Pullback § 3.2, p. 13

A map $T \to A \times_C B$ corresponds to a pair of maps $T \to A$ and $T \to B$ with equal composites to C.

Pushout § 3.5, p. 16

A map $A \amalg_C B \to T$ corresponds to a pair of maps $A \to T$ and $B \to T$ with equal restrictions to C.

Yoneda evaluation Thm. 4.3, p. 22

A natural map $\mathrm{Hom}_{\mathcal{C}}(-, C) \to F$, for a presheaf F, is determined by the image of id_C in $F(C)$.

Universal elements § 4.4.2, p. 23

An element $u \in F(C)$ of a covariant functor represents F when every $x \in F(T)$ has a unique expression $F(f)(u)$ for $f\colon C \to T$.

Free monoid Prop. 5.2, p. 27

A function from an alphabet to a monoid extends uniquely to a monoid map by evaluation on words.

D.2 Algebra and symmetry

Natural numbers object § 6.1, p. 41

Given a zero $x_0\colon 1 \to X$ and a step $t\colon X \to X$, one map $u\colon N \to X$ preserves zero and successor.

Group completion Thm. 7.4, p. 48

Every monoid homomorphism from a commutative monoid to an abelian group extends uniquely along the map to its group of formal differences.

Free group Prop. 8.2, p. 54

A function from a set of generators to the underlying set of a group extends uniquely to a group homomorphism.

Orbits and fixed points § 9.4, p. 64

The orbit quotient receives invariant maps from the action, while the fixed-point inclusion receives maps into the action from trivial actions. Thus $(-)/G \dashv \Delta \dashv (-)^G$. See also the Burnside trace derivation, § 9.8.4, p. 72.

Free associative algebra Prop. 10.3, p. 75

Sending the variables of $k\langle S\rangle$ to arbitrary elements of a k-algebra determines one k-algebra map.

Localization of rings Thm. 11.8, p. 85

A ring map $R \to A$ taking S to units extends uniquely along $R \to S^{-1}R$.

Free module Prop. 12.3, p. 89

A function $S \to U(M)$ extends uniquely to an R-linear map $R^{(S)} \to M$, where U takes the underlying set.

Tensor product § 14.1, p. 99

For modules over a commutative ring, linear maps $M \otimes_R N \to P$ correspond to bilinear maps $M \times N \to P$.

Extension of scalars Prop. 15.1, p. 107

For $R \to S$, the extension $S \otimes_R -$ is left adjoint to restriction of S-modules to R-modules.

Character pairing adjunction § 5.6, p. 39

Homomorphisms $G \to DH$ and $H \to DG$ correspond to pairings $G \times H \to \mathbb{T}$ additive in both entries. Its evaluation unit is invertible for finite abelian groups.

D.3 Topology, analysis, and learning

Pontryagin duality § 16.6, p. 117

Sending a locally compact abelian group to its continuous character group gives a contravariant equivalence via evaluation. The dual of a quotient is the annihilator of the subgroup.

Initial and final topologies § 16.2, p. 113

A map into the initial topology is continuous exactly when its composites with the defining maps are continuous. For the final topology, § 16.3, p. 114, the test is on maps out of the space.

Open-set classifier § 16.5, p. 117

Continuous maps from X to the Sierpiński space correspond to open subsets of X by inverse image of $\{1\}$.

Opens and points Thm. 17.2, p. 122

A continuous map $X \to \mathrm{Pt}(L)$ corresponds to a frame map $L \to \mathcal{O}(X)$.

Metric completion Thm. 18.5, p. 129

A 1-Lipschitz map $X \to Y$ to a complete metric space extends uniquely to a 1-Lipschitz map $\widehat{X} \to Y$.

Finite-sample representers § 38.4, p. 253

The adjoint of evaluation T synthesizes sample features. The Gram matrix factors as TT^*. For a sample loss with a strictly increasing norm penalty, every minimizer lies in the image of T^*.

Sharing a parameter § 39.2, p. 259

Reindexing along $\Delta_P = \langle \mathrm{id}_P, \mathrm{id}_P \rangle$ makes two parameter slots use a single value from P.

D.4 Logic and type formation

Subobject classifier § 23.1, p. 160

Each subobject of X is the pullback of truth $1 \to \Omega$ along a unique characteristic map $X \to \Omega$.

Logic and semantics § 22.1, p. 152

$K \subseteq \mathrm{Mod}(T)$ exactly when $T \subseteq \mathrm{Th}(K)$. The structure–semantics adjunction, § 22.4, p. 156 exchanges interpretations over **Set** with maps of theories.

Quantifiers § 24.1, p. 166

Existential and universal quantification give the left and right adjoints to inverse image of predicates. The pullback square, § 24.2, p. 167 describes their behavior under substitution.

Kan extensions § 25.2, p. 173

Left and right Kan extension are adjoint to restriction on diagrams. Where the required limits and colimits exist, compute them pointwise over the respective comma categories.

Exponential § 26.1, p. 178

Transposition exchanges maps $T \times A \to B$ with maps $T \to B^A$. Evaluation recovers the original map.

Dependent sums and products § 27.5, p. 188

In a locally cartesian closed category, pullback p^* between slices has left adjoint Σ_p and right adjoint Π_p. The base-change diagram, § 27.6, p. 188 compares binding with substitution.

Inductive type § 28.1, p. 190

An initial F-algebra has one algebra map into each target algebra. Its commuting square gives the recursion equations.

D.5 Recursion and behavior

Expression syntax § 31.1, p. 208

An interpretation of constructors as an algebra extends uniquely to a fold on well-formed expressions.

Least fixed point Thm. 32.2, p. 217

Starting from the bottom element and iterating a continuous map on an ω-complete partial order produces the least fixed point as their supremum. It is the initial algebra in that order.

Final coalgebra § 34.2, p. 227

A coalgebra with output in O and a transition has one coalgebra map to $O^{\mathbb{N}}$, taking a state to its output stream.

D.6 Local objects and derived algebra

Sheafification Thm. 44.4, p. 288

Maps from a presheaf to a sheaf factor uniquely through the sheafification unit.

Bundle descent Thm. 46.2, p. 299

Bundles over an open union are equivalent to local bundles equipped with overlap isomorphisms satisfying the cocycle law.

Affine base change § 47.4, p. 306

Maps into the fiber product of affines correspond to compatible maps into both factors. In coordinate rings, the fiber product is represented by the tensor product over the base ring.

Classifying topos § 49.2, p. 318

For a Grothendieck topos $\mathcal{E}$, pulling back the universal model gives $\mathrm{Geom}(\mathcal{E}, \mathcal{B}_{\mathbb{T}}) \simeq \mathrm{Mod}_{\mathbb{T}}(\mathcal{E})$.

Homology § 51.1, p. 328

Maps out of the cycle object that vanish on boundaries factor uniquely through homology.

Derived localization § 52.4, p. 338

For bounded-above complexes, the derived category universally inverts the quasi-isomorphisms of the homotopy category.

D.7 Spaces and higher geometry

Fundamental groupoid of a union Thm. 54.2, p. 353

With the basepoints specified in the theorem, the fundamental groupoid of an open union is the pushout of the local groupoids along the groupoid of the intersection.

Homotopy pullback § 57.2, p. 371

A map into a homotopy pullback supplies two component maps and a path between their composites to the base.

Homotopy pushout § 57.4, p. 373
Maps out of the double mapping cylinder supply two maps and a homotopy between the two restrictions to the overlap.

Suspension spectrum § 58.3, p. 376
Maps of spectra $\Sigma^\infty X \to E$ form the space of pointed maps $X \to \Omega^\infty E$.

Čech descent § 59.2, p. 383
A sheaf of spaces takes a covering nerve to a limit of local sections, their comparisons, and higher compatibility data.

Stack descent § 60.2, p. 391
The nerve of the stack on the union of two opens is the homotopy pullback of the nerves on the two opens over the overlap.

Univalent family § 61.5, p. 401
Maps into the universe classify small families. Its paths between types correspond to equivalences of fibers.

Higher inductive circle § 61.6, p. 402
A map out of S^1 is specified by a point of the target and a loop at that point.

Derived affine intersection § 62.2, p. 405
A map from $A \otimes^{\mathbf{L}}_R B$ into a test algebra supplies maps from A and B with a path between their restrictions to R.

Cotangent complex § 62.5, p. 407
At a k-point, maps from the pulled-back cotangent complex to a connective k-module M represent derived lifts of the point to $k \oplus M$.

APPENDIX E

Index of symbols and notation

Symbols are grouped by subject. When a symbol carries several meanings, its entries give the relevant contexts. Select a page number to open the corresponding passage.

Symbol	Meaning	Page
Arrows, diagrams, and categories		
$f\colon X \to Y$	An arrow with source X and target Y.	410
$x \mapsto f(x)$	The value assigned to x by a function.	410
$g \circ f$	Composition that applies f first and then g.	410
id_X	The identity arrow of X.	1
$X \hookrightarrow Y$	An inclusion or a monomorphism, depending on the category.	410, 146
$X \twoheadrightarrow Y$	A surjection or an epimorphism, depending on the category.	410, 411
$X \dashrightarrow Y$	A factorization or filler whose existence is stated in the accompanying property.	410, 27
$\lrcorner$	The corner mark that denotes a pullback square.	410, 13
$X \rightrightarrows Y$	A pair of parallel arrows from X to Y.	12, 15
$F \Rightarrow G$	A natural transformation between functors F and G.	21
$X \cong Y$	An isomorphism between objects, including a natural isomorphism between functors.	411, 21
$\mathcal{C}^{\mathrm{op}}$	The category with the same objects as $\mathcal{C}$ and arrows reversed.	411
$\mathcal{K}^{\mathrm{co}}$	The 2-category formed by reversing the 2-cells of $\mathcal{K}$.	412
$\mathrm{Hom}_{\mathcal{C}}(X, Y)$	The collection of arrows $X \to Y$, also written $\mathcal{C}(X, Y)$.	1, 412
$\mathrm{Nat}(F, G)$	The set of natural transformations from F to G.	416
Set, **FinSet**	The categories of sets and finite sets with functions as arrows.	1

Symbol	Meaning	Page
$[\mathcal{A}, \mathcal{E}]$	The category of functors $\mathcal{A} \to \mathcal{E}$ and natural transformations.	172
$0,\ 1$	An initial object and a terminal object, respectively.	8
$X \times Y$	The product, with a projection to each factor.	6
$X \amalg Y$	The coproduct, with a map from each summand.	7
$\langle a, b\rangle,\ [a, b]$	The pairing into a product and the copairing out of a coproduct.	6, 7
$X \times_Z Y$	The pullback of the given maps $X \to Z$ and $Y \to Z$.	13
$X \amalg_Z Y$	The pushout of the given maps $Z \to X$ and $Z \to Y$.	16
$\Delta_X,\ \Delta T$	The diagonal $X \to X \times X$ and the diagram constant at T, respectively.	14, 411
$h_A = \mathcal{C}(-, A)$	The presheaf represented by A, with restriction by precomposition.	21
$L \dashv R$	The functor L is left adjoint to R.	29
$\eta,\ \varepsilon$	The unit and counit of an adjunction.	30
$\lim D,\ \operatorname{colim} D$	The vertex of the universal cone and the vertex of the universal cocone, respectively.	17, 411
$\mathcal{U},\ \mathcal{V}$	Grothendieck universes, with $\mathcal{V}$ large enough to contain $\mathcal{U}$.	415
$\prod_i X_i,\ \coprod_i X_i$	The product and coproduct of an indexed family.	417
Algebra, spaces, and probability		
A^*	The free monoid of finite words in the alphabet A.	27
$K(M)$	The abelian group completion of the commutative monoid M.	47
$\langle S \mid R\rangle$	A group presented by generators S and relations R.	55
$BG,\ G \ltimes X$	The one-object category of G and the groupoid of its action on X.	61, 63
$Gx,\ G_x$	The orbit and stabilizer of a point x.	62
$X^G,\ X^g$	The points fixed by all of G, or by the individual element g.	64, 65
$X/G,\ G/H$	The orbit set of X and the set of left cosets of H in G.	64, 62
$k[G]$	The group algebra with basis $[g]$ and multiplication $[g][h] = [gh]$.	70

Symbol	Meaning	Page
R-**Mod**	The category of unital left modules over R.	87
$k\langle S\rangle,\ k[S]$	The free associative and free commutative algebras on S, respectively.	75, 76
$S^{-1}R$	The localization of R that makes every element of S a unit.	84
R/I	The ring obtained by setting every element of the ideal I equal to zero.	81
$\ker f,\ \operatorname{coker} f$	The kernel and cokernel of f.	92
$\bigoplus_i M_i$	The direct sum of modules, with tuples of finite support.	90
$R^{(S)}$	The free R-module with basis indexed by S.	89
$M \otimes_R N$	The tensor product, representing R-bilinear maps out of $M \times N$ for commutative R.	99
$M^\vee,\ \bigwedge^n M$	The linear dual and the nth exterior power of M.	102, 104
$\mathcal{O}(X),\ \mathrm{Pt}(L)$	The frame of opens of X and the space of points of the frame L.	120, 121
$\mathbb{T},\ \mathbb{T}[n]$	The circle group $\mathbb{R}/\mathbb{Z}$ and its subgroup killed by n.	59, 117
$\widehat{G},\ \eta_G$	Continuous character group and its natural evaluation map $G \to \widehat{\widehat{G}}$.	117
$H^\perp$	Annihilator of H, consisting of characters vanishing on H.	117, 159
$\widehat{X}$	The metric completion of X.	128
$Df_x,\ Tf$	The derivative of f at x and the induced map on tangent bundles.	133, 134
$\mathcal{D}X,\ \delta_x$	The finitely supported probability distributions on X and the point mass at x.	139
$k(y \mid x),\ \ell \odot k$	The conditional weight in a stochastic map and composition by mixing.	140
Logic and types		
$\mathrm{Sub}(X),\ \mathcal{C}/X$	The poset of subobjects of X and the slice category of arrows into X.	146
f^*	Reindexing by pullback. For subsets and open sets it acts by inverse image.	148, 120
$\top,\ \bot$	Truth and falsity, given by the greatest and least subobjects in their context.	147, 162
$P \wedge Q,\ P \vee Q$	The conjunction and disjunction of predicates, given by meet and join.	147, 162

Symbol	Meaning	Page
$P \Rightarrow Q,\ \neg P$	Implication and negation. Negation is implication to falsity.	161, 162
$m \models s$	The interpretation m satisfies the sentence s.	152
$\mathrm{Mod}(T),\ \mathrm{Th}(K)$	The models of a theory T and the sentences true in all models in K.	152
$\mathrm{Str}(U),\ \mathrm{Sem}(L)$	The theory of natural operations of U and the model category of L with its forgetful functor to sets.	155, 156
$\Omega,\ \chi_m$	A subobject classifier and the characteristic map of the monomorphism m.	160
$B^A,\ \mathrm{ev}$	An exponential and its evaluation map $B^A \times A \to B$.	160, 178
$\exists_f,\ \forall_f$	Existential and universal quantification, the left and right adjoints to f^*.	166, 168, 169
J^*	Restriction of diagrams by precomposition with J.	172
$\mathrm{Lan}_J,\ \mathrm{Ran}_J$	Left and right Kan extension along J.	172, 173
$\Gamma \vdash t : A$	Judgment that t has type A in context Γ.	179
$\lambda x.t$	Abstraction that binds x in t.	179
$t[\sigma]$	Term obtained by substituting the entries of σ into t.	179, 184
$\Gamma.A$	Context extended by a variable of type A over Γ.	184
$\mathrm{Fam}(\Gamma)$	Category of families over Γ and fiberwise maps.	184
$W,\ \sup(a, u)$	Type of well-founded trees and the constructor with label a and subtrees u.	193
$\Sigma_{x:A} B$	Dependent sum with entries $a : A$ and $b : B(a)$.	185, 398
$\Pi_{x:A} B$	Dependent product whose terms give an element of $B(a)$ for each $a : A$.	186, 398
$\mathrm{Id}_A(x, y),\ x =_A y$	Type of identifications of x and y in A.	197, 399
$t \equiv u$	Judgmental equality used by the conversion rules for terms and types.	197
$\mathrm{refl}_x,\ J_C$	Reflexivity constructor and identity eliminator with motive C.	197, 399
tr_p^B	Transport in the family B along the path p.	198, 399
$\mathcal{U}_i,\ \mathrm{El}_i(a)$	Universe of type codes at level i and the type decoded from a.	201
$\mathrm{idtoequiv},\ \mathsf{ua}(e)$	Transport from universe paths to equivalences, and its inverse in a univalent universe.	201, 401

Symbol	Meaning	Page
$\widetilde{\mathcal{U}} \to \mathcal{U}$	Universal small family of spaces over the universe of types.	401
Computation and learning		
fold_α	Interpretation of expression trees in the algebra with structure map α.	208, 191
$[\![e]\!]_\rho$	Integer denoted by expression e in environment ρ.	209, 210
$\bot,\ X_\bot$	Least element for an undefined result, and the flat lift that adjoins it to X.	215
$\mathrm{fix}(F)$	Least fixed point of a continuous map on a pointed ω-CPO.	216
(T, η, μ)	Monad with endofunctor T, unit η, and multiplication μ.	221
$f^\sharp$	Kleisli extension sending a computation in TA to one in TB for $f : A \to TB$.	222
$\gg=$	Bind, which sequences an effectful value with the next computation.	222
$\mathrm{Kl}(T)$	Kleisli category for T, with arrows represented by maps $A \to TB$.	222
$f(x)\downarrow,\ f(x)\uparrow$	Defined and undefined application of a partial function, respectively.	233
φ_e	Partial computable function with program number e.	234
$\mathcal{K}_1$	Kleene's first partial combinatory algebra, with application given by running a numbered program.	239
$P \vdash_X Q$	Entailment tracked by a single program for all instances over X.	239
$\widehat{p}_z$	Empirical law of sample z, obtained by assigning equal mass to its indices.	246, 278
$R_p(h),\ \widehat{R}_z(h)$	Mean loss of h under the population law p or on the sample z.	247, 278
$k(x, z)$	Kernel evaluated at inputs x, z, represented by the inner product of their features.	253
$K_0 = J_0 J_0^*$	Empirical neural tangent kernel matrix at initialization.	255
$\mathrm{Para}(\mathcal{C})$	Category of parameterized maps in a category $\mathcal{C}$ with finite products.	258
A_{ij}	Attention weight assigning part of the query i to key position j.	261

Symbol	Meaning	Page
$R[f]_x$	Reverse derivative, the dual of Df_x acting on covectors.	263
$\nabla L(w),\ G_\eta(w)$	Euclidean gradient and the gradient update with stepsize η.	268
$p_y(h)$	Posterior mass of hypothesis h conditional on observation y.	272
Descent, homology, and higher geometry		
$F_x,\ [s]_x$	Stalk of a presheaf at x and the germ of a section s there.	286
$\mathrm{Sh}(\mathcal{C}, J)$	Category of set-valued sheaves on the site $(\mathcal{C}, J)$.	292
F^+	Matching-family presheaf identifying data that agree after refining the cover.	293
$f^* \dashv f_*$	Inverse image and direct image of a geometric morphism.	296
θ_{ij}	Transition isomorphism from local bundle j to local bundle i on their overlap.	298
$\mathrm{Spec}\, A$	Affine spectrum of a commutative ring A, with its prime-ideal points and structure sheaf.	303, 304
$D_A(f)$	Distinguished open of $\mathrm{Spec}\, A$ consisting of primes that omit f.	303
G_k	Absolute Galois group of a field k, computed using a chosen separable closure.	310
$\mathcal{B}_{\mathbb{T}},\ U_{\mathbb{T}}$	Classifying topos and universal model of the geometric theory $\mathbb{T}$.	318
$Z_n(C),\ B_n(C)$	Cycles and boundaries of a chain complex C in degree n.	328
$H_n(C),\ H^n(K)$	Degree n homology of a chain complex and cohomology of a cochain complex.	328, 334
$X[r],\ C(f)$	Cochain shift by r and the mapping cone of a cochain map f.	336
$\mathbf{D}^-(R)$	Derived category of bounded-above complexes of R-modules.	338
$\otimes^{\mathbf{L}}_R$	Derived tensor product over R.	340, 405
$\mathrm{Tor}^R_i(A, B)$	Degree i homology of the derived tensor of two R-modules.	340, 406
$\Gamma(X, -),\ R\Gamma(X, -)$	Global sections and the right derived global sections functor on X.	343, 345
$H^q(X, \mathcal{F})$	Degree q sheaf cohomology of $\mathcal{F}$ on X.	345
$\Pi_1(X)$	Fundamental groupoid of a space, with points as objects and path classes as arrows.	350, 367

Symbol	Meaning	Page
$\pi_n(X, x)$	Degree n homotopy group of X based at x, for $n \geq 1$.	362, 378
$\mathbf{\Delta},\ \Delta^n$	Simplex category of finite nonempty ordinals, and the simplicial set represented by $[n]$.	363
Λ^n_i	Union of the faces of the standard n-simplex except the face opposite i.	363
$\mathrm{N}G,\ \mathrm{Sing}\, X$	Nerve of a category G and singular simplicial set of a topological space X.	367, 363
$\mathrm{Map}_{\mathcal{S}}(X, Y)$	Mapping space between spaces X and Y.	364, 370
$X \times^{\mathrm{h}}_Z Y$	Homotopy pullback of the maps from X and Y to Z.	371
$B \amalg^{\mathrm{h}}_A C$	Homotopy pushout of the maps from A to B and C.	373
$\Sigma X,\ \Omega Y$	Reduced suspension of a pointed space X and based loop space of Y.	373, 375
Sp	Infinity-category of spectra built from compatible loop equivalences.	375
$\Sigma^\infty X,\ \Omega^\infty E$	Suspension spectrum of X and zeroth space of a spectrum E.	376, 375
$\mathrm{Shv}_{\mathcal{S}}(\mathcal{C})$	Infinity-category of space-valued sheaves on $\mathcal{C}$.	382
$\mathcal{X}^\wedge$	Hypercompletion of the infinity-topos $\mathcal{X}$.	388
$[X/G]$	Quotient stack of a group action, tested by torsors equipped with an equivariant map to X.	394
$\mathrm{Spec}^{\mathrm{der}}\, A$	Derived affine spectrum represented by the simplicial commutative algebra A.	405
$\mathrm{Der}_{k,a}(D, M)$	Space of derived derivations of D into M at the augmentation $a : D \to k$.	407
$X \simeq Y$	An equivalence of spaces or categories, according to context.	366, 375
$\pi_0(X)$	The set of path components of a space.	362
$\tau_{\leq n} X$	The truncation of an object in degrees at most n.	367, 386
$\mathbb{S}$	The sphere spectrum, obtained by stabilizing the pointed zero-sphere.	378

APPENDIX F

Index of concepts and definitions

Entries are filed alphabetically. The short descriptions identify each concept. Select a page number for its definition, statement, or treatment in the text.

Concept	Definition or use	Page
A		
Abelian category	An additive category with kernels and cokernels where every canonical coimage-to-image map is invertible.	323
Absolute Galois group	The automorphism group of a separable closure over the ground field, with its profinite topology.	310
Action groupoid	Objects are points. Arrows $x \to y$ are the group elements sending x to y.	63
Adjunction	Natural bijections $\mathrm{Hom}(FX, Y) \cong \mathrm{Hom}(X, GY)$ for a pair of functors.	29
Affine scheme	The prime spectrum of a commutative ring with its Zariski topology and structure sheaf.	303, 304
Algebra action	An action given by an algebra homomorphism into linear endomorphisms.	68
Algebra over a commutative ring	A unital ring equipped with a scalar map from the base ring into its center.	74
Annihilator	Characters that vanish on a subgroup.	159, 117
Assembly	A set with a nonempty set of numerical realizers assigned to each element.	241
Attention	A weighted average of values whose weights are determined by queries and keys.	261
B		
Backpropagation	Reverse differentiation on a computation graph, accumulating covectors at shared nodes.	265
Bayes' rule	The reverse conditional law obtained by dividing the joint mass by a positive marginal mass.	142

Concept	Definition or use	Page
Beck–Chevalley condition	The comparison for changing base in a pullback square is invertible.	167, 188, 421
Beta and eta laws	Application computes after abstraction, and abstracting application to a fresh variable recovers the function.	180
Bicategory	Objects and hom categories with composition functors and coherent associativity and unit isomorphisms.	359
Bimodule	An abelian group with commuting left and right scalar actions.	88
Biproduct	A finite direct sum that serves as both a product and a coproduct.	90
Bisimulation	A relation pairing states with matching observations and related successors.	229
Brown–Comenetz dual	The spectrum $F(X, I_{\mathbb{Q}/\mathbb{Z}})$ with character groups in reversed homotopy degrees.	380
Burnside's lemma	For a finite group acting on a finite set, the orbit count is the average fixed-point count.	65, 72
C		
Cartesian closed category	A category with finite products and an exponential B^A for each pair of objects.	178
Catalan numbers	Counts of ordered full binary trees graded by their internal vertices.	195
Category	Objects and arrows with associative composition and identities.	1
Cauchy sequence	A sequence whose sufficiently late terms are pairwise within any prescribed positive distance.	126, 128
Čech nerve	The simplicial object of iterated fiber products of a map over its target.	383
Chain complex	A sequence with differentials lowering degree whose consecutive composites are zero.	328
Chain homotopy	A degree-raising family expressing a difference of chain maps as $dh + hd$.	329
Character	A homomorphism into the circle group, continuous for topological groups.	59, 117
Chinese remainder theorem	For comaximal ideals, the quotient by their intersection is the product of the separate quotients.	83
Classical truncation	The ordinary affine scheme obtained by taking the zeroth homotopy ring of a derived affine.	405
Classifying stack	The stack of torsors for a fixed group.	394

Concept	Definition or use	Page
Classifying topos	Represents the categories of models of a geometric theory by geometric morphisms into it.	318
Coalgebra	An object X with a structure map $X \to FX$ that exposes one step of behavior.	226
Cochain complex	A sequence with differentials raising degree whose consecutive composites are zero.	334
Coequalizer	Universal arrow out of the target that identifies two parallel maps.	15
Coextension of scalars	The right adjoint to restriction of scalars, given by a module hom functor.	108
Cofiber	In a pointed ∞-category, the pushout of a map along the map from its source to zero.	377
Cokernel	Universal map out of the target of a linear map that sends its image to zero.	92
Colimit	A cocone through which every other cocone from the diagram factors uniquely.	17
Commutative diagram	A diagram in which all parallel routes under comparison have equal composites.	410
Compiler correctness	Compiled code pushes the value of the source expression onto the input stack.	210
Conditional independence	The joint observation law factors as a product at each fixed hypothesis.	273
Constructive choice	Pointwise pairs of witnesses and proofs yield a witnessing function by first projection.	207
Context extension	The context of a family together with its variable, given by the total space over the base.	184
Coproduct	An object with injections universal among families of maps out of its summands.	7
Cotangent complex	The complex representing derived derivations into square-zero extensions.	407
Curry–Howard correspondence	Propositions are read as types and their proofs as terms of those types.	203, 204
D		
Dependent product	The type of sections of a family, with one output in the fiber at each input.	186, 188
Dependent sum	The type of pairs (a, b) with b in the fiber over a.	185, 188
Dependent type	A type family over a context, interpreted by an object in a slice category.	184, 188

Concept	Definition or use	Page
Derivative	The linear part of a map at a point, with a remainder of smaller order than the displacement.	133
Derived category	The localization of the homotopy category of complexes at the quasi-isomorphisms.	338
Derived fiber product	The homotopy pullback of derived schemes, computed affinely by the derived tensor product.	405
Derived tensor product	Tensor product computed after a suitable projective or flat replacement.	340, 406
Descent datum	Local objects with overlap isomorphisms satisfying identity and cocycle laws.	297, 298
Determinant	The scalar action of an endomorphism on the top exterior power of a finite free module.	105
Disintegration	Factorization of a joint distribution into a marginal and a conditional law.	142
Distinguished triangle	In the derived category, a triangle isomorphic to a mapping-cone triangle.	339
Distribution shift	A difference between the law used to draw training samples and the law at deployment.	282
Dual module	The module of linear maps from the given module to its scalar ring.	102
Dual numbers	Scalars with an infinitesimal generator ϵ satisfying $\epsilon^2 = 0$.	136
E		
Effective descent	Restriction from global objects to their category of descent data is an equivalence.	297, 299
Effective topos	The regular exact completion of assemblies over Kleene's first partial combinatory algebra.	243
Elementary topos	A category with finite limits, exponentials, and a subobject classifier.	160
Empirical distribution	The uniform law on sample positions, sent to the space of observations.	246
Empirical risk	The average loss across the observations in the training sample.	247
Empirical risk minimization	Selection of a predictor that minimizes empirical risk within a class.	249
Epimorphism	A map e such that $u \circ e = v \circ e$ implies $u = v$.	4
Equalizer	Universal arrow into the source on which two parallel maps agree.	12
Étale fundamental group	The automorphism group of the chosen geometric fiber functor on finite étale covers.	314

Concept	Definition or use	Page
Exact completion	The construction that adds effective quotients of internal equivalence relations.	243
Exact functor	An additive functor between abelian categories that preserves kernels and cokernels.	326
Existential quantification	The left adjoint to reindexing predicates, which takes the image along the map.	166, 170
Extension of scalars	Tensoring with the new scalar ring, left adjoint to restriction of scalars.	107
Exterior power	The module representing alternating multilinear maps of a specified degree.	104
F		
Fiber functor	Assigns to a finite étale cover its fiber over a chosen geometric point.	310, 314
Final coalgebra	A coalgebra admitting a unique morphism from each coalgebra for the same endofunctor.	227, 228
Final topology	The finest topology making the specified incoming maps continuous.	114
Finite étale algebra	Over a field, a finite product of finite separable extensions of that field.	309
Finite-sample representer property	For a sample-dependent loss with strictly increasing norm penalty, minimizers lie in the sample feature span.	253
Fixed point	A point fixed by every element of the acting group.	64
Fixed-point fusion	A strict continuous map intertwining endomaps carries the first least fixed point to the second.	218
Fold	The unique algebra morphism interpreting the constructors of an initial algebra.	191, 208
Frame	A complete lattice in which finite meets distribute over arbitrary joins.	120
Free associative algebra	The algebra spanned by words, universal for assignments to its generators.	75
Free cartesian closed category	The category of terms modulo equations freely interpreting the declared base types.	180
Free group	The group universal for assignments from a set of generators.	54
Free module	The module universal for arbitrary assignments of its basis to another module.	89
Free monoid	The monoid of words on an alphabet, with concatenation and the empty word.	27

Concept	Definition or use	Page
Free presentation	An object universal among interpretations of specified generators satisfying specified equations.	153, 34, 37
Function extensionality	The equivalence between paths of dependent functions and pointwise dependent paths.	400
Functor	An assignment on objects and arrows that preserves identities and composition.	20
Functor of points	The functor of maps from test objects to a given space or scheme.	308, 303
Fundamental group	The automorphism group of a basepoint in its fundamental groupoid.	351
Fundamental groupoid	The groupoid of points and endpoint-preserving homotopy classes of paths.	351
G		
Generalization gap	The difference between the population risk and the empirical risk of a predictor.	278, 280
Generators and relations	Data specifying maps out by assignments to generators that satisfy the chosen equations.	32, 77
Geometric formula	Built from atomic formulas using finite conjunctions, small disjunctions, and existential quantification.	317
Geometric morphism	A morphism of toposes given by an inverse-image adjunction whose left adjoint preserves finite limits.	296
Geometric syntactic site	The site built from formulas in context and provably functional relations of a geometric theory.	319
Germ	A local section at a point, taken up to agreement on some neighborhood of that point.	286
Gradient descent	An update that subtracts a step size times the gradient from the parameter.	268
Grothendieck topology	A choice of covering sieves satisfying maximality, pullback stability, and transitivity.	291
Grothendieck topos	A category equivalent to sheaves of sets on a small site.	295
Grothendieck universe	A transitive set containing $\mathbb{N}$ and closed under pairs, power sets, and small indexed unions.	415
Group	A monoid in which each element has an inverse.	53
Group action	A homomorphism from a group to the permutations of a set.	61

Concept	Definition or use	Page
Group algebra	The free module on a group with multiplication extending the group multiplication.	70
Group completion	The universal abelian group receiving a homomorphism from a commutative monoid.	48
Group presentation	A free group modulo the normal subgroup generated by its relators.	55
Groupoid	A category in which every arrow is invertible.	63, 199
H		
Halting problem	The undecidable problem of whether a given program halts on a specified input.	236
Heyting implication	In the subobject order, $P \Rightarrow Q$ is the greatest R with $R \wedge P \leq Q$.	161
Higher inductive circle	A type generated by a point and a path from that point to itself.	402
Higher topos	An accessible left exact localization of a presheaf ∞-category of spaces.	385
Hoeffding bound	An exponential bound on the deviation of a mean of independent bounded losses from its expectation.	279
Homology	In each degree, cycles modulo boundaries.	328, 331
Homotopy cardinality	A sum over components with alternating factors from their finite homotopy groups.	395
Homotopy fiber	The homotopy pullback of a map along a chosen point of its target.	372
Homotopy limit	An object representing coherent cones by equivalences of mapping spaces.	370
Homotopy pullback	A pullback whose mapping test includes a path comparing the two composites.	371
Homotopy pushout	A pushout whose mapping test includes a homotopy between the two restrictions.	373
Hypercompletion	The localization of an ∞-topos at its ∞-connective morphisms.	388
Hypercover	An augmented simplicial object whose matching maps are effective epimorphisms.	388
Hyperdescent	The gluing condition tested against all hypercovers.	388
I		
Ideal	An additive subgroup closed under multiplication by arbitrary ring elements on both sides.	81

Concept	Definition or use	Page
Identity type	The type of witnesses that two terms of the same type are equal.	197
Image and coimage	The submodule of values and the source modulo the kernel, canonically isomorphic for modules.	94
Inductive type	A type generated by constructors, with dependent elimination supplied by induction.	190, 192
Initial algebra	An algebra with a unique algebra morphism into every algebra for the same endofunctor.	190
Initial object	An object with exactly one map to each object.	8
Initial topology	The coarsest topology making the specified outgoing maps continuous.	113
Injective resolution	An exact augmented cochain complex with injective terms resolving its input object.	345
Interchange law	Compatibility of horizontal and vertical composition of 2-cells.	356
Internal logic	Predicates are interpreted as subobjects, with pullback as substitution.	160, 164
Isomorphism	A morphism with a two-sided inverse.	3
K		
Kan complex	A simplicial set in which every horn admits a filler.	363
Kernel	Universal map into the source of a linear map whose composite with it is zero.	92
Kleene fixed-point theorem	On a pointed ω-CPO, iteration of a continuous endomap from bottom yields its least fixed point.	216
Kleisli category	The category with arrows $X \to TY$, composed using the multiplication of the monad T.	222
Koszul complex	The tensor product of two-term complexes given by multiplication by individual ring elements.	423
L		
Lawvere theory	A category of operations on specified finite powers of one object.	153
Left Kan extension	The left adjoint to restriction of diagrams, computed pointwise by colimits when they exist.	172, 173
Likelihood	The probability of the observed outcome, viewed as a function of the hypothesis.	272

Concept	Definition or use	Page
Limit	A cone through which every other cone to the diagram factors uniquely.	17
Locale	A frame regarded as an object in the opposite of the category of frames.	120
Localization	Universal ring under a given ring in which the specified scalars become units.	85
Locally cartesian closed category	A category with finite limits in which the slice over each object is cartesian closed.	188
Locally small category	A category whose individual hom-sets are small in the chosen universe.	416
Logic–semantics adjunction	The adjunction between taking models of sentences and taking sentences true in every model.	152, 36
Loop space	The space of based loops, given by the homotopy pullback of a basepoint with itself.	372
M		
Mapping cone	For a cochain map $f\colon X \to Y$, the complex $Y \oplus X[1]$ with differential coupled by f.	336
Mapping space	The space of maps between two objects, including their homotopies and higher comparisons.	364
Markov kernel	A family of output probability measures depending measurably on the input.	144
Mate	A transformation corresponding to another by transposing across adjunctions.	420
Mayer–Vietoris sequence	Relates cohomology on two opens, their intersection, and their union in a long exact sequence.	348
Metric completion	A complete metric space universal for 1-Lipschitz maps from the original space.	129
Möbius inversion	Recovers a function on a finite poset from its sums over intervals.	10, 367
Module	An abelian group with an associative, unital scalar action additive in both arguments.	68
Module presentation	A quotient of a free module specified by generators and linear relations.	96
Monad	An endofunctor with a unit and associative multiplication satisfying unit laws.	221
Monodromy	The action of based loops on a covering fiber by path lifting.	353, 393
Monoid	A set with an associative multiplication and a two-sided unit.	27

Concept	Definition or use	Page
Monomorphism	A map m such that $m \circ u = m \circ v$ implies $u = v$.	4
Morphism	An arrow with a source and target in a category.	1
N		
Natural numbers object	A zero-and-successor triple initial among all such triples.	41
Natural transformation	Component maps between functors satisfying the naturality squares.	21
Neural tangent kernel	The Gram kernel of the parameter derivatives of a network's scalar predictions.	255
Norm map	For a finite group action, the map from coinvariants to invariants given by summing translates.	422
O		
ω-complete partial order	A partial order in which every ascending sequence has a least upper bound.	215
Opposite category	The category formed by reversing the directions of arrows and composition.	411
Orbit	All translates of an element under a group action.	62
Orbit–stabilizer theorem	An orbit is in bijection with the left cosets of its stabilizer.	62
P		
Parameter sharing	Two parameter slots receive the same value via a diagonal map.	259
Parameterized map	A parameter object P together with a map $P \times X \to Y$.	258
Partial computable function	A program computes the value on its domain and does not halt outside the domain.	233
Path induction	The identity eliminator extends a construction on reflexivity to arbitrary identity proofs.	197
Path object	A factorization of the diagonal through an equivalence and then a fibration in the spaces model.	399
Pontryagin duality	Evaluation identifies a locally compact abelian group with its double character group.	117, 315, 341, 380, 387
Polynomial ring	The free commutative algebra on its variables, with a monomial basis.	76
Population risk	Expected loss of a predictor under the population law.	250, 278

Concept	Definition or use	Page
Positive semidefinite kernel	A symmetric real function whose Gram matrices on finite lists are positive semidefinite.	253
Posterior distribution	The conditional law of a hypothesis after observing data.	272
Posterior predictive distribution	The law of a future observation obtained by averaging its kernel against the posterior.	274
Presheaf	A contravariant functor from a category to sets.	21
Product	An object with projections universal among families of maps into its factors.	6
Program specialization	A fixed input parameter can be compiled into the code by a total computable operation.	235
Propositional truncation	A type is replaced by a proposition recording only whether it is inhabited.	206
Pullback	Limit of a cospan, universal among pairs of maps agreeing over its shared target.	13
Pushout	Colimit of a span, universal among pairs of maps agreeing on its shared source.	16
Q		
Quasi-isomorphism	A map of complexes inducing isomorphisms on all homology or cohomology groups.	335
Quasicategory	A simplicial set in which every inner horn admits a filler.	363
Quotient ring	The ring universal for maps that annihilate the specified ideal.	82
Quotient stack	Classifies torsors equipped with equivariant maps to the given space.	394
R		
Realizability predicate	An assignment of a set of numerical realizers to each point in a context.	239
Representable functor	A functor naturally isomorphic to a hom functor $\mathrm{Hom}(A, -)$ or $\mathrm{Hom}(-, A)$.	21, 25
Reverse derivative	The dual of the derivative, pulling a covector back from the output to the input.	263
Rice's theorem	A nontrivial property depending only on the computed partial function is undecidable from code.	237
Ridge regression	Least-squares regression with a squared norm penalty on coefficients.	252
Right Kan extension	The right adjoint to restriction of diagrams, computed pointwise by limits when they exist.	172, 173

Concept	Definition or use	Page
Ring	An abelian group under addition with associative unital multiplication distributing over addition.	74
Ring action	A unital ring homomorphism into the additive endomorphisms of an abelian group.	68
S		
Scheme	A locally ringed space covered by open affine schemes.	307
Semantic closure	All consequences of a theory, or the models of all sentences true in a given class.	152
Sheaf cohomology	The right derived functors of global sections on abelian sheaves.	345
Sheaf of sets	A presheaf in which each compatible family of local sections has a unique gluing.	285, 292
Sheaf of spaces	A presheaf of spaces whose sections are equivalent to coherent matching families on every cover.	382
Sheafification	The left adjoint to the inclusion of sheaves into presheaves.	288, 293
Short exact sequence	A sequence $A \xrightarrow{i} B \xrightarrow{p} C$ where i is a kernel of p and p a cokernel of i.	324
Sierpiński space	A two-point space whose continuous characteristic maps classify open subsets.	117
Sieve	A family of arrows with common target that is closed under precomposition.	291
Simplicial commutative algebra	A simplicial object in commutative algebras.	404
Simplicial set	A presheaf of sets on the simplex category.	363
Simply typed lambda calculus	Typed terms with variables, abstraction and application, here also with unit and products.	179, 180
Site	A category equipped with a Grothendieck topology.	291
Skyscraper sheaf	A sheaf with value A on opens containing a specified point and zero on all other opens.	344
Snake lemma	Produces an exact sequence of kernels and cokernels from a map between short exact sequences.	324, 325
Sober space	A space in which each nonempty irreducible closed set has a unique generic point.	123
Span	A pair of arrows with a common source, composed with other spans by pullback.	357

Concept	Definition or use	Page
Spanning tree	A connected acyclic subgraph containing all the vertices.	98
Spectrum	A sequence of pointed spaces E_n with compatible equivalences $E_n \simeq \Omega E_{n+1}$.	375
Square-zero extension	An algebra extension with a kernel ideal whose pairwise products vanish.	407
Stabilizer	The subgroup of group elements fixing a chosen element of the set acted on.	62
Stable ∞-category	A pointed ∞-category with finite limits and colimits in which pushout and pullback squares coincide.	377
Stack in groupoids	A groupoid-valued pseudofunctor satisfying descent for objects and isomorphisms.	390, 391
Stalk	The colimit of the section sets over neighborhoods of a point.	286
State monad	Its computations $S \to X \times S$ produce a result and an updated state.	223
Stochastic gradient descent	A gradient step using a random sample or minibatch to estimate the empirical gradient.	270
Stochastic map	An assignment of a finitely supported output distribution to each input.	140
Strict 2-category	A category enriched in categories, with strictly associative and unital composition.	356
Structure–semantics adjunction	Interpreting a theory over sets corresponds to mapping it into the theory of natural operations.	156
Structure sheaf	The sheaf of rings on a scheme, with local stalks and localizations on affine basic opens.	304
Subobject	An equivalence class of monomorphisms into a fixed object.	146
Subobject classifier	An object Ω whose truth map $1 \to \Omega$ classifies subobjects by pullback.	160
Substitution	Reindexing a predicate or type by pullback along a map of contexts.	148, 184
Suspension spectrum	The spectrum freely associated to a pointed space by the left adjoint of evaluation at level zero.	376
T		
Tangent functor	The functor carrying a smooth map to its action on points and tangent vectors.	134
Tensor product	A module representing bilinear maps out of the two input modules.	99

Concept	Definition or use	Page
Terminal object	An object with exactly one map from each object.	8
Topology	A family of opens containing the empty and whole sets, closed under unions and finite intersections.	112
Tor	The left derived functors of the tensor product of modules.	340, 406
Torsor	A locally trivial family with a free and transitive group action on each fiber.	392
Transfer matrix	Counts one-step transitions by their endpoints and counts longer runs by taking powers.	232
Transport	The operation of moving a term in a family from one fiber to another along an identity witness.	198
U		
Uniform entailment	One program converts any realizer of the premise into a realizer of the conclusion at every point.	239
Unit and counit	For $F \dashv U$, the transformations $1 \Rightarrow UF$ and $FU \Rightarrow 1$ induced by identities.	30
Univalence	The map from identities of types in a universe to equivalences between them is an equivalence.	201, 401
Universal element	An element that induces a natural isomorphism from a representable functor to F.	23, 25
Universal evaluator	A partial computable function that runs a coded program on a supplied input.	234
Universal model	The model in a classifying topos whose inverse images yield the models it classifies.	318
Universal quantification	The right adjoint to reindexing predicates, which tests all points in a fiber in sets.	166, 168
Universe of types	A type of codes for small types, equipped with a family that decodes the codes.	201, 401
V		
Validation	Evaluating a trained predictor on separately drawn data.	282
Van Kampen theorem	Under cover and basepoint hypotheses, computes the fundamental groupoid by a pushout.	353
VC dimension	Supremum of the finite input-set sizes on which the class realizes every binary labeling.	283

Concept	Definition or use	Page
W		
Weak homotopy equivalence	A map inducing a bijection on components and isomorphisms on all based homotopy groups.	362
Well-founded tree type	An inductive type of trees whose node labels specify the index set of their branches.	193
Y		
Yoneda lemma	Evaluation at id_A identifies maps $\mathrm{Hom}(-, A) \Rightarrow F$ with $F(A)$.	22
Z		
Zariski topology	The topology on a prime spectrum with basic opens where a specified function does not vanish.	303

Bibliography

[1] Tom Leinster. *Basic Category Theory*. Cambridge University Press, 2014. https://arxiv.org/abs/1612.09375.

[2] Emily Riehl. *Category Theory in Context*. Dover Publications, 2016. https://emilyriehl.github.io/files/context.pdf.

[3] Martin Speirs. *Math 225A: Model Theory*. Course notes, University of California, Berkeley, 2013, §6.1. https://math.berkeley.edu/~scanlon/225af13lectures/ModelTheoryNotes.pdf.

[4] F. William Lawvere. *Functorial Semantics of Algebraic Theories*. Proceedings of the National Academy of Sciences 50(5), 869–872, 1963. https://www.sas.rochester.edu/mth/sites/doug-ravenel/otherpapers/lawvere.pdf.

[5] Tom Leinster. *Supplement to Rethinking Set Theory*. 2012. https://webhomes.maths.ed.ac.uk/~tl/rst_supp.pdf.

[6] Charles A. Weibel. *The K-book. An Introduction to Algebraic K-theory*. Graduate Studies in Mathematics 145, American Mathematical Society, 2013. Chapter II, §1. https://sites.math.rutgers.edu/~weibel/Kbook.html.

[7] Thomas W. Judson. *Abstract Algebra. Theory and Applications*. Group actions and permutation groups. https://judsonbooks.org/abstract-algebra-theory-and-applications/.

[8] Pavel Etingof, Oleg Golberg, Sebastian Hensel, Tiankai Liu, Alex Schwendner, Dmitry Vaintrob, and Elena Yudovina. *Introduction to Representation Theory*. https://math.mit.edu/~etingof/reprbook.pdf.

[9] The Stacks Project Authors. *Commutative Algebra*. https://stacks.math.columbia.edu/tag/00A0.

[10] James R. Munkres. *Topology*. 2nd ed., Pearson. Chapter 2 covers subspace, product, and quotient topologies. https://www.pearson.com/en-us/subject-catalog/p/topology/P200000006299/9780137848669.

[11] Jorge Picado and Aleš Pultr. *Frames and Locales. Topology without Points*. Birkhäuser, 2012. https://link.springer.com/book/10.1007/978-3-0348-0154-6.

[12] MIT OpenCourseWare. *Construction of* $\mathbb{R}$. Notes for 18.100B Analysis I, Fall 2010. https://ocw.mit.edu/courses/18-100b-analysis-i-fall-2010/resources/mit18_100bf10_const_of_r/.

[13] John M. Lee. *Introduction to Smooth Manifolds*. 2nd ed., Springer, 2013. Chapter 3 treats tangent vectors and differentials. https://doi.org/10.1007/978-1-4419-9982-5.

[14] Bart Jacobs. *Introduction to Coalgebra. Towards Mathematics of States and Observation*. Cambridge University Press, 2016. https://doi.org/10.1017/CBO9781316823187.

[15] Tobias Fritz. *A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics*. Advances in Mathematics 370, 107239, 2020. https://arxiv.org/abs/1908.07021.

[16] Saunders Mac Lane and Ieke Moerdijk. *Sheaves in Geometry and Logic. A First Introduction to Topos Theory*. Springer, 1992. Corrected reprint, 1994. https://doi.org/10.1007/978-1-4612-0927-0.

[17] F. William Lawvere. *Quantifiers and Sheaves*. Actes du Congrès International des Mathématiciens, Nice 1970, vol. 1, 329–334. Gauthier-Villars, 1971. https://lawverearchives.com/wp-content/uploads/2024/12/1970-quantifiers-and-sheaves.pdf.

[18] Joachim Lambek and Philip J. Scott. *Introduction to Higher-Order Categorical Logic*. Cambridge University Press, 1986. Part II treats cartesian closed categories and the typed lambda calculus. https://assets.cambridge.org/97805213/56534/toc/9780521356534_toc.pdf.

[19] Martin Hofmann. *Syntax and Semantics of Dependent Types*. In Andrew M. Pitts and Peter Dybjer, eds., *Semantics and Logics of Computation*, 79–130. Cambridge University Press, 1997. https://www.cambridge.org/core/books/abs/semantics-and-logics-of-computation/syntax-and-semantics-of-dependent-types/119C8085C6A1A0CD7F24928EF866748F.

[20] Martin Hofmann and Thomas Streicher. *The Groupoid Interpretation of Type Theory*. In Giovanni Sambin and Jan M. Smith, eds., *Twenty-Five Years of Constructive Type Theory*, 83–111. Oxford University Press, 1998. https://homotopytypetheory.org/references/hofmann-streicher-the-groupoid-interpretation-of-type-theory/.

[21] The Univalent Foundations Program. *Homotopy Type Theory. Univalent Foundations of Mathematics*. Institute for Advanced Study, 2013. https://homotopytypetheory.org/book/.

[22] Glynn Winskel. *The Formal Semantics of Programming Languages. An Introduction*. MIT Press, 1993. https://mitpress.mit.edu/9780262731034/the-formal-semantics-of-programming-languages/.

[23] Eugenio Moggi. *Notions of Computation and Monads*. Information and Computation 93(1), 55–92, 1991. https://person.dibris.unige.it/moggi-eugenio/ftp/ic91.pdf.

[24] Alan M. Turing. *On Computable Numbers, with an Application to the Entscheidungsproblem*. Proceedings of the London Mathematical Society, second series, 42, 230–265, 1937. https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/plms/s2-42.1.230.

[25] Henry G. Rice. *Classes of Recursively Enumerable Sets and Their Decision Problems*. Transactions of the American Mathematical Society 74(2), 358–366, 1953. https://www.di.ubi.pt/~desousa/DLPC/rice.pdf.

[26] J. Martin E. Hyland. *The Effective Topos*. In A. S. Troelstra and D. van Dalen, eds., *The L. E. J. Brouwer Centenary Symposium*, 165–216. North-Holland, 1982. https://www.dpmms.cam.ac.uk/~jmeh1/Research/Pub81-90/hyland-effectivetopos.pdf.

[27] Jaap van Oosten. *Realizability. An Introduction to its Categorical Side*. Studies in Logic and the Foundations of Mathematics 152, Elsevier, 2008. https://shop.elsevier.com/books/realizability/van-oosten/978-0-444-51584-1.

[28] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning. From Theory to Algorithms*. Cambridge University Press, 2014. https://www.cambridge.org/core/books/understanding-machine-learning/3059695661405D25673058E43C8BE2A6.

[29] Bernhard Schölkopf, Ralf Herbrich, and Alexander J. Smola. *A Generalized Representer Theorem*. In *Computational Learning Theory*, 416–426, Springer, 2001. https://link.springer.com/chapter/10.1007/3-540-44581-1_27.

[30] Arthur Jacot, Franck Gabriel, and Clément Hongler. *Neural Tangent Kernel. Convergence and Generalization in Neural Networks*. In *Advances in Neural Information Processing Systems* 31, 2018. https://proceedings.neurips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

[31] Brendan Fong, David I. Spivak, and Rémy Tuyéras. *Backprop as Functor. A Compositional Perspective on Supervised Learning*. 2017. https://arxiv.org/abs/1711.10455.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. In *Advances in Neural Information Processing Systems* 30, 2017. https://arxiv.org/abs/1706.03762.

[33] Atılım Güneş Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. *Automatic Differentiation in Machine Learning. A Survey*. Journal of Machine Learning Research 18(153), 1–43, 2018. https://jmlr.org/papers/v18/17-468.html.

[34] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. *Optimization Methods for Large-Scale Machine Learning*. SIAM Review 60(2), 223–311, 2018. https://arxiv.org/abs/1606.04838.

[35] Kenta Cho and Bart Jacobs. *Disintegration and Bayesian Inversion via String Diagrams*. Mathematical Structures in Computer Science 29(7), 938–971, 2019. https://arxiv.org/abs/1709.00322.

[36] Wassily Hoeffding. *Probability Inequalities for Sums of Bounded Random Variables*. Journal of the American Statistical Association 58(301), 13–30, 1963. https://www.jstor.org/stable/2282952.

[37] The Stacks Project Authors. *Descent*. https://stacks.math.columbia.edu/tag/0238.

[38] The Stacks Project Authors. *Schemes*. https://stacks.math.columbia.edu/tag/01H8.

[39] The Stacks Project Authors. *Fundamental Groups of Schemes*. https://stacks.math.columbia.edu/tag/0BQ6.

[40] Olivia Caramello. *Theories, Sites, Toposes. Relating and Studying Mathematical Theories through Topos-Theoretic Bridges*. Oxford University Press, 2018. https://academic.oup.com/book/26735.

[41] Charles A. Weibel. *An Introduction to Homological Algebra*. Cambridge University Press, 1994. https://www.cambridge.org/core/books/an-introduction-to-homological-algebra/AAA3F16482097015CD12D4376D505282.

[42] The Stacks Project Authors. *Homological Algebra*. https://stacks.math.columbia.edu/tag/00ZU.

[43] The Stacks Project Authors. *Derived Categories*. https://stacks.math.columbia.edu/tag/05QI.

[44] The Stacks Project Authors. *Cohomology of Sheaves*. https://stacks.math.columbia.edu/tag/01DW.

[45] The Stacks Project Authors. *Cohomology of Schemes*. Lemma 30.2.2, affine vanishing. https://stacks.math.columbia.edu/tag/01XB.

[46] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002. https://pi.math.cornell.edu/~hatcher/AT/ATpage.html.

[47] Ronald Brown. *Groupoids and van Kampen's Theorem*. Proceedings of the London Mathematical Society 17(3), 385–401, 1967. https://groupoids.org.uk/pdffiles/brownvk.pdf.

[48] Tom Leinster. *Basic Bicategories*. 1998. https://arxiv.org/abs/math/9810017.

[49] Stephen Lack. *A 2-Categories Companion*. In *Towards Higher Categories*, 105–191. Springer, 2010. https://arxiv.org/abs/math/0702535.

[50] Jacob Lurie. *Kerodon*. https://kerodon.net/.

[51] Jacob Lurie. *Higher Topos Theory*. Annals of Mathematics Studies 170, Princeton University Press, 2009. https://www.math.ias.edu/~lurie/papers/HTT.pdf.

[52] Jacob Lurie. *Higher Algebra*. 2017. https://www.math.ias.edu/~lurie/papers/HA.pdf.

[53] Bastiaan Cnossen. *Stable Homotopy Theory and Higher Algebra*, §4.3. https://bastiaancnossen.com/notes/stable-homotopy-higher-algebra/section/spectra-and-stable-infinity-categories/stabilization/index.html.

[54] The Stacks Project Authors. *Stacks in Groupoids*. https://stacks.math.columbia.edu/tag/02ZH.

[55] The Stacks Project Authors. *Quotients by Group Actions*. https://stacks.math.columbia.edu/tag/04UV.

[56] Krzysztof Kapulkin and Peter LeFanu Lumsdaine. *The Simplicial Model of Univalent Foundations (after Voevodsky)*. Journal of the European Mathematical Society 23(6), 2071–2126, 2021. https://ems.press/journals/jems/articles/274693.

[57] Bertrand Toën. *Derived Algebraic Geometry*. EMS Surveys in Mathematical Sciences 1(2), 153–240, 2014. https://arxiv.org/abs/1401.1044.

[58] Bertrand Toën and Gabriele Vezzosi. *Homotopical Algebraic Geometry II. Geometric Stacks and Applications*. Memoirs of the American Mathematical Society 193(902), 2008. https://arxiv.org/abs/math/0404373.

[59] The Stacks Project Authors. *Koszul Regular Sequences*. Lemma 15.31.2. https://stacks.math.columbia.edu/tag/062D.

[60] Sidney A. Morris. *Pontryagin Duality and the Structure of Locally Compact Abelian Groups*. Cambridge University Press, 1977, Chapters 4–7. https://www.cambridge.org/core/books/pontryagin-duality-and-the-structure-of-locally-compact-abelian-groups/854007D98ACCC770CCD9432D3D5D68D0.

[61] David W. Roeder. *Category theory applied to Pontryagin duality*. Pacific Journal of Mathematics 52(2), 519–527, 1974. https://msp.org/pjm/1974/52-2/pjm-v52-n2-p21-p.pdf.

[62] Norbert Hoffmann and Markus Spitzweck. *Homological algebra with locally compact abelian groups*. Advances in Mathematics 212(2), 504–524, 2007. https://arxiv.org/abs/math/0510345.

[63] Dinakar Ramakrishnan and Robert J. Valenza. *Fourier Analysis on Number Fields*. Springer, 1999, Chapter 3. https://link.springer.com/book/10.1007/978-1-4757-3085-2.

[64] Terence Tao. *245C, Notes 2: The Fourier transform*. Course notes, 2009. https://terrytao.wordpress.com/2009/04/06/the-fourier-transform/.

[65] Jürgen Neukirch, Alexander Schmidt, and Kay Wingberg. *Cohomology of Number Fields*. 2nd ed., Springer, § I.1. https://www.mathi.uni-heidelberg.de/~schmidt/NSW2e/.

[66] Agnès Beaudry and Jonathan A. Campbell. *A Guide for Computing Stable Homotopy Groups*. 2018, § 2.8. https://arxiv.org/abs/1801.07530.

[67] Peter Scholze. *Lectures on Condensed Mathematics*. Course notes, 2019. https://people.mpim-bonn.mpg.de/scholze/Condensed.pdf.

[68] Alexander Postnikov. *Algebraic Combinatorics*. MIT OpenCourseWare, 2019. https://ocw.mit.edu/courses/18-212-algebraic-combinatorics-spring-2019/resources/lecture-notes/.

[69] Yufei Zhao. *Probabilistic Methods in Combinatorics*. MIT OpenCourseWare, 2022, Lectures 1–2. https://ocw.mit.edu/courses/18-226-probabilistic-methods-in-combinatorics-fall-2022/mit18_226_f22_lec01-02.pdf.

[70] Richard P. Stanley. *Enumerative Combinatorics*, Volume 1. 2nd ed., Cambridge University Press, 2012. https://math.mit.edu/~rstan/ec/ec1/.

[71] Norbert Sauer. *On the density of families of sets*. Journal of Combinatorial Theory, Series A 13(1), 145–147, 1972. https://doi.org/10.1016/0097-3165(72)90019-2.

[72] Imma Gálvez-Carrillo, Joachim Kock, and Andrew Tonks. *Homotopy linear algebra*. Proceedings of the Royal Society of Edinburgh Section A 148(2), 293–325, 2018. https://arxiv.org/abs/1602.05082.

[73] The Open Logic Project. *First-order Logic*. Open Logic Text. https://builds.openlogicproject.org/content/first-order-logic/first-order-logic.pdf.

[74] F. William Lawvere. *Diagonal arguments and cartesian closed categories*. Lecture Notes in Mathematics 92, 134–145, 1969. Reprints in Theory and Applications of Categories 15, 1–13, 2006. https://www.tac.mta.ca/tac/reprints/articles/15/tr15.pdf.

[75] Saharon Shelah. *A combinatorial problem; stability and order for models and theories in infinitary languages*. Pacific Journal of Mathematics 41(1), 247–261, 1972. https://msp.org/pjm/1972/41-1/pjm-v41-n1-p21-p.pdf.