

The Science of Functional Programming

A tutorial with code examples in Scala

by Sergei Winitzki, Ph.D.

Draft version of April 26, 2020

Published by **lulu.com** in 2020

Copyright © 2018-2020 by Sergei Winitzki.

Published at <http://www.lulu.com/content/paperback-book/24915714>

ISBN: 978-0-359-76877-6

Source hash: 508a9a1260f662c08dc787d3b3cfbb5a7ee18e6aa6e1ecdfec0d40d97d32414a

Git commit: 91c2fb76b7b852a2ba685881d375838c8b3c82d8

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled “GNU Free Documentation License” (Appendix F).

A *Transparent* copy of the source code for the book (LyX, LaTeX, graphics files, and build scripts), together with a full-color hyper-linked PDF file, is available at <https://github.com/winitzki/sofp> as `sofp-src/sofp.pdf`. The source code may be also included as a “file attachment” named `sofp-src.tar.bz2` within the PDF file. To extract, run the command `pdftk sofp.pdf unpack_files output .` and then `tar jxvf sofp-src.tar.bz2`. See the file `README.md` for compilation instructions.

This book is a pedagogical in-depth tutorial and reference on the functional programming (FP) paradigm as practiced in the early 21st century. Starting from issues found in practical coding, the book builds up the theoretical intuition, knowledge, and techniques that programmers use for rigorous reasoning about types and code. Examples are given in Scala, but most of the material applies equally to other FP languages.

The book’s topics include working with collections; recursive functions and types; the Curry-Howard correspondence; laws, structural analysis, and code for functors, monads, and other typeclasses; techniques of symbolic derivation and proof; parametricity theorems; and free type constructions.

Long and difficult, yet boring explanations are logically developed in excruciating detail and are accompanied by 98 full step-by-step derivations and 179 solved examples tested in the Scala interpreter, as well as 192 exercises. Discussions further build upon each chapter’s material.

Beginners in FP will find clear explanations about the `map/reduce` programming style, type parameters, disjunctive types, and higher-order functions. For advanced readers, the book shows the practical uses of the Curry-Howard correspondence; proves that all the standard monads (e.g., `State` or `Continuation`) satisfy the monad laws; derives lawful instances of `Functor` and other typeclasses from types; shows that monad transformers need 18 laws; explains the use of Yoneda identities for reasoning about the Church encoding and the free functor constructions; and proves the parametricity theorems without unnecessary jargon.

Readers should have a working knowledge of programming; e.g., be able to write code that prints the number of words in a small text file. The difficulty of this book’s mathematical derivations is at the level of high-school calculus, similar to that of simplifying the expressions

$$\frac{1}{x-2} - \frac{1}{x+2} \quad \text{and} \quad \frac{d}{dx} ((x+1)e^{-x}) \quad .$$

Sergei Winitzki received a Ph.D. in theoretical physics. After a career in academic research, he currently works as a software engineer.

Contents

Preface	1
Formatting conventions used in this book	2
I Beginner level	3
1 Mathematical formulas as code. I. Nameless functions	4
1.1 Translating mathematics into code	4
1.1.1 First examples	4
1.1.2 Nameless functions	5
1.1.3 Nameless functions and bound variables	7
1.2 Aggregating data from sequences	8
1.3 Filtering and truncating a sequence	10
1.4 Solved examples	11
1.4.1 Aggregation	11
1.4.2 Transformation	13
1.5 Summary	13
1.6 Exercises	14
1.6.1 Aggregation	14
1.6.2 Transformation	15
1.7 Discussion	15
1.7.1 Functional programming as a paradigm	15
1.7.2 Functional programming languages	16
1.7.3 The mathematical meaning of “variables”	16
1.7.4 Iteration without loops	17
1.7.5 Nameless functions in mathematical notation	18
1.7.6 Named and nameless expressions and their uses	19
1.7.7 Historical perspective on nameless functions	20
2 Mathematical formulas as code. II. Mathematical induction	21
2.1 Tuple types	21
2.1.1 Examples of using tuples	21
2.1.2 Pattern matching for tuples	22
2.1.3 Using tuples with collections	23
2.1.4 Treating dictionaries as collections	24
2.1.5 Solved examples: Tuples and collections	27
2.1.6 Reasoning about type parameters in collections	31
2.1.7 Exercises: Tuples and collections	32
2.2 Converting a sequence into a single value	33
2.2.1 Inductive definitions of aggregation functions	34
2.2.2 Implementing functions by recursion	35
2.2.3 Tail recursion	35
2.2.4 Implementing general aggregation (<code>foldLeft</code>)	39
2.2.5 Solved examples: using <code>foldLeft</code>	41
2.2.6 Exercises: Using <code>foldLeft</code>	44

2.3	Converting a single value into a sequence	45
2.4	Transforming a sequence into another sequence	47
2.5	Summary	48
2.5.1	Solved examples	48
2.5.2	Exercises	55
2.6	Discussion and further developments	58
2.6.1	Total and partial functions	58
2.6.2	Scope and shadowing of pattern matching variables	59
2.6.3	Lazy values and sequences. Iterators and streams	59
3	The logic of types. I. Disjunctive types	63
3.1	Scala's case classes	63
3.1.1	Tuple types with names	63
3.1.2	Case classes with type parameters	65
3.1.3	Tuples with one part and with zero parts	66
3.1.4	Pattern matching for case classes	67
3.2	Disjunctive types	67
3.2.1	Motivation and first examples	67
3.2.2	Solved examples: Pattern matching for disjunctive types	69
3.2.3	Standard disjunctive types: Option, Either, Try	72
3.3	Lists and trees: recursive disjunctive types	78
3.3.1	Lists	78
3.3.2	Tail recursion with List	80
3.3.3	Binary trees	83
3.3.4	Rose trees	84
3.3.5	Regular-shaped trees	84
3.3.6	Abstract syntax trees	86
3.4	Summary	88
3.4.1	Solved examples	88
3.4.2	Exercises	91
3.5	Discussion and further developments	92
3.5.1	Disjunctive types as mathematical sets	92
3.5.2	Disjunctive types in other programming languages	94
3.5.3	Disjunctions and conjunctions in formal logic	95
II	Intermediate level	96
4	The logic of types. II. Curried functions	97
4.1	Functions that return functions	97
4.1.1	Motivation and first examples	97
4.1.2	Curried and uncurried functions	98
4.1.3	Equivalence of curried and uncurried functions	99
4.2	Fully parametric functions	100
4.2.1	Examples. Function composition	101
4.2.2	Laws of function composition	103
4.2.3	Example: A function that is <i>not</i> fully parametric	105
4.3	Symbolic calculations with nameless functions	106
4.3.1	Calculations with curried functions	106
4.3.2	Solved examples: Deriving a function's type from its code	108
4.4	Summary	111
4.4.1	Solved examples	111

4.4.2	Exercises	116
4.5	Discussion and further developments	117
4.5.1	Higher-order functions	117
4.5.2	Name shadowing and the scope of bound variables	118
4.5.3	Operator syntax for function applications	118
4.5.4	Deriving a function's code from its type	120
5	The logic of types. III. The Curry-Howard correspondence	121
5.1	Values computed by fully parametric functions	121
5.1.1	Motivation	121
5.1.2	Type notation and \mathcal{CH} -propositions for standard type constructions	122
5.1.3	Solved examples: Type notation	125
5.1.4	Exercises: Type notation	128
5.2	The logic of \mathcal{CH} -propositions	128
5.2.1	Motivation and first examples	128
5.2.2	Example: Failure of Boolean logic for type reasoning	130
5.2.3	The rules of proof for \mathcal{CH} -propositions	131
5.2.4	Example: Proving a \mathcal{CH} -proposition and deriving code	135
5.3	Solved examples: Equivalence of types	138
5.3.1	Logical identity does not correspond to type equivalence	138
5.3.2	Arithmetic identity corresponds to type equivalence	141
5.3.3	Type cardinalities and type equivalence	145
5.3.4	Type equivalence involving function types	147
5.4	Summary	153
5.4.1	Solved examples	154
5.4.2	Exercises	163
5.5	Discussion and further developments	164
5.5.1	Using the Curry-Howard correspondence for writing code	164
5.5.2	Implications for designing new programming languages	165
5.5.3	Uses of the void type (<i>Nothing</i>)	166
5.5.4	Relationship between Boolean logic and constructive logic	167
6	Functors, contrafunctors, and profunctors	170
6.1	Practical use	170
6.1.1	Motivation: Type constructors that wrap data	170
6.1.2	Example: Option and the identity law	171
6.1.3	Motivation for the composition law	172
6.1.4	Functors: definition and examples	173
6.1.5	Functor block expressions	177
6.1.6	Examples of non-functors	180
6.1.7	Contrafunctors	185
6.1.8	Subtyping, covariance, and contravariance	187
6.1.9	Solved examples: functors and contrafunctors	189
6.1.10	Exercises: functors and contrafunctors	193
6.2	Laws and structure	194
6.2.1	Reformulations of laws	194
6.2.2	Bifunctors	195
6.2.3	Constructions of functors	197
6.2.4	Constructions of contrafunctors	204
6.2.5	Solved examples: How to recognize functors and contrafunctors	206
6.3	Summary	209
6.3.1	Exercises: Functor and contrafunctor constructions	209

6.4	Further developments	210
6.4.1	Profunctors	210
6.4.2	Subtyping with injective or surjective conversion functions	211
7	Reasoning about code. Techniques of symbolic derivation	213
7.1	Mathematical code notation	213
7.1.1	The nine constructions of purely functional code	213
7.1.2	Function composition and the pipe notation	216
7.1.3	Functor and contrafunctor liftings	217
7.2	Derivation techniques	218
7.2.1	Auxiliary functions for handling products	218
7.2.2	Deriving laws for functions with known implementations	218
7.2.3	Working with disjunctive functions in matrix notation	219
7.2.4	Derivations involving unknown functions with laws	222
7.2.5	Exercises	224
8	Typeclasses and functions of types	225
8.1	Motivation and first examples	225
8.1.1	Constraining type parameters	225
8.1.2	Functions of types and values	225
8.1.3	Partial functions of types and values	226
8.2	Implementing typeclasses	227
8.2.1	Creating a partial function at type level	227
8.2.2	Scala's <code>implicit</code> values	229
8.2.3	Implementing typeclasses by making instances <code>implicit</code>	230
8.2.4	Extension methods	231
8.2.5	Solved examples: Implementing typeclasses in practice	232
8.2.6	Typeclasses for type constructors	238
8.3	Deriving typeclass instances via structural analysis of types	239
8.3.1	Extractors	239
8.3.2	Equality comparison — the <code>Eq</code> typeclass	244
8.3.3	Semigroups	247
8.3.4	Monoids	251
8.3.5	Pointed functors: motivation and laws	255
8.3.6	Pointed functors: structural analysis	257
8.3.7	Co-pointed functors	260
8.3.8	Pointed contrafunctors	264
8.4	Summary	266
8.4.1	Solved examples	267
8.4.2	Exercises	274
8.5	Further developments	276
8.5.1	The existence of values for recursive types	276
8.5.2	Proofs of associativity of <code>concat</code> for lists and arrays	278
8.5.3	“Kinds” and higher-order type functions	281
8.5.4	Inductive typeclasses and their properties	281
8.5.5	Typeclasses with more than one type parameter (type relations)	284
8.5.6	Inheritance and automatic conversions of typeclasses	285
9	Computations in functor blocks. I. Filterable functors and contrafunctors	288
9.1	Practical uses of filtering	288
9.1.1	Examples and intuitions for the filtering operation	289
9.1.2	Motivation for and derivation of the laws of filtering	290
9.1.3	Examples of non-filterable functors	292

9.1.4	Solved examples: Programming with filterable functors	293
9.1.5	Exercises: Programming with filterable functors	298
9.2	Laws and structure	299
9.2.1	Simplifying the filtering laws: Motivation for <code>deflate</code>	299
9.2.2	Equivalence of <code>filter</code> and <code>deflate</code>	301
9.2.3	Motivation and laws for <code>liftOpt</code>	306
9.2.4	Constructions of filterable functors	311
9.2.5	Filterable contrafunctors: motivation and examples	321
9.2.6	Constructions of filterable contrafunctors	324
9.3	Summary	328
9.3.1	Solved examples	328
9.3.2	Exercises	333
9.4	Further developments	334
9.4.1	Naturality laws and natural transformations	334
9.4.2	Generalizing the laws of liftings. Kleisli functions	337
9.4.3	Motivation for using category theory	338
10	Computations in functor blocks. II. Semimonads and monads	342
10.1	Practical use of monads	342
10.1.1	Motivation for semimonads: Nested iteration	342
10.1.2	List-like monads	344
10.1.3	Pass/fail monads	352
10.1.4	Tree-like semimonads and monads	354
10.1.5	The <code>Reader</code> monad	358
10.1.6	The <code>Writer</code> monad	361
10.1.7	The <code>State</code> monad	362
10.1.8	The eager/lazy evaluation monad (<code>Eval</code>)	363
10.1.9	The continuation monad (Cont)	364
10.1.10	Exercises	367
10.2	Laws of semimonads and monads	368
10.2.1	Intuition behind the semimonad laws	368
10.2.2	The laws of <code>flatten</code>	369
10.2.3	Verifying the associativity law via <code>flatten</code>	371
10.2.4	Motivation for monad laws	375
10.2.5	The monad identity laws in terms of <code>pure</code> and <code>flatten</code>	376
10.2.6	Monad laws in terms of Kleisli functions	376
10.2.7	Verifying the monad laws using Kleisli functions	379
10.2.8	Structural analysis of semimonads and monads	380
10.2.9	Exercises: laws and structure of monads	393
10.3	Discussion and further developments	394
10.3.1	Why monads must be functors	394
10.3.2	Equivalence of a natural transformation and a “lifting”	396
10.3.3	Monads, effects, and runners	397
10.3.4	Monads in category theory. Monad morphisms	399
10.3.5	Constructions of polynomial monads	403
10.3.6	Constructions of M -filterable functors and contrafunctors	403
11	Applicative functors, contrafunctors, and profunctors	405
11.1	Slides, Part I	405
11.1.1	Exercises I	406
11.2	Slides, Part II	406
11.2.1	Exercises	414

11.3 Practical use	415
11.4 Laws and structure	415
12 Traversable functors and profunctors	416
12.1 Slides	416
12.1.1 Exercises	419
12.2 Discussion	420
III Advanced level	421
13 “Free” type constructions	422
13.1 Slides	422
13.1.1 Exercises	433
13.2 Discussion	433
14 Computations in functor blocks. III. Monad transformers	434
14.1 Practical use	434
14.1.1 Combining monadic effects via functor composition	434
14.1.2 Combining monads via monad transformers	436
14.1.3 Monad transformers for standard monads	438
14.1.4 Combining more than two monads	444
14.1.5 Lift operations for monad stacks	446
14.2 Laws of monad transformers	447
14.2.1 Motivation for the laws of lift functions	447
14.2.2 Motivation for the laws of runner functions	448
14.2.3 Category-theoretic properties of lifts and runners	451
14.2.4 Summary of the laws of monad transformers	454
14.2.5 Examples of trivially incorrect monad transformers	455
14.2.6 Examples of failure to define a generic monad transformer	455
14.2.7 Functor composition with transformed monads	457
14.2.8 Stacking two monads	457
14.2.9 Stacking any number of monads	460
14.3 Monad transformers via functor composition: General properties	460
14.3.1 Motivation for the <code>swap</code> function	461
14.3.2 Deriving the necessary laws for <code>swap</code>	462
14.3.3 Intuition behind the laws of <code>swap</code>	466
14.3.4 Deriving <code>swap</code> from <code>flatten</code>	466
14.3.5 Monad transformer identity law: Proofs	469
14.3.6 Monad transformer lifting laws: Proofs	470
14.3.7 Monad transformer runner laws: Proofs	471
14.3.8 Summary of results	475
14.4 Composed-inside transformers: Linear monads	475
14.4.1 Definitions of <code>swap</code> and <code>flatten</code>	476
14.4.2 Laws of <code>swap</code>	477
14.4.3 Composition of transformers for linear monads	482
14.5 Composed-outside transformers: Rigid monads	482
14.5.1 Rigid monad construction 1: choice	483
14.5.2 Rigid monad construction 2: composition	495
14.5.3 Rigid monad construction 3: product	500
14.5.4 Rigid monad construction 4: selector	500
14.5.5 Rigid functors	500

14.6 Recursive monad transformers	506
14.6.1 Transformer for the free monad <code>FreeT</code>	506
14.6.2 Transformer for the list monad <code>ListT</code>	506
14.7 Monad transformers for monad constructions	506
14.7.1 Product of monad transformers	506
14.7.2 Free pointed monad transformer	508
14.8 Irregular and incomplete monad transformers	513
14.8.1 The state monad transformer (<code>StateT</code>)	513
14.8.2 The continuation monad transformer (<code>ContT</code>)	517
14.8.3 The choice monad transformer	518
14.8.4 The co-density monad transformer (<code>CodT</code>)	518
14.9 Summary and discussion	519
14.9.1 Some properties of monad morphisms	519
14.9.2 Exercises	520
IV Discussions	523
15 Sample problems	524
16 “Applied functional type theory”: A proposal	526
16.1 AFTT is not covered by computer science curricula	526
16.2 AFTT is not category theory, type theory, or formal logic	527
17 Essay: Software engineers and software artisans	529
17.1 Engineering disciplines	529
17.2 Artisanship: Trades and crafts	529
17.3 Programmers today are artisans, not engineers	530
17.3.1 No requirement of formal study	530
17.3.2 No mathematical formalism guides software development	531
17.3.3 Programmers avoid academic terminology	532
17.4 Towards true engineering in software	532
17.5 Does software need engineers, or are artisans good enough?	534
18 Essay: Towards functional data engineering with Scala	535
18.1 Data is math	535
18.2 Functional programming is math	535
18.3 The power of abstraction	536
18.4 Scala is Java on math	537
18.5 Summary	537
V Appendixes	538
A Notations	539
A.1 Summary of notations for types and code	539
A.2 Detailed explanations	540
B Glossary of terms	545
B.1 On the current misuse of the term “algebra”	547
C Inferring code from types with the LJT algorithm	548
C.1 Slides	548

D Parametricity theorem and naturality laws	553
D.1 Commutativity laws for profunctors and bifunctors	553
D.1.1 Proof of the profunctor commutativity law	554
D.1.2 Commutativity laws for bifunctors and bi-contrafunctors	556
D.2 Naturality laws for purely functional transformations	557
D.2.1 Dinatural transformations between profunctors	557
D.2.2 Composition of natural and dinatural transformations	559
D.2.3 Proof of the parametricity theorem	563
D.2.4 Uniqueness of functor and contrafunctor typeclass instances	568
D.3 Summary	569
E A humorous disclaimer	570
F GNU Free Documentation License	571
F.0.0 Applicability and definitions	571
F.0.1 Verbatim copying	571
F.0.2 Copying in quantity	571
F.0.3 Modifications	571
List of Tables	573
List of Figures	574
Index	575

Preface

This book is at once a reference text and a tutorial that teaches functional programmers how to reason mathematically about types and code, in a manner directly relevant to software practice.

The material ranges from introductory to advanced. The book assumes a certain amount of mathematical experience, at the level of familiarity with high-school algebra or calculus.

The presentation is self-contained, defining and explaining all required ideas, notations, and Scala language features from scratch. The goal is to make all mathematical notions and derivations understandable. To achieve a clearer presentation of the material, the book uses some non-standard notations (Appendix A) and terminology (Appendix B).

The vision of this book is to explain the mathematical principles that guide the practice of functional programming — principles that help us write code. So, all mathematical developments in this book are motivated by practical programming issues and are accompanied by Scala code that illustrates their usage. For instance, the laws for standard typeclasses (functors, monads, etc.) are first motivated heuristically through code examples. Only then the laws are formulated as mathematical equations and verified by explicit step-by-step derivations.

Each concept or technique is motivated and explained to make it as simple as possible and also clarified via solved examples and exercises, which the readers will be able to solve after reading each chapter. More difficult examples and exercises are marked by an asterisk (*).

A software engineer needs to know only a few fragments of mathematical theory; namely, the fragments that answer questions arising in the practice of functional programming. So this book keeps theoretical material at the minimum; *ars longa, vita brevis*. (Chapter 16 discusses the scope of the required theory.) Mathematical generalizations are not pursued beyond proven practical relevance or immediate pedagogical usefulness. This reduces the required mathematical knowledge to first notions of category theory, type theory, and formal logic. Concepts such as functors or natural transformations arise organically from the practice of reasoning about code and are introduced without reference to category theory. This book does not use “introduction/elimination rules”, “strong normalization”, “complete partial order domains”, “adjoint functors”, “pullbacks”, or “topoi”, because learning these concepts will not help a programmer write code.

This book is also *not* an introduction to current theoretical research in functional programming. Instead, the focus is on material known to be practically useful — including constructions such as the “filterable functor” and “applicative contrafunctor”, but excluding a number of theoretical developments that do not (yet?) appear to have significant applications.

The first part of the book is introductory and may be suitable for beginners in programming.

Readers already familiar with functional programming could skim the glossary (Appendix B) to see the unfamiliar terminology and then begin the book at Chapter 5.

Chapters 5–6 begin showing derivations in the code notation, such as Eq. (6.15). If that notation appears difficult to follow, readers could benefit from working through Chapter 7 where the code notation is summarized and clarified with additional examples.

All code examples are intended only for explanation and illustration. As a rule, the code is not optimized for performance or stack safety.

The author thanks Joseph Kim and Jim Kleck for doing some of the exercises and reporting some errors in earlier versions of this book. The author also thanks Bill Venners for many helpful comments on the draft.

Formatting conventions used in this book

- Text in boldface indicates a new concept or term that is being defined at that place in the text. Italics means logical emphasis. Example:

An **aggregation** is a function from a sequence of values to a *single* value.

- Equations are numbered per chapter: Eq. (1.2). Statements, examples, and exercises are numbered per subsection: Example 1.4.1.6 is in subsection 1.4.1, which belongs to Chapter 1.
- Scala code is written inline using a small monospaced font, such as `flatMap` or `val a = "xyz"`. Longer code examples are written in separate code blocks and may also show the Scala interpreter's output for certain lines:

```
val s = (1 to 10).toList
scala> s.product
res0: Int = 3628800
```

- In the introductory chapters, type expressions and code examples are written in the syntax of Scala. Starting from Chapters 4–5, the book introduces a new notation for types where, e.g., the Scala type expression `((A, B)) => Option[A]` is written as $A \times B \rightarrow 1 + A$. Also, a new notation for code is introduced and developed in Chapters 5–7 for efficient reasoning about typeclass laws. For example, the functor composition law is written in the code notation as

$$f^{\uparrow L} ; g^{\uparrow L} = (f ; g)^{\uparrow L} ,$$

where L is a functor and $f: A \rightarrow B$ and $g: B \rightarrow C$ are arbitrary functions of the specified types. The notation $f^{\uparrow L}$ denotes the function f lifted to the functor L and replaces Scala's syntax `x.map(f)` where x is of type `L[A]`. The symbol $;$ denotes the forward composition of functions (Scala's `andThen` method). Appendix A summarizes the notation conventions for types and code.

- Frequently used methods of standard typeclasses, such as `pure`, `flatMap`, `flatten`, `filter`, etc., are denoted by shorter words and are labeled by the type constructor they belong to. For instance, the text talks about typeclass methods `pure`, `flatten`, and `flatMap` for a monad M but denotes the same methods by pu_M , ftn_M , and flm_M when writing code formulas.
- Derivations are written in a two-column format where the right column contains formulas in the code notation and the left column gives a line-by-line explanation or indicates the property or law used to derive the expression at right. A green underline in an expression shows the parts to be rewritten using the law or equation indicated in the *next* line:

$$\begin{aligned} \text{expect to equal } \text{pu}_M : & \quad \underline{\text{pu}_M^{\text{Id}}} ; \text{pu}_M ; \text{ftn}_M \\ \text{lifting to the identity functor : } & = \text{pu}_M ; \underline{\text{pu}_M} ; \text{ftn}_M \\ \text{left identity law of } M : & = \text{pu}_M . \end{aligned}$$

A green underline is sometimes also used at the *last* step of a derivation, to indicate the part of the expression that resulted from the most recent rewriting. Other than providing hints that help remember the steps of a derivation, the green underlines *play no role* in symbolic calculations.

Part I

Beginner level

1 Mathematical formulas as code. I. Nameless functions

1.1 Translating mathematics into code

1.1.1 First examples

We begin by writing Scala code for some computational tasks.

Example 1.1.1.1: Factorial of 10 Find the product of integers from 1 to 10 (the **factorial** of 10).

First, we write a mathematical formula for the result:

$$\prod_{k=1}^{10} k \quad .$$

We can then write Scala code in a way that resembles this formula:

```
scala> (1 to 10).product
res0: Int = 3628800
```

The Scala interpreter indicates that the result is the value 3628800 of type `Int`. To define a name for this value, we use the “`val`” syntax:

```
scala> val fac10 = (1 to 10).product
fac10: Int = 3628800

scala> fac10 == 3628800
res1: Boolean = true
```

The code `(1 to 10).product` is an **expression**, which means that (1) the code can be evaluated (e.g., using the Scala interpreter) and yields a value, and (2) the code can be inserted into a larger expression. For example, we could write

```
scala> 100 + (1 to 10).product + 100 // This code contains '(1 to 10).product' as a sub-expression.
res0: Int = 3629000
```

Example 1.1.1.2: Factorial as a function Define a function that takes an integer n and computes the factorial of n .

A mathematical formula for this function can be written as

$$f(n) = \prod_{k=1}^n k \quad .$$

The corresponding Scala code is

```
def f(n: Int) = (1 to n).product
```

In Scala’s `def` syntax, we need to specify the type of a function’s argument; in this case, we write `n: Int`. In the usual mathematical notation, types of arguments are either not written at all, or written separately from the formula:

$$f(n) = \prod_{k=1}^n k, \quad \forall n \in \mathbb{N} \quad .$$

This indicates that n must be from the set of positive integers, denoted by \mathbb{N} in mathematics. This is similar to specifying the type `Int` in the Scala code. So, the argument's type in the code specifies the **domain** of a function (the set of admissible arguments).

Having defined the function `f`, we can now apply it to an integer argument:

```
scala> f(10)
res6: Int = 3628800
```

It is an error to apply `f` to a non-integer value:

```
scala> f("abc")
<console>:13: error: type mismatch;
  found   : String("abc")
  required: Int
        f("abc")
        ^
```

1.1.2 Nameless functions

The formula and the code, as written above, both involve *naming* the function as “ f ”. Sometimes a function does not really need a name, — say, if the function is used only once. “Nameless” mathematical functions may be denoted using the symbol \rightarrow (pronounced “maps to”) like this:

$$x \rightarrow (\text{some formula}) \quad .$$

So the mathematical notation for the nameless factorial function is

$$n \rightarrow \prod_{k=1}^n k \quad .$$

This reads as “a function that maps n to the product of all k where k goes from 1 to n ”. The Scala expression implementing this mathematical formula is

```
(n: Int) => (1 to n).product
```

This expression shows Scala's syntax for a **nameless function**. Here,

```
n: Int
```

is the function's **argument**, while

```
(1 to n).product
```

is the function's **body**. The arrow symbol `=>` separates the argument from the body.¹

Functions in Scala (whether named or nameless) are treated as values, which means that we can also define a Scala value as

```
scala> val fac = (n: Int) => (1 to n).product
fac: Int => Int = <function1>
```

We see that the value `fac` has the type `Int => Int`, which means that the function `fac` takes an integer (`Int`) argument and returns an integer result value. What is the value of the function `fac` *itself*? As we have just seen, the Scala interpreter prints `<function1>` as the “value” of `fac`. An alternative Scala interpreter² called `ammonite` prints something like this,

```
scala@ val fac = (n: Int) => (1 to n).product
fac: Int => Int = ammonite.$sess.cmd0$$Lambda$1675/2107543287@1e44b638
```

¹This book denotes functions using the arrow symbol \rightarrow in mathematical formulas and the symbols `=>` in Scala code. Many programming languages use the symbols `->` for the function arrow; see Table 1.2.

²<https://ammonite.io/>

The long number could indicate an address in memory.

One may imagine that a “function value” represents a block of compiled code that will actually run and evaluate the function’s body when the function is applied to its argument.

Once defined, a function can be applied to an argument like this:

```
scala> fac(10)
res1: Int = 3628800
```

However, functions can be used without naming them. We can directly apply a nameless factorial function to an integer argument 10 instead of writing `fac(10)`:

```
scala> ((n: Int) => (1 to n).product)(10)
res2: Int = 3628800
```

One would not often write code like this because there is no advantage in creating a nameless function and then applying it right away to an argument. This is because we can evaluate the expression

```
((n: Int) => (1 to n).product)(10)
```

by substituting 10 instead of `n` in the function body, which gives us

```
(1 to 10).product
```

If a nameless function uses the argument several times, for example

```
((n: Int) => n*n*n + n*n)(12345)
```

it is still better to substitute the argument and to eliminate the nameless function. We could write

```
12345*12345*12345 + 12345*12345
```

but, of course, it is better to avoid repeating the value 12345. To achieve that, we may define `n` as a value in an **expression block** like this:

```
scala> { val n = 12345; n*n*n + n*n }
res3: Int = 322687002
```

Defined in this way, the value `n` is visible only within the expression block. Outside the block, another value named `n` could be defined independently of this `n`. For this reason, the definition of `n` is called a **locally scoped** definition.

Nameless functions are convenient when they are themselves arguments of other functions, as we will see next.

Example 1.1.2.1: prime numbers Define a function that takes an integer argument n and determines whether n is a prime number.

A simple mathematical formula for this function can be written as

$$\text{isPrime}(n) = \forall k \in [2, n-1]. (n \% k) \neq 0 \quad . \quad (1.1)$$

This formula has two clearly separated parts: first, a range of integers from 2 to $n-1$, and second, a requirement that all these integers should satisfy a given condition, $(n \% k) \neq 0$. Formula (1.1) is translated into Scala code as

```
def isPrime(n: Int) = (2 to n-1).forall(k => n % k != 0)
```

In this code, the two parts of the mathematical formula are implemented in a way that is closely similar to the mathematical notation, except for the arrow after k .

We can now apply the function `isPrime` to some integer values:

```
scala> isPrime(12)
res3: Boolean = false

scala> isPrime(13)
res4: Boolean = true
```

As we can see from the output above, the function `isPrime` returns a value of type `Boolean`. Therefore, the function `isPrime` has type `Int => Boolean`.

A function that returns a `Boolean` value is called a **predicate**.

In Scala, it is strongly recommended (although often not mandatory) to specify the return type of named functions. The required syntax looks like this,

```
def isPrime(n: Int): Boolean = (2 to n-1).forall(k => n % k != 0)
```

However, we do not need to specify the type `Int` for the argument `k` of the nameless function `k => n % k != 0`. The Scala compiler knows that `k` is going to iterate over the *integer* elements of the range `(2 to n-1)`, which effectively forces `k` to be of type `Int`.

1.1.3 Nameless functions and bound variables

The code for `isPrime` differs from the mathematical formula (1.1) in two ways.

One difference is that the interval $[2, n - 1]$ is in front of `forall`. Another is that the Scala code uses a nameless function `(k => n % k != 0)`, while Eq. (1.1) does not seem to involve any functions.

To understand the first difference, we need to keep in mind that the Scala syntax such as `(2 to n-1).forall(k => ...)` means to apply a function called `forall` to *two* arguments: the first argument is the range `(2 to n-1)`, and the second argument is the nameless function `(k => ...)`. In Scala, the **method** syntax `x.f(z)`, and the equivalent **infix** syntax `x f z`, means that a function `f` is applied to its *two* arguments, `x` and `z`. In the ordinary mathematical notation, this would be $f(x, z)$. Infix notation is often easier to read and is widely used: for instance, we write $x + y$ rather than something like `plus(x, y)`.

A single-argument function could be also defined as a method, and then the syntax is `x.f`, as in the expression `(1 to n).product` we have seen before.

The methods `product` and `forall` are already provided in the Scala standard library, so it is natural to use them. If we want to avoid the method syntax, we could define a function `forall` with two arguments and write code like this,

```
forall(2 to n-1, k => n % k != 0)
```

This would have brought the syntax somewhat closer to the formula (1.1).

However, there still remains the second difference: The symbol `k` is used as an *argument* of a nameless function `(k => n % k != 0)` in the Scala code, while the mathematical formula

$$\forall k \in [2, n - 1]. (n \% k) \neq 0 \quad (1.2)$$

does not seem to use any functions but defines the symbol `k` that goes over the range $[2, n - 1]$. The variable `k` is then used for writing the predicate $(n \% k) \neq 0$.

Let us investigate the role of `k` more closely. The mathematical variable `k` is actually defined *only inside* the expression “ $\forall k : \dots$ ” and makes no sense outside that expression. This becomes clear by looking at Eq. (1.1): The variable `k` is not present in the left-hand side and could not possibly be used there. The name “`k`” is defined only in the right-hand side, where it is first mentioned as the arbitrary element $k \in [2, n - 1]$ and then used in the sub-expression “ $n \% k$ ”.

So, the mathematical notation in Eq. (1.2) says two things: First, we use the name `k` for integers from 2 to $n - 1$. Second, for each of those `k` we evaluate the expression $n \neq 0 \bmod k$, which can be viewed as a certain given *function of k* that returns a `Boolean` value. Translating the mathematical notation into code, it is therefore natural to use the nameless function

$$k \rightarrow (n \% k) \neq 0$$

and to write Scala code applying this nameless function to each element of the range $[2, n - 1]$ and checking that all result values be `true`:

```
(2 to n-1).forall(k => n % k != 0)
```

Just as the mathematical notation defines the variable `k` only in the right-hand side of Eq. (1.1), the argument `k` of

the nameless Scala function `k => n % k != 0` is defined only within that function's body and cannot be used in any code outside the expression `n % k != 0`.

Variables that are defined only inside an expression and are invisible outside are called **bound variables**, or “variables bound in an expression”. Variables that are used in an expression but are defined outside it are called **free variables**, or “variables occurring free in an expression”. These concepts apply equally well to mathematical formulas and to Scala code. For example, in the mathematical expression $k \rightarrow (n \% k) \neq 0$ (which is a nameless function), the variable k is bound (it is defined only within that expression) but the variable n is free (it is defined outside that expression).

The main difference between free and bound variables is that bound variables can be *locally renamed* at will, unlike free variables. To see this, consider that we could rename k to z and write instead of Eq. (1.1) an equivalent definition

$$\text{isPrime}(n) = \forall z \in [2, n - 1]. (n \% z) \neq 0 ,$$

or in Scala code,

```
(2 to n-1).forall(z => n % z != 0)
```

The argument z in the nameless function `z => n % z != 0` may be renamed without changing the result of the entire program. No code outside that function needs to be changed after renaming z . But the value n is defined outside and cannot be renamed “locally” (i.e., only within the sub-expression). If we wanted to rename n in the sub-expression `z => n % z != 0`, we would also need to change all other code that defines and uses n *outside* that expression, or else the program would become incorrect.

Mathematical formulas use bound variables in various constructions such as $\forall k : p(k)$, $\exists k : p(k)$, $\sum_{k=a}^b f(k)$, $\int_0^1 k^2 dk$, $\lim_{n \rightarrow \infty} f(n)$, and $\text{argmax}_k f(k)$. When translating mathematical expressions into code, we need to recognize the presence of bound variables, which the mathematical notation does not make quite so explicit. For each bound variable, we need to create a nameless function whose argument is that variable, e.g., `k=>p(k)` or `k=>f(k)` for the examples just shown. Only then will our code correctly reproduce the behavior of bound variables in mathematical expressions.

As an example, the mathematical formula $\forall k \in [1, n]. p(k)$ has a bound variable k and is translated into Scala code as

```
(1 to n).forall(k => p(k))
```

At this point we can apply a simplification trick to this code. The nameless function $k \rightarrow p(k)$ does exactly the same thing as the (named) function p : It takes an argument, which we may call k , and returns $p(k)$. So, we can simplify the Scala code above to

```
(1 to n).forall(p)
```

The simplification of $x \rightarrow f(x)$ to just f is always possible for functions f of a single argument.³

1.2 Aggregating data from sequences

Consider the task of counting how many even numbers there are in a given list L of integers. For example, the list `[5, 6, 7, 8, 9]` contains *two* even numbers: 6 and 8.

A mathematical formula for this task can be written like this,

$$\begin{aligned} \text{countEven}(L) &= \sum_{k \in L} \text{isEven}(k) , \\ \text{isEven}(k) &= \begin{cases} 1 & \text{if } (k \% 2) = 0 \\ 0 & \text{otherwise} \end{cases} . \end{aligned}$$

³Certain features of Scala allow programmers to write code that looks like `f(x)` but actually involves additional implicit or default arguments of the function `f`, or an implicit type conversion for its argument `x`. In those cases, replacing the code `x => f(x)` by `f` will fail to compile. But these complications do not arise when working with simple functions.

Here we defined a helper function `isEven` in order to write more easily a formula for `countEven`. In mathematics, complicated formulas are often split into simpler parts by defining helper expressions.

We can write the Scala code similarly. We first define the helper function `isEven`; the Scala code can be written in a style quite similar to the mathematical formula:

```
def isEven(k: Int): Int = (k % 2) match {
  case 0 => 1 // First, check if it is zero.
  case _ => 0 // The underscore matches everything else.
}
```

For such a simple computation, we could also write shorter code using a nameless function,

```
val isEven = (k: Int) => if (k % 2 == 0) 1 else 0
```

Given this function, we now need to translate into Scala code the expression $\sum_{k \in L} \text{is_even}(k)$. We can represent the list L using the data type `List[Int]` from the Scala standard library.

To compute $\sum_{k \in L} \text{is_even}(k)$, we must apply the function `isEven` to each element of the list L , which will produce a list of some (integer) results, and then we will need to add all those results together. It is convenient to perform these two steps separately. This can be done with the functions `map` and `sum`, defined in the Scala standard library as methods for the data type `List`.

The method `sum` is similar to `product` and is defined for any `List` of numerical types (`Int`, `Float`, `Double`, etc.). It computes the sum of all numbers in the list:

```
scala> List(1, 2, 3).sum
res0: Int = 6
```

The method `map` needs more explanation. This method takes a *function* as its second argument, applies that function to each element of the list, and puts all the results into a *new* list, which is then returned as the result value:

```
scala> List(1, 2, 3).map(x => x*x + 100*x)
res1: List[Int] = List(101, 204, 309)
```

In this example, the argument of `map` is the nameless function $x \rightarrow x^2 + 100x$. This function will be used repeatedly by `map` to transform each integer from `List(1, 2, 3)`, creating a new list as a result.

It is equally possible to define the transforming function separately, give it a name, and then use it as the argument to `map`:

```
scala> def func1(x: Int): Int = x*x + 100*x
func1: (x: Int)Int

scala> List(1, 2, 3).map(func1)
res2: List[Int] = List(101, 204, 309)
```

Short functions are often defined inline, while longer functions are defined separately with a name.

A method, such as `map`, can be also used with a “dotless” (infix) syntax:

```
scala> List(1, 2, 3) map func1
res3: List[Int] = List(101, 204, 309)
```

If the transforming function `func1` is used only once, and especially for a simple operation such as $x \rightarrow x^2 + 100x$, it is easier to work with a nameless function.

We can now combine the methods `map` and `sum` to define `countEven`:

```
def countEven(s: List[Int]) = s.map(isEven).sum
```

This code can be also written using a nameless function instead of `isEven`:

```
def countEven(s: List[Int]): Int = s
  .map { k => if (k % 2 == 0) 1 else 0 }
  .sum
```

It is customary in Scala to use methods when chaining several operations. For instance `s.map(...).sum` means first apply `s.map(...)`, which returns a *new* list, and then apply `sum` to that list. To make the

code more readable, we put each of the chained methods on a new line.

To test this code, let us run it in the Scala interpreter. In order to let the interpreter work correctly with code entered line by line, the dot character needs to be at the *end* of the line. (In compiled code, the dots may be at the beginning of line since the compiler reads the entire code at once.)

```
scala> def countEven(s: List[Int]): Int = s.
           map { k => if (k % 2 == 0) 1 else 0 }.
           sum
countEven: (s: List[Int])Int

scala> countEven(List(1,2,3,4,5))
res0: Int = 2

scala> countEven( List(1,2,3,4,5).map(x => x * 2) )
res1: Int = 5
```

Note that the Scala interpreter prints the types differently for named functions (i.e., functions declared using `def`). It prints `(s: List[Int])Int` for a function of type `List[Int] => Int`.

1.3 Filtering and truncating a sequence

In addition to the methods `sum`, `product`, `map`, `forall` that we have already seen, the Scala standard library defines many other useful methods. We will now take a look at using the methods `max`, `min`, `exists`, `size`, `filter`, and `takeWhile`.

The methods `max`, `min`, and `size` are self-explanatory:

```
scala> List(10, 20, 30).max
res2: Int = 30

scala> List(10, 20, 30).min
res3: Int = 10

scala> List(10, 20, 30).size
res4: Int = 3
```

The methods `forall`, `exists`, `filter`, and `takeWhile` require a predicate as an argument. The `forall` method returns `true` if and only if the predicate returns `true` for all values in the list; the `exists` method returns `true` if and only if the predicate holds (returns `true`) for at least one value in the list. These methods can be written as mathematical formulas like this:

$$\begin{aligned} \text{forall } (S, p) &= \forall k \in S. p(k) = \text{true} \\ \text{exists } (S, p) &= \exists k \in S. p(k) = \text{true} \end{aligned}$$

However, there is no mathematical notation for operations such as “removing elements from a list”, so we will focus on the Scala syntax for these functions.

The `filter` method returns a list that contains only the values for which the predicate returns `true`:

```
scala> List(1, 2, 3, 4, 5).filter(k => k % 3 != 0)
res5: List[Int] = List(1, 2, 4, 5)
```

The `takeWhile` method truncates a given list, returning a new list with the initial portion of values from the original list for which predicate remains `true`:

```
scala> List(1, 2, 3, 4, 5).takeWhile(k => k % 3 != 0)
res6: List[Int] = List(1, 2)
```

In all these cases, the predicate's argument, `k`, must be of the same type as the elements in the list. In the examples shown above, the elements are integers (i.e., the lists have type `List[Int]`), therefore `k` must be of type `Int`.

The methods `max`, `min`, `sum`, and `product` are defined on lists of *numeric types*, such as `Int`, `Double`, and `Long`. The other methods are defined on lists of all types.

Using these methods, we can solve many problems that involve transforming and aggregating data stored in lists (as well as in arrays, sets, or other similar data structures). In this context, a **transformation** is a function taking a list of values and returning another list of values; examples of transformation functions are `filter` and `map`. An **aggregation** is a function taking a list of values and returning a *single* value; examples of aggregation functions are `max` and `sum`.

Writing programs by chaining together various methods of transformation and aggregation is known as programming in the **map/reduce style**.

1.4 Solved examples

1.4.1 Aggregation

Example 1.4.1.1 Improve the code for `isPrime` by limiting the search to $k^2 \leq n$:

$$\text{isPrime}(n) = \forall k \in [2, n-1] \text{ such that if } k^2 \leq n \text{ then } (n \% k) \neq 0 \quad .$$

Solution: Use `takeWhile` to truncate the initial list when $k^2 \leq n$ becomes false:

```
def isPrime(n: Int): Boolean =
  (2 to n-1)
    .takeWhile(k => k*k <= n)
    .forall(k => n % k != 0)
```

Example 1.4.1.2 Compute this product of absolute values: $\prod_{k \in [1,10]} |\sin(k+2)|$.

Solution

```
(1 to 10)
  .map(k => math.abs(math.sin(k + 2)))
  .product
```

Example 1.4.1.3 Compute $\sum_{k \in [1,10]; \cos k > 0} \sqrt{\cos k}$.

Solution

```
(1 to 10)
  .filter(k => math.cos(k) > 0)
  .map(k => math.sqrt(math.cos(k)))
  .sum
```

It is safe to compute $\sqrt{\cos k}$, because we have first filtered the list by keeping only values k for which $\cos k > 0$. Let us check that this is so:

```
scala> (1 to 10).toList.filter(k => math.cos(k) > 0).map(x => math.cos(x))
res0: List[Double] = List(0.5403023058681398, 0.28366218546322625, 0.9601702866503661,
 0.7539022543433046)
```

Example 1.4.1.4 Compute the average of a non-empty list of type `List[Double]`,

$$\text{average}(s) = \frac{1}{n} \sum_{i=0}^{n-1} s_i \quad .$$

Solution We need to divide the sum by the length of the list:

```
def average(s: List[Double]): Double = s.sum / s.size

scala> average(List(1.0, 2.0, 3.0))
res0: Double = 2.0
```

Example 1.4.1.5 Given n , compute the Wallis product⁴ truncated up to $\frac{2n}{2n+1}$:

$$\text{wallis}(n) = \frac{2}{1} \frac{2}{3} \frac{4}{3} \frac{4}{5} \frac{6}{5} \frac{6}{7} \cdots \frac{2n}{2n+1} \quad .$$

Solution Define the helper function `wallis_frac(i)` that computes the i^{th} fraction. The method `toDouble` converts integers to `Double` numbers.

```
def wallis_frac(i: Int): Double = (2*i).toDouble/(2*i - 1)*(2*i)/(2*i + 1)

def wallis(n: Int) = (1 to n).map(wallis_frac).product

scala> math.cos(wallis(10000)) // Should be close to 0.
res0: Double = 3.9267453954401036E-5

scala> math.cos(wallis(100000)) // Should be even closer to 0.
res1: Double = 3.926966362362075E-6
```

The limit of the Wallis product is $\frac{\pi}{2}$, so the cosine of `wallis(n)` tends to zero in the limit of large n .

Example 1.4.1.6 Check numerically that $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$. First, define a function of n that computes a partial sum of that series until $k = n$. Then compute the partial sum for a large value of n and compare with the limit value.

Solution

```
def euler_series(n: Int): Double = (1 to n).map(k => 1.0/k/k).sum

scala> euler_series(100000)
res0: Double = 1.6449240668982423

scala> val pi = 4*math.atan(1)
pi: Double = 3.141592653589793

scala> pi*pi/6
res1: Double = 1.6449340668482264
```

Example 1.4.1.7 Check numerically the infinite product formula

$$\prod_{k=1}^{\infty} \left(1 - \frac{x^2}{k^2}\right) = \frac{\sin \pi x}{\pi x} \quad .$$

Solution Compute this product up to $k = n$ for $x = 0.1$ with a large value of n , say $n = 10^5$, and compare with the right-hand side:

```
def sine_product(n: Int, x: Double): Double = (1 to n).map(k => 1.0 - x*x/k/k).product

scala> sine_product(n = 100000, x = 0.1) // Arguments may be named, for clarity.
res0: Double = 0.9836317414461351

scala> math.sin(pi*0.1)/pi/0.1
res1: Double = 0.9836316430834658
```

Example 1.4.1.8 Define a function p that takes a list of integers and a function `f: Int => Int`, and returns the largest value of $f(x)$ among all x in the list.

Solution

```
def p(s: List[Int], f: Int => Int): Int = s.map(f).max
```

Here is a test for this function:

```
scala> p(List(2, 3, 4, 5), x => 60 / x)
res0: Int = 30
```

⁴https://en.wikipedia.org/wiki/Wallis_product

Mathematical notation	Scala code
$x \rightarrow \sqrt{x^2 + 1}$	<code>x => math.sqrt(x*x + 1)</code>
list $[1, 2, \dots, n]$	<code>(1 to n)</code>
list $[f(1), \dots, f(n)]$	<code>(1 to n).map(k => f(k))</code>
$\sum_{k=1}^n k^2$	<code>(1 to n).map(k => k*k).sum</code>
$\prod_{k=1}^n f(k)$	<code>(1 to n).map(f).product</code>
$\forall k \in [1, \dots, n]. p(k) \text{ holds}$	<code>(1 to n).forall(k => p(k))</code>
$\exists k \in [1, \dots, n]. p(k) \text{ holds}$	<code>(1 to n).exists(k => p(k))</code>
$\sum_{k \in S \text{ such that } p(k) \text{ holds}} f(k)$	<code>s.filter(p).map(f).sum</code>

Table 1.1: Translating mathematics into code.

1.4.2 Transformation

Example 1.4.2.1 Given a list of lists, `s: List[List[Int]]`, select the inner lists of size at least 3. The result must be again of type `List[List[Int]]`.

Solution To “select the inner lists” means to compute a *new* list containing only the desired inner lists. We use `filter` on the outer list `s`. The predicate for the filter is a function that takes an inner list and returns `true` if the size of that list is at least 3. Write the predicate as a nameless function, `t => t.size >= 3`, where `t` is of type `List[Int]`:

```
def f(s: List[List[Int]]): List[List[Int]] = s.filter(t => t.size >= 3)

scala> f(List(List(1,2), List(1,2,3), List(1,2,3,4)))
res0: List[List[Int]] = List(List(1, 2, 3), List(1, 2, 3, 4))
```

The Scala compiler deduces the type of `t` from the code; no other type would work since we apply `filter` to a *list of lists* of integers.

Example 1.4.2.2 Find all integers $k \in [1, 10]$ such that there are at least three different integers j , where $1 \leq j \leq k$, each j satisfying the condition $j^2 > 2k$.

Solution

```
scala> (1 to 10).toList.filter(k => (1 to k).filter(j => j*j > 2*k).size >= 3)
res0: List[Int] = List(6, 7, 8, 9, 10)
```

The argument of the outer `filter` is a nameless function that also uses a `filter`. The inner expression `(1 to k).filter(j => j*j > 2*k).size >= 3` (shown at left) computes the list of j ’s that satisfy the condition $j^2 > 2k$, and then compares the size of that list with 3. In this way, we impose the requirement that there should be at least 3 values of j . We can see how the Scala code closely follows the mathematical formulation of the task.

1.5 Summary

Functional programs are mathematical formulas translated into code. Table 1.1 shows how to implement some often used mathematical constructions in Scala.

What problems can one solve with this knowledge?

- Compute mathematical expressions involving sums, products, and quantifiers, based on integer ranges, such as $\sum_{k=1}^n f(k)$ etc.

- Transform and aggregate data from lists using `map`, `filter`, `sum`, and other methods from the Scala standard library.

What are examples of problems that are *not* solvable with these tools?

- Example 1: Compute the smallest $n \geq 1$ such that

$$f(f(f(\dots f(0)\dots)) > 1000 \quad ,$$

where the given function f is applied n times.

- Example 2: Given a list s of numbers, compute the list r of running averages:

$$r_n = \frac{1}{n} \sum_{k=0}^{n-1} s_k \quad .$$

- Example 3: Perform binary search over a sorted list of integers.

These computations involve *mathematical induction*, which we have not yet learned to translate into code in the general case.

Library functions we have seen so far, such as `map` and `filter`, implement a restricted class of iterative operations on lists: namely, operations that process each element of a given list independently and accumulate results. For instance, when computing `s.map(f)`, the number of function applications is given by the size of the initial list. However, Example 1 requires applying a function f repeatedly until a given condition holds — that is, repeating for an *initially unknown* number of times. So it is impossible to write an expression containing `map`, `filter`, `takeWhile`, etc., that solves Example 1. We could write the solution of Example 1 as a formula by using mathematical induction, but we have not yet seen how to implement that in Scala code.

Example 2 can be formulated as a definition of a new list r by induction,

$$r_0 = s_0 \quad ; \quad r_i = s_i + r_{i-1} \text{ for } i = 1, 2, 3, \dots$$

However, operations such as `map` and `filter` cannot compute r_i depending on the value of r_{i-1} .

Example 3 defines the search result by induction: the list is split in half, and search is performed by inductive hypothesis in the half that contains the required value. This computation requires an initially unknown number of steps.

Chapter 2 explains how to implement these tasks by translating mathematical induction into code using recursion.

1.6 Exercises

1.6.1 Aggregation

Exercise 1.6.1.1 Machin's formula⁵ converges to π faster than Example 1.4.1.5:

$$\begin{aligned} \frac{\pi}{4} &= 4 \arctan \frac{1}{5} - \arctan \frac{1}{239} \quad , \\ \arctan \frac{1}{n} &= \frac{1}{n} - \frac{1}{3} \frac{1}{n^3} + \frac{1}{5} \frac{1}{n^5} - \dots = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} n^{-2k-1} \quad . \end{aligned}$$

Implement a function that computes the series for $\arctan \frac{1}{n}$ up to a given number of terms, and compute an approximation of π using this formula. Show that about 12 terms of the series are already sufficient for a full-precision `Double` approximation of π .

⁵<http://turner.faculty.swau.edu/mathematics/materialslibrary/pi/machin.html>

Exercise 1.6.1.2 Using the function `isPrime`, check numerically the Euler product formula⁶ for the Riemann's zeta function $\zeta(4)$. It is known⁷ that $\zeta(4) = \frac{\pi^4}{90}$:

$$\zeta(4) = \prod_{k \geq 2; k \text{ is prime}} \frac{1}{1 - \frac{1}{p^4}} = \frac{\pi^4}{90} .$$

1.6.2 Transformation

Exercise 1.6.2.1 Define a function `add_20` of type `List[List[Int]] => List[List[Int]]` that adds 20 to every element of every inner list. A sample test:

```
scala> add_20( List( List(1), List(2, 3) ) )
res0: List[List[Int]] = List(List(21), List(22, 23))
```

Exercise 1.6.2.2 An integer n is called a “3-factor” if it is divisible by only three different integers j such that $2 \leq j < n$. Compute the set of all “3-factor” integers n among $n \in [1, \dots, 1000]$.

Exercise 1.6.2.3 Given a function `f: Int => Boolean`, an integer n is called a “3- f ” if there are only three different integers $j \in [1, \dots, n]$ such that $f(j)$ returns `true`. Define a function that takes f as an argument and returns a sequence of all “3- f ” integers among $n \in [1, \dots, 1000]$. What is the type of that function? Implement Exercise 1.6.2.2 using that function.

Exercise 1.6.2.4 Define a function `see100` of type `List[List[Int]] => List[List[Int]]` that selects only those inner lists whose largest value is at least 100. Test with:

```
scala> see100( List( List(0, 1, 100), List(60, 80), List(1000) ) )
res0: List[List[Int]] = List(List(0, 1, 100), List(1000))
```

Exercise 1.6.2.5 Define a function of type `List[Double] => List[Double]` that “normalizes” the list: it finds the element having the largest absolute value and, if that value is nonzero, divides all elements by that factor and returns a new list; otherwise returns the original list.

1.7 Discussion

1.7.1 Functional programming as a paradigm

Functional programming (FP) is a **paradigm** of programming — an approach that guides programmers to write code in specific ways, applicable to a wide range of tasks.

The main idea of FP is to write code *as a mathematical expression or formula*. This approach allows programmers to derive code through logical reasoning rather than through guessing, similarly to how books on mathematics reason about mathematical formulas and derive results systematically, without guessing or “debugging.” Like mathematicians and scientists who reason about formulas, functional programmers can *reason about code* systematically and logically, based on rigorous principles. This is possible only because code is written as a mathematical formula.

Mathematical intuition is productive because it is backed by the vast experience accumulated while working with data over millennia of human history. It took centuries to invent flexible and powerful notation such as $\sum_{k \in S} p(k)$ and to develop the corresponding rules of calculation. Functional programmers are fortunate to have these reasoning tools at their disposal.

As we have seen, the Scala code for certain computational tasks corresponds quite closely to mathematical formulas. (Scala conventions and syntax do require programmers to write out some details that are omitted in the mathematical notation.) Just as in mathematics, large code expressions may be split into smaller expressions when needed. Expressions can be easily reused, composed in various

⁶https://en.wikipedia.org/wiki/Proof_of_the_Euler_product_formula_for_the_Riemann_zeta_function

⁷<https://tinyurl.com/yxey4tsd>

ways, and written independently from each other. Over the years, the FP community has developed a toolkit of functions (such as `map`, `filter`, etc.) that proved to be especially useful in real-life programming, although many of them are not standard in mathematical literature.

Mastering FP involves learning how to translate the mathematics into code in various cases, practicing to reason about programs as formulas, building up the specific kind of applied mathematical intuition, and getting familiar with concepts adapted to a programmer's needs. The FP community has discovered a number of specific design patterns founded on mathematical principles but driven by practical necessities of programming. This book explains these mathematical design patterns in detail, starting from examples of Scala code and using heuristic ideas to build up the techniques of rigorous reasoning.

1.7.2 Functional programming languages

It is possible to apply the FP paradigm while writing code in any programming language. However, some languages lack certain features that make FP techniques much easier to use in practice. For example, in a language such as Python or Ruby, one can productively use only a limited number of FP idioms, such as the `map/reduce` operations. More advanced FP constructions are impractical in these languages because the required code becomes too hard to read and to write without errors, which negates the advantages of rigorous reasoning about functional programs.

Some programming languages, such as Haskell and OCaml, were designed specifically for advanced use in the FP paradigm. Other languages, such as ML, F#, Scala, Swift, Elm, and PureScript, have different design goals but still support enough FP features to be considered FP languages. This book uses Scala, but the same constructions may be implemented in other FP languages in a similar way. At the level of detail needed in this book, the differences between languages such as ML, OCaml, Haskell, F#, Scala, Swift, Elm, or PureScript do not play a significant role.

1.7.3 The mathematical meaning of “variables”

The usage of variables in functional programming is similar to how mathematical literature uses variables. In mathematics, **variables** are used first of all as *arguments* of functions; e.g., the formula

$$f(x) = x^2 + x$$

contains the variable x and defines a function f that takes x as its argument (to be definite, let us assume that x is an integer) and computes the value $x^2 + x$. The body of the function is the expression $x^2 + x$.

Mathematics has the convention that a variable, such as x , does not change its value within a formula. Indeed, there is no mathematical notation even to talk about “changing” the value of x *inside* the formula $x^2 + x$. It would be quite confusing if a mathematics textbook said “before adding the last x in the formula $x^2 + x$, we change that x by adding 4 to it”. If the “last x ” in $x^2 + x$ needs to have a 4 added to it, a mathematics textbook will just write the formula $x^2 + x + 4$.

Arguments of nameless functions are also immutable. Consider, for example,

$$f(n) = \sum_{k=0}^n (k^2 + k) \quad .$$

Here, n is the argument of the function f , while k is the argument of the nameless function $k \rightarrow k^2 + k$. Neither n nor k can be “modified” in any sense within the expressions where they are used. The symbols k and n stand for some integer values, and these values are immutable. Indeed, it is meaningless to say that we “modified the integer 4”. In the same way, we cannot modify k .

So, a variable in mathematics remains constant *within the expression* where it is defined; in that expression, a variable is essentially a “named constant”. Of course, a function f can be applied to different values x , to compute a different result $f(x)$ each time. However, a given value of x will remain unmodified within the body of the function f while $f(x)$ is being computed.

Functional programming adopts this convention from mathematics: variables are immutable named constants. (Scala also has *mutable* variables, but we will not consider them in this book.)

In Scala, function arguments are immutable within the function body:

```
def f(x: Int) = x * x + x // Cannot modify 'x' here.
```

The *type* of each mathematical variable (such as integer, vector, etc.) is also fixed. Each variable is a value from a specific set (e.g., the set of all integers, the set of all vectors, etc.). Mathematical formulas such as $x^2 + x$ do not express any “checking” that x is indeed an integer and not, say, a vector, in the middle of evaluating $x^2 + x$. The types of all variables are checked in advance.

Functional programming adopts the same view: Each argument of each function must have a **type** that represents *the set of possible allowed values* for that function argument. The programming language’s compiler will automatically check the types of all arguments *before* the program runs. A program that calls functions on arguments of incorrect types will not compile.

The second usage of **variables** in mathematics is to denote expressions that will be reused. For example, one writes: let $z = \frac{x-y}{x+y}$ and now compute $\cos z + \cos 2z + \cos 3z$. Again, the variable z remains immutable, and its type remains fixed.

In Scala, this construction (defining an expression to be reused later) is written with the “`val`” syntax. Each variable defined using “`val`” is a named constant, and its type and value are fixed at the time of definition. Type annotations for “`val`”s are optional in Scala: for instance we could write

```
val x: Int = 123
```

or we could omit the type annotation `:Int` and write more concisely

```
val x = 123
```

because it is clear that this x is an integer. However, it is often helpful to write out types. If we do so, the compiler will check that the types match correctly and give an error message whenever wrong types are used. For example, a type error is detected when using a `String` instead of an `Int`:

```
scala> val x: Int = "123"
<console>:11: error: type mismatch;
  found   : String("123")
  required: Int
         val x: Int = "123"
               ^
```

1.7.4 Iteration without loops

Another distinctive feature of the FP paradigm is handling of iteration without writing loops.

Iterative computations are ubiquitous in mathematics. As an example, consider the formula for the standard deviation estimated from a sample,

$$\sigma(s) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \sum_{j=1}^n s_i s_j - \frac{1}{n(n-1)} \left(\sum_{i=1}^n s_i \right)^2} .$$

These expressions are computed by iterating over values of i and j . And yet, no mathematics textbook uses “loops” or says “now repeat this formula ten times”. Indeed, it would be pointless to evaluate a formula such as $x^2 + x$ ten times, because the result of $x^2 + x$ remains the same every time. It is meaningless to “repeat” an equation such as $(x - 1)(x^2 + x + 1) = x^3 - 1$.

Instead of loops, mathematicians write *expressions* such as $\sum_{i=1}^n s_i$, where symbols such as $\sum_{i=1}^n$ or $\prod_{i=1}^n$ denote iterative computations. Such computations are defined using mathematical induction. The functional programming paradigm has developed rich tools for translating mathematical induction into code. In this chapter, we have seen methods such as `map`, `filter`, and `sum`, which implement certain kinds of iterative computations. These and other operations can be combined in very flexible ways, which allows programmers to write iterative code without loops.

The programmer can avoid writing loops because the iteration is delegated to the library functions `map`, `filter`, `sum`, and so on. It is the job of the library and the compiler to translate these functions into machine code. The machine code most likely *will* contain loops; but the functional programmer does not need to see that code or to reason about it.

1.7.5 Nameless functions in mathematical notation

Functions in mathematics are mappings from one set to another. A function does not necessarily *need* a name; the mapping just needs to be defined. However, nameless functions have not been widely used in the conventional mathematical notation. It turns out that nameless functions are important in functional programming because, in particular, they allow programmers to write code with a straightforward and consistent syntax.

Nameless functions contain bound variables that are invisible outside the function's scope. This property is directly reflected by the prevailing mathematical conventions. Compare the formulas

$$f(x) = \int_0^x \frac{dx}{1+x} \quad ; \quad f(x) = \int_0^x \frac{dz}{1+z} \quad .$$

The mathematical convention is that one may rename the integration variable at will, and so these formulas define the same function f .

In programming, the only situation when a variable "may be renamed at will" is when the variable represents an argument of a function. It follows that the notations $\frac{dx}{1+x}$ and $\frac{dz}{1+z}$ correspond to a nameless function whose argument was renamed from x to z . In FP notation, this nameless function would be denoted as $z \rightarrow \frac{1}{1+z}$, and the integral rewritten as code such as

```
integration(0, x, { z => 1.0 / (1 + z) } )
```

Now consider the traditional mathematical notations for summation, for instance,

$$\sum_{k=0}^x \frac{1}{1+k} \quad .$$

In that sum, the bound variable k is introduced under the \sum symbol; but in integrals, the bound variable follows the special symbol " d ". This notational inconsistency could be removed if we were to use nameless functions explicitly, for example:

$$\begin{aligned} &\text{denote summation by } \sum_0^x \left(k \rightarrow \frac{1}{1+k} \right) \text{ instead of } \sum_{k=0}^x \frac{1}{1+k} \quad , \\ &\text{denote integration by } \int_0^x \left(z \rightarrow \frac{1}{1+z} \right) \text{ instead of } \int_0^x \frac{dz}{1+z} \quad . \end{aligned}$$

In this notation, the new summation symbol \sum_0^x does not mention the name " k " but takes a function as an argument. Similarly, the new integration symbol \int_0^x does not mention " z " and does not use the special symbol " d " but now takes a function as an argument. Written in this way, the operations of summation and integration become *functions* that take functions as arguments. The above summation may be written in a consistent and straightforward manner as a Scala function:

```
summation(0, x, { y => 1.0 / (1 + y) } )
```

We could implement `summation(a, b, g)` as

```
def summation(a: Int, b: Int, g: Int => Double): Double = (a to b).map(g).sum

scala> summation(1, 10, x => math.sqrt(x))
res0: Double = 22.4682781862041
```

Integration requires longer code since the computations are more complicated. Simpson's rule⁸ is an algorithm for approximate numerical integration, defined by the formulas

$$\text{integration}(a, b, g, \varepsilon) = \frac{\delta}{3} (g(a) + g(b) + 4s_1 + 2s_2) \quad ,$$

where $n = 2 \left\lfloor \frac{b-a}{\varepsilon} \right\rfloor$, $\delta_x = \frac{b-a}{n}$,

$$s_1 = \sum_{k=1,3,\dots,n-1} g(a + k\delta_x) \quad , \quad s_2 = \sum_{k=2,4,\dots,n-2} g(a + k\delta_x) \quad .$$

A straightforward line-by-line translation of these formulas into Scala is

```
def integration(a: Double, b: Double, g: Double => Double, eps: Double): Double = {
  // First, we define some helper values and functions corresponding
  // to the definitions "where n = ..." in the mathematical formulas.
  val n: Int = 2 * ((b - a) / eps).toInt
  val delta_x = (b - a) / n
  val s1 = (1 to (n - 1) by 2).map { k => g(a + k * delta_x) }.sum
  val s2 = (2 to (n - 2) by 2).map { k => g(a + k * delta_x) }.sum
  // Now we can write the expression for the final result.
  delta_x / 3 * (g(a) + g(b) + 4 * s1 + 2 * s2)
}

scala> integration(0, 5, x => x*x*x*x, eps = 0.01)      // The exact answer is 625.
res0: Double = 625.0000000004167

scala> integration(0, 7, x => x*x*x*x*x*x, eps = 0.01) // The exact answer is 117649.
res1: Double = 117649.00000014296
```

The entire code is one large *expression*, with a few sub-expressions (s_1 , s_2 , etc.) defined for convenience in the local scope of the function. In other words, this code is written in the FP paradigm.

1.7.6 Named and nameless expressions and their uses

It is a significant advantage if a programming language supports unnamed (or “nameless”) expressions. To see this, consider a familiar situation where we take the absence of names for granted.

In today's programming languages, we may directly write expressions such as $(x+123)*y/(4+x)$. Note that the entire expression does not need to have a name. Parts of that expression (e.g., the sub-expressions $x+123$ or $4+x$) also do not have separate names. It would be inconvenient if we *needed* to assign a name to each sub-expression. The code for $(x+123)*y/(4+x)$ would then look like this:

```
{
  val r0 = 123
  val r1 = x + r0
  val r2 = r1 * y
  val r3 = 4
  val r4 = r3 + x
  val r5 = r2 / r4      // Do we still remember what 'r2' means?
  r5
}
```

This style of programming resembles assembly languages, where *every* sub-expression — that is, every step of every calculation — must be assigned a separate memory address or a CPU register.

Programmers become more productive when their programming language supports nameless expressions. This is also common practice in mathematics; names are assigned when needed, but most expressions remain nameless.

It is similarly useful if data structures can be created without names. For instance, a **dictionary** (also called a “map”) is created in Scala with this code:

⁸https://en.wikipedia.org/wiki/Simpson%27s_rule

```
Map("a" -> 1, "b" -> 2, "c" -> 3)
```

This is a nameless expression whose value is a dictionary. In programming languages that do not have such a construction, programmers have to write repetitive code that creates an initially empty dictionary and then fills it step by step with values:

```
// Scala code creating a dictionary:
Map("a" -> 1, "b" -> 2, "c" -> 3)

// Shortest Java code for the same:
new HashMap<String, Integer>() {{
    put("a", 1);
    put("b", 2);
    put("c", 3);
}}
```

Nameless functions are useful for the same reason as nameless values of other types: they allow us to build larger programs from simpler parts in a uniform way.

1.7.7 Historical perspective on nameless functions

Nameless functions were first used in 1936 in a theoretical programming language called “ λ -calculus”. In that language,⁹ all functions are nameless and have a single argument. The letter λ is a syntax separator denoting function arguments in nameless functions. For example, the nameless function $x \rightarrow x + 1$ could be written as $\lambda x. add\ x\ 1$ in λ -calculus, if it had a function *add* for adding integers (which it does not).

In most programming languages that were in use until around 1990, all functions required names. But by 2015, most languages added support for nameless functions, because programming in the map/reduce style (which invites frequent use of nameless functions) turned out to be immensely productive. Table 1.2 shows the year when nameless functions were introduced in each language.

What this book calls a “nameless function” is also called anonymous function, function expression, function literal, closure, lambda function, lambda expression, or just a “lambda”.

Language	Year	Code for $k \rightarrow k + 1$
λ -calculus	1936	$\lambda k. add\ k\ 1$
typed λ -calculus	1940	$\lambda k : \text{int}. add\ k\ 1$
LISP	1958	$(\lambda \text{lambda}\ (k)\ (+\ k\ 1))$
Standard ML	1973	$\text{fn}\ (k : \text{int}) \Rightarrow k + 1$
Caml	1985	$\text{fun}\ (k : \text{int}) \rightarrow k + 1$
Haskell	1990	$\lambda k \rightarrow k + 1$
Oz	1991	$\text{fun}\ \{\$\ K\} K + 1$
R	1993	$\text{function}(k)\ k + 1$
Python 1.0	1994	$\lambda \text{lambda}\ k : \text{int}\ k + 1$
JavaScript	1995	$\text{function}(k) \{ \text{return}\ k + 1; \}$
Mercury	1995	$\text{func}(K) = K + 1$
Ruby	1995	$\lambda \text{lambda}\ \{ k \ k + 1 \}$
Lua 3.1	1998	$\text{function}(k) \text{return}\ k + 1 \text{ end}$
Scala	2003	$(k : \text{Int}) \Rightarrow k + 1$
F#	2005	$\text{fun}\ (k : \text{int}) \rightarrow k + 1$
C# 3.0	2007	$\text{delegate}(\text{int}\ k) \{ \text{return}\ k + 1; \}$
Clojure	2009	$\text{fn}\ [k]\ (+\ k\ 1)$
C++ 11	2011	$\lambda\ (int\ k) \{ \text{return}\ k + 1; \}$
Go	2012	$\text{func}(k\ \text{int}) \{ \text{return}\ k + 1 \}$
Julia	2012	$\text{function}(k:: \text{Int}) k + 1 \text{ end}$
Kotlin	2012	$\{ k : \text{Int} \rightarrow k + 1 \}$
Swift	2014	$\{ (k:\text{int}) \rightarrow \text{int} \text{ in } \text{return}\ k + 1 \}$
Java 8	2014	$(\text{int}\ k) \rightarrow k + 1$
Rust	2015	$ k: \text{i32} k + 1$

Table 1.2: Nameless functions in programming languages.

⁹Although called a “calculus,” it is a (drastically simplified) programming language, not related to differential or integral calculus. Also, the letter λ has no particular significance; it plays a purely syntactic role in the λ -calculus. Practitioners of functional programming usually do not need to study any λ -calculus. All practically relevant knowledge related to λ -calculus is explained in Chapter 4 of this book.

2 Mathematical formulas as code. II. Mathematical induction

We will now study more flexible ways of working with data collections in the functional programming paradigm. The Scala standard library has methods for performing general iterative computations, that is, computations defined by induction. Translating mathematical induction into code is the focus of this chapter.

First, we need to become fluent in using tuple types with Scala collections.

2.1 Tuple types

2.1.1 Examples of using tuples

Many standard library methods in Scala work with tuple types. A simple example of a tuple is a *pair* of values, e.g., a pair of an integer and a string. The Scala syntax for this type of pair is

```
val a: (Int, String) = (123, "xyz")
```

The type expression `(Int, String)` denotes the type of this pair.

A **triple** is defined in Scala like this:

```
val b: (Boolean, Int, Int) = (true, 3, 4)
```

Pairs and triples are examples of tuples. A **tuple** can contain any number of values, which may be called **parts** of a tuple (they are also called **fields** of a tuple). The parts of a tuple can have different types, but the type of each part is fixed once and for all. Also, the number of parts in a tuple is fixed. It is a **type error** to use incorrect types in a tuple, or an incorrect number of parts of a tuple:

```
scala> val bad: (Int, String) = (1,2)
<console>:11: error: type mismatch;
  found   : Int(2)
  required: String
         val bad: (Int, String) = (1,2)
                           ^

scala> val bad: (Int, String) = (1,"a",3)
<console>:11: error: type mismatch;
  found   : (Int, String, Int)
  required: (Int, String)
         val bad: (Int, String) = (1,"a",3)
                           ^
```

Parts of a tuple can be accessed by number, starting from 1. The Scala syntax for **tuple accessor** methods looks like `._1`, for example:

```
scala> val a = (123, "xyz")
a: (Int, String) = (123,xyz)

scala> a._1
res0: Int = 123

scala> a._2
res1: String = xyz
```

It is a type error to access a tuple part that does not exist:

```
scala> a._0
<console>:13: error: value _0 is not a member of (Int, String)
      a._0
      ^

scala> a._5
<console>:13: error: value _5 is not a member of (Int, String)
      a._5
      ^
```

Type errors are detected at compile time, before any computations begin.

Tuples can be **nested**: any part of a tuple can be itself a tuple:

```
scala> val c: (Boolean, (String, Int), Boolean) = (true, ("abc", 3), false)
c: (Boolean, (String, Int), Boolean) = (true,abc,3),false)

scala> c._1
res0: Boolean = true

scala> c._2
res1: (String, Int) = (abc,3)
```

To define functions whose arguments are tuples, we could use the tuple accessors. An example of such a function is

```
def f(p: (Boolean, Int), q: Int): Boolean = p._1 && (p._2 > q)
```

The first argument, `p`, of this function, has a tuple type. The function body uses accessor methods (`_1` and `_2`) to compute the result value. Note that the second part of the tuple `p` is of type `Int`, so it is valid to compare it with an integer `q`. It would be a type error to compare the *tuple* `p` with an *integer* using the expression `p > q`. It would be also a type error to apply the function `f` to an argument `p` that has a wrong type, e.g., the type `(Int, Int)` instead of `(Boolean, Int)`.

2.1.2 Pattern matching for tuples

Instead of using accessor methods when working with tuples, it is often convenient to use **pattern matching**. Pattern matching occurs in two situations in Scala:

- destructuring definition: `val pattern = ...`
- `case` expression: `case pattern => ...`

```
scala> val g = (1, 2, 3)
g: (Int, Int, Int) = (1,2,3)

scala> val (x, y, z) = g
x: Int = 1
y: Int = 2
z: Int = 3
```

An example of a **destructuring definition** is shown at left. The value `g` is a tuple of three integers. After defining `g`, we define the three variables `x`, `y`, `z` *at once* in a single `val` definition. We imagine that this definition “destructures” the data structure contained in `g` and decomposes it into three parts, then assigns the names `x`, `y`, `z` to these parts. The types of `x`, `y`, `z` are also assigned automatically.

In the example above, the left-hand side of the destructuring definition contains a tuple pattern `(x, y, z)` that looks like a tuple, except that its parts are names `x`, `y`, `z` that are so far *undefined*. These names are called **pattern variables**. The destructuring definition checks whether the structure of the value of `g` “matches” the given pattern. (If `g` does not contain a tuple with exactly three parts, the definition will fail.) This computation is called **pattern matching**.

Pattern matching is often used for working with tuples. The expression `{case (a, b, c) => ...}` called a **case expression** (shown at left) performs pattern matching on its argument. The pattern matching will

“destructure” (i.e., decompose) a tuple and try to match it to the given pattern `(a, b, c)`. In this pattern, `a`, `b`, `c` are as yet undefined new variables, — that is, they are pattern variables. If the pattern

matching succeeds, the pattern variables `a`, `b`, `c` are assigned their values, and the function body can proceed to perform its computation. In this example, the pattern variables `a`, `b`, `c` will be assigned values 1, 2, and 3, and so the expression evaluates to 6.

Pattern matching is especially convenient for nested tuples. Here is an example where a nested tuple `p` is destructured by pattern matching:

```
def t1(p: (Int, (String, Int))): String = p match {
  case (x, (str, y)) => str + (x + y).toString
}

scala> t((10, ("result is ", 2)))
res0: String = result is 12
```

The type structure of the argument `(Int, (String, Int))` is visually repeated in the pattern `(x, (str, y))`, making it clear that `x` and `y` become integers and `str` becomes a string after pattern matching.

If we rewrite the code of `t1` using the tuple accessor methods instead of pattern matching, the code will look like this:

```
def t2(p: (Int, (String, Int))): String = p._2._1 + (p._1 + p._2._2).toString
```

This code is shorter but harder to read. For example, it is not immediately clear that `p._2._1` is a string. It is also harder to modify this code: Suppose we want to change the type of the tuple `p` to `((Int, String), Int)`. Then the new code is

```
def t3(p: ((Int, String), Int)): String = p._1._2 + (p._1._1 + p._2).toString
```

It takes time to verify, by going through every accessor method, that the function `t3` computes the same expression as `t2`. In contrast, the code is changed easily when using the pattern matching expression instead of the accessor methods:

```
def t4(p: ((Int, String), Int)): String = p match {
  case ((x, str), y) => str + (x + y).toString
}
```

The only change in the function body, compared to `t1`, is in the pattern matcher. So it is visually clear that `t4` computes the same expression as `t1`.

Sometimes we do not need some of the tuple parts in a pattern match. The following syntax is used to make this intention clear:

```
scala> val (x, _, _, z) = ("abc", 123, false, true)
x: String = abc
z: Boolean = true
```

The underscore symbol `(_)` denotes the parts of the pattern that we want to ignore. The underscore will always match any value regardless of its type.

A shorter syntax for functions such as `{case (x, y) => y}` that extract elements from tuples `(t => t._2)`, as illustrated here:

```
scala> val p: ((Int, Int)) => Int = { case (x, y) => y }
p: ((Int, Int)) => Int = <function1>

scala> p((1, 2))
res0: Int = 2

scala> val q: ((Int, Int)) => Int = (t => t._2)
q: ((Int, Int)) => Int = <function1>

scala> q((1, 2))
res1: Int = 2

scala> Seq( (1,10), (2,20), (3,30) ).map(t => t._2)
res2: Seq[Int] = List(10, 20, 30)
```

2.1.3 Using tuples with collections

Tuples can be combined with any other types without restrictions. For instance, we can define a tuple of functions,

```
val q: (Int => Int, Int => Int) = (x => x + 1, x => x - 1)
```

We can create a list of tuples,

```
val r: List[(String, Int)] = List(("apples", 3), ("oranges", 2), ("pears", 0))
```

We could define a tuple of lists of tuples of functions, or any other combination.

Here is an example of using the standard method `map` to transform a list of tuples. The argument of `map` must be a function taking a tuple as its argument. It is convenient to use pattern matching for writing such functions:

```
scala> val basket: List[(String, Int)] = List(("apples", 3), ("pears", 2), ("lemons", 0))
basket: List[(String, Int)] = List((apples,3), (pears,2), (lemons,0))

scala> basket.map { case (fruit, count) => count * 2 }
res0: List[Int] = List(6, 4, 0)

scala> basket.map { case (fruit, count) => count * 2 }.sum
res1: Int = 10
```

In this way, we can use the standard methods such as `map`, `filter`, `max`, `sum` to manipulate sequences of tuples. The names of the pattern variables “`fruit`”, “`count`” are chosen to help us remember the meaning of the parts of tuples.

We can easily transform a list of tuples into a list of values of a different type:

```
scala> basket.map { case (fruit, count) =>
  val isAcidic = (fruit == "lemons")
  (fruit, isAcidic)
}
res2: List[(String, Boolean)] = List((apples,false), (pears,false), (lemons,true))
```

In the Scala syntax, a nameless function written with braces `{ ... }` can define local values in its body. The return value of the function is the last expression written in the function body. In this example, the return value of the nameless function is the tuple `(fruit, isAcidic)`.

2.1.4 Treating dictionaries as collections

In the Scala standard library, tuples are frequently used as types of intermediate values. For instance, tuples are used when iterating over dictionaries. The Scala type `Map[K, V]` represents a dictionary with keys of type `K` and values of type `V`. Here `K` and `V` are **type parameters**. Type parameters represent unknown types that will be chosen later, when working with values having specific types.

In order to create a dictionary with given keys and values, we can write

```
Map(("apples", 3), ("oranges", 2), ("pears", 0))
```

The same result is obtained by first creating a sequence of key/value *pairs* and then converting that sequence into a dictionary via the method `toMap`:

```
List(("apples", 3), ("oranges", 2), ("pears", 0)).toMap
```

The same method works for other collection types such as `Seq`, `Vector`, `Stream`, and `Array`.

The Scala library defines a special infix syntax for pairs via the arrow symbol `->`. The expression `x -> y` is equivalent to the pair `(x, y)`:

```
scala> "apples" -> 3
res0: (String, Int) = (apples,3)
```

With this syntax, the code for creating a dictionary is easier to read:

```
Map("apples" -> 3, "oranges" -> 2, "pears" -> 0)
```

The method `toSeq` converts a dictionary into a sequence of pairs:

```
scala> Map("apples" -> 3, "oranges" -> 2, "pears" -> 0).toSeq
res20: Seq[(String, Int)] = ArrayBuffer((apples,3), (oranges,2), (pears,0))
```

The `ArrayBuffer` is one of the many list-like data structures in the Scala library. All these data structures are subtypes of the common “sequence” type `Seq`. The methods defined in the Scala standard library sometimes return different implementations of the `Seq` type for reasons of performance.

The standard library has several useful methods that use tuple types, such as `map` and `filter` (with dictionaries), `toMap`, `zip`, and `zipWithIndex`. The methods `flatten`, `flatMap`, `groupBy`, and `sliding` also work with most collection types, including dictionaries and sets. It is important to become familiar with these methods, because it will help writing code that uses sequences, sets, and dictionaries. Let us now look at these methods one by one.

The `map` and `toMap` methods Chapter 1 showed how the `map` method works on sequences: the expression `xs.map(f)` applies a given function `f` to each element of the sequence `xs`, gathering the results in a new sequence. In this sense, we can say that the `map` method “iterates over” sequences. The `map` method works similarly on dictionaries, except that iterating over a dictionary of type `Map[K, V]` when applying `map` looks like iterating over a sequence of *pairs*, `Seq[(K, V)]`. If `d: Map[K, V]` is a dictionary, the argument `f` of `d.map(f)` must be a function operating on tuples of type `(K, V)`. Typically, such functions are written using `case` expressions:

```
val fruitBasket = Map("apples" -> 3, "pears" -> 2, "lemons" -> 0)

scala> fruitBasket.map { case (fruit, count) => count * 2 }
res0: Seq[Int] = ArrayBuffer(6, 4, 0)
```

When using `map` to transform a dictionary into a sequence of pairs, the result is again a dictionary. But when any intermediate result is not a sequence of pairs, we may need to use `toMap`:

```
scala> fruitBasket.map { case (fruit, count) => (fruit, count * 2) }
res1: Map[String,Int] = Map(apples -> 6, pears -> 4, lemons -> 0)

scala> fruitBasket.map { case (fruit, count) => (fruit, count, count*2) }.
      map { case (fruit, _, count2) => (fruit, count2 / 2) }.toMap
res2: Map[String,Int] = Map(apples -> 3, pears -> 2, lemons -> 0)
```

The `filter` method works on dictionaries by iterating on key/value pairs. The filtering predicate must be a function of type `((K, V)) => Boolean`. For example:

```
scala> fruitBasket.filter { case (fruit, count) => count > 0 }.toMap
res2: Map[String,Int] = Map(apples -> 3, pears -> 2)
```

The `zip` and `zipWithIndex` methods The `zip` method takes *two* sequences and produces a sequence of pairs, taking one element from each sequence:

```
scala> val s = List(1, 2, 3)
s: List[Int] = List(1, 2, 3)

scala> val t = List(true, false, true)
t: List[Boolean] = List(true, false, true)

scala> s.zip(t)
res3: List[(Int, Boolean)] = List((1,true), (2,false), (3,true))

scala> s zip t
res4: List[(Int, Boolean)] = List((1,true), (2,false), (3,true))
```

In the last line, the equivalent “dotless” infix syntax (`s zip t`) is shown to illustrate a syntax convention of Scala that we will sometimes use.

The `zip` method works equally well on dictionaries: in that case, dictionaries are automatically converted to sequences of pairs before applying `zip`.

The `zipWithIndex` method transforms a sequence into a sequence of pairs, where the second part of the pair is the zero-based index:

```
scala> List("a", "b", "c").zipWithIndex
res5: List[(String, Int)] = List((a,0), (b,1), (c,2))
```

The flatten method converts nested sequences to “flattened” ones:

```
scala> List(List(1, 2), List(2, 3), List(3, 4)).flatten
res6: List[Int] = List(1, 2, 2, 3, 3, 4)
```

The “flattening” operation computes the concatenation of the inner sequences. In Scala, sequences are concatenated using the operation `++`, e.g.:

```
scala> List(1, 2, 3) ++ List(4, 5, 6) ++ List(0)
res7: List[Int] = List(1, 2, 3, 4, 5, 6, 0)
```

So the `flatten` method inserts the operation `++` between all the inner sequences.

Keep in mind that `flatten` removes *only one* level of nesting, which is at the “outside” of the data structure. If applied to a `List[List[List[Int]]]`, the `flatten` method returns a `List[List[Int]]`:

```
scala> List(List(List(1), List(2)), List(List(2), List(3))).flatten
res8: List[List[Int]] = List(List(1), List(2), List(2), List(3))
```

The flatMap method is closely related to `flatten` and can be seen as a shortcut, equivalent to first applying `map` and then `flatten`:

```
scala> List(1,2,3,4).map(n => (1 to n).toList)
res9: List[List[Int]] = List(List(1), List(1, 2), List(1, 2, 3), List(1, 2, 3, 4))

scala> List(1,2,3,4).map(n => (1 to n).toList).flatten
res10: List[Int] = List(1, 1, 2, 1, 2, 3, 1, 2, 3, 4)

scala> List(1,2,3,4).flatMap(n => (1 to n).toList)
res11: List[Int] = List(1, 1, 2, 1, 2, 3, 1, 2, 3, 4)
```

The `flatMap` operation transforms a sequence by mapping each element to a potentially different number of new elements.

At first sight, it may be unclear why `flatMap` is useful. (Should we perhaps combine `filter` and `flatten` into a “`flatFilter`”, or combine `zip` and `flatten` into a “`flatZip`”?) However, we will see later in this book that the use of `flatMap`, which is related to “monads”, is one of the most versatile and powerful design patterns in functional programming. In this chapter, several examples and exercises will illustrate the use of `flatMap` for working on sequences.

The groupBy method rearranges a sequence into a dictionary where some elements of the original sequence are grouped together into subsequences. For example, given a sequence of words, we can group all words that start with the letter “`y`” into one subsequence, and all other words into another subsequence. This is accomplished by the following code,

```
scala> Seq("wombat", "xanthan", "yoghurt", "zebra").groupBy(s => if (s startsWith "y") 1 else 2)
res12: Map[Int, Seq[String]] = Map(1 -> List(yoghurt), 2 -> List(wombat, xanthan, zebra))
```

The argument of the `groupBy` method is a *function* that computes a “key” out of each sequence element. The key can have an arbitrarily chosen type. (In the current example, that type is `Int`.) The result of `groupBy` is a dictionary that maps each key to the sub-sequence of values that have that key. (In the current example, the type of the dictionary is therefore `Map[Int, Seq[String]]`.) The order of elements in the sub-sequences remains the same as in the original sequence.

As another example of using `groupBy`, the following code will group together all numbers that have the same remainder after division by 3:

```
scala> List(1,2,3,4,5).groupBy(k => k % 3)
res13: Map[Int, List[Int]] = Map(2 -> List(2, 5), 1 -> List(1, 4), 0 -> List(3))
```

The sliding method creates a sequence of sliding windows of a given width:

```
scala> (1 to 10).sliding(4).toList
res14: List[IndexedSeq[Int]] = List(Vector(1, 2, 3, 4), Vector(2, 3, 4, 5), Vector(3, 4, 5, 6),
  Vector(4, 5, 6, 7), Vector(5, 6, 7, 8), Vector(6, 7, 8, 9), Vector(7, 8, 9, 10))
```

After creating a nested sequence, we can apply an aggregation operation to the inner sequences. For example, the following code computes a sliding-window average with window width 50 over an array of 100 numbers:

```
scala> (1 to 100).map(x => math.cos(x)).sliding(50).map(_.sum / 50).take(5).toList
res15: List[Double] = List(-0.005153079196990285, -0.0011160413780774369, 0.003947079736951305,
 0.005381273944717851, 0.0018679497047270743)
```

The `sortBy` method sorts a sequence according to a sorting key. The argument of `sortBy` is a *function* that computes the sorting key from a sequence element. In this way, we can sort elements in an arbitrary way:

```
scala> Seq(1, 2, 3).sortBy(x => -x)
res0: Seq[Int] = List(3, 2, 1)

scala> Seq("xx", "z", "yyy").sortBy(word => word)           // Sort alphabetically.
res1: Seq[String] = List(xx, yyy, z)

scala> Seq("xx", "z", "yyy").sortBy(word => word.length) // Sort by word length.
res2: Seq[String] = List(z, xx, yyy)
```

Sorting by the elements themselves, as we have done here with `.sortBy(word => word)`, is only possible if the element's type has a well-defined ordering. For strings, this is the alphabetic ordering, and for integers, the standard arithmetic ordering. For such types, a convenience method `sorted` is defined, and works equivalently to `sortBy(x => x)`:

```
scala> Seq("xx", "z", "yyy").sorted
res3: Seq[String] = List(xx, yyy, z)
```

2.1.5 Solved examples: Tuples and collections

Example 2.1.5.1 For a given sequence x_i , compute the sequence of pairs $b_i = (\cos x_i, \sin x_i)$.

Hint: use `map`, assume `xs: Seq[Double]`.

Solution We need to produce a sequence that has a pair of values corresponding to each element of the original sequence. This transformation is exactly what the `map` method does. So the code is

```
xs.map { x => (math.cos(x), math.sin(x)) }
```

Example 2.1.5.2 Count how many times $\cos x_i > \sin x_i$ occurs in a sequence x_i .

Hint: use `count`, assume `xs: Seq[Double]`.

Solution The method `count` takes a predicate and returns the number of sequence elements for which the predicate is `true`:

```
xs.count { x => math.cos(x) > math.sin(x) }
```

We could also reuse the solution of Exercise 2.1.5.1 that computed the cosine and the sine values. The code would then become

```
xs.map { x => (math.cos(x), math.sin(x)) }
      .count { case (cosine, sine) => cosine > sine }
```

Example 2.1.5.3 For given sequences a_i and b_i , compute the sequence of differences $c_i = a_i - b_i$.

Hint: use `zip`, `map`, and assume `as` and `bs` are of type `Seq[Double]`.

Solution We can use `zip` on `as` and `bs`, which gives a sequence of pairs,

```
as.zip(bs): Seq[(Double, Double)]
```

We then compute the differences $a_i - b_i$ by applying `map` to this sequence:

```
as.zip(bs).map { case (a, b) => a - b }
```

Example 2.1.5.4 In a given sequence p_i , count how many times $p_i > p_{i+1}$ occurs.

Hint: use `zip` and `tail`.

Solution Given `ps: Seq[Double]`, we can compute `ps.tail`. The result is a sequence that is 1 element shorter than `ps`, for example:

```
scala> val ps = Seq(1,2,3,4)
ps: Seq[Int] = List(1, 2, 3, 4)

scala> ps.tail
res0: Seq[Int] = List(2, 3, 4)
```

Taking a `zip` of the two sequences `ps` and `ps.tail`, we get a sequence of pairs:

```
scala> ps.zip(ps.tail)
res1: Seq[(Int, Int)] = List((1,2), (2,3), (3,4))
```

Note that `ps.tail` is 1 element shorter than `ps`, and the resulting sequence of pairs is also 1 element shorter than `ps`. In other words, it is not necessary to truncate `ps` before computing `ps.zip(ps.tail)`. Now apply the `count` method:

```
ps.zip(ps.tail).count { case (a, b) => a > b }
```

Example 2.1.5.5 For a given $k > 0$, compute the sequence $c_i = \max(b_{i-k}, \dots, b_{i+k})$.

Solution Applying the `sliding` method to a list gives a list of nested lists:

```
scala> val bs = List(1,2,3,4,5)
bs: List[Int] = List(1, 2, 3, 4, 5)

scala> bs.sliding(3).toList
res0: List[List[Int]] = List(List(1, 2, 3), List(2, 3, 4), List(3, 4, 5))
```

For each b_i , we need to obtain a list of $2k + 1$ nearby elements $(b_{i-k}, \dots, b_{i+k})$. So we need to use `.sliding(2 * k + 1)` to obtain a window of the required size. Now we can compute the maximum of each of the nested lists by using the `map` method on the outer list, with the `max` method applied to the nested lists. So the argument of the `map` method must be the function `nested => nested.max`:

```
bs.sliding(2 * k + 1).map(nested => nested.max)
```

In Scala, this code can be written more concisely using the syntax

```
bs.sliding(2 * k + 1).map(_.max)
```

because the syntax `_.max` means the nameless function `x => x.max`.

Example 2.1.5.6 Create a 10×10 multiplication table as a dictionary of type `Map[(Int, Int), Int]`. For example, a 3×3 multiplication table would be given by this dictionary,

```
Map( (1, 1) -> 1, (1, 2) -> 2, (1, 3) -> 3, (2, 1) -> 2,
     (2, 2) -> 4, (2, 3) -> 6, (3, 1) -> 3, (3, 2) -> 6, (3, 3) -> 9 )
```

Hint: use `flatMap` and `toMap`.

Solution We are required to make a dictionary that maps pairs of integers (x, y) to $x * y$. Begin by creating the list of *keys* for that dictionary, which must be a list of pairs (x, y) of the form `List((1,1), (1,2), ..., (2,1), (2,2), ...)`. We need to iterate over a sequence of values of `x`; and for each `x`, we then need to iterate over another sequence to provide values for `y`. Try this computation:

```
scala> val s = List(1, 2, 3).map(x => List(1, 2, 3))
s: List[List[Int]] = List(List(1, 2, 3), List(1, 2, 3), List(1, 2, 3))
```

We would like to get `List((1,1), (1,2), (1,3))` etc., and so we use `map` on the inner list with a nameless function `y => (1, y)` that converts a number into a tuple,

```
scala> List(1, 2, 3).map { y => (1, y) }
res0: List[(Int, Int)] = List((1,1), (1,2), (1,3))
```

The curly braces in `{y => (1, y)}` are only for clarity; we could also use parentheses and write `(y => (1, y))`.

Now, we need to have (x, y) instead of $(1, y)$ in the argument of `map`, where x iterates over `List(1, 2, 3)` in the outside scope. Using this `map` operation, we obtain

```
scala> val s = List(1, 2, 3).map(x => List(1, 2, 3).map { y => (x, y) })
s: List[List[(Int, Int)]] = List(List((1,1), (1,2), (1,3)), List((2,1), (2,2), (2,3)), List((3,1), (3,2), (3,3)))
```

This is almost what we need, except that the nested lists need to be concatenated into a single list. This is exactly what `flatten` does:

```
scala> val s = List(1, 2, 3).map(x => List(1, 2, 3).map { y => (x, y) }).flatten
s: List[(Int, Int)] = List((1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3))
```

It is shorter to write `.flatMap(...)` instead of `.map(...).flatten`:

```
scala> val s = List(1, 2, 3).flatMap(x => List(1, 2, 3).map { y => (x, y) })
s: List[(Int, Int)] = List((1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3))
```

This is the list of keys for the required dictionary. The dictionary needs to map each *pair* of integers (x, y) to $x * y$. To create that dictionary, we will apply `toMap` to a sequence of pairs `(key, value)`, which in our case needs to be of the form of a nested tuple $((x, y), x * y)$. To achieve this, we use `map` with a function that computes the product and creates these nested tuples:

```
scala> val s = List(1, 2, 3).flatMap(x => List(1, 2, 3).map { y => (x, y) }).
  map { case (x, y) => ((x, y), x * y) }
s: List[((Int, Int), Int)] = List(((1,1),1), ((1,2),2), ((1,3),3), ((2,1),2), ((2,2),4), ((2,3),6),
  ((3,1),3), ((3,2),6), ((3,3),9))
```

We can simplify this code if we notice that we are first mapping each y to a tuple (x, y) , and later map each tuple (x, y) to a nested tuple $((x, y), x * y)$. Instead, the entire computation can be done in the inner `map` operation:

```
scala> val s = List(1, 2, 3).flatMap(x => List(1, 2, 3).map { y => ((x, y), x * y) })
s: List[((Int, Int), Int)] = List(((1,1),1), ((1,2),2), ((1,3),3), ((2,1),2), ((2,2),4), ((2,3),6),
  ((3,1),3), ((3,2),6), ((3,3),9))
```

It remains to convert this list of tuples to a dictionary with `toMap`. Also, for better readability, we can use Scala's pair syntax, `key -> value`, which is equivalent to writing the tuple `(key, value)`:

```
(1 to 10).flatMap(x => (1 to 10).map { y => (x, y) -> x * y }).toMap
```

Example 2.1.5.7 For a given sequence x_i , compute the maximum of all of the numbers $x_i, x_i^2, \cos x_i, \sin x_i$. Hint: use `flatMap`, `max`.

Solution We will compute the required value if we take `max` of a list containing all of the numbers. To do that, first map each element of the list `xs: Seq[Double]` into a sequence of three numbers:

```
scala> val xs = List(0.1, 0.5, 0.9)          // An example list of some 'Double' numbers.
xs: List[Double] = List(0.1, 0.5, 0.9)

scala> xs.map { x => Seq(x, x*x, math.cos(x), math.sin(x)) }
res0: List[Seq[Double]] = List(List(0.1, 0.01000000000000002, 0.9950041652780258,
  0.09983341664682815), List(0.5, 0.25, 0.8775825618903728, 0.479425538604203), List(0.9, 0.81,
  0.6216099682706644, 0.7833269096274834))
```

This list is almost what we need, except we need to `flatten` it:

```
scala> res0.flatten
res1: List[Double] = List(0.1, 0.01000000000000002, 0.9950041652780258, 0.09983341664682815, 0.5,
  0.25, 0.8775825618903728, 0.479425538604203, 0.9, 0.81, 0.6216099682706644, 0.7833269096274834)
```

It remains to take the maximum of the resulting numbers:

```
scala> res1.max
res2: Double = 0.9950041652780258
```

The final code (starting from a given sequence `xs`) is

```
xs.flatMap { x => Seq(x, math.cos(x), math.sin(x)) }.max
```

Example 2.1.5.8 From a dictionary of type `Map[String, String]` mapping names to addresses, and assuming that the addresses do not repeat, compute a dictionary of type `Map[String, String]` mapping the addresses back to names.

Solution Keep in mind that iterating over a dictionary looks like iterating over a list of `(key, value)` pairs:

```
dict.map { case (name, addr) => (addr, name) }
```

Example 2.1.5.9 Write the solution of Example 2.1.5.8 as a function with type parameters `Name` and `Addr` instead of the fixed type `String`.

Solution In Scala, the syntax for type parameters in a function definition is

```
def rev[Name, Addr](...) = ...
```

The type of the argument is `Map[Name, Addr]`, while the type of the result is `Map[Addr, Name]`. So we use the type parameters `Name` and `Addr` in the type signature of the function. The final code is

```
def rev[Name, Addr](dict: Map[Name, Addr]): Map[Addr, Name] =  
  dict.map { case (name, addr) => (addr, name) }
```

The body of the function `rev` remains the same as in Example 2.1.5.8; only the type signature changed. This is because the procedure for reversing a dictionary works in the same way for dictionaries of any type. So the body of the function `rev` does not actually need to know the types of the keys and values in the dictionary. For this reason, it was easy for us to change the specific type `String` into type parameters in that function.

When the function `rev` is applied to a dictionary of a specific type, the Scala compiler will automatically set the type parameters `Name` and `Addr` that fit the required types of the dictionary's keys and values. For example, if we apply `rev` to a dictionary of type `Map[Boolean, Seq[String]]`, the type parameters will be set automatically as `Name = Boolean` and `Addr = Seq[String]`:

```
scala> val d = Map(true -> Seq("x", "y"), false -> Seq("z", "t"))  
d: Map[Boolean, Seq[String]] = Map(true -> List(x, y), false -> List(z, t))  
  
scala> rev(d)  
res0: Map[Seq[String], Boolean] = Map(List(x, y) -> true, List(z, t) -> false)
```

Type parameters can be also set explicitly when using the function `rev`. If the type parameters are chosen incorrectly, the program will not compile:

```
scala> rev[Boolean, Seq[String]](d)  
res1: Map[Seq[String], Boolean] = Map(List(x, y) -> true, List(z, t) -> false)  
  
scala> rev[Int, Double](d)  
<console>:14: error: type mismatch;  
  found   : Map[Boolean, Seq[String]]  
  required: Map[Int, Double]  
        rev[Int, Double](d)  
               ^
```

Example 2.1.5.10* Given a sequence `words: Seq[String]` of some “words”, compute a sequence of type `Seq[Seq[String], Int]`, where each inner sequence should contain all the words having the same length, paired with the integer value showing that length. The resulting sequence must be ordered by increasing length of words. So, the input `Seq("the", "food", "is", "good")` should produce

```
Seq((Seq("is"), 2), (Seq("the"), 3), (Seq("food", "good"), 4))
```

Solution Begin by grouping the words by length. The library method `groupBy` takes a function that computes a grouping key from each element of a sequence. To group by word length, which is

computed with the method `length` if applied to a string, we write

```
words.groupBy { word => word.length }
```

or, more concisely, `words.groupBy(_.length)`. The result of this expression is a dictionary that maps each length to the list of words having that length:

```
scala> words.groupBy(_.length)
res0: scala.collection.immutable.Map[Int,Seq[String]] = Map(2 -> List(is), 4 -> List(food, good), 3 -> List(the))
```

This is close to what we need. If we convert this dictionary to a sequence, we will get a list of pairs

```
scala> words.groupBy(_.length).toSeq
res1: Seq[(Int, Seq[String])] = ArrayBuffer((2,List(is)), (4,List(food, good)), (3,List(the)))
```

It remains to swap the length and the list of words and to sort the result by increasing length. We can do this in any order: first sort, then swap; or first swap, then sort. The final code is

```
words
  .groupBy(_.length)
  .toSeq
  .sortBy { case (len, words) => len }
  .map { case (len, words) => (words, len) }
```

This can be written somewhat shorter if we use the code `_.1` (equivalent to `x => x._1`) for selecting the first parts from pairs and `swap` for swapping the two elements of a pair:

```
words.groupBy(_.length).toSeq.sortBy(_.1).map(_.swap)
```

However, the program may now be harder to read and to modify.

2.1.6 Reasoning about type parameters in collections

In Example 2.1.5.10 we have applied a chain of operations to a sequence. Let us add comments showing the type of the intermediate result after each operation:

```
words // Seq[String]
  .groupBy(_.length) // Map[Int, Seq[String]]
  .toSeq // Seq[(Int, Seq[String])]
  .sortBy { case (len, words) => len } // Seq[(Int, Seq[String])]
  .map { case (len, words) => (words, len) } // Seq[(Seq[String], Int)]
```

In computations like this, the Scala compiler verifies at each step that the operations are applied to values of the correct type.

For instance, `sortBy` is defined for sequences but not for dictionaries, so it would be a type error to apply `sortBy` to a dictionary without first converting it to a sequence using `toSeq`. The type of the intermediate result after `toSeq` is `Seq[(Int, Seq[String])]`, and the `sortBy` operation is applied to that sequence. So the sequence element matched by `{ case (len, words) => len }` is a tuple `(Int, Seq[String])`, which means that the pattern variables `len` and `words` must have types `Int` and `Seq[String]` respectively. It would be a type error to use the sorting key function `{ case (len, words) => words }`: the sorting key can be an integer `len`, but not a string sequence `words` (because sorting by string sequences is not defined).

If we visualize how the type of the sequence should change at every step, we can more quickly understand how to implement the required task. Begin by writing down the intermediate types that would be needed during the computation:

```
words: Seq[String] // Need to group by word length.
Map[Int, Seq[String]] // Need to sort by word length; can't sort a dictionary!
// Need to convert this dictionary to a sequence:
Seq[(Int, Seq[String])] // Now sort this by the 'Int' value. Sorting does not change the types.
// It remains to swap the parts of all tuples in the sequence:
Seq[(Seq[String], Int)] // We are done.
```

Having written down these types, we are better assured that the computation can be done correctly. Writing the code becomes straightforward, since we are guided by the already known types of the intermediate results:

```
words.groupBy(_.length).toSeq.sortBy(_.length).map(_.swap)
```

This example illustrates the main benefits of reasoning about types: it gives direct guidance about how to organize the computation, together with a greater assurance in the correctness of the code.

2.1.7 Exercises: Tuples and collections

Exercise 2.1.7.1 Find all pairs i, j within $(0, 1, \dots, 9)$ such that $i + 4 * j > i * j$.

Hint: use `flatMap` and `filter`.

Exercise 2.1.7.2 Same task as in Exercise 2.1.7.1, but for i, j, k and the condition $i + 4 * j + 9 * k > i * j * k$.

Exercise 2.1.7.3 Given two sequences $p: Seq[String]$ and $q: Seq[Boolean]$ of equal length, compute a $Seq[String]$ with those elements of p for which the corresponding element of q is `true`.

Hint: use `zip`, `map`, `filter`.

Exercise 2.1.7.4 Convert a $Seq[Int]$ into a $Seq[(Int, Boolean)]$ where the Boolean value is `true` when the element is followed by a larger value. For example, the input sequence $Seq(1, 3, 2, 4)$ is to be converted into $Seq((1, true), (3, false), (2, true), (4, false))$. (The last element, 4, has no following element.)

Exercise 2.1.7.5 Given $p: Seq[String]$ and $q: Seq[Int]$ of equal length and assuming that values in q do not repeat, compute a $Map[Int, String]$ mapping numbers from q to the corresponding strings from p .

Exercise 2.1.7.6 Write the solution of Exercise 2.1.7.5 as a function with type parameters P and Q instead of the fixed types `String` and `Int`. Test it with $P = Boolean$ and $Q = Set[Int]$.

Exercise 2.1.7.7 Given $p: Seq[String]$ and $q: Seq[Int]$ of equal length, compute a $Seq[String]$ that contains the strings from p ordered according to the corresponding numbers from q . For example, if $p = Seq("a", "b", "c")$ and $q = Seq(10, -1, 5)$ then the result must be $Seq("b", "c", "a")$.

Exercise 2.1.7.8 Write the solution of Exercise 2.1.7.7 as a function with type parameter s instead of the fixed type `String`. The required type signature and a sample test:

```
def reorder[S](p: Seq[S], q: Seq[Int]): Seq[S] = ??? // In Scala, ??? means "not yet implemented".  
scala> reorder(Seq(6.0, 2.0, 8.0, 4.0), Seq(20, 10, 40, 30))  
res0: Seq[Double] = List(2.0, 6.0, 4.0, 8.0)
```

Exercise 2.1.7.9 Given a $Seq[(String, Int)]$ showing a list of purchased items (where item names may repeat), compute a $Map[String, Int]$ showing the total counts: e.g., for the input

```
Seq(("apple", 2), ("pear", 3), ("apple", 5), ("lemon", 2), ("apple", 3))
```

the output must be $Map("apple" \rightarrow 10, "pear" \rightarrow 3, "lemon" \rightarrow 2)$.

Hint: use `groupBy`, `map`, `sum`.

Exercise 2.1.7.10 Given a $Seq[List[Int]]$, compute a new $Seq[List[Int]]$ where each inner list contains three largest elements from the initial inner list (or fewer than three if the initial inner list is shorter).

Hint: use `map`, `sortBy`, `take`.

Exercise 2.1.7.11 (a) Given two sets, $p: Set[Int]$ and $q: Set[Int]$, compute a set of type $Set[(Int, Int)]$ as the Cartesian product of the sets p and q ; that is, the set of all pairs (x, y) where x is an element from the set p and y is an element from the set q .

(b) Implement this computation as a function with type parameters I, J instead of `Int`. The required type signature and a sample test:

```
def cartesian[I, J](p: Set[I], q: Set[J]): Set[(I, J)] = ???
```

```
scala> cartesian(Set("a", "b"), Set(10, 20))
res0: Set[(String, Int)] = Set((a,10), (a,20), (b,10), (b,20))
```

Hint: use `flatMap` and `map` on sets.

Exercise 2.1.7.12* Given a `Seq[Map[Person, Amount]]`, showing the amounts various people paid on each day, compute a `Map[Person, Seq[Amount]]`, showing the sequence of payments for each person. Assume that `Person` and `Amount` are type parameters. The required type signature and a sample test:

```
def payments[Person, Amount](data: Seq[Map[Person, Amount]]): Map[Person, Seq[Amount]] = ???
// On day 1, Tarski paid 10 and Gödel paid 20. On day 2, Church paid 100 and Gentzen paid 50, etc.
scala> payments(Seq(Map("Tarski" -> 10, "Gödel" -> 20), Map("Church" -> 100, "Gentzen" -> 50),
  Map("Tarski" -> 50), Map("Banach" -> 15, "Gentzen" -> 35)))
res0: Map[String, Seq[Int]] = Map(Gentzen -> List(50, 35), Church -> List(100), Banach -> List(15),
  Tarski -> List(10, 50), Gödel -> List(20))
```

Hint: use `flatMap`, `groupBy`, `mapValues` on dictionaries.

2.2 Converting a sequence into a single value

Until this point, we have been working with sequences using methods such as `map` and `zip`. These techniques are powerful but still insufficient for solving certain problems.

A simple computation that is impossible to do using `map` is obtaining the sum of a sequence of numbers. The standard library method `sum` already does this; but we cannot re-implement `sum` ourselves by using `map`, `zip`, or `filter`. These operations always compute *new sequences*, while we need to compute a single value (the sum of all elements) from a sequence.

We have seen a few library methods such as `count`, `length`, and `max` that compute a single value from a sequence; but we still cannot implement `sum` using these methods. What we need is a more general way of converting a sequence to a single value, such that we could ourselves implement `sum`, `count`, `max`, and other similar computations.

Another task not solvable with `map`, `sum`, etc., is to compute a floating-point number from a given sequence of decimal digits (including a “dot” character):

```
def digitsToDouble(ds: Seq[Char]): Double = ???
scala> digitsToDouble(Seq('2', '0', '4', '.', '5'))
res0: Double = 204.5
```

Why is it impossible to implement this function using `map`, `sum`, and other methods we have seen so far? In fact, the same task for *integer* numbers (instead of floating-point numbers) can be implemented via `length`, `map`, `sum`, and `zip`:

```
def digitsToInt(ds: Seq[Int]): Int = {
  val n = ds.length
  // Compute a sequence of powers of 10, e.g., [1000, 100, 10, 1].
  val powers: Seq[Int] = (0 to n - 1).map(k => math.pow(10, n - 1 - k).toInt)
  // Sum the powers of 10 with coefficients from 'ds'.
  (ds zip powers).map { case (d, p) => d * p }.sum
}

scala> digitsToInt(Seq(2,4,0,5))
res0: Int = 2405
```

This task is doable because the required computation can be written as the formula

$$r = \sum_{k=0}^{n-1} d_k * 10^{n-1-k} .$$

The sequence of powers of 10 can be computed separately and “zipped” with the sequence of digits d_k . However, for floating-point numbers, the sequence of powers of 10 depends on the position of

the “dot” character. Methods such as `map` or `zip` cannot compute a sequence whose next elements depend on previous elements, where the dependence is described by some custom function.

2.2.1 Inductive definitions of aggregation functions

Mathematical induction is a general way of expressing the dependence of next values on previously computed values. To define a function from a sequence to a single value (e.g., an aggregation function $f: Seq[Int] \Rightarrow Int$) via mathematical induction, we need to specify two computations:

- (The **base case** of the induction.) We need to specify what value the function f returns for an empty sequence, `Seq()`. The standard method `isEmpty` can be used to detect empty sequences. In case the function f is only defined for non-empty sequences, we need to specify what the function f returns for a one-element sequence such as `Seq(x)`, with any x .
- (The **inductive step**.) Assuming that the function f is already computed for some sequence xs (the **inductive assumption**), how to compute the function f for a sequence with one more element x ? The sequence with one more element is written as $xs :+ x$. So, we need to specify how to compute $f(xs :+ x)$ assuming that $f(xs)$ is already known.

Once these two computations are specified, the function f is defined (and can in principle be computed) for an arbitrary input sequence. This is how induction works in mathematics, and it works in the same way in functional programming. With this approach, the inductive definition of the method `sum` looks like this:

- The sum of an empty sequence is 0. That is, `Seq().sum == 0`.
- If the result $xs.sum$ is already known for a sequence xs , and we have a sequence that has one more element x , the new result is equal to $xs.sum + x$. In code, this is `(xs :+ x).sum == xs.sum + x`.

The inductive definition of the function `digitsToInt` is:

- For an empty sequence of digits, `Seq()`, the result is 0. This is a convenient base case, even if we never call `digitsToInt` on an empty sequence.
- If `digitsToInt(xs)` is already known for a sequence xs of digits, and we have a sequence $xs :+ x$ with one more digit x , then

```
digitsToInt(xs :+ x) = digitsToInt(xs) * 10 + x
```

Let us write inductive definitions for methods such as `length`, `max`, and `count`:

- The length of a sequence:
 - for an empty sequence, `Seq().length == 0`
 - if $xs.length$ is known then `(xs :+ x).length == xs.length + 1`
- Maximum element of a sequence (undefined for empty sequences):
 - for a one-element sequence, `Seq(x).max == x`
 - if $xs.max$ is known then `(xs :+ x).max == math.max(xs.max, x)`
- Count the sequence elements satisfying a predicate p :
 - for an empty sequence, `Seq().count(p) == 0`
 - if $xs.count(p)$ is known then `(xs :+ x).count(p) == xs.count(p) + c`, where we set $c = 1$ when $p(x) == true$ and $c = 0$ otherwise

There are two main ways of translating mathematical induction into code. The first way is to write a recursive function. The second way is to use a standard library function, such as `foldLeft` or `reduce`. Most often it is better to use the standard library functions, but sometimes the code is more transparent when using explicit recursion. So let us consider each of these ways in turn.

2.2.2 Implementing functions by recursion

A **recursive function** is any function that calls itself somewhere within its own body. The call to itself is the **recursive call**.

When the body of a recursive function is evaluated, it may repeatedly call itself with different arguments until the result value can be computed *without* any recursive calls. The last recursive call corresponds to the base case of the induction. It is an error if the base case is never reached, as in this example:

```
scala> def infiniteLoop(x: Int): Int = infiniteLoop(x+1)
infiniteLoop: (x: Int)Int

scala> infiniteLoop(2) // You will need to press Ctrl-C to stop this.
```

We translate mathematical induction into code by first writing a condition to decide whether we have the base case or the inductive step. As an example, let us define `sum` by recursion. The base case returns 0, and the inductive step returns a value computed from the recursive call. In this case,

```
def sum(s: Seq[Int]): Int = if (s.isEmpty) 0 else {
  val x = s.head // To split s = x ++ xs, compute x
  val xs = s.tail // and xs.
  sum(xs) + x // Call sum(...) recursively.
}
```

In this example, the `if/else` expression will separate the base case from the inductive step. In the inductive step, it is convenient to split the given sequence `s` into its first element `x`, or the “head” of `s`, and the remainder

(“tail”) sequence `xs`. So, we split `s` as `s = x ++ xs` rather than as `s = xs ++ x`.¹

For computing the sum of a numerical sequence, the order of summation does not matter. However, the order of operations *will* matter for many other computational tasks. We need to choose whether the inductive step should split the sequence as `s = x ++ xs` or as `s = xs ++ x`, depending on the task at hand.

Consider the implementation of `digitsToInt` according to the inductive definition shown in the previous subsection:

```
def digitsToInt(s: Seq[Int]): Int = if (s.isEmpty) 0 else {
  val x = s.last // To split s = xs ++ x, compute x
  val xs = s.take(s.length - 1) // and xs.
  digitsToInt(xs) * 10 + x // Call digitsToInt(...) recursively.
}
```

In this example, it is important to split the sequence `s` into `xs ++ x` and not into `x ++ xs`. The reason is that digits increase their numerical value from right to left, so the

correct result is computed if we multiply the value of the *left* subsequence, `digitsToInt(xs)`, by 10.

These examples show how mathematical induction is converted into recursive code. This approach often works but has two technical problems. The first problem is that the code will fail due to the “stack overflow” when the input sequence `s` is long enough. In the next subsection, we will see how this problem is solved (at least in some cases) using “tail recursion”.

The second problem is that each inductively defined function repeats the code for checking the base case and the code for splitting the sequence `s` into the subsequence `xs` and the extra element `x`. This repeated common code can be put into a library function, and the Scala library provides such functions. We will look at using them in Section 2.2.4.

2.2.3 Tail recursion

The code of `lengthS` will fail for large enough sequences. To see why, consider an inductive definition of the `length` method as a function `lengthS`:

```
def lengthS(s: Seq[Int]): Int =
  if (s.isEmpty) 0
  else 1 + lengthS(s.tail)

scala> lengthS((1 to 1000).toList)
```

¹It is easier to remember the meaning of `x ++ xs` and `xs ++ x` if we note that the colon (`:`) always points to the collection (`xs`) and the plus (`+`) to a single element (`x`).

```

res0: Int = 1000

scala> val s = (1 to 100000).toList
s: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22,
  23, 24, 25, 26, 27, 28, 29, 30, 31, 32, ...
  ...

```

The problem is not due to insufficient main memory: we *are* able to compute and hold in memory the entire sequence `s`. The problem is with the code of the function `lengths`. This function calls itself *inside* an expression `1 + lengths(...)`. So we can visualize how the computer evaluates this code:

```

lengths(Seq(1, 2, ..., 100000))
  = 1 + lengths(Seq(2, ..., 100000))
  = 1 + (1 + lengths(Seq(3, ..., 100000)))
  = ...

```

The function body of `lengths` will evaluate the inductive step, that is, the “`else`” part of the “`if/else`”, about 100,000 times. Each time, the sub-expression with nested computations `1 + (1 + (...))` will get larger. This intermediate sub-expression needs to be held somewhere in memory, until at

some point the function body goes into the base case and returns a value. When that happens, the entire intermediate sub-expression will contain about 100,000 nested function calls still waiting to be evaluated. This sub-expression is held in a special area of memory called **stack memory**, where the not-yet-evaluated nested function calls are held in the order of their calls on a stack. Due to the way computer memory is managed, the stack memory has a fixed size and cannot grow automatically. So, when the intermediate expression becomes large enough, it causes an overflow of the stack memory and crashes the program.

One way to avoid stack overflows is to use a trick called **tail recursion**. Using tail recursion means rewriting the code so that all recursive calls occur at the end positions (at the “tails”) of the function body. In other words, each recursive call must be *itself* the last computation in the function body, rather than placed inside other computations. Here is an example of tail-recursive code:

```

def lengthT(s: Seq[Int], res: Int): Int =      In this code, one of the branches of the if/else returns
  if (s.isEmpty) res          a fixed value without doing any recursive calls, while
  else lengthT(s.tail, res + 1)      the other branch returns the result of recursive call to

```

`lengthT(...)`. In the code of `lengthT`, recursive calls never occur within any sub-expressions.

It is not a problem that the recursive call to `lengthT` has some sub-expressions such as `res + 1` as its arguments, because all these sub-expressions will be computed *before* `lengthT` is recursively called. The recursive call to `lengthT` is the *last* computation performed by this branch of the `if/else`. A tail-recursive function can have many `if/else` or `match/case` branches, with or without recursive calls; but all recursive calls must be always the last expressions returned.

The Scala compiler has a feature for checking automatically that a function’s code is tail-recursive: the `@tailrec` annotation. If a function with a `@tailrec` annotation is not tail-recursive, or is not recursive at all, the program will not compile.

```

@tailrec def lengthT(s: Seq[Int], res: Int): Int =
  if (s.isEmpty) res
  else lengthT(s.tail, res + 1)

```

Let us trace the evaluation of this function on an example:

```

lengthT(Seq(1, 2, 3), 0)
  = lengthT(Seq(2, 3), 0 + 1) // = lengthT(Seq(2, 3), 1)
  = lengthT(Seq(3), 1 + 1)   // = lengthT(Seq(3), 2)
  = lengthT(Seq(), 2 + 1)    // = lengthT(Seq(), 3)
  = 3

```

All sub-expressions such as `1 + 1` and `2 + 1` are computed *before* recursive calls to `lengthT`. Because of that, sub-expressions do not grow within the stack memory. This is the main benefit of tail recursion.

How did we rewrite the code of `lengths` into the tail-recursive code of `lengthT`? An important difference between `lengths` and `lengthT` is the additional argument, `res`, called the **accumulator argument**.

This argument is equal to an intermediate result of the computation. The next intermediate result (`res + 1`) is computed and passed on to the next recursive call via the accumulator argument. In the base case of the recursion, the function now returns the accumulated result, `res`, rather than 0, because at that time the computation is finished.

Rewriting code by adding an accumulator argument to achieve tail recursion is called the **accumulator technique** or the “accumulator trick”.

One consequence of using the accumulator trick is that the function `lengthT` now always needs a value for the accumulator argument. However, our goal is to implement a function such as `length(s)` with just one argument, `s: Seq[Int]`. We can define `length(s) = lengthT(s, ???)` if we supply an initial accumulator value. The correct initial value for the accumulator is 0, since in the base case (an empty sequence `s`) we need to return 0.

So, a tail-recursive implementation of `lengthT` requires us to define *two* functions: the tail-recursive `lengthT` and the main function that will set the initial value of the accumulator argument. To emphasize that `lengthT` is a helper function, one could define it *inside* the main function:

```
def length[A](xs: Seq[A]): Int = {
  @tailrec def lengthT(s: Seq[A], res: Int): Int = {
    if (s.isEmpty) res
    else lengthT(s.tail, res + 1)
  }
  lengthT(xs, 0)
}
```

When `length` is implemented like that, users will not be able to call `lengthT` directly, because it is only visible within the body of the `length` function.

Another possibility in Scala is to use a **default value** for the `res` argument:

```
@tailrec def length[A](s: Seq[A], res: Int = 0): Int =
  if (s.isEmpty) res
  else length(s.tail, res + 1)
```

Giving a default value for a function argument is the same as defining *two* functions: one with that argument and one without. For example, the syntax

```
def f(x: Int, y: Boolean = false): Int = ... // Function body.
```

is equivalent to defining two functions (with the same name):

```
def f(x: Int, y: Boolean) = ...      // Define the function body here.
def f(x: Int): Int = f(Int, false)  // Call the function defined above.
```

Using a default argument, we can define the tail-recursive helper function and the main function at once, making the code shorter.

The accumulator trick works in a large number of cases, but it may be far from obvious how to introduce the accumulator argument, what its initial value must be, and how to define the inductive step for the accumulator. In the example with the `lengthT` function, the accumulator trick works because of the special mathematical property of the expression being computed:

$$1 + (1 + (1 + (\dots + 0))) = (((0 + 1) + 1) + \dots) + 1 \quad .$$

This equation follows from the **associativity law** of addition. So, the computation can be rearranged to group all additions to the left. In code, it means that intermediate expressions are computed immediately before making recursive calls; this avoids the growth of the intermediate expressions.

Usually, the accumulator trick works because some associativity law is present. In that case, we are able to rearrange the order of recursive calls so that these calls always occur outside all other sub-expressions — that is, in tail positions. However, not all computations obey a suitable associativity law. Even if a code rearrangement exists, it may not be immediately obvious how to find it.

As an example, consider a tail-recursive re-implementation of the function `digitsToInt` from the previous subsection, where the recursive call is within a sub-expression `digitsToInt(xs) * 10 + x`. To

transform the code into a tail-recursive form, we need to rearrange the main computation,

$$r = d_{n-1} + 10 * (d_{n-2} + 10 * (d_{n-3} + 10 * (\dots + 10 * d_0))) \quad ,$$

so that the operations group to the left. We can do this by rewriting r as

$$r = ((d_0 * 10 + d_1) * 10 + \dots) * 10 + d_{n-1} \quad .$$

It follows that the digit sequence s must be split into the *leftmost* digit and the rest, $s = s.\text{head} ++ s.\text{tail}$. So, a tail-recursive implementation of the above formula is

```
@tailrec def fromDigits(s: Seq[Int], res: Int = 0): Int =
  // 'res' is the accumulator.
  if (s.isEmpty) res
  else fromDigits(s.tail, 10 * res + s.head)
```

Despite a certain similarity between this code and the code of `digitsToInt` from the previous subsection, the implementation `fromDigits` cannot be directly derived from the inductive definition of `digitsToInt`. One needs a separate proof that `fromDigits(s, 0)` computes the same result as `digitsToInt(s)`. This holds due to the following property:

Statement 2.2.3.1 For any $xs: \text{Seq[Int]}$ and $r: \text{Int}$, we have

$$\text{fromDigits}(xs, r) = \text{digitsToInt}(xs) + r * \text{math.pow}(10, s.\text{length})$$

Proof We prove this by induction. To shorten the proof, denote sequences by $[1, 2, 3]$ instead of `Seq(1, 2, 3)` and temporarily write $d(s)$ instead of `digitsToInt(s)` and $f(s, r)$ instead of `fromDigits(s, r)`. Then an inductive definition of $f(s, r)$ is

$$f([], r) = r \quad , \quad f([x]++s, r) = f(s, 10 * r + x) \quad . \quad (2.1)$$

Denoting the length of a sequence s by $|s|$, we reformulate Statement 2.2.3.1 as

$$f(s, r) = d(s) + r * 10^{|s|} \quad , \quad (2.2)$$

We prove Eq. (2.2) by induction. For the base case $s = []$, we have $f([], r) = r$ and $d([]) + r * 10^0 = r$ since $d([]) = 0$ and $|s| = 0$. The resulting equality $r = r$ proves the base case.

To prove the inductive step, we assume that Eq. (2.2) holds for a given sequence s ; then we need to prove that

$$f([x]++s, r) = d([x]++s) + r * 10^{|s|+1} \quad . \quad (2.3)$$

We will transform the left-hand side and the right-hand side separately, hoping to obtain the same expression. The left-hand side of Eq. (2.3):

$$\begin{aligned} & f([x]++s, r) \\ \text{use Eq. (2.1)}: &= f(s, 10 * r + x) \\ \text{use Eq. (2.2)}: &= d(s) + (10 * r + x) * 10^{|s|} \quad . \end{aligned}$$

The right-hand side of Eq. (2.3) contains $d([x]++s)$, which we somehow need to simplify. Assuming that $d(s)$ correctly calculates a number from its digits, we use a property of decimal notation: a digit x in front of n other digits has the value $x * 10^n$. This property can be formulated as an equation,

$$d([x]++s) = x * 10^{|s|} + d(s) \quad . \quad (2.4)$$

So, the right-hand side of Eq. (2.3) can be rewritten as

$$\begin{aligned} & d([x]++s) + r * 10^{|s|+1} \\ \text{use Eq. (2.4)}: &= x * 10^{|s|} + d(s) + r * 10^{|s|+1} \\ \text{factor out } 10^{|s|}: &= d(s) + (10 * r + x) * 10^{|s|} \quad . \end{aligned}$$

We have successfully transformed both sides of Eq. (2.3) to the same expression.

We have not yet proved that the function d satisfies the property in Eq. (2.4). The proof uses induction and begins by writing the code of d in a short notation,

$$d([]) = 0 \quad , \quad d(s++[y]) = d(s) * 10 + y \quad . \quad (2.5)$$

The base case is Eq. (2.4) with $s = []$. It is proved by

$$x = d([]++[x]) = d([x]++) = x * 10^0 + d([]) = x \quad .$$

The inductive step assumes Eq. (2.4) for a given x and a given sequence s , and needs to prove that for any y , the same property holds with $s++[y]$ instead of s :

$$d([x]++s++[y]) = x * 10^{|s|+1} + d(s++[y]) \quad . \quad (2.6)$$

The left-hand side of Eq. (2.6) is transformed into its right-hand side like this:

$$\begin{aligned} & d([x]++s++[y]) \\ \text{use Eq. (2.5)} : &= d([x]++s) * 10 + y \\ \text{use Eq. (2.4)} : &= (x * 10^{|s|} + d(s)) * 10 + y \\ \text{expand parentheses} : &= x * 10^{|s|+1} + d(s) * 10 + y \\ \text{use Eq. (2.5)} : &= x * 10^{|s|+1} + d(s++[y]) \quad . \end{aligned}$$

This demonstrates Eq. (2.6) and so concludes the proof.

2.2.4 Implementing general aggregation (foldLeft)

An **aggregation** converts a sequence of values into a single value. In general, the type of the result may be different from the type of sequence elements. To describe that general situation, we introduce type parameters, A and B , so that the input sequence is of type $\text{Seq}[A]$ and the aggregated value is of type B . Then an inductive definition of any aggregation function $f: \text{Seq}[A] \Rightarrow B$ looks like this:

- (Base case.) For an empty sequence, we have $f(\text{Seq}()) = b_0$, where $b_0: B$ is a given value.
- (Inductive step.) Assuming that $f(xs) = b$ is already computed, we define $f(xs :+ x) = g(x, b)$ where g is a given function with type signature $g: (A, B) \Rightarrow B$.

The code implementing f is written using recursion:

```
def f[A, B](s: Seq[A]): B =
  if (s.isEmpty) b0
  else g(s.last, f(s.take(s.length - 1)))
```

We can now refactor this code into a generic utility function, by making b_0 and g into parameters. A possible implementation is

```
def f[A, B](s: Seq[A], b: B, g: (A, B) => B): B =
  if (s.isEmpty) b
  else g(s.last, f(s.take(s.length - 1), b, g))
```

However, this implementation is not tail-recursive. Applying f to a sequence of, say, three elements, $\text{Seq}(x, y, z)$, will create an intermediate expression $g(z, g(y, g(x, b)))$. This expression will grow with the length of s , which is not acceptable. To rearrange the computation into a tail-recursive form, we need to start the base case at the innermost call $g(x, b)$, then compute $g(y, g(x, b))$ and continue. In other words, we need to traverse the sequence starting from its *leftmost* element x , rather than starting from the right. So, instead of splitting the sequence s into $s.take(s.length - 1) :+ s.last$ as we did in the code of f , we need to split s into $s.head +: s.tail$. Let us also exchange the order of the arguments of g , in order to be more consistent with the way this code is implemented in the Scala library. The resulting code is tail-recursive:

```
@tailrec def leftFold[A, B](s: Seq[A], b: B, g: (B, A) => B): B =
  if (s.isEmpty) b
  else leftFold(s.tail, g(b, s.head), g)
```

We call this function a “left fold” because it aggregates (or “folds”) the sequence starting from the leftmost element.

In this way, we have defined a general method of computing any inductively defined aggregation function on a sequence. The function `leftFold` implements the logic of aggregation defined via mathematical induction. Using `leftFold`, we can write concise implementations of methods such as `sum`, `max`, and many other aggregation functions. The method `leftFold` already contains all the code necessary to set up the base case and the inductive step. The programmer just needs to specify the expressions for the initial value `b` and for the updater function `g`.

As a first example, let us use `leftFold` for implementing the `sum` method:

```
def sum(s: Seq[Int]): Int = leftFold(s, 0, (x, y) => x + y)
```

To understand in detail how `leftFold` works, let us trace the evaluation of this function when applied to `Seq(1, 2, 3)`:

```
sum(Seq(1, 2, 3)) == leftFold(Seq(1, 2, 3), 0, g)
  // Here, g = (x, y) => x + y, so g(x, y) = x + y.
== leftFold(Seq(2, 3), g(0, 1), g)           // g(0, 1) = 1.
== leftFold(Seq(2, 3), 1, g)      // Now expand the code of 'leftFold'.
== leftFold(Seq(3), g(1, 2), g)    // g(1, 2) = 3; expand the code.
== leftFold(Seq(), g(3, 3), g)    // g(3, 3) = 6; expand the code.
== 6
```

The second argument of `leftFold` is the accumulator argument. The initial value of the accumulator is specified when first calling `leftFold`. At each iteration, the new accumulator value is computed by calling the updater function `g`, which uses the previous accumulator value and the value of the next sequence element. To visualize the process of recursive evaluation, it is convenient to write a table showing the sequence elements and the accumulator values as they are updated:

Current element x	Old accumulator value	New accumulator value
1	0	1
2	1	3
3	3	6

We implemented `leftFold` only as an illustration. Scala’s library has a method called `foldLeft` implementing the same logic using a slightly different type signature. To see this difference, compare the implementation of `sum` using our `leftFold` function and using the standard `foldLeft` method:

```
def sum(s: Seq[Int]): Int = leftFold(s, 0, (x, y) => x + y)

def sum(s: Seq[Int]): Int = s.foldLeft(0) { (x, y) => x + y }
```

The syntax of `foldLeft` makes it more convenient to use a nameless function as the updater argument of `foldLeft`, since curly braces separate that argument from others. We will use the standard `foldLeft` method from now on.

In general, the type of the accumulator value can be different from the type of the sequence elements. An example is an implementation of `count`:

```
def count[A](s: Seq[A], p: A => Boolean): Int =
  s.foldLeft(0) { (x, y) => x + (if (p(y)) 1 else 0) }
```

method works in the same way for all types of accumulators and all types of sequence elements.

The method `foldLeft` is available in the Scala library for all collections, including dictionaries and sets. Since `foldLeft` is tail-recursive, no stack overflows will occur even for very large sequences.

The accumulator is of type `Int`, while the sequence elements can have an arbitrary type, parameterized by `A`. The `foldLeft`

The Scala library contains several other methods similar to `foldLeft`, such as `foldRight` and `reduce`. (However, `foldRight` is not tail-recursive!)

2.2.5 Solved examples: using `foldLeft`

It is important to gain experience using the `foldLeft` method.

Example 2.2.5.1 Use `foldLeft` for implementing the `max` function for integer sequences. Return the special value `Int.MinValue` for empty sequences.

Solution Begin by writing an inductive formulation of the `max` function for sequences. Base case: For an empty sequence, return `Int.MinValue`. Inductive step: If `max` is already computed on a sequence `xs`, say `max(xs) = b`, the value of `max` on a sequence `xs :+ x` is `math.max(b, x)`. So, the code is:

```
def max(s: Seq[Int]): Int = s.foldLeft(Int.MinValue) { (b, x) => math.max(b, x) }
```

If we are sure that the function will never be called on empty sequences, we can implement `max` in a simpler way by using the `reduce` method:

```
def max(s: Seq[Int]): Int = s.reduce { (x, y) => math.max(x, y) }
```

Example 2.2.5.2 Implement the `count` method on sequences of type `Seq[A]`.

Solution Use the inductive definition of the function `count` as shown in Section 2.2.1 and write

```
def count[A](s: Seq[A], p: A => Boolean): Int =
  s.foldLeft(0) { (b, x) => b + (if (p(x)) 1 else 0) }
```

Example 2.2.5.3 Implement the function `digitsToInt` using `foldLeft`.

Solution The inductive definition of `digitsToInt` is directly translated into code:

```
def digitsToInt(d: Seq[Int]): Int = d.foldLeft(0) { (n, x) => n * 10 + x }
```

Example 2.2.5.4 For a given non-empty sequence `xs: Seq[Double]`, compute the minimum, the maximum, and the mean as a tuple $(x_{\min}, x_{\max}, x_{\text{mean}})$. The sequence should be traversed only once; i.e., the entire code must be `xs.foldLeft(...)`, using `foldLeft` only once.

Solution Without the requirement of using a single traversal, we would write

```
(xs.min, xs.max, xs.sum / xs.length)
```

However, this code traverses `xs` at least three times, since each of the aggregations `xs.min`, `xs.max`, and `xs.sum` iterates over `xs`. We need to combine the four inductive definitions of `min`, `max`, `sum`, and `length` into a single inductive definition of some function. What is the type of that function's return value? We need to accumulate intermediate values of *all four* numbers (`min`, `max`, `sum`, and `length`) in a tuple. So the required type of the accumulator is `(Double, Double, Double, Double)`. To avoid repeating a long type expression, we can define a type alias for it, say, `D4`:

```
scala> type D4 = (Double, Double, Double, Double)
defined type alias D4
```

The updater updates each of the four numbers according to the definitions of their inductive steps:

```
def update(p: D4, x: Double): D4 = p match {
  case (min, max, sum, length) =>
    (math.min(x, min), math.max(x, max), x + sum, length + 1)
}
```

Now we can write the code of the required function:

```
def f(xs: Seq[Double]): (Double, Double, Double) = {
  val init: D4 = (Double.PositiveInfinity, Double.NegativeInfinity, 0, 0)
  val (min, max, sum, length) = xs.foldLeft(init)(update)
  (min, max, sum/length)
}
```

```
scala> f(Seq(1.0, 1.5, 2.0, 2.5, 3.0))
res0: (Double, Double, Double) = (1.0,3.0,2.0)
```

Example 2.2.5.5* Implement the function `digitsToDouble` using `foldLeft`. The argument is of type `Seq[Char]`. As a test, the expression `digitsToDouble(Seq('3', '4', '.', '2', '5'))` must evaluate to 34.25. Assume that all input characters are either digits or a dot (so, negative numbers are not supported).

Solution The evaluation of a `foldLeft` on a sequence of digits will visit the sequence from left to right. The updating function should work the same as in `digitsToInt` until a dot character is found. After that, we need to change the updating function. So, we need to remember whether a dot character has been seen. The only way for `foldLeft` to “remember” any data is to hold that data in the accumulator value. We can choose the type of the accumulator according to our needs. So, for this task we can choose the accumulator to be a *tuple* that contains, for instance, the floating-point result constructed so far and a `Boolean` flag showing whether we have already seen the dot character.

To see what `digitsToDouble` must do, let us consider how the evaluation of `digitsToDouble(Seq('3', '4', '.', '2', '5'))` should go. We can write a table showing the intermediate result at each iteration. This will hopefully help us figure out what the accumulator and the updater function `g(...)` must be:

Current digit c	Previous result n	New result $n' = g(n, c)$
'3'	0.0	3.0
'4'	3.0	34.0
'.'	34.0	34.0
'2'	34.0	34.2
'5'	34.2	34.25

While the dot character was not yet seen, the updater function multiplies the previous result by 10 and adds the current digit. After the dot character, the updater function must add to the previous result the current digit divided by a factor that represents increasing powers of 10. In other words, the update computation $n' = g(n, c)$ must be defined by:

$$g(n, c) = \begin{cases} n * 10 + c & , \text{ if the digit is before the dot character.} \\ n + c/f & , \text{ if after the dot character, where } f = 10, 100, 1000, \dots \text{ for each new digit.} \end{cases}$$

The updater function g has only two arguments: the current digit and the previous accumulator value. So, the changing factor f must be *part of* the accumulator value, and must be multiplied by 10 at each digit after the dot. If the factor f is not a part of the accumulator value, the function g will not have enough information for computing the next accumulator value correctly. So, the updater computation must be $n' = g(n, c, f)$, not $n' = g(n, c)$.

For this reason, we choose the accumulator type as a tuple `(Double, Boolean, Double)` where the first number is the result n computed so far, the `Boolean` flag indicates whether the dot was already seen, and the third number is f , that is, the power of 10 by which the current digit will be divided if the dot was already seen. Initially, the accumulator tuple will be equal to `(0.0, false, 10.0)`. Then the updater function is implemented like this:

```
def update(acc: (Double, Boolean, Double), c: Char): (Double, Boolean, Double) =
  acc match { case (num, flag, factor) =>
    if (c == '.') (num, true, factor) // Set flag to 'true' after a dot character was seen.
    else {
      val digit = c - '0'
      if (flag) (num + digit / factor, flag, factor * 10) // This digit is after the dot.
      else (num * 10 + digit, flag, factor) // This digit is before the dot.
    }
  }
```

Now we can implement `digitsToDouble` as follows:

```

def digitsToDouble(d: Seq[Char]): Double = {
  val initAcc = (0.0, false, 10.0)
  val (num, _, _) = d.foldLeft(initAcc)(update)
  num
}

scala> digitsToDouble(Seq('3', '4', '.', '2', '5'))
res0: Double = 34.25

```

The result of calling `d.foldLeft` is a tuple `(num, flag, factor)`, in which only the first part, `num`, is needed. In Scala's pattern matching syntax, the underscore `(_)` denotes pattern variables whose values are not needed in the code. We could get the first part using the accessor method `._1`, but the code will be more readable if we show all parts of the tuple `(num, _, _)`.

Example 2.2.5.6 Implement the `map` method for sequences by using `foldLeft`. The input sequence should be of type `Seq[A]` and the output sequence of type `Seq[B]`, where `A` and `B` are type parameters. The required type signature of the function and a sample test:

```

def map[A, B](xs: Seq[A])(f: A => B): Seq[B] = ???

scala> map(List(1, 2, 3)) { x => x * 10 }
res0: Seq[Int] = List(10, 20, 30)

```

Solution The required code should build a new sequence by applying the function `f` to each element. How can we build a new sequence using `foldLeft`? The evaluation of `foldLeft` consists of iterating over the input sequence and accumulating some result value, which is updated at each iteration. Since the result of a `foldLeft` is always equal to the last computed accumulator value, it follows that the new sequence should be the accumulator value. So, we need to update the accumulator by appending the value `f(x)`, where `x` is the current element of the input sequence:

```
def map[A, B](xs: Seq[A])(f: A => B): Seq[B] = xs.foldLeft(Seq[B]()) { (acc, x) => acc :+ f(x) }
```

Example 2.2.5.7 Implement a function `toPairs` that converts a sequence of type `Seq[A]` to a sequence of pairs, `Seq[(A, A)]`, by putting together the adjacent elements pairwise. If the initial sequence has an odd number of elements, a given default value of type `A` is used to fill the last pair. The required type signature and an example test:

```

def toPairs[A](xs: Seq[A], default: A): Seq[(A, A)] = ???

scala> toPairs(Seq(1, 2, 3, 4, 5, 6), -1)
res0: Seq[(Int, Int)] = List((1,2), (3,4), (5,6))

scala> toPairs(Seq("a", "b", "c"), "<nothing>")
res1: Seq[(String, String)] = List((a,b), (c,<nothing>))

```

Solution We need to accumulate a sequence of pairs, and each pair needs two values. However, we iterate over values in the input sequence one by one. So, a new pair can be made only once every two iterations. The accumulator needs to hold the information about the current iteration being even or odd. For odd-numbered iterations, the accumulator also needs to store the previous element that is still waiting for its pair. Therefore, we choose the type of the accumulator to be a tuple `(Seq[(A, A)], Seq(A))`. The first sequence is the intermediate result, and the second sequence is the “holdover”: it holds the previous element for odd-numbered iterations and is empty for even-numbered iterations. Initially, the accumulator should be empty. An example evaluation is:

Current element x	Previous accumulator	Next accumulator
"a"	(Seq(), Seq())	(Seq(), Seq("a"))
"b"	(Seq(), Seq("a"))	(Seq(("a", "b")), Seq())
"c"	(Seq(("a", "b")), Seq())	(Seq(("a", "b")), Seq("c"))

Now it becomes clear how to implement the updater function:

```

type Acc = (Seq[(A, A)], Seq[A])      // Type alias, for brevity.
def updater(acc: Acc, x: A): Acc = acc match {

```

```

    case (result, Seq())      => (result, Seq(x))
    case (result, Seq(prev))  => (result :+ ((prev, x)), Seq())
}

```

We will call `foldLeft` with this updater and then perform some post-processing to make sure we create the last pair in case the last iteration is odd-numbered, i.e., when the “holdover” is not empty after `foldLeft` is finished. In this implementation, we use pattern matching to decide whether a sequence is empty:

```

def toPairs[A](xs: Seq[A], default: A): Seq[(A, A)] = {
  type Acc = (Seq[(A, A)], Seq[A])      // Type alias, for brevity.
  def init: Acc = (Seq(), Seq())
  def updater(acc: Acc, x: A): Acc = acc match {
    case (result, Seq())      => (result, Seq(x))
    case (result, Seq(prev))  => (result :+ ((prev, x)), Seq())
  }
  val (result, holdover) = xs.foldLeft(init)(updater)
  holdover match {      // May need to append the last element to the result.
    case Seq()      => result
    case Seq(x)    => result :+ ((x, default))
  }
}

```

This code shows examples of partial functions that are applied safely. One of these partial functions is used in the expression

```

holdover match {
  case Seq()      => ...
  case Seq(a)    => ...
}

```

This code works when `holdover` is empty or has length 1 but fails for longer sequences. In the implementation of `toPairs`, the value of `holdover` will always be a sequence of length at most 1, so it is safe to use this partial function.

2.2.6 Exercises: Using `foldLeft`

Exercise 2.2.6.1 Implement a function `fromPairs` that performs the inverse transformation to the `toPairs` function defined in Example 2.2.5.7. The required type signature and a sample test are:

```

def fromPairs[A](xs: Seq[(A, A)]): Seq[A] = ???

scala> fromPairs(Seq((1, 2), (3, 4)))
res0: Seq[Int] = List(1, 2, 3, 4)

```

Hint: This can be done with `foldLeft` or with `flatMap`.

Exercise 2.2.6.2 Implement the `flatten` method for sequences by using `foldLeft`. The required type signature and a sample test are:

```

def flatten[A](xxs: Seq[Seq[A]]): Seq[A] = ???

scala> flatten(Seq(Seq(1, 2, 3), Seq(), Seq(4)))
res0: Seq[Int] = List(1, 2, 3, 4)

```

Exercise 2.2.6.3 Use `foldLeft` to implement the `zipWithIndex` method for sequences. The required type signature and a sample test:

```

def zipWithIndex[A](xs: Seq[A]): Seq[(A, Int)] = ???

scala> zipWithIndex(Seq("a", "b", "c", "d"))
res0: Seq[String] = List((a, 0), (b, 1), (c, 2), (d, 3))

```

Exercise 2.2.6.4 Use `foldLeft` to implement a function `filterMap` that combines `map` and `filter` for sequences. The required type signature and a sample test:

```

def filterMap[A, B](xs: Seq[A])(pred: A => Boolean)(f: A => B): Seq[B] = ???

```

```
scala> filterMap(Seq(1, 2, 3, 4)) { x => x > 2 } { x => x * 10 }
res0: Seq[Int] = List(30, 40)
```

Exercise 2.2.6.5* Split a sequence into subsequences (“batches”) of length not larger than a given maximum length n . The required type signature and a sample test:

```
def byLength[A](xs: Seq[A], length: Int): Seq[Seq[A]] = ???

scala> byLength(Seq("a", "b", "c", "d"), 2)
res0: Seq[Seq[String]] = List(List(a, b), List(c, d))

scala> byLength(Seq(1, 2, 3, 4, 5, 6, 7), 3)
res1: Seq[Seq[Int]] = List(List(1, 2, 3), List(4, 5, 6), List(7))
```

Exercise 2.2.6.6* Split a sequence into batches by “weight” computed via a given function. The total weight of items in any batch should not be larger than a given maximum weight. The required type signature and a sample test:

```
def byWeight[A](xs: Seq[A], maxW: Double)(w: A => Double): Seq[Seq[A]] = ???

scala> byWeight((1 to 10).toList, 5.75){ x => math.sqrt(x) }
res0: Seq[Seq[Int]] = List(List(1, 2, 3), List(4, 5), List(6, 7), List(8), List(9), List(10))
```

Exercise 2.2.6.7* Use `foldLeft` to implement a `groupBy` function. The type signature and a test:

```
def groupBy[A, K](xs: Seq[A])(by: A => K): Map[K, Seq[A]] = ???

scala> groupBy(Seq(1, 2, 3, 4, 5)){ x => x % 2 }
res0: Map[Int, Seq[Int]] = Map(1 -> List(1, 3, 5), 0 -> List(2, 4))
```

Hints: The accumulator should be of type `Map[K, Seq[A]]`. To work with dictionaries, you will need to use the methods `getOrElse` and `updated`. The method `getOrElse` fetches a value from a dictionary by key, and returns the given default value if the dictionary does not contain that key:

```
scala> Map("a" -> 1, "b" -> 2).getOrElse("a", 300)
res0: Int = 1

scala> Map("a" -> 1, "b" -> 2).getOrElse("c", 300)
res1: Int = 300
```

The method `updated` produces a new dictionary that contains a new value for the given key, whether or not that key already exists in the dictionary:

```
scala> Map("a" -> 1, "b" -> 2).updated("c", 300) // Key is new.
res0: Map[String,Int] = Map(a -> 1, b -> 2, c -> 300)

scala> Map("a" -> 1, "b" -> 2).updated("a", 400) // Key already exists.
res1: Map[String,Int] = Map(a -> 400, b -> 2)
```

2.3 Converting a single value into a sequence

An aggregation converts or “folds” a sequence into a single value; the opposite operation (“unfolding”) converts a single value into a sequence. An example of this task is to compute the sequence of decimal digits for a given integer:

```
def digitsOf(x: Int): Seq[Int] = ???

scala> digitsOf(2405)
res0: Seq[Int] = List(2, 4, 0, 5)
```

We cannot implement `digitsOf` using `map`, `zip`, or `foldLeft`, because these methods work only if we *already have* a sequence; but the function `digitsOf` needs to create a new sequence. We could create a sequence via the expression `(1 to n)` if the required length of the sequence were known in advance. However, the function `digitsOf` must produce a sequence whose length is determined by a condition that we cannot easily evaluate in advance.

A general “unfolding” operation needs to build a sequence whose length is not determined in advance. This kind of sequence is called a **stream**. The elements of a stream are computed only when necessary (unlike the elements of `List` or `Array`, which are all computed in advance). The unfolding operation will compute the next element on demand; this creates a stream. We can then apply `takeWhile` to the stream, in order to stop it when a certain condition holds. Finally, if required, the truncated stream may be converted to a list or another type of sequence. In this way, we can generate a sequence of initially unknown length according to any given requirements.

The Scala library has a general stream-producing function `Stream.iterate`. This function has two arguments, the initial value and a function that computes the next value from the previous one:

```
scala> Stream.iterate(2) { x => x + 10 }
res0: Stream[Int] = Stream(2, ?)
```

The stream is ready to start computing the next elements of the sequence (so far, only the first element, 2, has been computed). In order to see the next elements, we need to stop the stream at a finite size and then convert the result to a list:

```
scala> Stream.iterate(2) { x => x + 10 }.take(6).toList
res1: List[Int] = List(2, 12, 22, 32, 42, 52)
```

If we try to evaluate `toList` on a stream without first limiting its size via `take` or `takeWhile`, the program will keep producing more elements of the stream until it runs out of memory and crashes.

Streams are similar to sequences, and methods such as `map`, `filter`, and `flatMap` are also defined for streams. For instance, the method `drop` skips a given number of initial elements:

```
scala> Seq(10, 20, 30, 40, 50).drop(3)
res2: Seq[Int] = List(40, 50)
```

```
scala> Stream.iterate(2) { x => x + 10 }.drop(3)
res3: Stream[Int] = Stream(32, ?)
```

This example shows that in order to evaluate `drop(3)`, the stream had to compute its elements up to 32 (but the subsequent elements are still not computed).

To figure out the code for `digitsOf`, we first write this function as a mathematical formula. To compute the digits for, say, $n = 2405$, we need to divide n repeatedly by 10, getting a sequence n_k of intermediate numbers ($n_0 = 2405, n_1 = 240, \dots$) and the corresponding sequence of last digits, $n_k \bmod 10$ (in this example: 5, 0, ...). The sequence n_k is defined using mathematical induction:

- Base case: $n_0 = n$, where n is the given initial integer.
- Inductive step: $n_{k+1} = \lfloor \frac{n_k}{10} \rfloor$ for $k = 1, 2, \dots$

Here $\lfloor \frac{n_k}{10} \rfloor$ is the mathematical notation for the integer division by 10. Let us tabulate the evaluation of the sequence n_k for $n = 2405$:

$k =$	0	1	2	3	4	5	6
$n_k =$	2405	240	24	2	0	0	0
$n_k \bmod 10 =$	5	0	4	2	0	0	0

The numbers n_k will remain all zeros after $k = 4$. It is clear that the useful part of the sequence is before it becomes all zeros. In this example, the sequence n_k needs to be stopped at $k = 4$. The sequence of digits then becomes [5, 0, 4, 2], and we need to reverse it to obtain [2, 4, 0, 5]. For reversing a sequence, the Scala library has the standard method `reverse`. A complete implementation for `digitsOf` is thus

```
def digitsOf(n: Int): Seq[Int] =
  if (n == 0) Seq(0) else { // n == 0 is a special case.
    Stream.iterate(n) { nk => nk / 10 }
      .takeWhile { nk => nk != 0 }
      .map { nk => nk % 10 }
      .toList.reverse
  }
```

We can shorten the code by using the syntax such as `(_ % 10)` instead of `{ nk => nk % 10 }`,

```
def digitsOf(n: Int): Seq[Int] =
  if (n == 0) Seq(0) else { // n == 0 is a special case.
```

```

Stream.iterate(n)(_ / 10)
  .takeWhile(_ != 0)
  .map(_ % 10)
  .toList.reverse
}

```

The type signature of the method `Stream.iterate` can be written as

```
def iterate[A](init: A)(next: A => A): Stream[A]
```

and shows a close correspondence to a definition by mathematical induction. The base case is the first value, `init`, and the inductive step is a function, `next`, that computes the next element from the previous one. It is a general way of creating sequences whose length is not determined in advance.

2.4 Transforming a sequence into another sequence

We have seen methods such as `map` and `zip` that transform sequences into sequences. However, these methods cannot express a general transformation where the elements of the new sequence are defined by induction and depend on previous elements. An example of this kind is computing the partial sums of a given sequence x_i , say $b_k = \sum_{i=0}^{k-1} x_i$. This formula defines $b_0 = 0$, $b_1 = x_0$, $b_2 = x_0 + x_1$, $b_3 = x_0 + x_1 + x_2$, etc. A definition via mathematical induction may be written like this:

- Base case: $b_0 = 0$.
- Inductive step: Given b_k , we define $b_{k+1} = b_k + x_k$ for $k = 0, 1, 2, \dots$

The Scala library method `scanLeft` implements a general sequence-to-sequence transformation defined in this way. The code implementing the partial sums is

```

def partialSums(xs: Seq[Int]): Seq[Int] = xs.scanLeft(0){ (x, y) => x + y }

scala> partialSums(Seq(1, 2, 3, 4))
res0: Seq[Int] = List(0, 1, 3, 6, 10)

```

The first argument of `scanLeft` is the base case, and the second argument is an updater function describing the inductive step. In general, the type of elements of the second sequence is different from that of the first sequence. The updater function takes an element of the first sequence and a previous element of the second sequence, and returns the next element of the second sequence. Note that the result of `scanLeft` is one element longer than the original sequence, because the base case provides an initial value.

Until now, we have seen that `foldLeft` is sufficient to re-implement almost every method that work on sequences, such as `map`, `filter`, or `flatten`. Let us show, as an illustration, how to implement the method `scanLeft` via `foldLeft`. In the implementation, the accumulator contains the previous element of the second sequence together with a growing fragment of that sequence, which is updated as we iterate over the first sequence. The code is

```

1 def scanLeft[A, B](xs: Seq[A])(b0: B)(next: (B, A) => B)
2   : Seq[B] = {
3   val init: (B, Seq[B]) = (b0, Seq(b0))
4   val (_, result) = xs.foldLeft(init) {
5     case ((b, seq), x) =>
6       val newB = next(b, x)
7       (newB, seq :+ newB)
8   }
9   result
10 }

```

To implement the (nameless) updater function for `foldLeft` in lines 5–8, we used the Scala feature that makes it easier to define functions with several arguments containing tuples. In our case, the updater function in `foldLeft` has two arguments: the first is a tuple `(B, Seq[B])`, the second is a value of type `A`. The pattern expression `case ((b, seq), x) =>` appears

to match a nested tuple. In reality, this expression matches the two arguments of the updater function and, at the same time, destructures the tuple argument as `(b, seq)`.

2.5 Summary

We have seen a broad overview of translating mathematical induction into Scala code.

What problems can we solve now?

- Compute mathematical expressions involving arbitrary recursion.
- Use the accumulator trick to enforce tail recursion.
- Implement functions with type parameters.
- Use arbitrary inductive (i.e., recursive) formulas to:
 - convert sequences to single values (aggregation or “folding”);
 - create new sequences from single values (“unfolding”);
 - transform existing sequences into new sequences.

Definition by induction	Scala code example
$f([]) = b ; f(s++[x]) = g(f(s), x)$	<code>f(xs) = xs.foldLeft(b)(g)</code>
$x_0 = b ; x_{k+1} = g(x_k)$	<code>xs = Stream.iterate(b)(g)</code>
$y_0 = b ; y_{k+1} = g(y_k, x_k)$	<code>ys = xs.scanLeft(b)(g)</code>

Table 2.1: Implementing mathematical induction.

Table 2.1 shows Scala code implementing those tasks. Iterative calculations are implemented by translating mathematical induction directly into code. In the functional programming paradigm, the programmer does not need to write any loops or use array indices. Instead, the programmer reasons about sequences as mathematical

values: “Starting from this value, we get that sequence, then transform it into this other sequence,” etc. This is a powerful way of working with sequences, dictionaries, and sets. Many kinds of programming errors (such as an incorrect array index) are avoided from the outset, and the code is shorter and easier to read than conventional code written using loops.

What tasks are not possible with these tools? We cannot implement a non-tail-recursive function without stack overflow (i.e., without unlimited growth of intermediate expressions). The accumulator trick does not always work! In some cases, it is impossible to implement tail recursion in a given recursive computation. An example of such a computation is the “merge-sort” algorithm where the function body must contain two recursive calls within a single expression. (It is impossible to rewrite *two* recursive calls as one.)

What if our recursive code cannot be transformed into tail-recursive code via the accumulator trick, but the recursion depth is so large that stack overflows occur? There exist special tricks (e.g., “continuations” and “trampolines”) that convert non-tail-recursive code into iterative code without stack overflows. Those techniques are beyond the scope of this chapter.

2.5.1 Solved examples

Example 2.5.1.1 Compute the smallest n such that $f(f(f(\dots f(1)\dots)) \geq 1000$, where the function f is applied n times. Write this as a function taking f , 1, and 1000 as arguments. Test with $f(x) = 2x + 1$.

Solution Define a stream of values $[1, f(1), f(f(1)), \dots]$ and use `takeWhile` to stop the stream when the values reach 1000. The number n is then found as the length of the resulting sequence plus 1:

```
scala> Stream.iterate(1)(x => 2 * x + 1).takeWhile(x => x < 1000).toList
res0: List[Int] = List(1, 3, 7, 15, 31, 63, 127, 255, 511)

scala> 1 + Stream.iterate(1)(x => 2 * x + 1).takeWhile(x => x < 1000).length
res1: Int = 10
```

Example 2.5.1.2 (a) For a given `Stream[Int]`, compute the stream of the largest values seen so far.

(b) Compute the stream of k largest values seen so far (k is a given integer parameter).

Solution: We cannot use `max` or sort the entire stream, since the length of the stream is not known in advance. So we need to use `scanLeft`, which will build the output stream one element at a time.

(a) Maintain the largest value seen so far in the accumulator of the `scanLeft`:

```
def maxSoFar(xs: Stream[Int]): Stream[Int] =
  xs.scanLeft(xs.head) { (max, x) => math.max(max, x) }.drop(1)
```

We use `.drop(1)` to remove the initial value, `xs.head`, because it is not useful for our result but is always produced by `scanLeft`.

To test this function, let us define a stream whose values go up and down:

```
val s = Stream.iterate(0)(x => 1 - 2 * x)

scala> s.take(10).toList
res0: List[Int] = List(0, 1, -1, 3, -5, 11, -21, 43, -85, 171)

scala> maxSoFar(s).take(10).toList
res1: List[Int] = List(0, 1, 1, 3, 3, 11, 11, 43, 43, 171)
```

(b) We again use `scanLeft`, where now the accumulator needs to keep the largest k values seen so far. There are two ways of maintaining this accumulator: First, to have a sequence of k values that we sort and truncate each time. Second, to use a specialized data structure such as a priority queue that automatically keeps values sorted and its length bounded. For the purposes of this tutorial, let us avoid using specialized data structures:

```
def maxKSoFar(xs: Stream[Int], k: Int): Stream[Seq[Int]] = {
  // The initial value of the accumulator is an empty Seq() of type Seq[Int].
  xs.scanLeft(Seq[Int]()) { (seq, x) =>
    // Sort in descending order, and take the first k values.
    (seq :+ x).sorted.reverse.take(k)
  }.drop(1) // Skip the undesired first value.
}

scala> maxKSoFar(s, 3).take(10).toList
res2: List[Seq[Int]] = List(List(0), List(1, 0), List(1, 0, -1), List(3, 1, 0), List(3, 1, 0),
  List(11, 3, 1), List(11, 3, 1), List(43, 11, 3), List(43, 11, 3), List(171, 43, 11))
```

Example 2.5.1.3 Find the last element of a non-empty sequence. (Hint: use `reduce`.)

Solution This function is available in the Scala library as the standard method `last` on sequences. Here we need to re-implement it using `reduce`. Begin by writing an inductive definition:

- (Base case.) `last(Seq(x)) == x`.
- (Inductive step.) `last(x +: xs) == last(xs)` assuming `xs` is non-empty.

The `reduce` method implements an inductive aggregation similarly to `foldLeft`, except that for `reduce` the base case always returns `x` for a 1-element sequence `Seq(x)`. This is exactly what we need here, so the inductive definition is directly translated into code, with the updater function $g(x, y) = y$:

```
def last[A](xs: Seq[A]): A = xs.reduce { (x, y) => y }
```

Example 2.5.1.4 (a) Count the occurrences of each distinct word in a string:

```
def countWords(s: String): Map[String, Int] = ???

scala> countWords("a quick a quick a brown a fox")
res0: Map[String, Int] = Map(a -> 4, quick -> 2, brown -> 1, fox -> 1)
```

(b) Count the occurrences of each distinct element in a sequence of type `Seq[A]`.

Solution (a) We split the string into an array of words via `s.split(" ")`, and apply a `foldLeft` to that array, since the computation is a kind of aggregation over the array of words. The accumulator of the aggregation will be the dictionary of word counts for all the words seen so far:

```
def countWords(s: String): Map[String, Int] = {
```

```
val init: Map[String, Int] = Map()
s.split(" ").foldLeft(init) { (dict, word) =>
  val newCount = dict.getOrElse(word, 0) + 1
  dict.updated(word, newCount)
}
}
```

An alternative, shorter implementation of the same function is

```
def countWords(s: String): Map[String, Int] = s.split(" ").groupBy(w => w).mapValues(_.length)
```

The `groupBy` creates a dictionary in one function call rather than one entry at a time. But the resulting dictionary contains word lists instead of word counts, so we use `mapValues`:

```
scala> "a a b b b c".split(" ").groupBy(w => w)
res0: Map[String,Array[String]] = Map(b -> Array(b, b, b), a -> Array(a, a), c -> Array(c))

scala> res0.mapValues(_.length)
res1: Map[String,Int] = Map(b -> 3, a -> 2, c -> 1)
```

(b) The main code of `countWords` does not depend on the fact that words are of type `String`. It will work in the same way for any other type of keys for the dictionary. So we keep the same code (except for renaming `word` to `x`) and replace `String` by a type parameter `A` in the type signature:

```
def countValues[A](xs: Seq[A]): Map[A, Int] =
  xs.foldLeft(Map[A, Int]()) { (dict, x) =>
    val newCount = dict.getOrElse(x, 0) + 1
    dict.updated(x, newCount)
}

scala> countValues(Seq(100, 100, 200, 100, 200, 200, 100))
res0: Map[Int,Int] = Map(100 -> 4, 200 -> 3)
```

Example 2.5.1.5 **(a)** Implement the binary search algorithm² for a sorted sequence `xs: Seq[Int]` as a function returning the index of the requested value `goal` (assume that `xs` always contains `goal`):

```
@tailrec def binSearch(xs: Seq[Int], goal: Int): Int = ???

scala> binSearch(Seq(1, 3, 5, 7), 5)
res0: Int = 2
```

(b) Re-implement `binSearch` using `Stream.iterate` without writing explicitly recursive code.

Solution **(a)** The binary search algorithm splits the array into two halves and may continue the search recursively in one of the halves. We need to write the solution as a tail-recursive function with an additional accumulator argument. So we expect that the code should look like this:

```
@tailrec def binSearch(xs: Seq[Int], goal: Int, acc: _ = ???): Int = {
  if (????) acc // This condition must decide whether we are finished.
  else {
    // Determine which half of the sequence contains 'goal'.
    // Then update the accumulator accordingly.
    val newAcc = ???
    binSearch(xs, goal, newAcc) // Tail-recursive call.
  }
}
```

We will first decide the type and the initial value of the accumulator, then implement the updater.

The information required for the recursive call must show the segment of the sequence where the target number is present. That segment is defined by two indices i, j representing the left and the right bounds of the sub-sequence, such that the target element is x_n with $x_i \leq x_n < x_{j-1}$. It follows that the accumulator should be a pair of two integers (i, j) . The initial value of the accumulator is the pair $(0, N)$, where N is the length of the entire sequence. The search is finished when $i + 1 = j$.

²https://en.wikipedia.org/wiki/Binary_search_algorithm

For convenience, let us introduce *two* accumulator values representing (i, j) :

```
@tailrec def binSearch(xs: Seq[Int], goal: Int)(left: Int = 0, right: Int = xs.length): Int = {
  // Check whether 'goal' is at one of the boundaries.
  if (right - left <= 1 || xs(left) == goal) left
  else {
    val middle = (left + right) / 2
    // Determine which half of the array contains 'target'.
    // Update the accumulator accordingly.
    val (newLeft, newRight) =
      if (goal < xs(middle)) (left, middle)
      else (middle, right)
    binSearch(xs, goal)(newLeft, newRight) // Tail-recursive call.
  }
}

scala> binSearch(0 to 10, 3)() // Default accumulator values.
res0: Int = 3
```

Here we used a feature of Scala that allows us to set `xs.length` as a default value for the argument `right` of `binSearch`. This works because `right` is in a different **argument list** from `xs`. Default values in an argument list may depend on arguments in a *previous* argument list. However, the code

```
def binSearch(xs: Seq[Int], goal: Int, left: Int = 0, right: Int = xs.length)
```

will generate an error: the arguments in the same argument list cannot depend on each other. (The error will say `not found: value xs`.)

(b) We can visualize the binary search as a procedure that generates a sequence of progressively tighter bounds for the location of `goal`. The initial bounds are $(0, xs.length)$, and the final bounds are $(k, k+1)$ for some k . We can generate the sequence of bounds using `Stream.iterate` and stop the sequence when the bounds become sufficiently tight. To make the use of `takeWhile` more convenient, we add an extra sequence element where the bounds (k, k) are equal. The code becomes

```
def binSearch(xs: Seq[Int], goal: Int): Int = {
  type Acc = (Int, Int)
  val init: Acc = (0, xs.length)
  val updater: Acc => Acc = { case (left, right) =>
    if (right - left <= 1) (left, left) // Extra element (k, k) in the stream.
    else if (xs(left) == goal) (left, left + 1)
    else {
      val middle = (left + right) / 2
      // Determine which half of the array contains 'target'.
      // Update the accumulator accordingly.
      if (goal < xs(middle)) (left, middle)
      else (middle, right)
    }
  }
  Stream.iterate(init)(updater)
    .takeWhile{ case (left, right) => right > left }
    .last._1 // Take the 'left' boundary from the last element.
}
```

This code is clearer because recursion is delegated to `Stream.iterate` and cleanly separated from the “business logic” (i.e., implementing the base case, the inductive step, and the post-processing).

Example 2.5.1.6 For a given positive $n: \text{Int}$, compute the sequence $[s_0, s_1, s_2, \dots]$ defined by $s_0 = SD(n)$ and $s_k = SD(s_{k-1})$ for $k > 0$, where $SD(x)$ is the sum of the decimal digits of the integer x , e.g., $SD(123) = 6$. Stop the sequence s_i when the numbers begin repeating. For example, $SD(99) = 18$, $SD(18) = 9$, $SD(9) = 9$. So, for $n = 99$, the sequence s_i must be computed as $[99, 18, 9]$.

Hint: use `Stream.iterate`; compute the digits in the reverse order since their sum will be the same.

Solution We need to implement a function `sdSeq` having the type signature

```
def sdSeq(n: Int): Seq[Int]
```

First we need to implement $SD(x)$. The sum of digits is obtained similarly to Section 2.3:

```
def SD(n: Int): Int = if (n == 0) 0 else Stream.iterate(n)(_ / 10).takeWhile(_ != 0).map(_ % 10).sum
```

Now we can try evaluating SD on some numbers to see its behavior:

```
scala> (1 to 15).toList.map(SD)
res0: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 2, 3, 4, 5, 6)
```

It is clear that $SD(n) < n$ as long as $n \geq 10$. So the sequence elements s_i will not repeat until they become smaller than 10, and then they will always repeat. This seems to be an easy way of stopping the sequence. Let us try that:

```
scala> Stream.iterate(99)(SD).takeWhile(x => x >= 10).toList
res1: List[Int] = List(99, 18)
```

We are missing the last element of the sequence, $SD(18) = 9$, because `takeWhile` stops the sequence too early. In order to obtain the correct sequence, we need to compute one more element. To fix this, we can generate a stream of pairs:

```
scala> Stream.iterate((0, 99)){ case (prev, x) => (x, SD(x)) }.
  takeWhile{ case (prev, x) => prev >= 10 || x >= 10 }.toList
res2: List[(Int, Int)] = List((0,99), (99,18), (18,9))
```

This looks right; it remains to remove the first parts of the tuples:

```
def sdSeq(n: Int): Seq[Int] =
  Stream.iterate((0, n)){ case (prev, x) => (x, SD(x)) } // Stream[(Int, Int)]
    .takeWhile{ case (prev, x) => prev >= 10 || x >= 10 } // Stream[(Int, Int)]
    .map(_._2) // Stream[Int]
    .toList // List[Int]

scala> sdSeq(99)
res3: Seq[Int] = List(99, 18, 9)
```

Example 2.5.1.7 Implement a function `unfold` with the type signature

```
def unfold[A](init: A)(next: A => Option[A]): Stream[A]
```

The function should create a stream of values of type `A` with the initial value `init`. Next elements are computed from previous ones via the function `next` until it returns `None`. An example test:

```
scala> unfold(0) { x => if (x > 5) None else Some(x + 2) }
res0: Stream[Int] = Stream(0, ?)

scala> res0.toList
res1: List[Int] = List(0, 2, 4, 6)
```

Solution We can formulate the task as an inductive definition of a stream. If `next(init) == None`, the stream must stop at `init`. (This is the base case of the induction). Otherwise, `next(init) == Some(x)` yields a new value `x` and indicates that we need to continue to “unfold” the stream with `x` instead of `init`. (This is the inductive step.) Streams can be created from individual values via the Scala standard library method `Stream.cons` that constructs a stream from a single value and a tail:

```
def unfold[A](init: A)(next: A => Option[A]): Stream[A] = next(init) match {
  case None      => Stream(init) // A stream containing a single value ‘init’.
  case Some(x)   => Stream.cons(init, unfold(x)(next)) // ‘init’ followed by the tail of stream.
}
```

Example 2.5.1.8 For a given stream $[s_0, s_1, s_2, \dots]$ of type `Stream[T]`, compute the “half-speed” stream $h = [s_0, s_0, s_1, s_1, s_2, s_2, \dots]$. The half-speed sequence h is defined as $h_{2k} = h_{2k+1} = s_k$ for $k = 0, 1, 2, \dots$

Solution We use `map` to replace each element s_i by a sequence containing two copies of s_i . Let us try this on a sample sequence:

```
scala> Seq(1,2,3).map( x => Seq(x, x))
```

```
res0: Seq[Seq[Int]] = List(List(1, 1), List(2, 2), List(3, 3))
```

The result is almost what we need, except we need to `flatten` the nested list:

```
scala> Seq(1,2,3).map( x => Seq(x, x)).flatten
res1: Seq[Seq[Int]] = List(1, 1, 2, 2, 3, 3)
```

The composition of `map` and `flatten` is `flatMap`, so the final code is

```
def halfSpeed[T](str: Stream[T]): Stream[T] = str.flatMap(x => Seq(x, x))

scala> halfSpeed(Seq(1,2,3).toStream)
res2: Stream[Int] = Stream(1, ?)

scala> halfSpeed(Seq(1,2,3).toStream).toList
res3: List[Int] = List(1, 1, 2, 2, 3, 3)
```

Example 2.5.1.9 (The loop detection problem.) Stop a given stream $[s_0, s_1, s_2, \dots]$ at a place k where the sequence repeats itself; that is, an element s_k equals some earlier element s_i with $i < k$.

Solution The trick is to create a half-speed sequence h_i out of s_i and then find an index $k > 0$ such that $h_k = s_k$. (The condition $k > 0$ is needed because we will always have $h_0 = s_0$.) If we find such an index k , it would mean that either $s_k = s_{k/2}$ or $s_k = s_{(k-1)/2}$; in either case, we will have found an element s_k that equals an earlier element.

As an example, for an input sequence $s = [1, 3, 5, 7, 9, 3, 5, 7, 9, \dots]$ we obtain the half-speed sequence $h = [1, 1, 3, 3, 5, 5, 7, 7, 9, 9, 3, 3, \dots]$. Looking for an index $k > 0$ such that $h_k = s_k$, we find that $s_7 = h_7 = 7$. The element s_7 indeed repeats an earlier element (although s_7 is not the first such repetition).

There are in principle two ways of finding an index $k > 0$ such that $h_k = s_k$: First, to iterate over a list of indices $k = 1, 2, \dots$ and evaluate the condition $h_k = s_k$ as a function of k . Second, to build a sequence of pairs (h_i, s_i) and use `takeWhile` to stop at the required index. In the present case, we cannot use the first way because we do not have a fixed set of indices to iterate over. Also, the condition $h_k = s_k$ cannot be directly evaluated as a function of k because s and h are streams that compute elements on demand, not lists whose elements are computed in advance and ready for use.

So the code must iterate over a stream of pairs (h_i, s_i) :

```
def stopRepeats[T](str: Stream[T]): Stream[T] = {
  val halfSpeed = str.flatMap(x => Seq(x, x))
  val result = halfSpeed.zip(str) // Stream[(T, T)]
  .drop(1) // Enforce the condition k > 0.
  .takeWhile { case (h, s) => h != s } // Stream[(T, T)]
  .map(_._2) // Stream[T]
  str.head +: result // Prepend the first element that was dropped.
}

scala> stopRepeats(Seq(1, 3, 5, 7, 9, 3, 5, 7, 9).toStream).toList
res0: List[Int] = List(1, 3, 5, 7, 9, 3, 5)
```

Example 2.5.1.10 Reverse each word in a string, but keep the order of words:

```
def revWords(s: String): String = ???

scala> revWords("A quick brown fox")
res0: String = A kciuq nworb xof
```

Solution The standard method `split` converts a string into an array of words:

```
scala> "pa re ci vo mu".split(" ")
res0: Array[String] = Array(pa, re, ci, vo, mu)
```

Each word is reversed with `reverse`; the resulting array is concatenated into a string with `mkString`:

```
def revWords(s: String): String = s.split(" ").map(_.reverse).mkString(" ")
```

Example 2.5.1.11 Remove adjacent repeated characters from a string:

```
def noDups(s: String): String = ???

scala> noDups("abbcdeeeeefddgggggh")
res0: String = abcdefdgh
```

Solution A string is automatically converted into a sequence of characters when we use methods such as `map` or `zip` on it. So, we can use `s.zip(s.tail)` to get a sequence of pairs (s_k, s_{k+1}) where s_k is the k -th character of the string s . A `filter` will then remove elements s_k for which $s_{k+1} = s_k$:

```
scala> val s = "abbcd"
s: String = abbcd

scala> s.zip(s.tail).filter { case (sk, skPlus1) => sk != skPlus1 }
res0: IndexedSeq[(Char, Char)] = Vector((a,b), (b,c), (c,d))
```

It remains to convert this sequence of pairs into the string `"abcd"`. One way of doing this is to project the sequence of pairs onto the second parts of the pairs,

```
scala> res0.map(_._2).mkString
res1: String = bcd
```

We just need to add the first character, `'a'`. The resulting code is

```
def noDups(s: String): String = if (s == "") "" else {
  val pairs = s.zip(s.tail).filter { case (x, y) => x != y }
  pairs.head._1 +: pairs.map(_._2).mkString
}
```

The method `:+` prepends an element to a sequence, so `x +: xs` is equivalent to `Seq(x) ++ xs`.

Example 2.5.1.12 For a given sequence of type `Seq[A]`, find the longest subsequence that does not contain any adjacent duplicate values.

```
def longestNoDups[A](xs: Seq[A]): Seq[A] = ???

scala> longestNoDups(Seq(1, 2, 2, 5, 4, 4, 4, 8, 2, 3, 3))
res0: Seq[Int] = List(4, 8, 2, 3)
```

Solution This is a dynamic programming³ problem. Many such problems are solved with a single `foldLeft`. The accumulator represents the current “state” of the dynamic programming solution, and the “state” is updated with each new element of the input sequence.

We first need to determine the type of the accumulator value, or the “state”. The task is to find the longest subsequence without adjacent duplicates. So the accumulator should represent the longest subsequence found so far, as well as any required extra information about other subsequences that might grow as we iterate over the elements of `xs`. What is that extra information in our case?

Imagine creating the set of *all* subsequences that have no adjacent duplicates. For the input sequence `[1, 2, 2, 5, 4, 4, 4, 8, 2, 3, 3]`, this set of all subsequences will be $\{[1, 2], [2, 5, 4], [4, 8, 2, 3]\}$. We can build this set incrementally in the accumulator value of a `foldLeft`. To visualize how this set would be built, consider the partial result after seeing the first 8 elements of the input sequence, `[1, 2, 2, 5, 4, 4, 4, 8]`. The partial set of non-repeating subsequences is $\{[1, 2], [2, 5, 4], [4, 8]\}$. When we see the next element, 2, we will update that partial set to $\{[1, 2], [2, 5, 4], [4, 8, 2]\}$.

It is now clear that the subsequence `[1, 2]` has no chance of being the longest subsequence, since `[2, 5, 4]` is already longer. However, we do not yet know whether `[2, 5, 4]` or `[4, 8, 2]` is the winner, because the subsequence `[4, 8, 2]` could still grow and become the longest one (and it does become `[4, 8, 2, 3]` later). At this point, we need to keep both of these two subsequences in the accumulator, but we may already discard `[1, 2]`.

We have deduced that the accumulator needs to keep only *two* sequences: the first sequence is already terminated and will not grow, the second sequence ends with the current element and may

³https://en.wikipedia.org/wiki/Dynamic_programming

yet grow. The initial value of the accumulator is empty. The first subsequence is discarded when it becomes shorter than the second. The code can be written now:

```
def longestNoDups[A](xs: Seq[A]): Seq[A] = {
  val init: (Seq[A], Seq[A]) = (Seq(), Seq())
  val (first, last) = xs.foldLeft(init) { case ((first, current), x) =>
    // If 'current' is empty, 'x' is not considered to be repeated.
    val xWasRepeated = current != Seq() && current.last == x
    val firstIsLongerThanCurrent = first.length > current.length
    // Compute the new pair '(first, current)'.
    // Keep 'first' only if it is longer; otherwise replace it by 'current'.
    val newFirst = if (firstIsLongerThanCurrent) first else current
    // Append 'x' to 'current' if 'x' is not repeated.
    val newCurrent = if (xWasRepeated) Seq(x) else current :+ x
    (newFirst, newCurrent)
  }
  // Return the longer of the two subsequences; prefer 'first'.
  if (first.length >= last.length) first else last
}
```

2.5.2 Exercises

Exercise 2.5.2.1 Compute the sum of squared digits of a given integer; e.g., `dsq(123) = 14` (see Example 2.5.1.6). Generalize the solution to take as an argument an function `f: Int => Int` replacing the squaring operation. The required type signature and a sample test:

```
def digitsMapSum(x: Int)(f: Int => Int): Int = ???

scala> digitsMap(123){ x => x * x }
res0: Int = 14

scala> digitsMap(123){ x => x * x * x }
res1: Int = 36
```

Exercise 2.5.2.2 Compute the **Collatz sequence** c_i as a stream defined by

$$c_0 = n \quad ; \quad c_{k+1} = \begin{cases} c_k/2 & \text{if } c_k \text{ is even,} \\ 3 * c_k + 1 & \text{if } c_k \text{ is odd.} \end{cases}$$

Stop the stream when it reaches 1 (as we would expect⁴ it will).

Exercise 2.5.2.3 For a given integer n , compute the sum of cubed digits, then the sum of cubed digits of the result, etc.; stop the resulting sequence when it repeats itself, and so determine whether it ever reaches 1. (Use Exercise 2.5.2.1.)

```
def cubes(n: Int): Stream[Int] = ???

scala> cubes(123).take(10).toList
res0: List[Int] = List(123, 36, 243, 99, 1458, 702, 351, 153, 153, 153)

scala> cubes(2).take(10).toList
res1: List[Int] = List(2, 8, 512, 134, 92, 737, 713, 371, 371, 371)

scala> cubes(4).take(10).toList
res2: List[Int] = List(4, 64, 280, 520, 133, 55, 250, 133, 55, 250)

def cubesReach1(n: Int): Boolean = ???

scala> cubesReach1(10)
res3: Boolean = true
```

⁴https://en.wikipedia.org/wiki/Collatz_conjecture

```
scala> cubesReach1(4)
res4: Boolean = false
```

Exercise 2.5.2.4 For a, b, c of type `Set[Int]`, compute the set of all sets of the form `Set(x, y, z)` where x is from a , y from b , and z from c . The required type signature and a sample test:

```
def prod3(a: Set[Int], b: Set[Int], c: Set[Int]): Set[Set[Int]] = ???

scala> prod3(Set(1,2), Set(3), Set(4,5))
res0: Set[Set[Int]] = Set(Set(1,3,4), Set(1,3,5), Set(2,3,4), Set(2,3,5))
```

Hint: use `flatMap`.

Exercise 2.5.2.5* Same task as in Exercise 2.5.2.4 for a set of sets: instead of just three sets a, b, c , a `Set[Set[Int]]` is given. The required type signature and a sample test:

```
def prodSet(si: Set[Set[Int]]): Set[Set[Int]] = ???

scala> prodSet(Set(Set(1,2), Set(3), Set(4,5), Set(6)))
res0: Set[Set[Int]] = Set(Set(1,3,4,6), Set(1,3,5,6), Set(2,3,4,6), Set(2,3,5,6))
```

Hint: use `foldLeft` and `flatMap`.

Exercise 2.5.2.6* In a sorted array `xs: Array[Int]` where no values are repeated, find all pairs of values whose sum equals a given number n . Use tail recursion. A type signature and a sample test:

```
def pairs(goal: Int, xs: Array[Int]): Set[(Int, Int)] = ???

scala> pairs(10, Array(1, 2, 3, 4, 5, 6, 7, 8))()
res0: Set[(Int, Int)] = Set((2,8), (3,7), (4,6), (5,5))
```

Exercise 2.5.2.7 Reverse a sentence's word order, but keep the words unchanged:

```
def revSentence(s: String): String = ???

scala> revSentence("A quick brown fox")
res0: String = "fox brown quick A"
```

Exercise 2.5.2.8 (a) Reverse an integer's digits (see Example 2.5.1.6) as shown:

```
def revDigits(n: Int): Int = ???

scala> revDigits(12345)
res0: Int = 54321
```

(b) A **palindrome integer** is an integer number n such that `revDigits(n) == n`. Write a predicate function of type `Int => Boolean` that checks whether a given positive integer is a palindrome.

Exercise 2.5.2.9 Define a function `findPalindrome: Long => Long` performing the following computation: First define $f(n) = \text{revDigits}(n) + n$ for a given integer n , where the function `revDigits` was defined in Exercise 2.5.2.8. If $f(n)$ is a palindrome integer, `findPalindrome` returns that integer. Otherwise, it keeps applying the same transformation and computes $f(n), f(f(n)), \dots$, until a palindrome integer is eventually found (this is mathematically guaranteed). A sample test:

```
scala> findPalindrome(10101)
res0: Long = 10101

scala> findPalindrome(123)
res0: Long = 444

scala> findPalindrome(83951)
res1: Long = 869363968
```

Exercise 2.5.2.10 Transform a given sequence `xs: Seq[Int]` into a sequence `Seq[(Int, Int)]` of pairs that skip one neighbor. Implement this transformation as a function `skip1` with a type parameter `A` instead of the type `Int`. The required type signature and a sample test:

```
def skip1[A](xs: Seq[A]): Seq[(A, A)] = ???

scala> skip1(List(1,2,3,4,5))
res0: List[Int] = List((1,3), (2,4), (3,5))
```

Exercise 2.5.2.11 (a) For a given integer interval $[n_1, n_2]$, find the largest integer $k \in [n_1, n_2]$ such that the decimal representation of k does *not* contain any of the digits 3, 5, or 7. (b) For a given integer interval $[n_1, n_2]$, find the integer $k \in [n_1, n_2]$ with the largest sum of decimal digits. (c) A positive integer n is called a **perfect number** if it is equal to the sum of its divisors (other integers k such that $k < n$ and n/k is an integer). For example, 6 is a perfect number because its divisors are 1, 2, and 3, and $1 + 2 + 3 = 6$, while 8 is not a perfect number because its divisors are 1, 2, and 4, and $1 + 2 + 4 = 7 \neq 8$. Write a function that determines whether a given number n is perfect. Determine all perfect numbers up to one million.

Exercise 2.5.2.12 Remove adjacent repeated elements from a sequence of type `Seq[A]` when they are repeated more than k times. Repetitions up to k times should remain unchanged. The required type signature and a sample test:

```
def removeDups[A](s: Seq[A], k: Int): Seq[A] = ???

scala> removeDups(Seq(1, 1, 1, 1, 5, 2, 2, 5, 5, 5, 5, 5, 1), 3)
res0: Seq[Int] = List(1, 1, 1, 5, 2, 2, 5, 5, 5, 1)
```

Exercise 2.5.2.13 Implement a function `unfold2` with the type signature

```
def unfold2[A,B](init: A)(next: A => Option[(A,B)]): Stream[B]
```

The function should create a stream of values of type `B` by repeatedly applying the given function `next` until it returns `None`. At each iteration, `next` should be applied to the value of type `A` returned by the previous call to `next`. An example test:

```
scala> unfold2(0) { x => if (x > 5) None else Some((x + 2, s"had $x")) }
res0: Stream[String] = Stream(had 0, ?)

scala> res0.toList
res1: List[String] = List(had 0, had 2, had 4)
```

Exercise 2.5.2.14* (a) Remove repeated elements (whether adjacent or not) from a sequence of type `Seq[A]`. (This re-implements the standard library's method `distinct`.)

(b) For a sequence of type `Seq[A]`, remove all elements that are repeated (whether adjacent or not) more than k times:

```
def removeK[A](k: Int, xs: Seq[A]): Seq[A] = ???

scala> removeK(2, Seq("a", "b", "a", "b", "b", "c", "b", "a"))
res0: Seq[String] = List(a, b, a, b, c)
```

Exercise 2.5.2.15* For a given sequence `xs: Seq[Double]`, find a subsequence that has the largest sum of values. The sequence `xs` is not sorted, and its values may be positive or negative. The required type signature and a sample test:

```
def maxsub(xs: Seq[Double]): Seq[Double] = ???

scala> maxsub(Seq(1.0, -1.5, 2.0, 3.0, -0.5, 2.0, 1.0, -10.0, 2.0))
res0: Seq[Double] = List(2.0, 3.0, -0.5, 2.0, 1.0)
```

Hint: use dynamic programming and `foldLeft`.

Exercise 2.5.2.16* Using tail recursion, find all common integers between two *sorted* sequences:

```
@tailrec def commonInt(xs: Seq[Int], ys: Seq[Int]): Seq[Int] = ???  
  
scala> commonInt(Seq(1, 3, 5, 7), Seq(2, 3, 4, 6, 7, 8))  
res0: Seq[Int] = List(3, 7)
```

2.6 Discussion and further developments

2.6.1 Total and partial functions

In Scala, functions can be total or partial. A **total** function will always compute a result value, while a **partial** function may fail to compute its result for certain values of its arguments.

A simple example of a partial function in Scala is the `max` method: it only works for non-empty sequences. Trying to evaluate it on an empty sequence generates an error called an “exception”:

```
scala> Seq(1).tail  
res0: Seq[Int] = List()  
scala> res0.max  
java.lang.UnsupportedOperationException: empty.max  
at scala.collection.TraversableOnce$class.max(TraversableOnce.scala:229)  
at scala.collection.AbstractTraversable.max(Traversable.scala:104)  
... 32 elided
```

This kind of error may crash the entire program at run time. Unlike the type errors we saw before, which occur at compilation time (i.e., before the program can start), **run-time errors** occur while the program is running, and only when some partial function happens to get an incorrect input. The incorrect input may occur at any point after the program started running, which may crash the entire program in the middle of a long computation.

So, it seems clear that we should write code that does not generate such errors. For instance, it is safe to apply `max` to a sequence if we know that it is non-empty.

Sometimes, a function that uses pattern matching turns out to be a partial function because its pattern matching code fails on certain input data.

If a pattern matching expression fails, the code will throw an exception and stop running. In functional programming, we usually want to avoid this situation because it makes it much harder to reason about program correctness. In most cases, programs can be written to avoid the possibility of match errors. An example of an unsafe pattern matching expression is

```
def h(p: (Int, Int)): Int = p match { case (x, 0) => x }  
  
scala> h( (1,0) )  
res0: Int = 1  
  
scala> h( (1,2) )  
scala.MatchError: (1,2) (of class scala.Tuple2$mcII$sp)  
at .h(<console>:12)  
... 32 elided
```

Here the pattern contains a pattern variable `x` and a constant `0`. This pattern only matches tuples whose second part is equal to `0`. If the second argument is nonzero, a match error occurs and the program crashes. So, `h` is a partial function.

Pattern matching failures never happen if we match a tuple of correct size with a pattern such as `(x, y, z)`, because a pattern variable will always match a value. So, pattern matching with a pattern such as `(x, y, z)` is **infallible** (never fails at run time) when applied to a tuple with 3 elements.

Another way in which pattern matching can be made infallible is by including a pattern that matches everything:

```
p match {  
  case (x, 0) => ... // This only matches some tuples.
```

```
case _          => ... // This matches everything.
}
```

If the first pattern `(x, 0)` fails to match the value `p`, the second pattern will be tried (and will always succeed). The `case` patterns in a `match` expression are tried in the order they are written. So, a `match` expression may be made infallible by adding a “match-all” underscore pattern.

2.6.2 Scope and shadowing of pattern matching variables

Pattern matching introduces **locally scoped** variables — that is, variables defined only on the right-hand side of the pattern match expression. As an example, consider this code:

```
def f(x: (Int, Int)): Int = x match { case (x, y) => x + y }

scala> f( (2,4) )
res0: Int = 6
```

The argument of `f` is the variable `x` of a tuple type `(Int, Int)`, but there is also a pattern variable `x` in the case expression. The pattern variable `x` matches the first part of the tuple and has type `Int`. Because variables are locally scoped, the pattern variable `x` is only defined within the expression `x + y`. The argument `x: (Int, Int)` is a completely different variable whose value has a different type.

The code works correctly but is confusing to read because of the name clash between the two quite different variables, both named `x`. Another negative consequence of the name clash is that the argument `x: (Int, Int)` is *invisible* within the case expression: if we write “`x`” in that expression, we will get the pattern variable `x: Int`. One says that the argument `x: (Int, Int)` has been **shadowed** by the pattern variable `x` (which is a “bound variable” inside the case expression).

The problem is easy to avoid: we can give the pattern variable another name. Since the pattern variable is locally scoped, it can be renamed within its scope without affecting any other code:

```
def f(x: (Int, Int)): Int = x match { case (a, b) => a + b }

scala> f( (2,4) )
res0: Int = 6
```

2.6.3 Lazy values and sequences. Iterators and streams

We have used streams to create sequences whose length is not known in advance. An example is a stream containing a sequence of increasing positive integers:

```
scala> val p = Stream.iterate(1)(_ + 1)
p: Stream[Int] = Stream(1, ?)
```

At this point, we have not defined a stopping condition for this stream. In some sense, streams may be seen as “infinite” sequences, although in practice a stream is always finite because computers cannot run infinitely long. Also, computers cannot store infinitely many values in memory.

More precisely, streams are “partially computed” rather than “infinite”. The main difference between arrays and streams is that a stream’s elements are computed on demand and not all initially available, while an array’s elements are all computed in advance and are immediately available.

Generally, there are four possible ways a value could be available:

Availability	Explanation	Example Scala code
“eager”	computed immediately	<code>val z = f(123)</code>
“lazy”	computed upon first use	<code>lazy val z = f(123)</code>
“on-call”	computed each time it is used	<code>def z = f(123)</code>
“never”	cannot be computed due to errors	<code>val (x, y) = "abc"</code>

A **lazy value** (declared as `lazy val` in Scala) is computed only when it is needed in some other expression. Once computed, a lazy value stays in memory and will not be re-computed.

An “on-call” value is re-computed every time it is used. In Scala, a `def` declaration does that.

Most collection types in Scala (such as `List`, `Array`, `Set`, and `Map`) are **eager**: all elements of an eager collection are already evaluated.

A stream is a **lazy collection**. Elements of a stream are computed when first needed; after that, they remain in memory and will not be computed again:

```
scala> val str = Stream.iterate(1)(_ + 1)
str: Stream[Int] = Stream(1, ?)

scala> str.take(10).toList
res0: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

scala> str
res1: Stream[Int] = Stream(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ?)
```

In many cases, it is not necessary to keep previous values of a sequence in memory. For example:

```
scala> (1L to 1000000000L).sum           // Compute the sum of integers from 1 to 1 billion.
res0: Long = 500000000500000000
```

We do not actually need to put a billion numbers in memory if we only want to compute their sum. Indeed, the computation just shown does *not* put all the numbers in memory. The computation will fail if we use a list or a stream:

```
scala> (1L to 1000000000L).toList.sum
java.lang.OutOfMemoryError: GC overhead limit exceeded
```

The code `(1L to 1000000000L).sum` works because `(1 to n)` produces a sequence whose elements are computed whenever needed but do not remain in memory. This can be seen as a sequence with the “on-call” availability of elements. Sequences of this sort are called **iterators**:

```
scala> 1 to 5
res0: scala.collection.immutable.Range.Inclusive = Range(1, 2, 3, 4, 5)

scala> 1 until 5
res1: scala.collection.immutable.Range = Range(1, 2, 3, 4)
```

The types `Range` and `Range.Inclusive` are defined in the Scala standard library and are iterators. They behave as collections and support the usual methods (`map`, `filter`, etc.), but they do not store previously computed values in memory.

The `.view` method Eager collections such as `List` or `Array` can be converted to iterators by using the `.view` method. This is necessary when intermediate collections consume too much memory when fully evaluated. For example, consider the computation of Example 2.1.5.7 where we used `flatMap` to replace each element of an initial sequence by three new numbers before computing `max` of the resulting collection. If instead of three new numbers we wanted to compute *three million* new numbers each time, the intermediate collection created by `flatMap` would require too much memory, and the computation would crash:

```
scala> (1 to 10).flatMap(x => 1 to 3000000).max
java.lang.OutOfMemoryError: GC overhead limit exceeded
```

Even though the range `(1 to 10)` is an iterator, a subsequent `flatMap` operation creates an intermediate collection that is too large

for our computer’s memory. We can use `.view` to avoid this:

```
scala> (1 to 10).view.flatMap(x => 1 to 3000000).max
res0: Int = 3000000
```

The choice between using streams and using iterators is dictated by memory constraints. Except for that, streams and iterators behave similarly to other sequences. We may write programs in the map/reduce style, applying standard methods such as `map`, `filter`, etc., to streams and iterators. Mathematical reasoning about transforming a sequence is the same, whether the sequence is eager, lazy, or on-call.

The Iterator class The Scala library class `Iterator` has methods such as `Iterator.iterate` and others, similarly to `Stream`. However, `Iterator` does not behave as a *value* in the mathematical sense:

```
scala> val iter = (1 until 10).toIterator
iter: Iterator[Int] = non-empty iterator

scala> iter.toList // Look at the elements of 'iter'.
res0: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9)

scala> iter.toList // Look at those elements again...??
res1: List[Int] = List()

scala> iter
res2: Iterator[Int] = empty iterator
```

Evaluating the expression `iter.toList` two times produces a different result the second time. As we see from the Scala output, the value `iter` has become “empty” after the first use.

This situation is impossible in mathematics: if x is some value, such as 100, and f is some function, such as $f(x) = \sqrt{x}$, then $f(x)$ will be the same, $f(100) = \sqrt{100} = 10$, no matter how many times we compute $f(x)$. For instance, we can compute $f(x) + f(x) = 20$ and obtain the correct result. We could also set $y = f(x)$ and compute $y + y = 20$, with the same result. This property is called **referential transparency** or **functional purity** of the function f . After applying a pure function, we can be sure that, for instance, no hidden values in memory have been modified.

When we set $x = 100$ and compute $f(x) + f(x)$, the number 100 does not “become empty” after the first use; its value remains the same. This behavior is called the **value semantics** of numbers. One says that integers “are values” in the mathematical sense. Alternatively, one says that numbers are **immutable**, i.e., cannot be changed. (What would it mean to “modify” the number 10?)

In programming, a type has value semantics if a given computation applied to it always gives the same result. Usually, this means that the type contains immutable data, and the computation is referentially transparent. We can see that Scala’s `Range` has value semantics and is immutable:

```
scala> val x = 1 until 10
x: scala.collection.immutable.Range = Range(1, 2, 3, 4, 5, 6, 7, 8, 9)

scala> x.toList
res0: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9)

scala> x.toList
res1: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9)
```

Collections such as `List`, `Map`, or `Stream` are immutable. Some elements of a `Stream` may not be evaluated yet, but this does not affect its value semantics:

```
scala> val str = (1 until 10).toStream
str: scala.collection.immutable.Stream[Int] = Stream(1, ?)

scala> str.toList
res0: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9)

scala> str.toList
res1: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9)
```

Iterators produced by applying the `view` method to collections will have value semantics:

```
scala> val v = (1 until 10).view
v: scala.collection.SeqView[Int,IndexedSeq[Int]] = SeqView(...)

scala> v.toList
res0: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9)

scala> v.toList
res1: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9)
```

Due to the lack of value semantics, programs written using `Iterator` may not obey the usual rules of mathematical reasoning. This makes it easy to write wrong code that looks correct.

To illustrate the problem, let us re-implement Example 2.5.1.9 by keeping the same code but using `Iterator` instead of `Stream`:

```
def stopRepeatsBad[T](iter: Iterator[T]): Iterator[T] = {
  val halfSpeed = iter.flatMap(x => Seq(x, x))
  halfSpeed.zip(iter) // Do not prepend the first element. It won't help.
  .drop(1).takeWhile { case (h, s) => h != s }
  .map(_._2)
}

scala> stopRepeatsBad(Seq(1, 3, 5, 7, 9, 3, 5, 7, 9).toIterator).toList
res0: List[Int] = List(5, 9, 3, 7, 9)
```

The result `[5, 9, 3, 7, 9]` is incorrect, but not in an obvious way: the sequence *was* stopped at a repetition, as we wanted, but some of the elements of the given sequence are missing (while other elements are present). It is difficult to debug a program that produces *partially* correct numbers.

The error in this code occurs in the expression `halfSpeed.zip(iter)` due to the fact that `halfSpeed` was itself defined via `iter`. The result is that `iter` is used twice in this code, which leads to errors because `iter` is mutable and does not behave as a value. Creating an `Iterator` and using it twice in the same expression can give wrong results or even fail with an exception:

```
scala> val s = (1 until 10).toIterator
s: Iterator[Int] = non-empty iterator

scala> val t = s.zip(s).toList
java.util.NoSuchElementException: next on empty iterator
```

It is surprising and counter-intuitive that a variable cannot be used twice in some expression. Intuitively, we expect code such as `s.zip(s)` to work correctly even though the variable `s` is used twice. When we read the expression `s.zip(s)`, we imagine a given sequence `s` being “zipped” with itself. So we reason that `s.zip(s)` should produce a sequence of pairs. But Scala’s `Iterator` is **mutable** (can be modified during use), which breaks the usual ways of mathematical reasoning about code.

The self-modifying behavior of `Iterator` is an example of a side effect. A function has a **side effect** if the function’s code performs some action in addition to computing the result value. Examples of side effects are: starting and stopping external processes; modifying values stored in memory; writing files; printing; sending or receiving data over a network; and playing sounds. Functions with side effects do not have value semantics. Calling such a function twice produces the side effect twice, which is not the same as calling the function once and simply re-using the result value. On the other hand, pure functions have no side effects and have value semantics.

An `Iterator` can be converted to a `Stream` using the `toStream` method. This restores the value semantics, since streams are values:

```
scala> val iter = (1 until 10).toIterator
iter: Iterator[Int] = non-empty iterator

scala> val str = iter.toStream
str: Stream[Int] = Stream(1, ?)

scala> str.toList
res0: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9)

scala> str.toList
res1: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9)

scala> str.zip(str).toList
res2: List[(Int, Int)] = List((1,1), (2,2), (3,3), (4,4), (5,5), (6,6), (7,7), (8,8), (9,9))
```

Instead of `Iterator`, we can use `Stream` and `view` when lazy or on-call collections are required.

Libraries such as `scalaz` and `fs2` also provide lazy and on-call streams with correct value semantics.

3 The logic of types. I. Disjunctive types

Disjunctive types describe values that belong to a disjoint set of alternatives.

To see how Scala implements disjunctive types, we need to begin by looking at “case classes”.

3.1 Scala’s case classes

3.1.1 Tuple types with names

It is often helpful to use names for the different parts of a tuple. Suppose that some program represents the size and the color of socks with the tuple type `(Double, String)`. What if the same tuple type `(Double, String)` is used in another place in the program to mean the amount paid and the payee? A programmer could mix the two values by mistake, and it would be hard to find out why the program incorrectly computes, say, the total amount paid.

```
def totalAmountPaid(ps: Seq[(Double, String)]): Double = ps.map(_.1).sum
val x = (10.5, "white")           // Sock size and color.
val y = (25.0, "restaurant")    // Payment amount and payee.

scala> totalAmountPaid(Seq(x, y)) // Nonsense.
res0: Double = 35.5
```

We would prevent this kind of mistake if we could use two *different* types, with names such as `MySock` and `Payment`, for the two kinds of data. There are three basic ways of defining a new named type in Scala: using a type alias, using a class (or “trait”), and using an opaque type.

Opaque types (hiding a type under a new name) is a feature of a future version of Scala 3; so we focus on type aliases and case classes.

A **type alias** is an alternative name for an existing (already defined) type. We could use type aliases in our example to add clarity to the code:

```
type MySockTuple = (Double, String)
type PaymentTuple = (Double, String)

scala> val s: MySockTuple = (10.5, "white")
s: MySockTuple = (10.5,white)

scala> val p: PaymentTuple = (25.0, "restaurant")
p: PaymentTuple = (25.0,restaurant)
```

But type aliases do not prevent mix-up errors:

```
scala> totalAmountPaid(Seq(s, p)) // Nonsense again.
res1: Double = 35.5
```

Scala’s **case classes** can be seen as “tuples with names”. A case class is equivalent to a tuple type that has a name chosen when we define the case class. Also, each part of the case class will have a separate name that we must choose. This is how to define case classes for the example with socks and payments:

```
case class MySock(size: Double, color: String)
case class Payment(amount: Double, name: String)

scala> val sock = MySock(10.5, "white")
sock: MySock = MySock(10.5,white)
```

3 The logic of types. I. Disjunctive types

```
scala> val paid = Payment(25.0, "restaurant")
paid: Payment = Payment(25.0,restaurant)
```

This code defines new types named `MySock` and `Payment`. Values of type `MySock` are written as `MySock(10.5, "white")`, which is similar to writing the tuple `(10.5, "white")` except for adding the name `MySock` in front of the tuple.

To access the parts of a case class, we use the part names:

```
scala> sock.size
res2: Double = 10.5

scala> paid.amount
res3: Double = 25.0
```

The mix-up error is now a type error detected by the compiler:

```
def totalAmountPaid(ps: Seq[Payment]): Double = ps.map(_.amount).sum

scala> totalAmountPaid(Seq(paid, paid))
res4: Double = 50.0

scala> totalAmountPaid(Seq(sock, paid))
<console>:19: error: type mismatch;
  found   : MySock
  required: Payment
        totalAmountPaid(Seq(sock, paid))
```

A function whose argument is of type `MySock` cannot be applied to an argument of type `Payment`. Case classes with different names are *different types*, even if they contain the same parts.

Just as tuples can have any number of parts, case classes can have any number of parts, but the part names must be distinct, for example:

```
case class Person(firstName: String, lastName: String, age: Int)

scala> val noether = Person("Emmy", "Noether", 137)
einstein: Person = Person(Emmy,Noether,137)

scala> noether.firstName
res5: String = Emmy

scala> noether.age
res6: Int = 137
```

This data type carries the same information as a tuple `(String, String, Int)`. However, the declaration of a `case class` `Person` gives the programmer several features that make working with the tuple's data more convenient and less error-prone.

Some (or all) part names may be specified when creating a case class value:

```
scala> val poincaré = Person(firstName = "Henri", lastName = "Poincaré", 165)
poincaré: Person = Person(Henri,Poincaré,165)
```

It is a type error to use wrong types with a case class:

```
scala> val p = Person(140, "Einstein", "Albert")
<console>:13: error: type mismatch;
  found   : Int(140)
  required: String
        val p = Person(140, "Einstein", "Albert")

<console>:13: error: type mismatch;
  found   : String("Albert")
  required: Int
        val p = Person(140, "Einstein", "Albert")
```

This error is due to an incorrect order of parts when creating a case class value. However, parts can be specified in any order when using part names:

```
scala> val p = Person(age = 137, lastName = "Noether", firstName = "Emmy")
p: Person = Person(Emmy,Noether,137)
```

A part of a case class can have the type of another case class, creating a type similar to a nested tuple:

```
case class BagOfSocks(sock: MySock, count: Int)
val bag = BagOfSocks(MySock(10.5, "white"), 6)

scala> bag.sock.size
res7: Double = 10.5
```

3.1.2 Case classes with type parameters

Type classes can be defined with type parameters. As an example, consider a generalization of `MySock` where, in addition to the size and color, an “extended sock” holds another value. We could define several specialized case classes,

```
case class MySock_Int(size: Double, color: String, value: Int)
case class MySock_Boolean(size: Double, color: String, value: Boolean)
```

but it is better to define a single parameterized case class

```
case class MySockX[A](size: Double, color: String, value: A)
```

This case class can accommodate every type `A`. We may now create values of `MySockX` containing a value of any given type,

```
scala> val s = MySockX(10.5, "white", 123)
s: MySockX[Int] = MySockX(10.5,white,123)
```

Because 123 has type `Int`, the type parameter `A` in `MySockX[A]` was automatically set to the type `Int`.

Each time we create a value of type `MySockX`, a specific type will have to be used instead of the type parameter `A`. In other words, we can only create values of types `MySockX[Int]`, `MySockX[String]`, etc. If we want to be explicit, we may write

```
scala> val s = MySockX[String](10.5, "white", "last pair")
s: MySockX[String] = MySockX(10.5,white,last pair)
```

However, we can write **parametric code** working with `MySockX[A]`, that is, keeping the type parameter `A` in the code. For example, a function that checks whether a sock of type `MySockX[A]` fits the author's foot can be written as

```
def fits[A](sock: MySockX[A]): Boolean = (sock.size >= 10.5 && sock.size <= 11.0)
```

This function is defined for all types `A` at once, because its code works in the same way regardless of what `A` is. Scala will set the type parameter automatically when we apply `fits` to an argument:

```
scala> fits(MySockX(10.5, "blue", List(1,2,3))) // Type parameter A = List[Int].
res0: Boolean = true
```

This code forces the type parameter `A` to be `List[Int]`, and so we may omit the type parameter of `fits`. When types become more complicated, it may be helpful to write out some type parameters. The compiler can detect a mismatch between the type parameter `A = List[Int]` used in the “sock” value and the type parameter `A = Int` in the function `fits`:

```
scala> fits[Int](MySockX(10.5, "blue", List(1,2,3)))
<console>:15: error: type mismatch;
  found   : List[Int]
  required: Int
          fits[Int](MySockX(10.5, "blue", List(1,2,3)))
                           ^
```

Case classes may have several type parameters, and the types of the parts may use these type parameters. Here is an artificial example of a case class using type parameters in different ways,

```
case class Complicated[A,B,C,D](x: (A, A), y: (B, Int) => A, z: C => C)
```

This case class contains parts of different types that use the type parameters `A`, `B`, `C` in tuples and functions. The type parameter `D` is not used at all; this is allowed (and occasionally useful).

A type with type parameters, such as `MySockX` or `Complicated`, is called a **type constructor**. A type constructor “constructs” a new type, such as `MySockX[Int]`, from a given type parameter `Int`. Values of type `MySockX` cannot be created without setting the type parameter. So, it is important to distinguish the type constructor, such as `MySockX`, from a type that can have values, such as `MySockX[Int]`.

3.1.3 Tuples with one part and with zero parts

Let us compare tuples and case classes more systematically.

Parts of a case class are accessed by name with a dot syntax, for example `sock.color`. Parts of a tuple are accessed with the accessors such as `x._1`. This syntax is the same as that for a case class whose parts have names `_1`, `_2`, etc. So, it appears that tuple parts *do* have names in Scala, although those names are always automatically chosen as `_1`, `_2`, etc. Tuple types are also automatically named in Scala as `Tuple2`, `Tuple3`, etc., and they are parameterized, since each part of the tuple may be of any chosen type. A tuple type expression such as `(Int, String)` is just a special syntax for the parameterized type `Tuple2[Int, String]`. One could define the tuple types as case classes like this,

```
case class Tuple2[A, B](_1: A, _2: B)
case class Tuple3[A, B, C](_1: A, _2: B, _3: C) // And so on with Tuple4, Tuple5...
```

if these types were not already defined in the Scala library.

Proceeding systematically, we ask whether tuple types can have just one part or even no parts. Indeed, Scala defines `Tuple1[A]` (which is rarely used in practice) as a tuple with a single part.

The tuple with zero parts also exists and is called `Unit` (rather than “`Tuple0`”). The syntax for the value of the `Unit` type is the empty tuple, `()`. It is clear that there is *only one* value, `()`, of this type; this explains the name “unit”.

At first sight, the `Unit` type — a tuple that contains no data — may appear to be useless. It turns out, however, that the `Unit` type is important in functional programming. It is used as a type *guaranteed* to have only a single distinct value. This chapter will show some examples of using `Unit`.

Case classes may have one part or zero parts, similarly to the one-part and zero-part tuples:

```
case class B(z: Int)      // Tuple with one part.
case class C()           // Tuple with no parts.
```

The following table summarizes the correspondence between tuples and case classes:

Tuples	Case classes
<code>(123, "xyz")</code> : <code>Tuple2[Int, String]</code>	<code>case class A(x: Int, y: String)</code>
<code>(123,)</code> : <code>Tuple1[Int]</code>	<code>case class B(z: Int)</code>
<code>()</code> : <code>Unit</code>	<code>case class C()</code>

Scala has an alternative syntax for empty case classes:

```
case object C // Similar to 'case class C()'.
```

There are two main differences between `case class C()` and `case object C`:

- A `case object` cannot have type parameters, while we may define a `case class C[x, y, z]()` with type parameters `x`, `y`, `z` if needed.
- A `case object` is allocated in memory only once, while new values of a `case class C()` will be allocated in memory each time `C()` is evaluated.

Other than that, `case class C()` and `case object c` have the same meaning: a named tuple with zero parts, which we may also view as a “named `Unit`” type. This book will not use `case objects` because `case classes` are sufficient.

3.1.4 Pattern matching for case classes

Scala performs pattern matching in two situations:

- destructuring definition: `val pattern = ...`
- `case` expression: `case pattern => ...`

Case classes can be used in both situations. A destructuring definition can be used in a function whose argument is of case class type `BagOfSocks`:

```
case class MySock(size: Double, color: String)
case class BagOfSocks(sock: MySock, count: Int)

def printBag(bag: BagOfSocks): String = {
  val BagOfSocks(MySock(size, color), count) = bag // Destructure the 'bag'.
  s"bag has $count $color socks of size $size"
}

val bag = BagOfSocks(MySock(10.5, "white"), 6)

scala> printBag(bag)
res0: String = bag has 6 white socks of size 10.5
```

A `case` expression will destructure a value and compute a result:

```
def fits(bag: BagOfSocks): Boolean = bag match {
  case BagOfSocks(MySock(size, _), _) => (size >= 10.5 && size <= 11.0)
}
```

In the code of this function, we match the `bag` value against the pattern `BagOfSocks(MySock(size, _), _)`. This pattern will always match and will define `size` as a pattern variable of type `Double`.

The syntax for pattern matching expressions with case classes is similar to the syntax for pattern matching of tuples, except for the presence of the *names* of the case classes. For example, removing the case class names from the pattern

```
case BagOfSocks(MySock(size, _), _) => ...
```

we obtain the nested tuple pattern

```
case ((size, _), _) => ...
```

that could be used for values of type `((Double, String), Int)`. We see that within pattern matching expressions, case classes behave exactly as tuple types with added names.

Scala’s “case classes” got their name from their use in `case` expressions. It is usually more convenient to use `match/case` expressions with case classes than to use destructuring.

3.2 Disjunctive types

3.2.1 Motivation and first examples

In many situations, it is useful to have several different shapes of data within the same type. As a first example, suppose we are looking for real roots of a quadratic equation $x^2 + bx + c = 0$. There are three cases: no real roots, one real root, and two real roots. It is convenient to have a type that represents “the real roots of a quadratic equation”; call it `RootsOfQ`. Inside that type, we distinguish between the three cases, but outside it looks like a single type.

Another example is the binary search algorithm that looks for an integer x in a sorted array. Either the algorithm finds the index of x , or it determines that the array does not contain x . It is convenient if the algorithm could return a single value of a type (call it `SearchResult`) that represents *either* an index at which x is found, *or* the absence of an index.

More generally, we may have computations that *either* return a value *or* generate an error and fail to produce a result. It is then convenient to return a value of type (call it `Result`) that represents either a correct result or an error message.

In certain computer games, one has different types of “rooms”, each room having certain properties depending on its type. Some rooms are dangerous because of monsters, other rooms contain useful objects, certain rooms allow you to finish the game, and so on. We want to represent all the different kinds of rooms uniformly, as a type `Room`, so that a value of type `Room` automatically stores the correct properties in each case.

In all these situations, data comes in several mutually exclusive shapes. This data can be represented by a single type if that type is able to describe a mutually exclusive set of cases:

- `RootsOfQ` must be either the empty tuple `()`, or `Double`, or a tuple `(Double, Double)`
- `SearchResult` must be either `Int` or the empty tuple `()`
- `Result` must be either an `Int` value or a `String` message

We see that the empty tuple, also known as the `Unit` type, is natural to use in these situations. It is also helpful to assign names to each of the cases:

- `RootsOfQ` is “no roots” with value `()`, or “one root” with value `Double`, or “two roots” with value `(Double, Double)`
- `SearchResult` is “index” with value `Int`, or “not found” with value `()`
- `Result` is “value” of type `Int` or “error message” of type `String`

Scala’s case classes provide exactly what we need here — named tuples with zero, one, two or more parts. So it is natural to use case classes instead of tuples:

- `RootsOfQ` is a value of type `case class NoRoots()`, or a value of type `case class OneRoot(x: Double)`, or a value of type `case class TwoRoots(x: Double, y: Double)`
- `SearchResult` is a value of type `case class Index(Int)` or a value of type `case class NotFound()`
- `Result` is a value of type `case class Value(x: Int)` or a value of type `case class Error(message: String)`

Our three examples are now described as types that select one case class out of a given set. It remains to see how Scala defines such types. For instance, the definition of `RootsOfQ` needs to indicate that the case classes `NoRoots`, `OneRoot`, and `TwoRoots` are exactly the three alternatives described by the type `RootsOfQ`. The Scala syntax for that definition looks like this:

```
sealed trait RootsOfQ
final case class NoRoots() extends RootsOfQ
final case class OneRoot(x: Double) extends RootsOfQ
final case class TwoRoots(x: Double, y: Double) extends RootsOfQ
```

In the definition of `SearchResult`, we have two cases:

```
sealed trait SearchResult
final case class Index(i: Int) extends SearchResult
final case class NotFound() extends SearchResult
```

The definition of the `Result` type is parameterized, so that we can describe results of any type (while error messages are always of type `String`):

```
sealed trait Result[A]
final case class Value[A](x: A) extends Result[A]
final case class Error[A](message: String) extends Result[A]
```

The “`sealed trait / final case class`” syntax defines a type that represents a choice of one case class from a fixed set of case classes. This kind of type is called a **disjunctive type** in this book.

3.2.2 Solved examples: Pattern matching for disjunctive types

Our first examples of disjunctive types are `RootsOfQ`, `SearchResult`, and `Result[A]` defined in the previous section. We will now look at the Scala syntax for creating values of disjunctive types and for using the created values.

Consider the disjunctive type `RootsOfQ` having three case classes (`NoRoots`, `OneRoot`, `TwoRoots`). The only way of creating a value of type `RootsOfQ` is to create a value of one of these case classes. This is done by writing expressions such as `NoRoots()`, `OneRoot(2.0)`, or `TwoRoots(1.0, -1.0)`. Scala will accept these expressions as having the type `RootsOfQ`:

```
scala> val x: RootsOfQ = OneRoot(2.0)
x: RootsOfQ = OneRoot(2.0)
```

Given a value `x: RootsOfQ`, how can we use it, say, as a function argument? The main tool for working with values of disjunctive types is pattern matching with `match/case` expressions. In Chapter 2, we used pattern matching to destructure tuples with syntax such as `{ case (x, y) => ... }`. To use `match/case` expressions with disjunctive types, we may have to write *more than one case* pattern in a `match` expression, because we need to match several possible cases of the disjunctive type:

```
def f(r: RootsOfQ): String = r match {
  case NoRoots()      => "no real roots"
  case OneRoot(r)     => s"one real root: $r"
  case TwoRoots(x, y) => s"real roots: ($x, $y)"
}

scala> f(x)
res0: String = "one real root: 2.0"
```

If we only need to recognize certain cases of a disjunctive type, we can match all other cases with an underscore:

```
scala> x match {
  case OneRoot(r)  => s"one real root: $r"
  case _           => "have something else"
}
res1: String = one real root: 2.0
```

The `match/case` expression represents a choice over possible values of a given type. Note the similarity with this code:

```
def f(x: Int): Int = x match {
  case 0    => println(s"error: must be nonzero"); -1
  case 1    => println(s"error: must be greater than 1"); -1
  case _    => x
}
```

The values `0` and `1` are some possible values of type `Int`, just as `OneRoot(1.0)` is a possible value of type `RootsOfQ`. When used with disjunctive types, `match/case` expressions will usually contain a complete list of possibilities. If the list of cases is incomplete, the Scala compiler will print a warning:

```
scala> def g(x: RootsOfQ): String = x match {
  case OneRoot(r) => s"one real root: $r"
}
<console>:14: warning: match may not be exhaustive.
It would fail on the following inputs: NoRoots(), TwoRoots(_, _)
```

3 The logic of types. I. Disjunctive types

This code defines a *partial* function `g` that can be applied only to values of the form `OneRoot(...)` and will fail for other values.

Let us look at more examples of using the disjunctive types we just defined.

Example 3.2.2.1 Given a sequence of quadratic equations, compute a sequence containing their real roots as values of type `RootsOfQ`.

Solution Define a case class representing a quadratic equation $x^2 + bx + c = 0$:

```
case class QEqu(b: Double, c: Double)
```

The following function determines how many real roots an equation has:

```
def solve(quadraticEqu: QEqu): RootsOfQ = {
  val QEqu(b, c) = quadraticEqu // Destructure QEqu.
  val d = b * b / 4 - c
  if (d > 0) {
    val s = math.sqrt(d)
    TwoRoots(b / 2 - s, b / 2 + s)
  } else if (d == 0.0) OneRoot(b / 2)
  else NoRoots()
}
```

Test the `solve` function:

```
scala> solve(QEqu(1,1))
res1: RootsOfQ = NoRoots()

scala> solve(QEqu(1,-1))
res2: RootsOfQ = TwoRoots(-0.6180339887498949,1.618033988749895)

scala> solve(QEqu(6,9))
res3: RootsOfQ = OneRoot(3.0)
```

We can now implement the function `findRoots`,

```
def findRoots(equs: Seq[QEqu]): Seq[RootsOfQ] = equs.map(solve)
```

If the function `solve` will not be used often, we may want to write it inline as a nameless function:

```
def findRoots(equs: Seq[QEqu]): Seq[RootsOfQ] = equs.map { case QEqu(b, c) =>
  (b * b / 4 - c) match {
    case d if d > 0    =>
      val s = math.sqrt(d)
      TwoRoots(b / 2 - s, b / 2 + s)
    case 0.0            => OneRoot(b / 2)
    case _              => NoRoots()
  }
}
```

This code depends on some features of Scala syntax. We can use the partial function `{ case QEqu(b, c) => ... }` directly as the argument of `map` instead of defining this function separately. This avoids having to destructure `QEqu` at a separate step. The `if/else` expression is replaced by an “embedded” `if` within the `case` expression, which is easier to read. Test the final code:

```
scala> findRoots(Seq(QEqu(1,1), QEqu(2,1)))
res4: Seq[RootsOfQ] = List(NoRoots(), OneRoot(1.0))
```

Example 3.2.2.2 Given a sequence of values of type `RootsOfQ`, compute a sequence containing only the single roots. Example test:

```
def singleRoots(rs: Seq[RootsOfQ]): Seq[Double] = ???

scala> singleRoots(Seq(TwoRoots(-1, 1), OneRoot(3.0), OneRoot(1.0), NoRoots()))
res5: Seq[Double] = List(3.0, 1.0)
```

Solution We apply `filter` and `map` to the sequence of roots:

```
def singleRoots(rs: Seq[RootsOfQ]): Seq[Double] = rs.filter {
  case OneRoot(x) => true
  case _              => false
}.map { case OneRoot(x) => x }
```

In the `map` operation, we need to cover only the one-root case because the two other possibilities have been excluded (“filtered out”) by the preceding `filter` operation.

Example 3.2.2.3 Implement binary search returning a `SearchResult`. Modify the binary search implementation from Example 2.5.1.5(b) so that it returns a `NotFound` value when appropriate.

Solution The code from Example 2.5.1.5(b) will return *some* index even if the given number is not present in the array:

```
scala> binSearch(Array(1, 3, 5, 7), goal = 5)
res6: Int = 2

scala> binSearch(Array(1, 3, 5, 7), goal = 4)
res7: Int = 1
```

In that case, the array’s element at the computed index will not be equal to `goal`. We should return `NotFound()` in that case. The new code can be written as a `match/case` expression for clarity:

```
def safeBinSearch(xs: Seq[Int], goal: Int): SearchResult =
  binSearch(xs, goal) match {
    case n if xs(n) == goal    => Index(n)
    case _                      => NotFound()
  }
```

To test:

```
scala> safeBinSearch(Array(1, 3, 5, 7), 5)
res8: SearchResult = Index(2)

scala> safeBinSearch(Array(1, 3, 5, 7), 4)
res9: SearchResult = NotFound()
```

Example 3.2.2.4 Use the disjunctive type `Result[Int]` to implement “safe integer arithmetic”, where a division by zero or a square root of a negative number will give an error message. Define arithmetic operations directly for values of type `Result[Int]`. When errors occur, abandon further computations.

Solution Begin by implementing the square root:

```
def sqrt(r: Result[Int]): Result[Int] = r match {
  case Value(x) if x >= 0  => Value(math.sqrt(x).toInt)
  case Value(x)            => Error(s"error: sqrt($x)")
  case Error(m)            => Error(m) // Keep the error message.
}
```

The square root is computed only if we have the `Value(x)` case, and only if $x \geq 0$. If the argument `r` was already an `Error` case, we keep the error message and perform no further computations.

To implement the addition operation, we need a bit more work:

```
def add(rx: Result[Int], ry: Result[Int]): Result[Int] = (rx, ry) match {
  case (Value(x), Value(y)) => Value(x + y)
  case (Error(m), _)         => Error(m) // Keep the first error message.
  case (_, Error(m))        => Error(m) // Keep the second error message.
}
```

This code illustrates nested patterns that match the tuple `(rx, ry)` against various possibilities. In this way, the code is clearer than code written with nested `if/else` expressions.

Implementing the multiplication operation results in almost the same code:

```
def mul(rx: Result[Int], ry: Result[Int]): Result[Int] = (rx, ry) match {
  case (Value(x), Value(y)) => Value(x * y)
  case (Error(m), _)         => Error(m)
```

```
  case (_, Error(m))      => Error(m)
}
```

To avoid repetition, we may define a general function that “lifts” operations on integers to operations on `Result[Int]` types:

```
def do2(rx: Result[Int], ry: Result[Int])(op: (Int, Int) => Int): Result[Int] =
  (rx, ry) match {
    case (Value(x), Value(y)) => Value(op(x, y))
    case (Error(m), _)        => Error(m)
    case (_, Error(m))       => Error(m)
}
```

Now we can easily “lift” any binary operation on integers to a binary operation on `Result[Int]`, assuming that the operation never generates an error:

```
def sub(rx: Result[Int], ry: Result[Int]): Result[Int] = do2(rx, ry){ (x, y) => x - y }
```

Custom code is still needed for operations that *may* generate errors:

```
def div(rx: Result[Int], ry: Result[Int]): Result[Int] = (rx, ry) match {
  case (Value(x), Value(y)) if y != 0 => Value(x / y)
  case (Value(x), Value(y))          => Error(s"error: $x / $y")
  case (Error(m), _)               => Error(m)
  case (_, Error(m))              => Error(m)
}
```

We can now test the “safe arithmetic” on simple calculations:

```
scala> add(Value(1), Value(2))
res10: Result[Int] = Value(3)

scala> div(add(Value(1), Value(2)), Value(0))
res11: Result[Int] = Error(error: 3 / 0)
```

We see that indeed all further computations are abandoned once an error occurs. An error message shows only the immediate calculation that generated the error. For instance, the error message for $20 + 1/0$ never mentions 20:

```
scala> add(Value(20), div(Value(1), Value(0)))
res12: Result[Int] = Error(error: 1 / 0)

scala> add(sqrt(Value(-1)), Value(10))
res13: Result[Int] = Error(error: sqrt(-1))
```

3.2.3 Standard disjunctive types: Option, Either, Try

The Scala library defines the disjunctive types `Option`, `Either`, and `Try` because they are used often. We now look at each of them in turn.

The Option type is a disjunctive type with two cases: the empty tuple and a one-element tuple. The names of the two case classes are `None` and `Some`. If the `Option` type were not already defined in the standard library, one could define it with the code

```
sealed trait Option[T]
final case object None      extends Option[Nothing]
final case class Some[T](t: T) extends Option[T]
```

This code is similar to the type `SearchResult` defined in Section 3.2.1, except that `Option` has a type parameter instead of a fixed type `Int`. Another difference is the use of a `case object` for the empty case instead of an empty case class, such as `None()`. Since Scala’s `case objects` cannot have type parameters, the type parameter in the definition of `None` must be set to the special type `Nothing`, which is a type with *no* values (also called the **void type**).

An alternative (implemented in libraries such as `scalaz`) is to define the empty option value as

```
final case class None[T]() extends Option[T]
```

In that implementation, the empty option `None[T]()` has a type parameter.

Several consequences follow from the Scala library's decision to define `None` without a type parameter. One consequence is that `None` can be reused as a value of type `Option[A]` for any type `A`:

```
scala> val y: Option[Int] = None
y: Option[Int] = None

scala> val z: Option[String] = None
z: Option[String] = None
```

Typically, `Option` is used in situations where a value may be either present or missing, especially when a missing value is *not an error*. The missing-value case is represented by `None`, while `Some(x)` means that a value `x` is present.

Example 3.2.3.1 Information about “subscribers” must include a name and an email address, but a telephone number is optional. To represent this information, we define a case class like this,

```
case class Subscriber(name: String, email: String, phone: Option[Long])
```

What if we represent the missing telephone number by a special value such as `-1` and use the simpler type `Long` instead of `Option[Long]`? The disadvantage is that we would need to *remember* to check for the special value `-1` in all functions that take the telephone number as an argument. Looking at a function such as `sendSMS(phone: Long)` at a different place in the code, a programmer might forget that the telephone number is actually optional. In contrast, the type signature `sendSMS(phone: Option[Long])` unambiguously indicates that the telephone number might be missing and helps the programmer to remember to handle both cases.

Pattern-matching code involving `Option` can handle the two cases like this:

```
def getDigits(phone: Option[Long]): Option[Seq[Long]] = phone match {
  case None          => None          // Cannot obtain digits, return None.
  case Some(number)  => Some(digitsOf(number))
} // The function 'digitsOf' was defined in Section 2.3
```

At the two sides of `case None => None`, the value `None` has different types, namely `Option[Long]` and `Option[Seq[Long]]`. Since these types are declared in the type signature of the function `getDigits`, the Scala compiler is able to figure out the types of all expressions in the `match/case` construction. So, pattern-matching code can be written without explicit type annotations such as `(None: Option[Long])`.

If we now need to compute the number of digits, we can write

```
def numberofDigits(phone: Option[Long]): Option[Long] = getDigits(phone) match {
  case None          => None
  case Some(digits)  => Some(digits.length)
}
```

These examples perform a computation when an `Option` value is non-empty, and leave it empty otherwise. This design pattern is used often. To avoid repeating the code, we can implement this design pattern as a function that takes the computation as an argument `f`:

```
def doComputation(x: Option[Long], f: Long => Long): Option[Long] = x match {
  case None          => None
  case Some(i)       => Some(f(i))
}
```

It is then natural to generalize this function to arbitrary types using type parameters instead of a fixed type `Long`. The resulting function is usually called `fmap`:

```
def fmap[A, B](f: A => B): Option[A] => Option[B] = {
  case None          => None
  case Some(a)       => Some(f(a))
}
```

3 The logic of types. I. Disjunctive types

```
scala> fmap(digits0f)(Some(4096))
res0: Option[Seq[Long]] = Some(List(4, 0, 9, 6))

scala> fmap(digits0f)(None)
res1: Option[Seq[Long]] = None
```

We say that the `fmap` operation **lifts** a given function of type `A => B` to the type `Option[A] => Option[B]`.

It is important to keep in mind that the code `case Some(a) => Some(f(a))` changes the type of the option value. On the left side of the arrow, the type is `Option[A]`, while on the right side it is `Option[B]`. The Scala compiler knows this from the given type signature of `fmap`, so an explicit type parameter, `Some[B](f(a))`, is not needed.

The Scala library implements an equivalent function as a method on the `Option` class, with the syntax `x.map(f)` rather than `fmap(f)(x)`. We can concisely rewrite the previous code using these methods,

```
def getDigits(phone: Option[Long]): Option[Seq[Long]] = phone.map(digits0f)
def number0fDigits(phone: Option[Long]): Option[Long] = phone.map(digits0f).map(_.length)
```

We see that the `map` operation for the `Option` type is analogous to the `map` operation for sequences.

The similarity between `Option[A]` and `Seq[A]` is clearer if we view `Option[A]` as a special kind of “sequence” whose length is restricted to be either 0 or 1. So, `Option[A]` can have all the operations of `Seq[A]` except operations such as `concat` that may grow the sequence beyond length 1. The standard operations defined on `Option` include `map`, `filter`, `forall`, `exists`, `flatMap`, and `foldLeft`.

Example 3.2.3.2 Given a phone number as `Option[Long]`, extract the country code if it is present. (Assume that the country code is any digits in front of the 10-digit number; for the phone number 18004151212, the country code is 1.) The result must be again of type `Option[Long]`.

Solution If the phone number is a positive integer n , we may compute the country code simply as $n/10000000000L$. However, if the result of that division is zero, we should return an empty `Option` (i.e. the value `None`) rather than 0. To implement this logic, we may begin by writing this code,

```
def countryCode(phone: Option[Long]): Option[Long] = phone match {
  case None      => None
  case Some(n)   =>
    val countryCode = n / 10000000000L
    if (countryCode != 0L) Some(countryCode) else None
}
```

We may notice that we have reimplemented the design pattern similar to `map` in this code, namely “if `None` then return `None`, else do a computation”. So we may try to rewrite the code as

```
def countryCode(phone: Option[Long]): Option[Long] = phone.map { n =>
  val countryCode = n / 10000000000L
  if (countryCode != 0L) Some(countryCode) else None
} // Type error: the result is Option[Option[Long]], not Option[Long].
```

This code does not compile: we are returning an `Option[Long]` within a function lifted via `map`, so the resulting type is `Option[Option[Long]]`. We may use `flatten` to convert `Option[Option[Long]]` to the required type `Option[Long]`,

```
def countryCode(phone: Option[Long]): Option[Long] = phone.map { n =>
  val countryCode = n / 10000000000L
  if (countryCode != 0L) Some(countryCode) else None
}.flatten // Types are correct now.
```

Since the `flatten` follows a `map`, we can rewrite the code using `flatMap`:

```
def countryCode(phone: Option[Long]): Option[Long] = phone.flatMap { n =>
  val countryCode = n / 10000000000L
  if (countryCode != 0L) Some(countryCode) else None
} // Types are correct now.
```

Another way of implementing this example is to notice the design pattern “if condition does not hold, return `None`, otherwise keep the value”. For an `Option` type, this is equivalent to the `filter`

operation (recall that `filter` returns an empty sequence if the predicate never holds). The code is

```
def countryCode(phone: Option[Long]): Option[Long] = phone.map(_ / 10000000000L).filter(_ != 0L)

scala> countryCode(Some(18004151212L))
res0: Option[Long] = Some(1)

scala> countryCode(Some(8004151212L))
res1: Option[Long] = None
```

Example 3.2.3.3 Add a new requirement to Example 3.2.3.2: if the country code is not present, we should return the default country code 1.

Solution This is an often used code pattern: “if empty, substitute a default value”. The Scala library has the method `getOrElse` for this purpose:

```
scala> Some(100).getOrElse(1)
res2: Int = 100

scala> None.getOrElse(1)
res3: Int = 1
```

So we can implement the new requirement as

```
scala> countryCode(Some(8004151212L)).getOrElse(1L)
res4: Long = 1
```

Using Option with collections Many Scala library methods return an `Option` as a result. The main examples are `find`, `headOption`, and `lift` for sequences, and `get` for dictionaries.

The `find` method returns the first element satisfying a predicate:

```
scala> (1 to 10).find(_ > 5)
res0: Option[Int] = Some(6)

scala> (1 to 10).find(_ > 10) // No element is > 10.
res1: Option[Int] = None
```

The `lift` method returns the element of a sequence at a given index:

```
scala> (10 to 100).lift(0)
res2: Option[Int] = Some(10)

scala> (10 to 100).lift(1000) // No element at index 1000.
res3: Option[Int] = None
```

The `headOption` method returns the first element of a sequence, unless the sequence is empty. This is equivalent to `lift(0)`:

```
scala> Seq(1,2,3).headOption
res4: Option[Int] = Some(1)

scala> Seq(1,2,3).filter(_ > 10).headOption
res5: Option[Int] = None
```

Applying `.find(p)` computes the same result as `.filter(p).headOption`, but `.find(p)` may be faster.

The `get` method for a dictionary returns the value if it exists for a given key, and returns `None` if the key is not in the dictionary:

```
scala> Map(10 -> "a", 20 -> "b").get(10)
res6: Option[String] = Some(a)

scala> Map(10 -> "a", 20 -> "b").get(30)
res7: Option[String] = None
```

The `get` method is a safe by-key access to dictionaries, unlike the direct access that may fail:

```
scala> Map(10 -> "a", 20 -> "b")(10)
```

```
res8: String = a

scala> Map(10 -> "a", 20 -> "b")(30)
java.util.NoSuchElementException: key not found: 30
  at scala.collection.MapLike$class.default(MapLike.scala:228)
  at scala.collection.AbstractMap.default(Map.scala:59)
  ... 32 elided
```

Similarly, `lift` is a safe by-index access to collections, unlike the direct access that may fail:

```
scala> Seq(10,20,30)(0)
res9: Int = 10

scala> Seq(10,20,30)(5)
java.lang.IndexOutOfBoundsException: 5
  at scala.collection.LinearSeqOptimized$class.apply(LinearSeqOptimized.scala:65)
  at scala.collection.immutable.List.apply(List.scala:84)
  ... 32 elided
```

The Either type The standard disjunctive type `Either[A, B]` has two type parameters and is often used for computations that report errors. By convention, the *first* type (`A`) is the type of error, and the *second* type (`B`) is the type of the (non-error) result. The names of the two cases are `Left` and `Right`. A possible definition of `Either` may be written as

```
sealed trait Either[A, B]
final case class Left[A, B](value: A) extends Either[A, B]
final case class Right[A, B](value: B) extends Either[A, B]
```

By convention, a value `Left(x)` represents an error, and a value `Right(y)` represents a valid result.

As an example, the following function substitutes a default value and logs the error information:

```
def logError(x: Either[String, Int], default: Int): Int = x match {
  case Left(error)  => println(s"Got error: $error"); default
  case Right(res)   => res
}
```

To test:

```
scala> logError(Right(123), -1)
res1: Int = 123

scala> logError(Left("bad result"), -1)
Got error: bad result
res2: Int = -1
```

Why use `Either` instead of `Option` for computations that may fail? A failing computation such as `"xyz".toInt` cannot return a result, and sometimes we might use `None` to indicate that a result is not available. However, when the result is a requirement for further calculations, we will usually need to know exactly *which* error prevented the result from being available. The `Either` type may provide detailed information about such errors, which `Option` cannot do.

The `Either` type generalizes the type `Result` defined in Section 3.2.1 to an arbitrary error type instead of `String`. We have seen its usage in Example 3.2.2.4, where the design pattern was “if value is present, do a computation, otherwise keep the error”. This design pattern is implemented by the `map` method on `Either`:

```
1  scala> Right(1).map(_ + 1)
2  res0: Either[Nothing, Int] = Right(2)
3
4  scala> Left[String, Int]("error").map(_ + 1)
5  res1: Either[String, Int] = Left("error")
```

The type `Nothing` was filled in by the Scala compiler because we did not specify the first type parameter of `Right` in line 1.

The methods `flatMap`, `fold`, and `getOrElse` are also defined for `Either`, with the same convention that

a `Left` value represents an error.¹

Exceptions and the `Try` type When computations fail for any reason, Scala generates an **exception** instead of returning a value. An exception means that the evaluation of some expression was stopped without returning a result.

As an example, exceptions are generated when the available memory is too small to store the resulting data (as we saw in Section 2.6.3), or if a stack overflow occurs during the computation (as we saw in Section 2.2.3). Exceptions may also occur due to programmer's error: when a pattern matching operation fails, when a requested key does not exist in a dictionary, or when the `head` operation is applied to an empty list.

Motivated by these examples, we may distinguish “planned” and “unplanned” exceptions.

A **planned** exception is generated by programmer's code via the `throw` syntax:

```
scala> throw new Exception("This is a test... this is only a test.")
java.lang.Exception: This is a test... this is only a test.
```

The Scala library contains a `throw` operation in various places, such as in the code for applying the `head` method to an empty sequence, as well as in other situations where exceptions are generated due to programmer's errors. These exceptions are generated deliberately and in well-defined situations. Although these exceptions indicate errors, these errors are anticipated in advance and so may be handled by the programmer.

For example, many Java libraries will generate exceptions when function arguments have unexpected values, when a network operation takes too long or a network connection is unexpectedly broken, when a file is not found or cannot be read due to access permissions, and in many other situations. All these exceptions are “planned” because they are generated explicitly by library code such as `throw new FileNotFoundException(...)`. The programmer's code is expected to catch these exceptions, to handle the error, and to continue the evaluation of the program.

An **unplanned** exception is generated by the Java runtime system when critical errors occur, such as an out-of-memory error. It is rare that a programmer writes `val y = f(x)` while *expecting* that an out-of-memory exception will sometimes occur at that point.² An unplanned exception indicates a serious and unforeseen problem with memory or another critically important resource, such as the operating system's threads or file handles. Such problems usually cannot be fixed and will prevent the program from running any further. It is reasonable that the program should abruptly stop (or “crash” as programmers say) after such an error.

The use of planned exceptions assumes that the programmer will write code to handle each exception. This assumption makes it significantly harder to write programs correctly: it is hard to figure out and to keep in mind all the possible exceptions that a given library function may `throw` in its code (and in the code of all other libraries on which it depends). Instead of using exceptions for indicating errors, Scala programmers can write functions that return a disjunctive type, such as `Either`, describing both a correct result and an error condition. Users of these functions will *have* to do pattern matching on the result values. This helps programmers to remember and to handle all relevant error situations.

However, programmers will often need to use Java or Scala libraries that `throw` exceptions. To help write code for these situations, the Scala library contains a helper function called `Try()` and a disjunctive type also called `Try`. The type `Try[A]` can be seen as similar to `Either[Throwable, A]`, where `Throwable` is the general type of all exceptions (i.e. values to which a `throw` operation can be applied). The two parts of the disjunctive type `Try[A]` are called `Failure` and `Success[A]` (instead of `Left[Throwable]` and `Right[A]` in the `Either` type). The function `Try(expr)` will catch all exceptions thrown while the expression `expr` is evaluated. If the evaluation of `expr` succeeds and returns a value `x:A`, the value of `Try(expr)` will be `Success(x)`. Otherwise it will be `Failure(t)`, where `t:Throwable` is the value associated with the generated exception. Here is an example of using `Try`:

```
import scala.util.{Try, Success, Failure}
```

¹These methods are available in Scala 2.12 or a later version.

²Just once in the author's experience, an out-of-memory condition had to be anticipated in an Android app.

```
scala> val p = Try("xyz".toInt)
p: Try[Int] = Failure(java.lang.NumberFormatException: For input string: "xyz")

scala> val q = Try("0002".toInt)
q: Try[Int] = Success(2)
```

The code `Try("xyz".toInt)` does not generate any exceptions and will not crash the program. Any computation that may throw an exception can be enclosed in a `Try()`, and the exception will be caught and encapsulated within the disjunctive type as a `Failure(...)` value.

The methods `map`, `filter`, `flatMap`, `foldLeft` are defined for the `Try` class similarly to the `Either` type. One additional feature of `Try` is to catch exceptions generated by the function arguments of `map`, `filter`, `flatMap`, and other standard methods:

```
scala> val y = q.map(y => throw new Exception("ouch"))
y: Try[Int] = Failure(java.lang.Exception: ouch)

scala> val z = q.filter(y => throw new Exception("huh"))
z: Try[Int] = Failure(java.lang.Exception: huh)
```

matching on the values `y` and `z` and examine those exceptions. However, it is important that these exceptions were caught and the program did not crash, so the other code is *able* to run.

While the standard types `Try` and `Either` will cover many use cases, programmers can also define custom disjunctive types in order to represent all the anticipated failures or errors in the business logic of a particular application. Representing all errors in the types helps assure that the program will not crash because of an exception that we forgot to handle or did not even know about.

In this example, the values `y` and `z` were computed *successfully* even though exceptions were thrown while the function arguments of `map` and `filter` were evaluated. Other code can use pattern

3.3 Lists and trees: recursive disjunctive types

Consider this code defining a disjunctive type `NInt`:

```
sealed trait NInt
final case class N1(x: Int)      extends NInt
final case class N2(n: NInt)      extends NInt
```

The type `NInt` has two disjunctive parts, `n1` and `n2`. But the definition of the case class `n1` refers to the type `NInt` as if it were already defined.

A type whose definition uses that same type is called a **recursive type**. The type `NInt` is an example of a recursive disjunctive type.

We might imagine defining a disjunctive type `x` whose parts recursively refer to the same type `x` (and/or to each other) in complicated ways. What kind of data would be represented by such a type `x`, and in what situations would `x` be useful? For instance, the simple definition

```
final case class Bad(x: Bad)
```

is useless: we can create a value of type `Bad` only if we already have a value of type `Bad`. This is an example of an infinite loop in type recursion. We will never be able to create any values of type `Bad`, which means that the type `Bad` is void (has no values, like the the special type `Nothing`).

Section 8.5.1 will derive conditions for recursive types to be non-void. For now, we will look at the main examples of recursive disjunctive types that are used most often: lists and trees.

3.3.1 Lists

A list of values of type `A` is either empty, or has one value of type `A`, or two values of type `A`, etc. We can visualize the type `List[A]` as a disjunctive type defined by

```
sealed trait List[A]
final case class List0[A]()           extends List[A]
final case class List1[A](x: A)       extends List[A]
final case class List2[A](x1: A, x2: A) extends List[A]
???                                // Need an infinitely long definition.
```

However, this definition is not practical: we cannot define a separate case class for *each* possible length. Instead, we define the type `List[A]` via mathematical induction on the length of the list:

- Base case: empty list, `case class List0[A]()`.
- Inductive step: given a list of a previously defined length, say `Listn-1`, define a new case class `Listn` describing a list with one more element of type `A`. So we could define `Listn = (Listn-1, A)`.

Let us try to write this inductive definition as code:

```
sealed trait ListI[A]           // Inductive definition of a list.
final case class List0[A]()      extends ListI[A]
final case class List1[A](prev: List0[A], x: A)  extends ListI[A]
final case class List2[A](prev: List1[A], x: A)  extends ListI[A]
???
           // Still need an infinitely long definition.
```

To avoid writing an infinitely long type definition, we need to use a trick. Notice that all definitions of `List1`, `List2`, etc., have a similar form (while `List0` is not similar). We can replace all the definitions `List1`, `List2`, etc., by a single definition if we use the type `ListI[A]` recursively inside the case class:

```
sealed trait ListI[A]           // Inductive definition of a list.
final case class List0[A]()      extends ListI[A]
final case class ListN[A](prev: ListI[A], x: A)  extends ListI[A]
```

The type definition has become recursive. For this trick to work, it is important to use `ListI[A]` and not `ListN[A]` inside the definition `ListN[A]`; or else we would have created an infinite loop in type recursion similar to `case class Bad` shown previously.

Since we obtained the type definition of `ListI` via a trick, let us verify that the code actually defines the disjunctive type we wanted.

To create a value of type `ListI[A]`, we must use one of the two available case classes. Using the first case class, we may create a value `List0()`. Since this empty case class does not contain any values of type `A`, it effectively represents an empty list (the base case of the induction). Using the second case class, we may create a value `ListN(prev, x)` where `x` is of type `A` and `prev` is some previously constructed value of type `ListI[A]`. This represents the inductive step, because the case class `ListN` is a named tuple containing `ListI[A]` and `A`. Now, the same consideration recursively applies to constructing the value `prev`, which must be either an empty list or a pair containing another list and an element of type `A`. The assumption that the value `prev:ListI[A]` is already constructed is equivalent to the inductive assumption that we already have a list of a previously defined length. So, we have verified that `ListI[A]` implements the inductive definition shown above.

Examples of values of type `ListI` are the empty list `List0()`, a one-element list `ListN(List0(), x)`, and a two-element list `ListN(ListN(List0(), x), y)`.

To illustrate writing pattern-matching code using this type, let us implement the method `headOption`:

```
@tailrec def headOption[A]: ListI[A] => Option[A] = {
  case List0()          => None
  case ListN(List0(), x) => Some(x)
  case ListN(prev, _)    => headOption(prev)
}
```

The Scala library already defines the type `List[A]` in an equivalent but different way: its case classes are named differently, and the second case class (with the special name `::`) places the value of type `A` *before* the previously constructed list,

```
sealed trait List[A]
final case object Nil extends List[Nothing]
final case class ::[A](head: A, tail: List[A]) extends List[A]
```

Because “operator-like” case class names, such as `::`, support the infix syntax, we may write expressions such as `head :: tail` instead of `::(head, tail)`. This syntax can be also used in pattern matching on `List` values, with code that looks like this:

```
def headOption[A]: List[A] => Option[A] = {
  case Nil          => None
  case head :: tail => Some(head)
}
```

Examples of values created using Scala's standard `List` type are the empty list `Nil`, a one-element list `x :: Nil`, and a two-element list `x :: y :: Nil`. The same syntax `x :: y :: Nil` is used both for creating values of type `List` and for pattern-matching on such values.

The Scala library also defines the helper function `List()`, so that `List()` is the same as `Nil` and `List(1, 2, 3)` is the same as `1 :: 2 :: 3 :: Nil`. Lists are easier to use in the syntax `List(1, 2, 3)`. Pattern matching can also use that syntax when convenient:

```
val x: List[Int] = List(1, 2, 3)

x match {
  case List(a)      => ...
  case List(a, b, c) => ...
  case _             => ...
}
```

3.3.2 Tail recursion with List

Because the `List` type is defined by induction, it is straightforward to implement iterative computations with the `List` type using recursion.

A first example is the `map` function. We use reasoning by induction in order to figure out the implementation of `map`. The required type signature is

```
def map[A, B](xs: List[A])(f: A => B): List[B] = ???
```

The base case is an empty list, and we return again an empty list:

```
def map[A, B](xs: List[A])(f: A => B): List[B] = xs match {
  case Nil => Nil
  ...
}
```

In the inductive step, we have a pair `(head, tail)` in the case class `::`, with `head:A` and `tail>List[A]`. The pair can be pattern-matched with the syntax `head :: tail`. The `map` function should apply the argument `f` to the head value, which will give the first element of the resulting list. The remaining elements are computed by the induction assumption, i.e. by a recursive call to `map`:

```
def map[A, B](xs: List[A])(f: A => B): List[B] = xs match {
  case Nil      => Nil
  case head :: tail => f(head) :: map(tail)(f) // Not tail-recursive.
```

While this implementation is straightforward and concise, it is not tail-recursive. This will be a problem for large enough lists.

Instead of implementing the often-used methods such as `map` or `filter` one by one, let us implement `foldLeft`, because most of the other methods can be expressed via `foldLeft`.

The required type signature is

```
def foldLeft[A, R](xs: List[A])(init: R)(f: (R, A) => R): R = ???
```

Reasoning by induction, we start with the base case `xs == Nil`, where the only possibility is to return the value `init`:

```
def foldLeft[A, R](xs: List[A])(init: R)(f: (R, A) => R): R = xs match {
  case Nil      => init
  ...
}
```

The inductive step for `foldLeft` says that, given the values `head:A` and `tail>List[A]`, we need to apply the updater function to the previous accumulator value. That value is `init`. So we apply `foldLeft`

recursively to the tail of the list once we have the updated accumulator value:

```
@tailrec def foldLeft[A, R](xs: List[A])(init: R)(f: (R, A) => R): R =
  xs match {
    case Nil          => init
    case head :: tail =>
      val newInit = f(init, head) // Update the accumulator.
      foldLeft(tail)(newInit)(f) // Recursive call to 'foldLeft'.
  }
```

This implementation is tail-recursive because the recursive call to `foldLeft` is the last expression returned in a `case` branch.

Another example is a function for reversing a list. The Scala library defines the `reverse` method for this task, but we will show an implementation using `foldLeft`. The updater function *prepends* an element to a previous list:

```
def reverse[A](xs: List[A]): List[A] =
  xs.foldLeft(Nil: List[A])((prev, x) => x :: prev)

scala> reverse(List(1, 2, 3))
res0: List[Int] = List(3, 2, 1)
```

Without the explicit type annotation `Nil:List[A]`, the Scala compiler will decide that `Nil` has type `List[Nothing]`, and the types will not match later in the code. In Scala, one often finds that the initial value for `foldLeft` needs an explicit type annotation.

The `reverse` function can be used to obtain a tail-recursive implementation of `map` for `List`. The idea is to first use `foldLeft` to accumulate transformed elements:

```
scala> Seq(1, 2, 3).foldLeft(Nil:List[Int])((prev, x) => x*x :: prev)
res0: List[Int] = List(9, 4, 1)
```

The result is a reversed `.map(x => x*x)`, so we need to apply `reverse`:

```
def map[A, B](xs: List[A])(f: A => B): List[B] =
  xs.foldLeft(Nil: List[B])((prev, x) => f(x) :: prev).reverse

scala> map(List(1, 2, 3))(x => x*x)
res2: List[Int] = List(1, 4, 9)
```

This achieves stack safety at the cost of traversing the list twice. (This code is shown only as an example. The Scala library implements `map` using low-level tricks for better performance.)

Example 3.3.2.1 A definition of the **non-empty list** is similar to `List` except that the empty-list case is replaced by a 1-element case:

```
sealed trait NEL[A]
final case class Last[A](head: A)           extends NEL[A]
final case class More[A](head: A, tail: NEL[A]) extends NEL[A]
```

Values of a non-empty list look like this:

```
scala> val xs: NEL[Int] = More(1, More(2, Last(3))) // [1, 2, 3]
xs: NEL[Int] = More(1,More(2,Last(3)))

scala> val ys: NEL[String] = Last("abc") // One element, ["abc"].
ys: NEL[String] = Last(abc)
```

To create non-empty lists more easily, we implement a conversion function `toNEL` from an ordinary list. To guarantee that a non-empty list can be created, we give `toNEL` *two* arguments:

```
def toNEL[A](x: A, rest: List[A]): NEL[A] = rest match {
  case Nil          => Last(x)
  case y :: tail   => More(x, toNEL(y, tail))
} // Not tail-recursive: 'toNEL()' is used inside 'More(...)'.
```

To test:

```
scala> toNEL(1, List()) // Result = [1].
res0: NEL[Int] = Last(1)

scala> toNEL(1, List(2, 3)) // Result = [1, 2, 3].
res1: NEL[Int] = More(1,More(2,Last(3)))
```

The `head` method is safe for non-empty lists, unlike `head` for an ordinary `List`:

```
def head[A]: NEL[A] => A = {
  case Last(x)      => x
  case More(x, _)    => x
}
```

We can also implement a tail-recursive `foldLeft` function for non-empty lists:

```
@tailrec def foldLeft[A, R](n: NEL[A])(init: R)(f: (R, A) => R): R = n match {
  case Last(x)      => f(init, x)
  case More(x, tail) => foldLeft(tail)(f(init, x))(f)
}

scala> foldLeft(More(1, More(2, Last(3))))(0)(_ + _)
res2: Int = 6
```

Example 3.3.2.2 Use `foldLeft` to implement a `reverse` function for the type `NEL`. The required type signature and a sample test:

```
def reverse[A]: NEL[A] => NEL[A] = ???

scala> reverse(toNEL(10, List(20, 30))) // Result must be [30, 20, 10].
res3: NEL[Int] = More(30,More(20,Last(10)))
```

Solution We will use `foldLeft` to build up the reversed list as the accumulator value. It remains to choose the initial value of the accumulator and the updater function. We have already seen the code for reversing the ordinary list via the `foldLeft` method (Section 3.3.2),

```
def reverse[A](xs: List[A]): List[A] = xs.foldLeft(Nil: List[A])((prev, x) => x :: prev)
```

However, we cannot reuse the same code for non-empty lists by writing `More(x, prev)` instead of `x :: prev`, because the `foldLeft` operation works with non-empty lists differently. Since lists are always non-empty, the updater function is always applied to an initial value, and the code works incorrectly:

```
def reverse[A](xs: NEL[A]): NEL[A] =
  foldLeft(xs)(Last(head(xs)): NEL[A])((prev, x) => More(x, prev))

scala> reverse(toNEL(10, List(20, 30))) // Result = [30, 20, 10, 10].
res4: NEL[Int] = More(30,More(20,More(10,Last(10))))
```

The last element, 10, should not have been repeated. It was repeated because the initial accumulator value already contained the head element 10 of the original list. However, we cannot set the initial accumulator value to an empty list, since a value of type `NEL[A]` must be non-empty. It seems that we need to handle the case of a one-element list separately. So we begin by matching on the argument of `reverse`, and apply `foldLeft` only when the list is longer than 1 element:

```
def reverse[A]: NEL[A] => NEL[A] = {
  case Last(x)      => Last(x)    // 'reverse' is trivial.
  case More(x, tail) =>           // Use foldLeft on 'tail'.
    foldLeft(tail)(Last(x): NEL[A])((prev, x) => More(x, prev))
}

scala> reverse(toNEL(10, List(20, 30))) // Result = [30, 20, 10].
res5: NEL[Int] = More(30,More(20,Last(10)))
```

Exercise 3.3.2.3 Implement a function `toList` that converts a non-empty list into an ordinary Scala `List`. The required type signature and a sample test:

```
def toList[A](nel: NEL[A]): List[A] = ???

scala> toList(More(1, More(2, Last(3)))) // This is [1, 2, 3].
res6: List[Int] = List(1, 2, 3)
```

Exercise 3.3.2.4 Implement a `map` function for the type `NEL`. Type signature and a sample test:

```
def map[A,B](xs: NEL[A])(f: A => B): NEL[B] = ???

scala> map[Int, Int](toNEL(10, List(20, 30)))(_ + 5) // Result = [15, 25, 35].
res7: NEL[Int] = More(15, More(25, Last(35)))
```

Exercise 3.3.2.5 Implement a function `concat` that concatenates two non-empty lists:

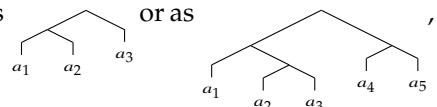
```
def concat[A](xs: NEL[A], ys: NEL[A]): NEL[A] = ???

scala> concat(More(1, More(2, Last(3))), More(4, Last(5))) // Result is [1, 2, 3, 4, 5].
res8: NEL[Int] = More(1, More(2, More(3, More(4, Last(5)))))
```

3.3.3 Binary trees

We will consider four kinds of trees defined as recursive disjunctive types: binary trees, rose trees, regular-shaped trees, and abstract syntax trees.

Examples of a **binary tree** with leaves of type `A` can be drawn as



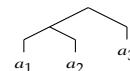
where a_i are some values of type `A`.

An inductive definition says that a binary tree is either a leaf with a value of type `A` or a branch containing *two* previously defined binary trees. Translating this definition into code, we get

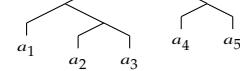
```
sealed trait Tree2[A]
final case class Leaf[A](a: A) extends Tree2[A]
final case class Branch[A](x: Tree2[A], y: Tree2[A]) extends Tree2[A]
```

Here are some examples of code expressions and the corresponding trees that use this definition:

```
Branch(Branch(Leaf("a1"), Leaf("a2")), Leaf("a3"))
```



```
Branch(Branch(Leaf("a1"), Branch(Leaf("a2"), Leaf("a3"))), Branch(Leaf("a4"), Leaf("a5")))
```



Recursive functions on trees are translated into concise code. For instance, the function `foldLeft` for trees of type `Tree2` is implemented as

```
def foldLeft[A, R](t: Tree2[A])(init: R)(f: (R, A) => R): R = t match {
  case Leaf(a)      => f(init, a)
  case Branch(t1, t2) =>
    val r1 = foldLeft(t1)(init)(f) // Fold the left branch.
    foldLeft(t2)(r1)(f) // Starting from 'r1', fold the right branch.
}
```

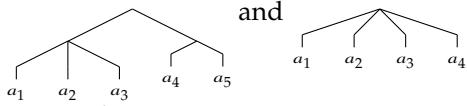
Note that this function *cannot* be made tail-recursive using the accumulator trick, because `foldLeft` needs to call itself twice in the `Branch` case. To test:

```
val t: Tree2[String] = Branch(Branch(Leaf("a1"), Leaf("a2")), Leaf("a3"))

scala> foldLeft(t)("")(_ + " " + _)
res0: String = " a1 a2 a3"
```

3.3.4 Rose trees

A **rose tree** is similar to the binary tree except the branches contain a non-empty list of trees. Because of that, a rose tree can fork into arbitrarily many branches at each node, rather than always into two branches as the binary tree does; for example,



A possible definition of a data type for the rose tree is

```
sealed trait TreeN[A]
final case class Leaf[A](a: A) extends TreeN[A]
final case class Branch[A](ts: NEL[TreeN[A]]) extends TreeN[A]
```

Since we used a non-empty list `NEL`, a `Branch()` value is guaranteed to have at least one branch. If we used an ordinary `List` instead, we could (by mistake) create a tree with no leaves and no branches.

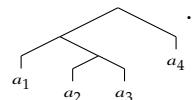
Exercise 3.3.4.1 Define the function `foldLeft` for a rose tree, using `foldLeft` for the type `NEL`. Type signature and a test:

```
def foldLeft[A, R](t: TreeN[A])(init: R)(f: (R, A) => R): R = ???

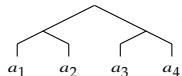
scala> foldLeft(Branch(More(Leaf(1), More(Leaf(2), Last(Leaf(3))))))(0)(_ + _)
res0: Int = 6
```

3.3.5 Regular-shaped trees

Binary trees and rose trees may choose to branch or not to branch at any given node, resulting in structures that may have different branching depths at different nodes, such as



regular-shaped tree always branches in the same way at every node until a chosen total depth, e.g.,



, where all nodes at depth 0 and 1 always branch into two, while nodes at depth 2 do not branch. The branching number is fixed for a given type of a regular-shaped tree; in this example, the branching number is 2, so it is a regular-shaped *binary* tree.

How can we define a data type representing a regular-shaped binary tree? We need a type that is either a single value, or a pair of values, or a pair of pairs, etc. Begin with the non-recursive (but, of course, impractical) definition

```
sealed trait RTree[A]
final case class Leaf[A](x: A) extends RTree[A]
final case class Branch1[A](xs: (A, A)) extends RTree[A]
final case class Branch2[A](xs: ((A, A), (A, A))) extends RTree[A]
???
// Need an infinitely long definition.
```

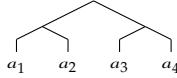
The case `Branch1` describes a regular-shaped tree with total depth 1, the case `Branch2` has total depth 2, and so on. Now, we cannot rewrite this definition as a recursive type because the case classes do not have the same structure. The non-trivial trick is to notice that each case class `Branchn` uses the previous case class's data structure *with the type parameter set to `(A, A)` instead of `A`*. So we can rewrite this definition as

```
sealed trait RTree[A]
final case class Leaf[A](x: A) extends RTree[A]
final case class Branch1[A](xs: Leaf[(A, A)]) extends RTree[A]
final case class Branch2[A](xs: Branch1[(A, A)]) extends RTree[A]
???
// Need an infinitely long definition.
```

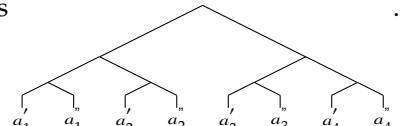
We can now apply the type recursion trick: replace the type `Branchn-1[(A, A)]` in the definition of `Branchn` by the recursively used type `RTree[(A, A)]`. Now we can define a regular-shaped binary tree:

```
sealed trait RTree[A]
final case class Leaf[A](x: A) extends RTree[A]
final case class Branch[A](xs: RTree[(A, A)]) extends RTree[A]
```

Since we used some tricks to figure out the definition of `RTree[A]`, let us verify that this definition actually describes the recursive disjunctive type we wanted. The only way to create a structure of type `RTree[A]` is to create a `Leaf[A]` or a `Branch[A]`. A value of type `Leaf[A]` is a correct regularly-shaped tree. It remains to consider the case of `Branch[A]`. Creating a `Branch[A]` requires a previously created `RTree` with values of type `(A, A)` instead of `A`. By the inductive assumption, the previously created `RTree[A]` would have the correct shape. Now, it is clear that if we replace the type parameter `A` by the pair `(A, A)`, a regular-shaped tree such as



remains regular-shaped but becomes one level deeper, which can be drawn (replacing each a_i by a pair a'_i, a''_i) as



We see that `RTree[A]` is a correct definition of a regular-shaped binary tree.

Example 3.3.5.1 Define a (non-tail-recursive) `map` function for a regular-shaped binary tree. The required type signature and a test:

```
def map[A, B](t: RTree[A])(f: A => B): RTree[B] = ???

scala> map(Branch(Branch(Leaf(((1,2),(3,4))))))(_ * 10)
res0: RTree[Int] = Branch(Branch(Leaf(((10,20),(30,40)))))
```

Solution Begin by pattern-matching on the tree:

```
def map[A, B](t: RTree[A])(f: A => B): RTree[B] = t match {
  case Leaf(x)      => ???
  case Branch(xs)   => ???
}
```

In the base case, we have no choice but to return `Leaf(f(x))`.

```
def map[A, B](t: RTree[A])(f: A => B): RTree[B] = t match {
  case Leaf(x)      => Leaf(f(x))
  case Branch(xs)   => ???
}
```

In the inductive step, we are given a previous tree value `xs:RTree[(A, A)]`. It is clear that we need to apply `map` recursively to `xs`. Let us try:

```
def map[A, B](t: RTree[A])(f: A => B): RTree[B] = t match {
  case Leaf(x)      => Leaf(f(x))
  case Branch(xs)   => Branch(map(xs)(f)) // Type error!
}
```

Here, `map(xs)(f)` has an incorrect type of the function `f`. Since `xs` has type `RTree[(A, A)]`, the recursive call `map(xs)(f)` requires `f` to be of type `((A, A)) => (B, B)` instead of `A => B`.

So, we need to provide a function of the correct type instead of `f`. A function of type `((A, A)) => (B, B)` will be obtained out of `f: A => B` if we apply `f` to each part of the tuple `(A, A)`; the code for that function is `{ case (x, y) => (f(x), f(y)) }`. Therefore, we can implement `map` as

```
def map[A, B](t: RTree[A])(f: A => B): RTree[B] = t match {
  case Leaf(x)      => Leaf(f(x))
  case Branch(xs)   => Branch(map(xs){ case (x, y) => (f(x), f(y)) })
}
```

This code is not tail-recursive since it calls `map` inside an expression.

Exercise 3.3.5.2 Using tail recursion, compute the depth of a regular-shaped binary tree of type `RTree`. (An `RTree` of depth n has 2^n leaf values.) The required type signature and a test:

```
@tailrec def depth[A](t: RTree[A]): Int = ???

scala> depth(Branch(Branch(Leaf(("a", "b"), ("c", "d")))))
res2: Int = 2
```

Exercise 3.3.5.3* Define a tail-recursive function `foldLeft` for a regular-shaped binary tree. The required type signature and a test:

```
@tailrec def foldLeft[A, R](t: RTree[A])(init: R)(f: (R, A) => R): R = ???

scala> foldLeft(Branch(Branch(Leaf((1,2),(3,4))))) (0) (_ + _)
res0: Int = 10

scala> foldLeft(Branch(Branch(Leaf(("a", "b"), ("c", "d"))))) ("") (_ + _)
res1: String = abcd
```

3.3.6 Abstract syntax trees

Expressions in formal languages are represented by abstract syntax trees. An **abstract syntax tree** (or **AST** for short) is defined as either a leaf of one of the available leaf types, or a branch of one of the available branch types. All the available leaf and branch types must be specified as part of the definition of an AST. In other words, one must specify the data carried by leaves and branches, as well as the branching numbers.

To illustrate how ASTs are used, let us rewrite Example 3.2.2.4 via an AST. We view Example 3.2.2.4 as a small sub-language that deals with “safe integers” and supports the “safe arithmetic” operations `Sqrt`, `Add`, `Mul`, and `Div`. Example calculations in this sub-language are $\sqrt{16} * (1 + 2) = 12$; $20 + 1/0 = \text{error}$; and $10 + \sqrt{-1} = \text{error}$.

We can implement this sub-language in two stages. The first stage will create a data structure (an AST) that represents an unevaluated expression in the sub-language. The second stage will evaluate that AST to obtain either a number or an error message.

A straightforward way of defining a data structure for an AST is to use a disjunctive type whose parts describe all the possible operations of the sub-language. We will need one case class for each of `Sqrt`, `Add`, `Mul`, and `Div`. An additional operation, `Num`, will lift ordinary integers into “safe integers”. So, we define the disjunctive type for “arithmetic sub-language expressions” as

```
sealed trait Arith
final case class Num(x: Int) extends Arith
final case class Sqrt(x: Arith) extends Arith
final case class Add(x: Arith, y: Arith) extends Arith
final case class Mul(x: Arith, y: Arith) extends Arith
final case class Div(x: Arith, y: Arith) extends Arith
```

A value of type `Arith` is either a `Num(x)` for some integer `x`, or an `Add(x, y)` where `x` and `y` are previously defined `Arith` expressions, or another operation.

This type definition is similar to the binary tree type

```
sealed trait Tree
final case class Leaf(x: Int) extends Tree
final case class Branch(x: Tree, y: Tree) extends Tree
```

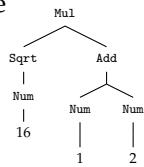
if we rename `Leaf` to `Num` and `Branch` to `Add`. However, the `Arith` type contains four different types of “branches”, some with branching number 1 and others with branching number 2.

This example illustrates the structure of an AST: it is a tree of a general shape, where leaves and branches are chosen from a specified set of allowed possibilities. In this example, we have a single allowed type of leaf (`Num`) and four allowed types of branches (`Sqrt`, `Add`, `Mul`, and `Div`).

This completes the first stage of implementing the sub-language. We may now use the disjunctive type `Arith` to create expressions in the sub-language. For example, $\sqrt{16} * (1 + 2)$ is represented by

```
scala> val x: Arith = Mul(Sqrt(Num(16)), Add(Num(1), Num(2)))
x: Arith = Mul(Sqrt(Num(16)),Add(Num(1),Num(2)))
```

We can visualize x as the abstract syntax tree



The expressions $20 + 1/0$ and $10 * \sqrt{-1}$ are represented by

```
scala> val y: Arith = Add(Num(20), Div(Num(1), Num(0)))
y: Arith = Add(Num(20),Div(Num(1),Num(0)))

scala> val z: Arith = Add(Num(10), Sqrt(Num(-1)))
z: Arith = Add(Num(10),Sqrt(Num(-1)))
```

As we see, the expressions x , y , and z *remain unevaluated*; each of them is a data structure that encodes a tree of operations of the sub-language. These operations will be evaluated at the second stage of implementing the sub-language.

To evaluate expressions in the “safe arithmetic”, we can implement a function with type signature `run: Arith => Either[String, Int]`. That function plays the role of an **interpreter** or “runner” for programs written in the sub-language. The runner will destructure the expression tree and execute all the operations, taking care of possible errors.

To implement `run`, we need to define required arithmetic operations on the type `Either[String, Int]`. For instance, we need to be able to add or multiply values of that type. Instead of custom code from Example 3.2.4, we can use the standard `map` and `flatMap` methods defined on `Either`. For example, addition and multiplication of two “safe integers” is implemented as

```
def add(x: Either[String, Int], y: Either[String, Int]): Either[String, Int] = x.flatMap { r1 => y.map(r2 => r1 + r2) }
def mul(x: Either[String, Int], y: Either[String, Int]): Either[String, Int] = x.flatMap { r1 => y.map(r2 => r1 * r2) }
```

while the “safe division” is

```
def div(x: Either[String, Int], y: Either[String, Int]): Either[String, Int] = x.flatMap { r1 => y.flatMap { r2 =>
  if (r2 == 0) Left(s"error: $r1 / $r2") else Right(r1 / r2) }
}
```

With this code, we can implement the runner as a recursive function,

```
def run: Arith => Either[String, Int] = {
  case Num(x)      => Right(x)
  case Sqrt(x)     => run(x).flatMap { r =>
    if (r < 0) Left(s"error: sqrt($r)") else Right(math.sqrt(r).toInt)
  }
  case Add(x, y)   => add(run(x), run(y))
  case Mul(x, y)   => mul(run(x), run(y))
  case Div(x, y)   => div(run(x), run(y))
}
```

Test it with the values x , y , z defined previously:

```
scala> run(x)
res0: Either[String, Int] = Right(12)

scala> run(y)
res1: Either[String, Int] = Left("error: 1 / 0")

scala> run(z)
res2: Either[String, Int] = Left("error: sqrt(-1)")
```

3.4 Summary

What problems can we solve now?

- Represent values from a disjoint domain by a custom-defined disjunctive type.
- Use disjunctive types instead of exceptions to indicate failures.
- Use standard disjunctive types `Option`, `Try`, `Either` and methods defined for them.
- Define recursive disjunctive types (e.g., lists and trees) and implement recursive functions that work with them.

The following examples and exercises illustrate these tasks.

3.4.1 Solved examples

Example 3.4.1.1 Define a disjunctive type `DayOfWeek` representing the seven days.

Solution Since there is no information other than the label on each day, we use empty case classes:

```
sealed trait DayOfWeek
final case class Sunday()      extends DayOfWeek
final case class Monday()     extends DayOfWeek
final case class Tuesday()    extends DayOfWeek
final case class Wednesday()  extends DayOfWeek
final case class Thursday()   extends DayOfWeek
final case class Friday()     extends DayOfWeek
final case class Saturday()   extends DayOfWeek
```

This data type is analogous to an enumeration type in C or C++:

```
typedef enum { Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday } DayOfWeek;
```

Example 3.4.1.2 Modify `DayOfWeek` so that the values additionally represent a restaurant name and total amount for Fridays and a wake-up time on Saturdays.

Solution For the days where additional information is given, we use non-empty case classes:

```
sealed trait DayOfWeekX
final case class Sunday()          extends DayOfWeekX
final case class Monday()         extends DayOfWeekX
final case class Tuesday()        extends DayOfWeekX
final case class Wednesday()      extends DayOfWeekX
final case class Thursday()       extends DayOfWeekX
final case class Friday(restaurant: String, amount: Int) extends DayOfWeekX
final case class Saturday(wakeUpAt: java.time.LocalTime) extends DayOfWeekX
```

This data type is no longer equivalent to an enumeration type.

Example 3.4.1.3 Define a disjunctive type that describes the real roots of the equation $ax^2 + bx + c = 0$, where a, b, c are arbitrary real numbers.

Solution Begin by solving the equation and enumerating all the possible cases. It may happen that $a = b = c = 0$, and then all x are roots. If $a = b = 0$ but $c \neq 0$, the equation is $c = 0$, which has no roots. If $a = 0$ but $b \neq 0$, the equation becomes $bx + c = 0$, having a single root. If $a \neq 0$ and $b^2 > 4ac$, we have two distinct real roots. If $a \neq 0$ and $b^2 = 4ac$, we have one real root. If $b^2 < 4ac$, we have no real roots. The resulting type definition can be written as

```
sealed trait RootsOfQ2
final case class AllRoots()          extends RootsOfQ2
final case class ConstNoRoots()      extends RootsOfQ2
final case class Linear(x: Double)  extends RootsOfQ2
final case class NoRealRoots()       extends RootsOfQ2
```

```
final case class OneRootQ(x: Double)           extends RootsOfQ2
final case class TwoRootsQ(x: Double, y: Double) extends RootsOfQ2
```

This disjunctive type contains six parts, among which three parts are empty tuples and two parts are single-element tuples; but this is not a useless redundancy. We would lose information if we reused `Linear` for the two cases $a = 0, b \neq 0$ and $a \neq 0, b^2 = 4ac$, or if we reused `NoRoots()` for representing all three different no-roots cases.

Example 3.4.1.4 Define a function `rootAverage` that computes the average value of all real roots of a general quadratic equation, where the set of roots is represented by the type `RootsOfQ2` defined in Example 3.4.1.3. The required type signature is

```
val rootAverage: RootsOfQ2 => Option[Double] = ???
```

The function should return `None` if the average is undefined.

Solution The average is defined only in cases `Linear`, `OneRootQ`, and `TwoRootsQ`. In all other cases, we must return `None`. We implement this via pattern matching:

```
val rootAverage: RootsOfQ2 => Option[Double] = roots => roots match {
  case Linear(x)      => Some(x)
  case OneRootQ(x)    => Some(x)
  case TwoRootsQ(x, y) => Some((x + y) * 0.5)
  case _               => None
}
```

We do not need to enumerate all other cases since the underscore (`_`) matches everything that the previous cases did not match.

The often-used code pattern of the form `x => x match { case ... }` can be shortened to the nameless function syntax `{ case ... }`. The code then becomes

```
val rootAverage: RootsOfQ2 => Option[Double] = {
  case Linear(x)      => Some(x)
  case OneRootQ(x)    => Some(x)
  case TwoRootsQ(x, y) => Some((x + y) * 0.5)
  case _               => None
}
```

Test it:

```
scala> Seq(NoRealRoots(), OneRootQ(1.0), TwoRootsQ(1.0, 2.0), AllRoots()).map(rootAverage)
res0: Seq[Option[Double]] = List(None, Some(1.0), Some(1.5), None)
```

Example 3.4.1.5 Generate 100 quadratic equations $x^2 + bx + c = 0$ with random coefficients b, c (uniformly distributed between -1 and 1) and compute the mean of the largest real roots from all these equations.

Solution Use the type `QEqu` and the `solve` function from Example 3.2.2.1. A sequence of equations with random coefficients is created by applying the method `Seq.fill`:

```
def random(): Double = scala.util.Random.nextDouble() * 2 - 1
val coeffs: Seq[QEqu] = Seq.fill(100)(QEqu(random(), random()))
```

Now we can use the `solve` function to compute all roots:

```
val solutions: Seq[RootsOfQ] = coeffs.map(solve)
```

For each set of roots, compute the largest root:

```
scala> val largest: Seq[Option[Double]] = solutions.map {
  case OneRoot(x)      => Some(x)
  case TwoRoots(x, y)  => Some(math.max(x, y))
  case _               => None
}
largest: Seq[Option[Double]] = List(None, Some(0.9346072365885472), Some(1.1356234869160806),
  Some(0.9453181931646322), Some(1.1595052441078866), None, Some(0.5762252742788) ...
```

It remains to remove the `None` values and to compute the mean of the resulting sequence. The Scala library defines the `flatten` method that removes `Nones` and transforms `Seq[Option[A]]` into `Seq[A]`:

```
scala> largest.flatten
res0: Seq[Double] = List(0.9346072365885472, 1.1356234869160806, 0.9453181931646322,
1.1595052441078866, 0.5762252742788...)
```

Now compute the mean of the last sequence. Since the `flatten` operation is preceded by `map`, we can replace it by a `flatMap`. The final code is

```
val largest = Seq.fill(100)(QEqu(random(), random()))
  .map(solve)
  .flatMap {
    case OneRoot(x)      => Some(x)
    case TwoRoots(x, y)  => Some(math.max(x, y))
    case _                => None
  }

scala> largest.sum / largest.size
res1: Double = 0.7682649774589514
```

Example 3.4.1.6 Implement a function with type signature

```
def f1[A, B]: Option[Either[A, B]] => Either[A, Option[B]] = ???
```

The function should preserve information as much as possible.

Solution Begin by pattern matching on the argument:

```
1 def f1[A, B]: Option[Either[A, B]] => Either[A, Option[B]] = {
2   case None          => ???
3   case Some(eab:Either[A, B]) => ???
4 }
```

of the pattern variable `eab` from the fact that we are matching a value of type `Option[Either[A, B]]`.

In the scope of line 2, we need to return a value of type `Either[A, Option[B]]`. A value of that type must be either a `Left(x)` for some `x:A`, or a `Right(y)` for some `y:Option[B]`, where `y` must be either `None` or `Some(z)` with a `z:B`. However, in our case the code is of the form `case None => ???`, and we cannot produce any values `x:A` or `z:B` since `A` and `B` are arbitrary, unknown types. The only remaining possibility is to return `Right(y)` with `y = None`, and so the code must be

```
...
case None => Right(None) // No other choice here.
```

In the next scope, we can perform pattern matching on the value `eab`:

```
...
case Some(eab: Either[A, B]) = eab match {
  case Left(a)  => ???
  case Right(b) => ???
}
```

we have a value of type `A` but no values of type `B`. So we have two possibilities: to return `Left(a)` or to return `Right(None)`. If we decide to return `Left(a)`, the code is

```
1 def f1[A, B]: Option[Either[A, B]] => Either[A, Option[B]] = {
2   case None          => Right(None) // No other choice here.
3   case Some(eab)    => eab match {
4     case Left(a)    => Left(a)   // Could return Right(None) here.
5     case Right(b)   => ???
6   }
7 }
```

i.e. we will lose information. So we return `Left(a)` in line 4.

Reasoning similarly for line 5, we find that we may return `Right(None)` or `Right(Some(b))`. The first

In line 3, we wrote the **type annotation** `:Either[A, B]` only for clarity; it is not required here because the Scala compiler can deduce the type

It remains to figure out what expressions to write in each case. In the case `Left(a) => ???`, we have a value of type `A`, and we need to compute a value of type `Either[A, Option[B]]`. We execute the same argument as before: The return value must be `Left(x)` for some `x:A`, or `Right(y)` for some `y:Option[B]`. At this point,

Let us decide whether to return `Left(a)` or `Right(None)` in line 4. Both choices will satisfy the required return type `Either[A, Option[B]]`. However, if we return `Right(None)` in that line, we will ignore the given value `a:A`,

choice ignores the given value of $b:B$. To preserve information, we make the second choice:

```

1 def f1[A, B]: Option[Either[A, B]] => Either[A, Option[B]] = {
2   case None      => Right(None)
3   case Some(eab) => eab match {
4     case Left(a)   => Left(a)
5     case Right(b)  => Right(Some(b))
6   }
7 }
```

We can now refactor this code into a somewhat more readable form by using nested patterns:

```

def f1[A, B]: Option[Either[A, B]] => Either[A, Option[B]] = {
  case None      => Right(None)
  case Some(Left(a)) => Left(a)
  case Some(Right(b)) => Right(Some(b))
}
```

Example 3.4.1.7 Implement a function with the type signature

```
def f2[A, B]: (Option[A], Option[B]) => Option[(A, B)] = ???
```

The function should preserve information as much as possible.

Solution Begin by pattern matching on the argument:

```

1 def f2[A, B]: (Option[A], Option[B]) => Option[(A, B)] = {
2   case (Some(a), Some(b)) => ???
3   ...
4 }
```

where we would need to obtain values $x:A$ and $y:B$. Since A and B are arbitrary types, we cannot produce new values x and y from scratch. The only way of obtaining $x:A$ and $y:B$ is to set $x = a$ and $y = b$. So, our choices are to return `Some((a, b))` or `None`. We reject returning `None` since that would unnecessarily lose information. Thus, we continue writing code as

```

1 def f2[A, B]: (Option[A], Option[B]) => Option[(A, B)] = {
2   case (Some(a), Some(b)) => Some((a, b))
3   case (Some(a), None)    => ???
4   ...
5 }
```

In line 3, we have a value $a:A$ but no values of type B . Since the type B is arbitrary, we cannot produce any values of type B to return a value of the form `Some((x, y))`. So, `None` is the only computable value of type `Option[(A, B)]` in line 3. We continue to write the code:

```

1 def f2[A, B]: (Option[A], Option[B]) => Option[(A, B)] = {
2   case (Some(a), Some(b)) => Some((a, b))
3   case (Some(a), None)    => None // No other choice here.
4   case (None, Some(b))   => ???
5   case (None, None)      => ???
6 }
```

In lines 4–5, we find that there is no choice other than returning `None`. So we can simplify the code:

```

def f2[A, B]: (Option[A], Option[B]) => Option[(A, B)] = {
  case (Some(a), Some(b)) => Some((a, b))
  case _                  => None // No other choice here.
}
```

In line 2, we have values $a:A$ and $b:B$, and we need to compute a value of type `Option[(A, B)]`. A value of that type is either `None` or `Some((x, y))`

3.4.2 Exercises

Exercise 3.4.2.1 Define a disjunctive type `CellState` representing the visual state of one cell in the *Minesweeper*³ game: A cell can be closed (showing nothing), or show a bomb, or be open and show the number of bombs in neighbor cells.

³[https://en.wikipedia.org/wiki/Minesweeper_\(video_game\)](https://en.wikipedia.org/wiki/Minesweeper_(video_game))

Exercise 3.4.2.2 Define a function from `Seq[Seq[CellState]]` to `Int`, counting the total number of cells with zero neighbor bombs shown.

Exercise 3.4.2.3 Define a disjunctive type `RootOfLinear` representing all possibilities for the solution of the equation $ax + b = 0$ for arbitrary real a, b . (The possibilities are: no roots; one root; all x are roots.) Implement the solution as a function `solve1` with type signature

```
def solve1: ((Double, Double)) => RootOfLinear = ???
```

Exercise 3.4.2.4 Given a `Seq[(Double, Double)]` containing pairs (a, b) of the coefficients of $ax + b = 0$, produce a `Seq[Double]` containing the roots of that equation when a unique root exists. Use the type `RootOfLinear` and the function `solve1` defined in Exercise 3.4.2.3.

Exercise 3.4.2.5 The case class `Subscriber` was defined in Example 3.2.3.1. Given a `Seq[Subscriber]`, compute the sequence of email addresses for all subscribers that did *not* provide a phone number.

Exercise 3.4.2.6* In this exercise, a “procedure” is a function of type `Unit => Unit`; an example of a procedure is `{ () => println("hello") }`. Define a disjunctive type `Proc` for an abstract syntax tree representing three operations on procedures: 1) `Func[A](f)`, create a procedure from a function `f` of type `Unit => A`, where `A` is a type parameter. (Note that the type `Proc` does not have any type parameters.) 2) `Sequ(p1, p2)`, execute two procedures sequentially. 3) `Para(p1, p2)`, execute two procedures in parallel. Then implement a “runner” that converts a `Proc` into a `Future[Unit]`, running the computations either sequentially or in parallel as appropriate. Test with this code:

```
sealed trait Proc; final case class Func[A](???) // And so on.
def runner: Proc => Future[Unit] = ???
val proc1: Proc = Func{_ => Thread.sleep(200); println("hello1")}
val proc2: Proc = Func{_ => Thread.sleep(400); println("hello2")}

scala> runner(Sequ(Para(proc2, proc1), proc2))
hello1
hello2
hello2
```

Exercise 3.4.2.7 Implement functions that have a given type signature, preserving information as much as possible:

```
def f1[A, B]: Option[(A, B)] => (Option[A], Option[B]) = ???
def f2[A, B]: Either[A, B] => (Option[A], Option[B]) = ???
def f3[A, B, C]: Either[A, Either[B, C]] => Either[Either[A, B], C] = ???
```

Exercise 3.4.2.8 Define a parameterized type `EvenList[A]` representing a list of values of type `A` that is guaranteed to have an even number of elements (zero, two, four, etc.). Implement functions `foldLeft` and `map` for `EvenList`.

3.5 Discussion and further developments

3.5.1 Disjunctive types as mathematical sets

To understand the properties of disjunctive types from the mathematical point of view, consider a function whose argument is a disjunctive type, such as

```
def isDoubleRoot(r: RootsOfQ) = ...
```

The type of the argument `r: RootsOfQ` represents the mathematical domain of the function, that is, the set of admissible values of the argument `r`. What kind of domain is that? The set of real roots of a quadratic equation $x^2 + bx + c = 0$ can be empty, or it can contain a single real number x , or a pair of real numbers (x, y) . Geometrically, a number x is pictured as a point on a line (a one-dimensional space), and pair of numbers (x, y) is pictured as a point on a Cartesian plane (a two-dimensional

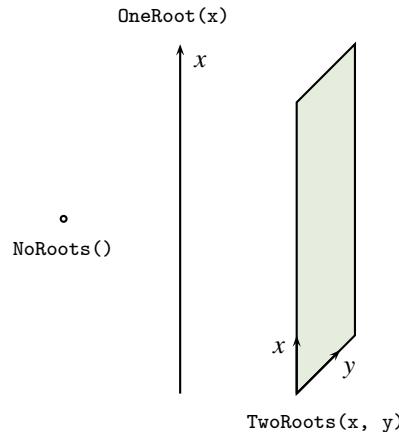


Figure 3.1: The disjoint domain represented by the type `RootsOfQ`.

space). The no-roots case corresponds to a zero-dimensional space, which is pictured as a single point (see Figure 3.1). The point, the line, and the plane do not intersect (have no common points); together, they form the domain of the possible roots. Such domains are called **disjoint**. So, the set of real roots of a quadratic equation $x^2 + bx + c = 0$ is a disjoint domain containing three parts.

In the mathematical notation, a one-dimensional real space is denoted by \mathbb{R} , a two-dimensional space by \mathbb{R}^2 , and a zero-dimensional space by \mathbb{R}^0 . At first, we may think that the mathematical representation of the type `RootsOfQ` is a union of the three sets, $\mathbb{R}^0 \cup \mathbb{R}^1 \cup \mathbb{R}^2$. But an ordinary union of sets would not always work correctly because we need to distinguish the parts of the union unambiguously, even if some parts have the same type. For instance, the disjunctive type shown in Example 3.4.1.3 cannot be correctly represented by the mathematical union

$$\mathbb{R}^0 \cup \mathbb{R}^0 \cup \mathbb{R}^1 \cup \mathbb{R}^0 \cup \mathbb{R}^1 \cup \mathbb{R}^2$$

because $\mathbb{R}^0 \cup \mathbb{R}^0 = \mathbb{R}^0$ and $\mathbb{R}^1 \cup \mathbb{R}^1 = \mathbb{R}^1$, so

$$\mathbb{R}^0 \cup \mathbb{R}^0 \cup \mathbb{R}^1 \cup \mathbb{R}^0 \cup \mathbb{R}^1 \cup \mathbb{R}^2 = \mathbb{R}^0 \cup \mathbb{R}^1 \cup \mathbb{R}^2 .$$

This representation has lost the distinction between e.g., `Linear(x)` and `OneRootQ(x)`.

In the Scala code, each part of a disjunctive type must be distinguished by a unique name such as `NoRoots`, `OneRoot`, and `TwoRoots`. To represent this mathematically, we can attach a distinct label to each part of the union. Labels are symbols without any special meaning, and we can just assume that labels are names of Scala case classes. Parts of the union are then represented by sets of pairs such as $(\text{OneRoot}, x)_{x \in \mathbb{R}^1}$. Then the domain `RootsOfQ` is expressed as

$$\text{RootsOfQ} = (\text{NoRoots}, u)_{u \in \mathbb{R}^0} \cup (\text{OneRoot}, x)_{x \in \mathbb{R}^1} \cup (\text{TwoRoots}, (x, y))_{(x, y) \in \mathbb{R}^2} .$$

This is an ordinary union of mathematical sets, but each of the sets has a unique label, so no two values from different parts of the union could possibly be equal. This kind of labeled union is called a **disjoint union** or “tagged union”. Each element of the disjoint union is a pair of the form `(label, data)`, where the label uniquely identifies the part of the union, and the data can have any chosen type such as \mathbb{R}^1 . If we use disjoint unions, we cannot confuse different parts of the union even if their data have the same type, because labels are required to be distinct.

Disjoint unions are not often used in mathematics, but they are needed in software engineering because real-life data often belongs to disjoint domains.

Named Unit types At first sight, it may seem strange that the zero-dimensional space is represented by a set containing *one* point. Why should we not use an empty set (rather than a set with one point) to represent the case where the equation has no real roots? The reason is that we are required to represent not only the values of the roots but also the information *about* the existence of the roots. The case with no real roots needs to be represented by some *value* of type `Roots0fQ`. This value cannot be missing, which would happen if we used an empty set to represent the no-roots case. It is natural to use the named empty tuple `NoRoots()` to represent this case, just as we used a named 2-tuple `TwoRoots(x, y)` to represent the case of two roots.

Consider the value u used by the mathematical set $(\text{NoRoots}, u)_{u \in \mathbb{R}^0}$. Since \mathbb{R}^0 consists of a single point, there is only *one* possible value of u . Similarly, the `Unit` type in Scala has only one distinct value, written as `()`. A case class with no parts, such as `NoRoots`, has only one distinct value, written as `NoRoots()`. This Scala value is fully analogous to the mathematical notation $(\text{NoRoots}, u)_{u \in \mathbb{R}^0}$.

So, case classes with no parts are quite similar to `Unit` except for an added name, e.g., `NoRoots()` is the `Unit` value `()` with name `NoRoots`. For this reason, they can be viewed as “named unit” types.

3.5.2 Disjunctive types in other programming languages

Disjunctive types and pattern matching turns out to be one of the defining features of functional programming languages. Languages that were not designed for functional programming do not support these features, while ML, OCaml, Haskell, F#, Scala, Swift, Elm, and PureScript support disjunctive types and pattern matching as part of the language design.

It is remarkable that named tuple types (also called “structs” or “records”) are provided in almost every programming language, while disjunctive types are almost never present except in languages designed for the FP paradigm. (Ada and Pascal are the only languages that support disjunctive types without other FP features.⁴)

The `union` types in C and C++ are not disjunctive types because it is not possible to determine which part of the union is represented by a given value. A `union` declaration in C looks like this,

```
union { int x; double y; long z; } di;
```

Without a label, we (and the compiler) will not know whether a given value of type `di` represents an `int`, a `double`, or a `long`. This will lead to errors that are hard to detect.

Programming languages of the C family (C, C++, Objective C, Java) support `enumeration` (`enum`) types, which are a limited form of disjunctive types. An `enum` type declaration in Java looks like this:

```
enum Color { RED, GREEN, BLUE; }
```

In Scala, this is equivalent to a disjunctive type containing three *empty* tuples,

```
sealed trait Color
final case class RED()    extends Color
final case class GREEN()  extends Color
final case class BLUE()   extends Color
```

If `enum` types were “enriched” with extra data, so that the tuples could be non-empty, we would obtain the full functionality of disjunctive types. A definition of `Roots0fQ` could then look like this:

```
enum Roots0fQ {      // This is not valid in Java!
  NoRoots(), OneRoot(Double x), TwoRoots(Double x, Double y);
}
```

A future version of Scala 3 will have a syntax for disjunctive types⁵ that resembles “enriched `enum`”:

```
enum Roots0fQ {
  case NoRoots;  case OneRoot(x: Double);   case TwoRoots(x: Double, y: Double)
}
```

⁴[https://en.wikipedia.org/wiki/Comparison_of_programming_languages_\(basic_instructions\)#Other_types](https://en.wikipedia.org/wiki/Comparison_of_programming_languages_(basic_instructions)#Other_types)

⁵<https://dotty.epfl.ch/docs/reference/enums/adts.html>

For comparison, the syntax for a disjunctive type equivalent to `RootsOfQ` in OCaml and Haskell is

```
(* OCaml *)
type RootsOfQ = NoRoots | OneRoot of float | TwoRoots of float*float

-- Haskell
data RootsOfQ = NoRoots | OneRoot Double | TwoRoots (Double, Double)
```

This is more concise than the Scala syntax. When reasoning about disjunctive types, it is inconvenient to write out long type definitions. Chapter 5 defines a mathematical notation designed for efficient reasoning about types.

3.5.3 Disjunctions and conjunctions in formal logic

In logic, a **proposition** is a logical formula that could be true or false. A **disjunction** of propositions A, B, C is denoted by $A \vee B \vee C$ and is true if and only if *at least one* of A, B, C is true. A **conjunction** of A, B, C is denoted by $A \wedge B \wedge C$ and is true if and only if *all* of the propositions A, B, C are true.

There is a similarity between a disjunctive data type and a logical disjunction of propositions. A value of the disjunctive data type `RootsOfQ` can be constructed only if we have one of the values `NoRoots()`, `OneRoot(x)`, or `TwoRoots(x, y)`. Let us now rewrite the previous sentence as a logical formula. Denote by $\mathcal{CH}(A)$ the logical proposition “this Code \mathcal{H} as a value of type A ”, where “this code” refers to a particular function in a program. So, the proposition “the function *can* return a value of type `RootsOfQ`” is denoted by $\mathcal{CH}(\text{RootsOfQ})$. We can then write the above sentence about `RootsOfQ` as the logical formula

$$\mathcal{CH}(\text{RootsOfQ}) = \mathcal{CH}(\text{NoRoots}) \vee \mathcal{CH}(\text{OneRoot}) \vee \mathcal{CH}(\text{TwoRoots}) . \quad (3.1)$$

There is also a similarity between logical *conjunctions* and tuple types. Consider the named tuple (i.e. a case class) `TwoRoots(x: Double, y: Double)`. When can we have a value of type `TwoRoots`? Only if we have two values of type `Double`. Rewriting this sentence as a logical formula, we get

$$\mathcal{CH}(\text{TwoRoots}) = \mathcal{CH}(\text{Double}) \wedge \mathcal{CH}(\text{Double}) .$$

Formal logic admits the simplification

$$\mathcal{CH}(\text{Double}) \wedge \mathcal{CH}(\text{Double}) = \mathcal{CH}(\text{Double}) .$$

However, no such simplification will be available in the general case, e.g.

```
case class Data3(x: Int, y: String, z: Double)
```

For this type, we will have the formula

$$\mathcal{CH}(\text{Data3}) = \mathcal{CH}(\text{Int}) \wedge \mathcal{CH}(\text{String}) \wedge \mathcal{CH}(\text{Double}) . \quad (3.2)$$

We find that tuples are related to logical conjunctions in the same way as disjunctive types are related to logical disjunctions. This is the main reason for choosing the name “disjunctive types”.⁶

The correspondence between disjunctions, conjunctions, and data types is explained in more detail in Chapter 5. For now, we note that the operations of conjunction and disjunction are not sufficient to produce all possible logical expressions. To obtain a complete logic, it is also necessary to have a logical negation $\neg A$ (“ A is not true”) or, equivalently, a logical implication $A \rightarrow B$ (“if A is true then B is true”). It turns out that the logical implication $A \rightarrow B$ is related to the function type `A => B`. In Chapter 4, we will study function types in depth.

⁶These types are also called “variants”, “sum types”, “co-product types”, and “tagged union types”.

Part II

Intermediate level

4 The logic of types. II. Curried functions

4.1 Functions that return functions

4.1.1 Motivation and first examples

Consider the task of preparing a logger function that prints messages with a configurable prefix.

A simple logger function can be a value of type `String => Unit`, such as

```
val logger: String => Unit = { message => println(s"INFO: $message") }

scala> logger("hello world")
INFO: hello world
```

This function prints any given message with the logging prefix "INFO".

The standard library function `println(...)` always returns a `Unit` value after printing its arguments. As we already know, there is only a single value of type `Unit`, and that value is denoted by `()`. To see that `println` returns `Unit`, run this code:

```
scala> val x = println(123)
123
x: Unit = ()
```

The task is to make the logging prefix configurable. A simple solution is to implement a function `logWith` that takes a prefix as an argument and returns a new logger containing that prefix. Note that the function `logWith` returns a new *function*, i.e., a new value of type `String => Unit`:

```
def logWith(prefix: String): (String => Unit) = {
  message => println(s"$prefix: $message")
}
```

The body of `logWith` consists of a nameless function `message => println(...)`, which is a value of type `String => Unit`. This value will be returned when we evaluate `logWith("...")`.

We can now use `logWith` to create some logger functions:

```
scala> val info = logWith("INFO")
info: String => Unit = <function1>

scala> val warn = logWith("WARN")
warn: String => Unit = <function1>
```

The created loggers are then usable as ordinary functions:

```
scala> info("hello")
INFO: hello

scala> warn("goodbye")
WARN: goodbye
```

The values `info` and `warn` can be used by any code that needs a logging function.

It is important that the prefix is "baked into" functions created by `logWith`. A logger such as `warn` will always print messages with the prefix "WARN", and the prefix cannot be changed any more. This is because the value `prefix` is treated as a local constant within the body of the nameless function computed and returned by `logWith`. For instance, the body of the function `warn` is equivalent to

```
{ val prefix = "WARN"; (message => s"$prefix: $message") }
```

Whenever a new function is created using `logWith(prefix)`, the (immutable) value of `prefix` is stored within the body of the newly created function. This is a general feature of nameless functions: the function's body keeps copies of all the outer-scope values it uses. One sometimes says that the function's body “closes over” those values; for this reason, nameless functions are also called “**closures**”. It would be clearer to say that nameless functions “capture” values from outer scopes.

As another example of the capture of values, consider this code:

```
val f: (Int => Int) = {
  val p = 10
  val q = 20
  x => p + q * x
}
```

The body of the function `f` is equivalent to `{ x => 10 + 20 * x }` because the values `p = 10` and `q = 20` were captured.

The parentheses around a function's type, such as `(Int => Int)`, are optional in Scala.

4.1.2 Curried and uncurried functions

Reasoning mathematically about the code

```
val info = logWith("INFO")
info("hello")
```

we would expect that `info` is *the same value* as `logWith("INFO")`, and so the code `info("hello")` should have the same effect as the code `logWith("INFO")("hello")`. This is indeed so:

```
scala> logWith("INFO")("hello")
INFO: hello
```

The syntax `logWith("INFO")("hello")` looks like the function `logWith` applied to *two* arguments. Yet, `logWith` was defined as a function with a single argument of type `String`. This is not a contradiction because `logWith("INFO")` returns a function that accepts an additional argument. So, expressions `logWith("INFO")` and `logWith("INFO")("hello")` are both valid. In this sense, we are allowed to apply `logWith` to one argument at a time.

A function that can be applied to arguments in this way is called a **curried** function.

While a curried function can be applied to one argument at a time, an **uncurried** function must be applied to all arguments at once, e.g.

```
def prefixLog(prefix: String, message: String): Unit = println(s"$prefix: $message")
```

The type of the curried function `logWith` is `String => (String => Unit)`. By Scala's syntax conventions, the function arrow (`=>`) groups to the *right*. So the parentheses in the type expression `String => (String => Unit)` are not needed; the function's type can be written as `String => String => Unit`.

The type `String => String => Unit` is different from `(String => String) => Unit`, which is the type of a function returning `Unit` and having a function of type `String => String` as its argument. When an argument's type is a function type, e.g., `String => String`, it must be enclosed in parentheses.

In general, a curried function takes an argument and returns another function that again takes an argument and returns another function, and so on, until finally a non-function type is returned. So, the type signature of a curried function generally looks like `A => B => C => ... => R => S`, where `A, B, ..., R` are the **curried arguments** and `S` is the “final” result type.

For example, in the type expression `A => B => C => D` the types `A, B, C` are the types of curried arguments, and `D` is the final result type. It takes time to get used to reading this kind of syntax.

In Scala, functions defined with multiple argument groups (enclosed in multiple pairs of parentheses) are curried functions. We have seen examples of curried functions before:

```
def map[A, B](xs: Seq[A])(f: A => B): Seq[B]
def fmap[A, B](f: A => B)(xs: Option[A]): Option[B]
def foldLeft[A, R](xs: Seq[A])(init: R)(update: (R, A) => R): R
```

The type signatures of these functions can be also written equivalently as

```
def map[A, B]: Seq[A] => (A => B) => Seq[B]
def fmap[A, B]: (A => B) => Option[A] => Option[B]
def foldLeft[A, R]: Seq[A] => R => ((R, A) => R) => R
```

Curried arguments of a *function type*, such as $(A \Rightarrow B)$, need parentheses.

To summarize, a curried function such as `logWith` can be defined in three equivalent ways in Scala:

```
1 def logWith1(prefix: String)(message: String): Unit = println(s"$prefix: $message")
2 def logWith2(prefix: String): String => Unit = { message => println(s"$prefix: $message") }
3 def logWith3: String => String => Unit = { prefix => message => println(s"$prefix: $message") }
```

We will sometimes enclose nameless functions in parentheses or curly braces to improve clarity.

Line 3 above shows that the arrow symbols \Rightarrow group to the right within the *code* of nameless functions: $x \Rightarrow y \Rightarrow \text{expr}$ means $\{x \Rightarrow \{y \Rightarrow \text{expr}\}\}$, a nameless function taking an argument x and returning a nameless function that takes an argument y and returns an expression expr . This syntax convention is helpful since the code $x \Rightarrow y \Rightarrow z$ visually corresponds to the curried function's type signature $A \Rightarrow B \Rightarrow C$, which uses the same syntax convention. Also, the syntax $(x \Rightarrow y) \Rightarrow z$ could not possibly work for a nameless function because it would be an argument match with the pattern $x \Rightarrow y$. If we matched a function such as $\{t \Rightarrow t + 20\}$ against the pattern $x \Rightarrow y$ by setting $x = t$ and $y = t + 20$, we would have no value for the bound variable t . (What would be the integer value of y ?) So $x \Rightarrow (y \Rightarrow z)$ is the only sensible way of inserting parentheses into $x \Rightarrow y \Rightarrow z$.

Although the code $(x \Rightarrow y) \Rightarrow z$ is invalid, a type expression $(A \Rightarrow B) \Rightarrow C$ is valid. A nameless function of type $(A \Rightarrow B) \Rightarrow C$ is written as $f \Rightarrow \text{expr}$ where $f: A \Rightarrow B$ is its argument and expr its body.

4.1.3 Equivalence of curried and uncurried functions

We defined the curried function `logWith` in order to be able to create logger functions such as `info` and `warn`. However, some curried functions, such as `foldLeft`, are almost always applied to all possible arguments. A curried function applied to all its possible arguments is equivalent to an uncurried function that takes all those arguments at once. Let us look at this equivalence in more detail.

Consider a curried function with type signature `Int => Int => Int`. This function takes an integer and returns an (uncurried) function taking an integer and returning an integer. An example of such a curried function is

```
def f1(x: Int): Int => Int = { y => x - y }
```

The function takes an integer x and returns the expression $y \Rightarrow x - y$, which is a function of type `Int => Int`. The code of `f1` can be written equivalently as

```
val f1: Int => Int => Int = { x => y => x - y }
```

Let us compare the function `f1` with a function that takes its two arguments at once, e.g.

```
def f2(x: Int, y: Int): Int = x - y
```

The function `f2` has type signature `(Int, Int) => Int`.

The syntax for using the functions `f1` and `f2` is different:

```
scala> f1(20)(4)
res0: Int = 16

scala> f2(20, 4)
res1: Int = 16
```

The main difference between the usage of `f1` and `f2` is that `f2` must be applied at once to both arguments, while `f1` applied to just the first argument, 20. The result of evaluating `f1(20)` is a function that can be later applied to another argument:

```
scala> val r1 = f1(20)
r1: Int => Int = <function1>
```

```
1  scala> r1(4)
2  res2: Int = 16
```

Applying a curried function to some but not all possible arguments is called a **partial application**. Applying a curried function to all possible arguments is called a **saturated application**.

If we need to partially apply an *uncurried* function, we can use the underscore (`_`) symbol:

```
1  scala> val r2: Int => Int = f2(20, _)
2  r2: Int => Int = <function1>
3
4  scala> r2(4)
5  res3: Int = 16
```

(The type annotation `Int => Int` is required in line 1.) This code creates a function `r2` by partially applying `f2` to the first argument but not to the second. Other than that, `r2` is the same function as `r1` defined above; i.e., `r2` returns the same values for the same arguments as `r1`. A more straightforward syntax for a partial application is

```
1  scala> val r3: Int => Int = { x => f2(20, x) }    // Same as r2 above.
2  r3: Int => Int = <function1>
3
4  scala> r3(4)
5  res4: Int = 16
```

We can see that a curried function, such as `f1`, is better adapted for partial application than `f2`, because the syntax is shorter. However, the functions `f1` and `f2` are **computationally equivalent** in the sense that given `f1` we can reconstruct `f2` and vice versa:

```
1  def f2new(x: Int, y: Int): Int = f1(x)(y)           // f2new is equal to f2
2  def f1new: Int => Int => Int = { x => y => f2(x, y) } // f1new is equal to f1
```

It is clear that the function `f1new` computes the same results as `f1`, and that the function `f2new` computes the same results as `f2`. The computational equivalence of the functions `f1` and `f2` is not *equality* — these functions are *different*; but each of them can be reconstructed from the other if necessary.

More generally, a curried function has a type signature of the form `A => B => C => ... => R => S`, where `A, B, C, ..., S` are some types. A function with this type signature is computationally equivalent to an uncurried function with type signature `(A, B, C, ..., R) => S`. The uncurried function takes all arguments at once, while the curried function takes one argument at a time. Other than that, these two functions compute the same results given the same arguments.

We have seen how a curried function can be converted to an equivalent uncurried one, and vice versa. The Scala library defines the methods `curried` and `uncurried` that convert between these forms of functions. To convert between `f2` and `f1`:

```
1  scala> val f1c = (f2 _).curried
2  f1c: Int => (Int => Int) = <function1>
3
4  scala> val f2u = Function.uncurried(f1c)
5  f2u: (Int, Int) => Int = <function2>
```

The syntax `(f2 _)` is needed in Scala to convert methods to function values. Recall that Scala has two ways of defining a function: one as a method (defined using `def`), another as a function value (defined using `val`). The extra underscore will become unnecessary in Scala 3.

The methods `curried` and `uncurried` are easy to implement, as we will show in Section 4.2.1.

4.2 Fully parametric functions

We have seen that some functions are declared with type parameters, which are set only when the function is applied to specific arguments. Examples of such functions are `map` and `filter`, written as

```
1  def map[A, B](xs: Seq[A])(f: A => B): Seq[B]
2  def filter[A](xs: Seq[A])(p: A => Boolean): Seq[A]
```

Such functions can be applied to arguments of different types without changing the function's code. It is clearly better to implement a single function with type parameters instead of writing several functions with repeated code but working with different types.

When we apply the function `map` as `map(xs)(f)` to a specific value `xs` of type, e.g., `Seq[Int]`, and a specific function `f` of type, say, `Int => String`, the Scala compiler will automatically set the type parameters `A = Int` and `B = String` in the code of `map`. We may also set type parameters explicitly and write, for example, `map[Int, String](xs)(f)`. This syntax shows a certain similarity between type parameters such as `Int, String` and “value parameters” (arguments) `xs` and `f`. Setting type parameters, e.g., `map[Int, String]`, means substituting `A = Int, B = String` into the type signature of the function, similarly to how setting value parameters means substituting specific values into the function body.

In the functions `map` and `filter` as just shown, some types are parameters while others are specific types, such as `Seq` and `Boolean`. It is sometimes possible to replace *all* specific types in the type signature of a function by type parameters. The result is a “fully parametric” function.

We call a function **fully parametric** if its arguments have types described by type parameters, and the code of the function treats all types as type parameters. In other words, fully parametric functions do not use any values of specific types, such as `Int` or `String`, in their code. A fully parametric function does not use any information about its argument types, other than assuming that those types correctly match the type signature.

What kind of functions are fully parametric? To build up intuition, let us compare the following two functions that have the same type signature:

```
def cos_sin(p: (Double, Double)): (Double, Double) = p match {
  case (x, y) =>
    val r = math.sqrt(x * x + y * y)
    (x / r, y / r) // Return cos and sin of the angle.
}

def swap(p: (Double, Double)): (Double, Double) = p match {
  case (x, y) => (y, x)
}
```

We can introduce type parameters into the type signature of `swap` to make it fully parametric:

```
def swap[A, B](p: (A, B)): (B, A) = p match {
  case (x, y) => (y, x)
}
```

Converting `swap` into a fully parametric function is possible because the operation of swapping two parts of a tuple `(A, B)` works in the same way for all types `A, B`. No changes were made in

the body of the function. The specialized version of `swap` working on `(Double, Double)` can be obtained from the fully parametric version of `swap` if we set the type parameters as `A = Double, B = Double`.

In contrast, the function `cos_sin` performs a computation that is specific to the type `Double` and cannot be generalized to an arbitrary type parameter `A` instead of `Double`. So, `cos_sin` cannot be generalized to a fully parametric function.

Typically, a fully parametric function has all its arguments typed with type parameters or with some combinations of type parameters, i.e., type expressions such as `(A, B)` or `X => Either[X, Y]`.

The `swap` operation for pairs is already defined in the Scala library:

```
scala> (1, "abc").swap
res0: (String, Int) = (abc,1)
```

If needed, other swapping functions can be implemented for tuples with more elements, e.g.

```
def swap12[A,B,C]: ((A, B, C)) => (B, A, C) = { case (x, y, z) => (y, x, z) }
```

The Scala syntax requires the double parentheses around tuple types of arguments but not around the tuple type of a function’s result. So, the function `cos_sin` may be written as a value of type

```
val cos_sin: ((Double, Double)) => (Double, Double) = ...
```

4.2.1 Examples. Function composition

Further examples of fully parametric functions are the identity function, the constant function, the function composition methods, and the `curry` / `uncurry` conversions.

The identity function (available in the Scala library as `identity[T]`) is

```
def id[T]: T => T = (t => t)
```

The constant function (available in the Scala library as `Function.const[c, x]`) takes an argument `c:c` and returns a new function that always returns `c`:

```
def const[C, X](c: C): X => C = (_ => c)
```

The syntax `_ => c` is used to emphasize that the function ignores its argument.

Function composition Consider two functions `f: Int => Double` and `g: Double => String`. We can apply `f` to an integer argument `x:Int` and get a result `f(x)` of type `Double`. Applying `g` to that result gives a `String` value `g(f(x))`. The transformation from the original integer `x:Int` to the final `String` value `g(f(x))` can be viewed as a new function of type `Int => String`. That new function is called the **forward composition** of the two functions `f` and `g`. In Scala, this operation is written as `f andThen g`:

```
val f: Int => Double = (x => 5.67 + x)
val g: Double => String = (x => f"Result x = $x%3.2f")

scala> val h = f andThen g      // h(x) is defined as g(f(x)).
h: Int => String = <function1>

scala> h(40)
res36: String = Result x = 45.67
```

The Scala compiler derives the type of `h` automatically as `Int => String`.

The forward composition is denoted by `;` (pronounced “before”) and can be defined as

$$f ; g \triangleq x \rightarrow g(f(x)) . \quad (4.1)$$

The symbol \triangleq means “is defined as”.

We could implement the operation of forward composition as a fully parametric function,

```
def andThen[X, Y, Z](f: X => Y)(g: Y => Z): X => Z = { x => g(f(x)) }
```

This type signature requires the types of the function arguments to match in a certain way, or else the composition is undefined (and the code would produce a type error).

The method `andThen` is an example of a function that *both* returns a new function *and* takes other functions as arguments.

The **backward composition** of two functions `f` and `g` works in the opposite order: first `g` is applied and then `f`. Using the symbol `o` (pronounced “after”) for this operation, we can write

$$f \circ g \triangleq x \rightarrow f(g(x)) . \quad (4.2)$$

In Scala, the backward composition is called `compose` and used as `f compose g`. This method may be implemented as a fully parametric function

```
def compose[X, Y, Z](f: Y => X)(g: Z => Y): Z => X = { z => f(g(z)) }
```

We have already seen the methods `curried` and `uncurried` defined by the Scala library. As an illustration, let us write our own code for converting curried functions to uncurried:

```
def uncurry[A, B, R](f: A => B => R): ((A, B)) => R = { case (a, b) => f(a)(b) }
```

We conclude from these examples that fully parametric functions perform operations so general that they work in the same way for all types. Some arguments of fully parametric functions may have complicated types such as `A => B => R`, which are type expressions built up from type parameters. But fully parametric functions do not use values of specific types such as `Int` or `String`.

Functions with type parameters are sometimes called “generic”. This book uses the term “**fully parametric**” to designate a certain restricted kind of generic functions.

4.2.2 Laws of function composition

The operations of function composition, introduced in Section 4.2.1, have three important properties or “laws” that follow directly from the definitions. These laws are:

- The two **identity laws**: the composition of any function f with the identity function will give again the function f .
- The **associativity law**: the consecutive composition of three functions f, g, h does not depend on the order in which the pairs are composed.

These laws hold equally for the forward and the backward composition, since those are just syntactic variants of the same operation. Let us write these laws rigorously as equations and prove them.

Proofs with forward composition The composition of the identity function with an arbitrary function f can be $\text{id} \circ f$ with the identity function to the left of f , or $f \circ \text{id}$ with the identity function to the right of f . In both cases, the result must be equal to the function f . The resulting two laws are

$$\begin{aligned} \text{left identity law of composition : } \text{id} \circ f &= f \quad , \\ \text{right identity law of composition : } f \circ \text{id} &= f \quad . \end{aligned}$$

To show that these laws always hold, we need to show that both sides of the laws, which are functions, give the same result when applied to an arbitrary value x . Let us first clarify how the type parameters must be set for the laws to have consistently matching types.

The laws must hold for an arbitrary function f . So we may assume that f has the type signature $A \rightarrow B$, where A and B are arbitrary type parameters. Consider the left identity law. The function $(\text{id} \circ f)$ is, by definition (4.1), a function that takes an argument x , applies id to that x , and then applies f to the result:

$$\text{id} \circ f = (x \rightarrow f(\text{id}(x))) \quad .$$

If f has type $A \rightarrow B$, its argument must be of type A , or else the types will not match. Therefore, the identity function must have type $A \rightarrow A$, and the argument x must have type A . With these choices of the type parameters, the function $(x \rightarrow f(\text{id}(x)))$ will have type $A \rightarrow B$, as it must since the right-hand side of the law is f . We add type annotations to the code as *superscripts*,

$$\text{id}^{A \rightarrow A} \circ f^{A \rightarrow B} = (x^A \rightarrow f(\text{id}(x)))^{A \rightarrow B} \quad .$$

In the Scala syntax, this formula may be written as

```
id[A] andThen (f: A => B) == { x: A => f(id(x)) }: A => B
```

We will follow the convention where type parameters are single uppercase letters, as is common in Scala code (although this convention is not enforced by the Scala compiler). The colon symbol ($:$) in the superscript x^A means a type annotation, as in Scala code $x:A$. Superscripts without a colon, such as id^A , denote type parameters, as in Scala code $\text{identity}[A]$. Since the function $\text{identity}[A]$ has type $A \Rightarrow A$, we can write id^A or equivalently (but more verbosely) $\text{id}^{A \rightarrow A}$ to denote that function.

By definition of the identity function, we have $\text{id}(x) = x$, and so

$$\text{id} \circ f = (x \rightarrow f(\text{id}(x))) = (x \rightarrow f(x)) = f \quad .$$

The last step works since $x \rightarrow f(x)$ is a function taking an argument x and applying f to that argument; i.e., $x \rightarrow f(x)$ is an **expanded form** of the same function f .

Now consider the right identity law, $f \circ \text{id} = f$. Write out the left-hand side:

$$f \circ \text{id} = (x \rightarrow \text{id}(f(x))) \quad .$$

To check that the types match, assume that $f:A \rightarrow B$. Then x must have type A , and the identity function must have type $B \rightarrow B$. The result of $\text{id}(f(x))$ will also have type B . With these choices of type parameters, all types match:

$$f:A \rightarrow B ; \text{id}:B \rightarrow B = (x:A \rightarrow \text{id}(f(x))):A \rightarrow B .$$

Since $\text{id}(f(x)) = f(x)$, we find that

$$f ; \text{id} = (x \rightarrow f(x)) = f .$$

In this way, we have demonstrated that both identity laws hold.

The associativity law is written as an equation like this:

$$\text{associativity law of function composition : } (f ; g) ; h = f ; (g ; h) . \quad (4.3)$$

Let us verify that the types match here. The types of the functions f , g , and h must be such that all the function compositions match. If f has type $A \rightarrow B$ for some type parameters A and B , then the argument of g must be of type B ; so we must have $g:B \rightarrow C$, where C is another type parameter. The composition $f ; g$ has type $A \rightarrow C$, so h must have type $C \rightarrow D$ for some type D . Assuming the types as $f:A \rightarrow B$, $g:B \rightarrow C$, and $h:C \rightarrow D$, we find that the types in all the compositions $f ; g$, $g ; h$, $(f ; g) ; h$, and $f ; (g ; h)$ match. We can rewrite Eq. (4.3) with type annotations,

$$(f:A \rightarrow B ; g:B \rightarrow C) ; h:C \rightarrow D = f:A \rightarrow B ; (g:B \rightarrow C ; h:C \rightarrow D) . \quad (4.4)$$

After checking the types, we are ready to verify the associativity law. Note that both sides of the law (4.4) are functions of type $A \rightarrow D$. To prove that two functions are equal means to prove that they always return the same results when applied to the same arguments. So, let us apply both sides of Eq. (4.4) to an arbitrary value $x:A$. Using the definition (4.1) of the forward composition, we find:

$$\begin{aligned} ((f ; g) ; h)(x) &= h((f ; g)(x)) = h(g(f(x))) , \\ (f ; (g ; h))(x) &= (g ; h)(f(x)) = h(g(f(x))) . \end{aligned}$$

Both sides of the law are now equal when applied to an arbitrary value x .

Because of the associativity law, we do not need parentheses when writing the expression $f ; g ; h$. The function $(f ; g) ; h$ is equal to the function $f ; (g ; h)$.

In the proof, we have omitted the type annotations since we already checked that all types match. Checking the types beforehand allows us to write shorter derivations.

Proofs with backward composition This book uses the **forward composition** $f ; g$ rather than the backward composition $g \circ f$. If necessary, all equations can be converted from one notation to the other by reversing the order of compositions,

$$f ; g = g \circ f$$

for any functions $f:A \rightarrow B$ and $g:B \rightarrow C$. Let us see how to prove the composition laws in the backward notation. We will just need to reverse the order of function compositions in the proofs above.

The left identity and right identity laws are

$$f \circ \text{id} = f , \quad \text{id} \circ f = f .$$

To match the types, we need to choose the type parameters as

$$f:A \rightarrow B ; \text{id}:A \rightarrow A = f:A \rightarrow B , \quad \text{id}:B \rightarrow B ; f:A \rightarrow B = f:A \rightarrow B .$$

We now apply both sides of the laws to an arbitrary value $x:A$. For the left identity law, we find

$$\text{use definition (4.2) : } f \circ \text{id} = (x \rightarrow f(\text{id}(x))) = (x \rightarrow f(x)) = f .$$

Similarly for the right identity law,

$$\text{id} \circ f = (x \rightarrow \text{id}(f(x))) = (x \rightarrow f(x)) = f \quad .$$

The associativity law,

$$h \circ (g \circ f) = (h \circ g) \circ f \quad ,$$

is proved by applying both sides to an arbitrary value x of a suitable type:

$$\begin{aligned} (h \circ (g \circ f))(x) &= h((g \circ f)(x)) = h(g(f(x))) \quad , \\ ((h \circ g) \circ f)(x) &= (h \circ g)(f(x)) = h(g(f(x))) \quad . \end{aligned}$$

The types are checked by assuming that f has the type $f:A \rightarrow B$. The types in $g \circ f$ match only when $g:B \rightarrow C$, and then $g \circ f$ is of type $A \rightarrow C$. The type of h must be $h:C \rightarrow D$ for the types in $h \circ (g \circ f)$ to match. We can write the associativity law with type annotations as

$$h:C \rightarrow D \circ (g:B \rightarrow C \circ f:A \rightarrow B) = (h:C \rightarrow D \circ g:B \rightarrow C) \circ f:A \rightarrow B \quad . \quad (4.5)$$

The associativity law allows us to omit parentheses in the expression $h \circ g \circ f$.

The length of calculations is the same in the forward and the backward notation. One difference is that types of function compositions are more visually clear in the forward notation: it is harder to check that types match in Eq. (4.5) than in Eq. (4.4). To make the backward composition easier to work with, one could write¹ the function types in reverse as, e.g., $g:C \leftarrow B \circ f:B \leftarrow A$.

4.2.3 Example: A function that is *not* fully parametric

Fully parametric functions should not make any decisions based on the actual types of arguments. As an example of code that is not fully parametric, consider the following “fake identity” function:

```
def fakeId[A]: A => A = {
  case x: Int => (x - 1).asInstanceOf[A]      // Special code for A = Int.
  case x        => x                          // Common code for all other types A.
}
```

This function’s type signature is the same as that of `id[A]`, and its behavior is the same for all types `A` except for `A = Int`:

```
scala> fakeId("abc")
res0: String = abc

scala> fakeId(true)
res1: Boolean = true

scala> fakeId(0)
res2: Int = -1
```

While Scala allows us to write this kind of code, the resulting function does not appear to be useful. In any case, `fakeId` is not a fully parametric function.

The identity laws of composition will not hold if we use `fakeId[A]` instead of the correct function `id[A]`. For example, consider the composition of `fakeId` with a simple function `f_1` defined by

```
def f_1: Int => Int = { x => x + 1 }
```

The composition `(f_1 andThen fakeId)` has type `Int => Int`. Since `f_1` has type `Int => Int`, Scala will automatically set the type parameter `A = Int` in `fakeId[A]`,

```
scala> def f_2 = f_1 andThen fakeId
f_2: Int => Int
```

The identity law says that $f_2 = f_1 \circ \text{id} = f_1$. But we can check that `f_1` and `f_2` are not the same:

¹As in the book “Program design by calculation”, see <http://www4.di.uminho.pt/~jno/ps/pdbc.pdf>

```
scala> f_1(0)
res3: Int = 1

scala> f_2(0)
res4: Int = 0
```

It is important that we are able to detect that `fakeId` is not a fully parametric function by checking whether some equation holds, without need to examine the code of `fakeId`. In this book, we will always formulate the desired properties through equations or “laws”. To verify that a law holds, we will perform symbolic calculations similar to the proofs in Section 4.2.2. These calculations are **symbolic** in the sense that we are manipulating symbols (such as x, f, g, h) without substituting any specific values for these symbols but using only the general properties of functions. This is similar to symbolic calculations in polynomial algebra, such as $(x - y)(x + y) = x^2 - y^2$. In the next section, we will get more experience with symbolic calculations relevant to functional programming.

4.3 Symbolic calculations with nameless functions

4.3.1 Calculations with curried functions

In mathematics, functions are evaluated by substituting their argument values into their body. Nameless functions are evaluated in the same way. For example, applying the nameless function $x \rightarrow x + 10$ to an integer 2, we substitute 2 instead of x in “ $x + 10$ ” and get “2 + 10”, which we then evaluate to 12. The computation is written like this,

$$(x \rightarrow x + 10)(2) = 2 + 10 = 12 .$$

To run this computation in Scala, we need to add a type annotation to the nameless function as in $(x:\text{Int} \rightarrow x + 10)(2)$:

```
scala> ((x: Int) => x + 10)(2)
res0: Int = 12
```

Curried function calls such as $f(x)(y)$ or $(x \rightarrow \text{expr}(x))(y)(z)$ may look unfamiliar and confusing; we need to gain some experience working with them.

Consider a curried nameless function being applied to arguments, such as $(x \rightarrow y \rightarrow x - y)(20)(4)$, and compute the result of this function application. Begin with the argument 20; applying a nameless function of the form $(x \rightarrow \dots)$ to 20 means substituting $x = 20$ into the body of the function. After that substitution, we obtain the expression $y \rightarrow 20 - y$, which is again a nameless function. Applying that function to the remaining argument 4 means substituting $y = 4$ into the body of $y \rightarrow 20 - y$. We get the expression $20 - 4$, which equals 16. Check the result with Scala:

```
scala> ((x: Int) => (y: Int) => x - y)(20)(4)
res1: Int = 16
```

Applying a curried function such as $x \rightarrow y \rightarrow z \rightarrow \text{expr}(x, y, z)$ to three curried arguments 10, 20, and 30 means substituting $x = 10$, $y = 20$, and $z = 30$ into the expression $\text{expr}(x, y, z)$. In this way, we can easily apply a curried function to any number of curried arguments.

This calculation is helped by the convention that $f(g)(h)$ means first applying f to g and then applying the result to h . In other words, function application groups to the *left*: $f(g)(h) = (f(g))(h)$. It would be confusing if function application grouped to the right and $f(g)(h)$ meant first applying g to h and then applying f to the result. If *that* were the syntax convention, it would be harder to reason about applying a curried function to its arguments.

We see that the right grouping of the function arrow \Rightarrow is well adapted to the left grouping of function applications. All functional languages follow these syntactic conventions.

To make calculations shorter, we will write code in a mathematical notation rather than in the Scala syntax. Type annotations are written with a colon in the superscript, for example: $x:\text{Int} \rightarrow x + 10$ instead of the code $((x:\text{Int}) \Rightarrow x + 10)$.

The symbolic evaluation of the Scala code $((x:\text{Int}) \Rightarrow (y:\text{Int}) \Rightarrow x - y)(20)(4)$ can be written as

$$(x:\text{Int} \rightarrow y:\text{Int} \rightarrow x - y)(20)(4)$$

apply function and substitute $x = 20$: $= (y:\text{Int} \rightarrow 20 - y)(4)$

apply function and substitute $y = 4$: $= 20 - 4 = 16 .$

In the above step-by-step calculation, the colored underlines and comments at left are added for clarity. A colored underline indicates a sub-expression that is going to be rewritten at the *next* step.

Here we performed calculations by substituting an argument into a function at each step. A compiled Scala program is evaluated in a similar way at run time.

Nameless functions are *values* and can be used as part of larger expressions, just as any other values. For instance, nameless functions can be arguments of other functions (nameless or not). Here is an example of applying a nameless function $f \rightarrow f(9)$ to a nameless function $x \rightarrow x\%4$:

$$(f \rightarrow f(9))(x \rightarrow x\%4)$$

substitute $f = (x \rightarrow x\%4)$: $= (x \rightarrow x\%4)(9)$

substitute $x = 9$: $= 9\%4 = 1$.

In the nameless function $f \rightarrow f(9)$, the argument f has to be itself a function, otherwise the expression $f(9)$ would make no sense. The argument x of $f(x)$ must be an integer, or else we would not be able to compute $x\%4$. The result of computing $f(9)$ is 1, an integer. We conclude that f must have type $\text{Int} \rightarrow \text{Int}$, or else the types will not match. To verify this result in Scala, we need to specify a type annotation for f :

```
scala> ((f: Int => Int) => f(9))(x => x % 4)
res2: Int = 1
```

No type annotation is needed for $x \rightarrow x\%4$ since the Scala compiler already knows the type of f and figures out that x in $x \rightarrow x\%4$ must have type Int .

To summarize the syntax conventions for curried nameless functions:

- Function expressions group everything to the right: $x \rightarrow y \rightarrow z \rightarrow e$ means $x \rightarrow (y \rightarrow (z \rightarrow e))$.
- Function calls group everything to the left. So, $f(x)(y)(z)$ means $((f(x))(y))(z)$, i.e., $f(x)$ gives a new function that is applied to y , giving again a new function that is finally applied to z .
- Function applications group stronger than infix operations, so $f(x) + y$ means $(f(x)) + y$, as usual in mathematics, and not $f(x + y)$.

Here are some more examples of performing function applications symbolically. Types are omitted for brevity; every non-function value is of type Int .

$$(x \rightarrow x * 2)(10) = 10 * 2 = 20 .$$

$$(p \rightarrow z \rightarrow z * p)(t) = (z \rightarrow z * t) .$$

$$(p \rightarrow z \rightarrow z * p)(t)(4) = (z \rightarrow z * t)(4) = 4 * t .$$

Some results of these computation are integer values such as 20; other results are nameless functions such as $z \rightarrow z * t$. Verify this in Scala:

```
scala> ((x:Int) => x*2)(10)
res3: Int = 20

scala> ((p:Int) => (z:Int) => z*p)(10)
res4: Int => Int = <function1>

scala> ((p:Int) => (z:Int) => z*p)(10)(4)
res5: Int = 40
```

In the following examples, some arguments are themselves functions. Consider an expression that uses the nameless function $(g \rightarrow g(2))$ as an argument:

$$(f \rightarrow p \rightarrow f(p))(g \rightarrow g(2)) \tag{4.6}$$

substitute $f = (g \rightarrow g(2))$: $= p \rightarrow (g \rightarrow g(2))(p)$

substitute $g = p$: $= p \rightarrow p(2) .$ (4.7)

The result of this expression is a function $p \rightarrow p(2)$ that will apply *its* argument p to the value 2. A possible value for p is the function $x \rightarrow x + 4$. So, let us apply the expression in Eq. (4.6) to $p \rightarrow p(2)$:

$$\begin{aligned}
 & (f \rightarrow p \rightarrow f(p))(g \rightarrow g(2))(x \rightarrow x + 4) \\
 \text{use Eq. (4.7)} : & = (p \rightarrow p(2))(x \rightarrow x + 4) \\
 \text{substitute } p = (x \rightarrow x + 4) : & = (x \rightarrow x + 4)(2) \\
 \text{substitute } x = 2 : & = 2 + 4 = 6 \quad .
 \end{aligned}$$

To verify this calculation in Scala, we need to add appropriate type annotations for f and p . To figure out the types, we reason like this:

We know that the function $f \rightarrow p \rightarrow f(p)$ is being applied to the arguments $f = (g \rightarrow g(2))$ and $p = (x \rightarrow x + 4)$. So, the argument f in $f \rightarrow p \rightarrow f(p)$ must be a function that takes p as an argument.

The variable x in $x \rightarrow x + 4$ must be of type `Int`. Thus, the type of the expression $x \rightarrow x + 4$ is `Int → Int`, and so must be the type of the argument p . We write $p : \text{Int} \rightarrow \text{Int}$.

Finally, we need to make sure that the types match in the function $f \rightarrow p \rightarrow f(p)$. Types match in $f(p)$ if the type of f 's argument is the same as the type of p , which is `Int → Int`. So f 's type must be $(\text{Int} \rightarrow \text{Int}) \rightarrow A$ for some type A . Since in our example $f = (g \rightarrow g(2))$, types match only if g has type `Int → Int`. But then $g(2)$ has type `Int`, and so we must have $A = \text{Int}$. Thus, the type of f is $(\text{Int} \rightarrow \text{Int}) \rightarrow \text{Int}$. We know enough to write the Scala code now:

```
scala> ((f: (Int => Int) => Int) => p => f(p))(g => g(2))(x => x + 4)
res6: Int = 6
```

Type annotations for p , g , and x may be omitted because the Scala compiler will

figure out the missing types from the given type of f . However, it is never an error to specify more type annotations when writing code; it just makes the code longer.

4.3.2 Solved examples: Deriving a function's type from its code

Checking that the types match is an important part of the functional programming paradigm, both in the practice of writing code and in theoretical derivations of laws for various functions. For instance, in the derivations of the composition laws (Section 4.2.2), we were able to deduce the possible type parameters for f , g , and h in the expression $f \circ g \circ h$. This worked because the composition operation `andThen` (denoted by the symbol `;`) is fully parametric. Given a fully parametric function, one can derive the most general type signature that matches the body of that function. The same type-deriving procedure may also help in converting a given function to a fully parametric form.

Let us look at some examples of doing this.

Example 4.3.2.1 The functions `const` and `id` were defined in Section 4.2.1. What is the value `const(id)` and what is its type? Determine the most general type parameters in the expression `const(id)`.

Solution We need to treat the functions `const` and `id` as values, since our goal is to apply `const` to `id`. Write the code of these functions in a short notation:

$$\begin{aligned}
 \text{const}^{C,X} &\triangleq c : C \rightarrow _ : X \rightarrow c \quad , \\
 \text{id}^A &\triangleq a : A \rightarrow a \quad .
 \end{aligned}$$

The types will match in the expression `const(id)` only if the argument of the function `const` has the same type as the type of `id`. Since `const` is a curried function, we need to look at its *first* curried argument, which is of type C . The type of `id` is $A \rightarrow A$, where A is so far an arbitrary type. So, the type parameter C in $\text{const}^{C,X}$ must be equal to $A \rightarrow A$:

$$C = A \rightarrow A \quad .$$

The type parameter X in $\text{const}^{C,X}$ is not constrained, so we keep it as X . The result of applying `const` to `id` is of type $X \rightarrow C$, which equals $X \rightarrow A \rightarrow A$. In this way, we find

$$\text{const}^{A \rightarrow A, X}(\text{id}^A) : X \rightarrow A \rightarrow A \quad .$$

The types A and X can be arbitrary. The type $X \rightarrow A \rightarrow A$ is the most general type for the expression `const(id)` because we have not made any assumptions about the types except requiring that all functions must be always applied to arguments of the correct types.

To compute the value of `const(id)`, it remains to substitute the code of `const` and `id`. Since we already checked the types, we may omit all type annotations:

$$\begin{aligned}
 \text{const(id)} & \\
 \text{definition of const :} &= (c \rightarrow x \rightarrow c)(id) \\
 \text{apply function, substitute } c = id : &= x \rightarrow id \\
 \text{definition of id :} &= x \rightarrow a \rightarrow a \quad .
 \end{aligned}$$

The function $(x \rightarrow a \rightarrow a)$ takes an argument $x:X$ and returns the identity function $a:A \rightarrow a$. It is clear that the argument x is ignored by this function, so we can rewrite it equivalently as

$$\text{const(id)} = _ : X \rightarrow a : A \rightarrow a \quad .$$

Example 4.3.2.2 Implement a function `twice` that takes a function $f: \text{Int} \Rightarrow \text{Int}$ as its argument and returns a function that applies f twice. For instance, if the function f is `{ x => x + 3 }`, the result of `twice(f)` should be equal to the function `x => x + 6`. Test this with the expression `twice(x => x + 3)(10)`. After implementing the function `twice`, generalize it to a fully parametric function.

Solution According to the requirements, the function `twice` must return a new function of type `Int => Int`. So the type signature of `twice` is

```
def twice(f: Int => Int): Int => Int = ???
```

Since `twice(f)` must be a new function with an integer argument, we begin the code of `twice` by writing a new nameless function `{ (x: Int) => ... }`,

```
def twice(f: Int => Int): Int => Int = { (x: Int) => ??? }
```

The new function must apply f twice to its argument, that is, it must return $f(f(x))$. We can finish the implementation now:

```
def twice(f: Int => Int): Int => Int = { x => f(f(x)) }
```

The type annotation `(x: Int)` can be omitted. Let us verify that `twice(x => x+3)(10)` equals $10 + 6$:

```
scala> val g = twice(x => x + 3) // Expect g to be equal to the function { x => x + 6 }.
g: Int => Int = <function1>

scala> g(10) // Expect twice(x => x + 3)(10) to be equal to (x => x+6)(10) = 16.
res0: Int = 16
```

To transform `twice` into a fully parametric function means replacing its type signature by a fully parameterized type signature while keeping the function body unchanged,

```
def twice[A, B, ...](f: ...): ... = { x => f(f(x)) }
```

To determine the type signature and the possible type parameters A, B, \dots , we need to determine the most general type that matches the function body. The function body is the expression $x \rightarrow f(f(x))$. Assume that x has type A ; for types to match in the sub-expression $f(x)$, we need f to have type $A \rightarrow B$ for some type B . The sub-expression $f(x)$ will then have type B . For types to match in $f(f(x))$, the argument of f must have type B ; but we already assumed $f:A \rightarrow B$. This is consistent only if $A = B$. In this way, $x:A$ implies $f:A \rightarrow A$, and the expression $x \rightarrow f(f(x))$ has type $A \rightarrow A$. We can now write the type signature of `twice`,

```
def twice[A](f: A => A): A => A = { x => f(f(x)) }
```

This fully parametric function has only one independent type parameter, A , and can be equivalently written in the code notation as

$$\text{twice}^A \triangleq f:A \rightarrow A \rightarrow x:A \rightarrow f(f(x)) = f:A \rightarrow A \rightarrow f \circ f \quad . \quad (4.8)$$

The procedure of deriving the most general type for a given code is called **type inference**. In Example 4.3.2.2, the presence of the type parameter A and the type signature $(A \rightarrow A) \rightarrow A \rightarrow A$ have been “inferred” from the code $f \rightarrow x \rightarrow f(f(x))$.

Example 4.3.2.3 Consider the fully parametric function `twice` defined in Example 4.3.2.2. What is the most general type of `twice(twice)`, and what computation does it perform? Test your answer on the expression `twice(twice)(x => x+3)(10)`. What are the type parameters in that expression?

Solution Note that `twice(twice)` means that the function `twice` is used as *its own* argument, i.e., this is `twice(f)` with `f = twice`. We begin by assuming unknown type parameters as `twice[A]` (`twice[B]`). The function `twice[A]` of type $(A \rightarrow A) \rightarrow A \rightarrow A$ can be applied to the argument `twice[B]` only if `twice[B]` has type $A \rightarrow A$. But `twice[B]` is of type $(B \rightarrow B) \rightarrow B \rightarrow B$. The symbol \rightarrow groups to the right, so we have

$$(B \rightarrow B) \rightarrow B \rightarrow B = (B \rightarrow B) \rightarrow (B \rightarrow B) \quad .$$

This can match with $A \rightarrow A$ only if we set $A = (B \rightarrow B)$. So the most general type of `twice(twice)` is

$$\text{twice}^{B \rightarrow B}(\text{twice}^B) : (B \rightarrow B) \rightarrow B \rightarrow B \quad . \quad (4.9)$$

After checking that types match, we may omit types from further calculations.

Example 4.3.2.2 defined `twice` with the `def` syntax. To use `twice` as an argument in the expression `twice(twice)`, it is convenient to define `twice` as a value, `val twice = ...`. However, the function `twice` needs type parameters, and Scala 2 does not directly support `val` definitions with type parameters. Scala 3 will support type parameters appearing together with arguments in a nameless function:

```
val twice = [A] => (f: A => A) => (x: A) => f(f(x)) // Valid only in Scala 3.
```

Keeping this in mind, we use the definition of `twice` from Eq. (4.8): $\text{twice}(f) = f \circ f$, which omits the curried argument $x:A$ and makes the calculation shorter. Substituting that into `twice(twice)`, we find

$$\begin{aligned} \text{twice}(\text{twice}) &= \text{twice} \circ \text{twice} \\ \text{expand function composition:} &= f \rightarrow \text{twice}(\text{twice}(f)) \quad . \\ \text{definition of } \text{twice}(f) &= f \rightarrow \text{twice}(f \circ f) \\ \text{definition of } \text{twice} &= f \rightarrow f \circ f \circ f \circ f \quad . \end{aligned}$$

This clearly shows that `twice(twice)` is a function applying its (function-typed) argument *four* times.

The types in `twice(twice)(x => x+3)` follow from Eq. (4.9): since `x => x+3` has type `Int => Int`, types will match only if we set $B = \text{Int}$. The result is `twice[Int => Int](twice[Int])`. To test, we need to write at least one type parameter in the code, or else Scala cannot correctly infer the types in `twice(twice)`:

```
scala> twice(twice[Int])(x => x + 3)(10) // Or write 'twice[Int => Int](twice)(x => x + 3)(10)' .
res0: Int = 22
```

This confirms that `twice(twice)(x => x+3)` equals the function `x => x + 12`, unlike `twice(twice(x => x+3))`, which equals the function `x => x + 6` as shown in Example 4.3.2.2.

Example 4.3.2.4 (a) For the given function `p`, infer a general type signature with type parameter(s):

```
def p[...]: ... = { f => f(2) }
```

(b) Could we choose the type parameters in the expression `p(p)` such that the types match?

Solution (a) In the nameless function $f \rightarrow f(2)$, the argument f must be itself a function with an argument of type `Int`, otherwise the sub-expression `f(2)` is ill-typed. So, types will match if f has type $\text{Int} \rightarrow \text{Int}$ or $\text{Int} \rightarrow \text{String}$ or similar. The most general case is when f has type $\text{Int} \rightarrow A$, where A is an arbitrary type (i.e., a type parameter); then the value $f(2)$ has type A . Since the nameless function $f \rightarrow f(2)$ has an argument f of type $\text{Int} \rightarrow A$ and a result $f(2)$ of type A , we find that the type of p must be $(\text{Int} \rightarrow A) \rightarrow A$. With this type assignment, all types match. The type parameter A remains undetermined and is added to the type signature of the function `p`. The code is

```
def p[A]: (\text{Int} \rightarrow A) \rightarrow A = { f => f(2) }
```

Notation	Scala syntax	Comments
$x:A$	<code>x: A</code>	a value or an argument of type A
$f:A \rightarrow B$	<code>f: A => B</code>	a function of type A \rightarrow B
$x:\text{Int} \rightarrow f(x)$	<code>{ x: Int => f(x) }</code>	a nameless function with argument x
$f^{A,B} \triangleq \dots$	<code>def f[A, B] = ...</code>	define a function with type parameters
id^A , also $\text{id}^{A \rightarrow A}$	<code>identity[A]</code>	the standard “identity” function
$A \rightarrow B \rightarrow C$	<code>A => B => C</code>	type of a curried function
$f \circ g$	<code>f andThen g</code>	forward composition of functions
$g \circ f$	<code>g compose f</code>	backward composition of functions

Table 4.1: Some notation for symbolic reasoning about code.

(b) The expression $p(p)$ applies p to itself, just as `twice(twice)` did in Example 4.3.2.3. Begin by writing $p(p)$ with unknown type parameters: $p[A](p[B])$. Then try to choose A and B so that the types match in that expression. Does the type of $p[B]$, which is $(\text{Int} \rightarrow \text{B}) \rightarrow \text{B}$, match the type of the argument of $p[A]$, which is $\text{Int} \rightarrow \text{A}$, with some choice of A and B? A function type $P \rightarrow Q$ matches $X \rightarrow Y$ only if $P = X$ and $Q = Y$. So $(\text{Int} \rightarrow \text{B}) \rightarrow \text{B}$ can match $\text{Int} \rightarrow \text{A}$ only if $\text{Int} \rightarrow \text{B}$ matches Int and if $B = A$. But it is impossible for $\text{Int} \rightarrow \text{B}$ to match Int , no matter how we choose B.

We conclude that the expression $p[A](p[B])$ has a problem: for any choice of A and B, some type will be mismatched. One says that the expression $p(p)$ is **not well-typed** and does not **typecheck**. Such expressions contain a type error and are rejected by the Scala compiler.

We have seen that, for a code fragment containing function expressions, one can infer the most general type that makes all functions match their arguments, unless the code does not typecheck. The Damas-Hindley-Milner algorithm² performs type inference (or determines that there is a type error) for a large class of expressions containing functions, tuples, and disjunctive types.

4.4 Summary

Table 4.1 shows the notations introduced in this chapter.

What can we do with this chapter’s techniques?

- Implement functions that return new functions and/or take functions as arguments.
- Simplify function expressions symbolically, when functions are applied to arguments.
- Infer the most general type for a given code expression (perform type inference).
- Convert functions to a fully parametric form when possible.

The following solved examples and exercises illustrate these techniques.

4.4.1 Solved examples

Example 4.4.1.1 Implement a function that applies a given function f repeatedly to an initial value x_0 , until a given function `cond` returns `true`:

```
def converge[X](f: X => X, x0: X, cond: X => Boolean): X = ???
```

²https://en.wikipedia.org/wiki/Hindley-Milner_type_system#Algorithm_W

Solution We create an iterator that keeps applying the function f , and use `find` to stop the sequence when the condition first holds:

```
def converge[X](f: X => X, x0: X, cond: X => Boolean): X =
  Stream.iterate(x0)(f)    // Type is Stream[X].
  .find(cond)              // Type is Option[X].
  .get                      // Type is X.
```

The method `get` is a partial function that can be applied only to non-empty `Option` values. It is safe to call `get` here, because the stream is unbounded and, if the condition `cond` never becomes `true`, the program will run out of memory (since `Stream.iterate` keeps all computed values in memory) or the user will run out of patience. So `.find(cond)` can never return an empty `Option` value. Of course, it is not satisfactory that the program crashes when the sequence does not converge. Exercise 4.4.2.2 will implement a safer version of this function by limiting the allowed number of iterations.

A tail-recursive implementation that works in constant memory is

```
@tailrec def converge[X](f: X => X, x0: X, cond: X => Boolean): X =
  if (cond(x0)) x0 else converge(f, f(x0), cond)
```

To test this code, compute an approximation to \sqrt{q} by Newton's method. The iteration function f is

$$f(x) = \frac{1}{2} \left(x + \frac{q}{x} \right) .$$

We iterate $f(x)$ starting with $x_0 = q/2$ until we obtain a given precision:

```
def approx_sqrt(q: Double, precision: Double): Double = {
  def cond(x: Double): Boolean = math.abs(x * x - q) <= precision
  def iterate_sqrt(x: Double): Double = 0.5 * (x + q / x)
  converge(iterate_sqrt, q / 2, cond)
}
```

Newton's method for \sqrt{q} is guaranteed to converge when $q \geq 0$. Test it:

```
scala> approx_sqrt(25, 1.0e-8)
res0: Double = 5.000000000016778
```

Example 4.4.1.2 Using both `def` and `val`, define a Scala function that takes an integer x and returns a function that adds x to its argument.

Solution Let us first write down the required type signature: the function must take an integer argument $x: \text{Int}$, and the return value must be a function of type `Int => Int`.

```
def add_x(x: Int): Int => Int = ???
```

We are required to return a function that adds x to its argument. Let us call that argument z , to avoid confusion with the x . So, we are required to return the function $\{ z => z + x \}$. Since functions are values, we return a new function by writing a nameless function expression:

```
def add_x(x: Int): Int => Int = { z => z + x }
```

To implement the same function by using a `val`, we first convert the type signature of `add_x` to the equivalent curried type `Int → Int → Int`. Now we can write the Scala code of a function `add_x_v`:

```
val add_x_v: Int => Int => Int = { x => z => z + x }
```

The function `add_x_v` is equal to `add_x` except for using the `val` syntax instead of `def`. It is not necessary to specify the type of the arguments x and z because we already specified the type `Int → Int → Int` for the value `add_x_v`.

Example 4.4.1.3 Using both `def` and `val`, implement a curried function `prime_f` that takes a function f and an integer x , and returns `true` when $f(x)$ is a prime number. Use the function `isPrime` defined in Section 1.1.2.

Solution First, determine the required type signature of `prime_f`. The value $f(x)$ must have type `Int`, or else we cannot check whether it is prime. So, f must have type `Int → Int`. Since `prime_f` should be a curried function, we need to put each argument into its own set of parentheses:

```
def prime_f(f: Int => Int)(x: Int): Boolean = ???
```

To implement `prime_f`, we need to return the result of `isPrime` applied to $f(x)$. A simple solution is

```
def prime_f(f: Int => Int)(x: Int): Boolean = isPrime(f(x))
```

To implement the same function as a `val`, rewrite its type signature as

```
val prime_f: (Int => Int) => Int => Boolean = ???
```

(The parentheses around `Int => Int` are mandatory since `Int => Int => Int => Boolean` would be a completely different type.) The implementation is

```
val prime_f: (Int => Int) => Int => Boolean = { f => x => isPrime(f(x)) }
```

The code `isPrime(f(x))` is a forward composition of the functions f and `isPrime`, so we can write

```
val prime_f: (Int => Int) => Int => Boolean = (f => f andThen isPrime)
```

A nameless function of the form $f \Rightarrow f.\text{something}$ is equivalent to a shorter Scala syntax `(_.something)`. So we finally rewrite the code of `prime_f` as

```
val prime_f: (Int => Int) => Int => Boolean = (_ andThen isPrime)
```

Example 4.4.1.4 Implement a function `choice(x,p,f,g)` that takes a value x , a predicate p , and two functions f and g . The return value must be $f(x)$ if $p(x)$ returns `true`; otherwise the return value must be $g(x)$. Infer the most general type for this function.

Solution The code of this function must be

```
def choice[...](x,p,f,g) = if (p(x)) f(x) else g(x)
```

To infer the most general type for this code, begin by assuming that x has type A , where A is a type parameter. Then the predicate p must have type `A => Boolean`. Since p is an arbitrary predicate, the value $p(x)$ will be sometimes `true` and sometimes `false`. So, `choice(x,p,f,g)` will sometimes compute $f(x)$ and sometimes $g(x)$. It follows that type A must be the argument type of both f and g , which means that the most general types so far are $f:A \rightarrow B$ and $g:A \rightarrow C$, yielding the type signature

$$\text{choice}(x:A, p:A \rightarrow \text{Boolean}, f:A \rightarrow B, g:A \rightarrow C) \quad .$$

What could be the return type of `choice(x,p,f,g)`? If $p(x)$ returns `true`, the function `choice` returns $f(x)$, which is of type B . Otherwise, `choice` returns $g(x)$, which is of type C . However, the type signature of `choice` must be fixed in advance (at compile time) and cannot depend on the value $p(x)$ computed at run time. So, the types of $f(x)$ and of $g(x)$ must be the same, $B = C$. The type signature of `choice` will thus have only two type parameters, A and B :

```
def choice[A, B](x: A, p: A => Boolean, f: A => B, g: A => B): B = if (p(x)) f(x) else g(x)
```

Example 4.4.1.5 Infer the most general type for the fully parametric function

```
def q[...]: ... = { f => g => g(f) }
```

What types are inferred for the expressions `q(q)` and `q(q(q))`?

Solution Begin by assuming $f:A$ with a type parameter A . In the sub-expression $g \rightarrow g(f)$, the curried argument g must itself be a function, because it is being applied to f as $g(f)$. So we assign types as $f:A \rightarrow g:A \rightarrow B \rightarrow g(f)$, where A and B are type parameters. Then the final returned value $g(f)$ has type B . Since there are no other constraints on the types, the types A and B remain arbitrary, so we add them to the type signature:

```
def q[A, B]: A => (A => B) => B = { f => g => g(f) }
```

To match types in the expression $q(q)$, we first assume arbitrary type parameters and write $q[A, B](q[C, D])$. We need to introduce new type parameters C, D because these type parameters may need to be set differently from A, B when we try to match the types in the expression $q(q)$.

The type of the first curried argument of $q[A, B]$, which is A , must match the entire type of $q[C, D]$, which is $C \rightarrow (C \rightarrow D) \rightarrow D$. So we must set the type parameter A as

$$A = C \rightarrow (C \rightarrow D) \rightarrow D \quad .$$

The type of $q(q)$ becomes

$$q^{A,B}(q^{C,D}) : ((C \rightarrow (C \rightarrow D) \rightarrow D) \rightarrow B) \rightarrow B \quad , \quad \text{where } A = C \rightarrow (C \rightarrow D) \rightarrow D \quad .$$

There are no other constraints on the type parameters B, C, D .

We use this result to infer the most general type for $q(q(q))$. We may denote $r \triangleq q(q)$ for brevity; then, as we just found, r has type $((C \rightarrow (C \rightarrow D) \rightarrow D) \rightarrow B) \rightarrow B$. To infer types in the expression $q(r)$, we introduce new type parameters E, F and write $q[E, F](r)$. The type of the argument of $q[E, F]$ is E , and this must be the same as the type of r . This gives the constraint

$$E = ((C \rightarrow (C \rightarrow D) \rightarrow D) \rightarrow B) \rightarrow B \quad .$$

Other than that, the type parameters are arbitrary. The type of the expression $q(q(q))$ is $(E \rightarrow F) \rightarrow F$. We conclude that the most general type of $q(q(q))$ is

$$\begin{aligned} q^{E,F}(q^{A,B}(q^{C,D})) &: (((C \rightarrow (C \rightarrow D) \rightarrow D) \rightarrow B) \rightarrow B) \rightarrow F \rightarrow F \quad , \\ \text{where } A &= C \rightarrow (C \rightarrow D) \rightarrow D \\ \text{and } E &= ((C \rightarrow (C \rightarrow D) \rightarrow D) \rightarrow B) \rightarrow B \quad . \end{aligned}$$

It is clear from this derivation that expressions such as $q(q(q(q)))$, $q(q(q(q(q))))$, etc., are well-typed.

Let us test these results in Scala, renaming the type parameters for clarity to A, B, C, D :

```
scala> def qq[A, B, C]: ((A => (A => B) => B) => C) => C = q(q)
qq: [A, B, C] => (((A => (A => B) => B)) => C) => C

scala> def qqq[A, B, C, D]: (((A => (A => B) => B) => C) => C) => D = q(q(q))
qqq: [A, B, C, D] => (((((A => (A => B) => B)) => C) => C) => D) => D
```

We did not need to write any type parameters within the expressions $q(q)$ and $q(q(q))$ because the full type signature was declared for each of these expressions. Since the Scala compiler did not print any error messages, we are assured that the types match correctly.

Example 4.4.1.6 Infer types in the code expression

$$(f \rightarrow g \rightarrow g(f))(f \rightarrow g \rightarrow g(f))(f \rightarrow f(10))$$

and simplify the code through symbolic calculations.

Solution The given expression is a curried function $f \rightarrow g \rightarrow g(f)$ applied to two curried arguments. The plan is to consider each of these sub-expressions in turn, assigning types for them using type parameters, and then to figure out how to set the type parameters so that all types match.

Begin by renaming the shadowed variables (f and g) to remove shadowing:

$$(f \rightarrow g \rightarrow g(f))(x \rightarrow y \rightarrow y(x))(h \rightarrow h(10)) \quad . \quad (4.10)$$

As we have seen in Example 4.4.1.5, the sub-expression $f \rightarrow g \rightarrow g(f)$ is typed as $f:A \rightarrow g:A \rightarrow B \rightarrow g(f)$, where A and B are some type parameters. The sub-expression $x \rightarrow y \rightarrow y(x)$ is the same function as $f \rightarrow g \rightarrow g(f)$ but with possibly different type parameters, say, $x:C \rightarrow y:C \rightarrow D \rightarrow y(x)$. The types A, B, C, D are so far unknown.

Finally, the variable h in the sub-expression $h \rightarrow h(10)$ must have type $\text{Int} \rightarrow E$, where E is another type parameter. So, the sub-expression $h \rightarrow h(10)$ is a function of type $(\text{Int} \rightarrow E) \rightarrow E$.

The types must match in the entire expression (4.10):

$$(f:A \rightarrow g:A \rightarrow B \rightarrow g(f))(x:C \rightarrow y:C \rightarrow D \rightarrow y(x))(h:\text{Int} \rightarrow E \rightarrow h(10)) . \quad (4.11)$$

It follows that f must have the same type as $x \rightarrow y \rightarrow y(x)$, while g must have the same type as $h \rightarrow h(10)$. The type of g , which we know as $A \rightarrow B$, will match the type of $h \rightarrow h(10)$, which we know as $(\text{Int} \rightarrow E) \rightarrow E$, only if $A = (\text{Int} \rightarrow E)$ and $B = E$. It follows that f has type $\text{Int} \rightarrow E$. At the same time, the type of f must match the type of $x \rightarrow y \rightarrow y(x)$, which is $C \rightarrow (C \rightarrow D) \rightarrow D$. This can work only if $C = \text{Int}$ and $E = (C \rightarrow D) \rightarrow D = (\text{Int} \rightarrow D) \rightarrow D$.

In this way, we have found all the relationships between the type parameters A, B, C, D, E in Eq. (4.11). The type D remains arbitrary, while the type parameters A, B, C, E are expressed as

$$A = \text{Int} \rightarrow (\text{Int} \rightarrow D) \rightarrow D , \quad (4.12)$$

$$B = E = (\text{Int} \rightarrow D) \rightarrow D , \quad (4.13)$$

$$C = \text{Int} .$$

The entire expression in Eq. (4.11) is a saturated application of a curried function, and thus has the same type as the “final” result expression $g(f)$, which has type B . So, the entire expression in Eq. (4.11) has type $B = (\text{Int} \rightarrow D) \rightarrow D$.

Having established that types match, we can now omit the type annotations and rewrite the code:

$$(f \rightarrow g \rightarrow g(f))(x \rightarrow y \rightarrow y(x))(h \rightarrow h(10))$$

substitute $f = x \rightarrow y \rightarrow y(x)$ and $g = h \rightarrow h(10)$: $= (h \rightarrow h(10))(x \rightarrow y \rightarrow y(x))$

substitute h : $= (x \rightarrow y \rightarrow y(x))(10)$

substitute x : $= y \rightarrow y(10) .$

The type of this expression is $(\text{Int} \rightarrow D) \rightarrow D$ with a type parameter D . Since the argument y is an arbitrary function, we cannot simplify $y(10)$ or $y \rightarrow y(10)$ any further. We conclude that the final simplified form of Eq. (4.10) is $y:\text{Int} \rightarrow D \rightarrow y(10)$.

To test this, we first define the function $f \rightarrow g \rightarrow g(f)$ as in Example 4.4.1.5,

```
def q[A, B]: A => (A => B) => B = { f => g => g(f) }
```

We also define the function $h \rightarrow h(10)$ with a general type $(\text{Int} \rightarrow E) \rightarrow E$,

```
def r[E]: (Int => E) => E = { h => h(10) }
```

To help Scala evaluate Eq. (4.11), we need to set the type parameters for the first q function as $q[A, B]$ where A and B are given by Eqs. (4.12)–(4.13):

```
scala> def s[D] = q[Int => (Int => D) => D, (Int => D) => D](q)(r)
s: [D]> (Int => D) => D
```

To verify that the function s^D indeed equals $y:\text{Int} \rightarrow D \rightarrow y(10)$, we apply s^D to some functions of type $\text{Int} \rightarrow D$, say, for $D = \text{Boolean}$ or $D = \text{Int}$:

```
scala> s(_ > 0) // Set D = Boolean and evaluate (10 > 0).
res6: Boolean = true

scala> s(_ + 20) // Set D = Int and evaluate (10 + 20).
res7: Int = 30
```

Example 4.4.1.7 Compute $(x \rightarrow y \rightarrow x(x(y))) \circ (p \rightarrow p(2)) \circ (z \rightarrow z + 3)$ symbolically and infer types.

Solution The forward composition $f \circ g$ substitutes the *body* of f into the argument of g :

substitute $y = f(x) : (x \rightarrow f(x)) \circ (y \rightarrow g(y)) = (x \rightarrow g(f(x))) .$

This allows us to compute the forward compositions left to right:

$$(x \rightarrow y \rightarrow x(x(y))) \circ (p \rightarrow p(2)) = x \rightarrow (y \rightarrow x(x(y)))(2) = x \rightarrow x(x(2)) \quad .$$

$$(x \rightarrow x(x(2))) \circ (z \rightarrow z + 3) = x \rightarrow x(x(2)) + 3 \quad .$$

Computing the pairwise combinations in another order, we get:

$$(p \rightarrow p(2)) \circ (z \rightarrow z + 3) = p \rightarrow p(2) + 3 \quad .$$

$$(x \rightarrow y \rightarrow x(x(y))) \circ (p \rightarrow p(2) + 3) = x \rightarrow (y \rightarrow x(x(y)))(2) + 3 = x \rightarrow x(x(2)) + 3 \quad .$$

Types are inferred as $(x \rightarrow y \rightarrow x(x(y)))^{:(\text{Int} \rightarrow \text{Int}) \rightarrow (\text{Int} \rightarrow \text{Int})} \circ (p \rightarrow p(2))^{:(\text{Int} \rightarrow \text{Int}) \rightarrow \text{Int}} \circ (z \rightarrow z + 3)^{:\text{Int} \rightarrow \text{Int}}$.

4.4.2 Exercises

Exercise 4.4.2.1 Revise the function from Exercise 1.6.2.4, implementing it as a curried function and replacing the hard-coded number 100 by a *curried* first argument. The type signature should become `Int => List[List[Int]] => List[List[Int]]`.

Exercise 4.4.2.2 Implement the function `converge` from Example 4.4.1.1 as a curried function with an additional argument to set the maximum number of iterations, returning `Option[Double]` as the final result type. The new version of `converge` should return `None` if the convergence condition is not satisfied after the given maximum number of iterations. The type signature and an example test:

```
@tailrec def convergeN[X](cond: X => Boolean)(x0: X)(maxIter: Int)(f: X => X): Option[X] = ???

scala> convergeN[Int](_ < 0)(0)(10)(_ + 1) // This does not converge.
res0: Option[Int] = None

scala> convergeN[Double]{ x => math.abs(x * x - 25) < 1e-8 }(1.0)(10) { x => 0.5 * (x + 25 / x) }
res1: Option[Double] = Some(5.00000000053722)
```

Exercise 4.4.2.3 Implement a fully parametric, information-preserving, curried function that recovers an error using a given function argument. The type signature and an example test:

```
def recover[E, A]: Option[Either[E, A]] => (E => A) => Option[A] = ???

scala> recover(Some(Left("error"))){ _ => 123 }
res0: Option[Int] = Some(123)
```

Exercise 4.4.2.4 For `id` and `const` as defined above, what are the types of `id(id)`, `id(id)(id)`, `id(id(id))`, `id(const)`, and `const(const)`? Simplify these code expressions by symbolic calculations.

Exercise 4.4.2.5 For the function `twice` from Example 4.3.2.2, infer the most general type for the function `twice(twice(twice))`. What does that function do? Test your answer on an example.

Exercise 4.4.2.6 Define a function `thrice` similarly to `twice` except it should apply a given function 3 times. What does the function `thrice(thrice(thrice))` do?

Exercise 4.4.2.7 Define a function `ence` similarly to `twice` except it should apply a given function n times, where n is an additional curried argument.

Exercise 4.4.2.8 Define a fully parametric function `flip(f)` that swaps arguments for any given function `f` having two arguments. To test:

```
def f(x: Int, y: Int) = x - y    // Expect f(10, 2) == 8.
val g = flip(f)                 // Now expect g(2, 10) == 8.

scala> assert( f(10, 2) == 8 && g(2, 10) == 8 )
```

Exercise 4.4.2.9 Write a function `curry2` converting an uncurried function of type `(Int, Int) => Int` into an equivalent curried function of type `Int => Int => Int`.

Exercise 4.4.2.10 Apply the function $(x \rightarrow _ \rightarrow x)$ to the value $(z \rightarrow z(q))$ where $q:Q$ is a given value of type Q . Infer types in these expressions.

Exercise 4.4.2.11 Infer types in the following expressions and test in Scala:

- (a) $p \rightarrow q \rightarrow p(t \rightarrow t(q))$. (b) $p \rightarrow q \rightarrow q(x \rightarrow x(p(q)))$.

Exercise 4.4.2.12 Show that the following expressions cannot be well-typed:

- (a) $p \rightarrow p(q \rightarrow q(p))$. (b) $p \rightarrow q \rightarrow q(x \rightarrow p(q(x)))$.

Exercise 4.4.2.13 Infer types and simplify the following code expressions by symbolic calculations:

- (a) $q \rightarrow (x \rightarrow y \rightarrow z \rightarrow x(z)(y(z)))(a \rightarrow a)(b \rightarrow b(q))$.
 (b) $(f \rightarrow g \rightarrow h \rightarrow f(g(h)))(x \rightarrow x)$.
 (c) $(x \rightarrow y \rightarrow x(y))(x \rightarrow y \rightarrow x)$.
 (d) $(x \rightarrow y \rightarrow x(y))(x \rightarrow y \rightarrow y)$.
 (e) $x \rightarrow (f \rightarrow y \rightarrow f(y)(x))(z \rightarrow _ \rightarrow z)$.
 (f) $z \rightarrow (x \rightarrow y \rightarrow x)(x \rightarrow x(z))(y \rightarrow y(z))$.

Exercise 4.4.2.14 Infer types and simplify the following code expressions by symbolic calculations:

- (a) $(z \rightarrow z + 1) \circ (x \rightarrow y \rightarrow x/y) \circ (p \rightarrow p(2))$.
 (b) $(p \rightarrow q \rightarrow p + q + 1) \circ (f \rightarrow f \circ f) \circ (x \rightarrow x(1))$.

Exercise 4.4.2.15* In the following statements, the types A and B are fixed, and functions are *not* assumed to be fully parametric in A or B .

(a) Given a function $h:A \rightarrow B$ that satisfies the law $f:A \rightarrow A \circ h:A \rightarrow B = h:A \rightarrow B$ for any $f:A \rightarrow A$, prove that the function h must ignore its argument and return a fixed value of type B .

(b) We are given two functions $g:A \rightarrow A$ and $h:B \rightarrow B$. We know only that g and h satisfy the law $f:A \rightarrow B \circ h:B \rightarrow B = g:A \rightarrow A \circ f:A \rightarrow B$ for any function $f:A \rightarrow B$. Prove that both g and h must be equal to identity functions of suitable types: $g:A \rightarrow A = \text{id}^A$ and $h:B \rightarrow B = \text{id}^B$.

Hint: choose f to be a suitable *constant* function and substitute f into the given laws.

4.5 Discussion and further developments

4.5.1 Higher-order functions

The **order** of a function is the number of function arrows (\Rightarrow) contained in the type signature of that function. If a function's type signature contains more than one arrow, the function is called a **higher-order** function. Higher-order functions take functions as arguments and/or return functions.

The methods `andThen`, `compose`, `curried`, and `uncurried` are examples of higher-order functions that take other functions as arguments *and* return new functions.

The following examples illustrate the concept of a function's order. Consider the code

```
def f1(x: Int): Int = x + 10
```

The function `f1` has type signature `Int => Int` and order 1, so it is *not* a higher-order function.

```
def f2(x: Int): Int => Int = (z => z + x)
```

The function `f2` has type signature `Int => Int => Int` and is a higher-order function of order 2.

```
def f3(g: Int => Int): Int = g(123)
```

The function `f3` has type signature `(Int => Int) => Int` and is a higher-order function of order 2.

Note that `f2` is a higher-order function only because its return value is of a function type. An equivalent computation can be performed by an uncurried function that is not higher-order:

```
scala> def f2u(x: Int, z: Int): Int = z + x // Type signature (Int, Int) => Int
```

Unlike `f2`, the function `f3` *cannot* be converted to a non-higher-order function because `f3` has an argument of a function type. Converting to an uncurried form cannot eliminate such arguments.

4.5.2 Name shadowing and the scope of bound variables

Bound variables are introduced in nameless functions whenever an argument is defined. For example, in the nameless function $x \rightarrow y \rightarrow x + y$, the bound variables are the curried arguments x and y . The variable y is only defined within the scope ($y \rightarrow x + y$) of the inner function; the variable x is defined within the entire scope of $x \rightarrow y \rightarrow x + y$.

Another way of introducing bound variables in Scala is to write a `val` or a `def` within curly braces:

```
val x = {  
  val y = 10           // Bound variable 'y'.  
  y + y * y  
} // Same as 'val x = 10 + 10 * 10'.
```

A bound variable is invisible outside the scope that defines it. So, it is easy to rename a bound variable: no outside code could possibly use it and depend on its value.

However, outside code may define a variable that (by chance) has the same name as a bound variable inside the scope. Consider this example from calculus: In the integral

$$f(x) = \int_0^x \frac{dx}{1+x} ,$$

two bound variables named x are defined in two scopes: one in the scope of f , another in the scope of the nameless function $x \rightarrow \frac{1}{1+x}$. The convention in mathematics is to treat these two x 's as two *completely different* variables that just happen to have the same name. In sub-expressions where both of these bound variables are visible, priority is given to the bound variable defined in the smaller inner scope. The outer definition of x is then **shadowed** (hidden) by the inner definition of x . For this reason, evaluating $f(10)$ will give

$$f(10) = \int_0^{10} \frac{dx}{1+x} = \log_e(11) \approx 2.398 ,$$

rather than $\int_0^{10} \frac{dx}{1+10} = \frac{10}{11}$. The outer definition $x = 10$ is shadowed within the expression $\frac{1}{1+x}$ by the definition of x in the smaller local scope of $x \rightarrow \frac{1}{1+x}$.

Since this is the standard mathematical convention, the same convention is adopted in functional programming. A variable defined in a function scope (i.e., a bound variable) will shadow any outside definitions of a variable with the same name.

Name shadowing is not advisable in practical programming, because it usually decreases the clarity of code and so invites errors. Consider the nameless function

$$x \rightarrow x \rightarrow x ,$$

and let us decipher this confusing syntax. The symbol \rightarrow groups to the right, so $x \rightarrow x \rightarrow x$ is the same as $x \rightarrow (x \rightarrow x)$. It is a function that takes x and returns $x \rightarrow x$. Since the argument x in $(x \rightarrow x)$ may be renamed to y without changing the function, we can rewrite the code to

$$x \rightarrow (y \rightarrow y) .$$

Having removed name shadowing, we can more easily understand this code and reason about it. For instance, it becomes clear that this function ignores its argument x and always returns the same value (the identity function $y \rightarrow y$). So we can rewrite $(x \rightarrow x \rightarrow x)$ as $(_ \rightarrow y \rightarrow y)$, which is clearer.

4.5.3 Operator syntax for function applications

In mathematics, function applications are sometimes written without parentheses, for instance $\cos x$ or $\sin z$. Formulas such as $2 \sin x \cos x$ imply parentheses as $2 \cdot \sin(x) \cdot \cos(x)$. Functions such as $\cos x$ are viewed as “operators” that are applied to their arguments without parentheses, similar to the operators of summation $\sum_k f(k)$ and differentiation $\frac{d}{dx} f(x)$.

Many programming languages (such as ML, OCaml, F#, Haskell, Elm, PureScript) have adopted this “operator syntax”, making parentheses optional for function arguments so that $f x$ means the same as $f(x)$. Parentheses are still used where necessary to avoid ambiguity or for readability.³

The conventions for nameless functions in the operator syntax become:

- Function expressions group to the right, so $x \rightarrow y \rightarrow z \rightarrow e$ means $x \rightarrow (y \rightarrow (z \rightarrow e))$.
- Function applications group to the left, so $f x y z$ means $((f x) y) z$.
- Function applications group stronger than infix operations, so $f x + y$ means $(f x) + y$, just as in mathematics “ $\cos x + y$ ” groups “ $\cos x$ ” stronger than the infix “ $+$ ” operation.

Thus, $x \rightarrow y \rightarrow a b c + p q$ means $x \rightarrow (y \rightarrow ((a b) c) + (p q))$. When this notation becomes hard to read correctly, one needs to add parentheses, e.g., writing $f(x \rightarrow g h)$ instead of $f x \rightarrow g h$.

This book will not use the “operator syntax” when reasoning about code. Scala does not support the parentheses-free operator syntax; parentheses are needed around each curried argument.

In programming language theory, curried functions are “simpler” because they always have a *single* argument (but may return a function that will consume further arguments). From the point of view of programming practice, curried functions are often harder to read and to write.

In the operator syntax, a curried function f is applied to curried arguments as, e.g., $f 20 4$. This departs further from the mathematical tradition and requires some getting used to. If the two arguments are more complicated than just 20 and 4, the resulting expression may become harder to read, compared with the syntax where commas are used to separate the arguments. (Consider, for instance, the expression $f(g(10), h(20) + 30)$.) To improve readability of code, programmers may prefer to define names for complicated expressions and then use those names as curried arguments.

In Scala, the choice of whether to use curried or uncurried function signatures is largely a matter of syntactic convenience. Most Scala code tends to be written with uncurried functions, while curried functions are used when they produce more easily readable code.

One of the syntactic features for curried functions in Scala is the ability to specify a curried argument using the curly brace syntax. Compare the two definitions of the function `summation` described in Section 1.7.5:

```
def summation1(a: Int, b: Int, g: Int => Int): Int = (a to b).map(g).sum
def summation2(a: Int, b: Int)(g: Int => Int): Int = (a to b).map(g).sum
```

These functions are applied to arguments like this:

```
scala> summation1(1, 10, { x => x**** + 2*x })
res0: Int = 3135

scala> summation2(1, 10) { x => x**** + 2*x }
res1: Int = 3135
```

bodies to contain local definitions (`val` or `def`) of new bound variables.

Another feature of Scala is the “dotless” method syntax: for example, `xs map f` is equivalent to `xs.map(f)` and `f andThen g` is equivalent to `f.andThen(g)`. The “dotless” syntax is available only for infix methods, such as `map`, defined on specific types such as `Seq`. In Scala 3, the “dotless” syntax will only work for methods having a special `@infix` annotation. Do not confuse Scala’s “dotless” method syntax with the operator syntax used in Haskell and other languages.

The code that calls `summation2` is easier to read because the curried argument is syntactically separated from the rest of the code by curly braces. This is especially useful when the curried argument is itself a function with a complicated body, since Scala’s curly braces syntax allows function

³The operator syntax has a long history in programming. It is used in Unix shell commands, for example `cp file1 file2`. In LISP-like languages, function applications are enclosed in parentheses but the arguments are space-separated, for example `(f 10 20)`. Operator syntax is also used in some programming languages such as Tcl, Groovy, and Coffeescript.

4.5.4 Deriving a function's code from its type

We have seen how the procedure of type inference derives the type signature from a function's code. A well-known algorithm for type inference is the Damas-Hindley-Milner algorithm,⁴ with a Scala implementation available.⁵

It is remarkable that one can sometimes perform “code inference”: derive a function's *code* from the function's type signature. We will now look at some examples of this.

Consider a fully parametric function that performs partial applications for arbitrary other functions. A possible type signature is

```
def pa[A, B, C](x: A)(f: (A, B) => C): B => C = ???
```

The function `pa` substitutes a fixed argument value `x:A` into another given function `f`.

How can we implement `pa`? Since `pa(x)(f)` must return a function of type `B => C`, we have no choice other than to begin writing a nameless function in the code,

```
def pa[A, B, C](x: A)(f: (A, B) => C): B => C = { y: B =>
  ??? // Need to compute a value of type C in this scope.
}
```

In the inner scope, we need to compute a value of type `C`, and we have values `x:A`, `y:B`, and `f: (A, B) => C`. How can we compute a value of type `C`? If we knew that `C = Int` when `pa(x)(f)` is applied, we could have simply selected a fixed integer value, say, `1`, as the value of type `C`. If we knew that `C = String`, we could have selected a fixed string, say, `"hello"`, as the value of type `C`. But a fully parametric function cannot use any knowledge of the types of its actual arguments.

So, a fully parametric function cannot produce a value of an arbitrary type `C` from scratch. The only way of producing a value of type `C` is by applying the function `f` to arguments of types `A` and `B`. Since the types `A` and `B` are arbitrary, we cannot obtain any values of these types other than `x:A` and `y:B`. So, the only way of getting a value of type `C` is to compute `f(x, y)`. Thus, the body of `pa` must be

```
def pa[A, B, C](x: A)(f: (A, B) => C): B => C = { y => f(x, y) }
```

In this way, we have *unambiguously* derived the body of this function from its type signature, by assuming that the function must be fully parametric.

Another example is the operation of forward composition $f \circ g$ viewed as a fully parametric function with type signature

```
def before[A, B, C](f: A => B, g: B => C): A => C = ???
```

To implement `before`, we need to create a nameless function of type `A => C`,

```
def before[A, B, C](f: A => B, g: B => C): A => C = { x: A =>
  ??? // Need to compute a value of type C in this scope.
}
```

In the inner scope, we need to compute a value of type `C` from the values $x:A$, $f:A \rightarrow B$, and $g:B \rightarrow C$. Since the type `C` is arbitrary, the only way of obtaining a value of type `C` is by applying `g` to an argument of type `B`. In turn, the only way of obtaining a value of type `B` is to apply `f` to an argument of type `A`. Finally, we have only one value of type `A`, namely $x:A$. So, the only way of obtaining the required result is to compute $g(f(x))$.

We have unambiguously inferred the body of the function from its type signature:

```
def before[A, B, C](f: A => B, g: B => C): A => C = { x => g(f(x)) }
```

In Chapter 5 and in Appendix C, we will see how code can be derived from type signatures for a wide range of fully parametric functions.

⁴https://en.wikipedia.org/wiki/Hindley%20%93Milner_type_system

⁵<http://dysphoria.net/2009/06/28/hindley-milner-type-inference-in-scala/>

5 The logic of types. III. The Curry-Howard correspondence

Fully parametric functions (introduced in Section 4.2) perform operations so general that their code does not depend on values of any specific data types such as `Int` or `String`. An example of a fully parametric function is

```
def before[A, B, C](f: A => B, g: B => C): A => C = { x => g(f(x)) }
```

We have also seen in Section 4.5.4 that for certain functions of this kind, the code can be derived unambiguously from the type signature.

There exists a mathematical theory (called the **Curry-Howard correspondence**) that gives precise conditions for the possibility of deriving a function's code from its type, and a systematic derivation algorithm. Technical details about the algorithm are in Appendix C. This chapter describes the main results and applications of this theory to functional programming.

5.1 Values computed by fully parametric functions

5.1.1 Motivation

Consider possible Scala code for a fully parametric function,

```
def f[A, B, ...]: ... = {  
  val x: Either[A, B] = ... // Some expression here.  
  ... }
```

It is sometimes *impossible* to compute a value of a certain type within the body of a fully parametric function. For example, the fully parametric function `fmap` shown in Section 3.2.3.1 cannot compute any values of type `A`,

```
def fmap[A, B](f: A => B): Option[A] => Option[B] = {  
  val x: A = ??? // Cannot compute x here!  
  ... }
```

function that returns values of type `A`. In `fmap`, no values of type `A` are given as arguments; the given function `f: A => B` returns values of type `B` and not `A`. The code of `fmap` must perform pattern matching on a value of type `Option[A]`:

```
def fmap[A, B](f: A => B): Option[A] => Option[B] = {  
  case None      =>  
    val x: A = ??? // Cannot compute x here!  
    ...  
  case Some(a)   =>  
    val x: A = a   // Can compute x in this scope.  
    ... }
```

result value. This requires computing `x` in all cases, not just within one part of the `match` expression.

The body of `fmap` also cannot compute any values of type `B`. Since no arguments of type `B` are given, the only way of obtaining a value of type `B` would be to apply the function `f: A => B` to *some* value of type `A`; but we just saw that the body of `fmap` cannot compute any values of type `A`.

Another example where one cannot compute a value of a certain type is in the following code:

If this program compiles without type errors, it means that the types match and, in particular, that the function `f` is able to compute a value `x` of type `Either[A, B]`.

The reason is that a fully parametric function cannot compute values of type `A` from scratch without using previously given values of type `A` and without applying a function

Since the case `None` has no values of type `A`, we are unable to compute a value `x` in that scope (as long as `fmap` remains a fully parametric function).

Being able to compute `x:A` "within the body of a function" means that, if needed, the function should be able to *return* `x` as a

```
def before[A, B, C](f: A => B, g: B => C): A => C = {
  // val h: C => A = ??? // Cannot compute h here!
  a => g(f(a)) // Can compute a value of type A => C.
}
```

matter what code we try to write. The reason is that the body of `before` has no given values of type `A` and no functions that return values of type `A`, so a nameless function such as `{c:c => ???}` cannot compute its return value of type `A`. Since a fully parametric function cannot create values of an arbitrary type `A` from scratch, we see no possibility of computing `h` within the body of `before`.

Can we prove rigorously that a value of type `c => A` cannot be computed within the body of `before`? Or, perhaps, a clever trick *could* produce a value of that type? So far, we only saw informal arguments about whether values of certain types can be computed. To make the arguments rigorous, we need to translate statements such as “*a fully parametric function before can compute a value of type `c => A`*” into mathematical formulas, with rigorous rules for proving them true or false.

In Section 3.5.3, we denoted by $\mathcal{CH}(A)$ the proposition “the Code \mathcal{H} has a value of type A ”. By “the code” we now mean the body of a given fully parametric function. So, the notation $\mathcal{CH}(A)$ is not completely adequate because the validity of the proposition $\mathcal{CH}(A)$ depends not only on the choice of the type A but also on the place in the code fragment where the value of type A needs to be computed. What exactly is this additional dependency? In the above examples, we used the *types* of a function’s arguments when reasoning about getting a value of a given type A . Thus, a precise description of the proposition $\mathcal{CH}(A)$ is

\mathcal{CH} -proposition : a fully parametric function having arguments of types
 X, Y, \dots, Z can compute a value of type A . (5.1)

Here X, Y, \dots, Z, A may be either type parameters or more complicated type expressions such as $B \rightarrow C$ or $(C \rightarrow D) \rightarrow E$, built from other type parameters.

If arguments of types X, Y, \dots, Z are given, it means we already have values of these types. So, the propositions $\mathcal{CH}(X), \mathcal{CH}(Y), \dots, \mathcal{CH}(Z)$ will be true. Thus, proposition (5.1) is equivalent to “ $\mathcal{CH}(A)$ assuming $\mathcal{CH}(X), \mathcal{CH}(Y), \dots, \mathcal{CH}(Z)$ ”. In mathematical logic, a statement of this form is called a **sequent** and is denoted by

$\mathcal{CH}(X), \mathcal{CH}(Y), \dots, \mathcal{CH}(Z) \vdash \mathcal{CH}(A)$. (5.2)

The assumptions $\mathcal{CH}(X), \mathcal{CH}(Y), \dots, \mathcal{CH}(Z)$ are called **premises** and the proposition $\mathcal{CH}(A)$ is called the **goal**. Showing rigorously the possibility of computing values in functions means proving sequents of the form (5.2). Our previous examples are denoted by the following sequents:

fmap for Option : $\mathcal{CH}(A \rightarrow B) \vdash \mathcal{CH}(\text{Option}[A] \rightarrow \text{Option}[B])$
the function before : $\mathcal{CH}(A \rightarrow B), \mathcal{CH}(B \rightarrow C) \vdash \mathcal{CH}(A \rightarrow C)$
value of type A within fmap : $\mathcal{CH}(A \rightarrow B), \mathcal{CH}(\text{Option}[A]) \vdash \mathcal{CH}(A)$
value of type C → A within before : $\mathcal{CH}(A \rightarrow B), \mathcal{CH}(B \rightarrow C) \vdash \mathcal{CH}(C \rightarrow A)$

Calculations in formal logic are called **proofs**. So, in this section we gave informal arguments towards proving the first two sequents and disproving the last two. We will now develop tools for rigorous reasoning about sequents.

A proposition $\mathcal{CH}(A)$ may be true for one set of premises such as $\mathcal{CH}(X), \mathcal{CH}(Y), \dots, \mathcal{CH}(Z)$ but false for another. Here and in the following sections, we will be reasoning about \mathcal{CH} -propositions within the body of a *chosen* fully parametric function, i.e., with a fixed set of premises. We will then temporarily omit the premises and use the shorter notation $\mathcal{CH}(A)$.

5.1.2 Type notation and \mathcal{CH} -propositions for standard type constructions

In Section 3.5.3 we saw examples of reasoning about \mathcal{CH} -propositions for case classes and for disjunctive types. We will now extend this reasoning systematically to all type constructions that pro-

The body of `before` may only use the arguments `f` and `g`. We can compose `f` and `g` to get a value of type `A => C`; but it is impossible to compute a value `h` of type `C => A`, no

matter what code we try to write. The reason is that the body of `before` has no given values of type `A` and no functions that return values of type `A`, so a nameless function such as `{c:c => ???}` cannot compute its return value of type `A`. Since a fully parametric function cannot create values of an arbitrary type `A` from scratch, we see no possibility of computing `h` within the body of `before`.

Can we prove rigorously that a value of type `c => A` cannot be computed within the body of `before`? Or, perhaps, a clever trick *could* produce a value of that type? So far, we only saw informal arguments about whether values of certain types can be computed. To make the arguments rigorous, we need to translate statements such as “*a fully parametric function before can compute a value of type `c => A`*” into mathematical formulas, with rigorous rules for proving them true or false.

In Section 3.5.3, we denoted by $\mathcal{CH}(A)$ the proposition “the Code \mathcal{H} has a value of type A ”. By “the code” we now mean the body of a given fully parametric function. So, the notation $\mathcal{CH}(A)$ is not completely adequate because the validity of the proposition $\mathcal{CH}(A)$ depends not only on the choice of the type A but also on the place in the code fragment where the value of type A needs to be computed. What exactly is this additional dependency? In the above examples, we used the *types* of a function’s arguments when reasoning about getting a value of a given type A . Thus, a precise description of the proposition $\mathcal{CH}(A)$ is

\mathcal{CH} -proposition : a fully parametric function having arguments of types
 X, Y, \dots, Z can compute a value of type A . (5.1)

Here X, Y, \dots, Z, A may be either type parameters or more complicated type expressions such as $B \rightarrow C$ or $(C \rightarrow D) \rightarrow E$, built from other type parameters.

If arguments of types X, Y, \dots, Z are given, it means we already have values of these types. So, the propositions $\mathcal{CH}(X), \mathcal{CH}(Y), \dots, \mathcal{CH}(Z)$ will be true. Thus, proposition (5.1) is equivalent to “ $\mathcal{CH}(A)$ assuming $\mathcal{CH}(X), \mathcal{CH}(Y), \dots, \mathcal{CH}(Z)$ ”. In mathematical logic, a statement of this form is called a **sequent** and is denoted by

$\mathcal{CH}(X), \mathcal{CH}(Y), \dots, \mathcal{CH}(Z) \vdash \mathcal{CH}(A)$. (5.2)

The assumptions $\mathcal{CH}(X), \mathcal{CH}(Y), \dots, \mathcal{CH}(Z)$ are called **premises** and the proposition $\mathcal{CH}(A)$ is called the **goal**. Showing rigorously the possibility of computing values in functions means proving sequents of the form (5.2). Our previous examples are denoted by the following sequents:

fmap for Option : $\mathcal{CH}(A \rightarrow B) \vdash \mathcal{CH}(\text{Option}[A] \rightarrow \text{Option}[B])$
the function before : $\mathcal{CH}(A \rightarrow B), \mathcal{CH}(B \rightarrow C) \vdash \mathcal{CH}(A \rightarrow C)$
value of type A within fmap : $\mathcal{CH}(A \rightarrow B), \mathcal{CH}(\text{Option}[A]) \vdash \mathcal{CH}(A)$
value of type C → A within before : $\mathcal{CH}(A \rightarrow B), \mathcal{CH}(B \rightarrow C) \vdash \mathcal{CH}(C \rightarrow A)$

Calculations in formal logic are called **proofs**. So, in this section we gave informal arguments towards proving the first two sequents and disproving the last two. We will now develop tools for rigorous reasoning about sequents.

A proposition $\mathcal{CH}(A)$ may be true for one set of premises such as $\mathcal{CH}(X), \mathcal{CH}(Y), \dots, \mathcal{CH}(Z)$ but false for another. Here and in the following sections, we will be reasoning about \mathcal{CH} -propositions within the body of a *chosen* fully parametric function, i.e., with a fixed set of premises. We will then temporarily omit the premises and use the shorter notation $\mathcal{CH}(A)$.

5.1.2 Type notation and \mathcal{CH} -propositions for standard type constructions

In Section 3.5.3 we saw examples of reasoning about \mathcal{CH} -propositions for case classes and for disjunctive types. We will now extend this reasoning systematically to all type constructions that pro-

grams could use. A special type notation explained in this section will help us write type expressions more concisely. (See Appendix A for reference on the type notation.)

There are six **standard type constructions** supported by all functional languages: primitive types (including `Unit` type and the void type, called `Nothing` in Scala), product (tuple) types, co-product (disjunctive) types, function types, parameterized types, and recursive types. We will now derive the rules for writing \mathcal{CH} -propositions for each of these type constructions (except recursive types).

1a) Rule for `Unit` type The `Unit` type has only a single value `()`, and this value (an “empty tuple”) can be *always* computed since it does not need any previous data:

```
def f[...]: ... = {
  ...
  val x: Unit = () // We can always compute a 'Unit' value.
  ...
}
```

So, the proposition $\mathcal{CH}(\text{Unit})$ is always true. In the type notation, the `Unit` type is denoted by `1`.

Named unit types also have a single value that is always possible to compute. For example,

```
final case class N1()
```

defines a named unit type; we can compute

```
val x: N1 = N1()
```

So, the proposition $\mathcal{CH}(N1)$ is always true. Named unit types are denoted by `1`, just as the `Unit` type itself.

1b) Rule for the void type The Scala type `Nothing` has no values, so the proposition $\mathcal{CH}(\text{Nothing})$ is always false. The type `Nothing` is denoted by `0` in the type notation.

1c) Rule for primitive types For a specific primitive (or library-defined) type such as `Int` or `String`, the corresponding \mathcal{CH} -proposition is *always true* because we could use any constant value, e.g.

```
def f[...]: ... = {
  ...
  val x: String = "abc" // We can always compute a 'String' value.
  ...
}
```

So, the rule for primitive types is the same as the rule for the `Unit` type.

2) Rule for tuple types To compute a value of a tuple type (A, B) requires computing a value of type A *and* a value of type B . This is expressed by the logic formula $\mathcal{CH}((A, B)) = \mathcal{CH}(A) \wedge \mathcal{CH}(B)$. A similar formula holds for case classes, as Eq. (3.2) shows. In the type notation, the tuple (A, B) is written as $A \times B$. Tuples and case classes with more than two parts are denoted similarly as $A \times B \times \dots \times C$. For example, the Scala definition

```
case class Person(firstName: String, lastName: String, age: Int)
```

is written in the type notation as $\text{String} \times \text{String} \times \text{Int}$. So, the rule for tuple types is

$$\mathcal{CH}(A \times B \times \dots \times C) = \mathcal{CH}(A) \wedge \mathcal{CH}(B) \wedge \dots \wedge \mathcal{CH}(C) .$$

3) Rule for disjunctive types A disjunctive type may consist of several case classes. Having a value of a disjunctive type means to have a value of (at least) one of those case classes. An example of translating this relationship into a formula was shown by Eq. (3.1). For the standard disjunctive type `Either[A, B]`, we have the logical formula $\mathcal{CH}(\text{Either}[A, B]) = \mathcal{CH}(A) \vee \mathcal{CH}(B)$. In the type notation, the Scala type `Either[A, B]` is written as $A + B$. A longer example: the Scala definition

```
sealed trait RootsOfQ
final case class NoRoots() extends RootsOfQ
final case class OneRoot(x: Double) extends RootsOfQ
final case class TwoRoots(x: Double, y: Double) extends RootsOfQ
```

is translated to the type notation as

$$\text{RootsOfQ} = 1 + \text{Double} + \text{Double} \times \text{Double} .$$

The type notation is significantly shorter because it omits all case class names and part names from the type definitions. In this notation, the rule for disjunctive types is

$$CH(A + B + \dots + C) = CH(A) \vee CH(B) \vee \dots \vee CH(C) .$$

4) Rule for function types Consider now a function type such as $A \Rightarrow B$. (This type is written in the type notation as $A \rightarrow B$.) To compute a value of that type, we need to write code such as

```
val f: A => B = { (a: A) =>
  ??? // Compute a value of type B in this scope.
}
```

The inner scope of the function needs to compute a value of type B , and the given value $a: A$ may be used for that. So, $CH(A \rightarrow B)$ is true if and only if we are able to compute a value of type B when we are given a value of type A . To translate this statement into the language of logical propositions, we need to use the logical **implication**, $CH(A) \Rightarrow CH(B)$, which means that $CH(B)$ can be proved if $CH(A)$ already holds. So the rule for function types is

$$CH(A \rightarrow B) = CH(A) \Rightarrow CH(B) .$$

5) Rule for parameterized types Consider a function with type parameters, e.g.

```
def f[A, B]: A => (A => B) => B = { x => g => g(x) }
```

Being able to define the body of such a function is equivalent to being able to compute a value of type $A \Rightarrow (A \Rightarrow B) \Rightarrow B$ for *all* possible types A and B . In the notation of formal logic, this is written as

$$CH(\forall(A, B). A \rightarrow (A \rightarrow B) \rightarrow B)$$

and is equivalent to

$$\forall(A, B). CH(A \rightarrow (A \rightarrow B) \rightarrow B) .$$

The code notation for the parameterized function f is

$$f^{A, B} : A \rightarrow (A \rightarrow B) \rightarrow B ,$$

and its type can be written as

$$\forall(A, B). A \rightarrow (A \rightarrow B) \rightarrow B .$$

The symbol \forall means “for all” and is known as the **universal quantifier** in logic.

In Scala, longer type expressions can be named and their names (called **type aliases**) can be used to make code shorter. Type aliases may also contain type parameters. Defining and using a type alias for the type of the function f looks like this,

```
type F[A, B] = A => (A => B) => B
def f[A, B]: F[A, B] = { x => g => g(x) }
```

This is written in the type notation as

$$\begin{aligned} F^{A, B} &\triangleq A \rightarrow (A \rightarrow B) \rightarrow B , \\ f^{A, B} : F^{A, B} &\triangleq x : A \rightarrow g : A \rightarrow B \rightarrow g(x) , \end{aligned}$$

or equivalently (although somewhat less readably)

$$f : (\forall(A, B). F^{A, B}) \triangleq \forall(A, B). x : A \rightarrow g : A \rightarrow B \rightarrow g(x) .$$

In Scala 3, the function f can be written as a value via the syntax

```
val f: [A, B] => A => (A => B) => B = { // Valid only in Scala 3.
  [A, B] => (x: A) => (g: A => B) => g(x)
}
```

Type construction	Scala syntax	Type notation	\mathcal{CH} -proposition
type parameter	<code>[A]</code>	A	$\mathcal{CH}(A)$
product type (tuple)	<code>(A, B)</code>	$A \times B$	$\mathcal{CH}(A) \wedge \mathcal{CH}(B)$
disjunctive type	<code>Either[A, B]</code>	$A + B$	$\mathcal{CH}(A) \vee \mathcal{CH}(B)$
function type	<code>A => B</code>	$A \rightarrow B$	$\mathcal{CH}(A) \Rightarrow \mathcal{CH}(B)$
unit or a “named unit” type	<code>Unit</code>	1	$\mathcal{CH}(1) = \text{True}$
primitive type	<code>Int, String, ...</code>	<code>Int, String, ...</code>	$\mathcal{CH}(\text{Int}) = \text{True}$
void type	<code>Nothing</code>	\emptyset	$\mathcal{CH}(\emptyset) = \text{False}$
value parameterized by type	<code>def f[A]: F[A]</code>	$f^A : F^A$	$\forall A. \mathcal{CH}(F^A)$
type with quantifier	<code>[A] => F[A] (Scala 3)</code>	$\forall A. F^A$	$\forall A. \mathcal{CH}(F^A)$

Table 5.1: The correspondence between type constructions and \mathcal{CH} -propositions.

This syntax corresponds more closely to the mathematical notation shown above.

So, the rule for parameterized types with the type notation F^A is

$$\mathcal{CH}(\forall A. F^A) = \forall A. \mathcal{CH}(F^A) .$$

Case classes and disjunctive types use *names* for the types and their parts. However, those names only add convenience for programmers and do not affect the computational properties of types. The type notation is designed to support nameless type expressions.

Table 5.1 summarizes the type notation and also shows how to translate it into logic formulas with propositions of the form $\mathcal{CH}(\dots)$.

The precedence of operators in the type notation is chosen to have fewer parentheses in the type expressions that are frequently used. The rules of precedence are:

- The type product operator (\times) groups stronger than the disjunctive operator ($+$), so that type expressions such as $A + B \times C$ have the same operator precedence as in standard arithmetic. That is, $A + B \times C$ means $A + (B \times C)$. This convention makes type expressions easier to reason about (for people familiar with arithmetic).
- The function type arrow (\rightarrow) groups weaker than the operators $+$ and \times , so that often-used types such as $A \rightarrow 1 + B$ (representing `A => Option[B]`) or $A \times B \rightarrow C$ (representing `((A, B)) => C`) can be written without any parentheses. Type expressions such as $(A \rightarrow B) \times C$ will require parentheses but are needed less often.
- The type quantifiers group weaker than all other operators, so we can write types such as $\forall A. A \rightarrow A \rightarrow A$ without parentheses. Type quantifiers are most often placed outside a type expression. When this is not the case, parentheses are necessary, e.g., in the type expression $(\forall A. A \rightarrow A \rightarrow A) \rightarrow 1 + 1$.

5.1.3 Solved examples: Type notation

From now on, we will prefer to write types in the type notation rather than in the Scala syntax. The type notation allows us to write nameless type expressions and, in particular, makes the structure of disjunctive types and their parts more clear, compared with the Scala syntax. Names of types and parts of types are, of course, helpful for reminding the programmer of the meaning of data in a program. However, writing names for every part of every type is not helpful for reasoning about the

properties of types. Type notation makes reasoning about types easier, as we will see throughout this chapter. Once the programmer has finished deriving the necessary types and verifying their properties, the type expressions can be straightforwardly translated from the type notation into Scala code.

Let us get some experience converting between type notation and Scala code.

Example 5.1.3.1 Define a function `delta` taking an argument `x` and returning the pair `(x, x)`. Derive the most general type for this function. Write the type signature of `delta` in the type notation, and translate it into a \mathcal{CH} -proposition. Simplify the \mathcal{CH} -proposition if possible.

Solution Begin by writing the code of the function:

```
def delta(x: ...) = (x, x)
```

To derive the most general type for `delta`, first assume `x:A`, where `A` is a type parameter; then the tuple `(x, x)` has type `(A, A)`. We do not see any constraints on the type parameter `A`. So the type parameter represents an arbitrary type and needs to be added to the type signature of `delta`:

```
def delta[A](x: A): (A, A) = (x, x)
```

We find that the most general type of `delta` is `A => (A, A)`. We also note that there is only one way of implementing a fully parametric function with type signature `A => (A, A)`: the function must duplicate its given argument.

It is convenient to use the letter Δ for the function `delta`. In the type notation, the type signature of Δ is written as

$$\Delta^A : A \rightarrow A \times A \quad .$$

So the proposition $\mathcal{CH}(\Delta)$ (meaning “the function Δ can be implemented”) is

$$\mathcal{CH}(\Delta) = \forall A. \mathcal{CH}(A \rightarrow A \times A) \quad .$$

In the type expression $A \rightarrow A \times A$, the product symbol (\times) binds stronger than the function arrow (\rightarrow), so the parentheses in $A \rightarrow (A \times A)$ may be omitted.

Using the rules for transforming \mathcal{CH} -propositions, we rewrite

$$\begin{aligned} \mathcal{CH}(A \rightarrow A \times A) \\ \text{rule for function types :} &= \mathcal{CH}(A) \Rightarrow \mathcal{CH}(A \times A) \\ \text{rule for tuple types :} &= \mathcal{CH}(A) \Rightarrow (\mathcal{CH}(A) \wedge \mathcal{CH}(A)) \quad . \end{aligned}$$

Thus the proposition $\mathcal{CH}(\Delta)$ is equivalent to

$$\mathcal{CH}(\Delta) = \forall A. \mathcal{CH}(A) \Rightarrow (\mathcal{CH}(A) \wedge \mathcal{CH}(A)) \quad .$$

Example 5.1.3.2 The standard disjunctive types `Either[A, B]` and `Option[A]` are written in the type notation as

$$\text{Either}^{A,B} \triangleq A + B \quad , \quad \text{Opt}^A \triangleq \mathbb{1} + A \quad .$$

The type `Either[A, B]` is written as $A + B$ by definition of the disjunctive type notation (+). The type `Option[A]` has two disjoint cases, `None` and `Some[A]`. The case class `None` is a “named Unit” and is denoted by $\mathbb{1}$. The case class `Some[A]` contains a single value of type `A`. So, the type notation for `Option[A]` is $\mathbb{1} + A$.

Example 5.1.3.3 The Scala definition of the disjunctive type `UserAction`,

```
sealed trait UserAction
final case class SetName(first: String, last: String) extends UserAction
final case class SetEmail(email: String) extends UserAction
final case class SetUserId(id: Long) extends UserAction
```

is written in the type notation as

$$\text{UserAction} \triangleq \text{String} \times \text{String} + \text{String} + \text{Long} \quad . \quad (5.3)$$

The type operation \times groups stronger than $+$, as in arithmetic. To derive the type notation (5.3), we first drop all names from case classes and get three nameless tuples $(\text{String}, \text{String})$, (String) , and (Long) . Each of these tuples is then converted into a product using the operator \times , and all products are “summed” in the type notation using the operator $+$.

Example 5.1.3.4 The parameterized disjunctive type `Either3` is a generalization of `Either`:

```
sealed trait Either3[A, B, C]
final case class Left[A, B, C](x: A) extends Either3[A, B, C]
final case class Middle[A, B, C](x: B) extends Either3[A, B, C]
final case class Right[A, B, C](x: C) extends Either3[A, B, C]
```

This disjunctive type is written in the type notation as

$$\text{Either3}^{A,B,C} \triangleq A + B + C \quad .$$

Example 5.1.3.5 Define a Scala type constructor `F[A]` corresponding to the type notation

$$F^A \triangleq \mathbb{1} + \text{Int} \times A \times A + \text{Int} \times (\text{Int} \rightarrow A) \quad .$$

Solution The formula for F^A defines a disjunctive type `F[A]` with three parts. To implement `F[A]` in Scala, we need to choose names for each of the disjoint parts, which will become case classes. For the purposes of this example, let us choose names `F1`, `F2`, and `F3`. Each of these case classes needs to have the same type parameter `A`. So we begin writing the code as

```
sealed trait F[A]
final case class F1[A](...) extends F[A]
final case class F2[A](...) extends F[A]
final case class F3[A](...) extends F[A]
```

Each of these case classes represents one part of the disjunctive type: `F1` represents $\mathbb{1}$, `F2` represents $\text{Int} \times A \times A$, and `F3` represents $\text{Int} \times (\text{Int} \rightarrow A)$. To define these case classes, we need to name their parts. The final code is

```
sealed trait F[A]
final case class F1[A]() extends F[A] // Named unit type.
final case class F2[A](n: Int, x1: A, x2: A) extends F[A]
final case class F3[A](n: Int, f: Int => A) extends F[A]
```

The names `n`, `x1`, `x2`, and `f` are chosen purely for convenience.

Example 5.1.3.6 Write the type signature of the function

```
def fmap[A, B](f: A => B): Option[A] => Option[B]
```

in the type notation.

Solution This is a curried function, so we first rewrite the type signature as

```
def fmap[A, B]: (A => B) => Option[A] => Option[B]
```

The type notation for `Option[A]` is $\mathbb{1} + A$. Now we can write the type signature of `fmap` as

$$\begin{aligned} \text{fmap}^{A,B} : (A \rightarrow B) \rightarrow \mathbb{1} + A \rightarrow \mathbb{1} + B \quad , \\ \text{or equivalently : } \text{fmap} : \forall(A, B). (A \rightarrow B) \rightarrow \mathbb{1} + A \rightarrow \mathbb{1} + B \quad . \end{aligned}$$

We do not put parentheses around $\mathbb{1} + A$ and $\mathbb{1} + B$ because the function arrow (\rightarrow) groups weaker than the other type operations. Parentheses around $(A \rightarrow B)$ are required.

We will usually prefer to write type parameters in superscripts rather than under type quantifiers. So, for example, we will write $\text{id}^A \triangleq x^{:A} \rightarrow x$ rather than $\text{id} \triangleq \forall A. x^{:A} \rightarrow x$.

5.1.4 Exercises: Type notation

Exercise 5.1.4.1 Define a Scala disjunctive type $Q[T, A]$ corresponding to the type notation

$$Q^{T, A} \triangleq \mathbb{1} + T \times A + \text{Int} \times (T \rightarrow T) + \text{String} \times A \quad .$$

Exercise 5.1.4.2 Rewrite `Either[(A, Int), Either[(A, Char), (A, Float)]]` in the type notation.

Exercise 5.1.4.3 Define a Scala type `OptE[A, B]` written in the type notation as $\text{OptE}^{A, B} \triangleq \mathbb{1} + A + B$.

Exercise 5.1.4.4 Write a Scala type signature for the fully parametric function

$$\text{flatMap}^{A, B} : \mathbb{1} + A \rightarrow (A \rightarrow \mathbb{1} + B) \rightarrow \mathbb{1} + B$$

and implement this function, preserving information as much as possible.

5.2 The logic of \mathcal{CH} -propositions

5.2.1 Motivation and first examples

So far, we were able to convert statements such as “*a fully parametric function can compute values of type A*” into logical propositions of the form $\mathcal{CH}(A)$ that we called \mathcal{CH} -propositions. The next step is to determine the proof rules suitable for reasoning about \mathcal{CH} -propositions.

Formal logic uses axioms and derivation rules for proving that certain formulas are true or false. A simple example of a true formula is “any proposition α is equivalent to itself”,

$$\forall \alpha. \alpha = \alpha \quad .$$

In logic, equivalence of propositions is usually understood as **implication** (\Rightarrow) in both directions: $\alpha = \beta$ means $(\alpha \Rightarrow \beta) \wedge (\beta \Rightarrow \alpha)$. So, the above formula is the same as

$$\forall \alpha. \alpha \Rightarrow \alpha \quad .$$

If the proposition α is a \mathcal{CH} -proposition, $\alpha \triangleq \mathcal{CH}(A)$ for some type A , we obtain the formula

$$\forall A. \mathcal{CH}(A) \Rightarrow \mathcal{CH}(A) \quad . \quad (5.4)$$

We expect true \mathcal{CH} -propositions to correspond to types that *can* be computed in a fully parametric function. Let us see if this example fits our expectations. We can rewrite Eq. (5.4) as

$$\begin{aligned} & \forall A. \mathcal{CH}(A) \Rightarrow \mathcal{CH}(A) \\ \text{rule for function types : } &= \underline{\forall A. \mathcal{CH}(A \rightarrow A)} \\ \text{rule for parameterized types : } &= \mathcal{CH}(\forall A. A \rightarrow A) \quad . \end{aligned}$$

The last line shows the \mathcal{CH} -proposition that corresponds to the function type $\forall A. A \rightarrow A$. Translating the type notation into a Scala type signature, we get

```
def f[A]: A => A
```

This type signature can be easily implemented,

```
def f[A]: A => A = { x => x }
```

So, in this example we see how we converted a true formula in logic into the type of a value f that can be implemented.

While the formula $\forall \alpha. \alpha = \alpha$ may be self-evident, the point of using formal logic is to have a set of axioms and proof rules that allow us to deduce *all* correct formulas systematically, without need for intuition or guessing. What axioms and proof rules are suitable for proving \mathcal{CH} -propositions?

A well-known set of logical rules is called Boolean logic. In that logic, each proposition is either *True* or *False*, and the implication operation (\Rightarrow) is *defined* by

$$(\alpha \Rightarrow \beta) \triangleq ((\neg\alpha) \vee \beta) \quad . \quad (5.5)$$

To verify a formula, substitute *True* or *False* into every variable and check if the formula has the value *True* in all possible cases. The result can be arranged into a truth table. The basic operations (disjunction, conjunction, negation, and implication) have the following truth tables:

α	β	$\alpha \vee \beta$	$\alpha \wedge \beta$	$\neg\alpha$	$\alpha \Rightarrow \beta$
<i>True</i>	<i>True</i>	<i>True</i>	<i>True</i>	<i>False</i>	<i>True</i>
<i>True</i>	<i>False</i>	<i>True</i>	<i>False</i>	<i>False</i>	<i>False</i>
<i>False</i>	<i>True</i>	<i>True</i>	<i>False</i>	<i>True</i>	<i>True</i>
<i>False</i>	<i>False</i>	<i>False</i>	<i>False</i>	<i>True</i>	<i>True</i>

The formula $\alpha \Rightarrow \alpha$ has the value *True* whether α itself is *True* or *False*. This check is sufficient to show that $\forall \alpha. \alpha \Rightarrow \alpha$ is true in Boolean logic.

Here is the truth table for the formula $\forall(\alpha, \beta). (\alpha \wedge \beta) \Rightarrow \alpha$; that formula is true in Boolean logic since all values in the last column are *True*:

α	β	$\alpha \wedge \beta$	$(\alpha \wedge \beta) \Rightarrow \alpha$
<i>True</i>	<i>True</i>	<i>True</i>	<i>True</i>
<i>True</i>	<i>False</i>	<i>False</i>	<i>True</i>
<i>False</i>	<i>True</i>	<i>False</i>	<i>True</i>
<i>False</i>	<i>False</i>	<i>False</i>	<i>True</i>

The formula $\forall(\alpha, \beta). \alpha \Rightarrow (\alpha \wedge \beta)$ is not true in Boolean logic, which we can see from the following truth table (one value in the last column is *False*):

α	β	$\alpha \wedge \beta$	$\alpha \Rightarrow (\alpha \wedge \beta)$
<i>True</i>	<i>True</i>	<i>True</i>	<i>True</i>
<i>True</i>	<i>False</i>	<i>False</i>	<i>False</i>
<i>False</i>	<i>True</i>	<i>False</i>	<i>True</i>
<i>False</i>	<i>False</i>	<i>False</i>	<i>True</i>

Table 5.2 shows more examples of logical formulas that are true in Boolean logic. Each formula is first given in terms of \mathcal{CH} -propositions (we denoted $\alpha \triangleq \mathcal{CH}(A)$ and $\beta \triangleq \mathcal{CH}(B)$ for brevity) and then into a Scala type signature of a function that can be implemented.

Table 5.3 some examples of formulas that are *not true* in Boolean logic. Translated into type formulas and then into Scala, these formulas yield type signatures that *cannot* be implemented by fully parametric functions.

At first sight, it appears from these examples that whenever a logical formula is true in Boolean logic, the corresponding type signature can be implemented in code, and vice versa. However, this is *incorrect*: the rules of Boolean logic are not suitable for reasoning about types in a functional language. Below we will see some examples of formulas that are true in Boolean logic but yield unimplementable type signatures.

Logic formula	Type formula	Scala code
$\forall\alpha. \alpha \Rightarrow \alpha$	$\forall A. A \rightarrow A$	<code>def id[A](x: A): A = x</code>
$\forall\alpha. \alpha \Rightarrow \text{True}$	$\forall A. A \rightarrow \mathbb{1}$	<code>def toUnit[A](x: A): Unit = ()</code>
$\forall(\alpha, \beta). \alpha \Rightarrow (\alpha \vee \beta)$	$\forall(A, B). A \rightarrow A + B$	<code>def toL[A, B](x: A): Either[A, B] = Left(x)</code>
$\forall(\alpha, \beta). (\alpha \wedge \beta) \Rightarrow \alpha$	$\forall(A, B). A \times B \rightarrow A$	<code>def first[A, B](p: (A, B)): A = p._1</code>
$\forall(\alpha, \beta). \alpha \Rightarrow (\beta \Rightarrow \alpha)$	$\forall(A, B). A \rightarrow (B \rightarrow A)$	<code>def const[A, B](x: A): B => A = (_ => x)</code>

Table 5.2: Examples of logical formulas that are true theorems in Boolean logic.

Logic formula	Type formula	Scala type signature
$\forall\alpha. \text{True} \Rightarrow \alpha$	$\forall A. \mathbb{1} \rightarrow A$	<code>def f[A](x: Unit): A</code>
$\forall(\alpha, \beta). (\alpha \vee \beta) \Rightarrow \alpha$	$\forall(A, B). A + B \rightarrow A$	<code>def f[A, B](x: Either[A, B]): A</code>
$\forall(\alpha, \beta). \alpha \Rightarrow (\alpha \wedge \beta)$	$\forall(A, B). A \rightarrow A \times B$	<code>def f[A, B](p: A): (A, B)</code>
$\forall(\alpha, \beta). (\alpha \Rightarrow \beta) \Rightarrow \alpha$	$\forall(A, B). (A \rightarrow B) \rightarrow A$	<code>def f[A, B](x: A => B): A</code>

Table 5.3: Examples of logical formulas that are *not* true in Boolean logic.

5.2.2 Example: Failure of Boolean logic for type reasoning

To see an explicit example of obtaining an incorrect result when using Boolean logic to reason about values computed by fully parametric functions, consider the following type,

$$\forall(A, B, C). (A \rightarrow B + C) \rightarrow (A \rightarrow B) + (A \rightarrow C) \quad , \quad (5.6)$$

which corresponds to the Scala type signature

```
def bad[A, B, C](g: A => Either[B, C]): Either[A => B, A => C] = ???
```

The function `bad` cannot be implemented as a fully parametric function. To see why, consider that the only available data is a function $g: A \rightarrow B + C$, which returns values of type B or C depending (in some unknown way) on the input value of type A . The function `bad` must return either a function of type $A \rightarrow B$ or a function of type $A \rightarrow C$. How can the code of `bad` make this decision? The only input data is the function g that takes an argument of type A . We could imagine applying g to various arguments of type A and to see whether g returns a B or a C . However, the type A is arbitrary, and a fully parametric function cannot produce a value of type A in order to apply g to it. So the decision about whether to return $A \rightarrow B$ or $A \rightarrow C$ must be independent of g ; that decision must be hard-coded in the function `bad`.

Suppose we hard-coded the decision to return a function of type $A \rightarrow B$. How can we create a function of type $A \rightarrow B$ in the body of `bad`? Given a value $x: A$ of type A , we would need to compute some value of type B . Since the type B is arbitrary (it is a type parameter), we cannot produce a value of type B from scratch. The only potential source of values of type B is the given function g . The only way of using g is to apply it to $x: A$. However, for some x , the value $g(x)$ may be of the form `Right(c)`, where c is of type C . In that case, we will have a value of type C , not B . So, in general, we cannot guarantee that we can always obtain a value of type B from a given value $x: A$. This means we cannot build a function of type $A \rightarrow B$ out of the function g . Similarly, we cannot build a function of type $A \rightarrow C$ out of g .

Whether we decide to return $A \rightarrow B$ or $A \rightarrow C$, we will not be able to return a value of the required type, as we just saw. We must conclude that we cannot implement `bad` as a fully parametric function.

We could try to switch between $A \rightarrow B$ and $A \rightarrow C$ depending on a given value of type A . This idea, however, means that we are working with a different type signature:

$$\forall(A, B, C). (A \rightarrow B + C) \rightarrow A \rightarrow (A \rightarrow B) + (A \rightarrow C) \quad .$$

This type signature *can* be implemented, for instance, by this Scala code:

```
def q[A, B, C](g: A => Either[B, C]): A => Either[A => B, A => C] = { a =>
  g(a) match {
    case Left(b) => Left(_ => b)
    case Right(c) => Right(_ => c)
  }
}
```

But this is not the required type signature (5.6).

Now let us convert the type signature (5.6) into a \mathcal{CH} -proposition:

$$\forall(\alpha, \beta, \gamma). (\alpha \Rightarrow (\beta \vee \gamma)) \Rightarrow ((\alpha \Rightarrow \beta) \vee (\alpha \Rightarrow \gamma)) \quad , \quad (5.7)$$

where $\alpha \triangleq \mathcal{CH}(A)$, $\beta \triangleq \mathcal{CH}(B)$, $\gamma \triangleq \mathcal{CH}(C)$.

It turns out that this formula is true in Boolean logic. To prove this, we need to show that Eq. (5.7) is equal to *True* for any Boolean values of the variables α, β, γ . One way is to rewrite the expression (5.7) using the rules of Boolean logic, such as Eq. (5.5):

$$\begin{aligned} \alpha \Rightarrow (\beta \vee \gamma) & \\ \text{definition of } \Rightarrow \text{ via Eq. (5.5)} : & = (\neg\alpha) \vee \beta \vee \gamma \quad , \\ & \underline{(\alpha \Rightarrow \beta) \vee (\alpha \Rightarrow \gamma)} \\ \text{definition of } \Rightarrow \text{ via Eq. (5.5)} : & = \underline{(\neg\alpha)} \vee \beta \vee \underline{(\neg\alpha)} \vee \gamma \\ \text{property } x \vee x = x \text{ in Boolean logic} : & = (\neg\alpha) \vee \beta \vee \gamma \quad , \end{aligned}$$

showing that $\alpha \Rightarrow (\beta \vee \gamma)$ is in fact *equal* to $(\alpha \Rightarrow \beta) \vee (\alpha \Rightarrow \gamma)$ in Boolean logic.

Let us also give a proof via truth-value reasoning. The only possibility for an implication $X \Rightarrow Y$ to be *False* is when $X = \text{True}$ and $Y = \text{False}$. So, Eq. (5.7) can be *False* only if $(\alpha \Rightarrow (\beta \vee \gamma)) = \text{True}$ and $(\alpha \Rightarrow \beta) \vee (\alpha \Rightarrow \gamma) = \text{False}$. A disjunction can be false only when both parts are false; so we must have both $(\alpha \Rightarrow \beta) = \text{False}$ and $(\alpha \Rightarrow \gamma) = \text{False}$. This is only possible if $\alpha = \text{True}$ and $\beta = \gamma = \text{False}$. But, with these value assignments, we find $(\alpha \Rightarrow (\beta \vee \gamma)) = \text{False}$ rather than *True* as we assumed. It follows that we cannot ever make Eq. (5.7) equal to *False*. So, Eq. (5.7) is true in Boolean logic.

5.2.3 The rules of proof for \mathcal{CH} -propositions

Section 5.2.2 shows that some true formulas in Boolean logic do not correspond to types of *implementable* fully parametric functions. However, we have also seen several other examples where Boolean logic does provide correct results: some true formulas correspond to implementable type signatures, while some false formulas correspond to non-implementable type signatures.

Instead of guessing whether the rules of Boolean logic are suitable in a given case, let us derive the suitable logical axioms and proof rules systematically.

The proposition $\mathcal{CH}(A)$ is true when a value of type A can be computed by a fully parametric function with a given type signature. To describe all possible ways of computing a value of type A , we need to enumerate all possible ways of writing code within a fully parametric function. The requirement of parametricity means that we are not allowed to use any specific types such as `Int` or `String`. We are only allowed to work with values of unknown types described by the given type parameters. We cannot use any concrete values such as `123` or `"hello"`, or any library functions that work with specific (non-parametric) types; however, we are permitted to use fully parametric types such as `Either[A, B]` or `Option[A]`. The allowed eight code constructs are illustrated in this code fragment:

```

def f[A, B, ...](a: A, b: B)... = { // (A given type signature.)
  val x1: Unit = ()           // 1) Use a value of type Unit.
  val x2: A = a               // 2) Use a given argument.
  val x3 = { x: A => ... }    // 3) Create a function.
  val x4: D = x3(x2)         // 4) Use a function.
  val x5: (A, B) = (a, b)    // 5) Create a tuple.
  val x6: B = x5._2          // 6) Use a tuple.
  val x7: Option[A] = Some(x2) // 7) Create values of a disjunctive type.
  val x8 = x7 match { ... }   // 8) Use values of a disjunctive type.
}

```

A value of type X can be computed (i.e., $\mathcal{CH}(X)$ is true) if and only if we can create a sequence of computed values such as x_1, x_2, \dots , each being the result of one of these eight code constructs, ending with a value of type X . So, each of the eight code constructs should correspond to a logical rule for proving a \mathcal{CH} -proposition.

A set of axioms and proof rules defines a **formal logic**. So, we will now find the proof rules that define the formal logic suitable for reasoning about \mathcal{CH} -propositions.

Because each proof rule will be obtained from a specific code construct, any \mathcal{CH} -proposition such as $\mathcal{CH}(X)$ proved by applying a sequence of these rules will automatically correspond to a code fragment that combines the relevant code constructs to compute a value of type X . Conversely, any fully parametric code computing a value of type X must be a combination of some of the eight code constructs, and that combination can be automatically translated into a sequence of applications of proof rules in the logic to produce a proof of the proposition $\mathcal{CH}(X)$.

Let us now write down the proof rules that follow from the eight code constructs. We will need to consider the full formulation (5.1) of \mathcal{CH} -propositions and write them as sequents such as Eq. (5.2). For brevity, we define $\alpha \triangleq \mathcal{CH}(A)$, $\beta \triangleq \mathcal{CH}(B)$, etc. It is also customary to use the letter Γ to denote a set of premises, such as $\mathcal{CH}(X)$, $\mathcal{CH}(Y)$, ..., $\mathcal{CH}(Z)$ in Eq. (5.2). So, we can write a shorter formula $\Gamma \vdash \alpha$ instead of the sequent (5.2).

With these notations, we will enumerate all the possible ways of proving that a \mathcal{CH} -proposition is true.

1) Use a `Unit` value At any place in the code, we may write the expression `()` of type `Unit`. This expression corresponds to a proof of the proposition $\mathcal{CH}(\mathbb{1})$ with any set Γ of premises (even with an empty set of premises). So, the sequent $\Gamma \vdash \mathcal{CH}(\mathbb{1})$ is always true. The code corresponding to the proof of this sequent is an expression that creates a value of the `Unit` type:

$$\text{Proof } (\Gamma \vdash \mathcal{CH}(\mathbb{1})) = 1 \quad ,$$

where we denoted by 1 the value `()`.

In formal logic, a sequent that is found to be always true, such as our $\Gamma \vdash \mathcal{CH}(\mathbb{1})$, is called an **axiom** and is written in the following notation,

$$\frac{}{\Gamma \vdash \mathcal{CH}(\mathbb{1})} \quad \text{(create unit)} \quad .$$

The “fraction with a label” represents a proof rule. The denominator of the “fraction” is the target sequent that we need to prove. The numerator of the “fraction” can have zero or more other sequents that need to be proved before the target sequent can be proved. In this case, the set of previous sequents is empty: the target sequent is an axiom and so requires no previous sequents for its proof. The label “create unit” is an arbitrary name used to refer to the rule.

2) Use a given argument At any place within the code of a fully parametric function, we may use one of the function’s arguments, say $x:A$. If some argument has type A , it means that $\alpha \triangleq \mathcal{CH}(A)$ belongs to the set of premises of the sequent we are trying to prove. To indicate this, we write the set of premises as “ Γ, α ”. The code construct `x:A` computes a value of type A , i.e., show that α is true, given these premises. This is expressed by the sequent $\Gamma, \alpha \vdash \alpha$. The proof of this sequent corresponds to an expression that returns one of the given arguments (which we here called $x:A$),

$$\text{Proof } (\Gamma, \alpha \vdash \alpha) = x:A \quad .$$

This sequent is an axiom since its proof requires no previous sequents. The formal logic notation for this axiom is

$$\frac{}{\Gamma, \alpha \vdash \alpha} \text{ (use arg)} .$$

3) Create a function At any place in the code, we may compute a nameless function of type, say, $A \rightarrow B$, by writing `(x:A) => expr` as long as a value `expr` of type B can be computed in the inner scope of the function. The code for `expr` is also required to be fully parametric; it may use `x` and/or other values visible in that scope. So we now need to answer the question of whether a fully parametric function can compute a value of type B , given an argument of type A as well as all other arguments previously given to the parent function. This question is answered by a sequent whose premises contain one more proposition, $\text{CH}(A)$, in addition to all previously available premises. Translating this into the language of CH -propositions, we find that we will prove the sequent

$$\Gamma \vdash \text{CH}(A \rightarrow B) = \Gamma \vdash \text{CH}(A) \Rightarrow \text{CH}(B) \triangleq \Gamma \vdash \alpha \Rightarrow \beta$$

if we can prove the sequent $\Gamma, \text{CH}(A) \vdash \text{CH}(B) = \Gamma, \alpha \vdash \beta$. In the notation of formal logic, this is a **derivation rule** (rather than an axiom) and is written as

$$\frac{\Gamma, \alpha \vdash \beta}{\Gamma \vdash \alpha \Rightarrow \beta} \text{ (create function)} .$$

The **turnstile** symbol, \vdash , groups weaker than other operators. So, we can write sequents such as $(\Gamma, \alpha) \vdash (\beta \Rightarrow \gamma)$ with fewer parentheses: $\Gamma, \alpha \vdash \beta \Rightarrow \gamma$.

What code corresponds to the “create function” rule? The proof of $\Gamma \vdash \alpha \Rightarrow \beta$ depends on a proof of another sequent. So, the corresponding code must be a *function* that takes a proof of the previous sequent as an argument and returns a proof of the new sequent. By the CH correspondence, a proof of a sequent corresponds to a code expression of the type given by the goal of the sequent; the expression may use arguments of types corresponding to the premises of the sequent. So, a proof of the sequent $\Gamma, \alpha \vdash \beta$ is an expression `exprB` of type B that may use a given value of type A as well as any other arguments given previously. Then we can write the proof code for the sequent $\Gamma \vdash \alpha \Rightarrow \beta$ as the nameless function `(x:A) => exprB`. This function has type $A \rightarrow B$ and requires us to already have a suitable `exprB`. This exactly corresponds to the proof rule “create function”. We may write the corresponding code as

$$\text{Proof}(\Gamma \vdash \text{CH}(A) \Rightarrow \text{CH}(B)) = x:A \rightarrow \text{Proof}(\Gamma, x:A \vdash \text{CH}(B)) .$$

Here we wrote $x:A$ instead of $\text{CH}(A)$ since the value $x:A$ is a proof of the proposition $\text{CH}(A)$. We will see in Section 5.2.4 how premises such as $\Gamma, x:A$ are implemented in code.

4) Use a function At any place in the code, we may apply an already defined function of type $A \rightarrow B$ to an already computed value of type A . The result will be a value of type B . This corresponds to assuming $\text{CH}(A \rightarrow B)$ and $\text{CH}(A)$, and then deriving $\text{CH}(B)$. The formal logic notation for this proof rule is

$$\frac{\Gamma \vdash \alpha \quad \Gamma \vdash \alpha \Rightarrow \beta}{\Gamma \vdash \beta} \text{ (use function)} .$$

The code corresponding to this proof rule takes previously computed values `x:A` and `f:A => B`, and writes the expression `f(x)`. This can be written as a function application,

$$\text{Proof}(\Gamma \vdash \beta) = \text{Proof}(\Gamma \vdash \alpha \Rightarrow \beta)(\text{Proof}(\Gamma \vdash \alpha)) .$$

5) Create a tuple If we have already computed some values `a:A` and `b:B`, we may write the expression `(a, b)` and so compute a value of the tuple type `(A, B)`. The proof rule is

$$\frac{\Gamma \vdash \alpha \quad \Gamma \vdash \beta}{\Gamma \vdash \alpha \wedge \beta} \text{ (create tuple)} .$$

Writing $a \times b$ to mean the pair (a, b) , we can write the corresponding code expression as

$$\text{Proof}(\Gamma \vdash \alpha \wedge \beta) = \text{Proof}(\Gamma \vdash \alpha) \times \text{Proof}(\Gamma \vdash \beta) .$$

This rule describes creating a tuple of 2 values. A larger tuple, such as (w, x, y, z) , can be expressed via nested pairs, e.g., as $(w, (x, (y, z)))$. So, it suffices to have a sequent rule for creating pairs; this rule can express the sequent rules for creating all other tuples, and we do not need to define separate rules for, say, $\Gamma \vdash \alpha \wedge \beta \wedge \gamma$.

6) Use a tuple If we already have a value $t : (A, B)$ of a tuple type $A \times B$, we can extract one of the parts of the tuple and obtain a value of type A or a value of type B . The code is $t._1$ and $t._2$ respectively, and the corresponding sequent proof rules are

$$\frac{\Gamma \vdash \alpha \wedge \beta}{\Gamma \vdash \alpha} \quad (\text{use tuple-1}) \quad \frac{\Gamma \vdash \alpha \wedge \beta}{\Gamma \vdash \beta} \quad (\text{use tuple-2}) .$$

The code can be written as

$$\begin{aligned} \text{Proof}(\Gamma \vdash \alpha) &= \pi_1(\text{Proof}(\Gamma \vdash \alpha \wedge \beta)) , \\ \text{Proof}(\Gamma \vdash \beta) &= \pi_2(\text{Proof}(\Gamma \vdash \alpha \wedge \beta)) , \end{aligned}$$

where we introduced the notation π_1 and π_2 to mean the Scala code $_{_1}$ and $_{_2}$.

Since all tuples can be expressed through pairs, it is sufficient to have proof rules for pairs.

7) Create a disjunctive value The type `Either[A, B]` corresponding to the disjunction $\alpha \vee \beta$ can be used to define any other disjunctive type; e.g., a disjunctive type with three parts can be expressed as `Either[A, Either[B, C]]`. So it is sufficient to have proof rules for a disjunction of *two* propositions.

There are two ways of creating a value of the type `Either[A, B]`: the code expressions are `Left(x:A)` and `Right(y:B)`. The values `x:A` or `y:B` must have been computed previously (and correspond to previously proved sequents). So, the sequent proof rules are

$$\frac{\Gamma \vdash \alpha}{\Gamma \vdash \alpha \vee \beta} \quad (\text{create Left}) \quad \frac{\Gamma \vdash \beta}{\Gamma \vdash \alpha \vee \beta} \quad (\text{create Right}) .$$

The corresponding code can be written using the case class names `Left` and `Right` as

$$\begin{aligned} \text{Proof}(\Gamma \vdash \alpha \vee \beta) &= \text{Left}(\text{Proof}(\Gamma \vdash \alpha)) , \\ \text{Proof}(\Gamma \vdash \alpha \vee \beta) &= \text{Right}(\text{Proof}(\Gamma \vdash \beta)) . \end{aligned}$$

8) Use a disjunctive value The basic way of using a value of the disjunctive type `Either[A, B]` is by pattern matching on it:

```
val result: C = (e: Either[A, B]) match {
  case Left(x:A)  => expr1(x)
  case Right(y:B) => expr2(y)
}
```

Here, `expr1(x)` must be an expression of some type `C`, computed using `x:A` and any previously available arguments (i.e., the premises Γ). Similarly, `expr2(y)` must be an expression of type `C` computed using `y:B` and previous arguments. It is clear that `expr1(x)` represents a

proof of a sequent with an additional premise of type `A`, i.e., $\Gamma, \alpha \vdash \gamma$, where we denoted $\gamma \triangleq \text{CH}(C)$. Similarly, `expr2(y)` is a proof of the sequent $\Gamma, \beta \vdash \gamma$. So, we can write the proof rule corresponding to the `match/case` expression as a rule with three previous sequents:

$$\frac{\Gamma \vdash \alpha \vee \beta \quad \Gamma, \alpha \vdash \gamma \quad \Gamma, \beta \vdash \gamma}{\Gamma \vdash \gamma} \quad (\text{use Either}) .$$

The code can be written as

$$\text{Proof}(\Gamma \vdash \gamma) = \text{Proof}(\Gamma \vdash \alpha \vee \beta) \text{ match } \begin{cases} \text{case } a:A \rightarrow 0:B + a:A \\ \text{case } b:B \rightarrow b:B + 0:A \end{cases} .$$

Table 5.4 summarizes the eight proof rules derived in this section. These proof rules define a logic known as the **intuitionistic propositional logic** or **constructive propositional logic**. We will call this logic “constructive” for short.

axioms : $\frac{}{\Gamma \vdash \mathcal{CH}(\mathbb{1})} \quad (\text{use unit})$ $\frac{\Gamma, \alpha \vdash \beta}{\Gamma \vdash \alpha \Rightarrow \beta} \quad (\text{create function})$	$\frac{}{\Gamma, \alpha \vdash \alpha} \quad (\text{use arg})$ $\frac{\Gamma \vdash \alpha \quad \Gamma \vdash \alpha \Rightarrow \beta}{\Gamma \vdash \beta} \quad (\text{use function})$ $\frac{\Gamma \vdash \alpha \quad \Gamma \vdash \beta}{\Gamma \vdash \alpha \wedge \beta} \quad (\text{create tuple})$
derivation rules : $\frac{\Gamma \vdash \alpha \wedge \beta}{\Gamma \vdash \alpha} \quad (\text{use tuple-1})$ $\frac{\Gamma \vdash \alpha \quad \Gamma \vdash \beta}{\Gamma \vdash \alpha \vee \beta} \quad (\text{create Left})$ $\frac{\Gamma \vdash \alpha \wedge \beta}{\Gamma \vdash \beta} \quad (\text{use tuple-2})$ $\frac{\Gamma \vdash \beta}{\Gamma \vdash \alpha \vee \beta} \quad (\text{create Right})$	$\frac{\Gamma \vdash \alpha \vee \beta \quad \Gamma, \alpha \vdash \gamma \quad \Gamma, \beta \vdash \gamma}{\Gamma \vdash \gamma} \quad (\text{use Either})$

Table 5.4: Proof rules for the constructive logic.

5.2.4 Example: Proving a \mathcal{CH} -proposition and deriving code

The task is to implement a fully parametric function

```
def f[A, B]: ((A => A) => B) => B = ???
```

Implementing this function is the same as being able to compute a value of type F , where F is defined as

$$F \triangleq \forall(A, B). ((A \rightarrow A) \rightarrow B) \rightarrow B \quad .$$

Since the type parameters A and B are arbitrary, the body of the fully parametric function f cannot use any previously defined values of types A or B . So, the task is formulated as computing a value of type F with *no* previously defined values. This is written as the sequent $\Gamma \vdash \mathcal{CH}(F)$, where the set Γ of premises is empty, $\Gamma = \emptyset$. Rewriting this sequent using the rules of Table 5.1, we get

$$\forall(\alpha, \beta). \emptyset \vdash ((\alpha \Rightarrow \alpha) \Rightarrow \beta) \Rightarrow \beta \quad , \quad (5.8)$$

where we denoted $\alpha \triangleq \mathcal{CH}(A)$ and $\beta \triangleq \mathcal{CH}(B)$.

The next step is to prove the sequent (5.8) using the logic proof rules of Section 5.2.3. For brevity, we will omit the quantifier $\forall(\alpha, \beta)$ since it will be present in front of every sequent.

Begin by looking for a proof rule whose “denominator” has a sequent similar to Eq. (5.8), i.e., has an implication $(p \Rightarrow q)$ in the goal. We have only one rule that can prove a sequent of the form $\Gamma \vdash (p \Rightarrow q)$; this is the rule “create function”. That rule requires us to already have a proof of the sequent $(\Gamma, p) \vdash q$. So, we use this rule with $\Gamma = \emptyset$, and we set $p \triangleq (\alpha \Rightarrow \alpha) \Rightarrow \beta$ and $q \triangleq \beta$:

$$\frac{(\alpha \Rightarrow \alpha) \Rightarrow \beta \vdash \beta}{\emptyset \vdash ((\alpha \Rightarrow \alpha) \Rightarrow \beta) \Rightarrow \beta} \quad .$$

We now need to prove the sequent $(\alpha \Rightarrow \alpha) \Rightarrow \beta \vdash \beta$, which we can write as $\Gamma_1 \vdash \beta$ where $\Gamma_1 \triangleq [(\alpha \Rightarrow \alpha) \Rightarrow \beta]$ denotes the set containing the single premise $(\alpha \Rightarrow \alpha) \Rightarrow \beta$.

There are no proof rules that derive a sequent with an explicit premise of the form of an implication $p \Rightarrow q$. However, we have a rule called “use function” that derives a sequent by assuming another sequent containing an implication. We would be able to use that rule,

$$\frac{\Gamma_1 \vdash \alpha \Rightarrow \alpha \quad \Gamma_1 \vdash (\alpha \Rightarrow \alpha) \Rightarrow \beta}{\Gamma_1 \vdash \beta} \quad ,$$

if we could prove the two sequents $\Gamma_1 \vdash \alpha \Rightarrow \alpha$ and $\Gamma_1 \vdash (\alpha \Rightarrow \alpha) \Rightarrow \beta$. To prove these sequents, note that the rule “create function” applies to $\Gamma_1 \vdash \alpha \Rightarrow \alpha$ like this,

$$\frac{\Gamma_1, \alpha \vdash \alpha}{\Gamma_1 \vdash \alpha \Rightarrow \alpha} \quad .$$

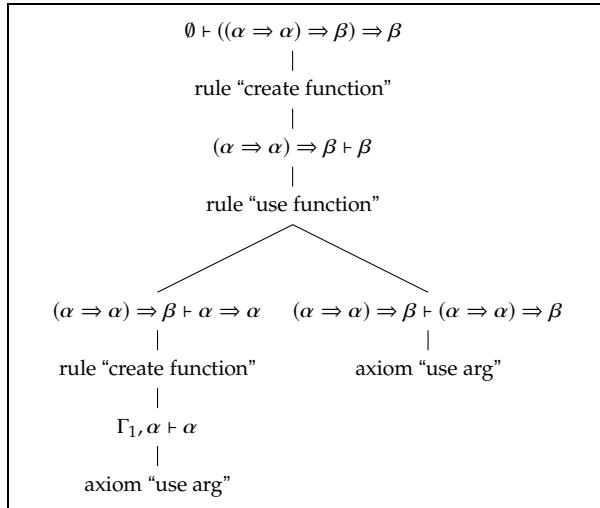


Figure 5.1: Proof tree for sequent (5.8).

The sequent $\Gamma_1, \alpha \vdash \alpha$ is proved directly by the axiom "use arg". The sequent $\Gamma_1 \vdash (\alpha \Rightarrow \alpha) \Rightarrow \beta$ is also proved by the axiom "use arg" because Γ_1 already contains $(\alpha \Rightarrow \alpha) \Rightarrow \beta$.

The proof of the sequent (5.8) is now complete and can be visualized as a tree (Figure 5.1). The next step is to extract the code from that proof.

To do that, we combine the code expressions that correspond to each of the proof rules we used. We need to retrace the proof backwards, starting from the leaves of the tree and going towards the root. We will then assemble the Proof(...) code expressions one by one.

Begin with the left-most leaf "use arg". This rule corresponds to the code $x^{:A}$,

$$\text{Proof}(\Gamma_1, \alpha \vdash \alpha) = x^{:A} .$$

Here $x^{:A}$ must be a proof of the premise α in the sequent $\Gamma_1, \alpha \vdash \alpha$. So, we need to use the same $x^{:A}$ when we write the code for the previous rule, "create function":

$$\text{Proof}(\Gamma_1 \vdash \alpha \Rightarrow \alpha) = (x^{:A} \rightarrow \text{Proof}(\Gamma_1, \alpha \vdash \alpha)) = (x^{:A} \rightarrow x) .$$

The right-most leaf "use arg" corresponds to the code $f^{:(A \rightarrow A) \rightarrow B}$, where f is the premise contained in Γ_1 . So we can write

$$\text{Proof}(\Gamma_1 \vdash (\alpha \Rightarrow \alpha) \Rightarrow \beta) = f^{:(A \rightarrow A) \rightarrow B} .$$

The previous rule, "use function", combines the two preceding proofs:

$$\begin{aligned} & \text{Proof}((\alpha \Rightarrow \alpha) \Rightarrow \beta \vdash \beta) \\ &= \text{Proof}(\Gamma_1 \vdash (\alpha \Rightarrow \alpha) \Rightarrow \beta) (\text{Proof}(\Gamma_1 \vdash \alpha \Rightarrow \alpha)) \\ &= f(x^{:A} \rightarrow x) . \end{aligned}$$

Going further backwards, we find that the rule applied before "use function" was "create function". We need to provide the same $f^{:(A \rightarrow A) \rightarrow B}$ as in the premise above, and so we obtain the code

$$\begin{aligned} & \text{Proof}(\emptyset \vdash ((\alpha \Rightarrow \alpha) \Rightarrow \beta) \Rightarrow \beta) \\ &= f^{:(A \rightarrow A) \rightarrow B} \rightarrow \text{Proof}((\alpha \Rightarrow \alpha) \Rightarrow \beta \vdash \beta) \\ &= f^{:(A \rightarrow A) \rightarrow B} \rightarrow f(x^{:A} \rightarrow x) . \end{aligned}$$

This is the final code expression that implements the type $((A \rightarrow A) \rightarrow B) \rightarrow B$. In this way, we have systematically derived the code from the type signature of a function. This function can be written in Scala as

```
def f[A, B]: ((A => A) => B) => B = { f => f(x => x) }
```

We found the proof tree in Figure 5.1 by guessing how to combine various proof rules. If we *somehow* find a proof tree for a sequent, we can prove the sequent and derive the corresponding code. However, it is not always obvious how to combine the proof rules to prove a given initial sequent. This is because the rules of Table 5.7 do not provide an algorithm for finding a proof tree automatically. It turns out that such an algorithm exists (the “LJT algorithm”, see Appendix C). That algorithm can find proofs and infer code from type signatures containing tuples, disjunctive types, and function types, if the given type signature can be implemented.

The library `curryhoward`¹ implements the LJT algorithm. Here are some examples of using this library for “code inference”. We will run the `ammonite`² shell to load the library more easily.

As a non-trivial (but artificial) example, consider the type signature

$$\forall(A, B). (((A \rightarrow B) \rightarrow A) \rightarrow B) \rightarrow B .$$

It is not immediately clear whether it is possible to implement a function with this type signature. It turns out that it *is* possible, and the code can be derived automatically by the LJT algorithm. The library does this via the method `implement`:

```
@ import $ivy.`io.chymyst:curryhoward:0.3.7`, io.chymyst.ch._

@ def f[A, B]: (((A => B) => A) => B) => B = implement
defined function f

@ println(f.lambdaTerm.prettyPrint)
a => a (b => b (c => a (d => c)))
```

The code $a \rightarrow a (b \rightarrow b (c \rightarrow a (d \rightarrow c)))$ was derived automatically for the function `f`. The function `f` has been compiled and is ready to be used in any subsequent code.

A compile-time error occurs when trying to use a type signature that cannot be implemented as a fully parametric function:

```
@ def g[A, B]: ((A => B) => A) => A = implement
cmd3.sc:1: type ((A => B) => A) => A cannot be implemented
def g[A, B]: ((A => B) => A) => A = implement
                                         ^
Compilation Failed
```

The logical formula corresponding to this type signature is

$$\forall(\alpha, \beta). ((\alpha \Rightarrow \beta) \Rightarrow \alpha) \Rightarrow \alpha . \quad (5.9)$$

This formula is known as “Peirce’s law”.³ It is another example showing that the logic of types in functional programming languages is not Boolean. Peirce’s law is a true theorem in Boolean logic but does not hold in the constructive logic (i.e., it cannot be derived using the proof rules of Table 5.7). If we try to implement `g[A, B]` with the type signature shown above, we will fail to write fully parametric code for `g` that compiles without type errors. This is because no such code exists, — not because we are insufficiently clever. The LJT algorithm can *prove* that the given type signature cannot be implemented; the `curryhoward` library will then print an error message, and compilation will fail.

As another example, let us verify that the type signature from Section 5.2.2 is not implementable by fully parametric functions:

```
@ def bad[A, B, C](g: A => Either[B, C]): Either[A => B, A => C] = implement
cmd4.sc:1: type (A => Either[B, C]) => Either[A => B, A => C] cannot be implemented
def bad[A, B, C](g: A => Either[B, C]): Either[A => B, A => C] = implement
                                         ^
Compilation Failed
```

The rules of constructive logic and the LJT algorithm define rigorously what it means to infer code “guided by the types”. However, in order to use the LJT algorithm productively, a programmer must know how to infer code from types by hand. We will practice doing that throughout the book.

¹<https://github.com/Chymyst/curryhoward>

²<http://ammonite.io/#Ammonite-Shell>

³https://en.wikipedia.org/wiki/Peirce%27s_law

5.3 Solved examples: Equivalence of types

We found a correspondence between types, code, logical propositions, and proofs, which is known as the **Curry-Howard correspondence**. An example of the CH correspondence is that a proof of the logical proposition

$$\forall(\alpha, \beta). \alpha \Rightarrow (\beta \Rightarrow \alpha) \quad (5.10)$$

corresponds to the code of the function

```
def f[A, B]: A => (B => A) = { x => _ => x }
```

With the CH correspondence in mind, we may say that the *existence* of the code `x => _ => x` with the type $A \rightarrow (B \rightarrow A)$ “is” a proof of the logical formula (5.10), because it shows how to compute a value of type $\forall(A, B). A \rightarrow B \rightarrow A$.

The Curry-Howard correspondence maps logic formulas such as $(\alpha \vee \beta) \wedge \gamma$ into type expressions such as $(A + B) \times C$. We have seen that types behave similarly to logic formulas in one respect: A logic formula is a true theorem of constructive logic when the corresponding type signature can be implemented as a fully parametric function, and vice versa.

It turns out that the similarity ends here. In other respects, type expressions behave as *arithmetic* expressions and not as logic formulas. For this reason, the type notation used in this book denotes disjunctive types by $A + B$ and tuples by $A \times B$, which is designed to remind us of arithmetic expressions (such as $1 + 2$ and 2×3) rather than of logical formulas (such as $A \vee B$ and $A \wedge B$).

An important use of the type notation is for writing equations with types. Can we use the arithmetic intuition for writing type equations such as

$$(A + B) \times C = A \times C + B \times C \quad ? \quad (5.11)$$

In this section, we will learn how to check whether one type expression is equivalent to another.

5.3.1 Logical identity does not correspond to type equivalence

The CH correspondence maps Eq. (5.11) into the logic formula

$$\forall(A, B, C). (A \vee B) \wedge C = (A \wedge C) \vee (B \wedge C) \quad . \quad (5.12)$$

This formula is the well-known “distributive law”⁴ valid in Boolean logic as well as in the constructive logic. Since a logical equation $P = Q$ means $P \Rightarrow Q$ and $Q \Rightarrow P$, the distributive law (5.12) means that the two formulas hold,

$$\forall(A, B, C). (A \vee B) \wedge C \Rightarrow (A \wedge C) \vee (B \wedge C) \quad , \quad (5.13)$$

$$\forall(A, B, C). (A \wedge C) \vee (B \wedge C) \Rightarrow (A \vee B) \wedge C \quad . \quad (5.14)$$

The CH correspondence maps these logical formulas to fully parametric functions with types

```
def f1[A, B, C]: ((Either[A, B], C)) => Either[(A, C), (B, C)] = ???  
def f2[A, B, C]: Either[(A, C), (B, C)] => (Either[A, B], C) = ???
```

In the type notation, these type signatures are written as

$$f_1^{A, B, C} : (A + B) \times C \rightarrow A \times C + B \times C \quad ,$$

$$f_2^{A, B, C} : A \times C + B \times C \rightarrow (A + B) \times C \quad .$$

Since the two logical formulas (5.13)–(5.14) are true theorems in constructive logic, we expect to be able to implement the functions `f1` and `f2`. It is not straightforward to guess how to combine the proof rules of Table 5.7 to obtain proofs of Eqs. (5.13)–(5.14). So, instead of deriving the implementations of `f1` and `f2` from the CH correspondence, we will write the Scala code directly.

To implement `f1`, we need to perform pattern matching on the argument:

⁴https://en.wikipedia.org/wiki/Distributive_property#Rule_of_replacement

```
def f1[A, B, C]: ((Either[A, B], C)) => Either[(A, C), (B, C)] = {
  case (Left(a), c)  => Left((a, c)) // No other choice here.
  case (Right(b), c) => Right((b, c)) // No other choice here.
}
```

In both cases, we have only one possible expression of the correct type.

Similarly, the implementation of `f2` leaves us no choices:

```
def f2[A, B, C]: Either[(A, C), (B, C)] => (Either[A, B], C) = {
  case Left((a, c))  => (Left(a), c) // No other choice here.
  case Right((b, c)) => (Right(b), c) // No other choice here.
}
```

orously as a requirement that an arbitrary value $x: \text{Either}[(A, B), C]$ be mapped by `f1` to some value $y: \text{Either}[(A, C), (B, C)]$ and then mapped by `f2` back to *the same* value x . Similarly, any value y of type `Either[(A, C), (B, C)]` should be transformed by `f2` and then by `f1` back to the same value y .

Let us write these conditions as equations,

$$\forall x^{(A+B) \times C}. f_2(f_1(x)) = x \quad , \quad \forall y^{A \times C + B \times C}. f_1(f_2(y)) = y \quad .$$

If these equations hold, it means that all the information in a value $x^{(A+B) \times C}$ is completely preserved inside the value $y \triangleq f_1(x)$; the original value x can be recovered as $x = f_2(y)$. Then the function `f1` is the **inverse** of `f2`. Conversely, all the information in a value $y^{A \times C + B \times C}$ is preserved inside $x \triangleq f_2(y)$ and can be recovered by applying `f1`. Since the values $x^{(A+B) \times C}$ and $y^{A \times C + B \times C}$ are arbitrary, it will follow that the *data types* themselves, $(A + B) \times C$ and $A \times C + B \times C$, carry equivalent information. Such types are called equivalent or isomorphic.

Generally, we say that types P and Q are **equivalent** or **isomorphic** (denoted $P \cong Q$) when there exist functions $f_1: P \rightarrow Q$ and $f_2: Q \rightarrow P$ that are inverses of each other. We can write these conditions using the notation $(f_1 \circ f_2)(x) \triangleq f_2(f_1(x))$ as

$$f_1 \circ f_2 = \text{id} \quad , \quad f_2 \circ f_1 = \text{id} \quad .$$

(In Scala, the forward composition $f_1 \circ f_2$ is the function `f1 andThen f2`. We omit type annotations since we already checked that the types match.) If these conditions hold, there is a one-to-one correspondence between values of types P and Q . This is the same as to say that the data types P and Q “carry equivalent information”.

To verify that the Scala functions `f1` and `f2` defined above are inverses of each other, we first check if $f_1 \circ f_2 = \text{id}$. Applying $f_1 \circ f_2$ means to apply `f1` and then to apply `f2` to the result. Begin by applying `f1` to an arbitrary value $x^{(A+B) \times C}$. A value x of that type can be in only one of the two disjoint cases: a tuple `(Left(a), c)` or a tuple `(Right(b), c)`, for some values $a:A$, $b:B$, and $c:C$. The Scala code of `f1` maps these tuples to `Left((a, c))` and to `Right((b, c))` respectively; we can see this directly from the code of `f1`. We then apply `f2` to those values, which maps them back to a tuple `(Left(a), c)` or a tuple `(Right(b), c)` respectively, according to the code of `f2`. These tuples are exactly the value x we started with. So, applying $f_1 \circ f_2$ to an arbitrary $x^{(A+B) \times C}$ does not change the value x ; this is the same as to say that $f_1 \circ f_2 = \text{id}$.

To check whether $f_2 \circ f_1 = \text{id}$, we apply `f2` to an arbitrary value $y^{A \times C + B \times C}$, which must be one of the two disjoint cases, `Left((a, c))` or `Right((b, c))`. The code of `f2` maps these two cases into tuples `(Left(a), c)` and `(Right(b), c)` respectively. Then we apply `f1` and map these tuples back to `Left((a, c))` and `Right((b, c))` respectively. It follows that applying `f2` and then `f1` will always recover the initial value y . In other words, $f_2 \circ f_1 = \text{id}$.

By looking at the code of `f1` and `f2`, we can directly observe that these functions are inverses of each other: the tuple pattern `(Left(a), c)` is mapped to `Left((a, c))`, and the pattern `(Right(b), c)` to `Right((b, c))`, or vice versa. It is visually clear that no information is lost and that the original values are restored by function compositions `f1 \circ f2` or `f2 \circ f1`.

The code of `f1` and `f2` never discards any given values; in other words, these functions appear to preserve information. We can formulate this property rig-

We find that the logical identity (5.12) leads to an equivalence of the corresponding types,

$$(A + B) \times C \cong A \times C + B \times C . \quad (5.15)$$

To get Eq. (5.15) from Eq. (5.12), we need convert a logical formula to an arithmetic expression by mentally replacing the disjunction operations \vee by $+$ and the conjunctions \wedge by \times everywhere.

Consider another example of a logical identity: the associativity law for conjunction,

$$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma) . \quad (5.16)$$

The corresponding types are $(A \times B) \times C$ and $A \times (B \times C)$; in Scala, $((A, B), C)$ and $(A, (B, C))$. We can define functions that convert between these types without information loss:

```
def f3[A, B, C]: (((A, B), C)) => (A, (B, C)) = { case ((a, b), c) => (a, (b, c)) }
def f4[A, B, C]: (A, (B, C)) => (((A, B), C)) = { case (a, (b, c)) => ((a, b), c) }
```

By applying these functions to arbitrary values of types $((A, B), C)$ and $(A, (B, C))$, it is easy to see that the functions $f3$ and $f4$ are inverses of each other. This is also directly visible in the code: the nested tuple pattern $((a, b), c)$ is mapped to the pattern $(a, (b, c))$ and back. So, the types $(A \times B) \times C$ and $A \times (B \times C)$ are equivalent, and we can write $A \times B \times C$ without parentheses.

Does a logical identity always correspond to an equivalence of types? This turns out to be *not* so. A simple example of a logical identity that does not correspond to a type equivalence is

$$True \vee \alpha = True . \quad (5.17)$$

Since the CH correspondence maps the logical constant *True* into the unit type $\mathbb{1}$, the type equivalence corresponding to Eq. (5.17) is $\mathbb{1} + A \cong \mathbb{1}$. The type denoted by $\mathbb{1} + A$ means `Option[A]` in Scala, so the corresponding equivalence is `Option[A] ≈ Unit`. Intuitively, this type equivalence should not hold: an `Option[A]` may carry a value of type `A`, which cannot possibly be stored in a value of type `Unit`. We can verify this intuition rigorously by proving that any fully parametric functions with type signatures $g_1 : \mathbb{1} + A \rightarrow \mathbb{1}$ and $g_2 : \mathbb{1} \rightarrow \mathbb{1} + A$ will not satisfy $g_1 \circ g_2 = \text{id}$. To verify this, we note that $g_2 : \mathbb{1} \rightarrow \mathbb{1} + A$ must have type signature

```
def g2[A]: Unit => Option[A] = ???
```

Such a function must always return `None`, since a fully parametric function cannot produce values of an arbitrary type `A` from scratch. Therefore, $g_1 \circ g_2$ is also a function that always returns `None`. The function $g_1 \circ g_2$ has type signature $\mathbb{1} + A \rightarrow \mathbb{1} + A$ or, in Scala syntax, `Option[A] => Option[A]`, and is not equal to the identity function, because the identity function does not *always* return `None`.

Another example of a logical identity without a type equivalence is the distributive law

$$\forall (A, B, C). (A \wedge B) \vee C = (A \vee C) \wedge (B \vee C) , \quad (5.18)$$

which is “dual” to the law (5.12), i.e., it is obtained from Eq. (5.12) by swapping all conjunctions (\wedge) with disjunctions (\vee). In logic, a dual formula to an identity is often also an identity. The CH correspondence maps Eq. (5.18) into the type equation

$$\forall (A, B, C). (A \times B) + C = (A + C) \times (B + C) . \quad (5.19)$$

However, the types $A \times B + C$ and $(A + C) \times (B + C)$ are *not* equivalent. To see why, look at the possible code of the function $g_3 : (A + C) \times (B + C) \rightarrow A \times B + C$:

```
1 def g3[A,B,C]: ((Either[A, C], Either[B, C])) => Either[(A, B), C] = {
2   case (Left(a), Left(b))    => Left((a, b)) // No other choice.
3   case (Left(a), Right(c))   => Right(c)      // No other choice.
4   case (Right(c), Left(b))   => Right(c)      // No other choice.
5   case (Right(c1), Right(c2)) => Right(c1)    // Must discard c1 or c2 here!
6 } // May return Right(c2) instead of Right(c1) in the last line.
```

In line 5, we have a choice of returning `Right(c1)` or `Right(c2)`. Whichever we choose, we will lose information because we will have discarded one of the given values `c1`, `c2`. After evaluating `g3`, we will not be able to restore *both* `c1` and `c2`, no matter what code we write for `g4`. So, the composition `g3 ∘ g4` cannot be equal to the identity function. The type equation (5.19) is incorrect.

We conclude that a logical identity $\mathcal{CH}(P) = \mathcal{CH}(Q)$ guarantees, via the CH correspondence, that we can implement *some* fully parametric functions of types $P \rightarrow Q$ and $Q \rightarrow P$. However, it is not guaranteed that these functions are inverses of each other, i.e., that the type conversions $P \rightarrow Q$ or $Q \rightarrow P$ have no information loss. So, the type equivalence $P \cong Q$ does not automatically follow from the logical identity $\mathcal{CH}(P) = \mathcal{CH}(Q)$.

The CH correspondence means that we can compute *some* value $x:X$ of a given type X when the proposition $\mathcal{CH}(X)$ holds. However, the CH correspondence does not guarantee that the computed value $x:X$ will satisfy any additional properties or laws.

5.3.2 Arithmetic identity corresponds to type equivalence

Looking at the examples of equivalent types, we notice that correct type equivalences correspond to *arithmetical* identities rather than *logical* identities. For instance, the logical identity in Eq. (5.12) leads to the type equivalence (5.15), which looks like a standard identity of arithmetic, such as

$$(1 + 10) \times 20 = 1 \times 20 + 10 \times 20 \quad .$$

The logical identity in Eq. (5.18), which does *not* yield a type equivalence, leads to an incorrect arithmetic equation 5.19, e.g., $(1 \times 10) + 20 \neq (1 + 20) \times (10 + 20)$. Similarly, the associativity law (5.16) leads to a type equivalence and to the arithmetic identity

$$(a \times b) \times c = a \times (b \times c) \quad ,$$

while the logical identity in Eq. (5.17), which does not yield a type equivalence, leads to an incorrect arithmetic statement $(\forall a. 1 + a = 1)$.

Table 5.5 summarizes these and other examples of logical identities, with the corresponding type equivalences. In all rows, quantifiers such as $\forall a$ or $\forall(A, B)$ are implied when necessary.

Because we chose the type notation to be similar to the ordinary arithmetic notation, it is easy to translate a possible type equivalence into an arithmetic equation. In all cases, valid arithmetic identities correspond to type equivalences, and failures to obtain a type equivalence correspond to incorrect arithmetic identities. With regard to type equivalence, types such as $A + B$ and $A \times B$ behave similarly to arithmetic expressions such as $10 + 20$ and 10×20 and not similarly to logical formulas such as $\alpha \vee \beta$ and $\alpha \wedge \beta$.

We already verified the first line and the last three lines of Table 5.5. Other identities are verified in a similar way. Let us begin with lines 3 and 4 of Table 5.5, which involve the proposition `False` and the corresponding void type `0` (Scala's `Nothing`). Reasoning about the void type needs a special technique that we will now develop while verifying the type isomorphisms $0 \times A \cong 0$ and $0 + A \cong A$.

Example 5.3.2.1 Verify the type equivalence $0 \times A \cong 0$.

Solution Recall that the type notation $0 \times A$ represents the Scala tuple type `(Nothing, A)`. To demonstrate that the type `(Nothing, A)` is equivalent to the type `Nothing`, we need to show that the type `(Nothing, A)` has *no* values. Indeed, how could we create a value of type, say, `(Nothing, Int)`? We would need to fill *both* parts of the tuple. We have values of type `Int`, but we can never get a value of type `Nothing`. So, regardless of the type `A`, it is impossible to create any values of type `(Nothing, A)`. In other words, the set of values of the type `(Nothing, A)` is empty; but that is the definition of the void type `Nothing`. The types `(Nothing, A)` (denoted by $0 \times A$) and `Nothing` (denoted by `0`) are both void and therefore equivalent.

Example 5.3.2.2 Verify the type equivalence $0 + A \cong A$.

Logical identity	Type equivalence (if it holds)
$True \vee \alpha = True$	$\mathbb{1} + A \not\cong \mathbb{1}$
$True \wedge \alpha = \alpha$	$\mathbb{1} \times A \cong A$
$False \vee \alpha = \alpha$	$\mathbb{0} + A \cong A$
$False \wedge \alpha = False$	$\mathbb{0} \times A \cong \mathbb{0}$
$\alpha \vee \beta = \beta \vee \alpha$	$A + B \cong B + A$
$\alpha \wedge \beta = \beta \wedge \alpha$	$A \times B \cong B \times A$
$(\alpha \vee \beta) \vee \gamma = \alpha \vee (\beta \vee \gamma)$	$(A + B) + C \cong A + (B + C)$
$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma)$	$(A \times B) \times C \cong A \times (B \times C)$
$(\alpha \vee \beta) \wedge \gamma = (\alpha \wedge \gamma) \vee (\beta \wedge \gamma)$	$(A + B) \times C \cong A \times C + B \times C$
$(\alpha \wedge \beta) \vee \gamma = (\alpha \vee \gamma) \wedge (\beta \vee \gamma)$	$(A \times B) + C \not\cong (A + C) \times (B + C)$

Table 5.5: Logic identities with disjunction and conjunction, and the corresponding equivalences of types.

Solution Recall that the type notation $\mathbb{0} + A$ represents the Scala type `Either[Nothing, A]`. We need to show that any value of that type can be mapped without loss of information to a value of type `A`, and vice versa. This means implementing functions $f_1 : \mathbb{0} + A \rightarrow A$ and $f_2 : A \rightarrow \mathbb{0} + A$ such that $f_1 \circ f_2 = \text{id}$ and $f_2 \circ f_1 = \text{id}$.

The argument of f_1 is of type `Either[Nothing, A]`. How can we create a value of that type? Our only choices are to create a `Left(x)` with $x:\text{Nothing}$, or to create a `Right(y)` with $y:A$. However, we cannot create a value x of type `Nothing` because the type `Nothing` has *no* values; so we cannot create a `Left(x)`. The only remaining possibility is to create a `Right(y)` with some value y of type `A`. So, any values of type $\mathbb{0} + A$ must be of the form `Right(y)`, and we can extract that y to obtain a value of type `A`:

```
def f1[A]: Either[Nothing, A] => A = {
  case Right(y) => y
  // No need for 'case Left(x) => ...' since no 'x' can ever be given in 'Left(x)'.
}
```

For the same reason, there is only one implementation of the function f_2 ,

```
def f2[A]: A => Either[Nothing, A] = { y => Right(y) }
```

It is clear from the code that the functions f_1 and f_2 are inverses of each other.

We have just seen that a value of type $\mathbb{0} + A$ is always a `Right(y)` with some $y:A$. Similarly, a value of type $A + \mathbb{0}$ is always a `Left(x)` with some $x:A$. So, we will use the notation $A + \mathbb{0}$ and $\mathbb{0} + A$ to denote the `Left` and the `Right` parts of the disjunctive type `Either`. This notation agrees with the behavior of the Scala compiler, which will infer the types `Either[A, Nothing]` or `Either[Nothing, A]` for these parts:

We can write the functions `toLeft` and `toRight` in a code notation as

$$\text{toLeft}^{A,B} = x:A \rightarrow x:A + \mathbb{0}:B \quad , \\ \text{toRight}^{A,B} = y:B \rightarrow \mathbb{0}:A + y:B \quad .$$

```
def toLeft[A, B]: A => Either[A, B] = x => Left(x)
def toRight[A, B]: B => Either[A, B] = y => Right(y)

scala> toLeft(123)
res0: Either[Int, Nothing] = Left(123)

scala> toRight("abc")
res1: Either[Nothing, String] = Right("abc")
```

In this notation, a value of the disjunctive type is shown without using Scala class names such as `Either`, `Right`, and `Left`. This shortens the writing and speeds up code reasoning.

The type annotation $\mathbb{0}^A$ is helpful to remind ourselves about the type parameter `A` used e.g., by the disjunctive value $\mathbb{0}^A + y^B$ in the body of `toRight[A, B]`. Without this type annotation, $\mathbb{0} + y^B$ means

a value of type `Either[A, B]` where the parameter A is left unspecified and should be determined by matching the types of other expressions.

In the notation $0 + y^B$, we use the symbol 0 rather than an ordinary zero (0), to avoid suggesting that 0 is a value of type 0. The void type 0 has no values, unlike the `Unit` type, 1 , which has a value denoted by 1 in the code notation.

Example 5.3.2.3 Verify the type equivalence $A \times 1 \cong A$.

Solution The corresponding Scala types are the tuple `(A, Unit)` and the type `A`. We need to implement functions $f_1 : \forall A. A \times 1 \rightarrow A$ and $f_2 : \forall A. A \rightarrow A \times 1$ and to demonstrate that they are inverses of each other. The Scala code for these functions is

```
def f1[A]: ((A, Unit)) => A = { case (a, ()) => a }
def f2[A]: A => (A, Unit) = { a => (a, ()) }
```

Let us first write a proof by reasoning directly with Scala code:

```
(f1 andThen f2)((a,())) == f2(f1((a,()))) == f2(a) == (a,())
(f2 andThen f1)(a) == f1(f2(a)) == f1((a,())) = a
```

Now let us write a proof in the code notation. The codes of f_1 and f_2 are

$$f_1 = a^A \times 1 \rightarrow a \quad , \\ f_2 = a^A \rightarrow a \times 1 \quad ,$$

where we denoted by 1 the value `()` of the `Unit` type. We find

$$(f_1 \circ f_2)(a^A \times 1) = f_2(f_1(a \times 1)) = f_2(a) = a \times 1 \quad , \\ (f_2 \circ f_1)(a^A) = f_1(f_2(a)) = f_1(a \times 1) = a \quad .$$

This shows that both compositions are identity functions. Another way of writing the proof is by computing the function compositions symbolically, without applying to a value a^A ,

$$f_1 \circ f_2 = (a \times 1 \rightarrow a) \circ (a \rightarrow a \times 1) = (a \times 1 \rightarrow a \times 1) = \text{id}^{A \times 1} \quad , \\ f_2 \circ f_1 = (a \rightarrow a \times 1) \circ (a \times 1 \rightarrow a) = (a \rightarrow a) = \text{id}^A \quad .$$

Example 5.3.2.4 Verify the type equivalence $A + B \cong B + A$.

Solution The corresponding Scala types are `Either[A, B]` and `Either[B, A]`. We use pattern matching to implement the functions required for the type equivalence:

```
def f1[A, B]: Either[A, B] => Either[B, A] = {
  case Left(a)    => Right(a) // No other choice here.
  case Right(b)   => Left(b)  // No other choice here.
}
def f2[A, B]: Either[B, A] => Either[A, B] = f1[B, A]
```

by using only a given value `a:A`. The only way of doing that is by returning `Right(a)`.

It is clear from the code that the functions f_1 and f_2 are inverses of each other. To verify that rigorously, we need show that f_1 `andThen` f_2 is equal to an identity function. The function f_1 `andThen` f_2 applies f_2 to the result of f_1 . The code of f_1 contains two `case ...` lines, each returning a result. So, we need to apply f_2 separately in each line. Evaluate the code symbolically:

```
(f1 andThen f2) == {
  case Left(a)    => f2(Right(a))
  case Right(b)   => f2(Left(b))
} == {
  case Left(a)    => Left(a)
  case Right(b)   => Right(b)
}
```

The functions f_1 and f_2 are implemented by code that can be derived unambiguously from the type signatures. For instance, the line `case Left(a) => ...` is required to return a value of type `Either[B, A]`.

The result is a function of type `Either[A, B] => Either[A, B]` that does not change its argument; so it is equal to the identity function.

Let us now write the function `f1` in the code notation and perform the same derivation. We will also develop a useful notation for functions operating on disjunctive types.

The pattern matching construction in the Scala code of `f1` contains a pair of functions with types `A => Either[B, A]` and `B => Either[B, A]`. One of these functions is chosen depending on whether the argument of `f1` has type $A + \mathbb{0}$ or $\mathbb{0} + B$. So, we may write the code of `f1` as

$$f1 \triangleq x^{A+B} \rightarrow \begin{cases} \text{if } x = a^{A+\mathbb{0}} + \mathbb{0}^{B+A} & : \mathbb{0}^B + a^A \\ \text{if } x = \mathbb{0}^{A+\mathbb{0}} + b^{B+A} & : b^B + \mathbb{0}^A \end{cases}$$

Since both the argument and the result of `f1` are disjunctive types with 2 parts each, it is convenient to write the code of `f1` as a 2×2 matrix that maps the input parts to the output parts:

```
def f1[A, B]: Either[A, B] => Either[B, A] = {
  case Left(a)    => Right(a)
  case Right(b)   => Left(b)
}
```

$$f1 \triangleq \begin{array}{c|cc} & B & A \\ \hline A & \mathbb{0} & a^A \rightarrow a \\ B & b^B \rightarrow b & \mathbb{0} \end{array} .$$

The rows of the matrix correspond to the `case` rows in the Scala code; there is one row for each part of the disjunctive type of the argument. The columns of the matrix correspond to the parts of the disjunctive type of the result. The double line marks the input types of the functions.

The code of `f2` is written similarly; let us rename arguments for clarity:

```
def f2[A, B]: Either[B, A] => Either[A, B] = {
  case Left(y)    => Right(y)
  case Right(x)   => Left(x)
}
```

$$f2 \triangleq \begin{array}{c|cc} & A & B \\ \hline B & \mathbb{0} & y^B \rightarrow y \\ A & x^A \rightarrow x & \mathbb{0} \end{array} .$$

The forward composition $f1 \circ f2$ is computed by the standard rules of row-by-column matrix multiplication.⁵ Any terms containing $\mathbb{0}$ are omitted, and the remaining functions are composed:

$$\begin{aligned} f1 \circ f2 &= \begin{array}{c|cc} & B & A \\ \hline A & \mathbb{0} & a^A \rightarrow a \\ B & b^B \rightarrow b & \mathbb{0} \end{array} \circ \begin{array}{c|cc} & A & B \\ \hline B & \mathbb{0} & y^B \rightarrow y \\ A & x^A \rightarrow x & \mathbb{0} \end{array} \\ \text{matrix multiplication : } &= \begin{array}{c|cc} & A & B \\ \hline A & (a^A \rightarrow a) \circ (x^A \rightarrow x) & \mathbb{0} \\ B & \mathbb{0} & (b^B \rightarrow b) \circ (y^B \rightarrow y) \end{array} \\ \text{function composition : } &= \begin{array}{c|cc} & A & B \\ \hline A & \text{id} & \mathbb{0} \\ B & \mathbb{0} & \text{id} \end{array} = \text{id}^{A+B \rightarrow A+B} . \end{aligned}$$

Several features of the matrix notation are helpful in such calculations. The parts of the code of `f1` are automatically composed with the corresponding parts of the code of `f2`. To check that the types match in the function composition, we just need to compare the types in the output row $\begin{array}{c|cc} & B & A \end{array}$ of

`f1` with the input column $\begin{array}{c|cc} B \\ A \end{array}$ of `f2`. Once we verified that all types match, we may omit the type

⁵https://en.wikipedia.org/wiki/Matrix_multiplication

annotations and write the same derivation more concisely as

$$\begin{aligned}
 f_1 \circ f_2 &= \left\| \begin{array}{cc} \mathbb{0} & a:A \rightarrow a \\ b:B \rightarrow b & \mathbb{0} \end{array} \right\| \left\| \begin{array}{cc} \mathbb{0} & y:B \rightarrow y \\ x:A \rightarrow x & \mathbb{0} \end{array} \right\| \\
 \text{matrix multiplication :} &= \left\| \begin{array}{cc} (a:A \rightarrow a) \circ (x:A \rightarrow x) & \mathbb{0} \\ \mathbb{0} & (b:B \rightarrow b) \circ (y:B \rightarrow y) \end{array} \right\| \\
 \text{function composition :} &= \left\| \begin{array}{cc} \text{id} & \mathbb{0} \\ \mathbb{0} & \text{id} \end{array} \right\| = \text{id} \quad .
 \end{aligned}$$

The identity function is represented by the diagonal matrix $\left\| \begin{array}{cc} \text{id} & \mathbb{0} \\ \mathbb{0} & \text{id} \end{array} \right\|$.

Exercise 5.3.2.5 Verify the type equivalence $A \times B \cong B \times A$.

Exercise 5.3.2.6 Verify the type equivalence $(A + B) + C \cong A + (B + C)$. Since Section 5.3.1 verified the equivalences $(A + B) + C \cong A + (B + C)$ and $(A \times B) \times C \cong A \times (B \times C)$, we may write $A + B + C$ and $A \times B \times C$ without any parentheses.

Exercise 5.3.2.7 Verify the type equivalence

$$(A + B) \times (A + B) = A \times A + 2 \times A \times B + B \times B \quad ,$$

where 2 denotes the `Boolean` type (defined as $2 \triangleq \mathbb{1} + \mathbb{1}$).

5.3.3 Type cardinalities and type equivalence

To understand why type equivalences are related to arithmetic identities, consider the question of how many different values a given type can have.

Begin by counting the number of distinct values for simple types. For example, the `Unit` type has only one distinct value; the type `Nothing` has zero values; the `Boolean` type has two distinct values, `true` and `false`; and the type `Int` has 2^{32} distinct values.

It is more difficult to count the number of distinct values in a type such as `String`, which is equivalent to a list of unknown length, `List[Char]`. However, each computer's memory is limited, so there will exist a maximum length for values of type `String`, and so the total number of possible different strings will be finite (at least, for any given computer).

For a given type A , let us denote by $|A|$ the number of distinct values of type A . The number $|A|$ is called the **cardinality** of type A ; this is the same as the number of elements in the set of all values of type A . Since any computer's memory is finite, and since we may assume that we are already working with the largest possible computer, then there will be *finitely* many different values of a given type A that can exist in the computer. So, we may assume that $|A|$ is always a finite integer value. This assumption will simplify our reasoning. We will not actually need to compute the precise number of, say, all the different possible strings; it is sufficient to know that the set of all strings is finite, so that we can denote its cardinality by $|\text{String}|$.

The next step is to consider the cardinality of types such as $A \times B$ and $A + B$. If the types A and B have cardinalities $|A|$ and $|B|$, it follows that the set of all distinct pairs (a, b) has $|A| \times |B|$ elements. So the cardinality of the type $A \times B$ is equal to the (arithmetic) product of the cardinalities of A and B . The set of all pairs

$$\{(a, b) : a \in A, b \in B\}$$

is also known as the **Cartesian product** of sets A and B , and is denoted by $A \times B$. For this reason, the tuple type is also called the **product type**. Accordingly, the type notation adopts the symbol \times for the product type.

The set of all distinct values of the type $A + B$, i.e., of the Scala type `Either[A, B]`, is a disjoint union of the set of values of the form `Left(a)` and the set of values of the form `Right(b)`. It is clear that the cardinalities of these sets are equal to $|A|$ and $|B|$ respectively. So the cardinality of the type `Either[A, B]` is equal to $|A| + |B|$. For this reason, disjunctive types such as `Either[A, B]` are also called **sum types**, and the type notation adopts the symbol $+$ for these types.

We can write our conclusions as

$$|A \times B| = |A| \times |B| \quad , \\ |A + B| = |A| + |B| \quad .$$

The type notation, $A \times B$ for pairs and $A + B$ for `Either[A, B]`, translates directly into type cardinalities.

The last step is to notice that two types can be equivalent, $P \cong Q$, only if their cardinalities are equal, $|P| = |Q|$. When the cardinalities are not equal, $|P| \neq |Q|$, it will be impossible to have a one-to-one correspondence between the sets of values of type P and values of type Q . So it will be impossible to convert values from type P to type Q and back without loss of information.

We conclude that types are equivalent when a logical identity *and* an arithmetic identity hold.

The presence of both identities does not automatically guarantee a useful type equivalence. The fact that information in one type can be identically stored in another type does not necessarily mean that it is helpful to do so in a given application.

For example, the types `Option[Option[A]]` and `Either[Boolean, A]` are equivalent because both types contain $2 + |A|$ distinct values. The short notation for these types is $1 + 1 + A$ and $2 + A$ respectively (the type `Boolean` is denoted by 2 since it has only two distinct values).

One could easily write code to convert between these types without loss of information:

```
def f1[A]: Option[Option[A]] => Either[Boolean, A] = {
  case None          => Left(false) // Or maybe Left(true)?
  case Some(None)    => Left(true)
  case Some(Some(x)) => Right(x)
}

def f2[A]: Either[Boolean, A] => Option[Option[A]] = {
  case Left(false)   => None
  case Left(true)    => Some(None)
  case Right(x)      => Some(Some(x))
}
```

A sign of trouble is the presence of an arbitrary choice in this code. In `f1`, we could map `None` to `Left(false)` or to `Left(true)`, and adjust the rest of the code accordingly; the type equivalence would still hold. So, formally speaking, these types *are* equivalent, but there is no “natural” choice of the conversion functions `f1` and `f2` that will work correctly in all applications, because the meaning of these data types is application-dependent. This type equivalence is “accidental”.

Example 5.3.3.1 Are the types `Option[A]` and `Either[Unit, A]` equivalent? Check whether the corresponding logic identity and arithmetic identity hold.

Solution Begin by writing the given types in the type notation: `Option[A]` is written as $1 + A$, and `Either[Unit, A]` is written also as $1 + A$. The notation already indicates that the types are equivalent. But let us verify explicitly that the type notation is not misleading here.

To establish type equivalence, we need to implement two fully parametric functions

```
def f1[A]: Option[A] => Either[Unit, A] = ???
def f2[A]: Either[Unit, A] => Option[A] = ???
```

such that $f_1 \circ f_2 = \text{id}$ and $f_2 \circ f_1 = \text{id}$. It is straightforward to implement `f1` and `f2`:

```
def f1[A]: Option[A] => Either[Unit, A] = {
  case None          => Left(())
  case Some(x)       => Right(x)
}
def f2[A]: Either[Unit, A] => Option[A] = {
```

```

  case Left()    => None
  case Right(x) => Some(x)
}

```

The code clearly shows that f_1 and f_2 are inverses of each other; this verifies the type equivalence.

The logic identity is $True \vee A = True \vee A$ and holds trivially. It remains to check the arithmetic identity, which relates the number of distinct values of types `Option[A]` and `Either[Unit, A]`. Assume that the number of distinct values of type `A` is $|A|$. Any possible value of type `Option[A]` must be either `None` or `Some(x)`, where `x` is a value of type `A`. So the number of distinct values of type `Option[A]` is $1 + |A|$. All possible values of type `Either[Unit, A]` are of the form `Left()` or `Right(x)`, where `x` is a value of type `A`. So the number of distinct values of type `Either[Unit, A]` is $1 + |A|$. We see that the arithmetic identity holds: the types `Option[A]` and `Either[Unit, A]` have equally many distinct values.

This example shows that the type notation is helpful for reasoning about type equivalences. The solution was found immediately when we wrote the type notation, $1 + A$, for the given types.

5.3.4 Type equivalence involving function types

Until now, we have looked at product types and disjunctive types. Let us now consider type constructions involving function types.

Consider two types A and B , whose cardinalities are known as $|A|$ and $|B|$. What is the cardinality of the set of all maps between given sets A and B ? In other words, how many distinct values does the function type $A \rightarrow B$ have? A function $f: A \Rightarrow B$ needs to select a value of type B for each possible value of type A . Therefore, the number of different functions $f: A \Rightarrow B$ is $|B|^{|A|}$ (the **numeric exponent**, $|B|$ to the power $|A|$).

For the types $A = B = \text{Int}$, we have $|A| = |B| = 2^{32}$, and so the estimate will give

$$|A \rightarrow B| = (2^{32})^{(2^{32})} = 2^{32 \times 2^{32}} = 2^{2^{37}} \approx 10^{4.1 \times 10^{10}} .$$

In fact, most of these functions will map integers to integers in a complicated (and practically useless) way and will be impossible to implement on a realistic computer because their code will be much longer than the available memory. So, the number of practically implementable functions of type $A \rightarrow B$ is often much smaller than $|B|^{|A|}$. Nevertheless, the estimate $|B|^{|A|}$ is useful since it shows the number of distinct functions that are possible in principle.

Let us now look for logic identities and arithmetic identities involving function types. Table 5.6 lists the available identities and the corresponding type equivalences. (In the last column, we defined $a \triangleq |A|$, $b \triangleq |B|$, and $c \triangleq |C|$ for brevity.)

It is notable that no logic identity is available for the formula $\alpha \Rightarrow (\beta \vee \gamma)$, and correspondingly no type equivalence is available for the type expression $A \rightarrow B + C$ (although there is an identity for $A \rightarrow B \times C$). The presence of type expressions of the form $A \rightarrow B + C$ makes type reasoning more complicated because they cannot be transformed into equivalent formulas with simpler parts.

We will now prove some of the type identities in Table 5.6.

Example 5.3.4.1 Verify the type equivalence $1 \rightarrow A \cong A$.

Solution Recall that the type notation $1 \rightarrow A$ means the Scala function type `Unit => A`. There is only one value of type `Unit`, so the choice of a function of the type `Unit => A` is the same as the choice of a value of type `A`. Thus, the type $1 \rightarrow A$ has $|A|$ distinct values, so the arithmetic identity holds.

To verify the type equivalence explicitly, we need to implement two functions

```

def f1[A]: (Unit => A) => A = ???
def f2[A]: A => Unit => A = ???

```

The first function needs to produce a value of type `A`, given an argument of the function type `Unit => A`. The only possibility is to apply that function to the value of type `Unit`; we can always produce that value as `()`:

```

def f1[A]: (Unit => A) => A = (h: Unit => A) => h(())

```

Logical identity (if holds)	Type equivalence	Arithmetic identity
$(True \Rightarrow \alpha) = \alpha$	$\mathbb{1} \rightarrow A \cong A$	$a^1 = a$
$(False \Rightarrow \alpha) = True$	$\mathbb{0} \rightarrow A \cong \mathbb{1}$	$a^0 = 1$
$(\alpha \Rightarrow True) = True$	$A \rightarrow \mathbb{1} \cong \mathbb{1}$	$1^a = 1$
$(\alpha \Rightarrow False) \neq False$	$A \rightarrow \mathbb{0} \not\cong \mathbb{0}$	$0^a \neq 0$
$(\alpha \vee \beta) \Rightarrow \gamma = (\alpha \Rightarrow \gamma) \wedge (\beta \Rightarrow \gamma)$	$A + B \rightarrow C \cong (A \rightarrow C) \times (B \rightarrow C)$	$c^{a+b} = c^a \times c^b$
$(\alpha \wedge \beta) \Rightarrow \gamma = \alpha \Rightarrow (\beta \Rightarrow \gamma)$	$A \times B \rightarrow C \cong A \rightarrow B \rightarrow C$	$c^{a \times b} = (c^b)^a$
$\alpha \Rightarrow (\beta \wedge \gamma) = (\alpha \Rightarrow \beta) \wedge (\alpha \Rightarrow \gamma)$	$A \rightarrow B \times C \cong (A \rightarrow B) \times (A \rightarrow C)$	$(b \times c)^a = b^a \times c^a$

Table 5.6: Logical identities with implication, and the corresponding type equivalences and arithmetic identities.

Implementing $f2$ is straightforward; we can just discard the `Unit` argument:

```
def f2[A]: A => Unit => A = (x: A) => _ => x
```

It remains to show that the functions $f1$ and $f2$ are inverses of each other. Let us perform the proof using Scala code and then using the code notation.

Writing Scala code, compute $f1(f2(x))$ for an arbitrary $x:A$. Substituting the code, we get

```
f1(f2(x)) == f1(_ => x) == (_ => x)(()) == x
```

Now compute $f2(f1(h))$ for arbitrary $h: Unit => A$ in Scala code:

```
f2(f1(h)) == f2(h(())) == { _ => h(()) }
```

How can we show that the function $\{ _ => h(())\}$ is equal to h ? Whenever we apply equal functions to equal arguments, they return equal results. In our case, the argument of h is of type `Unit`, so we only need to verify that the result of applying h to the value `()` is the same as the result of applying $\{ _ => h(())\}$ to `()`. In other words, we need to apply both sides to an additional argument `()`:

```
f2(f1(h)) == { _ => h(()) } () == h()
```

This completes the proof.

For comparison, let us show the same proof in the code notation. The functions $f1$ and $f2$ are

$$f1 \triangleq h: \mathbb{1} \rightarrow A \rightarrow h(1) ,$$

$$f2 \triangleq x:A \rightarrow 1 \rightarrow x .$$

Now write the function compositions in both directions:

$$\text{expect to equal id : } f1 \circ f2 = (h: \mathbb{1} \rightarrow A \rightarrow h(1)) \circ (x:A \rightarrow 1 \rightarrow x)$$

$$\text{compute composition : } = h: \mathbb{1} \rightarrow A \rightarrow 1 \rightarrow h(1)$$

$$\text{note that } 1 \rightarrow h(1) \text{ is the same as } h : = (h: \mathbb{1} \rightarrow A \rightarrow h) = \text{id} .$$

$$\text{expect to equal id : } f2 \circ f1 = (x:A \rightarrow 1 \rightarrow x) \circ (h: \mathbb{1} \rightarrow A \rightarrow h(1))$$

$$\text{compute composition : } = x:A \rightarrow (1 \rightarrow x)(1)$$

$$\text{apply function : } = (x:A \rightarrow x) = \text{id} .$$

The type $\mathbb{1} \rightarrow A$ is equivalent to the type A , but these types are not the same. The most important difference between these types is that a value of type A is available immediately, while a value of type $\mathbb{1} \rightarrow A$ is a function that still needs to be applied to an argument (of type $\mathbb{1}$) before a value

of type A is obtained. The type $\mathbb{1} \rightarrow A$ may represent an “on-call” value of type A ; that is, a value computed on demand every time. (See Section 2.6.3 for more details about “on-call” values.)

The void type $\mathbb{0}$ needs special reasoning, as the next examples show:

Example 5.3.4.2 Verify the type equivalence $\mathbb{0} \rightarrow A \cong \mathbb{1}$.

Solution What could be a function $f: \mathbb{0} \rightarrow A$ from the type $\mathbb{0}$ to a type A ? Since there exist no values of type $\mathbb{0}$, the function f will never be applied to any arguments and so *does not need* to compute any actual values of type A . So, f is a function whose body is “empty”; or at least it does not need to contain any expressions of type A . In Scala, such a function can be written as

```
def absurd[A]: Nothing => A = { ??? }
```

This code will compile without type errors. An equivalent code is

```
def absurd[A]: Nothing => A = { x => ??? }
```

The symbol `???` is defined in the Scala library and represents code that is “not implemented”. Trying to evaluate this symbol will produce an error:

```
scala> ???
scala.NotImplementedError: an implementation is missing
  scala.Predef$.qmark$qmark$qmark(Predef.scala:288)
```

Since the function `absurd` can never be applied to an argument, this error will never happen. So, we can pretend that the result value (which will never be computed) has any required type, e.g., type A .

Let us now verify that there exists *only one* distinct function of type $\mathbb{0} \rightarrow A$. Take any two functions of that type, $f: \mathbb{0} \rightarrow A$ and $g: \mathbb{0} \rightarrow A$. Are they different? The only way of showing that f and g are different is by producing a value $x: \mathbb{0}$ such that $f(x) \neq g(x)$. But there are *no values* of type $\mathbb{0}$, and for this reason, we will never be able to find the required value $x: \mathbb{0}$. It follows that any two functions f and g of type $\mathbb{0} \rightarrow A$ are equal, $f = g$. In other words, there exists only one distinct value of type $\mathbb{0} \rightarrow A$; i.e., the cardinality of the type $\mathbb{0} \rightarrow A$ is 1. So, the type $\mathbb{0} \rightarrow A$ is equivalent to the type $\mathbb{1}$.

Example 5.3.4.3 Show that $A \rightarrow \mathbb{0} \not\cong \mathbb{0}$ and $A \rightarrow \mathbb{0} \not\cong \mathbb{1}$.

Solution To prove that two types are *not* equivalent, it is sufficient to show that their type cardinalities are different. Let us determine the cardinality of the type $A \rightarrow \mathbb{0}$, assuming that the cardinality of A is known. We note that a function of type, say, $\text{Int} \rightarrow \mathbb{0}$ is impossible to implement. (If we had such a function $f: \text{Int} \rightarrow \mathbb{0}$, we could evaluate, say, $x \triangleq f(123)$ and obtain a value x of type $\mathbb{0}$, which is impossible by definition of the type $\mathbb{0}$. It follows that $|\text{Int} \rightarrow \mathbb{0}| = 0$. However, Example 5.3.4.2 shows that $\mathbb{0} \rightarrow \mathbb{0}$ has cardinality 1. So, the cardinality $|A \rightarrow \mathbb{0}| = 1$ if the type A is itself $\mathbb{0}$ but $|A \rightarrow \mathbb{0}| = 0$ for all other types A . We conclude that the type $A \rightarrow \mathbb{0}$ is not equivalent to $\mathbb{0}$ or $\mathbb{1}$ for all A . The type $A \rightarrow \mathbb{0}$ is equivalent to $\mathbb{0}$ only for non-void types A .

Example 5.3.4.4 Verify the type equivalence $A \rightarrow \mathbb{1} \cong \mathbb{1}$.

Solution There is only one fully parametric function that returns $\mathbb{1}$:

```
def f[A]: A => Unit = { _ => () }
```

The function f cannot use its argument of type A since nothing is known about that type. So the code of f *must* discard its argument and return the fixed value $\mathbb{0}$ of type `Unit`. In the code notation, this function is written as

$$f: A \rightarrow \mathbb{1} \triangleq (_ \rightarrow \mathbb{1}) \quad .$$

We can show that there exist only *one* distinct function of type $A \rightarrow \mathbb{1}$ (that is, the type $A \rightarrow \mathbb{1}$ has cardinality 1). Assume that f and g are two such functions, and try to find a value $x: A$ such that $f(x) \neq g(x)$. We cannot find any such x because $f(x) = \mathbb{1}$ and $g(x) = \mathbb{1}$ for all x . So, any two functions f and g of type $A \rightarrow \mathbb{1}$ must be equal to each other. The cardinality of the type $A \rightarrow \mathbb{1}$ is 1.

Any type having cardinality 1 is equivalent to the `Unit` type, $\mathbb{1}$. So $A \rightarrow \mathbb{1} \cong \mathbb{1}$.

Example 5.3.4.5 Verify the type equivalence

$$A + B \rightarrow C \cong (A \rightarrow C) \times (B \rightarrow C) \quad .$$

Solution Begin by implementing two functions with type signatures

```
def f1[A,B,C]: (Either[A, B] => C) => (A => C, B => C) = ???
def f2[A,B,C]: ((A => C, B => C)) => Either[A, B] => C = ???
```

The code can be derived unambiguously from the type signatures. For the first function, we need to produce a pair of functions of type $(A \Rightarrow C, B \Rightarrow C)$. Can we produce the first part of that pair? Computing a function of type $A \Rightarrow C$ means that we need to produce a value of type C given an arbitrary value $a:A$. The available data is a function of type $\text{Either}[A, B] \Rightarrow C$ called, say, h . We can apply that function to $\text{Left}(a)$ and obtain a value of type C as required. So, a function of type $A \Rightarrow C$ is computed as $a \Rightarrow h(\text{Left}(a))$. Similarly, we produce a function of type $B \Rightarrow C$. The code is

```
def f1[A,B,C]: (Either[A, B] => C) => (A => C, B => C) =
  (h: Either[A, B] => C) => (a => h(Left(a)), b => h(Right(b)))
```

A code notation for this function is

$$f_1 : (A + B \rightarrow C) \rightarrow (A \rightarrow C) \times (B \rightarrow C) , \\ f_1 \triangleq h^{A+B \rightarrow C} \rightarrow (a^A \rightarrow h(a + \mathbb{0}^B)) \times (b^B \rightarrow h(\mathbb{0}^A + b)) .$$

For the function f_2 , we need to apply pattern matching to both curried arguments and then return a value of type C . This can be achieved in only one way:

```
def f2[A,B,C]: ((A => C, B => C)) => Either[A, B] => C = { case (f, g) =>
  {
    case Left(a) => f(a)
    case Right(b) => g(b)
  }
}
```

A code notation for this function can be written as

$$f_2 : (A \rightarrow C) \times (B \rightarrow C) \rightarrow A + B \rightarrow C , \\ f_2 \triangleq f^{A \rightarrow C} \times g^{B \rightarrow C} \rightarrow \begin{array}{c|c|c} & & C \\ \hline & A & a \rightarrow f(a) \\ \hline & B & b \rightarrow g(b) \end{array} .$$

The matrix in the last line has only one column because the result type, C , is not known to be a disjunctive type. We may simplify the functions, e.g., $a \rightarrow f(a)$ into f , and write

$$f_2 \triangleq f^{A \rightarrow C} \times g^{B \rightarrow C} \rightarrow \begin{array}{c|c} & C \\ \hline A & f \\ \hline B & g \end{array} .$$

It remains to verify that $f_1 \circ f_2 = \text{id}$ and $f_2 \circ f_1 = \text{id}$. To compute $f_1 \circ f_2$, we write (omitting types)

$$f_1 \circ f_2 = (h \rightarrow (a \rightarrow h(a + \mathbb{0})) \times (b \rightarrow h(\mathbb{0} + b))) \circ \left(f \times g \rightarrow \begin{array}{c} f \\ g \end{array} \right)$$

compute composition : $= h \rightarrow \begin{array}{c} a \rightarrow h(a + \mathbb{0}) \\ b \rightarrow h(\mathbb{0} + b) \end{array} .$

To proceed, we need to simplify the expressions $h(a + \mathbb{0})$ and $h(\mathbb{0} + b)$. We rewrite the argument h (an

arbitrary function of type $A + B \rightarrow C$ in the matrix notation:

$$h \triangleq \begin{vmatrix} & & C \\ \hline A & a \rightarrow p(a) \\ B & b \rightarrow q(b) \end{vmatrix} = \begin{vmatrix} & & C \\ \hline A & p \\ B & q \end{vmatrix},$$

where $p: A \rightarrow C$ and $q: B \rightarrow C$ are new arbitrary functions. Since we already checked the types, we can omit all type annotations and write h as

$$h \triangleq \begin{vmatrix} p \\ q \end{vmatrix}.$$

To evaluate expressions such as $h(a + \mathbb{0})$ and $h(\mathbb{0} + b)$, we need to use one of the rows of the column matrix h . The correct row will be selected *automatically* by the rules of matrix multiplication if we place a row vector to the left of the matrix and use the convention of omitting terms containing $\mathbb{0}$:

$$\begin{vmatrix} a & \mathbb{0} \end{vmatrix} \triangleright \begin{vmatrix} p \\ q \end{vmatrix} = a \triangleright p, \quad \begin{vmatrix} \mathbb{0} & b \end{vmatrix} \triangleright \begin{vmatrix} p \\ q \end{vmatrix} = b \triangleright q.$$

Here we used the symbol \triangleright to separate an argument from a function when the argument is written to the *left* of the function. The symbol \triangleright (pronounced “pipe”) is defined by $x \triangleright f \triangleq f(x)$. In Scala, this operation is available as `x.pipe(f)` as of Scala 2.13.

We can write values of disjunctive types, such as $a + \mathbb{0}$, as row vectors $\begin{vmatrix} a & \mathbb{0} \end{vmatrix}$:

$$h(a + \mathbb{0}) = (a + \mathbb{0}) \triangleright h = \begin{vmatrix} a & \mathbb{0} \end{vmatrix} \triangleright h. \quad (5.20)$$

With these notations, we can compute further, by omitting terms containing $\mathbb{0}$:

$$\begin{aligned} h(a + \mathbb{0}) &= \begin{vmatrix} a & \mathbb{0} \end{vmatrix} \triangleright \begin{vmatrix} p \\ q \end{vmatrix} = a \triangleright p = p(a), \\ h(\mathbb{0} + b) &= \begin{vmatrix} \mathbb{0} & b \end{vmatrix} \triangleright \begin{vmatrix} p \\ q \end{vmatrix} = b \triangleright q = q(b). \end{aligned}$$

Now we can complete the proof of $f_1 \circ f_2 = \text{id}$:

$$\begin{aligned} f_1 \circ f_2 &= h \rightarrow \begin{vmatrix} a \rightarrow h(a + \mathbb{0}) \\ b \rightarrow h(\mathbb{0} + b) \end{vmatrix} \\ \text{previous equations :} &= \begin{vmatrix} p \\ q \end{vmatrix} \rightarrow \begin{vmatrix} a \rightarrow p(a) \\ b \rightarrow q(b) \end{vmatrix} \\ \text{simplify functions :} &= \begin{vmatrix} p \\ q \end{vmatrix} \rightarrow \begin{vmatrix} p \\ q \end{vmatrix} = \text{id}. \end{aligned}$$

To prove that $f_2 \circ f_1 = \text{id}$, use the notation (5.20):

$$\begin{aligned} f_2 \circ f_1 &= \left(f \times g \rightarrow \begin{vmatrix} f \\ g \end{vmatrix} \right) \circ (h \rightarrow (a \rightarrow (a + \mathbb{0}) \triangleright h) \times (b \rightarrow (\mathbb{0} + b) \triangleright h)) \\ \text{compute composition :} &= f \times g \rightarrow (a \rightarrow \begin{vmatrix} a & \mathbb{0} \end{vmatrix} \triangleright \begin{vmatrix} f \\ g \end{vmatrix}) \times (b \rightarrow \begin{vmatrix} \mathbb{0} & b \end{vmatrix} \triangleright \begin{vmatrix} f \\ g \end{vmatrix}) \end{aligned}$$

matrix notation : $= f \times g \rightarrow (a \rightarrow a \triangleright f) \times (b \rightarrow b \triangleright g)$
definition of \triangleright : $= f \times g \rightarrow (a \rightarrow f(a)) \times (b \rightarrow g(b))$
simplify functions : $= (f \times g \rightarrow f \times g) = \text{id}$.

In this way, we have proved that f_1 and f_2 are mutual inverses. The proofs appear long because we took time to motivate and introduce new notation for applying matrices to row vectors. Given this notation, the proof for $f_1 \circ f_2 = \text{id}$ can be written as

$$\begin{aligned}
 f_1 \circ f_2 &= (h \rightarrow (a \rightarrow (a + \mathbb{0}) \triangleright h) \times (b \rightarrow (\mathbb{0} + b) \triangleright h)) \circ \left(f \times g \rightarrow \begin{vmatrix} f \\ g \end{vmatrix} \right) \\
 \text{compute composition :} \quad &= h \rightarrow \begin{vmatrix} a \rightarrow \begin{vmatrix} a & \mathbb{0} \\ \mathbb{0} & b \end{vmatrix} \triangleright h \\ b \rightarrow \begin{vmatrix} \mathbb{0} & b \end{vmatrix} \triangleright h \end{vmatrix} = \begin{vmatrix} p \\ q \end{vmatrix} \rightarrow \begin{vmatrix} a \rightarrow \begin{vmatrix} a & \mathbb{0} \\ \mathbb{0} & b \end{vmatrix} \triangleright \begin{vmatrix} p \\ q \end{vmatrix} \\ &\quad b \rightarrow \begin{vmatrix} \mathbb{0} & b \end{vmatrix} \triangleright \begin{vmatrix} p \\ q \end{vmatrix} \end{vmatrix} \\
 \text{matrix notation :} \quad &= \begin{vmatrix} p \\ q \end{vmatrix} \rightarrow \begin{vmatrix} a \rightarrow a \triangleright p \\ b \rightarrow b \triangleright q \end{vmatrix} = \begin{vmatrix} p \\ q \end{vmatrix} \rightarrow \begin{vmatrix} p \\ q \end{vmatrix} = \text{id} .
 \end{aligned}$$

Proofs in the code notation are shorter than in Scala syntax since many names and keywords (such as `Left`, `Right`, `case`, `match`, etc.) are omitted. From now on, we will prefer to use the code notation in proofs, keeping in mind that one can always convert code notation to Scala and back.

Note that the function arrow (\rightarrow) binds weaker than the pipe operation (\triangleright), so the code notation $x \rightarrow y \triangleright z$ means $x \rightarrow (y \triangleright z)$. We will review the code notation more thoroughly in Chapter 7.

Example 5.3.4.6 Verify the type equivalence

$$A \times B \rightarrow C \cong A \rightarrow B \rightarrow C .$$

Solution Begin by implementing the two functions

```
def f1[A,B,C]: (((A, B)) => C) => A => B => C = ???
def f2[A,B,C]: (A => B => C) => ((A, B)) => C = ???
```

The Scala code can be derived from the type signatures unambiguously:

```
def f1[A,B,C]: (((A, B)) => C) => A => B => C = g => a => b => g((a, b))
def f2[A,B,C]: (A => B => C) => ((A, B)) => C = h => { case (a, b) => h(a)(b) }
```

Write these functions in the code notation:

$$\begin{aligned}
 f_1 &= g: A \times B \rightarrow C \rightarrow a:A \rightarrow b:B \rightarrow g(a \times b) , \\
 f_2 &= h: A \rightarrow B \rightarrow C \rightarrow (a \times b):A \times B \rightarrow h(a)(b) .
 \end{aligned}$$

We denote by $(a \times b):A \times B$ the argument of type (A, B) with pattern matching implied. This notation allows us to write shorter code formulas involving tupled arguments.

Compute the function composition $f_1 \circ f_2$:

$$\begin{aligned}
 f_1 \circ f_2 &= (g \rightarrow a \rightarrow b \rightarrow g(a \times b)) \circ (h \rightarrow a \times b \rightarrow h(a)(b)) \\
 \text{substitute } h = a \rightarrow b \rightarrow g(a \times b) : &= g \rightarrow a \times b \rightarrow g(a \times b) \\
 \text{simplify function :} &= (g \rightarrow g) = \text{id} .
 \end{aligned}$$

Compute the function composition $f_2 \circ f_1$:

$$\begin{aligned}
 f_2 \circ f_1 &= (h \rightarrow a \times b \rightarrow h(a)(b)) \circ (g \rightarrow a \rightarrow b \rightarrow g(a \times b)) \\
 \text{substitute } g = a \times b \rightarrow h(a)(b) : &= h \rightarrow a \rightarrow b \rightarrow h(a)(b) \\
 \text{simplify function } b \rightarrow h(a)(b) : &= h \rightarrow a \rightarrow h(a) \\
 \text{simplify function } a \rightarrow h(a) \text{ to } h : &= (h \rightarrow h) = \text{id} \quad .
 \end{aligned}$$

Exercise 5.3.4.7 Verify the type equivalence $(A \rightarrow B \times C) \cong (A \rightarrow B) \times (A \rightarrow C)$.

5.4 Summary

What tasks can we perform now?

- Convert a fully parametric type signature into a logical formula to:
 - Decide whether the type signature can be implemented in code.
 - If possible, derive the code using the CH correspondence.
- Use the type notation (Table 5.1) for reasoning about types to:
 - Decide type equivalence using the rules in Tables 5.5–5.6.
 - Simplify type expressions before writing code.
- Use the matrix notation and the pipe notation to write code that works on disjunctive types.

What tasks *cannot* be performed with these tools?

- Automatically generate code for a *recursive* function. (The CH correspondence is based on propositional logic, which cannot describe recursion. Accordingly, recursion is absent from the eight code constructions of Section 5.2.3.)
- Automatically generate code satisfying a property (e.g., isomorphism). We may generate the code, but it is not guaranteed that properties will hold. The workaround is to verify the required properties manually, after deriving the code.
- Express complicated conditions (e.g., “array is sorted”) in a type signature. This can be done using **dependent types** (i.e., types that depend on run-time values in an arbitrary way) — an advanced technique for which Scala has limited support. (Programming languages such as Coq, Agda, and Idris support full dependent types.)
- Generate code using type constructors with known properties (e.g., the `map` method).

As an example of using type constructors with properties, consider this type signature:

```
def q[A]: Array[A] => (A => Option[B]) => Array[Option[B]]
```

Can we generate the code of this function from its type signature? We know that the Scala library defines a `map` method on the `Array` type constructor, so the implementation of `q` is simple,

```
def q[A]: Array[A] => (A => Option[B]) => Array[Option[B]] = { arr => f => arr.map(f) }
```

However, it is hard to create an *algorithm* that can derive this implementation automatically from the type signature of `q` via the Curry-Howard correspondence. The algorithm would have to convert the type signature of `q` into the logical formula

$$CH(\text{Array}^A) \Rightarrow CH(A \rightarrow \text{Opt}^B) \Rightarrow CH(\text{Array}^{\text{Opt}^B}) . \quad (5.21)$$

To derive an implementation, the algorithm would need to use the available `map` method for `Array`. That method has the type signature

$$\text{map} : \forall(A, B). \text{Array}^A \rightarrow (A \rightarrow B) \rightarrow \text{Array}^B \quad .$$

To derive the \mathcal{CH} -proposition (5.21), the algorithm will need to assume that the \mathcal{CH} -proposition

$$\mathcal{CH}(\forall(A, B). \text{Array}^A \rightarrow (A \rightarrow B) \rightarrow \text{Array}^B) \quad (5.22)$$

already holds, i.e., that Eq. (5.22) is one of the premises of a sequent to be proved. Reasoning about propositions such as Eq. (5.22) requires **first-order logic** — a logic whose proof rules can handle quantified types such as $\forall(A, B)$ *inside* premises. However, first-order logic is **undecidable**: no algorithm can guarantee finding a proof or showing the absence of a proof in all cases.

The constructive propositional logic (with the rules listed in Section 5.2.3) is **decidable**, i.e., it has an algorithm that either finds a proof or disproves any given formula. However, that logic cannot handle premises containing type quantifiers such as $\forall(A, B)$ *inside*, because all the available rules have the quantifiers placed *outside* the premises.

So, code for functions such as `q` can only be derived by trial and error, informed by intuition. This book will help functional programmers to acquire the necessary intuition and technique.

5.4.1 Solved examples

Example 5.4.1.1 Find the cardinality of the type `P = Option[Option[Boolean] => Boolean]`. Write `P` in the type notation and simplify to an equivalent type.

Solution Begin with the type `Option[Boolean]`, which can be either `None` or `Some(x)` with an `x:Boolean`. Since the type `Boolean` has 2 possible values, the type `Option[Boolean]` has 3 values:

$$|\text{Opt}^{\text{Boolean}}| = |\mathbb{1} + \text{Boolean}| = 1 + |\text{Boolean}| = 3 \quad .$$

In the type notation, `Boolean` is denoted by the symbol `2`, and the type `Option[Boolean]` by `1 + 2`. So, the type notation `1 + 2` is consistent with the cardinality 3 of that type,

$$|\mathbb{1} + \text{Boolean}| = |\mathbb{1} + 2| = 1 + 2 = 3 \quad .$$

The function type `Option[Boolean] => Boolean` is denoted by `1 + 2 → 2`. Its cardinality is computed as the arithmetic power

$$|\text{Opt}^{\text{Boolean} \rightarrow \text{Boolean}}| = |\mathbb{1} + 2 \rightarrow 2| = |2|^{|\mathbb{1} + 2|} = 2^3 = 8 \quad .$$

Finally, the we write `P` in the type notation as `P = 1 + (1 + 2 → 2)` and find

$$|P| = |\mathbb{1} + (\mathbb{1} + 2 \rightarrow 2)| = 1 + |\mathbb{1} + 2 \rightarrow 2| = 1 + 8 = 9 \quad .$$

Example 5.4.1.2 Implement a Scala type `P[A]` for the type notation

$$P^A \triangleq 1 + A + \text{Int} \times A + (\text{String} \rightarrow A) \quad .$$

Solution To translate type notation into Scala code, begin by defining the disjunctive types as case classes (with names chosen for convenience). In this case, P^A is a disjunctive type with four parts, so we will need four case classes:

```
sealed trait P[A]
final case class P1[A](???) extends P[A]
final case class P2[A](???) extends P[A]
final case class P3[A](???) extends P[A]
final case class P4[A](???) extends P[A]
```

Each of the case classes represents one part of the disjunctive type. Now we write the contents for each of the case classes, in order to implement the data in each of the disjunctive parts:

```
sealed trait P[A]
final case class P1[A]() extends P[A]
final case class P2[A](x: A) extends P[A]
final case class P3[A](n: Int, x: A) extends P[A]
final case class P4[A](f: String => A) extends P[A]
```

Example 5.4.1.3 Find an equivalent disjunctive type for the type $P = (\text{Either}[A, B], \text{Either}[C, D])$.

Solution Begin by writing the given type in the type notation. The tuple becomes the product type, and `Either` becomes the disjunctive (or “sum”) type:

$$P \triangleq (A + B) \times (C + D) .$$

We can use the usual rules of arithmetic to expand brackets in this type expression and to obtain an equivalent type:

$$P \cong A \times C + A \times D + B \times C + B \times D .$$

This is a disjunctive type having 4 parts.

Example 5.4.1.4 Show that the following type equivalences do *not* hold: $A + A \not\cong A$ and $A \times A \not\cong A$, although the corresponding logical identities hold.

Solution Note that the arithmetic equalities do not hold, $A + A \neq A$ and $A \times A \neq A$. This already indicates that the types are not equivalent. To build further intuition, consider that a value of type $A + A$ (in Scala, `Either[A, A]`) is a `Left(a)` or a `Right(a)` for some $a:A$. In the code notation, it is either $a:A + 0$ or $0 + a:A$. So, a value of type $A + A$ contains a value of type A with the additional information about whether it is the first or the second part of the disjunctive type. We cannot represent that information in a single value of type A .

Similarly, a value of type $A \times A$ contains two (possibly different) values of type A , which cannot be represented by a single value of type A without loss of information.

However, the corresponding logical identities $\alpha \vee \alpha \Rightarrow \alpha$ and $\alpha \wedge \alpha \Rightarrow \alpha$ hold. To see that, we could derive the four formulas

$$\begin{aligned} \alpha \vee \alpha &\Rightarrow \alpha , & \alpha &\Rightarrow \alpha \vee \alpha , \\ \alpha \wedge \alpha &\Rightarrow \alpha , & \alpha &\Rightarrow \alpha \wedge \alpha , \end{aligned}$$

using the proof rules of Section 5.2.3. Alternatively, we may use the CH correspondence and show that the type signatures

$$\begin{aligned} \forall A. A + A &\rightarrow A , & \forall A. A &\rightarrow A + A , \\ \forall A. A \times A &\rightarrow A , & \forall A. A &\rightarrow A \times A \end{aligned}$$

can be implemented via fully parametric functions. For a programmer, it is easier to write code than to guess the correct sequence of proof rules. For the first pair of type signatures, we find

```
def f1[A]: Either[A, A] => A = {
  case Left(a)  => a  // No other choice here.
  case Right(a) => a  // No other choice here.
}
def f2[A]: A => Either[A, A] = { a => Left(a) } // Can be also Right(a).
```

The presence of an arbitrary choice, to return `Left(a)` or `Right(a)`, is a warning sign showing that additional information is required to create a value of type `Either[A, A]`. This is precisely the information present in the type $A + A$ but missing in the type A .

The code notation for these functions is

$$f_1 \triangleq \begin{array}{c|c} & A \\ \hline A & a \rightarrow a \\ A & a \rightarrow a \end{array} = \begin{array}{c|c} & A \\ \hline A & \text{id} \\ A & \text{id} \end{array} , \quad f_2 \triangleq a:A \rightarrow a + 0:A = \begin{array}{c|cc} & A & A \\ \hline A & a \rightarrow a & 0 \end{array} = \begin{array}{c|cc} & A & A \\ \hline A & \text{id} & 0 \end{array} .$$

The composition of these functions is not equal to identity:

$$f_1 \circ f_2 = \left\| \begin{array}{c} \text{id} \\ \text{id} \end{array} \right\| \circ \left\| \begin{array}{c} \text{id} \\ \text{id} \end{array} \right\| = \left\| \begin{array}{cc} \text{id} & 0 \\ \text{id} & 0 \end{array} \right\| \neq \text{id} = \left\| \begin{array}{cc} \text{id} & 0 \\ 0 & \text{id} \end{array} \right\| .$$

For the second pair of type signatures, the code is

```
def f1[A]: ((A, A)) => A = { case (a1, a2) => a1 } // Can be also 'a2'.
cef f2[A]: A => (A, A) = { a => (a, a) } // No other choice here.
```

It is clear that the first function loses information when it returns a_1 and discards a_2 (or vice versa).

The code notation for the functions f_1 and f_2 is

$$f_1 \triangleq a_1^A \times a_2^A \rightarrow a_1 = \pi_1^{A \times A \rightarrow A} , \quad f_2 \triangleq a^A \rightarrow a \times a = \Delta^{A \rightarrow A \times A} .$$

Computing the compositions of these functions, we find that $f_2 \circ f_1 = \text{id}$ while $f_1 \circ f_2 \neq \text{id}$:

$$\begin{aligned} f_1 \circ f_2 &= (a_1 \times a_2 \rightarrow a_1) \circ (a \rightarrow a \times a) \\ &= (a_1 \times a_2 \rightarrow a_1 \times a_1) \neq \text{id} = (a_1 \times a_2 \rightarrow a_1 \times a_2) . \end{aligned}$$

We have implemented all four type signatures as fully parametric functions, which shows that the corresponding logical formulas are all true (i.e., can be derived using the proof rules). However, the functions cannot be inverses of each other. So, the type equivalences do not hold.

Example 5.4.1.5 Show that $((A \wedge B) \Rightarrow C) \neq (A \Rightarrow C) \vee (B \Rightarrow C)$ in the constructive logic, but the equality holds in Boolean logic. (This is another example where Boolean reasoning about types fails.)

Solution Begin by rewriting the logical equality as two implications,

$$\begin{aligned} (A \wedge B \Rightarrow C) &\Rightarrow (A \Rightarrow C) \vee (B \Rightarrow C) , \\ ((A \Rightarrow C) \vee (B \Rightarrow C)) &\Rightarrow ((A \wedge B) \Rightarrow C) . \end{aligned}$$

It is sufficient to show that one of these implications is incorrect. Rather than looking for a proof tree in the constructive logic (which would be difficult, since we would need to demonstrate that *no* proof tree exists), let us use the CH correspondence. So the task is to implement fully parametric functions with the type signatures

$$\begin{aligned} (A \times B \rightarrow C) &\rightarrow (A \rightarrow C) + (B \rightarrow C) , \\ (A \rightarrow C) + (B \rightarrow C) &\rightarrow A \times B \rightarrow C . \end{aligned}$$

For the first type signature, the Scala code is

```
def f1[A,B,C]: (((A, B)) => C) => Either[A => C, B => C] = { k => ??? }
```

We are required to return either a `Left(g)` with $g: A \Rightarrow C$, or a `Right(h)` with $h: B \Rightarrow C$. The only given data is a function k of type $A \times B \rightarrow C$, so the decision of whether to return a `Left` or a `Right` must be hard-coded in the function f_1 independently of k . Can we produce a function g of type $A \Rightarrow C$? Given a value of type A , we would need to return a value of type C . The only way to obtain a value of type C is by applying k to some arguments. But to apply k , we need a value of type B , which we do not have. So we cannot produce a $g: A \Rightarrow C$. Similarly, we cannot produce a function h of type $B \Rightarrow C$.

To repeat the same argument in the type notation: Obtaining a value of type $(A \rightarrow C) + (B \rightarrow C)$ means to compute either $g^{A \rightarrow C} + \emptyset$ or $\emptyset + h^{B \rightarrow C}$. This decision must be hard-coded since the only data is a function $k^{A \times B \rightarrow C}$. We can compute $g^{A \rightarrow C}$ only by partially applying $k^{A \times B \rightarrow C}$ to a value of type B . However, we have no values of type B . Similarly, we cannot get an $h^{B \rightarrow C}$.

The inverse type signature *can* be implemented:

```
def f2[A,B,C]: Either[A=>C, B=>C] => ((A,B)) => C = {
  case Left(g)  => { case (a, b)  => g(a) }
  case Right(h) => { case (a, b)  => h(b) }
}
```

$$f_2 \triangleq \left| \begin{array}{c|c} & A \times B \rightarrow C \\ \hline A \rightarrow C & g: A \rightarrow C \rightarrow a \times b \rightarrow g(a) \\ B \rightarrow C & h: B \rightarrow C \rightarrow a \times b \rightarrow h(b) \end{array} \right| .$$

Let us now show that the logical identity

$$((\alpha \wedge \beta) \Rightarrow \gamma) = ((\alpha \Rightarrow \gamma) \vee (\beta \Rightarrow \gamma)) \quad (5.23)$$

holds in Boolean logic. A straightforward calculation is to simplify the Boolean expression using Eq. (5.5), which only holds in Boolean logic (but not in the constructive logic). We find

$$\begin{aligned} \text{left-hand side of Eq. (5.23)} : & (\alpha \wedge \beta) \Rightarrow \gamma \\ \text{use Eq. (5.5)} : & = \neg(\alpha \wedge \beta) \vee \gamma \\ \text{use de Morgan's law} : & = \neg\alpha \vee \neg\beta \vee \gamma . \\ \text{right-hand side of Eq. (5.23)} : & (\alpha \Rightarrow \gamma) \vee (\beta \Rightarrow \gamma) \\ \text{use Eq. (5.5)} : & = \neg\alpha \vee \gamma \vee \neg\beta \vee \gamma \\ \text{use identity } \gamma \vee \gamma = \gamma : & = \neg\alpha \vee \neg\beta \vee \gamma . \end{aligned}$$

Both sides of Eq. (5.23) are equal to the same formula, $\neg\alpha \vee \neg\beta \vee \gamma$, so the identity holds.

This calculation does not work in the constructive logic because its proof rules can derive neither the Boolean formula (5.5) nor the **law of de Morgan**, $\neg(\alpha \wedge \beta) = (\neg\alpha \vee \neg\beta)$.

Another way of proving the Boolean identity (5.23) is to enumerate all possible truth values for the variables α , β , and γ . The left-hand side, $(\alpha \wedge \beta) \Rightarrow \gamma$, can be *False* only if $\alpha \wedge \beta = \text{True}$ (that is, both α and β are *True*) and $\gamma = \text{False}$; for all other truth values of α , β , and γ , the formula $(\alpha \wedge \beta) \Rightarrow \gamma$ is *True*. Let us determine when the right-hand side, $(\alpha \Rightarrow \gamma) \vee (\beta \Rightarrow \gamma)$, can be *False*. This can happen only if both parts of the disjunction are *False*; that means $\alpha = \text{True}$, $\beta = \text{True}$, and $\gamma = \text{False}$. So, the two sides of the identity (5.23) are both *True* or both *False* with any choice of truth values of α , β , and γ . In Boolean logic, this is sufficient to prove the identity (5.23).

It is important to note that the proof rules of the constructive logic are not equivalent to checking whether some propositions are *True* or *False*. A general form of this statement was proved by Kurt Gödel in 1932.⁶ In this sense, constructive logic does not imply that every proposition is either *True* or *False*. This is not intuitive and requires getting used to.

The following example shows how to use the identities from Tables 5.5–5.6 to derive type equivalence for complicated type expressions, without need for proofs.

Example 5.4.1.6 Use known rules to verify the type equivalences:

- (a) $A \times (A + 1) \times (A + 1 + 1) \cong A \times (1 + 1 + A \times (1 + 1 + 1 + A))$.
- (b) $1 + A + B \rightarrow 1 \times B \cong (B \rightarrow B) \times (A \rightarrow B) \times B$.

Solution (a) We can expand brackets in the type expression as in arithmetic,

$$\begin{aligned} A \times (A + 1) &\cong A \times A + A \times 1 \cong A \times A + A , \\ A \times (A + 1) \times (A + 1 + 1) &\cong (A \times A + A) \times (A + 1 + 1) \\ &\cong A \times A \times A + A \times A + A \times A \times (1 + 1) + A \times (1 + 1) \\ &\cong A \times A \times A + A \times A \times (1 + 1 + 1) + A \times (1 + 1) . \end{aligned}$$

The result looks like a polynomial in A , which we can now rearrange into the required form:

$$A \times A \times A + A \times A \times (1 + 1 + 1) + A \times (1 + 1) \cong A \times (1 + 1 + A \times (1 + 1 + 1 + A)) .$$

⁶See plato.stanford.edu/entries/intuitionistic-logic-development/

(b) Keep in mind that the conventions of the type notation make the function arrow (\rightarrow) group weaker than other type operations. So, the type expression $1 + A + B \rightarrow 1 \times B$ means a function from $1 + A + B$ to $1 \times B$.

Begin by using the rule $1 \times B \cong B$ to obtain $1 + A + B \rightarrow B$. Now we use the rule

$$A + B \rightarrow C \cong (A \rightarrow C) \times (B \rightarrow C)$$

and derive the equivalence

$$1 + A + B \rightarrow B \cong (1 \rightarrow B) \times (A \rightarrow B) \times (B \rightarrow B) \quad .$$

Finally, we note that $1 \rightarrow B \cong B$ and that the type product is commutative, so we can rearrange the last type expression into the required form:

$$B \times (A \rightarrow B) \times (B \rightarrow B) \cong (B \rightarrow B) \times (A \rightarrow B) \times B \quad .$$

Example 5.4.1.7 Denote $\text{Read}^{E,T} \triangleq E \rightarrow T$ and implement fully parametric functions with types $A \rightarrow \text{Read}^{E,A}$ and $\text{Read}^{E,A} \rightarrow (A \rightarrow B) \rightarrow \text{Read}^{E,B}$.

Solution Begin by defining a type alias for the type constructor $\text{Read}^{E,T}$:

```
type Read[E, T] = E => T
```

The first type signature has only one implementation:

```
def p[E, A]: A => Read[E, A] = { x => _ => x }
```

We *must* discard the argument of type E ; we cannot use it for computing a value of type A given $x:A$.

The second type signature has three type parameters. It is the curried version of the function `map`:

```
def map[E, A, B]: Read[E, A] => (A => B) => Read[E, B] = ???
```

Expanding the type alias, we see that the two curried arguments are functions of types $E \rightarrow A$ and $A \rightarrow B$. The forward composition of these functions is a function of type $E \rightarrow B$, or $\text{Read}^{E,B}$, which is exactly what we are required to return. So the code can be written as

```
def map[E, A, B]: (E => A) => (A => B) => E => B = { r => f => r andThen f }
```

If we did not notice this shortcut, we would reason differently: We are required to compute a value of type B given *three* curried arguments $r:E \rightarrow A$, $f:A \rightarrow B$, and $e:E$. Write this requirement as

$$\text{map} \triangleq r:E \rightarrow A \rightarrow f:A \rightarrow B \rightarrow e:E \rightarrow ???:B \quad ,$$

The symbol $???:B$ is called a **typed hole**; it stands for a value that we are still figuring out how to compute, but whose type is already known. Typed holes are supported in Scala by an experimental compiler plugin.⁷ The plugin will print the known information about the typed hole.

To fill the typed hole $???:B$, we need a value of type B . Since no arguments have type B , the only way of getting a value of type B is to apply $f:A \rightarrow B$ to some value of type A . So we write

$$\text{map} \triangleq r:E \rightarrow A \rightarrow f:A \rightarrow B \rightarrow e:E \rightarrow f(???:A) \quad .$$

The only way of getting an A is to apply r to a value of type E ,

$$\text{map} \triangleq r:E \rightarrow A \rightarrow f:A \rightarrow B \rightarrow e:E \rightarrow f(r(???:E)) \quad .$$

We have exactly one value of type E , namely $e:E$. So the code must be

$$\text{map}^{E,A,B} \triangleq r:E \rightarrow A \rightarrow f:A \rightarrow B \rightarrow e:E \rightarrow f(r(e)) \quad .$$

Translate this to the Scala syntax:

```
def map[E, A, B]: (E => A) => (A => B) => E => B = { r => f => e => f(r(e)) }
```

We may now notice that the expression $e \rightarrow f(r(e))$ is a function composition $r \circ f$ applied to e , and simplify the code accordingly.

⁷<https://github.com/cb372/scala-typed-holes>

Example 5.4.1.8 Show that the type signature $\text{Read}[A, T] \Rightarrow (A \Rightarrow B) \Rightarrow \text{Read}[B, T]$ cannot be implemented as a fully parametric function.

Solution Expand the type signature and try implementing this function:

```
def m[A, B, T] : (A => T) => (A => B) => B => T = { r => f => b => ??? }
```

Given values $r:A \rightarrow T$, $f:A \rightarrow B$, and $b:B$, we need to compute a value of type T :

$$m = r:A \rightarrow T \rightarrow f:A \rightarrow B \rightarrow b:B \rightarrow ???^T .$$

The only way of getting a value of type T is to apply r to some value of type A ,

$$m = r:A \rightarrow T \rightarrow f:A \rightarrow B \rightarrow b:B \rightarrow r(???^A) .$$

However, we do not have any values of type A . We have a function $f:A \rightarrow B$ that *consumes* values of type A , and we cannot use f to produce any values of type A . So we seem to be unable to fill the typed hole $???^A$ and implement the function m .

In order to verify that m is unimplementable, we need to prove that the logical formula

$$\forall(\alpha, \beta, \tau). (\alpha \Rightarrow \tau) \Rightarrow (\alpha \Rightarrow \beta) \Rightarrow (\beta \Rightarrow \tau) \quad (5.24)$$

is not true in the constructive logic. We could use the `curryhoward` library for that:

```
@ def m[A, B, T] : (A => T) => (A => B) => B => T = implement
cmd1.sc:1: type (A => T) => (A => B) => B => T cannot be implemented
def m[A, B, T] : (A => T) => (A => B) => B => T = implement
^
Compilation Failed
```

Another way is to check whether this formula is true in Boolean logic. A formula that holds in constructive logic will always hold in Boolean logic, because all rules shown in Section 5.2.3 preserve Boolean truth values (see Section 5.5.4 for a proof). It follows that any formula that fails to hold in Boolean logic will also not hold in constructive logic.

It is relatively easy to check whether a given Boolean formula is always equal to *True*. Simplifying Eq. (5.24) with the rules of Boolean logic, we find

$$\begin{aligned} & (\alpha \Rightarrow \tau) \Rightarrow (\alpha \Rightarrow \beta) \Rightarrow (\beta \Rightarrow \tau) \\ \text{use Eq. (5.5)} : &= \neg(\alpha \Rightarrow \tau) \vee \neg(\alpha \Rightarrow \beta) \vee (\beta \Rightarrow \tau) \\ \text{use Eq. (5.5)} : &= \neg(\neg\alpha \vee \tau) \vee \neg(\neg\alpha \vee \beta) \vee (\neg\beta \vee \tau) \\ \text{use de Morgan's law} : &= (\alpha \wedge \neg\tau) \vee (\alpha \wedge \neg\beta) \vee \neg\beta \vee \tau \\ \text{use identity } (p \wedge q) \vee q = q : &= (\alpha \wedge \neg\tau) \vee \neg\beta \vee \tau \\ \text{use identity } (p \wedge \neg q) \vee q = p \vee q : &= \alpha \vee \neg\beta \vee \tau . \end{aligned}$$

This formula is not identically *True*: it is *False* when $\alpha = \tau = \text{False}$ and $\beta = \text{True}$. So, Eq. (5.24) is not true in Boolean logic, and thus is not true in constructive logic. By the CH correspondence, we conclude that the type signature of m cannot be implemented by a fully parametric function.

Example 5.4.1.9 Define the type constructor $P^A \triangleq \mathbb{1} + A + A$ and implement `map` for it,

$$\text{map}^{A,B} : P^A \rightarrow (A \rightarrow B) \rightarrow P^B .$$

To check that `map` preserves information, verify the law $\text{map}(p)(x \Rightarrow x) == p$ for all $p: P[A]$.

Solution It is implied that `map` should be fully parametric and information-preserving. Begin by defining a Scala type constructor for the notation $P^A \triangleq \mathbb{1} + A + A$:

```
sealed trait P[A]
final case class P1[A]() extends P[A]
final case class P2[A](x: A) extends P[A]
final case class P3[A](x: A) extends P[A]
```

Now we can write code to implement the required type signature. Each time we have several choices of an implementation, we will choose to preserve information as much as possible.

```
def map[A, B]: P[A] => (A => B) => P[B] =  
  p => f => p match {  
    case P1() => P1() // No other choice.  
    case P2(x) => ???  
    case P3(x) => ???  
  }
```

In the case $P2(x)$, we are required to produce a value of type P^B from a value $x:A$ and a function $f:A \rightarrow B$. Since P^B is a disjunctive type with three parts, we can produce a value of type P^B in three different ways: $P1()$, $P2(\dots)$, and $P3(\dots)$. If we return $P1()$, we will lose the information about the value x . If we return $P3(\dots)$, we will preserve the information about

x but lose the information that the input value was a $P2$ rather than a $P3$. By returning $P2(\dots)$ in that scope, we preserve the entire input information.

The value under $P2(\dots)$ must be of type B , and the only way of getting a value of type B is to apply f to x . So, we return $P2(f(x))$.

Similarly, in the case $P3(x)$, we should return $P3(f(x))$. The final code of `map` is

```
def map[A, B]: P[A] => (A => B) => P[B] = p => f => p match {  
  case P1() => P1() // No other choice here.  
  case P2(x) => P2(f(x)) // Preserve information.  
  case P3(x) => P3(f(x)) // Preserve information.  
}
```

To verify the given law, we first write a matrix notation for `map`:

$$\text{map}^{A,B} \triangleq p^{:\mathbb{1}+A+A} \rightarrow f^{:A \rightarrow B} \rightarrow p \triangleright \begin{array}{c|ccc} & \mathbb{1} & B & B \\ \hline \mathbb{1} & \text{id} & 0 & 0 \\ A & 0 & f & 0 \\ A & 0 & 0 & f \end{array} .$$

The required law is written as an equation (called the **identity law**)

$$\text{map}(p)(\text{id}) = p .$$

Substituting the code notation for `map`, we verify the law:

$$\begin{aligned} \text{expect to equal } p : & \quad \text{map}(p)(\text{id}) \\ \text{apply map}() \text{ to arguments : } & \quad = p \triangleright \begin{array}{ccc} \text{id} & 0 & 0 \\ 0 & \text{id} & 0 \\ 0 & 0 & \text{id} \end{array} \\ \text{identity function in matrix notation : } & \quad = p \triangleright \text{id} \\ \text{---notation : } & \quad = \text{id}(p) = p . \end{aligned}$$

Example 5.4.1.10 Implement `map` and `flatMap` for `Either[L, R]`, applied to the type parameter `L`.

Solution For a type constructor, say, P^A , the standard type signatures for `map` and `flatMap` are

$$\begin{aligned} \text{map} : P^A \rightarrow (A \rightarrow B) \rightarrow P^B , \\ \text{flatMap} : P^A \rightarrow (A \rightarrow P^B) \rightarrow P^B . \end{aligned}$$

If a type constructor has more than one type parameter, e.g., $P^{A,S,T}$, one can define the functions `map` and `flatMap` applied to a chosen type parameter. For example, when applied to the type parameter A , the type signatures are

$$\begin{aligned} \text{map} : P^{A,S,T} \rightarrow (A \rightarrow B) \rightarrow P^{B,S,T} , \\ \text{flatMap} : P^{A,S,T} \rightarrow (A \rightarrow P^{B,S,T}) \rightarrow P^{B,S,T} . \end{aligned}$$

Being “applied to the type parameter A ” means that the other type parameters S, T in $P^{A,S,T}$ remain fixed while the type parameter A is replaced by B in the type signatures of `map` and `flatMap`.

For the type `Either[L, R]` (i.e., $L + R$), we keep the type parameter R fixed while L is replaced by M . So we obtain the type signatures

$$\begin{aligned} \text{map} : L + R \rightarrow (L \rightarrow M) \rightarrow M + R & , \\ \text{flatMap} : L + R \rightarrow (L \rightarrow M + R) \rightarrow M + R & . \end{aligned}$$

Implementing these functions is straightforward:

```
def map[L,M,R]: Either[L, R] => (L => M) => Either[M, R] = e => f => e match {
  case Left(x) => Left(f(x))
  case Right(y) => Right(y)
}
def flatMap[L,M,R]: Either[L, R] => (L => Either[M, R]) => Either[M, R] = e => f => e match {
  case Left(x) => f(x)
  case Right(y) => Right(y)
}
```

The code notation for these functions is

$$\begin{aligned} \text{map} &\triangleq e^{L+R} \rightarrow f^{L \rightarrow M} \rightarrow e \triangleright \begin{array}{c|cc} & M & R \\ \hline L & f & \mathbb{0} \\ R & \mathbb{0} & \text{id} \end{array} , \\ \text{flatMap} &\triangleq e^{L+R} \rightarrow f^{L \rightarrow M+R} \rightarrow e \triangleright \begin{array}{c|c} & M+R \\ \hline L & f \\ R & y^{:R} \rightarrow \mathbb{0}^{:M} + y \end{array} . \end{aligned}$$

Note that we cannot split f into the M and R columns since $f(x:L)$ could return either part of the disjunctive type $M + R$.

Example 5.4.1.11* Define a type constructor $\text{State}^{S,A} \equiv S \rightarrow A \times S$ and implement the functions:

- (a) $\text{pure}^{S,A} : A \rightarrow \text{State}^{S,A}$.
- (b) $\text{map}^{S,A,B} : \text{State}^{S,A} \rightarrow (A \rightarrow B) \rightarrow \text{State}^{S,B}$.
- (c) $\text{flatMap}^{S,A,B} : \text{State}^{S,A} \rightarrow (A \rightarrow \text{State}^{S,B}) \rightarrow \text{State}^{S,B}$.

Solution It is assumed that all functions must be fully parametric and preserve as much information as possible. We define the type alias

```
type State[S, A] = S => (A, S)
```

- (a) The type signature is $A \rightarrow S \rightarrow A \times S$, and there is only one implementation,

```
def pure[S, A]: A => State[S, A] = a => s => (a, s)
```

In the code notation, this is written as

$$\text{pu}^{S,A} \triangleq a^{:A} \rightarrow s^{:S} \rightarrow a \times s .$$

- (b) The type signature is

$$\text{map}^{S,A,B} : (S \rightarrow A \times S) \rightarrow (A \rightarrow B) \rightarrow S \rightarrow B \times S .$$

Begin writing a Scala implementation:

```
def map[S, A, B]: State[S, A] => (A => B) => State[S, B] = { t => f => s => ??? }
```

We need to compute a value of $B \times S$ from the curried arguments $t^{S \rightarrow A \times S}$, $f^{A \rightarrow B}$, and s^{S} . We begin writing the code of `map` using a typed hole,

$$\text{map} \triangleq t^{S \rightarrow A \times S} \rightarrow f^{A \rightarrow B} \rightarrow s^{S} \rightarrow ???^{B \times S} \quad .$$

The only way of getting a value of type B is by applying f to a value of type A :

$$\text{map} \triangleq t^{S \rightarrow A \times S} \rightarrow f^{A \rightarrow B} \rightarrow s^{S} \rightarrow f(???^A) \times ???^S \quad .$$

The only possibility of filling the typed hole $???^A$ is to apply t to a value of type S ; we already have such a value, s^{S} . Computing $t(s)$ yields a pair of type $A \times S$, from which we may take the first part (of type A) to fill the typed hole $???^A$. The second part of the pair is a value of type S that we may use to fill the second typed hole, $???^S$. So the Scala code is

```
1 def map[S, A, B]: State[S, A] => (A => B) => State[S, B] = {
2   t => f => s =>
3     val (a, s2) = t(s)
4     (f(a), s2)    // We could also return '(f(a), s)' here.
5 }
```

($f(a)$, s) in line 4, we will have discarded the computed value $s2$, which is a loss of information.

To write the code notation for `map`, we need to destructure the pair that $t(s)$ returns. We can write explicit destructuring code like this:

$$\text{map} \triangleq t^{S \rightarrow A \times S} \rightarrow f^{A \rightarrow B} \rightarrow s^{S} \rightarrow (a^A \times s_2^S \rightarrow f(a) \times s_2)(t(s)) \quad .$$

If we temporarily denote by q the destructuring function

$$q \triangleq (a^A \times s_2^S \rightarrow f(a) \times s_2) \quad ,$$

we will notice that the expression $s \rightarrow q(t(s))$ is a function composition applied to s . So, we rewrite $s \rightarrow q(t(s))$ as the composition $t \circ q$ and obtain shorter code,

$$\text{map} \triangleq t^{S \rightarrow A \times S} \rightarrow f^{A \rightarrow B} \rightarrow t \circ (a^A \times s^S \rightarrow f(a) \times s) \quad .$$

Shorter formulas are often easier to reason about in derivations (although not necessarily easier to read when converted to program code).

(c) The required type signature is

$$\text{flatMap}^{S, A, B} : (S \rightarrow A \times S) \rightarrow (A \rightarrow S \rightarrow B \times S) \rightarrow S \rightarrow B \times S \quad .$$

We perform code reasoning with typed holes:

$$\text{flatMap} \triangleq t^{S \rightarrow A \times S} \rightarrow f^{A \rightarrow S \rightarrow B \times S} \rightarrow s^{S} \rightarrow ???^B \times ???^S \quad .$$

To fill $???^B$, we need to apply f to some arguments, since f is the only function that returns any values of type B . A saturated application of f will yield a value of type $B \times S$, which we can return without change:

$$\text{flatMap} \triangleq t^{S \rightarrow A \times S} \rightarrow f^{A \rightarrow S \rightarrow B \times S} \rightarrow s^{S} \rightarrow f(???^A)(???^S) \quad .$$

To fill the new typed holes, we need to apply t to an argument of type S . We have only one given value s^{S} of type S , so we must compute $t(s)$ and destructure it:

$$\text{flatMap} \triangleq t^{S \rightarrow A \times S} \rightarrow f^{A \rightarrow S \rightarrow B \times S} \rightarrow s^{S} \rightarrow (a \times s_2 \rightarrow f(a)(s_2))(t(s)) \quad .$$

Translating this notation into Scala code, we obtain

Why not return the original value s in the tuple $B \times S$, instead of the new value $s2$? The reason is that we would like to preserve information as much as possible. If we return

```

def flatMap[S, A, B]: State[S, A] => (A => State[S, B]) => State[S, B] = {
  t => f => s =>
    val (a, s2) = t(s)
    f(a)(s2)           // We could also return `f(a)(s)` here.
}

```

As before, in order to preserve information, we choose not to discard the computed value s_2 .

The code notation for this `flatMap` can be simplified to

$$\text{flatMap} \triangleq t: S \rightarrow A \times S \rightarrow f: A \rightarrow S \rightarrow B \times S \rightarrow t \circ (a \times s \rightarrow f(a)(s)) \quad .$$

5.4.2 Exercises

Exercise 5.4.2.1 Find the cardinality of the type `P = Option[Boolean] => Option[Boolean]`. Show that `P` is equivalent to `Option[Boolean] => Boolean`, and that the equivalence is accidental and not “natural”.

Exercise 5.4.2.2 Verify the type equivalences $A + A \cong 2 \times A$ and $A \times A \cong 2 \rightarrow A$, where 2 denotes the Boolean type.

Exercise 5.4.2.3 Show that $A \Rightarrow (B \vee C) \neq (A \Rightarrow B) \wedge (A \Rightarrow C)$ in constructive and Boolean logic.

Exercise 5.4.2.4 Use known rules to verify the type equivalences without need for proofs:

- (a) $(A + B) \times (A \rightarrow B) \cong A \times (A \rightarrow B) + (\mathbb{1} + A \rightarrow B)$.
- (b) $(A \times (\mathbb{1} + A) \rightarrow B) \cong (A \rightarrow B) \times (A \rightarrow A \rightarrow B)$.
- (c) $A \rightarrow (\mathbb{1} + B) \rightarrow C \times D \cong (A \rightarrow C) \times (A \rightarrow D) \times (A \times B \rightarrow C) \times (A \times B \rightarrow D)$.

Exercise 5.4.2.5 Write the type notation for `Either[(A, Int), Either[(A, Char), (A, Float)]]`. Transform this type into an equivalent type of the form $A \times (...)$.

Exercise 5.4.2.6 Define a type $\text{OptE}^{T,A} \triangleq \mathbb{1} + T + A$ and implement information-preserving `map` and `flatMap` for it, applied to the type parameter A . Get the same result using the equivalent type $(\mathbb{1} + A) + T$, i.e., `Either[Option[A], T]`. The required type signatures are

$$\begin{aligned} \text{map}^{A,B,T} : \text{OptE}^{T,A} \rightarrow (A \rightarrow B) \rightarrow \text{OptE}^{T,B} \quad , \\ \text{flatMap}^{A,B,T} : \text{OptE}^{T,A} \rightarrow (A \rightarrow \text{OptE}^{T,B}) \rightarrow \text{OptE}^{T,B} \quad . \end{aligned}$$

Exercise 5.4.2.7 Implement the `map` function for `P[A]` (see Example 5.4.1.2). The required type signature is $P^A \rightarrow (A \rightarrow B) \rightarrow P^B$. Preserve information as much as possible.

Exercise 5.4.2.8 For the type constructor $Q^{T,A}$ defined in Exercise 5.1.4.1, define the `map` function, preserving information as much as possible,

$$\text{map}^{T,A,B} : Q^{T,A} \rightarrow (A \rightarrow B) \rightarrow Q^{T,B} \quad .$$

Exercise 5.4.2.9 Define a recursive type constructor Tr_3 as $\text{Tr}_3^A \triangleq \mathbb{1} + A \times A \times A \times \text{Tr}_3^A$ and implement the `map` function for it, with the standard type signature: $\text{map}^{A,B} : \text{Tr}_3^A \rightarrow (A \rightarrow B) \rightarrow \text{Tr}_3^B$.

Exercise 5.4.2.10 Implement fully parametric, information-preserving functions with the types:

- (a) $Z + A \times A \rightarrow (A \rightarrow B) \rightarrow Z + B \times B$.
- (b) $A + Z \rightarrow B + Z \rightarrow (A \rightarrow B \rightarrow C) \rightarrow C + Z$.
- (c) $\text{flatMap}^{E,A,B} : \text{Read}^{E,A} \rightarrow (A \rightarrow \text{Read}^{E,B}) \rightarrow \text{Read}^{E,B}$.
- (d) $\text{State}^{S,A} \rightarrow (S \times A \rightarrow B) \rightarrow \text{State}^{S,B}$.

Exercise 5.4.2.11* Denote $\text{Cont}^{R,T} \triangleq (T \rightarrow R) \rightarrow R$ and implement the functions:

- (a) $\text{map}^{R,T,U} : \text{Cont}^{R,T} \rightarrow (T \rightarrow U) \rightarrow \text{Cont}^{R,U}$.
- (b) $\text{flatMap}^{R,T,U} : \text{Cont}^{R,T} \rightarrow (T \rightarrow \text{Cont}^{R,U}) \rightarrow \text{Cont}^{R,U}$.

Exercise 5.4.2.12* Denote $\text{Sel}^{Z,T} \triangleq (T \rightarrow Z) \rightarrow T$ and implement the functions:

- (a) $\text{map}^{Z,A,B} : \text{Sel}^{Z,A} \rightarrow (A \rightarrow B) \rightarrow \text{Sel}^{Z,B}$.
- (b) $\text{flatMap}^{Z,A,B} : \text{Sel}^{Z,A} \rightarrow (A \rightarrow \text{Sel}^{Z,B}) \rightarrow \text{Sel}^{Z,B}$.

5.5 Discussion and further developments

5.5.1 Using the Curry-Howard correspondence for writing code

This chapter shows how the CH correspondence performs two practically important reasoning tasks: checking whether a type signature can be implemented as a fully parametric function, and determining whether two types are equivalent. The first task is accomplished by mapping type expressions into formulas in the constructive logic and by applying the proof rules of that logic. The second task is accomplished by mapping type expressions into *arithmetic* formulas and applying the ordinary rules of arithmetic.

Fully parametric functions can be often derived from their type signatures alone. It is useful for a programmer to know that certain type signatures, such as

```
def f[A, B] : A => (A => B) => B
```

have only one possible implementation, while other type signatures, such as

```
def g[A, B] : A => (B => A) => B
def h[A, B] : ((A => B) => A) => A
```

cannot be implemented as fully parametric functions.

Although tools such as the `curryhoward` library can sometimes derive code from types, it is more beneficial if a programmer is able to derive an implementation by hand, or to recognize quickly that an implementation is impossible. Exercises in this chapter help to build up the required technique and intuition. For instance, we have seen heuristics such as “values of parametric types cannot be constructed from scratch” and “one must hard-code the decision to return a chosen part of a disjunctive type”. These heuristics can be justified by the rigorous rules of proof (Section 5.2.3).

Throughout this chapter, we required all functions to be fully parametric. The reason is that the CH correspondence becomes informative only with parameterized types and with fully parametric functions. For concrete types, e.g., `Int`, one can always produce *some* value even with no previous data, so the proposition $\text{CH}(\text{Int})$ is always true within any code.

Consider the function `(x:Int) => x + 1`. Its type signature, `Int => Int`, is insufficient to specify the code of the function, because there are many different functions with the same type signature, such as `x => x - 1`, `x => x * 2`, etc. So, deriving code from the type signature `Int => Int` is not a meaningful task. Only a fully parametric type signature, such as $A \rightarrow (A \rightarrow B) \rightarrow B$, gives enough information for possibly deriving the function’s code. If we permit functions that are not fully parametric, we will not be able to reason about implementability of type signatures or about code derivation.

Information about the implementability of type signatures is given by logical formulas involving CH -propositions. Validity of a CH -proposition $\text{CH}(T)$ means that we can compute *some* value of the given type T but does not give any information about the properties of that value, such as whether it satisfies any additional laws. This is why type equivalence (which requires the laws of isomorphism) is not determined by an equivalence of logical formulas.

It is useful for programmers to be able to reason about types and transform type expressions to equivalent simpler types before starting to write code. The type notation introduced in this book is designed to help programmers to recognize patterns in type expressions and to reason about them more easily. We have shown that a type equivalence corresponds to *each* standard arithmetic identity such as $(a + b) + c = a + (b + c)$, $(a \times b) \times c = a \times (b \times c)$, $1 \times a = a$, $(a + b) \times c = a \times c + b \times c$, etc. So, we are allowed to transform and simplify types as if they were arithmetic expressions, e.g., to rewrite

$$1 \times (A + B) \times C + D \cong D + A \times C + B \times C .$$

The type notation makes this reasoning more intuitive (for people familiar with arithmetic).

These results apply to all type expressions built up using product types, disjunctive types (also called “sum” types because they correspond to arithmetic sums), and function types (also called “exponential” types because they correspond to arithmetic exponentials). Type expressions that contain only products and sum types may be called **polynomial**. Type expressions that also contain function types may be called **exponential-polynomial**.⁸ This book focuses on exponential-polynomial types because they are sufficient for almost all design patterns used in functional programming.

There are no type constructions corresponding to subtraction or division, so equations such as

$$(1-t) \times (1+t) = 1 - t \times t \quad \text{or} \quad \frac{t+t \times t}{t} = 1+t$$

do not directly yield any type equivalences. However, consider this well-known formula,

$$\frac{1}{1-t} = 1 + t + t^2 + t^3 + \dots + t^n + \dots .$$

At first sight, this formula appears to involve subtraction, division, and an infinite series, and thus cannot be directly translated into a type equivalence. However, the formula can be rewritten as

$$\frac{1}{1-t} \triangleq L(t) = 1 + t + t^2 + t^3 + \dots + t^n \times L(t) , \quad (5.25)$$

which is finite and only contains additions and multiplications. So, Eq. (5.25) can be translated into a type equivalence:

$$L^A \cong 1 + A + A \times A + A \times A \times A + \dots + \underbrace{A \times \dots \times A}_{n \text{ times}} \times L^A . \quad (5.26)$$

This type formula (with $n = 1$) is equivalent to a recursive definition of the type constructor `List`,

$$\text{List}^A \triangleq 1 + A \times \text{List}^A .$$

The type equivalence (5.26) suggests that we may view the recursive type `List` as an “infinite disjunction” describing lists of zero, one, etc. elements.

5.5.2 Implications for designing new programming languages

The functional programming paradigm assumes that programmers will use the six standard type constructions (Section 5.1.2) and the eight standard code constructions (Section 5.2.3). These constructions are foundational in the sense that they are used to express all design patterns of functional programming. A language that does not directly support some of these constructions cannot be considered a functional programming language.

A remarkable consequence of the CH correspondence is that the type system of any programming language (functional or not) is mapped into a *certain logic*, i.e., a system of logical operations and proof rules. A logical operation will correspond to each of the type constructions available in the programming language; a proof rule will correspond to each of the available code constructions. Functional programming languages that support all the standard type and code constructions — for instance, OCaml, Haskell, F#, Scala, Swift, etc., — will be mapped into the constructive logic with all standard logical operations available (*True*, *False*, disjunction, conjunction, and implication).

Languages such as C, C++, Java, C# are mapped into logics that do not have the disjunction operation or the constants *True* and *False*. In other words, these languages are mapped into *incomplete* logics where some theorems will not be provable. (If $\text{CH}(A)$ is true but not provable, a value of type

⁸Polynomial types are often called “algebraic data types”.

A is not directly computable by programs, although it could have been.) Incompleteness of the logic of types will make a programming language unable to express certain computations, e.g., directly handle data that belongs to a disjoint domain.

Languages such as Python, JavaScript, Ruby, Clojure have no type checking and so are mapped to *inconsistent* logics where any proposition can be derived — even propositions normally considered *False* may be derived from *True*. The CH correspondence will map such derivations to code that appears to compute a certain value (since the *CH*-proposition appears to be *True*) although that value is not actually available. In practice, such code *crashes* because the computed value has a wrong type, is “null”, or is a pointer to an invalid memory location.

None of these errors will happen in a programming language whose logic of types is complete and consistent, provided that types are checked at compile time.

So, the CH correspondence gives a mathematically justified procedure for designing type systems in new programming languages. The procedure has the following steps:

- Choose a formal logic that is complete and free of inconsistencies.
- For each logical operation, provide a type construction in the language.
- For each proof rule and axiom, provide a code construction in the language.

Mathematicians have studied different logics: e.g., modal logic, temporal logic, or linear logic. Compared with the constructive logic, these other logics have some additional operations. (For instance, modal logic adds the operations “necessarily” and “possibly”, and temporal logic adds the operation “until”.) For each logic, mathematicians have determined the minimal complete sets of operations, axioms, and proof rules that do not lead to inconsistency. Programming language designers can choose a logic and translate it into a minimal programming language where the code is guaranteed *not to crash* as long as types match. This mathematical guarantee (known as **type safety**) is a powerful help for programmers since it automatically prevents a large set of programming errors. So, programmers will benefit if their programming language is designed using the CH correspondence.

Practically useful programming languages will, of course, introduce many more features than the minimal, mathematically necessary constructions derived from the chosen logic. Programmers will still benefit from type safety as long as their programs stay within the mathematically consistent subset of the language. For Scala, a “safe” subset is identified by the `scalazzi` project.⁹

At present, it is not fully understood whether a practical programming language can use e.g., modal or linear logic as its logic of types. Experience suggests that, at least, the operations of the plain constructive logic should be available. So, it appears that the six type constructions and the eight code constructions will remain available in all future languages of functional programming.

5.5.3 Uses of the void type (`Nothing`)

The void type (Scala’s `Nothing`) corresponds to the logical constant *False*. There are few practical uses of the void type. One use case is for a branch of a `match/case` expression that does not return a value because it throws an exception. A `throw` expression is defined as if it returns a value of type `Nothing`. We can then pretend to convert that “value” (which will never be actually computed) into a value of any other type. (See Example 5.3.4.2 that implements a function `absurd[A]: Nothing => A`.)

To see how this trick is used, consider this code defining a value `x`,

```
val x: Double = if (t >= 0.0) math.sqrt(t) else { throw new Exception("error") }
```

The `else` branch does not return a value, but `x` is declared to be of type `Double`. For this code to type-check, both branches must return values of the same type. So, the compiler needs to pretend that the `else` branch also returns a value of type `Double`. The compiler first assigns the type `Nothing` to the expression `throw ...` and then automatically uses the conversion `Nothing => Double` to convert that type to `Double`. In this way, types will match in the definition of the value `x`.

⁹<https://github.com/scalaz/scalazzi>

Constructive logic	Boolean logic
$\overline{\Gamma \vdash \mathcal{CH}(1)}$ (create unit)	$\neg \Gamma \vee \text{True} = \text{True}$
$\overline{\Gamma, \alpha \vdash \alpha}$ (use arg)	$\neg \Gamma \vee \neg \alpha \vee \alpha = \text{True}$
$\overline{\Gamma \vdash \alpha \Rightarrow \beta}$ (create function)	$(\neg \Gamma \vee \neg \alpha \vee \beta) = (\neg \Gamma \vee (\alpha \Rightarrow \beta))$
$\overline{\Gamma \vdash \alpha \quad \Gamma \vdash \alpha \Rightarrow \beta \quad \Gamma \vdash \beta}$ (use function)	$((\neg \Gamma \vee \alpha) \wedge (\neg \Gamma \vee (\alpha \Rightarrow \beta))) \Rightarrow (\neg \Gamma \vee \beta)$
$\overline{\Gamma \vdash \alpha \quad \Gamma \vdash \beta}$ (create tuple)	$(\neg \Gamma \vee \alpha) \wedge (\neg \Gamma \vee \beta) = (\neg \Gamma \vee (\alpha \wedge \beta))$
$\overline{\Gamma \vdash \alpha \wedge \beta}$ (use tuple-1)	$(\neg \Gamma \vee (\alpha \wedge \beta)) \Rightarrow (\neg \Gamma \vee \alpha)$
$\overline{\Gamma \vdash \alpha \quad \Gamma \vdash \alpha \vee \beta}$ (create Left)	$(\neg \Gamma \vee \alpha) \Rightarrow (\neg \Gamma \vee (\alpha \vee \beta))$
$\overline{\Gamma \vdash \alpha \vee \beta \quad \Gamma, \alpha \vdash \gamma \quad \Gamma, \beta \vdash \gamma}$ (use Either)	$((\neg \Gamma \vee \alpha \vee \beta) \wedge (\neg \Gamma \vee \neg \alpha \vee \gamma) \wedge (\neg \Gamma \vee \neg \beta \vee \gamma)) \Rightarrow (\neg \Gamma \vee \gamma)$

Table 5.7: Proof rules of constructive logic are true also in the Boolean logic.

We will not use exceptions in this book. The functional programming paradigm does not use exceptions because their presence significantly complicates reasoning about code.

As another example of using the void type, suppose an external library implements a function

```
def parallel_run[E, A, B](f: A => Either[E, B]) = ???
```

that performs some parallel computations using a given function f . In general, the library supports functions $f: A \rightarrow E + B$ that may return an error of type E or a result of type B . Suppose we know that a particular function f never fails to compute its result. To express that knowledge in code, we may explicitly set the type parameter E to the void type `Nothing` when applying `parallel_run`:

```
parallel_run[Nothing, A, B](f) // Types match only when values f(a) always are of the form Right(b).
```

Returning an error is now impossible (the type `Nothing` has no values). If the function `parallel_run` is fully parametric, it will work in the same way with all types E , including $E = \emptyset$. The code implements our intention via type parameters, giving a compile-time guarantee of correct results.

So far, none of our examples involved the logical **negation** operation. It is defined as

$$\neg \alpha \triangleq \alpha \Rightarrow \text{False} ,$$

and its practical use is as limited as that of `False` and the void type. However, logical negation plays an important role in Boolean logic, which we will discuss next.

5.5.4 Relationship between Boolean logic and constructive logic

We have seen that some true theorems of Boolean logic are not true in constructive logic. For example, the Boolean identities $\neg(\neg \alpha) = \alpha$ and $(\alpha \Rightarrow \beta) = (\neg \alpha \vee \beta)$ do not hold in the constructive logic. However, any theorem of constructive logic is also a theorem of Boolean logic. The reason is that all eight rules of constructive logic (Section 5.2.3) are also true in Boolean logic.

To verify that a formula is true in Boolean logic, we only need to check that the value of the formula is `True` for all possible truth values (`True` or `False`) of its variables. A sequent such as $\alpha, \beta \vdash \gamma$ is true in Boolean logic if and only if $\gamma = \text{True}$ under the assumption that $\alpha = \beta = \text{True}$. So, the sequent $\alpha, \beta \vdash \gamma$ is translated into the Boolean formula

$$\alpha, \beta \vdash \gamma = ((\alpha \wedge \beta) \Rightarrow \gamma) = (\neg \alpha \vee \neg \beta \vee \gamma) .$$

Table 5.7 translates all proof rules of Section 5.2.3 into Boolean formulas. The first two lines are axioms, while the subsequent lines are Boolean theorems that can be verified by calculation.

To simplify the calculations, note that all terms in the formulas contain the operation $(\neg \Gamma \vee \dots)$ corresponding to the context Γ . Now, if Γ is `False`, the entire formula becomes automatically `True`,

and there is nothing else to check. So, it remains to verify the formula in case $\Gamma = \text{True}$, and then we can simply omit all instances of $\neg\Gamma$ in the formulas. Let us show the Boolean derivations for the rules “use function” and “use Either”; other formulas are checked in a similar way.

$$\begin{aligned}
 \text{formula “use function” : } & (\alpha \wedge (\alpha \Rightarrow \beta)) \Rightarrow \beta \\
 \text{use Eq. (5.5) : } & = \neg(\alpha \wedge (\neg\alpha \vee \beta)) \vee \beta \\
 \text{de Morgan’s laws : } & = \neg\alpha \vee (\alpha \wedge \neg\beta) \vee \beta \\
 \text{identity } p \vee (\neg p \wedge q) = p \vee q \text{ with } p = \neg\alpha \text{ and } q = \beta : & = \neg\alpha \vee \neg\beta \vee \beta \\
 \text{axiom “use arg” : } & = \text{True} .
 \end{aligned}$$

$$\begin{aligned}
 \text{formula “use Either” : } & ((\alpha \vee \beta) \wedge (\alpha \Rightarrow \gamma) \wedge (\beta \Rightarrow \gamma)) \Rightarrow \gamma \\
 \text{use Eq. (5.5) : } & = \neg((\alpha \vee \beta) \wedge (\neg\alpha \vee \gamma) \wedge (\neg\beta \vee \gamma)) \vee \gamma \\
 \text{de Morgan’s laws : } & = (\neg\alpha \wedge \neg\beta) \vee (\alpha \wedge \neg\gamma) \vee (\beta \wedge \neg\gamma) \vee \gamma \\
 \text{identity } p \vee (\neg p \wedge q) = p \vee q : & = (\neg\alpha \wedge \neg\beta) \vee \alpha \vee \beta \vee \gamma \\
 \text{identity } p \vee (\neg p \wedge q) = p \vee q : & = \neg\alpha \vee \alpha \vee \beta \vee \gamma \\
 \text{axiom “use arg” : } & = \text{True} .
 \end{aligned}$$

Since each proof rule of the constructive logic is translated into a true formula in Boolean logic, it follows that a proof tree in the constructive logic will be translated into a tree of Boolean formulas that have value *True* for each axiom or proof rule. The result is that any constructive proof for a sequent such as $\emptyset \vdash f(\alpha, \beta, \gamma)$ is translated into a chain of Boolean implications that look like this,

$$\text{True} = (\dots) \Rightarrow (\dots) \Rightarrow \dots \Rightarrow f(\alpha, \beta, \gamma) .$$

Since $(\text{True} \Rightarrow \alpha) = \alpha$, this chain proves the Boolean formula $f(\alpha, \beta, \gamma)$.

For example, the proof tree shown in Figure 5.1 is translated into

$$\begin{aligned}
 \text{axiom “use arg” : } & \text{True} = \neg((\alpha \Rightarrow \alpha) \Rightarrow \beta) \vee \neg\alpha \vee \alpha \\
 \text{rule “create function” : } & \Rightarrow \neg((\alpha \Rightarrow \alpha) \Rightarrow \beta) \vee (\alpha \Rightarrow \alpha) . \\
 \text{axiom “use arg” : } & \text{True} = \neg((\alpha \Rightarrow \alpha) \Rightarrow \beta) \vee ((\alpha \Rightarrow \alpha) \Rightarrow \beta) . \\
 \text{rule “use function” : } & \text{True} \Rightarrow (\neg((\alpha \Rightarrow \alpha) \Rightarrow \beta) \vee \beta) \\
 \text{rule “create function” : } & \Rightarrow (((\alpha \Rightarrow \alpha) \Rightarrow \beta) \Rightarrow \beta) .
 \end{aligned}$$

It is easier to check Boolean truth than to find a proof tree in constructive logic (or to establish that no proof tree exists). So, if we find that a formula is not true in Boolean logic, we know it is also not true in constructive logic. This gives us a quick way of proving that some type signatures are not implementable as fully parametric functions. In addition to formulas shown in Table 5.3 (Section 5.2.1), further examples of formulas that are not true in Boolean logic are

$$\begin{aligned}
 \forall\alpha. \alpha & , \\
 \forall(\alpha, \beta). \alpha \Rightarrow \beta & , \\
 \forall(\alpha, \beta). (\alpha \Rightarrow \beta) \Rightarrow \beta & .
 \end{aligned}$$

Table 5.7 uses the Boolean identity $(\alpha \Rightarrow \beta) = (\neg\alpha \vee \beta)$, which does not hold in the constructive logic, to translate the constructive axiom “use arg” into the Boolean axiom $\neg\alpha \vee \alpha = \text{True}$. The formula $\neg\alpha \vee \alpha = \text{True}$ is known as the **law of excluded middle**¹⁰ and is equivalent to saying that any proposition α is either true or false. It is remarkable that the constructive logic *does not have* the law of excluded middle; it is neither an axiom nor a derived theorem of constructive logic.

¹⁰https://en.wikipedia.org/wiki/Law_of_excluded_middle

To see why, consider what it would mean for $\neg\alpha \vee \alpha = \text{True}$ to hold in the constructive logic. The negation operation, $\neg\alpha$, is defined as the implication $\alpha \Rightarrow \text{False}$. So, the logical formula $\forall\alpha. \neg\alpha \vee \alpha$ corresponds to the type $\forall A. (A \rightarrow \mathbb{0}) + A$. Can we compute a value of this type in a fully parametric function? We need to compute either a value of type $A \rightarrow \mathbb{0}$ or a value of type A ; this decision needs to be made in advance independently of A , because the code of a fully parametric function must operate in the same way for all types. Should we return A or $A \rightarrow \mathbb{0}$? We certainly cannot compute a value of type A from scratch, since A is an arbitrary type. As we have seen in Example 5.3.4.3, a value of type $A \rightarrow \mathbb{0}$ exists if the type A is itself $\mathbb{0}$; but we do not know whether $A = \mathbb{0}$. In any case, a fully parametric function needs to have the same code for all types A . Since there are no values of type $\mathbb{0}$, and the type parameter A could be, say, `Int`, we cannot compute a value of type $A \rightarrow \mathbb{0}$.

Example 5.3.4.3 showed that the type $A \rightarrow \mathbb{0}$ is equivalent to $\mathbb{0}$ if A is not itself void ($A \not\cong \mathbb{0}$), and to $\mathbb{1}$ otherwise. Surely, any type A is either void or not void. So, why exactly is it impossible to implement a value of the type $(A \rightarrow \mathbb{0}) + A$? We could say that if A is void then $(A \rightarrow \mathbb{0}) \cong \mathbb{1}$ is not void, and so one of the types in the disjunction $(A \rightarrow \mathbb{0}) + A$ should be non-void (i.e., have values).

However, this reasoning is incorrect. It is insufficient to show that a value “should exist”; the real requirement is to *compute* a value of type $(A \rightarrow \mathbb{0}) + A$ via fully parametric function. That function’s code may not decide what to do depending on whether A is void — the code be the same for all types A (void or not). As we have seen, that code is impossible to write.

In Boolean logic, it is sufficient to prove that a value “should exist” (or that the non-existence of a value is contradictory in some way). However, any practically useful program needs to “construct”, i.e., compute, actual values and return them. The “constructive” logic got its name from this requirement. So, it is the constructive logic (not the Boolean logic) that provides correct reasoning about the types of values computable by fully parametric functional programs.

Without the requirement of full parametricity, we *could* implement the law of excluded middle. Special features of Scala (“reflection”, “type tags”, and “type casts”) allow us to compare types as values and to determine what type was given to a type parameter when a function is applied:

```
import scala.reflect.runtime.universe._

def getType[T: TypeTag]: Type = weakTypeOf[T]    // Convert the type parameter T into a special value.
def equalTypes[A: TypeTag, B: TypeTag]: Boolean = getType[A] ==: getType[B] // Compare types A and B.

def excludedMiddle[A: TypeTag]: Either[A, A => Nothing] = // excludedMiddle has type \forall A. (A \rightarrow \mathbb{0}) + A.
  if (equalTypes[A, Nothing]) Right((identity _).asInstanceOf[A => Nothing])    // Return id: \mathbb{0} \rightarrow \mathbb{0}.
  else if (equalTypes[A, Int]) Left(123.asInstanceOf[A])                         // Produce some value of type Int.
  else if (equalTypes[A, Boolean]) Left(true.asInstanceOf[A]) // Produce some value of type Boolean.
  else ???                                                 // Need to write many more definitions to support all other Scala types.

scala> excludedMiddle[Int]
res0: Either[Int,Int => Nothing] = Left(123)

scala> excludedMiddle[Nothing]
res1: Either[Nothing,Nothing => Nothing] = Right(<function1>)
```

In this code, we check whether $A = \mathbb{0}$; if so, we can implement $A \rightarrow \mathbb{0}$ as an identity function of type $\mathbb{0} \rightarrow \mathbb{0}$. Otherwise, we know that A is one of the existing Scala types (`Int`, `Boolean`, etc.), which are not void and have values that we can simply write down one by one in the subsequent code.

Explicit **type casts**, such as `123.asInstanceOf[A]`, are needed because the Scala compiler cannot know that A is `Int` in the scope where we return `Left(123)`. Without a type cast, the compiler will not accept `123` as a value of type A in that scope.

The method `asInstanceOf` is dangerous: the code `x.asInstanceOf[T]` disables all type checking for the given value `x`, telling the Scala compiler to believe that `x` has type `T` even when the type `T` is inconsistent with the actually given code of `x`. Programs written in this way will compile but may give unexpected results or crash because of errors that would have been prevented by type checking. It is rare that a Scala program truly requires type casts or explicit comparisons of type parameters. In this book, we will avoid writing such code.

6 Functors, contrafunctors, and profunctors

Type constructors such as `Seq[A]` or `Array[A]` are data structures that hold or “wrap” zero or more values of a given type `A`. These data structures are fully parametric: they work in the same way for every type `A`. Working with parametric “data wrappers” or “data containers” turns out to be a powerful design pattern of functional programming. To fully realize its benefits, we will formalize the concept of data wrapping through a set of mathematical laws. We will then extend that design pattern to all data types for which the laws hold.

6.1 Practical use

6.1.1 Motivation: Type constructors that wrap data

How to formalize the idea of wrapped data? An intuitive view is that the data is “still there”, i.e. we should be able to manipulate the data held within the wrapper. In functional programming, to manipulate means to apply functions to data. So, if an integer value 123 is “wrapped”, we should be able somehow to apply a function such as `f(x => x * 2)` and obtain a “wrapped” value 246.

Let us look at some often used type constructors defined in the Scala standard library, such as `Seq[A]`, `Try[A]`, and `Future[A]`. We notice the common features:

- There are some methods for creating a data structure that wraps zero or more values of a given type. For example, the Scala code `List.fill(10)(0)` creates a list of ten zeros of type `Int`.
- There are some methods for reading the wrapped values, if they exist. For example, the `List` class has the method `headOption` that returns a non-empty `Option` when the first element exists.
- There are some methods for manipulating the wrapped values while *keeping* them wrapped. For example, `List(10, 20, 30).map(_ + 5)` evaluates to `List(15, 25, 35)`.

The data types `Seq[A]`, `Try[A]`, and `Future[A]` express quite different kinds of wrapping. The data structure implementing `Seq[A]` can hold a variable number of values of type `A`. The data structure `Try[A]` holds either a successfully computed value of type `A` or a failure. The data structure `Future[A]` implements a computation that has been scheduled to run but may not have finished yet, and may compute a value of type `A` (or fail) at a later time.

Since the meaning of the “wrappers” `Seq`, `Try`, and `Future` is quite different, the methods for creating and reading the wrapped values have different type signatures for each wrapper. However, the method `map` is similar in all three examples. We can say generally that the `map` method will apply a given function $f: A \rightarrow B$ to the data of type `A` held inside the wrapper, and the new data (of type `B`) will remain within a wrapper of the same type:

```
val a = List(x, y, z).map(f) // Result is List(f(x), f(y), f(z)).  
val b = Try(x).map(f)      // Result is Try(f(x)).  
val c = Future(x).map(f)   // Result is Future(f(x)).
```

This motivates us to use the `map` function as the requirement for the wrapping functionality: A type constructor `Wrap[A]` is a “wrapper” if there exists a function `map` with the type signature

```
def map[A, B]: Wrap[A] => (A => B) => Wrap[B]
```

We can see that `Seq`, `Try`, and `Future` are “wrappers” because they have a suitable `map` method. This chapter focuses on the properties of `map` that are common to *all* wrapper types. We will ignore all

other features — reading data out of the wrapper, inserting or deleting data, waiting until data becomes available etc., — implemented by different methods specific to each wrapper type.

6.1.2 Example: Option and the identity law

As another example of a “data wrapper”, consider the type constructor `Option[A]`, which is written in the type notation as

$$\text{Opt}^A \triangleq \mathbb{1} + A \quad .$$

The type signature of its `map` function is

$$\text{map}^{A,B} : \mathbb{1} + A \rightarrow (A \rightarrow B) \rightarrow \mathbb{1} + B \quad .$$

This function produces a new `Option[B]` value that wraps transformed data. We will now use this example to develop intuition about manipulating data in a wrapper.

Two possible implementations of `map` fit the type signature:

```
def mapX[A, B](oa: Option[A])(f: A => B): Option[B] = None

def mapY[A, B](oa: Option[A])(f: A => B): Option[B] =
  oa match {
    case None      => None
    case Some(x)   => Some(f(x))
  }
```

The code of `mapX` loses information since it always returns `None` and ignores all input. The implementation `mapY` is more useful since it preserves information.

How can we formulate this property of `mapY` in a rigorous way? The trick is to choose the argument $f: A \rightarrow B$ in the expression `map(oa)(f)` to be the identity function $\text{id}^{A \rightarrow A}$ (setting `map`'s type parameters as $A = B$, so that the types match). Applying an identity function to a value wrapped in an `Option[A]` should not change that value. To verify that, substitute the identity function instead of `f` into `mapY` and compute:

```
mapY[A, A](x: Option[A])(identity[A]: A => A): Option[A]
  == x match {
    case None      => None          // No change.
    case Some(x)   => Some(x)       // No change.
  } == x
```

The result is always equal to `x`. We can write that fact as an equation,

$$\forall x: \text{Opt}^A. \text{map}(x)(\text{id}) = x \quad .$$

This equation is called the **identity law** of `map`. The identity law is a formal way of expressing the information-preserving property of the `map` function. The implementation `mapX` violates the identity law since it always returns `None` and so `mapX(oa)(id) == None` and not equal to `oa` for arbitrary values of `oa`. A data wrapper should not unexpectedly lose information when we manipulate the wrapped data. So, the correct implementation of `map` is `mapY`. The code notation for `map` is

$$\text{map}^{A,B} \triangleq p: \mathbb{1} + A \rightarrow f: A \rightarrow B \rightarrow p \triangleright \begin{array}{c|cc} & \mathbb{1} & B \\ \hline \mathbb{1} & \text{id} & \mathbb{0} \\ A & \mathbb{0} & f \end{array} \quad .$$

When writing code, it is convenient to use the `map` method defined in the Scala library. However, when reasoning about the properties of `map`, it turns out to be more convenient to flip the order of the curried arguments and to use the equivalent function, called `fmap`, with the type signature

$$\text{fmap}^{A,B} : (A \rightarrow B) \rightarrow \mathbb{1} + A \rightarrow \mathbb{1} + B \quad .$$

The Scala implementation and the code notation for `fmap` are shorter than those for `map`:

```
def fmap[A, B](f: A => B): Option[A] => Option[B] = {
  case None      => None
  case Some(x)   => Some(f(x))
}
```

$$\text{fmap}(f: A \rightarrow B) \triangleq \begin{array}{c|cc} & \mathbb{1} & B \\ \hline \mathbb{1} & \text{id} & \mathbb{0} \\ A & \mathbb{0} & f \end{array} \quad . \quad (6.1)$$

The identity law also looks simpler if expressed in terms of `fmap`, namely $fmap(id) = id$. In writing $fmap(id) = id$, we omitted the type parameters A and B , which must be both equal.

Note that the type signature of `fmap` looks like a transformation from functions of type $A \Rightarrow B$ to functions of type `Option[A] \Rightarrow Option[B]`. This transformation is called **lifting** because it “lifts” a function $f:A \rightarrow B$ operating on simple values into a function operating on `Option`-wrapped values.

So, the identity law can be formulated as “a lifted identity function is also an identity function”. If we lift an identity function and apply the resulting function to a wrapper, we expect the wrapped data not to change. The identity law expresses this expectation in a mathematical equation.

6.1.3 Motivation for the composition law

The main feature of a “data wrapper” is to allow us to manipulate the data inside it by applying functions to that data. The corresponding Scala code is `p.map(f)`, where `p` is a value of a wrapper type. It is natural to expect that lifted functions behave in the same way as the “unlifted” ones. For example, suppose we need to increment a counter `c` of type `Option[Int]`. The `Option` type means that the counter may be empty or non-empty; if it is non-empty, we increment the integer value wrapped inside the `Option` using the incrementing function

$$\text{incr} \triangleq x:\text{Int} \rightarrow x + 1 \quad .$$

In order to apply a function to the counter `c`, we need to lift that function. The Scala code is

```
def incr: Int => Int = x => x + 1
val c: Option[Int] = Some(0)

scala> c.map(incr)
res0: Option[Int] = Some(1)
```

If we apply the lifted function twice, we expect that the counter will be incremented twice:

```
scala> c.map(incr).map(incr)
res1: Option[Int] = Some(2)
```

This result is the same as when applying a lifted function $x \rightarrow x + 2$:

`scala> c.map(x => x + 2)` It would be confusing and counter-intuitive if `c.map(x => x + 2)` did
`res2: Option[Int] = Some(2)` not give the same result as `c.map(incr).map(incr)`.

We can formulate this property more generally: liftings should preserve function composition for arbitrary functions $f:A \rightarrow B$ and $g:B \rightarrow C$. This is written as

$$c.fmap(f).fmap(g) == c.fmap(f \text{ andThen } g) == c.fmap(x \Rightarrow g(f(x)))$$

$$c^{F^A} \triangleright \text{fmap}(f:A \rightarrow B) \triangleright \text{fmap}(g:B \rightarrow C) = c \triangleright \text{fmap}(f) ; \text{fmap}(g) = c \triangleright \text{fmap}(f:A \rightarrow B ; g:B \rightarrow C) \quad .$$

This equation is called the **composition law**. The law has the form $c \triangleright p = c \triangleright q$ with some functions p and q , which is the same as $\forall c. p(c) = q(c)$. This means an equality between functions, $p = q$. So we may omit the argument c and rewrite the law in a shorter form as

$$\text{fmap}(f) ; \text{fmap}(g) = \text{fmap}(f ; g) \quad .$$

Let us verify the composition law of the `Option` type. To practice the code derivations, we will perform the calculations by using both the code notation and the Scala syntax.

The Scala code for the function `fmap` was given in Section 6.1.2. To evaluate $\text{fmap}(f ; g)$, we apply `fmap(f andThen g)`, where `f: A \Rightarrow B` and `g: B \Rightarrow C` are arbitrary functions, to an arbitrary value `oa:Option[A]`. In Scala code, it is convenient to use the method `map` and write `oa.map(f)` instead of the equivalent expression `fmap(f)(oa)`:

```
fmap(f andThen g)(oa) == oa.map(f andThen g) == oa match {
  case None      => None
  case Some(x)   => (f andThen g)(x)
}
```

Since $(f \text{ andThen } g)(x) == g(f(x))$, we rewrite the result as

```
oa.map(f andThen g) == oa match {
  case None      => None
  case Some(x)   => g(f(x))
}
```

Now we consider the left-hand side of the law, $\text{fmap}(f) \circ \text{fmap}(g)$, and write the Scala expressions:

```
oa.map(f).map(g) == (oa match {
  case None      => None
  case Some(x)   => f(x)
}) .map(g) == (oa match {
  case None      => None
  case Some(x)   => f(x)
}) match {
  case None      => None
  case Some(y)   => g(y)
} == oa match {
  case None      => None
  case Some(x)   => g(f(x))
}
```

We find that the two sides of the law have identical code.

The derivation is much shorter in the matrix notation; we use Eq. (6.1) as the definition of fmap and omit the types:

$$\text{fmap}(f) \circ \text{fmap}(g) = \begin{vmatrix} \text{id} & 0 \\ 0 & f \end{vmatrix} \circ \begin{vmatrix} \text{id} & 0 \\ 0 & g \end{vmatrix}$$

matrix composition : $= \begin{vmatrix} \text{id} \circ \text{id} & 0 \\ 0 & f \circ g \end{vmatrix} = \begin{vmatrix} \text{id} & 0 \\ 0 & f \circ g \end{vmatrix}$

definition of fmap : $= \text{fmap}(f \circ g)$.

These calculations prove that the `map` method of the `Option` type satisfies the composition law. If the composition law did not hold, we would not be able to understand how `map` manipulates data within the `Option` wrapper. Looking at the Scala code example above, we expect `c.map(incr).map(incr)` to increment the data wrapped by `c` two times. If the result of `c.map(incr).map(incr)` were not `Some(2)` but, say, `Some(1)` or `None`, our ordinary intuitions about data transformations would become incorrect. In other words, violations of the composition law prevent us from understanding the code via mathematical reasoning about transformation of data values.

The composition law is a rigorous formulation of the requirement that wrapped data should be transformed (by lifted functions) in the same way as ordinary data. For example, the following associativity property holds for lifted functions:

Statement 6.1.3.1 For arbitrary functions $f:A \rightarrow B$, $g:B \rightarrow C$, and $h:C \rightarrow D$, we have

$$\text{fmap}(f) \circ \text{fmap}(g \circ h) = \text{fmap}(f \circ g) \circ \text{fmap}(h) .$$

Proof The left-hand side is rewritten as

$$\begin{aligned} & \text{fmap}(f) \circ \text{fmap}(g \circ h) \\ \text{composition law for } (g \circ h) : &= \text{fmap}(f) \circ (\text{fmap}(g) \circ \text{fmap}(h)) \\ \text{associativity law (4.3)} : &= (\text{fmap}(f) \circ \text{fmap}(g)) \circ \text{fmap}(h) \\ \text{composition law for } (f \circ g) : &= \text{fmap}(f \circ g) \circ \text{fmap}(h) , \end{aligned}$$

which now equals the right-hand side. This proves the statement.

6.1.4 Functors: definition and examples

Separating the functionality of “data wrapper” from any other features of a data type, we obtain:

- A data type with a type parameter, e.g., `L[A]`. We will use the notation L^\bullet (in Scala, `L[_]`) for the type constructor itself when the name of the type parameter is not needed.
- A fully parametric function fmap with type signature

$$\text{fmap}_L : (A \rightarrow B) \rightarrow L^A \rightarrow L^B .$$

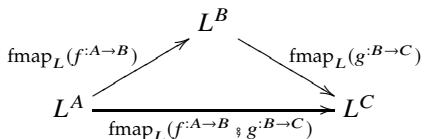
- The function fmap satisfies two laws:

$$\text{identity law of } L : \text{fmap}_L(\text{id}^{A \rightarrow A}) = \text{id}^{L^A \rightarrow L^A} , \quad (6.2)$$

$$\text{composition law of } L : \text{fmap}_L(f^{A \rightarrow B} \circ g^{B \rightarrow C}) = \text{fmap}_L(f^{A \rightarrow B}) \circ \text{fmap}_L(g^{B \rightarrow C}) . \quad (6.3)$$

A type constructor L^{\bullet} with these properties is called a **functor**. The laws (6.2)–(6.3) are the functor laws of identity and composition.

When a law involves function compositions, it is helpful to draw a type diagram to clarify how the functions transform various types involved in the law. A **type diagram** is a directed graph whose vertices are types and edges are functions mapping one type to another. Function composition corresponds to following a path in the diagram. A type diagram for the composition law (6.3) is shown



at left. There are two paths from L^A to L^C ; by Eq. (6.3), both paths must give the same result. Mathematicians call such diagrams **commutative**.

Type diagrams are easier to read when using the *forward* composition ($f \circ g$) because the order of edges is the same

as the order of functions in the composition. To see this, compare Eq. (6.3) and the type diagram above with the same law written using the backward composition,

$$\text{fmap}_L(g:B \rightarrow C \circ f:A \rightarrow B) = \text{fmap}_L(g:B \rightarrow C) \circ \text{fmap}_L(f:A \rightarrow B) .$$

The function `map` is computationally equivalent to `fmap` and can be defined through `fmap` by

$$\begin{aligned} \text{map}_L : L^A &\rightarrow (A \rightarrow B) \rightarrow L^B , \\ \text{map}_L(x:L^A)(f:A \rightarrow B) &= \text{fmap}_L(f:A \rightarrow B)(x:L^A) . \end{aligned}$$

Each of the type constructors `Option`, `Seq`, `Try`, and `Future` has its own definition of `map`; but the functor laws remain the same. We use the subscript L when writing `mapL` and `fmapL`, in order to indicate clearly the type constructor those functions work with.

We will now look at some examples of type constructors that are functors.

Standard data structures Many type constructors defined in the Scala library have a `map` method, and almost all of them are functors. The most often used functors are:

- The standard disjunctive types `Option`, `Try`, and `Either[A, B]` (where, by default, transformations apply to the type parameter `B`).
- The linear sequence `Seq` and its various derived classes such as `List`, `Range`, `Vector`, `IndexedSeq`, and `Stream`.
- The “task-like” constructors: `Future` and its alternatives: `Task` (provided by the `monix` library), `Async` and `Concurrent` (provided by the `cats-effect` library), `zio` (provided by the `zio` library).
- Dictionaries: `Map[K, V]` with respect to the type parameter `V`. The method is called `mapValues` instead of `map`: it transforms the values in the dictionary, leaving the keys unchanged.

Application-specific, custom type constructors defined by the programmer, such as case classes with type parameters, are often functors. Their structure is simple and helps build intuition for functors, so let us now consider some examples of case classes that are functors. In this book, they are called **polynomial functors**.

Polynomial functors Type constructors built with primitive types, type parameters, products, and disjunctions (or “sums”) are often used to represent application-specific data. Consider the code

```
case class Counted[A](n: Int, a: A) {
  def map[B](f: A => B): Counted[B] = Counted(n, f(a))
}
```

The data type `Counted[A]` may be used to describe `n` repetitions of a given value `a: A`. The code already defines the method `map` for the `Counted` class, which can be used like this,

```
scala> Counted(10, "abc").map(s => "prefix " + s)
res0: Counted[String] = Counted(10, prefix abc)
```

It is often more convenient to implement `map` as a class method rather than as a function such as

```
def map[A, B](c: Counted[A])(f: A => B): Counted[B] = c match {
  case Counted(n, a) => Counted(n, f(a))
}
```

The type notation for `Counted` is

$$\text{Counted}^A \triangleq \text{Int} \times A$$

showing that `Counted[_]` is a polynomial type constructor. The existence of a `map` method suggests that `Counted[_]` is a functor. We still need to check that the functor laws hold for it.

Example 6.1.4.1 Verify that the above implementation of `map` for `Counted` satisfies the functor laws.

Solution The implementation of `map` is fully parametric since it does not perform any type-specific operations; it uses the value `n: Int` as if `Int` were a type parameter. It remains to check that the laws hold. We will first verify the laws using the Scala syntax and then using the code notation.

The identity law means that for all `n: Int` and `a: A` we must have

```
Counted(n, a).map(identity) == Counted(n, a)
```

To verify this, we substitute the code of `map` and find

```
Counted(n, a).map(identity) == Counted(n, identity(a)) == Counted(n, a)
```

The composition law means that for all `n: Int`, `a: A`, `f: A => B`, and `g: B => C`, we must have

```
Counted(n, a).map(f).map(g) == Counted(n, a).map(f andThen g)
```

Substitute the Scala code of `map` into the left-hand side:

```
Counted(n, a).map(f).map(g) == Counted(n, f(a)).map(g) == Counted(n, g(f(a)))
```

The right-hand side can be transformed to the same expression:

```
Counted(n, a).map(f andThen g) == Counted(n, (f andThen g)(a)) == Counted(n, g(f(a)))
```

Let us now write a proof in the code notation, formulating the laws via the `fmap` method:

$$\text{fmap}_{\text{Counted}}(f: A \rightarrow B) \triangleq (n: \text{Int} \times a: A \rightarrow n \times f(a)) .$$

To verify the identity law, we write

$$\begin{aligned} \text{expect to equal } \text{id} : & \text{fmap}_{\text{Counted}}(\text{id}) \\ \text{definition of } \text{fmap}_{\text{Counted}} : & = (n \times a \rightarrow n \times \underline{\text{id}}(a)) \\ \text{definition of } \text{id} : & = (n \times a \rightarrow n \times a) = \text{id} . \end{aligned}$$

To verify the composition law,

$$\begin{aligned} \text{expect to equal } \text{fmap}_{\text{Counted}}(f \circ g) : & \text{fmap}_{\text{Counted}}(f) \circ \text{fmap}_{\text{Counted}}(g) \\ \text{definition of } \text{fmap}_{\text{Counted}} : & = (n \times a \rightarrow n \times f(a)) \circ (n \times b \rightarrow n \times g(b)) \\ \text{compute composition} : & = n \times a \rightarrow n \times \underline{g(f(a))} \\ \text{definition of } (f \circ g) : & = (n \times a \rightarrow n \times (f \circ g)(a)) = \text{fmap}_{\text{Counted}}(f \circ g) . \end{aligned}$$

We will prove later that all polynomial type constructors have a definition of `map` that satisfies the functor laws. It will be clear without proof that our definition of `map` for `Counted` is correct.

What would be an *incorrect* implementation of `map`? As an example, `map` could transform `Counted(n, a)` as before, except that the value `n` is now used to count the number of times `map` is applied:

```
def map_bad[A, B](c: Counted[A])(f: A => B): Counted[B] = c match {
  case Counted(n, a) => Counted(n + 1, f(a))
}
```

This implementation may appear reasonable. However, it violates both functor laws; for instance,

```
Counter(n, a) != map_bad(Counter(n, a))(identity) == Counter(n + 1, a)
```

The failure of functor laws leads to surprising behavior because a code refactoring changes the result:

```
Counter(n, a).map(incr).map(incr) != Counter(n, a).map(x => x + 2)
```

Let us look at some other simple examples of polynomial type constructors.

Example 6.1.4.2 Implement the `fmap` function for the type constructor

```
case class Vec3[A](x: A, y: A, z: A)
```

Solution Begin by implementing a fully parametric function:

```
def fmap[A, B](f: A => B): Vec3[A] => Vec3[B] = {
  case Vec3(x, y, z) => Vec3(f(x), f(y), f(z)) // Apply 'f' to all data of type 'A'.
}
```

Since all three values $f(x)$, $f(y)$, $f(z)$ have type B , the code of `fmap` would still satisfy the required type signature by returning, say, `Vec3(f(z), f(x), f(x))` or some other combination of these values. However, that implementation does not preserve information about the values x, y, z and about the ordering of these values in the original data `Vec(x, y, z)`. For this reason, we use the implementation of `fmap` shown first.

The type notation for the type constructor `Vec3[_]` is

$$\text{Vec}_3^A \triangleq A \times A \times A ,$$

and the code notation for `fmap` is

$$\text{fmap}_{\text{Vec}_3}(f: A \rightarrow B) \triangleq x:A \times y:A \times z:A \rightarrow f(x) \times f(y) \times f(z) .$$

Example 6.1.4.3 Implement the `fmap` function for the type constructor

$$\text{QueryResult}^A \triangleq \text{String} + \text{String} \times \text{Long} \times A .$$

Solution Begin by implementing the type constructor in Scala,

```
sealed trait QueryResult[A]
case class Error[A](message: String) extends QueryResult[A]
case class Success[A](name: String, time: Long, data: A) extends QueryResult[A]
```

Now implement a fully parametric, information-preserving function with the type signature of `fmap` for this type constructor:

```
def fmap[A, B](f: A => B): QueryResult[A] => QueryResult[B] = {
  case Error(message) => Error(message)
  case Success(name, time, data) => Success(name, time, f(data))
}
```

As in the previous example, we treat specific types (`Long`, `String`) as if they were type parameters. In this way, we obtain a correct implementation of `fmap` that satisfies the functor laws.

Recursive polynomial functors Recursive disjunctive type constructors shown in Section 3.3, such as lists and trees, are functors. Their `fmap` methods are recursive functions; they usually *cannot* be directly implemented with tail recursion.

Example Define a list of *odd* length as a recursive type LO^\bullet ,

$$\begin{aligned} \text{LO}^A &\triangleq A + A \times A \times \text{LO}^A \\ &\cong A + A \times A \times A + A \times A \times A \times A + \dots \end{aligned} \tag{6.4}$$

and implement `fmap` for it.

Solution The Scala definition of the type constructor `LO[_]` is

```
sealed trait LO[A]
final case class LO1[A](x: A) extends LO[A]
final case class LO2[A](x: A, y: A, tail: LO[A]) extends LO[A]
```

We can implement `fmap` as a recursive function:

```
def fmap[A, B](f: A => B): LO[A] => LO[B] = {
  case LO1(x)      => LO1[B](f(x))
  case LO2(x, y, tail) => LO2[B](f(x), f(y), fmap(f)(tail))
}
```

This code for `fmap` is not tail-recursive because `fmap` is called inside the case class constructor `LO2`.

The type constructor LO^* is a **recursive polynomial functor** because it is defined by a recursive type equation (6.4) that uses only polynomial type operations (sums and products) in its right-hand side. For the same reason, lists and trees are recursive polynomial functors.

6.1.5 Functor block expressions

Computations with wrapped values often require a chain of `map` methods, e.g.

```
scala> val result = Map(1 -> "one", 2 -> "two", 3 -> "three").
  map { case (i, name) => (i * i, name) }           // Compute i * i.
  map { case (x, name) => (x, s"$name * $name") }   // Compute product message.
  map { case (x, product) => s"$product is $x" }     // Compute final message.
result: Seq[String] = List(one * one is 1, two * two is 4, three * three is 9)
```

Such code can be rewritten equivalently in the **functor block** syntax:

```
val result = for {
  (i, name) <- Map(1 -> "one", 2 -> "two", 3 -> "three") // For each (i, name)...
  x = i * i           // define 'x' by computing i * i...
  product = s"$name * $name" // define 'product'...
} yield s"$product is $x" // and put these expressions into the 'result' sequence.
result: Seq[String] = List(one * one is 1, two * two is 4, three * three is 9)
```

Written in this way, the computations are easier to understand for two main reasons:

- There is less code to read and to write; no `map` or `case` and fewer curly braces.
- Values such as `name` and `x` need to be kept in tuples and passed from one `map` function to another, but any line in a functor block can directly reuse all values defined in previous lines.

The functor block is an important idiom in functional programming because it replaces a chain of `map` methods (as well as `filter` and `flatMap` methods, as we will see in later chapters) by a visually clearer sequence of definitions and expressions. Scala defines a functor block via the keywords `for` and `yield`. We will see many examples of functor blocks throughout this book. In this chapter, we only consider functor blocks that are equivalent to a chain of `map` operations on a functor value `p: LO[A]`. These functor blocks can be recognized because they contain *only one* left arrow (in the first line). Here is how to replace a chain of `map` operations by a functor block:

```
p.map(x => f(x)).map(y => g(y)).map(z => h(z)) == for {
  x <- p           // The first line must contain a left arrow before a functor value 'p'.
  y = f(x)         // Some computation involving 'x'.
  z = g(y)         // Another computation, uses 'y'.
} yield h(z)        // The 'yield h(z)' replaces the inner result of the last 'map'.
```

Translating functor blocks back into a chain of `map` operations is straightforward except for one complication: if some lines in the functor block make use of variables defined in earlier lines, the `map` operations may need to create some intermediate tuples that are not present in the functor block syntax. Consider the code

```
val result: L[B] = for {
  x <- p          // The first line must contain a left arrow before a functor value 'p'.
  y = f(x)        // Some computation involving 'x'.
  z = g(x, y)     // Another computation, uses 'x' and 'y'.
  ...
} yield q(x, y, z) // The 'yield' may use 'x', 'y', 'z', and any other defined variables.
```

The above functor block code assumes that $q(x, y, z)$ has type B , and is equivalent to

```
val result: L[B] = p
  .map { x => (x, f(x)) } // Create a tuple because we need to keep 'x' and 'f(x)'.
  .map { case (x, y) => (x, y, g(x, y)) } // Need to keep 'x', 'y', and 'g(x, y)'.
  ...
  .map { case (x, y, z) => q(x, y, z) } // Here, we can use 'x', 'y', and 'z'.
```

This code creates intermediate tuples only because the values x, y, z need to be used in later calculations. The functor block code is easier to read, write, and modify.

If desired, functor blocks may be written in a single line by using semicolons to separate the individual steps:

```
scala> for { x <- List(1, 2, 3); y = x * x; z = y + 2 } yield z
res0: List[Int] = List(3, 6, 11)
```

A confusing feature of the `for/yield` syntax is that, at first sight, functor blocks (such as the code shown at left) appear to compute (or to “yield”) the value $expr(x)$. However, this is not so. As the above examples show, if p is a sequence then the functor block also computes a *sequence*. In general, the result of a functor block is a “wrapped” value, where the type of the “wrapper” is determined by the first line of the functor block. The first line must have a left arrow followed by a “source”, which must be an expression of a functor type, i.e. of type $L[A]$ for some functor $L[.]$. The result’s type will be $L[B]$ where B is the type of the expression after the `yield` keyword.

For instance, the first line of the following functor block contains an `Option` value, `Some(123)`, as the “source”. Because of that, the value of the entire functor block expression will also be of type `Option`:

```
scala> for {
  x <- Some(123) // "Source" is Option[Int].
  y = (x - 3) / 10
} yield { if (y > 0) s"Have $y" else "Error" }
res1: Option[String] = Some(Have 12)
```

In this code, the `yield` keyword is followed by an expression of type `String`. So, the result of the entire functor block is of type `Option[String]`. Note that the expression after the “`yield`” can be a block of arbitrary code containing new `vals`, new `defs`, and/or

other `for/yield` functor blocks if needed.

Functor blocks can be used with any functor that has a `map` method, not only with library-defined type constructors such as `Seq` or `Option`. Here are some examples of defining the `map` methods and using functor blocks with disjunctive types.

The type constructor `QueryResult[.]` may define the `map` method on the trait itself and split its implementation between the case classes like this:

```
sealed trait QueryResult[A] {
  def map[B](f: A => B): QueryResult[B] // No implementation here.
}
case class Error[A](message: String) extends QueryResult[A] {
  def map[B](f: A => B): QueryResult[B] = Error(message)
}
case class Success[A](name: String, time: Long, data: A) extends QueryResult[A] {
  def map[B](f: A => B): QueryResult[B] = Success(name, time, f(data))
}
```

After these definitions, we can use `QueryResult` in functor blocks:

```
val q: QueryResult[Int] = Success("addresses", 123456L, 10)
scala> val result = for {
  x <- q
  y = x + 2
} yield y
```

```

} yield s"$y addresses instead of $x"
result: QueryResult[String] = Success(addresses,123456,12 addresses instead of 10)

```

As another example, let us define the `map` method on the `LO` trait (a recursive disjunctive type):

```

sealed trait LO[A] {
  def map[B](f: A => B): LO[B]
}

final case class LO1[A](x: A) extends LO[A] {
  def map[B](f: A => B): LO[B] = LO1[B](f(x))
}

final case class LO2[A](x: A, y: A, tail: LO[A]) extends LO[A] {
  def map[B](f: A => B): LO[B] = LO2[B](f(x), f(y), tail.map(f))
}

```

After these definitions, we may use values of type `LO[_]` in functor blocks:

```

scala> val result = for {
  x <- LO2("a", "quick", LO2("brown", "fox", LO1("jumped")))
  y = x.capitalize
  z = y + "/"
} yield (z, z.length)
result: LO[(String, Int)] = LO2((A,/2),(Quick/,6),LO2((Brown/,6),(Fox/,4),LO1((Jumped/,7))))

```

Functor blocks and functor laws There is an important connection between the functor laws and the properties of code in functor blocks. Consider the following code,

```

def f(x: Int) = x * x    // Some computations.
def g(x: Int) = x - 1    // More computations.

scala> for {
  x <- List(10, 20, 30)
  y = x
  z = f(y)    // Perform computations.
} yield g(z)
res0: List[Int] = List(99, 399, 899)

```

The code says that `x = y`, so it appears reasonable to eliminate `y` and simplify this code into

```

scala> for {
  x <- List(10, 20, 30)    // Eliminated 'y' from the code.
  z = f(x)    // Perform computations.
} yield g(z)
res1: List[Int] = List(99, 399, 899)

```

Another example of refactoring that appears reasonable is to combine transformations:

```

scala> for {
  x <- List(10, 20, 30)
  y = x + 1
  z = f(y)    // Perform computations.
} yield g(z)
res2: List[Int] = List(120, 440, 960)

```

The code says that `y = x + 1`, so we may want to replace `f(y)` by `f(x + 1)`:

```

scala> for {
  x <- List(10, 20, 30)
  z = f(x + 1)    // Eliminated 'y' from the code.
} yield g(z)
res3: List[Int] = List(120, 440, 960)

```

Looking at these code changes, we expect that the computed results will remain the same. Indeed, when the code directly states that `x = y`, it would be confusing and counter-intuitive if the result value changed after replacing `y` by `x`. When the code says that `y = x + 1`, ordinary mathematical reasoning suggests that `f(y)` can be replaced by `f(x + 1)` without affecting the results.

To see the connection with the functor laws, we translate the functor block syntax line by line into

Functor block syntax	Chains of <code>map</code> methods
<pre>for { // Code fragment 1a. x <- List(10, 20, 30) y = x z = f(y) } yield g(z)</pre>	<pre>List(10, 20, 30) // Code fragment 1b. .map(x => x).map(y => f(y)).map(z => g(z))</pre>
<pre>for { // Code fragment 2a. x <- List(10, 20, 30) z = f(x) } yield g(z)</pre>	<pre>List(10, 20, 30) // Code fragment 2b. .map(x => f(x)).map(z => g(z))</pre>
<pre>for { // Code fragment 3a. x <- List(10, 20, 30) y = x + 1 z = f(y) } yield g(z)</pre>	<pre>List(10, 20, 30) // Code fragment 3b. .map(x => x + 1).map(y => f(y)).map(z => g(z))</pre>
<pre>for { // Code fragment 4a. x <- List(10, 20, 30) z = f(x + 1) } yield g(z)</pre>	<pre>List(10, 20, 30) // Code fragment 4b. .map(x => f(x + 1)).map(z => g(z))</pre>

Table 6.1: Example translations of functor blocks into `map` methods.

chains of `map` methods. The resulting code fragments are shown in Table 6.1. The fragments using `map` methods were split into lines to emphasize their close correspondence to functor blocks.

We find that code fragments 1b and 2b are equal only if `.map(x => x)` does not modify the list to which it applies. This holds if the `map` method obeys the functor identity law, `p.map(identity) == p`, for all `p` of the appropriate type. We also find that code fragments 3b and 4b are equal if we can replace `.map(x => x + 1).map(f)` by `.map(x => f(x + 1))`. This replacement is justified as long as the `map` method obeys the functor composition law,

```
p.map(h).map(f) == p.map(x => f(h(x)))
```

for all `p` and functions `h` and `f` of appropriate types.

Functor laws guarantee that we can correctly understand and modify code written in functor blocks, reasoning about transformations of values as we do in mathematics.

6.1.6 Examples of non-functors

What properties of a data type make it a functor? To build an intuition, it is helpful to see examples of data types that are *not* functors.

There are several possibilities for a type constructor to fail being a functor:

- A `map` function's type signature cannot be implemented at all.
- A `map` function can be implemented but cannot satisfy the functor laws.
- A given `map` function is incorrect (does not satisfy the laws), although the error could be fixed: a different implementation of `map` satisfies the laws.
- A given `map[A, B]` function satisfies the laws for most types `A` and `B`, but violates the laws for certain specially chosen types.

We will now look at examples illustrating these possibilities.

Cannot implement `map`'s type signature Consider the type constructor H^* defined by

$$H^A \triangleq A \rightarrow \text{Int} \quad .$$

Scala code for this type notation can be

```
final case class H[A](r: A => Int)
```

The data type `H[A]` does not wrap data of type `A`; instead, it is a function that *consumes* data of type `A`. One cannot implement a fully parametric `map` function with the required type signature

$$\text{map}^{A,B} : (A \rightarrow \text{Int}) \rightarrow (A \rightarrow B) \rightarrow (B \rightarrow \text{Int}) \quad .$$

To see this, recall that a fully parametric function needs to treat all types as type parameters, including the primitive type `Int`. So the code

```
def map[A, B]: H[A] => (A => B) => H[B] = { r => f => C(_ => 123) }
```

satisfies the type signature of `map` but is not fully parametric because it returns a specific value `123` of type `Int`, which is not allowed. Replacing the type `Int` by a new type parameter `N`, we obtain the type signature

$$\text{map}^{A,B,N} : (A \rightarrow N) \rightarrow (A \rightarrow B) \rightarrow B \rightarrow N \quad .$$

We have seen in Example 5.4.1.8 that this type signature is not implementable. So, the type constructor H is not a functor.

Another important example of type constructors where the `map`'s type signature cannot be implemented are certain kinds of type constructors called **generalized algebraic data types** (GADTs). In this book, they are called **unfunctors** for short. An unfunctor is a type constructor having special values when its type parameter is set to certain specific types. An example of an unfunctor is

```
sealed trait ServerAction[R]
final case class GetResult[R](r: String => R) extends ServerAction[R]
final case class StoreId(x: Long, y: String) extends ServerAction[Boolean]
final case class StoreName(name: String) extends ServerAction[Int]
```

We see that some parts of the disjunctive type `ServerAction[R]` do not carry the type parameter `R` but instead set `R` to specific types, `R = Boolean` and `R = Int`. As a consequence, e.g., the case class `StoreName` has no type parameters and can only represent values of type `ServerAction[Int]` but not, say, `ServerAction[String]`. For this reason, `ServerAction[A]` cannot have a fully parametric `map` function,

```
def map[A, B]: ServerAction[A] => (A => B) => ServerAction[B]
```

To implement `map`, we are required to support any choice of the type parameters `A` and `B`. For example, with `A = Int`, we must be able to transform the value `StoreName("abc")` of type `ServerAction[Int]` to a value of type `ServerAction[B]` with any given `B`. However, the only way of creating a value of type `ServerAction[B]` with an arbitrary type `B` is to use the case class `GetResult[B]`. That requires us to create a function of type `String => B`. It is impossible for us to produce such a function out of `StoreName("abc")` and a function `f: Int => B` because the type `B` is unknown, and no fully parametric code could compute any values of type `Int` or of type `B` from the given value `StoreName("abc")`.

We are prevented from implementing `map` because some type parameters are already set in the definition of `ServerAction[R]`. One can say that the unfunctor `ServerAction[_]` fails to be fully parametric *in its type definition*. This behavior of unfunctors is intentional; unfunctors are only used in situations where the lack of `map` does not lead to problems (see Chapter 13).

Cannot implement a lawful `map` An example of a non-functor of the second kind is

$$Q^A \triangleq (A \rightarrow \text{Int}) \times A \quad .$$

Scala code for this type constructor is

```
final case class Q[A](q: A => Int, a: A)
```

A fully parametric `map` function with the correct type signature *can* be implemented (and there is only one such implementation):

$$\text{map}^{A,B} \triangleq q^{A \rightarrow \text{Int}} \times a^A \rightarrow f^{A \rightarrow B} \rightarrow (_ \rightarrow q(a))^{B \rightarrow \text{Int}} \times f(a) .$$

The corresponding Scala code is

```
def map[A, B]: Q[A] => (A => B) => Q[B] = { qa => f =>
  Q[B](_ => qa.q(qa.a), f(qa.a))
}
```

This `map` function is fully parametric (since it treats the type `Int` as a type parameter) and has the right type signature, but the functor laws do not hold. To show that the

identity law fails, we consider an arbitrary value $q^{A \rightarrow \text{Int}} \times a^A$ and compute:

$$\begin{aligned} \text{expect to equal } q \times a : & \text{ map}(q \times a)(\text{id}) \\ \text{definition of map} : & = (_ \rightarrow q(a)) \times \underline{\text{id}(a)} \\ \text{definition of id} : & = (_ \rightarrow q(a)) \times a \\ \text{expanded function, } q = (x \rightarrow q(x)) : & \neq q \times a = (x \rightarrow q(x)) \times a . \end{aligned}$$

The law must hold for arbitrary functions $q^{A \rightarrow \text{Int}}$, but the function $(_ \rightarrow q(a))$ always returns the same value $q(a)$ and thus is not equal to the original function q . So, the result of evaluating the expression $\text{map}(q \times a)(\text{id})$ is not always equal to the original value $q \times a$.

Since this `map` function is the only available implementation of the required type signature, we conclude that Q^\bullet is not a functor (we cannot implement `map` that satisfies the laws).

Mistakes in implementing `map` Non-functors of the third kind are type constructors with an incorrectly implemented `map`. An example is a type constructor $P^A \triangleq A \times A$ with the `map` function

$$\text{map} \triangleq x^A \times y^A \rightarrow f^{A \rightarrow B} \rightarrow f(y) \times f(x) .$$

Here is the Scala code corresponding to this code notation:

```
def map[A, B](p: (A, A))(f: A => B): (B, B) = p match { case (x, y) => (f(y), f(x)) }
```

This code swaps the values in the pair (x, y) ; we could say that it fails to preserve information about the order of those values. The functor identity law does not hold:

$$\begin{aligned} \text{expect to equal } x \times y : & \text{ map}(x^A \times y^A)(\text{id}^A) \\ \text{definition of map} : & = \underline{\text{id}(y)} \times \underline{\text{id}(x)} \\ \text{definition of id} : & = y \times x \neq x \times y . \end{aligned}$$

We should not have swapped the values in the pair. The correct implementation of `map`,

$$\text{map} \triangleq x^A \times y^A \rightarrow f^{A \rightarrow B} \rightarrow f(x) \times f(y) ,$$

preserves information and satisfies the functor laws.

Example 6.1.4.2 shows the type constructor Vec_3^\bullet with an incorrect implementation of `map` that reorders some parts of a tuple and duplicates other parts. The correct implementation preserves the order of parts in a tuple and does not duplicate or omit any parts.

Another case of an incorrect implementation is the following `map` function for `Option[_]`:

```
def map_bad[A, B]: Option[A] => (A => B) => Option[B] = { _ => _ => None }
```

This function always returns `None`, losing information and violating the identity law. However, we have already seen that `Option[_]` has a different implementation of `map` that satisfies the functor laws.

Similarly, one could define `map` for the `List[_]` type constructor to always return an empty list:

```
def map_bad[A, B]: List[A] => (A => B) => List[B] = { _ => _ => List() }
```

This implementation loses information and violates the functor laws. Of course, the Scala library provides a correct implementation of `map` for `List[_]`.

Example 6.1.4.1 is another situation where an incorrectly implemented `map` violates functor laws.

Functor laws will also be violated when `map` is not fully parametric. For instance, consider an implementation of `fmap[A, B](f)` that checks whether the two type parameters A and B are equal to each other as types, and if so, applies the function argument f twice. We need to use special features of Scala (run-time type reflection and `TypeTag`) for comparing two type parameters as types:

```
import scala.reflect.runtime.universe._
def getType[T: TypeTag]: Type = weakTypeOf[T]
def equalTypes[A: TypeTag, B: TypeTag]: Boolean = getType[A] ==: getType[B]

def fmap_bad[A: TypeTag, B: TypeTag](f: A => B)(oa: Option[A]): Option[B] = oa match {
  case None      => None
  case Some(x)   => // If A = B, compute f(f(x)), else compute f(x).
    val z: B = if (equalTypes[A, B]) f(f(x).asInstanceOf[A]) else f(x)
    Some(z)
}
```

Testing shows that this function works as designed:

```
scala> fmap_bad[Int, String](_ + " a")(Some(123))           // Appends " a" once.
res0: Option[String] = Some(123 a)

scala> fmap_bad[String, String](_ + " a")(Some("123")) // Appends " a" twice.
res1: Option[String] = Some(123 a a)
```

The function `fmap_bad[A, B]` satisfies the identity law but violates the composition law when $A = B$:

```
scala> fmap_bad[String, String](_ + " b")(Some("123 a a"))
res2: Option[String] = Some(123 a b b)

scala> fmap_bad[String, String](_ + " a b")(Some("123"))
res3: Option[String] = Some(123 a b a b)
```

In all these examples, we could implement a `map` function that would obey the laws. It is not precise to say that, e.g., the type constructor `Vec3[_]` is *by itself* a functor: being a functor depends on having a lawful `map` function. Keeping that in mind, we will say that the type constructor `Vec3[_]` “is” a functor, meaning that a suitable lawful implementation of `map` is known.

Laws hold for some types but not for others The Scala standard library contains `map` methods for the type constructors `Set` (transforming the values in a set) and `Map` (transforming both the keys and values in a dictionary). However, `Set[K]` and `Map[K, V]` fail to be lawful functors with respect to the type parameter K . The reason for the failure is complicated. A value of type `Set[K]` represents a set of zero or more values of type K , and it is enforced that all values in the set are distinct. So, the correct functionality of `Set` requires us to be able to check whether two values of type K are equal. A standard way of comparing values for equality is the `equals` method defined in the Scala library:

```
scala> List(1, 2, 3).equals(List(1, 2, 3))
res0: Boolean = true

scala> List(1, 2, 3).equals(List(1, 2, 3, 4))
res1: Boolean = false
```

However, an `equals` operation will work as expected only if it obeys the laws of **identity** (if $x = y$ then $f(x) = f(y)$ for any f), **symmetry** (if $x = y$ then $y = x$), **reflexivity** ($x = x$ for any x), and **transitivity** (if $x = y$ and $y = z$ then $x = z$). In most practical applications, the required type K (such as `String` or `Int`) will have a lawful `equals` method. In some cases, however, data types could redefine their `equals` method for application-specific purposes and violate some of the required laws.

Here are two examples of law-breaking (but potentially useful) code for `equals`. The first example¹ is a disjunctive type $A + B$ whose `equals` method allows only values of type $A + \emptyset$ to be equal:

¹This example is based on a comment by Paweł Szulc at <https://gist.github.com/tpolecat/7401433>

```
final case class OnlyA[A, B](eab: Either[A, B]) {
  override def equals(y: Any): Boolean = (eab, y) match {
    case (Left(a1), OnlyA(Left(a2)))  => a1 == a2 // Values Left(a1) and Left(a2) might be equal.
    case _                           => false      // Never equal unless both are 'Left'.
  }
}
```

This implementation of `equals` is mathematically invalid: it violates the reflexivity law ($\forall x. x = x$) because values of the form `OnlyA` are never equal to each other:

```
scala> OnlyA(Right(0)) equals OnlyA(Right(0))
res2: Boolean = false
```

As a result, the library code of `Set[OnlyA]` will fail to detect that, e.g., several values `OnlyA(Right(0))` are equal. The composition law of functors will fail when intermediate values of that type are used:

```
val f: OnlyA[Int, Int] => Int = { case OnlyA(Left(a)) => a; case OnlyA(Right(a)) => a }
val g: Int => OnlyA[Int, Int] = { a => OnlyA(Right(a)) }
val xs = Seq(0, 0, 0).map(g).toSet

scala> xs.map(f.andThen g) // 'Set' fails to detect identical values.
res3: Set[OnlyA[Int, Int]] = Set(OnlyA(Right(0)), OnlyA(Right(0)), OnlyA(Right(0)))

scala> xs.map(f).map(g) // 'Set' detects identical values.
res4: Set[OnlyA[Int, Int]] = Set(OnlyA(Right(0)))
```

The second example is a product type $A \times B$ whose `equals` method ignores the part of type B :

```
final case class IgnoreB[A, B](a: A, b: B) {
  override def equals(y: Any): Boolean = y match {
    case IgnoreB(a2, b2)  => a == a2 // Equal as long as the parts of type A are equal.
    case _                 => false   // Never equal to a value of another type (not IgnoreB).
  }
}

scala> IgnoreB(123, "abc") == IgnoreB(123, "def")
res5: Boolean = true
```

As a result, Scala's library code of `Set[IgnoreB]` will fail to detect that some values are different. This violates the functor composition law:

```
val f: IgnoreB[Int, Int] => IgnoreB[Int, Int] = { case IgnoreB(x, y) => IgnoreB(y, x) } // f ∘ f = id
val xs = Set(IgnoreB(0, 0), IgnoreB(1, 0))

scala> xs.map(f.andThen f) // This is equal to 'xs'.
res6: Set[IgnoreB[Int, Int]] = Set(IgnoreB(0,0), IgnoreB(1,0))

scala> xs.map(f).map(f) // This is not equal to 'xs'.
res7: Set[IgnoreB[Int, Int]] = Set(IgnoreB(0,0))
```

The functor laws for a type constructor L^\bullet do not require that the types A, B used in the function

$$\text{fmap}_L : (A \rightarrow B) \rightarrow L^A \rightarrow L^B$$

should have a mathematically lawful definition of the `equals` method (or of any other operation). The `map` method of a functor L^\bullet must be **lawful**, i.e., must satisfy the functor laws (6.2)–(6.3) for all types A, B . The functor laws must hold even if a type A 's implementation of some operations violate some other laws. For this reason, `Set[_]` cannot be considered a functor in a rigorous sense.

The `map` method for dictionaries has a similar problem: the keys of a dictionary must be distinct and will be compared using the `equals` method. So, the `map` method for `Map[K, V]` will violate the functor laws unless the type K has a lawful `equals` method.

The Scala standard library still provides the `map` and `flatMap` methods for sets `Set[K]` and dictionaries `Map[K, V]` because most applications will use types K that have lawful `equals` operations, and the functor laws will hold.

6.1.7 Contrafunctors

As we have seen in Section 6.1.6, the type constructor H^\bullet defined by $H^A \triangleq A \rightarrow \text{Int}$ is not a functor because it is impossible to implement the type signature of `map` as a fully parametric function,

$$\text{map}^{A,B} : (A \rightarrow \text{Int}) \rightarrow (A \rightarrow B) \rightarrow B \rightarrow \text{Int} \quad .$$

To see why, begin writing the code with a typed hole,

$$\text{map}(h^{A \rightarrow \text{Int}})(f^{A \rightarrow B})(b^B) = ???^{\text{Int}} \quad .$$

The only way of returning an `Int` in fully parametric code is by applying the function $h^{A \rightarrow \text{Int}}$. Since h consumes (rather than wraps) values of type A , we have no values of type A and cannot apply the function $h^{A \rightarrow \text{Int}}$. However, it would be possible to apply a function of type $B \rightarrow A$ since a value of type B is given as one of the curried arguments, b^B . So, we can implement a function called `contramap` with a different type signature where the function type is $B \rightarrow A$ instead of $A \rightarrow B$:

$$\text{contramap}^{A,B} : (A \rightarrow \text{Int}) \rightarrow (B \rightarrow A) \rightarrow B \rightarrow \text{Int} \quad .$$

The implementation of this function is written in the code notation as

$$\text{contramap} \triangleq h^{A \rightarrow \text{Int}} \rightarrow f^{B \rightarrow A} \rightarrow (f \circ h)^{B \rightarrow \text{Int}} \quad ,$$

and the corresponding Scala code is

```
def contramap[A, B](h: H[A])(f: B => A): H[B] = { f andThen h }
```

Flipping the order of the curried arguments in `contramap`, we define `cmap` as

$$\begin{aligned} \text{cmap}^{A,B} &: (B \rightarrow A) \rightarrow H^A \rightarrow H^B \quad , \\ \text{cmap} &\triangleq f^{B \rightarrow A} \rightarrow h^{A \rightarrow \text{Int}} \rightarrow (f \circ h)^{B \rightarrow \text{Int}} \quad . \end{aligned} \quad (6.5)$$

The type signature of `cmap` has the form of a “reverse lifting”: functions of type `B => A` are lifted into the type `H[A] => H[B]`. The Scala code for `cmap` is

```
def cmap[A, B](f: B => A): H[A] => H[B] = { h => f andThen h }
```

We can check that this `cmap` satisfies two laws analogous to the functor laws:

identity law : $\text{cmap}^{A,A}(\text{id}^{A \rightarrow A}) = \text{id}^{H^A \rightarrow H^A} \quad ,$

composition law : $\text{cmap}^{A,B}(f^{B \rightarrow A}) \circ \text{cmap}^{B,C}(g^{C \rightarrow B}) = \text{cmap}^{A,C}(g \circ f) \quad .$

$$\begin{array}{ccc} & H^B & \\ \text{cmap}_H(f^{B \rightarrow A}) & \nearrow & \searrow \text{cmap}_H(g^{C \rightarrow B}) \\ H^A & \xrightarrow{\text{cmap}_H(g^{C \rightarrow B} \circ f^{B \rightarrow A})} & H^C \end{array}$$

Since the function argument $f^{B \rightarrow A}$ has the reverse order of types, the composition law reverses the order of composition $(g \circ f)$ on one side; in this way, all types match. To verify the identity law:

expect to equal id : $\text{cmap}(\text{id})$

use Eq. (6.5) : $= h \rightarrow (\text{id} \circ h)$

definition of id : $= (h \rightarrow h) = \text{id} \quad .$

To verify the composition law:

expect to equal $\text{cmap}(g \circ f)$: $\text{cmap}(f) \circ \text{cmap}(g)$

use Eq. (6.5) : $= (h \rightarrow (f \circ h)) \circ (h \rightarrow (g \circ h))$

rename h to k for clarity : $= (h \rightarrow (f \circ h)) \circ (k \rightarrow (g \circ k))$

compute composition : $= (h \rightarrow g \circ f \circ h)$

use Eq. (6.5) : $= \text{cmap}(g \circ f) \quad .$

A type constructor with a fully parametric `cmap` is called a **contrafunctor** if the identity and the composition laws are satisfied.

Example 6.1.7.1 Show that the type constructor $D^A \triangleq A \rightarrow A \rightarrow \text{Int}$ is a contrafunctor.

Solution The required type signature for `contramap` is

```
def contramap[A, B](d: A => A => Int)(f: B => A): B => B => Int = ???
```

We begin implementing `contramap` by writing code with a typed hole:

$$\text{contramap}^{A,B} \triangleq d: A \rightarrow A \rightarrow \text{Int} \rightarrow f: B \rightarrow A \rightarrow b_1^B \rightarrow b_2^B \rightarrow ???^{\text{Int}} \quad .$$

To fill the typed hole, we need to compute a value of type `Int`. The only possibility is to apply `d` to two curried arguments of type `A`. We have two curried arguments of type `B`. So we apply $f: B \rightarrow A$ to those arguments, obtaining two values of type `A`. To avoid information loss, we need to preserve the order of the curried arguments. So the resulting expression is

$$\text{contramap}^{A,B} \triangleq d: A \rightarrow A \rightarrow \text{Int} \rightarrow f: B \rightarrow A \rightarrow b_1^B \rightarrow b_2^B \rightarrow d(f(b_1))(f(b_2)) \quad .$$

The corresponding Scala code is

```
def contramap[A, B](d: A => A => Int)(f: B => A): B => B => Int = { b1 => b2 => d(f(b1))(f(b2)) }
```

To verify the laws, it is easier to use the equivalent `cmap` defined by

$$\text{cmap}^{A,B}(f: B \rightarrow A) \triangleq d: A \rightarrow A \rightarrow \text{Int} \rightarrow b_1^B \rightarrow b_2^B \rightarrow d(f(b_1))(f(b_2)) \quad . \quad (6.6)$$

To verify the identity law:

$$\begin{aligned} \text{expect to equal id : } & \text{cmap(id)} \\ \text{use Eq. (6.6) : } & = d \rightarrow b_1 \rightarrow b_2 \rightarrow d(\underline{\text{id}(b_1)})(\underline{\text{id}(b_2)}) \\ \text{definition of id : } & = d \rightarrow b_1 \rightarrow b_2 \rightarrow d(b_1)(b_2) \\ \text{simplify curried function : } & = (d \rightarrow d) = \text{id} \quad . \end{aligned}$$

To verify the composition law, we rewrite its left-hand side into the right-hand side:

$$\begin{aligned} \text{cmap}(f); \text{cmap}(g) \\ \text{use Eq. (6.6) : } & = (d \rightarrow b_1 \rightarrow b_2 \rightarrow d(f(b_1))(f(b_2))) ; (d \rightarrow b_1 \rightarrow b_2 \rightarrow d(g(b_1))(g(b_2))) \\ \text{rename } d \text{ to } e : & = (d \rightarrow b_1 \rightarrow b_2 \rightarrow d(f(b_1))(f(b_2))) ; (e \rightarrow b_1 \rightarrow b_2 \rightarrow e(g(b_1))(g(b_2))) \\ \text{compute composition : } & = d \rightarrow b_1 \rightarrow b_2 \rightarrow d(f(g(b_1)))(f(g(b_2))) \\ \text{use Eq. (6.6) : } & = \text{cmap}(b \rightarrow f(g(b))) \\ \text{definition of } (g; f) : & = \text{cmap}(g; f) \quad . \end{aligned}$$

The type H^A represents a function that consumes a value of type A to produce an integer; the type D^A represents a curried function consuming *two* values of type A . These examples suggest the heuristic view that contrafunctors “consume” data while functors “wrap” data. By looking at the position of a given type parameter in a type expression such as $A \times \text{Int}$ or $A \rightarrow A \rightarrow \text{Int}$, we can see whether the type parameter is “consumed” or “wrapped”: A type parameter to the left of a function arrow is being “consumed”; a type parameter to the right of a function arrow (or used without a function arrow) is being “wrapped”. We will make this intuition precise in Section 6.2.

Type constructors that are not contrafunctors A type constructor that both consumes *and* wraps data is neither a functor nor a contrafunctor. An example of such a type constructor is

$$N^A \triangleq (A \rightarrow \text{Int}) \times (\mathbb{1} + A) \quad .$$

We can implement neither `map` nor `contramap` for N^A . Intuitively, the type parameter A is used both to the left of a function arrow (being “consumed”) and outside of a function (being “wrapped”).

Unfunctors (type constructors that lack full parametricity) also cannot be contrafunctors because the required type signature for `contramap` cannot be implemented by a fully parametric function. To show that `ServerAction[_]` cannot be a contrafunctor, we can straightforwardly adapt the reasoning used in Section 6.1.6 when we showed that `ServerAction[_]` cannot be a functor.

6.1.8 Subtyping, covariance, and contravariance

A type P is called a **subtype** of a type Q if there exists a designated **type conversion** function of type $P \rightarrow Q$ that the compiler will automatically use whenever necessary to match types. For instance, applying a function of type $Q \rightarrow Z$ to a value of type P is ordinarily a type error,

```
val h: Q => Z = ???
val p: P = ???
h(p) // Type error: the argument of h must be of type Q, not P.
```

However, this code will work when P is a subtype of Q because the compiler will automatically use the type conversion $P \rightarrow Q$ before applying the function h .

Different programming languages define subtyping differently because they make different choices of the type conversion functions and of types P, Q to which type conversions apply. Most often, the language designers choose the type conversion functions to be *identity* functions that merely reassign the types. Let us look at some examples of type conversion functions of that kind.

Within the focus of this book, the main example of subtyping is with disjunctive types. Consider this definition,

```
sealed trait AtMostTwo
final case class Zero() extends AtMostTwo
final case class One(x: Int) extends AtMostTwo
final case class Two(x: Int, y: Int) extends AtMostTwo
```

The corresponding type notation can be written as

$$\text{AtMostTwo} \triangleq \mathbb{1} + \text{Int} + \text{Int} \times \text{Int} \quad .$$

Each of the case classes (`Zero`, `One`, and `Two`) defines a type that is a subtype of `AtMostTwo`. To see that, we need to implement type conversion functions from each of the three case classes to `AtMostTwo`. The required functions reassign the types but perform no transformations on the data:

```
def f0: Zero => AtMostTwo = { case Zero() => Zero() }
def f1: One => AtMostTwo = { case One(x) => One(x) }
def f2: Two => AtMostTwo = { case Two(x, y) => Two(x, y) }
```

The implementation of these type conversion functions looks like the code of *identity* functions. In the matrix notation, we can write

$$\begin{array}{c} f_0 \triangleq \begin{array}{c|ccc} & \text{Zero} & \text{One} & \text{Two} \\ \hline \text{Zero} & \text{id} & 0 & 0 \end{array}, \quad f_0(\mathbb{1}^{\text{Zero}}) \triangleq \mathbb{1} + 0^{\text{One}} + 0^{\text{Two}} , \\ f_1 \triangleq \begin{array}{c|ccc} & \text{Zero} & \text{One} & \text{Two} \\ \hline \text{One} & 0 & \text{id} & 0 \end{array}, \quad f_1(x^{\text{Int}}) \triangleq 0^{\text{Zero}} + x^{\text{One}} + 0^{\text{Two}} , \\ f_2 \triangleq \begin{array}{c|ccc} & \text{Zero} & \text{One} & \text{Two} \\ \hline \text{Two} & 0 & 0 & \text{id} \end{array}, \quad f_2(x^{\text{Int}} \times y^{\text{Int}}) \triangleq 0^{\text{Zero}} + 0^{\text{One}} + (x \times y)^{\text{Two}} . \end{array}$$

This notation emphasizes that the code consists of identity functions with reassigned types.

Another example is a subtyping relation between function types. Consider the types

```
type P = (AtMostTwo => Int)
type Q = (Two => Int)
```

We can convert a function f of type P into a function g of type Q because f includes all the information necessary to define g . The Scala code for that type conversion is

```
def p2q(f: P): Q = { t: Two => f(t) }
```

This is written in the code notation as

$$\text{p2q}(f^{\text{AtMostTwo} \rightarrow \text{Int}}) \triangleq t^{\text{Two}} \rightarrow f(t) \quad .$$

Note that $t^{\text{Two}} \rightarrow f(t)$ is the same function as f , except applied to a subtype Two of AtMostTwo . So, the implementation of $\text{p2q}(f)$ is just f composed with an identity function with reassigned types.

In these cases, it is useful if the compiler could insert the appropriate conversion functions automatically whenever necessary. Any function that consumes an argument of type Q could be then automatically applicable to arguments of type P . The compiler could also remove the identity functions from the code, since they do not perform any data transformations. In this way, code involving subtypes becomes more concise with no decrease in performance.

To achieve this, we need to declare to the Scala compiler that certain types are in a subtyping relation. This can be done in one of three ways depending on the situation at hand:

1. Declaring a class that `extends` another class (as we have just seen).
2. Declaring type parameters with a “variance annotation” such as `L[+A]` or `L[-B]`.
3. Declaring type parameters with a “subtyping annotation” (`A <: B`).

Subtyping for disjunctive types A function with argument of type `AtMostTwo` can be applied to a value of type `Two` with no additional code written by the programmer:

```
def head: AtMostTwo => Option[Int] = {
  case Zero()      => None
  case One(x)      => Some(x)
  case Two(x, y)   => Some(x)
}

scala> head(Two(10, 20))
res0: Option[Int] = Some(10)
```

We may imagine that the compiler automatically used the type conversion function f_2 shown above to convert a value of the type `Two` into a value of the type `AtMostTwo`. Since the code of f_2 is equivalent to an identity function, the type conversion does not change any data and only reassigns the types of the given values. So the compiler does not need to insert any additional code, and the type conversion does not lead to any decrease in performance.

Subtyping for type constructors If a type constructor L^A is a functor, we can use its fmap_L method to lift a type conversion function $f : P \rightarrow Q$ into

$$\text{fmap}_L(f) : L^P \rightarrow L^Q \quad ,$$

which gives a type conversion function from L^P to L^Q . This gives a subtyping relation between the types L^P and L^Q because the code of the lifted function $\text{fmap}_L(f)$ is an identity function, due to functor L 's identity law, $\text{fmap}_L(\text{id}) = \text{id}$.

If a type constructor H^A is a contrafunctor, a type conversion function $f^{P \rightarrow Q}$ is lifted to

$$\text{cmap}_H(f) : H^Q \rightarrow H^P \quad ,$$

showing that H^Q is a subtype of H^P . The identity law of the contrafunctor H ,

$$\text{cmap}_H(\text{id}) = \text{id} \quad ,$$

shows that the lifted conversion function is an identity function with reassigned types.

A type constructor F is called **covariant** if F^A is a subtype of F^B whenever A is a subtype of B . A **contravariant** type constructor H has the subtype relation in the opposite direction: H^B is a subtype of H^A . In principle, all functors could be declared as covariant type constructors, and all contrafunctors as contravariant type constructors.² However, the Scala compiler does not automatically determine whether a given type constructor `F[A]` is covariant with respect to a given type parameter A . To indicate the covariance property, the programmer needs to use a **variance annotation**, which

²The name “contrafunctor” was chosen in this book as a shortened form of “contravariant functor”.

looks like $F[+A]$, on the relevant type parameters. For example, the type constructor `Counted[A]` defined in Section 6.1.4 is a functor and so is covariant in its type parameter A . If we use the variance annotation `Counted[+A]` in the definition, Scala will automatically consider the type `Counted[Two]` as a subtype of `Counted[AtMostTwo]`. So we may now apply a function to a value of type `Counted[Two]` as if it had type `Counted[AtMostTwo]`:

```
final case class Counted[+A](n: Int, a: A)

def total(c: Counted[AtMostTwo]): Int = c match {
  case Counted(n, Zero())      => 0
  case Counted(n, One(_))     => n
  case Counted(n, Two(_, _))   => n * 2
}

scala> total(Counted(2, Two(10, 20)))
res1: Int = 4
```

The contravariance property for contrafunctors can be annotated using the syntax `F[-A]`.

A given type constructor may have several type parameters and may be covariant with respect to some of them and contravariant with respect to others. As we have seen, the position of a type parameter in a type expression indicates whether the value is “wrapped” (used in a **covariant position**) or “consumed” (used in a **contravariant position**). Covariant positions are to the right of function arrows, or outside function arrows; contravariant positions are to the left of a function arrow. The next examples confirm this intuition, which will be made rigorous in Section 6.2.

6.1.9 Solved examples: functors and contrafunctors

Example 6.1.9.1 Consider this implementation of `map` for the type constructor `Option[_]`:

```
def map[A, B](oa: Option[A])(f: A => B): Option[B] = oa match {
  case None          => None
  case Some(x: Int)  => Some(f((x+1).asInstanceOf[A]))
  case Some(x)        => Some(f(x))
}
```

This code performs a non-standard computation if the type parameter A is set to `Int`. Show that this implementation of `map` violates the functor laws.

Solution If the type parameter A is not `Int`, or if the argument `oa` is `None`, the given code is the same as the standard (correct) implementation of `map` for `Option`. The function does something non-standard when e.g., `oa == Some(123)`. Substitute this value of `oa` into the identity law, `map(oa)(identity) == oa`, and compute symbolically (using Scala syntax)

```
map(oa)(identity) == Some(identity((123+1).asInstanceOf[Int])) == Some(124) != oa
```

This shows a violation of the functor identity law.

Example 6.1.9.2 Define case classes and implement `fmap` for the given type constructors:

- (a) $Data^A \triangleq String + A \times Int + A \times A \times A$.
- (b) $Data^A \triangleq 1 + A \times (Int \times String + A)$.
- (c) $Data^A \triangleq (String \rightarrow Int \rightarrow A) \times A + (Boolean \rightarrow Double \rightarrow A) \times A$.

Solution (a) Begin by defining a case class for each part of the disjunctive type:

```
sealed trait Data[A]
final case class Message[A](message: String)  extends Data[A]
final case class Have1[A](x: A, n: Int)        extends Data[A]
final case class Have3[A](x: A, y: A, z: A)     extends Data[A]
```

The names `Message`, `Have1`, `Have3`, `n`, `x`, `y`, `z` are chosen arbitrarily.

The function `fmap` must have the type signature

$$fmap^{A,B} : f^{A \rightarrow B} \rightarrow Data^A \rightarrow Data^B .$$

To implement `fmap` correctly, we need to transform each part of the disjunctive type `Data[A]` into the corresponding part of `Data[B]` without loss of information. To clarify where the transformation $f^{A \rightarrow B}$

need to be applied, let us write the type notation for Data^A and Data^B side by side:

$$\begin{aligned}\text{Data}^A &\triangleq \text{String} + A \times \text{Int} + A \times A \times A \quad , \\ \text{Data}^B &\triangleq \text{String} + B \times \text{Int} + B \times B \times B \quad .\end{aligned}$$

Now it is clear that we need to apply f to each value of type A present in Data^A , preserving the order of values. The Scala code is

```
def fmap[A, B](f: A => B): Data[A] => Data[B] = {
  case Message(message) => Message(message)
  case Have1(x, n)      => Have1(f(x), n)
  case Have3(x, y, z)   => Have3(f(x), f(y), f(z))
}
```

(b) It is convenient to define the disjunctive type $\text{Int} \times \text{String} + A$ separately as P^A :

```
sealed trait P[A]
final case class Message[A](code: Int, message: String) extends P[A]
final case class Value[A](x: A) extends P[A]
```

Now we notice that the type expression $(\mathbb{1} + \dots)$ can be encoded via the standard `Option` type. So, the Scala code for Data^A is

```
final case class Data[A](d: Option[(A, P[A])])
```

To help us implement `fmap` correctly, we write out the type expressions

$$\begin{aligned}\text{Data}^A &\triangleq \mathbb{1} + A \times (\text{Int} \times \text{String} + A) \quad , \\ \text{Data}^B &\triangleq \mathbb{1} + B \times (\text{Int} \times \text{String} + B) \quad ,\end{aligned}$$

and transform Data^A into Data^B by applying $f: A \rightarrow B$ at the correct places:

```
def fmap[A, B](f: A => B): Data[A] => Data[B] = {
  case Data(None)                      => Data(None)
  case Data(Some((x, Message(code, message)))) => Data(Some((f(x), Message(code, message))))
  case Data(Some((x, Value(y))))        => Data(Some((f(x), Value(f(y))))))
}
```

When deeply nested patterns become hard to read, we may handle the nested structure separately:

```
def fmap[A, B](f: A => B): Data[A] => Data[B] = {
  case Data(None)                      => Data(None)
  case Data(Some((x, p)))              =>
    val newP: P[B] = p match {
      case Message(code, message)    => Message(code, message)
      case Value(x)                 => Value(f(x))
    }
    Data(Some((f(x), newP)))
}
```

(c) Since the type structures $(\text{String} \rightarrow \text{Int} \rightarrow A) \times A$ and $(\text{Boolean} \rightarrow \text{Double} \rightarrow A) \times A$ have a similar pattern, let us define a parameterized type

$$Q^{X,Y,A} \triangleq (X \rightarrow Y \rightarrow A) \times A \quad ,$$

and express the given type expression as

$$\text{Data}^A \triangleq Q^{\text{String}, \text{Int}, A} + Q^{\text{Boolean}, \text{Double}, A} \quad .$$

It is then convenient to define `Data[A]` using the standard disjunctive type `Either`:

```
type Q[X, Y, A] = (X => Y => A, A)
type Data[A] = Either[Q[String, Int, A], Q[Boolean, Double, A]]
```

To make the code clearer, we will implement `fmap` separately for Q^\bullet and Data^\bullet .

To derive the code of `fmap` for Q^\bullet , we begin with the type signature

$$\text{fmap}_{Q^\bullet}^{A,B} : (A \rightarrow B) \rightarrow (X \rightarrow Y \rightarrow A) \times A \rightarrow (X \rightarrow Y \rightarrow B) \times B$$

and start writing the code using typed holes,

$$\text{fmap}_{Q^\bullet}(f: A \rightarrow B) \triangleq g: X \rightarrow Y \rightarrow A \times a: A \rightarrow \text{???}: X \rightarrow Y \rightarrow B \times \text{???}: B \quad .$$

The typed hole $\text{???}: B$ is filled by $f(a)$. To fill the remaining type hole, we write

$$\begin{aligned} \text{???}: X \rightarrow Y \rightarrow B \\ = x: X \rightarrow y: Y \rightarrow \text{???}: B \\ = x: X \rightarrow y: Y \rightarrow f(\text{???}: A) \quad . \end{aligned}$$

It would be wrong to fill the typed hole $\text{???}: A$ by $a: A$ because, to preserve information, a value of type $X \rightarrow Y \rightarrow B$ should be computed using the given data g of type $X \rightarrow Y \rightarrow A$. So we write

$$\text{???}: X \rightarrow Y \rightarrow B = x: X \rightarrow y: Y \rightarrow f(g(x)(y)) \quad .$$

The corresponding Scala code is

```
def fmap_Q[A, B, X, Y](f: A => B): Q[X, Y, A] => Q[X, Y, B] = {
  case (g, a) => (x => y => f(g(x)(y)), f(a))
  // Could also write the code as
  // case (g, a) => (x => g(x) andThen f, f(a))
}
```

Finally, we can write the code for `fmap` for Data^\bullet :

```
def fmap_Data[A, B](f: A => B): Data[A] => Data[B] = {
  case Left(q) => Left(fmap_Q(f)(q))
  case Right(q) => Right(fmap_Q(f)(q))
}
```

The Scala compiler will automatically infer the type parameters required by `fmap_Q` and check that all types match. With all inferred types written out, the code above would be

```
def fmap_Data[A, B](f: A => B): Data[A] => Data[B] = {
  case Left(q: Q[String, Int, A]) =>
    Left[Q[String, Int, B]](fmap_Q[A, B, String, Int](f)(q))
  case Right(q: Q[Boolean, Double, A]) =>
    Right[Q[Boolean, Double, B]](fmap_Q[A, B, Boolean, Double](f)(q))
}
```

When types become complicated, it may help to write out some of the type parameters in the code.

Example 6.1.9.3 Decide which of these types are functors or contrafunctors, and implement `fmap` or `cmap` as appropriate:

- (a) $\text{Data}^A \triangleq (A \rightarrow \text{Int}) + (A \rightarrow A \rightarrow \text{String}) \quad .$
- (b) $\text{Data}^{A,B} \triangleq (A + B) \times ((A \rightarrow \text{Int}) \rightarrow B) \quad .$

Solution (a) The type constructor Data^A uses its type parameter A always as an argument of some functions, i.e., to the left of function arrows:

```
type Data[A] = Either[A => Int, A => A => String]
```

So, Data^A consumes values of type A , and we expect that Data^A is a contrafunctor. Indeed, we can implement `cmap`:

```
def cmap[A, B](f: B => A): Data[A] => Data[B] = {
  case Left(a2Int) => Left(b => a2Int(f(b)))
  case Right(a2a2String) => Right(b1 => b2 => a2a2String(f(b1))(f(b2)))
}
```

(b) The type constructor $\text{Data}^{A,B}$ has *two* type parameters, and so we need to answer the question separately for each of them. Write the Scala type definition as

```
type Data[A, B] = (Either[A, B], (A => Int) => B)
```

Begin with the type parameter A and notice that a value of type $\text{Data}^{A,B}$ possibly contains a value of type A within `Either[A, B]`. In other words, A is “wrapped”, i.e., it is in a covariant position within the first part of the tuple. It remains to check the second part of the tuple, which is a higher-order function of type $(A \rightarrow \text{Int}) \rightarrow B$. That function consumes a function of type $A \rightarrow \text{Int}$, which in turn consumes a value of type A . Consumers of A are contravariant in A , but it turns out that a “consumer of a consumer of A ” is *covariant* in A . So we expect to be able to implement `fmap` that applies to the type parameter A of $\text{Data}^{A,B}$. Renaming the type parameter B to Z for clarity, we write the type signature for `fmap` like this,

$$\text{fmap}^{A,C,Z} : (A \rightarrow C) \rightarrow (A + Z) \times ((A \rightarrow \text{Int}) \rightarrow Z) \rightarrow (C + Z) \times ((C \rightarrow \text{Int}) \rightarrow Z) \quad .$$

We need to transform each part of the tuple separately. Transforming $A + Z$ into $C + Z$ is straightforward via the function

		C	Z
A		f	0
Z		0	id

This code notation corresponds to the following Scala code:

```
{
  case Left(x)    => Left(f(x))
  case Right(z)   => Right(z)
}
```

To derive code transforming $(A \rightarrow \text{Int}) \rightarrow Z$ into $(C \rightarrow \text{Int}) \rightarrow Z$, we use typed holes:

$$\begin{aligned}
 f^{A \rightarrow C} &\rightarrow g^{(A \rightarrow \text{Int}) \rightarrow Z} \rightarrow \underline{\text{???}^{(C \rightarrow \text{Int}) \rightarrow Z}} \\
 \text{nameless function :} &= f^{A \rightarrow C} \rightarrow g^{(A \rightarrow \text{Int}) \rightarrow Z} \rightarrow p^{C \rightarrow \text{Int}} \rightarrow \underline{\text{???}^Z} \\
 \text{get a } Z \text{ by applying } g : &= f^{A \rightarrow C} \rightarrow g^{(A \rightarrow \text{Int}) \rightarrow Z} \rightarrow p^{C \rightarrow \text{Int}} \rightarrow g(\underline{\text{???}^{A \rightarrow \text{Int}}}) \\
 \text{nameless function :} &= f^{A \rightarrow C} \rightarrow g^{(A \rightarrow \text{Int}) \rightarrow Z} \rightarrow p^{C \rightarrow \text{Int}} \rightarrow g(a^A \rightarrow \underline{\text{???}^{\text{Int}}}) \\
 \text{get an Int by applying } p : &= f^{A \rightarrow C} \rightarrow g^{(A \rightarrow \text{Int}) \rightarrow Z} \rightarrow p^{C \rightarrow \text{Int}} \rightarrow g(a^A \rightarrow p(\underline{\text{???}^C})) \\
 \text{get a } C \text{ by applying } f : &= f^{A \rightarrow C} \rightarrow g^{(A \rightarrow \text{Int}) \rightarrow Z} \rightarrow p^{C \rightarrow \text{Int}} \rightarrow g(a^A \rightarrow p(f(\underline{\text{???}^A}))) \\
 \text{use argument } a^A : &= f \rightarrow g \rightarrow p \rightarrow g(a \rightarrow p(f(a))) \quad .
 \end{aligned}$$

In the resulting Scala code for `fmap`, we write out some types for clarity:

```
def fmapA[A, Z, C](f: A => C): Data[A, Z] => Data[C, Z] = {
  case (e: Either[A, Z], g: ((A => Int) => Z)) =>
    val newE: Either[C, Z] = e match {
      case Left(x)    => Left(f(x))
      case Right(z)   => Right(z)
    }
    val newG: (C => Int) => Z = { p => g(a => p(f(a))) }
    (newE, newG) // This has type Data[C, Z].
}
```

This suggests that $\text{Data}^{A,Z}$ is covariant with respect to the type parameter A . The results of Section 6.2 will show rigorously that the functor laws hold for this implementation of `fmap`.

The analysis is simpler for the type parameter B because it is only used in covariant positions, never to the left of function arrows. So we expect $\text{Data}^{A,B}$ to be a functor with respect to B . Implementing the corresponding `fmap` is straightforward:

```

def fmapB[Z, B, C](f: B => C): Data[Z, A] => Data[Z, B] = {
  case (e: Either[Z, B], g: ((Z => Int) => B)) =>
    val newE: Either[Z, B] = e match {
      case Left(x) => Left(f(x))
      case Right(z) => Right(z)
    }
    val newG: (C => Int) => Z = { p => g(a => p(f(a))) }
    (newE, newG) // This has type Data[C, Z].
}

```

The code indicates that $\text{Data}^{A,B}$ is a functor with respect to both A and B .

Example 6.1.9.4 Rewrite the following code in the type notation; identify covariant and contravariant type usages; verify with the Scala compiler that the variance annotations are correct:

```

sealed trait Coi[A, B]
final case class Pa[A, B](b: (A, B), c: B => Int)      extends Coi[A, B]
final case class Re[A, B](d: A, e: B, c: Int)           extends Coi[A, B]
final case class Ci[A, B](f: String => A, g: B => A)   extends Coi[A, B]

```

Solution The type notation puts all parts of the disjunctive type into a single type expression:

$$\text{Coi}^{A,B} \triangleq A \times B \times (B \rightarrow \text{Int}) + A \times B \times \text{Int} + (\text{String} \rightarrow A) \times (B \rightarrow A) .$$

Now find which types are wrapped and which are consumed in this type expression. The type parameter A is wrapped and never consumed, but B is both wrapped and consumed (in $B \rightarrow A$). So, the type constructor `Coi` is covariant in A but neither covariant nor contravariant in B . We can check this by compiling the corresponding Scala code with variance annotations:

```

sealed trait Coi[+A, B]
case class Pa[+A, B](b: (A, B), c: B => Int)      extends Coi[A, B]
case class Re[+A, B](d: A, e: B, c: Int)           extends Coi[A, B]
case class Ci[+A, B](f: String => A, g: B => A)   extends Coi[A, B]

```

We could also replace the fixed types `Int` and `String` by type parameters `N` and `S`. A similar analysis shows that `N` is in covariant positions while `S` is in a contravariant position. We can then check that the Scala compiler accepts the following type definition with variance annotations:

```

sealed trait Coi2[+A, B, +N, -S]
case class Pa2[+A, B, +N, -S](b: (A, B), c: B => N)  extends Coi[A, B, N, S]
case class Re2[+A, B, +N, -S](d: A, e: B, c: N)        extends Coi[A, B, N, S]
case class Ci2[+A, B, +N, -S](f: S => A, g: B => A)   extends Coi[A, B, N, S]

```

6.1.10 Exercises: functors and contrafunctors

Exercise 6.1.10.1 An implementation of `fmap` for the type constructor `Either[A, A]` is given as

```

def fmap[A, B](f: A => B): Either[A, A] => Either[B, B] = {
  case Left(a)      => Right(f(a))
  case Right(a: Int) => Left(f(a + 1))
  case Right(a)      => Left(f(a))
}

```

Show that this implementation of `fmap` violates the functor laws. Implement `fmap` correctly for this type constructor and the given type signature.

Exercise 6.1.10.2 Define these type constructors in Scala, decide whether they are covariant or contravariant, and implement `fmap` or `cmap` as appropriate:

- (a) $\text{Data}^A \triangleq (\mathbb{1} + A) \times (\mathbb{1} + A) \times \text{String}$.
- (b) $\text{Data}^A \triangleq (A \rightarrow \text{Boolean}) \rightarrow (A \times (\text{Int} + A))$.
- (c) $\text{Data}^{A,B} \triangleq (A \rightarrow \text{Boolean}) \times (A + B \rightarrow \text{Int})$.
- (d) $\text{Data}^A \triangleq (\mathbb{1} + (A \rightarrow \text{Boolean})) \rightarrow (\mathbb{1} + (A \rightarrow \text{Int})) \rightarrow \text{Int}$.

(e) $\text{Data}^B \triangleq (B + (\text{Int} \rightarrow B)) \times (B + (\text{String} \rightarrow B))$.

Exercise 6.1.10.3 Rewrite the following code in the type notation; find covariant and contravariant positions of type parameters; add variance annotations and verify that the resulting code compiles:

```
sealed trait S[A, B]
final case class P[A, B](a: A, b: B, c: Int)      extends S[A, B]
final case class Q[A, B](d: Int => A, e: Int => B) extends S[A, B]
final case class R[A, B](f: A => A, g: A => B)    extends S[A, B]
```

6.2 Laws and structure

A type constructor is a functor if it admits a lawful `map` function. How can we recognize quickly that a given type constructor is a functor or perhaps a contrafunctor? For example, consider the type constructor $Z^{A,R}$ defined by

$$Z^{A,R} \triangleq ((A \rightarrow A \rightarrow R) \rightarrow R) \times A + (1 + R \rightarrow A + \text{Int}) + A \times A \times \text{Int} \times \text{Int} . \quad (6.7)$$

Is $Z^{A,R}$ a functor with respect to A , or perhaps with respect to R ? To answer these questions, we will systematically build up various type expressions for which the functor or contrafunctor laws hold.

6.2.1 Reformulations of laws

We begin by introducing a more convenient notation for the functor laws. The laws (6.2)–(6.3) were defined in terms of the function `fmap`. When written in terms of the curried function `map`, the structure of the laws becomes less clear:

$$\begin{aligned} \text{map}_L(x^{:L^A})(\text{id}^{:A \rightarrow A}) &= x , \\ \text{map}_L(x^{:L^A})(f^{:A \rightarrow B} ; g^{:B \rightarrow C}) &= \text{map}_L(\text{map}_L(x)(f))(g) . \end{aligned}$$

The laws again look clearer when using `map` as a class method:

```
x.map(identity) == x
x.map(f).map(g) == x.map(f andThen g)
```

To take advantage of this syntax, we can use the pipe notation where $x \triangleright \text{fmap}(f)$ means `x.map(f)`, and write the functor laws as

$$\begin{aligned} x \triangleright \text{fmap}_L(\text{id}) &= x , \\ x \triangleright \text{fmap}_L(f) \triangleright \text{fmap}_L(g) &= x \triangleright \text{fmap}_L(f ; g) . \end{aligned}$$

In later chapters of this book, we will find that the `map` methods (equivalently, the `fmap` function) are used so often in different contexts that the notation $\text{fmap}_L(f)$ becomes too verbose. To make code expressions visually easy to manipulate, we need a shorter notation. At the same time, it is important to show clearly the relevant type constructor L . Dropping the symbol L can lead to errors, since it will be sometimes unclear what type constructors are involved in an expression such as `x.map(f).map(g)` and whether we are justified in replacing that expression with `x.map(f andThen g)`.

For these reasons, we introduce the superscript notation \uparrow^L (pronounced “lifted to L ”) defined, for any function f , by

$$(f^{:A \rightarrow B})^{\uparrow^L} : L^A \rightarrow L^B , \quad f^{\uparrow^L} \triangleq \text{fmap}_L(f) .$$

Now we can choose the notation according to convenience and write

$$\text{map}_L(x)(f) = \text{fmap}_L(f)(x) = x \triangleright \text{fmap}_L(f) = x \triangleright f^{\uparrow^L} = f^{\uparrow^L}(x) .$$

In this notation, the identity and composition laws for a functor L are especially easy to use:

$$\text{id}^{\uparrow^L} = \text{id} , \quad (f ; g)^{\uparrow^L} = f^{\uparrow^L} ; g^{\uparrow^L} .$$

Applying a composition of lifted functions to a value looks like this,

$$x \triangleright (f \circ g)^{\uparrow L} = x \triangleright f^{\uparrow L} \circ g^{\uparrow L} = x \triangleright f^{\uparrow L} \triangleright g^{\uparrow L} .$$

This equation directly represents the Scala code syntax

```
x.map(f andThen g) == (_.map(f) andThen _.map(g))(x) == x.map(f).map(g)
```

since the piping symbol (\triangleright) groups weaker than the composition symbol (\circ).

Written in the *backward* notation ($f \circ g$), the functor composition law is

$$(g \circ f)^{\uparrow L} = g^{\uparrow L} \circ f^{\uparrow L} .$$

The analogous notation for a contrafunctor C^\bullet is

$$f^{\downarrow C} \triangleq \text{cmap}_C(f) .$$

The contrafunctor laws are then written as

$$\text{id}^{\downarrow C} = \text{id} , \quad (f \circ g)^{\downarrow C} = g^{\downarrow C} \circ f^{\downarrow C} , \quad (g \circ f)^{\downarrow C} = f^{\downarrow C} \circ g^{\downarrow C} .$$

We will mostly use the forward composition $f \circ g$ in this book, keeping in mind that one can straightforwardly and mechanically translate between forward and backward notations via

$$f \circ g = g \circ f , \quad x \triangleright f = f(x) .$$

6.2.2 Bifunctors

A type constructor can be a functor with respect to several type parameters. A **bifunctor** is a type constructor with *two* type parameters that satisfies the functor laws with respect to both type parameters at once.

As an example, consider the type constructor F defined by

$$F^{A,B} \triangleq A \times B \times B .$$

If we fix the type parameter B but let the parameter A vary, we get a type constructor that we can denote as $F^{\bullet,B}$. We see that the type constructor $F^{\bullet,B}$ is a functor, with the corresponding `fmap` function

$$\text{fmap}_{F^{\bullet,B}}(f:A \rightarrow C) \triangleq a:A \times b_1^B \times b_2^B \rightarrow f(a) \times b_1 \times b_2 .$$

Instead of saying that $F^{\bullet,B}$ is a functor, we can also say more verbosely that $F^{A,B}$ is a functor with respect to A .

If we now fix the type parameter A , we find that the type constructor $F^{A,\bullet}$ is a functor, with the `fmap` function

$$\text{fmap}_{F^{A,\bullet}}(g:B \rightarrow D) \triangleq a:A \times b_1^B \times b_2^B \rightarrow a \times g(b_1) \times g(b_2) .$$

Since the bifunctor $F^{\bullet,\bullet}$ is a functor with respect to each type parameter separately, we can transform a value of type $F^{A,B}$ to a value of type $F^{C,D}$ by applying the two `fmap` functions one after another. It is convenient to denote this transformation by a single operation called `bimap` that uses two functions $f:A \rightarrow C$ and $g:B \rightarrow D$ as arguments:

$$\begin{aligned} \text{bimap}_F(f:A \rightarrow C)(g:B \rightarrow D) &: F^{A,B} \rightarrow F^{C,D} , \\ \text{bimap}_F(f:A \rightarrow C)(g:B \rightarrow D) &\triangleq \text{fmap}_{F^{\bullet,B}}(f:A \rightarrow C) \circ \text{fmap}_{F^{A,\bullet}}(g:B \rightarrow D) . \end{aligned} \tag{6.8}$$

In the condensed notation, this is written as

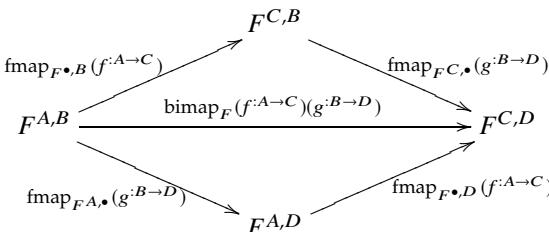
$$\text{bimap}_F(f:A \rightarrow C)(g:B \rightarrow D) \triangleq f^{\uparrow F^{\bullet,B}} \circ g^{\uparrow F^{A,\bullet}} ,$$

but in this case the longer notation in Eq. (6.8) is easier to reason about.

What if we apply the two fmap functions in the opposite order? Since these functions work with different type parameters, it is reasonable to expect that the transformation $F^{A,B} \rightarrow F^{C,D}$ should be independent of the order of application:

$$\text{fmap}_{F \bullet, B}(f^{A \rightarrow C}) \circ \text{fmap}_{F C, \bullet}(g^{B \rightarrow D}) = \text{fmap}_{F A, \bullet}(g^{B \rightarrow D}) \circ \text{fmap}_{F \bullet, D}(f^{A \rightarrow C}) \quad . \quad (6.9)$$

This equation is illustrated by the type diagram below.



Different paths in this diagram give the same results if they arrive at the same vertex (as mathematicians say, “the diagram commutes”). In this way, the diagram illustrates at once the commutativity law (6.9) and the definition (6.8) of bimap_F .

Let us verify the commutativity law for the bifunctor $F^{A,B} \triangleq A \times B \times B$:

left-hand side : $\text{fmap}_{F \bullet, B}(f^{A \rightarrow C}) \circ \text{fmap}_{F C, \bullet}(g^{B \rightarrow D})$

definitions of $\text{fmap}_{F \bullet, \bullet}$: $= (a^A \times b_1^B \times b_2^B \rightarrow f(a) \times b_1 \times b_2) \circ (c^C \times b_1^B \times b_2^B \rightarrow c \times g(b_1) \times g(b_2))$

compute composition : $= a^A \times b_1^B \times b_2^B \rightarrow f(a) \times g(b_1) \times g(b_2) \quad ,$

right-hand side : $\text{fmap}_{F A, \bullet}(g^{B \rightarrow D}) \circ \text{fmap}_{F \bullet, D}(f^{A \rightarrow C})$

definitions of $\text{fmap}_{F \bullet, \bullet}$: $= (a^A \times b_1^B \times b_2^B \rightarrow a \times g(b_1) \times g(b_2)) \circ (a^A \times d_1^D \times d_2^D \rightarrow f(a) \times d_1 \times d_2)$

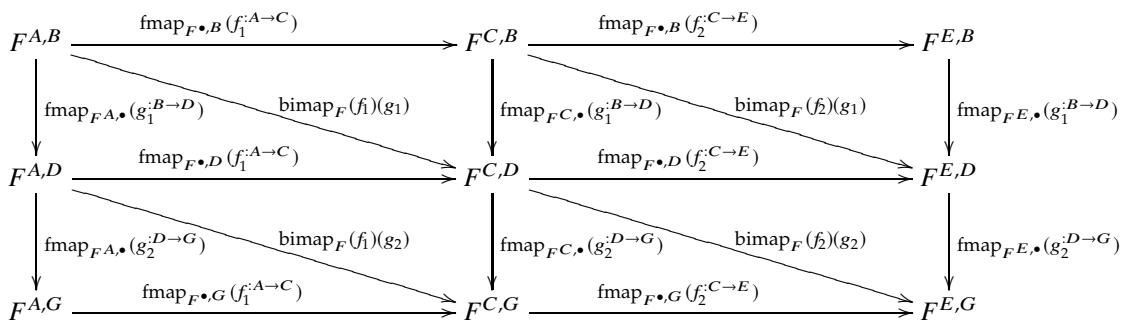
compute composition : $= a^A \times b_1^B \times b_2^B \rightarrow f(a) \times g(b_1) \times g(b_2) \quad .$

Both sides of the law are equal.

The commutativity law (6.9) leads to the composition law of bimap ,

$$\text{bimap}_F(f_1^{A \rightarrow C})(g_1^{B \rightarrow D}) \circ \text{bimap}_F(f_2^{C \rightarrow E})(g_2^{D \rightarrow G}) = \text{bimap}_F(f_1 \circ f_2)(g_1 \circ g_2) \quad . \quad (6.10)$$

The following type diagram shows the relationships between various bimap and fmap functions:



To derive the composition law from Eq. (6.9), write

$$\text{bimap}_F(f_1)(g_1) \circ \text{bimap}_F(f_2)(g_2)$$

use Eq. (6.8) : $= \text{fmap}_{F \bullet, B}(f_1) \circ \text{fmap}_{F C, \bullet}(g_1) \circ \text{fmap}_{F \bullet, D}(f_2) \circ \text{fmap}_{F E, \bullet}(g_2)$

commutativity law (6.9) : $= \text{fmap}_{F \bullet, B}(f_1) \circ \text{fmap}_{F \bullet, B}(f_2) \circ \text{fmap}_{F E, \bullet}(g_1) \circ \text{fmap}_{F E, \bullet}(g_2)$

composition laws : $= \text{fmap}_{F \bullet, B}(f_1 \circ f_2) \circ \text{fmap}_{F E, \bullet}(g_1 \circ g_2)$

use Eq. (6.8) : $= \text{bimap}_F(f_1 \circ f_2)(g_1 \circ g_2) \quad .$

Construction	Type notation	Comment
type parameter	$L^A \triangleq A$	the identity functor
product type	$L^A \triangleq P^A \times Q^A$	the functor product; P and Q must be functors
disjunctive type	$L^A \triangleq P^A + Q^A$	the functor co-product; P and Q must be functors
function type	$L^A \triangleq C^A \rightarrow P^A$	the functor exponential; P is a functor and C a contrafunctor
type parameter	$L^A \triangleq Z$	the constant functor; Z is a fixed type
type constructor	$L^A \triangleq P^Q^A$	functor composition; P and Q are both functors or both contrafunctors
recursive type	$L^A \triangleq S^{A,L^A}$	recursive functor; $S^{A,B}$ must be a functor w.r.t. both A and B

Table 6.2: Type constructions defining a functor L^A .

Conversely, we can derive Eq. (6.9) from the composition law (6.10). We write the composition law with specially chosen functions:

$$\text{bimap}_F(f:A \rightarrow C)(g:B \rightarrow D) = \text{bimap}_F(\text{id}:A \rightarrow A)(g:B \rightarrow D) ; \text{bimap}_F(f:A \rightarrow C)(\text{id}:D \rightarrow D) \quad . \quad (6.11)$$

Using Eq. (6.8), we find

$$\begin{aligned} \text{expect fmap}_{F^A, \bullet}(g) ; \text{fmap}_{F^B, \bullet}(f) &: \text{fmap}_{F^{\bullet}, B}(f:A \rightarrow C) ; \text{fmap}_{F^C, \bullet}(g:B \rightarrow D) \\ \text{use Eq. (6.8)} &: = \text{bimap}_F(f:A \rightarrow C)(g:B \rightarrow D) \\ \text{use Eq. (6.11)} &: = \text{bimap}_F(\text{id}:A \rightarrow A)(g:B \rightarrow D) ; \text{bimap}_F(f:A \rightarrow C)(\text{id}:D \rightarrow D) \\ \text{use Eq. (6.8)} &: = \underline{\text{fmap}_{F^{\bullet}, B}(\text{id})} ; \underline{\text{fmap}_{F^A, \bullet}(g)} ; \underline{\text{fmap}_{F^B, \bullet}(f)} ; \underline{\text{fmap}_{F^C, \bullet}(\text{id})} \\ \text{identity laws for } F &: = \text{fmap}_{F^A, \bullet}(g) ; \text{fmap}_{F^B, \bullet}(f) \quad . \end{aligned}$$

The identity law of `bimap` holds as well,

$$\begin{aligned} \text{expect to equal id} &: \text{bimap}_F(\text{id}:A \rightarrow A)(\text{id}:B \rightarrow B) \\ \text{use Eq. (6.8)} &: = \underline{\text{fmap}_{F^{\bullet}, B}(\text{id})} ; \underline{\text{fmap}_{F^C, \bullet}(\text{id})} \\ \text{identity laws for } F &: = \text{id} ; \text{id} = \text{id} \quad . \end{aligned}$$

If $F^{A,B}$ is known to be a functor separately with respect to A and B , will the commutativity law (6.9) always hold? The calculation for the example $F^{A,B} \triangleq A \times B \times B$ shows that the two `fmap` functions commute because they work on different parts of the data structure $F^{A,B}$. This turns out³ to be true in general: the commutativity law follows from the parametricity of the `fmap` functions. Because of that, we do not need to verify the `bimap` laws as long as $F^{\bullet, B}$ and $F^{A, \bullet}$ are lawful functors.

Type constructors with more than two type parameters have similar properties. It is sufficient to check the functor laws with respect to each type parameter separately.

In general, a type constructor may be a functor with respect to some type parameters and a contrafunctor with respect to others. Below we will see examples of such type constructors.

6.2.3 Constructions of functors

What type expressions will produce a functor? Functional programming languages support the six standard type constructions (see Section 5.1.2). This section will check whether each construction produces a new type that obeys the functor laws. The results are summarized in Table 6.2.

³Proof of that statement is beyond the scope of this chapter. See Section D.1 in Appendix D.

In each of these constructions, the `fmap` function for a new functor is defined either from scratch or by using the known `fmap` functions for previously defined type constructors. We will now derive the code for these constructions and prove their validity. We will use the code notation for brevity, occasionally showing the translation into the Scala syntax.

Statement 6.2.3.1 The type constructor $\text{Id}^A \triangleq A$ is a lawful functor (the **identity functor**).

Proof The `fmap` function is defined by

$$\begin{aligned} \text{fmap}_{\text{Id}} : (A \rightarrow B) &\rightarrow \text{Id}^A \rightarrow \text{Id}^B \cong (A \rightarrow B) \rightarrow A \rightarrow B , \\ \text{fmap}_{\text{Id}} \triangleq (f : A \rightarrow B \rightarrow f) &= \text{id}^{(A \rightarrow B) \rightarrow A \rightarrow B} . \end{aligned}$$

The identity function is the only fully parametric implementation of the type signature $(A \rightarrow B) \rightarrow A \rightarrow B$. Since the code of `fmap` is the identity function, the laws are satisfied automatically:

$$\begin{aligned} \text{identity law} : \text{fmap}_{\text{Id}}(\text{id}) &= \text{id}(\text{id}) = \text{id} , \\ \text{composition law} : \text{fmap}_{\text{Id}}(f \circ g) &= f \circ g = \text{fmap}_{\text{Id}}(f) \circ \text{fmap}_{\text{Id}}(g) . \end{aligned}$$

Statement 6.2.3.2 The type constructor $\text{Const}^{Z,A} \triangleq Z$ is a lawful functor (a **constant functor**) with respect to the type parameter A .

Proof The `fmap` function is defined by

$$\begin{aligned} \text{fmap}_{\text{Const}} : (A \rightarrow B) &\rightarrow \text{Const}^{Z,A} \rightarrow \text{Const}^{Z,B} \cong (A \rightarrow B) \rightarrow Z \rightarrow Z , \\ \text{fmap}_{\text{Const}}(f : A \rightarrow B \rightarrow f) &\triangleq (z : Z \rightarrow z) = \text{id}^{Z \rightarrow Z} . \end{aligned}$$

It is a constant function that ignores f and returns the identity $\text{id}^{Z \rightarrow Z}$. The laws are satisfied:

$$\begin{aligned} \text{identity law} : \text{fmap}_{\text{Const}}(\text{id}) &= \text{id} , \\ \text{composition law} : \text{fmap}_{\text{Const}}(f \circ g) &= \text{id} = \text{fmap}_{\text{Const}}(f) \circ \text{fmap}_{\text{Const}}(g) = \text{id} \circ \text{id} . \end{aligned}$$

The corresponding Scala code is

```
type Const[Z, A] = Z
def fmap[A, B](f: A => B): Const[Z, A] => Const[Z, B] = identity[Z]
```

The identity functor Id^{\bullet} and the constant functor $\text{Const}^{Z,\bullet}$ are not often used: their `fmap` implementations are identity functions, and so they rarely provide useful functionality.

We have seen that type constructors with product types, such as $L^A \triangleq A \times A \times A$, are functors. The next construction (the **functor product**) explains why.

Statement 6.2.3.3 If L^{\bullet} and M^{\bullet} are two functors then the product $P^A \triangleq L^A \times M^A$ is also a functor.

Proof The `fmap` function for P is defined by

```
def fmap[A, B](f: A => B): (L[A], M[A]) => (L[B], M[B]) = {
  case (la, ma) => (la.map(f), ma.map(f))
}
```

The corresponding code notation is

$$f^{\uparrow P} \triangleq l : L^A \times m : M^A \rightarrow f^{\uparrow L}(l) \times f^{\uparrow M}(m) .$$

Writing this code using the pipe (`▷`) operation makes it somewhat closer to the Scala syntax:

$$(l : L^A \times m : M^A) \triangleright f^{\uparrow P} \triangleq (l \triangleright f^{\uparrow L}) \times (m \triangleright f^{\uparrow M}) . \quad (6.12)$$

An alternative notation uses the **pair product** symbol \boxtimes defined by

$$\begin{aligned} p : A \rightarrow B \boxtimes q : C \rightarrow D : A \times C \rightarrow B \times D , \\ p \boxtimes q \triangleq a \times c \rightarrow p(a) \times q(c) , \\ (a \times c) \triangleright (p \boxtimes q) = (a \triangleright p) \times (b \triangleright q) . \end{aligned}$$

In this notation, the lifting for P is defined more concisely:

$$f^{\uparrow P} = f^{\uparrow L \times M} \triangleq f^{\uparrow L} \boxtimes f^{\uparrow M} \quad . \quad (6.13)$$

We need to verify the identity law and the composition law.

To verify the identity law of P , pipe an arbitrary value of type $L^A \times M^A$ into both sides of the law:

$$\begin{aligned} \text{expect to equal } l \times m : & (l^{\uparrow L^A} \times m^{\uparrow M^A}) \triangleright \text{id}^{\uparrow P} \\ \text{definition of } f^{\uparrow P} : & = (l \triangleright \text{id}^{\uparrow L}) \times (m \triangleright \text{id}^{\uparrow M}) \\ \text{identity laws of } L \text{ and } M : & = (l \triangleright \text{id}) \times (m \triangleright \text{id}) \\ \text{definition of id} : & = l \times m \quad . \end{aligned}$$

To verify the composition law of P , we need to show that

$$f^{\uparrow P} ; g^{\uparrow P} = (f ; g)^{\uparrow P} \quad .$$

Apply both sides of this equation to an arbitrary value of type $L^A \times M^A$:

$$\begin{aligned} \text{expect to equal } (l \times m) \triangleright (f ; g)^{\uparrow P} : & (l^{\uparrow L^A} \times m^{\uparrow M^A}) \triangleright f^{\uparrow P} ; g^{\uparrow P} \\ \triangleright \text{notation} : & = (l^{\uparrow L^A} \times m^{\uparrow M^A}) \triangleright f^{\uparrow P} \triangleright g^{\uparrow P} \\ \text{use Eq. (6.12)} : & = ((l \triangleright f^{\uparrow L}) \times (m \triangleright f^{\uparrow M})) \triangleright g^{\uparrow P} \\ \text{use Eq. (6.12)} : & = (l \triangleright f^{\uparrow L} \triangleright g^{\uparrow L}) \times (m \triangleright f^{\uparrow M} \triangleright g^{\uparrow M}) \\ \triangleright \text{notation} : & = (l \triangleright f^{\uparrow L} ; g^{\uparrow L}) \times (m \triangleright f^{\uparrow M} ; g^{\uparrow M}) \\ \text{composition laws of } L \text{ and } M : & = (l \triangleright (f ; g)^{\uparrow L}) \times (m \triangleright (f ; g)^{\uparrow M}) \\ \text{use Eq. (6.12)} : & = (l \times m) \triangleright (f ; g)^{\uparrow P} \quad . \end{aligned}$$

The calculations are shorter if we use the pair product operation:

$$\begin{aligned} \text{expect to equal } (f ; g)^{\uparrow P} : & f^{\uparrow P} ; g^{\uparrow P} = (f^{\uparrow L} \boxtimes f^{\uparrow M}) ; (g^{\uparrow L} \boxtimes g^{\uparrow M}) \\ \text{composition of functions under } \boxtimes : & = (f^{\uparrow L} ; g^{\uparrow L}) \boxtimes (f^{\uparrow M} ; g^{\uparrow M}) \\ \text{composition laws of } L \text{ and } M : & = (f ; g)^{\uparrow L} \boxtimes (f ; g)^{\uparrow M} = (f ; g)^{\uparrow P} \quad . \end{aligned}$$

For comparison, the same derivation using the Scala code syntax looks like this,

```
((1, m).map(f).map(g) == ((1.map(f), m.map(f))).map(g)
  == ((1.map(f).map(g), m.map(f).map(g)))
  == ((1.map(f) andThen g), m.map(f andThen g)))
```

assuming that the `map` method is defined on pairs by Eq. (6.12),

```
((1, m).map(f) == ((1.map(f), m.map(f))))
```

The proof written in the Scala syntax does not show the type constructors whose `map` methods are used in each expression. For instance, it is not indicated that the two `map` methods used in the expression `m.map(f).map(g)` belong to the *same* type constructor M and thus obey M 's composition law. The code notation shows this more concisely and more clearly, helping us in reasoning:

$$m \triangleright f^{\uparrow M} \triangleright g^{\uparrow M} = m \triangleright f^{\uparrow M} ; g^{\uparrow M} = m \triangleright (f ; g)^{\uparrow M} \quad .$$

By the convention of the pipe notation, it groups to the left, so we have

$$(x \triangleright f) \triangleright g = x \triangleright f \triangleright g = x \triangleright f ; g = x \triangleright (f ; g) = (f ; g)(x) = g(f(x)) \quad .$$

We will often use this notation in code derivations. (Chapter 7 gives an overview of the techniques of code derivation, including some more details about the pipe notation.)

Statement 6.2.3.4 If P^A and Q^A are functors then $L^A \triangleq P^A + Q^A$ is a functor, with `fmap` defined by

```
def fmap[A, B](f: A => B): Either[P[A], Q[A]] => Either[P[B], Q[B]] = {
  case Left(pa)  => Left(fmap_P(f)(pa))  // Use fmap for P.
  case Right(qa)  => Right(fmap_Q(f)(qa)) // Use fmap for Q.
}
```

The functor L^\bullet is the **functor co-product** of P^\bullet and Q^\bullet . The code notation for the `fmap` function is

$$\text{fmap}_L(f: A \rightarrow B) = f^{\uparrow L} \triangleq \begin{vmatrix} & P^B & Q^B \\ \hline P^A & f^{\uparrow P} & \mathbf{0} \\ Q^A & \mathbf{0} & f^{\uparrow Q} \end{vmatrix}.$$

Here we assume that lawful `fmap` functions are given for the functors P and Q .

Proof Omitting the type annotations, we write the code of $\text{fmap}_L(f)$ as

$$\text{fmap}_L(f) = f^{\uparrow L} = \begin{vmatrix} f^{\uparrow P} & \mathbf{0} \\ \mathbf{0} & f^{\uparrow Q} \end{vmatrix}. \quad (6.14)$$

To verify the identity law, use Eq. (6.14) and the identity laws for P and Q :

$$\begin{aligned} \text{expect to equal id : } \text{id}^{\uparrow L} &= \begin{vmatrix} \text{id}^{\uparrow P} & \mathbf{0} \\ \mathbf{0} & \text{id}^{\uparrow Q} \end{vmatrix} = \begin{vmatrix} \text{id} & \mathbf{0} \\ \mathbf{0} & \text{id} \end{vmatrix} \\ \text{identity function in matrix notation : } &= \text{id} \end{aligned}$$

To verify the composition law:

$$\begin{aligned} \text{expect to equal } (f \circ g)^{\uparrow L} : \quad f^{\uparrow L} \circ g^{\uparrow L} &= \begin{vmatrix} f^{\uparrow P} & \mathbf{0} \\ \mathbf{0} & f^{\uparrow Q} \end{vmatrix} \circ \begin{vmatrix} g^{\uparrow P} & \mathbf{0} \\ \mathbf{0} & g^{\uparrow Q} \end{vmatrix} \\ \text{matrix composition : } &= \begin{vmatrix} f^{\uparrow P} \circ g^{\uparrow P} & \mathbf{0} \\ \mathbf{0} & f^{\uparrow Q} \circ g^{\uparrow Q} \end{vmatrix} \\ \text{composition laws of } P \text{ and } Q : &= \begin{vmatrix} (f \circ g)^{\uparrow P} & \mathbf{0} \\ \mathbf{0} & (f \circ g)^{\uparrow Q} \end{vmatrix} = (f \circ g)^{\uparrow L} \end{aligned}$$

The last two statements show that a type constructor built up via primitive types, type parameters, products and co-products, such as $L^A \triangleq 1 + (\text{String} + A) \times A \times \text{Int} + A$, is a functor. Functors of this kind are called **polynomial functors** because they are analogous to ordinary arithmetic polynomial functions of a variable A . The type notation with its symbols ($+$, \times) makes this analogy visually clear.

Implementing `fmap` for a polynomial functor is straightforward: `fmap` replaces each occurrence of the A value of type A by the corresponding value of type B , leaving constant types unchanged and keeping the order of parts in all products and disjunctive types. Previously, our implementations of `fmap` for various type constructors (such as shown in Example 6.1.9.2) were guided by the idea of preserving information. Statements 6.2.3.3–6.2.3.4 explain why those implementations of the `fmap` are correct (i.e., obey the functor laws).

The next construction shows when a function type is a functor: the argument of the function must be a contrafunctor.

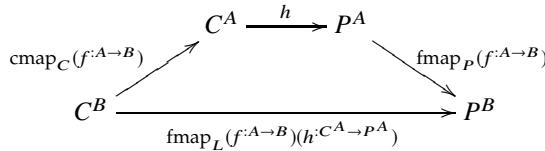
Statement 6.2.3.5 If C is a contrafunctor and P is a functor then $L^A \triangleq C^A \rightarrow P^A$ is a functor, called an **functor exponential**, with `fmap` defined by

$$\begin{aligned} \text{fmap}_L^{A,B}(f: A \rightarrow B) : (C^A \rightarrow P^A) &\rightarrow C^B \rightarrow P^B, \\ \text{fmap}_L(f: A \rightarrow B) = f^{\uparrow L} &\triangleq h: C^A \rightarrow P^A \rightarrow f^{\downarrow C} \circ h \circ f^{\uparrow P} \end{aligned} \quad (6.15)$$

The corresponding Scala code is

```
def fmap_L[A, B](f: A => B)(h: C[A] => P[A]): C[B] => P[B] = {
  cmap_C(f) andThen h andThen fmap_P(f)
}
```

A type diagram for $fmap_L$ can be drawn as



Proof Since the types are already checked, we can use Eq. (6.15) without type annotations,

$$h \triangleright f^{\uparrow L} = f^{\downarrow C} ; h ; f^{\uparrow P} . \quad (6.16)$$

To verify the identity law of L , show that $\text{id}^{\uparrow L}(h) = h$:

$$\begin{aligned}
 & \text{expect to equal } h : h \triangleright \text{id}^{\uparrow L} \\
 & \text{definition (6.16) of } \uparrow L : = \underline{\text{id}^{\downarrow C}} ; h ; \underline{\text{id}^{\uparrow P}} \\
 & \text{identity laws of } C \text{ and } P : = \underline{\text{id}} ; h ; \underline{\text{id}} \\
 & \text{definition of id} : = h .
 \end{aligned}$$

To verify the composition law of L , it helps to apply both sides of the law to an arbitrary $h^{\cdot L^A}$:

$$\begin{aligned}
 & \text{expect to equal } h \triangleright f^{\uparrow L} ; g^{\uparrow L} : h \triangleright (f ; g)^{\uparrow L} \\
 & \text{definition (6.16) of } \uparrow L : = (\underline{f} ; \underline{g})^{\downarrow C} ; h ; (\underline{f} ; \underline{g})^{\uparrow P} \\
 & \text{composition laws of } C \text{ and } P : = g^{\downarrow C} ; \underline{f^{\downarrow C}} ; h ; \underline{f^{\uparrow P}} ; g^{\uparrow P} \\
 & \text{definition (6.16) of } \uparrow L : = \underline{g^{\downarrow C}} ; (h \triangleright f^{\uparrow L}) ; \underline{g^{\uparrow P}} \\
 & \text{definition (6.16) of } \uparrow L : = (h \triangleright f^{\uparrow L}) \triangleright g^{\uparrow L} = h \triangleright f^{\uparrow L} ; g^{\uparrow L} .
 \end{aligned}$$

It is important for this proof that C is a contrafunctor and so the order of lifted function compositions is reversed, $(f ; g)^{\downarrow C} = g^{\downarrow C} ; f^{\downarrow C}$. If C were a functor, the proof would not work: we would have obtained $f^{\uparrow C} ; g^{\uparrow C}$ instead of $g^{\downarrow C} ; f^{\downarrow C}$, and we would not be able to group $f^{\downarrow C} ; h ; f^{\uparrow P}$ together (the order of composition cannot be permuted for arbitrary functions f, g).

Examples of functors obtained via the exponential construction are $L^A \triangleq Z \rightarrow A$ (with the contrafunctor C^A chosen as the constant contrafunctor Z , where Z is a fixed type) and $L^A \triangleq (A \rightarrow Z) \rightarrow A$ (with the contrafunctor $C^A \triangleq A \rightarrow Z$). Statement 6.2.3.5 generalizes those examples to arbitrary contrafunctors C^A used as arguments of function types.

Similarly, one can prove that $P^A \rightarrow C^A$ is a contrafunctor (Exercise 6.3.1.2). Together with Statements 6.2.3.3–6.2.3.5, this gives us the rules of reasoning about covariance and contravariance of type parameters in arbitrary type expressions. Every function arrow (\rightarrow) flips the variance from covariant to contravariant and back. For instance, the identity functor $L^A \triangleq A$ is covariant in A , while $A \rightarrow Z$ is contravariant in A , and $(A \rightarrow Z) \rightarrow Z$ is again covariant in A . As we have seen, $A \rightarrow A \rightarrow Z$ is contravariant in A , so any number of curried arrows count as one in this reasoning (and, in any case, $A \rightarrow A \rightarrow Z \cong A \times A \rightarrow Z$). Products and disjunctions do not change variance, so $(A \rightarrow Z_1) \times (A \rightarrow Z_2) + (A \rightarrow Z_3)$ is contravariant in A . This is shown in more detail in Section 6.2.5.

The remaining constructions set a type parameter to another type. The **functor composition** P^Q^A , written in Scala as $P[Q[A]]$, is analogous to a function composition such as $f(g(x))$ except for using type constructors. Viewed in this way, type constructors are **type-level functions** (i.e., maps on the set of types). So, functor composition may be denoted by $P \circ Q$, like the function composition $f \circ g$.

An example of functor composition in Scala is `List[Option[Int]]`. Since both `List` and `Option` have a `map` method, we may write code such as

```
val p: List[Option[Int]] = List(Some(1), None, Some(2), None, Some(3))

scala> p.map(_.map(x => x + 10))
res0: List[Option[Int]] = List(Some(11), None, Some(12), None, Some(13))
```

The code `p.map(_.map(f))` lifts an $f:A \rightarrow B$ into a function of type `List[Option[A]] \Rightarrow List[Option[B]]`. In this way, we may perform the `map` operation on the nested data type `List[Option[_]]`.

The next statement shows that this code always produces a lawful `map` function. In other words, the composition of functors is always a functor.

Statement 6.2.3.6 If P^A and Q^A are functors then $L^A \triangleq P^{Q^A}$ is also a functor, with `fmap` defined by

```
def fmap_L[A, B](f: A => B): P[Q[A]] => P[Q[B]] = fmap_P(fmap_Q(f))
```

Here we assumed that the functions $fmap_P$ and $fmap_Q$ are known and satisfy the functor laws.

In the code notation, $fmap_L$ is written equivalently as

type signature : $fmap_L : f:A \rightarrow B \rightarrow P^{Q^A} \rightarrow P^{Q^B}$,

implementation : $fmap_L(f) \triangleq fmap_P(fmap_Q(f))$,

equivalent code : $fmap_L \triangleq fmap_Q \circ fmap_P$,

in a shorter notation : $f^{\uparrow L} \triangleq (f^{\uparrow Q})^{\uparrow P} \triangleq f^{\uparrow Q \uparrow P}$.

(6.17)

(6.18)

Proof To verify the identity law of L , use the identity laws for P and Q :

$$\text{id}^{\uparrow L} = (\text{id}^{\uparrow Q})^{\uparrow P} = \text{id}^{\uparrow P} = \text{id} .$$

To verify the composition law of L , use the composition laws for P and Q :

$$(f \circ g)^{\uparrow L} = ((f \circ g)^{\uparrow Q})^{\uparrow P} = (f^{\uparrow Q} \circ g^{\uparrow Q})^{\uparrow P} = f^{\uparrow Q \uparrow P} \circ g^{\uparrow Q \uparrow P} .$$

Finally, we consider recursive data types such as lists and trees (Section 3.3). It is helpful to use the type notation for reasoning about those types. The list type,

```
sealed trait List[A]
final case class Empty() extends List[A]
final case class Head[A](head: A, tail: List[A]) extends List[A]
```

is written in type notation as

$$\text{List}^A \triangleq \mathbb{1} + A \times \text{List}^A .$$

The binary tree type,

```
sealed trait Tree2[A]
final case class Leaf[A](a: A) extends Tree2[A]
final case class Branch[A](x: Tree2[A], y: Tree2[A]) extends Tree2[A]
```

is defined by $\text{Tree}_2^A \triangleq A + \text{Tree}_2^A \times \text{Tree}_2^A$. Such definitions of recursive types look like “type equations”. We can generalize these examples to a recursive definition

$$L^A \triangleq S^{A, L^A} ,$$
(6.19)

where $S^{A, R}$ is a suitably chosen type constructor with two type parameters A, R . If the type constructor $S^{\bullet, \bullet}$ is given, the Scala code defining L^{\bullet} can be written as

```
type S[A, R] = ... // Must be defined previously as type alias, class, or trait.
final case class L[A](x: S[A, L[A]])
```

Description	Type definition	Bifunctor $S^{A,R}$
list	$L^A \triangleq \mathbb{1} + A \times L^A$	$S^{A,R} \triangleq \mathbb{1} + A \times R$
non-empty list	$NEL^A \triangleq A + A \times NEL^A$	$S^{A,R} \triangleq A + A \times R$
list of odd length	$L^A \triangleq A + A \times A \times L^A$	$S^{A,R} \triangleq A + A \times A \times R$
binary tree	$L^A \triangleq A + L^A \times L^A$	$S^{A,R} \triangleq A + R \times R$
rose tree	$L^A \triangleq A + NEL^{L^A}$	$S^{A,R} \triangleq A + NEL^R$
regular-shaped binary tree	$L^A \triangleq A + L^{A \times A}$	not possible
abstract syntax tree	$L^A \triangleq P^A + Q^{L^A}$	$S^{A,R} \triangleq P^A + Q^R$

Table 6.3: Recursive disjunctive types defined using type equations.

We must use a case class to define L because Scala does not support recursive type aliases:

```
scala> type L[A] = Either[A, L[A]]
<console>:14: error: illegal cyclic reference involving type L
          type L[A] = Either[A, L[A]]
                           ^

scala> final case class L[A](x: Either[A, L[A]])
defined class L
```

Table 6.3 summarizes our previous examples of recursive disjunctive types and shows the relevant choices of $S^{A,R}$, which turns out to be always a bifunctor. For abstract syntax trees, the functors P^{\bullet} and Q^{\bullet} must be given; they specify the available shapes of leaves and branches respectively.

We will now prove that Eq. (6.19) always defines a functor when $S^{\bullet,\bullet}$ is a bifunctor.

Statement 6.2.3.7 If $S^{A,B}$ is a bifunctor (a functor with respect to both type parameters A and B) then the recursively defined type constructor L^A is a functor,

$$L^A \triangleq S^{A,L^A} .$$

The `fmap` method for L is a recursive function implemented as

$$\text{fmap}_L(f: A \rightarrow B) \triangleq \text{bimap}_S(f)(\text{fmap}_L(f)) . \quad (6.20)$$

The corresponding Scala code is

```
final case class L[A](x: S[A, L[A]]) // The type constructor S[_, _] must be defined previously.

def bimap_S[A, B, C, D](f: A => C)(g: B => D): S[A, B] => S[C, D] = ??? // Must be defined.

def fmap_L[A, B](f: A => B): L[A] => L[B] = { case L(x) =>
  val newX: S[B, L[B]] = bimap_S(f)(fmap_L(f))(x) // Recursive call to fmap_L.
  L(newX) // Need to wrap the value of type S[B, L[B]] into the type constructor L.
}
```

Proof Usually, laws for a recursive function (such as fmap_L) must be proved by induction. In the recursive implementation of fmap_L , its code calls itself in some cases but returns without recursive calls in other cases. So, the base case of induction corresponds to the non-recursive evaluations in the code of fmap_L , and we need to prove that the law is then satisfied. The inductive step must prove that the code of fmap_L obeys the law under the inductive assumption that all recursive calls to fmap_L already obey that law. In the proof, we do not need to separate the base case from the inductive step; we just derive the law using the inductive assumption whenever needed.

For clarity, we add an overline to recursive calls in the code formula:

$$\text{fmap}_L(f) \triangleq \text{bimap}_S(f)(\overline{\text{fmap}_L(f)}) .$$

To prove the identity law:

expect to equal id : $\text{fmap}_L(\text{id})$
 definition of fmap_L : $= \text{bimap}_S(\text{id})(\overline{\text{fmap}_L}(\text{id}))$
 inductive assumption — the law holds for $\overline{\text{fmap}_L}$: $= \text{bimap}_S(\text{id})(\text{id})$
 identity law of S : $= \text{id}$.

To prove the composition law:

expect to equal $\text{fmap}_L(f \circ g)$: $\text{fmap}_L(f) \circ \text{fmap}_L(g)$
 definition of fmap_L : $= \text{bimap}_S(f)(\overline{\text{fmap}_L}(f)) \circ \text{bimap}_S(g)(\overline{\text{fmap}_L}(g))$
 composition law of S : $= \text{bimap}_S(f \circ g)(\overline{\text{fmap}_L}(f) \circ \overline{\text{fmap}_L}(g))$
 inductive assumption : $= \text{bimap}_S(f \circ g)(\overline{\text{fmap}_L}(f \circ g))$
 definition of fmap_L : $= \text{fmap}_L(f \circ g)$.

For the regular-shaped binary tree, the construction (6.19) is insufficient: no bifunctor S^{A,L^A} can replace the type argument A in L^A to obtain $L^{A \times A}$. To see that, consider that S^{A,L^A} is an application of a type-level function $S^{\bullet,\bullet}$ to its two type parameters, which are set to A and L^A . In Scala syntax, S^{A,L^A} is written as s[A,L[A]] . No matter how we define the type constructor S , the resulting type expression s[A,L[A]] will always use the type constructor L as L[A] and not as L[(A,A)] .

To describe regular-shaped trees, we need to modify the construction by adding another arbitrary functor, P^\bullet , in the type argument of L^\bullet :

$$L^A \triangleq S^A + L^{P^A} . \quad (6.21)$$

Regular-shaped binary trees are defined by Eq. (6.21) with $S^A \triangleq A$ and $P^A \triangleq A \times A$. The Scala code for these definitions is

```
type S[A] = A                                // The shape of a leaf.
type P[A] = (A, A)                            // The shape of a branch.
final case class L[A](s: Either[S[A], L[P[A]]]) // Or 'case class L[A](s: Either[A, L[(A, A)]])'.
```

Different choices of P will define regular-shaped trees with different kinds of branching.

6.2.4 Constructions of contrafunctors

The previous section performed **structural analysis** for functors: a systematic search for type constructions (product, co-product, etc.) that create new functors. *Mutatis mutandis*, similar constructions work for contrafunctors, as shown in Table 6.4. One difference with respect to Table 6.2 is the absence of the identity type constructor, $L^A \triangleq A$ (it is a functor, not a contrafunctor). However, the constant type constructor, $L^A \triangleq Z$, is a functor and a contrafunctor at the same time.

Let us now prove the validity of some of these constructions.

Statement 6.2.4.1 If Z is any fixed type, the constant type constructor $C^A \triangleq Z$ is a contrafunctor (the **constant contrafunctor**) whose cmap returns an identity function of type $Z \rightarrow Z$:

```
type Const[Z, A] = Z
def cmap[Z, A, B](f: B => A): Const[Z, A] => Const[Z, B] = identity[Z]
```

Proof All laws hold because cmap returns an identity function:

identity law : $\text{cmap}(\text{id}) = \text{id}$,
 composition law : $\text{cmap}(f) \circ \text{cmap}(g) = \text{id} \circ \text{id} = \text{id} = \text{cmap}(g \circ f)$.

Construction	Type notation	Comment
tuple	$C^A \triangleq P^A \times Q^A$	the product contrafunctor; P and Q must be contrafunctors
disjunctive type	$C^A \triangleq P^A + Q^A$	the co-product contrafunctor; P and Q must be contrafunctors
function type	$C^A \triangleq L^A \rightarrow H^A$	the exponential contrafunctor; L is a functor and H a contrafunctor
type parameter	$C^A \triangleq Z$	the constant contrafunctor; Z is a fixed type
type constructor	$C^A \triangleq P^Q^A$	the composition; P is a functor and Q a contrafunctor (or vice versa)
recursive type	$C^A \triangleq S^{A,C^A}$	$S^{A,B}$ must be a contrafunctor w.r.t. A and functor w.r.t. B

Table 6.4: Type constructions defining a contrafunctor C^A .

Statement 6.2.4.2 If P^A is a functor and Q^A is a contrafunctor then $L^A \triangleq P^Q^A$ is a contrafunctor with `cmap` defined by

```
def cmap[A, B](f: B => A): P[Q[A]] => P[Q[B]] = fmap_P(cmap_Q(f))
```

where lawful implementations of `fmap_P` and `cmap_Q` are assumed to be given.

Proof Convert the Scala implementation of `cmap_L` into the code notation:

$$\text{cmap}_L(f: B \rightarrow A) \triangleq \text{fmap}_P(\text{cmap}_Q(f)) .$$

It is easier to reason about this function if we rewrite it as

$$f^{\downarrow L} \triangleq (f^{\downarrow Q})^{\uparrow P} .$$

The contrafunctor laws for L are then proved like this:

$$\begin{aligned} \text{identity law : } \text{id}^{\downarrow L} &= (\text{id}^{\downarrow Q})^{\uparrow P} = \text{id}^{\uparrow P} = \text{id} . \\ \text{composition law : } f^{\downarrow L} ; g^{\downarrow L} &= (f^{\downarrow Q})^{\uparrow P} ; (g^{\downarrow Q})^{\uparrow P} \\ \text{use } P\text{'s composition law : } &= (f^{\downarrow Q} ; g^{\downarrow Q})^{\uparrow P} = ((g ; f)^{\downarrow Q})^{\uparrow P} = (g ; f)^{\downarrow L} . \end{aligned}$$

Finally, the recursive construction works for contrafunctors, except that the type constructor $S^{A,R}$ must be a contrafunctor in A (but still a functor in R). An example of such a type constructor is

$$S^{A,R} \triangleq (A \rightarrow \text{Int}) + R \times R . \quad (6.22)$$

The type constructor $S^{\bullet,\bullet}$ is not a bifunctor because it is contravariant in its first type parameter; so we cannot define a `bimap` function for it. However, we can define an analogous function called `xmap`, with the type signature

```
def xmap[A, B, Q, R](f: B => A)(g: Q => R): S[A, Q] => S[B, R]
```

$$\begin{aligned} \text{xmap}_S: (B \rightarrow A) \rightarrow (Q \rightarrow R) \rightarrow S^{A,Q} \rightarrow S^{B,R} , \\ \text{xmap}_S(f: B \rightarrow A)(g: Q \rightarrow R) \triangleq \text{fmap}_{S^{A,\bullet}}(g) ; \text{cmap}_{S^{\bullet,R}}(f) . \end{aligned}$$

The function `xmap` should obey the laws of identity and composition:

$$\text{identity law : } \text{xmap}_S(\text{id})(\text{id}) = \text{id} , \quad (6.23)$$

$$\text{composition law : } \text{xmap}_S(f_1)(g_1) ; \text{xmap}_S(f_2)(g_2) = \text{xmap}_S(f_2 ; f_1)(g_1 ; g_2) . \quad (6.24)$$

These laws are similar to the identity and composition laws for bifunctors (Section 6.2.2), except for inverting the order of the composition ($f_2 ; f_1$). The laws hold automatically whenever the functor

and contrafunctor methods for S ($\text{fmap}_{S^{A,\bullet}}$ and $\text{cmap}_{S^{\bullet,R}}$) are fully parametric. We omit the details since they are quite similar to what we saw in Section 6.2.2 for bifunctors.

If we define a type constructor L^\bullet using the recursive “type equation”

$$L^A \triangleq S^{A,L^A} \triangleq (A \rightarrow \text{Int}) + L^A \times L^A \quad ,$$

we obtain a contrafunctor in the shape of a binary tree whose leaves are functions of type $A \rightarrow \text{Int}$. The next statement shows that recursive type equations of this kind always define contrafunctors.

Statement 6.2.4.3 If $S^{A,R}$ is a contrafunctor with respect to A and a functor with respect to R then the recursively defined type constructor C^A is a contrafunctor,

$$C^A \triangleq S^{A,C^A} \quad .$$

Given the functions $\text{cmap}_{S^{\bullet,R}}$ and $\text{fmap}_{S^{A,\bullet}}$ for S , we implement cmap_C as

$$\begin{aligned} \text{cmap}_C(f: B \rightarrow A) : C^A \rightarrow C^B &\cong S^{A,C^A} \rightarrow S^{B,C^B} \quad , \\ \text{cmap}_C(f: B \rightarrow A) &\triangleq \text{xmap}_S(f)(\overline{\text{cmap}_C}(f)) \quad . \end{aligned}$$

The corresponding Scala code can be written as

```
final case class C[A](x: S[A, C[A]]) // The type constructor S[_,_] must be defined previously.

def xmap_S[A,B,Q,R](f: B => A)(g: Q => R): S[A, Q] => S[B, R] = ??? // Must be defined.

def cmap_C[A, B](f: B => A): C[A] => C[B] = { case C(x) =>
  val sbcb: S[B, C[B]] = xmap_S(f)(cmap_C(f))(x) // Recursive call to cmap_C.
  C(sbcb) // Need to wrap the value of type S[B, C[B]] into the type constructor C.
}
```

Proof The code of cmap is recursive, and the recursive call is marked by an overline:

$$\text{cmap}_C(f) \triangleq f^{\downarrow C} \triangleq \text{xmap}_S(f)(\overline{\text{cmap}_C}(f)) \quad .$$

To verify the identity law:

$$\begin{aligned} \text{expect to equal id : } \text{cmap}_C(\text{id}) &= \text{xmap}_S(\text{id})(\overline{\text{cmap}_C}(\text{id})) \\ \text{inductive assumption : } &= \text{xmap}_S(\text{id})(\text{id}) \\ \text{identity law of } \text{xmap}_S : &= \text{id} \quad . \end{aligned}$$

To verify the composition law:

$$\begin{aligned} \text{expect to equal } (g^{\downarrow C} ; f^{\downarrow C}) : (f: D \rightarrow B ; g: B \rightarrow A)^{\downarrow C} &= \text{xmap}_S(f ; g)(\overline{\text{cmap}_C}(f ; g)) \\ \text{inductive assumption : } &= \text{xmap}_S(f ; g)(\overline{\text{cmap}_C}(g) ; \overline{\text{cmap}_C}(f)) \\ \text{composition law of } \text{xmap}_S : &= \text{xmap}_S(g)(\overline{\text{cmap}_C}(g)) ; \text{xmap}_S(f)(\overline{\text{cmap}_C}(f)) \\ \text{definition of } \downarrow C : &= g^{\downarrow C} ; f^{\downarrow C} \quad . \end{aligned}$$

6.2.5 Solved examples: How to recognize functors and contrafunctors

Sections 6.2.3 and 6.2.4 describe how functors and contrafunctors are built from other type expressions. We can see from Tables 6.2 and 6.4 that *every* one of the six basic type constructions (unit type, type parameters, product types, co-product types, function types, recursive types) gives either a new functor or a new contrafunctor. The six type constructions generate all exponential-polynomial types, including recursive ones. So, we should be able to decide whether any given exponential-polynomial type expression is a functor or a contrafunctor. The decision algorithm is based on the results shown in Tables 6.2 and 6.4:

- Primitive types $\mathbb{1}$, `Int`, `String`, etc., can be viewed both as constant functors and as constant contrafunctors (since they do not contain type parameters).
- Polynomial type expressions (not containing any function arrows) are always functors with respect to every type parameter. Equivalently, we may say that all polynomial type constructors are covariant in every type parameter. For example, the type expression $A \times B + (A + \mathbb{1} + B) \times A \times C$ is covariant in each of the type parameters A, B, C .
- Type parameters to the right of a function arrow are in a covariant position. For example, $\text{Int} \rightarrow A$ is covariant in A .
- Each time a type parameter is placed to the left of an *uncurried* function arrow \rightarrow , the variance is reversed: covariant becomes contravariant and vice versa. For example,

this is covariant in A : $\mathbb{1} + A \times A$,
 this is contravariant in A : $(\mathbb{1} + A \times A) \rightarrow \text{Int}$,
 this is covariant in A : $((\mathbb{1} + A \times A) \rightarrow \text{Int}) \rightarrow \text{Int}$,
 this is contravariant in A : $((((\mathbb{1} + A \times A) \rightarrow \text{Int}) \rightarrow \text{Int}) \rightarrow \text{Int})$.

- Repeated curried function arrows work as one arrow: $A \rightarrow \text{Int}$ is contravariant in A , and $A \rightarrow A \rightarrow A \rightarrow \text{Int}$ is still contravariant in A . This is because the type $A \rightarrow A \rightarrow A \rightarrow \text{Int}$ is equivalent to $A \times A \times A \rightarrow \text{Int}$, which is of the form $F^A \rightarrow \text{Int}$ with a type constructor $F^A \triangleq A \times A \times A$. Exercise 6.3.1.1 will show that $F^A \rightarrow \text{Int}$ is contravariant in A .
- Nested type constructors combine their variances: e.g., if we know that F^A is contravariant in A then $F^{A \rightarrow \text{Int}}$ is covariant in A , while $F^{A \times A \times A}$ is contravariant in A .

For any exponential-polynomial type expression, such as Eq. (6.7),

$$Z^{A,R} \triangleq ((A \rightarrow A \rightarrow R) \rightarrow R) \times A + (\mathbb{1} + R \rightarrow A + \text{Int}) + A \times A \times \text{Int} \times \text{Int} ,$$

we mark the position of each type parameter as either covariant (+) or contravariant (-), according to the number of nested function arrows:

$$((A \underset{+}{\rightarrow} A \underset{+}{\rightarrow} R) \underset{+}{\rightarrow} R) \times A + (\mathbb{1} + R \underset{-}{\rightarrow} A \underset{+}{+} \text{Int}) + A \underset{+}{\times} A \underset{+}{\times} \text{Int} \times \text{Int} .$$

We find that A is always in covariant positions, while R is sometimes in covariant and sometimes in contravariant positions. So, we expect that $Z^{A,R}$ is a functor with respect to A , but not a functor (nor a contrafunctor) with respect to R .

To show that $Z^{A,R}$ is indeed a functor in the parameter A , we need to implement a suitable `map` method and verify that the functor laws hold. To do that from scratch, we could use the techniques explained in this and the previous chapters: starting from the type signature

$$\text{map}_Z : Z^{A,R} \rightarrow (A \rightarrow B) \rightarrow Z^{B,R} ,$$

we could derive a fully parametric, information-preserving implementation of `map`. We could then look for proofs of the identity and composition laws for that `map` function. This would require a lot of work for a complicated type constructor such as $Z^{A,R}$.

However, that work can be avoided if we find a way of building up $Z^{A,R}$ step by step via the known functor and contrafunctor constructions. Each step automatically provides both a fragment of the code of `map` and a proof that the functor laws hold up to that step. In this way, we will avoid the need to look for an implementation of `map` and proofs of laws for each new functor and contrafunctor. The next examples illustrate the procedure for a simpler type constructor.

Example 6.2.5.1 Rewrite this Scala definition in the type notation and decide whether it is covariant or contravariant with respect to each type parameter:

```
final case class G[A, Z](p: Either[Int, A], q: Option[Z => Int => Z => (Int, A)])
```

Solution The type notation for G is $G^{A,Z} \triangleq (\text{Int} + A) \times (\mathbb{1} + (Z \rightarrow \text{Int} \rightarrow Z \rightarrow \text{Int} \times A))$. Mark the covariant and the contravariant positions in this type expression:

$$(\text{Int} + A) \times (\mathbb{1} + (Z \rightarrow \text{Int} \rightarrow Z \rightarrow \text{Int} \times A)) .$$

All Z positions in the sub-expression $Z \rightarrow \text{Int} \rightarrow Z \rightarrow \text{Int} \times A$ are contravariant since the function arrows are curried rather than nested. We see that A is always in covariant positions (+) while Z is always in contravariant positions (-). It follows that $G^{A,Z}$ is covariant in A and contravariant in Z .

Example 6.2.5.2 Use known functor constructions to implement the `map` method with respect to A for the type $G[A, Z]$ from Example 6.2.5.1.

Solution We need to build $G^{A,Z}$ via step-by-step constructions that start from primitive types and type parameters. At the top level of its type expression, $G^{A,Z}$ is a product type. So, we begin by using the “functor product” construction (Statement 6.2.3.3):

$$G^{A,Z} \cong G_1^A \times G_2^{A,Z} ,$$

where $G_1^A \triangleq \text{Int} + A$ and $G_2^{A,Z} \triangleq \mathbb{1} + (Z \rightarrow \text{Int} \rightarrow Z \rightarrow \text{Int} \times A) .$

We continue with G_1^A , which is a co-product of Int (a constant functor) and A (the identity functor). The constant functor and the identity functor have lawful `map` implementations that are already known (Statements 6.2.3.1–6.2.3.2). Now, the “functor co-product” construction (Statement 6.2.3.3) produces a `map` implementation for G_1^A together with a proof that it satisfies the functor laws:

$$\text{fmap}_{G_1}(f: A \rightarrow B) = f^{\uparrow G_1} \triangleq \begin{array}{c|cc} & \text{Int} & B \\ \hline \text{Int} & \text{id} & \mathbb{0} \\ A & \mathbb{0} & f \end{array} .$$

Turning our attention to $G_2^{A,Z}$, we find that it is a disjunctive type containing a curried function type that ultimately returns the product type $\text{Int} \times A$. This tells us to use the functor constructions for “co-product”, “exponential”, and “product”. Write down the functor constructions needed at each step as we decompose $G_2^{A,Z}$:

$$\begin{aligned} G_2^{A,Z} &\triangleq \mathbb{1} + (Z \rightarrow \text{Int} \rightarrow Z \rightarrow \text{Int} \times A) . \\ \text{co-product : } G_2^{A,Z} &\cong \mathbb{1} + G_3^{A,Z} \quad \text{where } G_3^{A,Z} \triangleq Z \rightarrow \text{Int} \rightarrow Z \rightarrow \text{Int} \times A . \\ \text{exponential : } G_3^{A,Z} &\cong Z \rightarrow G_4^{A,Z} \quad \text{where } G_4^{A,Z} \triangleq \text{Int} \rightarrow Z \rightarrow \text{Int} \times A . \\ \text{exponential : } G_4^{A,Z} &\cong \text{Int} \rightarrow G_5^{A,Z} \quad \text{where } G_5^{A,Z} \triangleq Z \rightarrow \text{Int} \times A . \\ \text{exponential : } G_5^{A,Z} &\cong Z \rightarrow G_6^A \quad \text{where } G_6^A \triangleq \text{Int} \times A . \\ \text{product : } G_6^A &\cong \text{Int} \times A \cong \text{Const}^{\text{Int}, A} \times \text{Id}^A . \end{aligned}$$

Each of the type constructors G_1, \dots, G_6 is a functor in A because all of the functor constructions preserve functor laws. Therefore, $G^{A,Z}$ is a functor in A .

It remains to derive the code for the `fmap` method of G . Each of the functor constructions combines the `fmap` implementations from previously defined functors into a new `map` implementation, so we just need to combine the code fragments in the order of constructions. For brevity, we will use the

notations $f^{\uparrow L} \triangleq \text{fmap}_L(f)$ and $x \triangleright f^{\uparrow L}$ instead of the Scala code `x.map(f)` throughout the derivations:

$$\begin{aligned} \text{product: } G^{A,Z} &\cong G_1^A \times G_2^{A,Z} \quad , \quad (g_1 \times g_2) \triangleright f^{\uparrow G} = (g_1 \triangleright f^{\uparrow G_1}) \times (g_2 \triangleright f^{\uparrow G_2}) \quad . \\ \text{co-product: } G_1^A &\triangleq \text{Int} + A \quad , \quad f^{\uparrow G_1} = \begin{vmatrix} \text{id} & 0 \\ 0 & f \end{vmatrix} \quad . \\ \text{co-product: } G_2^{A,Z} &\triangleq \text{Int} + G_3^{A,Z} \quad , \quad f^{\uparrow G_2} = \begin{vmatrix} \text{id} & 0 \\ 0 & f^{\uparrow G_3} \end{vmatrix} \quad . \\ \text{exponential: } G_3^{A,Z} &\triangleq Z \rightarrow G_4^{A,Z} \quad , \quad g_3 \triangleright f^{\uparrow G_3} = g_3 ; f^{\uparrow G_4} = z^{\cdot Z} \rightarrow z \triangleright g_3 \triangleright f^{\uparrow G_4} \quad . \end{aligned}$$

The pipe symbol groups stronger than the function arrow. So, $z \rightarrow z \triangleright g \triangleright h$ means $z \rightarrow (z \triangleright g \triangleright h)$, or $z \rightarrow h(g(z))$. Applying the exponential construction three times, we finally obtain

$$\begin{aligned} G_3^{A,Z} &\triangleq Z \rightarrow \text{Int} \rightarrow Z \rightarrow G_6^A \quad , \quad g_3 \triangleright f^{\uparrow G_3} = z_1^{\cdot Z} \rightarrow n^{\cdot \text{Int}} \rightarrow z_2^{\cdot Z} \rightarrow g_3(z_1)(n)(z_2) \triangleright f^{\uparrow G_6} \quad . \\ G_6^A &\triangleq \text{Int} \times A \quad , \quad (i \times a) \triangleright f^{\uparrow G_6} = i \times f(a) \quad . \end{aligned}$$

We can now write the corresponding Scala code for `fmapG`:

```
def fmap_G[A, B, Z](f: A => B): G[A, Z] => G[B, Z] = { case G(p, q) =>
  val newP: Either[Int, B] = p.map(f)           // Use the standard map method for Either[Int, A].
  val newQ: Option[Z => Int => Z => (Int, B)] = q.map { // Use the map method for Option[_].
    (g3: Z => Int => Z => (Int, A)) =>
    z1 => n => z2 =>                      // The code of map for G_3.
    val (i, a) = g3(z1)(n)(z2)
    (i, f(a))                                // The code of map for G_6.
  }
  G(newP, newQ)                                // The code of map for G_1.
}
```

In this way, the code of fmap_F can be unambiguously *derived* for any functor F from the type expression of F^A , and similarly the code for cmap_C for any contrafunctor C .

6.3 Summary

What tasks can we perform with the techniques of this chapter?

- Quickly decide if a given type constructor is a functor, a contrafunctor, or neither.
- Implement a `fmap` or a `cmap` function that satisfies the appropriate laws.
- Use constructions to derive the correct code of `fmap` or `cmap` without trial and error.
- Use functor blocks to manipulate data wrapped in functors with more readable code.

6.3.1 Exercises: Functor and contrafunctor constructions

Exercise 6.3.1.1 If H^A is a contrafunctor and L^A is a functor, show that $C \triangleq L^A \rightarrow H^A$ is a contrafunctor with the `cmap` method defined by the following code:

```
def cmap[A, B](f: B => A)(c: L[A] => H[A]): L[B] => H[B] = {
  1b: L[B] => cmap_H(f)(c(fmap_L(f)(1b)))                                // Code notation:  $f^{\downarrow C} \triangleq c \rightarrow f^{\uparrow L} ; c ; f^{\downarrow H}$ 
}
```

Here, `cmap_H` and `fmap_L` are the methods already defined for H and L . Prove that the laws hold.

Exercise 6.3.1.2 Implement the required `fmap` or `cmap` function for the given type constructors L and prove that the appropriate laws hold. Write the implementations both in Scala and in the code notation. Assume that the given type constructors F and G already satisfy their respective laws.

- (a) $L^A \triangleq F^A \times G^A$ is a contrafunctor if F^A and G^A are contrafunctors.
- (b) $L^A \triangleq F^A + G^A$ is a contrafunctor if F^A and G^A are contrafunctors.
- (c) $L^A \triangleq F^{G^A}$ is a functor when both F and G are contrafunctors.
- (d) $L^A \triangleq F^{G^A}$ is a contrafunctor when F is a contrafunctor and G is a functor.

Exercise 6.3.1.3 Show that the type constructor L defined by Eq. (6.21) is a functor for any given bifunctor S and functor P .

Exercise 6.3.1.4 Show that $L^A \triangleq F^A \rightarrow G^A$ is, in general, neither a functor nor a contrafunctor when both F^A and G^A are functors or both are contrafunctors (an example of suitable F^A and G^A is sufficient).

Exercise 6.3.1.5 For each of the Scala type constructors defined below, formulate the definition in the type notation and decide whether the type constructors are functors, contrafunctors, or neither.

```
type F[A] = Int => (Option[A], Either[A, Int], Either[A, A])
type G[A] = ((Int, A)) => Either[Int, A]
type H[A] = Either[A, (A, Option[A])] => Int => Int
```

Exercise 6.3.1.6 Using the known constructions, determine which of the following are functors or contrafunctors (or neither) and implement `fmap` or `cmap` if appropriate. Answer this question with respect to each type parameter separately.

- (a) $F^A \triangleq \text{Int} \times A \times A + (\text{String} \rightarrow A) \times A$.
- (b) $G^{A,B} \triangleq (A \rightarrow \text{Int} \rightarrow \mathbb{1} + B) + (A \rightarrow \mathbb{1} + A \rightarrow \text{Int})$.
- (c) $H^{A,B,C} \triangleq (A \rightarrow A \rightarrow B \rightarrow C) \times C + (B \rightarrow A)$.
- (d) $P^{A,B} \triangleq (((A \rightarrow B) \rightarrow A) \rightarrow B) \rightarrow A$.

Exercise 6.3.1.7 Show that the recursive type constructor L^\bullet defined by

$$L^A \triangleq \mathbb{1} + A + L^A$$

is a functor, and implement a `map` or `fmap` function for L in Scala.

Exercise 6.3.1.8 Show that the regular-shaped tree L^\bullet defined by

$$L^A \triangleq A \times A + L^{A \times A \times A}$$

is a functor, and implement a `map` or `fmap` function for L in Scala.

6.4 Further developments

6.4.1 Profunctors

We have seen that some type constructors are neither functors nor contrafunctors because their type parameters appear both in covariant and contravariant positions. An example of such a type constructor is

$$P^A \triangleq A + (A \rightarrow \text{Int})$$

It is not possible to define either a `map` or a `cmap` function for P : the required type signatures cannot be implemented. However, we *can* implement a function called `xmap`, with the type signature

$$\text{xmap}_P : (B \rightarrow A) \rightarrow (A \rightarrow B) \rightarrow P^A \rightarrow P^B$$

To see why, let us temporarily rename the contravariant occurrence of A to Z and define a new type constructor \tilde{P} by

$$\tilde{P}^{Z,A} \triangleq A + (Z \rightarrow \text{Int})$$

The original type constructor P^A is expressed as $P^A = \tilde{P}^{A,A}$. Now, $\tilde{P}^{Z,A}$ is covariant in A and contravariant in Z . We can implement $\text{xmap}_{\tilde{P}}$ as a composition of fmap with respect to A and cmap with respect to Z , similarly to what we saw in the proof of Statement 6.2.4.3. The function $\text{xmap}_{\tilde{P}}$ will satisfy the identity and composition laws (6.23)–(6.24). Setting the type parameter $Z = A$, we will obtain the xmap_P function for P . The identity and composition laws for xmap_P will hold, since the laws of $\tilde{P}^{Z,A}$ hold for all type parameters:

$$\text{P's identity law : } \text{xmap}_P(\text{id}^{A \rightarrow A})(\text{id}^{A \rightarrow A}) = \text{id} \quad ,$$

$$\text{P's composition law : } \text{xmap}_P(f_1^{B \rightarrow A})(g_1^{A \rightarrow B}) \circ \text{xmap}_P(f_2^{C \rightarrow B})(g_2^{B \rightarrow C}) = \text{xmap}_P(f_2 \circ f_1)(g_1 \circ g_2) \quad .$$

A type constructor P^A with these properties ($P^A \cong \tilde{P}^{A,A}$ where $\tilde{P}^{Z,A}$ has a lawful $\text{xmap}_{\tilde{P}}$) is called a **profunctor**. Sometimes the type constructor $\tilde{P}^{Z,A}$ is also called a profunctor.

Consider an exponential-polynomial type constructor P^A , no matter how complicated, such as

$$P^A \triangleq (1 + A \times A \rightarrow A) \times A \rightarrow 1 + (A \rightarrow A + \text{Int}) \quad .$$

Each copy of the type parameter A will occur either in covariant or in a contravariant position because no other possibility is available in exponential-polynomial types. So, we can always rename all contravariant occurrences of the type parameter A to “ Z ” and so obtain a new type constructor $\tilde{P}^{Z,A}$, which will be covariant in A and contravariant in Z . Since $\tilde{P}^{A,Z}$ is a functor in A and a contrafunctor in Z , we will be able to define a function $\text{xmap}_{\tilde{P}}$ satisfying the identity and composition laws. Setting $Z = A$, we will obtain a lawful xmap_P , which makes P a profunctor. Thus, *every* exponential-polynomial type constructor is a profunctor.

An unfunctor, such as the disjunctive type `ServerAction[R]` shown in Section 6.1.6, cannot be made into a profunctor. The type signature of xmap cannot be implemented for `ServerAction[R]` because it is not a fully parametric type constructor (and so is not exponential-polynomial).

Profunctors are not often used in practical coding. We will see profunctors occasionally in later chapters where we need to reason about type constructors of arbitrary variance.

6.4.2 Subtyping with injective or surjective conversion functions

In some cases, P is a subtype of Q when the set of values of P is a *subset* of values of Q . In other words, the conversion function $P \rightarrow Q$ is injective and embeds all information from a value of type P into a value of type Q . This kind of subtyping works for parts of disjunctive types, such as `Some[A] <: Option[A]` (in the type notation, $\mathbb{0} + A \lesssim \mathbb{1} + A$). The set of all values of type `Some[A]` is a subset of the set of values of type `Option[A]`, and the conversion function is injective because it is an identity function, $\mathbb{0} + x:A \rightarrow \mathbb{0} + x$, that merely reassigns types.

However, subtyping does not necessarily imply that the conversion function is injective. An example of a subtyping relation with a *surjective* conversion function is between the function types $P \triangleq \mathbb{1} + A \rightarrow \text{Int}$ and $Q \triangleq \mathbb{0} + A \rightarrow \text{Int}$ (in Scala, $P = \text{Option[A] } \Rightarrow \text{Int}$ and $Q = \text{Some[A] } \Rightarrow \text{Int}$). We have $P \lesssim Q$ because $P \cong C^{\mathbb{1}+A}$ and $Q \cong C^{\mathbb{0}+A}$, where $C^X \triangleq X \rightarrow \text{Int}$ is a contrafunctor. The conversion function $P \rightarrow Q$ is an identity function that reassigns types,

```
def p2q[A](p: Option[A] => Int): Some[A] => Int = { x: Some[A] => p(x) }
```

In the code notation, $p \rightarrow x \rightarrow p(x)$ is easily seen to be the same as $p \rightarrow p$.

Nevertheless, it is not true that all information from a value of type P is preserved in a value of type Q ; the type P describes functions that also accept `None` as an argument, while functions of type Q do not. So, there is strictly more information in the type P than in Q . The conversion function $p2q : P \rightarrow Q$ is surjective.

We have now seen examples of injective and surjective type conversions. Suppose $P_1 \lesssim Q_1$ and $P_2 \lesssim Q_2$, and consider the product types $P_1 \times P_2$ and $Q_1 \times Q_2$. Since the product type is one of the functor constructions, the product $A \times B$ is covariant in both type parameters. It follows that $P_1 \times P_2 \lesssim Q_1 \times Q_2$. If $r_1 : P_1 \rightarrow Q_1$ is injective but $r_2 : P_2 \rightarrow Q_2$ is surjective, the pair product

$r_1 \boxtimes r_2 : P_1 \times P_2 \rightarrow Q_1 \times Q_2$ is neither injective nor surjective. So, type conversion functions are not necessarily injective or surjective; they can also be anything “in between”.

A property of functor liftings is that they preserve injectivity and surjectivity: if a function $f : A \rightarrow B$ is injective, it is lifted to an injective function $f^{\uparrow L} : L^A \rightarrow L^B$; and similarly for surjective functions f . Let us prove this property for injective functions; the proof for surjective functions is quite similar.

Statement 6.4.2.1 If L^A is a lawful functor and $f : A \rightarrow B$ is an injective function then $\text{fmap}_L(f)$ is also an injective function of type $L^A \rightarrow L^B$.

Proof We begin by noting that an injective function $f : A \rightarrow B$ must somehow embed all information from a value of type A into a value of type B . The **image** of f (the subset of all values of type B that can be obtained as $f(a)$ for some $a : A$) thus contains a distinct value of type B for each distinct value of type A . So, there exists a function that maps any b from the image of f back to a value $a : A$ it came from; call that function $g : B \rightarrow A$. The function g must satisfy

$$\forall a : A. \, g(f(a)) = a \quad ,$$

equivalently written as

$$g \circ f = \text{id} \quad .$$

It is important that g is a partial function. The function g is partial because it is defined only for a subset of values of type B , namely the values within the image of f . Despite the equation $g \circ f = \text{id}$, the function g is not an inverse for f . An inverse function for f must be a *total* (not a partial) function h satisfying $h \circ f = \text{id}$ and $f \circ h = \text{id}$. The function g is called a **left inverse** for f because $f \circ g \neq \text{id}$, since $f \circ g$ is only a partial function.

The fact that f has a left inverse is *equivalent* to the assumption that f is injective. Indeed, if any function f has a left inverse g , we can show that f is injective. Assume some x and y such that $f(x) = f(y)$; we will prove that f is injective if we show that $x = y$. Applying g to both sides of $f(x) = f(y)$, we get

$$x = g(f(x)) = g(f(y)) = y \quad .$$

Now we apply this trick to functions lifted into the functor L . To prove that $\text{fmap}_L(f)$ is injective, we need to show that it has a left inverse. We can lift both sides of the equation $g \circ f = \text{id}$ to get

$$\begin{aligned} & \text{fmap}_L(g) \circ \text{fmap}_L(f) \\ \text{composition law of } L : &= \text{fmap}_L(g \circ f) \\ \text{use } g \circ f = \text{id} : &= \text{fmap}_L(\text{id}) \\ \text{identity law of } L : &= \text{id} \quad . \end{aligned}$$

It follows that $\text{fmap}_L(g) \circ \text{fmap}_L(f) = \text{id}$, i.e., $\text{fmap}_L(g)$ is a left inverse for $\text{fmap}_L(f)$. Since $\text{fmap}_L(f)$ has a left inverse, it is injective.

7 Reasoning about code. Techniques of symbolic derivation

In previous chapters, we have performed symbolic derivations of some laws. To make those derivations more manageable, we gradually developed special notations and techniques of reasoning. This short chapter is a summary of these notations and techniques.

7.1 Mathematical code notation

7.1.1 The nine constructions of purely functional code

The eight basic constructions introduced in Section 5.2.3, together with recursion, serve as a foundation for **purely functional** coding style. All major techniques and design patterns of functional programming can be implemented using only these constructions, i.e., by purely functional programs. We will now define the code notation (summarized in Table 7.1) for each of the nine constructions.

1) Use a constant At any place in the code, we may use a fixed constant value of a primitive type, such as `Int`, `String`, or `Unit`. We may also use a “named unit”, e.g., `None` of type `Option[A]` for any type `A`. All named unit values are denoted by `1` and are viewed as having type `1`.

With this construction, we can create **constant functions** (functions that ignore their argument):

```
def c_1(x: String): Int = 123
```

$$c_1(x^{\text{String}}) \triangleq 123$$

2) Use a given argument In any expression that has a bound variable (e.g., an argument within a function’s body), we may use the bound variable at any place, as many times as we need.

```
def c_2[A](x: String, y: Int): Int = 123 + y + y
```

$$c_2(x^{\text{String}}, y^{\text{Int}}) \triangleq 123 + y + y$$

3) Create a function We can always make a nameless function `{ x => expr }` out of a variable, say `x`, and any expression `expr` that may use `x` as a free variable (i.e., a variable that must be already defined outside the expression). E.g., the expression `123 + y + y` uses `y` as a free variable because `123 + y + y` only makes sense if `y` is already defined. So, we can create a nameless function

```
{ y: Int => 123 + y + y }
```

$$y^{\text{Int}} \rightarrow 123 + y + y$$

If the expression `expr` already contains `x` as a bound variable, the function `{ x => expr }` will have a name clash. As an example, consider an expression `expr == { x => x }` that already contains a nameless function with bound variable `x`. If we want to make a function out of that expression, we could write `x => { x => x }`, but such code is confusing. It is helpful to avoid the name clash by renaming the bound variables inside `expr`, e.g., `expr == { z => z }`:

```
val f = { x: Int => { z: Int => z } }
```

$$f \triangleq x^{\text{Int}} \rightarrow z^{\text{Int}} \rightarrow z$$

4) Use a function If a function is already defined, we can use it by applying it to an argument.

```
val f = { y: Int => 123 + y + y }
f(100) // Evaluates to 323.
```

$$f \triangleq y^{\text{Int}} \rightarrow 123 + y + y$$
$$f(100) = 323$$

5) Create a tuple Given two values `a: A` and `b: B`, we can always create the tuple `(a, b)` as well as `(b, a)`. In the code notation, those tuples are written as `a × b` and `b × a`.

Constructions	Scala examples	Code notation
use a constant	<code>()</code> or <code>true</code> or <code>"abc"</code> or <code>123</code>	$1, \text{true}, \text{"abc"}, 123$
use a given argument	<code>def f(x: A) = { ... x ... }</code>	$f(x:A) \triangleq \dots x \dots$
create a function	<code>(x: A) => expr(x)</code>	$x:A \rightarrow \text{expr}(x)$
use a function	<code>f(x)</code> or <code>x.pipe(f)</code> (Scala 2.13)	$f(x)$ or $x \triangleright f$
create a tuple	<code>val p: (A, B) = (a, b)</code>	$p:A \times B \triangleq a \times b$
use a tuple	<code>p._1</code> or <code>p._2</code>	$p \triangleright \pi_1$ or $p \triangleright \pi_2$
create a disjunctive value	<code>Left[A, B](x)</code> or <code>Right[A, B](y)</code>	$x:A + \emptyset:B$ or $\emptyset:A + y:B$
use a disjunctive value	<code>val p: Either[A, B] = ...</code> <code>val q: C = p match {</code> <code> case Left(x) => f(x)</code> <code> case Right(y) => g(y)</code> <code>}</code>	$q:C \triangleq p:A + B \triangleright \begin{array}{c c} & C \\ \hline A & x:A \rightarrow f(x) \\ B & y:B \rightarrow g(y) \end{array}$
use a recursive call	<code>def f(x) = { ... f(y) ... }</code>	$f(x) \triangleq \dots \overline{f}(y) \dots$

Table 7.1: Mathematical notation for the nine basic code constructions.

6) Use a tuple Given a tuple `p == (a, b)`, we can extract each of the values via `p._1` and `p._2`. The corresponding code notation is $p \triangleright \pi_1$ and $p \triangleright \pi_2$. The auxiliary functions π_i (where $i = 1, 2, \dots$) may be used for tuples of any size. Example code defining these functions:

```
def pi_1[A, B]: ((A, B)) => A = {
  case (a, b) => a
} // Same as '_1'
def pi_2[A, B]: ((A, B)) => B = {
  case (a, b) => b
} // Same as '_2'
```

$$\begin{aligned}\pi_1^{A,B} &\triangleq a:A \times b:B \rightarrow a \quad , \\ \pi_2^{A,B} &\triangleq a:A \times b:B \rightarrow b \quad .\end{aligned}$$

The notation $a \times b$ is used in an *argument* of a function to destructure a tuple.

7) Create a disjunctive value Once a disjunctive type such as $A + B + C$ has been defined in Scala, its named “constructors” (i.e., case classes) are used to create values of the type:

```
sealed trait S
final case class P(w: Int, x: Int) extends S
final case class Q(y: String) extends S
final case class R(z: Int) extends S

val s: S = P(10, 20) // Create a value of type S.
val t: S = R(30) // Another value of type S.
```

$$S \triangleq \text{Int} \times \text{Int} + \text{String} + \text{Int} \quad ,$$

$$\begin{aligned}s:S &\triangleq 10 \times 20 + \emptyset:\text{String} + \emptyset:\text{Int} \quad , \\ t:S &\triangleq \emptyset:\text{Int} \times \text{Int} + \emptyset:\text{String} + 30 \quad .\end{aligned}$$

The code notation for disjunctive values, e.g., $\emptyset + \emptyset + x$, is more verbose than the Scala syntax such as `R(x)`. The advantage is that we may explicitly annotate all types and show clearly the part of the disjunction that we are creating. Another advantage is that the notation $\emptyset + \emptyset + x$ is similar to a row vector, $\begin{vmatrix} \emptyset & \emptyset & x \end{vmatrix}$, which is well adapted to the matrix notation for functions.

8) Use a disjunctive value Once created, disjunctive values can be used in a pattern matching expression (Scala’s `match/case`). Recall that functions that take a disjunctive value as an argument (“*disjunctive functions*”) may be written *without* the `match` keyword:

```
val compute: Option[Int] => Option[Int] = {
  case None      => Some(100)
  case Some(x)   => Some(x / 2)
}
```

$$\text{compute}^{:\text{Option}:\text{Int} \rightarrow \text{Option}:\text{Int}} \triangleq \begin{array}{c|c} & 1 & \text{Int} \\ \hline 1 & 0 & 1 \rightarrow 100 \\ \text{Int} & 0 & x \rightarrow \frac{x}{2} \end{array} \quad .$$

We will use this example to see how disjunctive functions are written in the matrix notation.

Each row of a matrix corresponds to a part of the disjunctive type matched by one of the `case` expressions. In this example, the disjunctive type `Option[Int]` has two parts, the named unit `None` (denoted by `1`) and the case class `Some[Int]`, which is equivalent to the type `Int`. So, the matrix has two rows labeled `1` and `Int`, showing that the function's argument type is `1 + Int`.

The columns of the matrix correspond to the parts of the disjunctive type *returned* by the function. In this example, the return type is also `Option[Int]`, that is, `1 + Int`, so the matrix has two columns labeled `1` and `Int`. If the return type is not disjunctive, the matrix will have one column.

What are the matrix elements? The idea of the matrix notation is to translate the `case` expressions line by line from the Scala code. Look at the first `case` line as if it were a standalone partial function,

```
{ case None => Some(100) }
```

Since `None` is a named unit, this function is written in the code notation as $1 \rightarrow \mathbb{0} + 100:\text{Int}$.

The second line is written in the form of a partial function as

```
{ case Some(x) => Some(x / 2) }
```

The pattern variable on the left side is `x`, so we can denote that function by $x:\text{Int} \rightarrow \mathbb{0} + (x/2):\text{Int}$.

To obtain the matrix notation, we may simply write the two partial functions in the two rows:

```
val compute: Option[Int] => Option[Int] = {
  case None      => Some(100)
  case Some(x)   => Some(x / 2)
}
```

$$\text{compute}^{1 + \text{Int} \rightarrow 1 + \text{Int}} \triangleq \begin{array}{c|c} & 1 + \text{Int} \\ \hline 1 & 1 \rightarrow \mathbb{0} + 100 \\ \text{Int} & x \rightarrow \mathbb{0} + \frac{x}{2} \end{array} .$$

This is already a valid matrix notation for the function f . So far, the matrix has two rows and one column. However, we notice that each row's return value is *known* to be in a specific part of the disjunctive type `1 + Int` (in this example, both rows happen to return values of type `0 + Int`). So, we can split the column into two and obtain a clearer and more useful notation for this function:

$$\text{compute}^{1 + \text{Int} \rightarrow 1 + \text{Int}} \triangleq \begin{array}{c|c|c} & 1 & \text{Int} \\ \hline 1 & \mathbb{0} & 1 \rightarrow 100 \\ \text{Int} & \mathbb{0} & x:\text{Int} \rightarrow \frac{x}{2} \end{array} .$$

The void type `0` is written symbolically to indicate that the disjunctive part in that column is not returned. In this way, the matrix shows the parts of disjunctive types that are being returned.

Partial functions are expressed in the matrix notation by writing `0` in the missing rows:

```
def get[A]: Option[A] => A = {
  case Some(x) => x
} // Partial function; fails on 'None'.
```

$$\text{get}^{1 + A \rightarrow A} \triangleq \begin{array}{c|c} & A \\ \hline 1 & \mathbb{0} \\ A & x:A \rightarrow x \end{array} = \begin{array}{c|c} & A \\ \hline 1 & \mathbb{0} \\ A & \text{id} \end{array} .$$

Scala's `match` expression is equivalent to an application of a disjunctive function:

```
val p: Option[Int] = Some(64)
val q: Option[Int] = p match {
  case None      => Some(100)
  case Some(x)   => Some(x / 2)
} // The value of q equals Some(32) .
```

$$p \triangleq \mathbb{0} + 64:\text{Int} \quad , \quad q \triangleq p \triangleright \begin{array}{c|c} & 1 & \text{Int} \\ \hline 1 & \mathbb{0} & 1 \rightarrow 100 \\ \text{Int} & \mathbb{0} & x \rightarrow \frac{x}{2} \end{array} .$$

It is convenient to put the argument p to the left of the disjunctive function, as in the Scala code.

Because only one part of a disjunctive type can ever be returned, a row can have at most one non-void element, which must be in the column corresponding to the part being returned.

The matrix notation allows us to compute such function applications directly. We view the disjunctive value `0 + 64:Int` as a “row vector” $\begin{array}{c} \mathbb{0} \\ 64 \end{array}$, written with a single left line to distinguish it from a function matrix. Calculations use the standard rules of a vector-matrix product:

$$(\mathbb{0} + 64) \triangleright \begin{array}{c|c} \mathbb{0} & 1 \rightarrow 100 \\ \hline \mathbb{0} & x \rightarrow \frac{x}{2} \end{array} = \begin{array}{c} \mathbb{0} \\ 64 \end{array} \triangleright \begin{array}{c|c} \mathbb{0} & 1 \rightarrow 100 \\ \hline \mathbb{0} & x \rightarrow \frac{x}{2} \end{array} = \begin{array}{c} \mathbb{0} \\ 64 \triangleright (x \rightarrow \frac{x}{2}) \end{array} = \begin{array}{c} \mathbb{0} \\ 32 \end{array} = (\mathbb{0} + 32) .$$

The pipe (\triangleright) operation plays the role of the “multiplication” of matrix elements, and we drop any terms containing \emptyset . We also omitted type annotations since we already checked that the types match.

9) Use a recursive call The last construction is to call a function recursively within its own definition. This construction was not shown in Section 5.2.3 because the constructive propositional logic (which was the main focus in that chapter) cannot represent a recursively defined value. However, this limitation of propositional logic means only that we do not have an algorithm for *automatic* derivation of recursive code. Similarly, no algorithm can automatically derive code that involves type constructors with known methods. Nevertheless, those derivations can be performed by hand.

Recursive code is used often, and we need to get some experience reasoning about it. In derivations, this book denotes recursive calls by an overline. For example, the standard `fmap` method for the `List` functor is defined as

$$\text{fmap}_{\text{List}}(f) = f^{\uparrow \text{List}} \triangleq \begin{array}{c|c|c} & 1 & A \times \text{List}^A \\ \hline 1 & \text{id} & \emptyset \\ A \times \text{List}^A & \emptyset & h^A \times t^{\text{List}^A} \rightarrow f(h) \times (t \triangleright \overline{\text{fmap}_{\text{List}}(f)}) \end{array} .$$

The recursive call to `fmap` is applied to a list’s tail (the value t).

In proofs of laws for recursive functions, it is necessary to use induction in the number of recursive self-calls. However, the proof does not need to separate the base case (no recursive calls) from the inductive step. In the proof, we write a symbolic calculation as usual, except that we may assume that the law already holds for any recursive calls to the same function.

For example, a proof of the identity law of `fmap`, which says $\text{fmap}_{\text{List}}(\text{id}) = \text{id}$, may proceed by replacing the recursive call $\overline{\text{fmap}_{\text{List}}(\text{id})}$ by id during the calculations:

$$\begin{aligned} \text{fmap}_{\text{List}}(\text{id}) &= \begin{array}{c|c} \text{id} & \emptyset \\ \hline \emptyset & h^A \times t^{\text{List}^A} \rightarrow \text{id}(h) \times (t \triangleright \overline{\text{fmap}_{\text{List}}(\text{id})}) \end{array} \\ \text{inductive assumption :} &= \begin{array}{c|c} \text{id} & \emptyset \\ \hline \emptyset & h \times t \rightarrow \text{id}(h) \times (t \triangleright \text{id}) \end{array} = \begin{array}{c|c} \text{id} & \emptyset \\ \hline \emptyset & h \times t \rightarrow h \times t \end{array} \\ \text{identity matrix :} &= \begin{array}{c|c} \text{id} & \emptyset \\ \hline \emptyset & \text{id} \end{array} = \text{id} . \end{aligned}$$

7.1.2 Function composition and the pipe notation

In addition to the basic code constructions, our derivations will often need to work with function compositions and lifted functions. It is often faster to perform calculations with functions when we do not write all of their arguments explicitly; e.g., writing the right identity law as $f \circ \text{id} = f$ instead of $\text{id}(f(x)) = f(x)$. This is known as calculating in **point-free** style (meaning “argument-free”). Many laws can be formulated and used more easily in the point-free form.

Calculations in point-free style almost always involve composing functions. This book prefers to use the *forward* function composition ($f \circ g$) defined for arbitrary $f: A \rightarrow B$ and $g: B \rightarrow C$ by

`f andThen g == { x => g(f(x)) }`

$f \circ g \triangleq x \rightarrow g(f(x))$.

A useful tool for calculations is the **pipe** operation, $x \triangleright f$, which places the argument (x) to the *left* of a function (f). It is then natural to apply further functions at *right*, for example $(x \triangleright f) \triangleright g$ meaning $g(f(x))$. In Scala, methods such as `map` and `filter` are often combined in this way:

`x.map(f).filter(p)`

$x \triangleright \text{fmap}(f) \triangleright \text{filt}(p)$.

To enable this common usage, the \triangleright operation is defined to group towards the left. So, the parentheses in $(x \triangleright f) \triangleright g$ are not needed, and we write $x \triangleright f \triangleright g$.

Since $x \triangleright f \triangleright g = g(f(x))$ by definition, it follows that the composition $f \circ g$ satisfies

$$x \triangleright f \triangleright g = x \triangleright (f \circ g) \quad .$$

Such formulas are needed often, so we follow the convention that the pipe operation (\triangleright) groups weaker than the composition operation (\circ). We can then omit parentheses: $x \triangleright (f \circ g) = x \triangleright f \circ g$.

Another common simplification occurs with function compositions of the form

$$(x \rightarrow t \triangleright f) \circ g = x \rightarrow g(t \triangleright f) = x \rightarrow (t \triangleright f \triangleright g) = x \rightarrow t \triangleright f \circ g \quad .$$

The function arrow groups weaker than the pipe operator: $x \rightarrow t \triangleright f \circ g = x \rightarrow (t \triangleright f \circ g)$.

How can we verify this and other similar computations where the operations \triangleright and \circ are combined in some way? Instead of memorizing a large set of identities, we can rely on knowing only one rule that says how arguments are symbolically substituted as parameters into functions, for example:

substitute x instead of a : $x \triangleright (a \rightarrow f(a)) = f(x)$.

substitute $f(x)$ instead of y : $(x \rightarrow f(x)) \circ (y \rightarrow g(y)) = x \rightarrow g(f(x))$.

Whenever there is a doubt (is $x \triangleright (f \triangleright g)$ or $(x ; f) \triangleright g$ the correct formula..?), we can always write functions in an expanded form, $x \rightarrow f(x)$ instead of f , and perform calculations more verbosely. After getting some experience with the \triangleright and $;$ operations, the reader will start using them more freely without writing functions in expanded form.

The matrix notation is adapted to the pipe operation and the forward function composition. As an example, let us write the composition of the functions `compute` and `get[Int]` shown above:

$$\text{compute} \circ \text{get} = \boxed{\boxed{1 \quad \text{Int}} \quad \boxed{0 \quad 1 \rightarrow 100} \quad \boxed{\boxed{1 \quad \text{Int}} \quad 0} \quad = \quad \boxed{1 \quad \text{Int}} \quad \boxed{(1 \rightarrow 100) \circ \text{id}} \quad = \quad \boxed{1 \quad \text{Int}} \quad \boxed{1 \rightarrow 100}} \quad .$$

$$\text{Int} \quad \boxed{0 \quad x \rightarrow \frac{x}{2}} \quad \text{Int} \quad \text{id} \quad \text{Int} \quad \boxed{(x \rightarrow \frac{x}{2}) \circ \text{id}} \quad \text{Int} \quad \boxed{x \rightarrow \frac{x}{2}} \quad .$$

In this computation, we used the composition (\cdot) instead of the “multiplication” of matrix elements.

Why does the rule for matrix multiplication work for function compositions? The reason is the equivalence $x \triangleright f \triangleright g = x \triangleright f \circ g$. We have defined the matrix form of functions to work with the “row-vector” form of disjunctive types, i.e., for the computation $x \triangleright f$ (where x is a row vector representing a value of a disjunctive type). The result of computing $x \triangleright f$ is again a row vector, which we can pipe into another matrix g as $x \triangleright f \triangleright g$. The standard rules of matrix multiplication make it associative; so, the result of $x \triangleright f \triangleright g$ is the same as the result of piping x into the matrix product of f and g . Therefore, the matrix product of f and g must yield the function $f \circ g$.

A “non-disjunctive” function (i.e., one not taking or returning disjunctive types) may be written as a 1×1 matrix, so its composition with disjunctive functions can be computed via the same rules.

7.1.3 Functor and contrafunctor liftings

Functions lifted to a functor (or a contrafunctor) and their compositions are used in derivations so often that we need shorter notation than $x \triangleright \text{fmap}_F(f)$ or its Scala analog `x.map(f)`. This book uses the notation $x \triangleright f^{\uparrow F}$ for functors F and $x \triangleright f^{\downarrow C}$ for contrafunctors C . This notation graphically emphasizes the function f being lifted and also shows the name of the relevant functor or the contrafunctor. Compositions of lifted functions are visually easy to recognize, for example:

$$f^{\downarrow H} \circ g^{\downarrow H} = (g \circ f)^{\downarrow H} \quad , \quad f^{\uparrow L} \circ g^{\uparrow L} \circ h^{\uparrow L} = (f \circ g \circ h)^{\uparrow L} \quad .$$

In these formulas, the labels H and L clearly indicate the possibility of pulling several functions under a single lifting. We can also split a lifted composition into a composition of liftings.

The lifting notation helps us recognize that these steps are possible by looking at the formula. Of course, it still remains to find a useful sequence of steps in a given derivation or proof.

7.2 Derivation techniques

7.2.1 Auxiliary functions for handling products

The functions denoted by π_1 , π_2 , Δ , and \boxtimes proved to be helpful in derivations that involve tuples. (However, the last two functions are unlikely to be frequently used in practical programming.)

We already saw the definition and the implementation of the functions π_1 and π_2 .

The “diagonal” function Δ is a right inverse for π_1 and π_2 :

$$\text{def delta}[A]: A \Rightarrow (A, A) = \{ x \Rightarrow (x, x) \} \quad \Delta^A : A \rightarrow A \times A \quad , \quad \Delta \triangleq a^A \rightarrow a \times a \quad .$$

It is clear that extracting any part of a pair $\text{delta}(x) == (x, x)$ will give back the original x . This property can be written as an equation or a “law”,

$$\text{delta}(x)._1 == x \quad \pi_1(\Delta(x)) = x \quad .$$

We can transform this law into a point-free equation by first using the pipe notation,

$$\pi_1(\Delta(x)) = (\Delta(x)) \triangleright \pi_1 = x \triangleright \Delta \triangleright \pi_1 = x \triangleright \Delta \circ \pi_1 \quad ,$$

and then rewriting the equation $x \triangleright \Delta \circ \pi_1 = x$ in a point-free form:

$$\Delta \text{ is a right inverse of } \pi_1 : \Delta \circ \pi_1 = \text{id} \quad . \quad (7.1)$$

The same property holds for π_2 .

The **pair product** operation $f \boxtimes g$ is defined for any functions $f^{A \rightarrow P}$ and $g^{B \rightarrow Q}$ by

$$\text{def pairProduct}[A,B,P,Q](f: A \Rightarrow P, g: B \Rightarrow Q): ((A, B)) \Rightarrow (P, Q) = \{ \text{case } (a, b) \Rightarrow (f(a), g(b)) \}$$

$$f \boxtimes g : A \times B \rightarrow P \times Q \quad , \quad f \boxtimes g \triangleq a \times b \rightarrow f(a) \times g(b) \quad .$$

Two properties of this operation follow directly from its definition:

$$\text{composition law} : (f^{A \rightarrow P} \boxtimes g^{B \rightarrow Q}) \circ (m^{P \rightarrow X} \boxtimes n^{Q \rightarrow Y}) = (f \circ m) \boxtimes (g \circ n) \quad , \quad (7.2)$$

$$\text{left and right projection laws} : (f^{A \rightarrow P} \boxtimes g^{B \rightarrow Q}) \circ \pi_1 = \pi_1 \circ f \quad , \quad (f \boxtimes g) \circ \pi_2 = \pi_2 \circ g \quad , \quad (7.3)$$

$$\text{identity law} : \text{id}^A \boxtimes \text{id}^B = \text{id}^{A \times B} \quad .$$

An equivalent way of defining $f \boxtimes g$ is via this Scala code,

$$\text{def pairProduct}[A,B,P,Q](f: A \Rightarrow P, g: B \Rightarrow Q)(p: (A, B)): (P, Q) = (f(p._1), g(p._2))$$

$$f \boxtimes g = p^{A \times B} \rightarrow f(p \triangleright \pi_1) \times g(p \triangleright \pi_2) = p \rightarrow (p \triangleright \pi_1 \triangleright f) \times (p \triangleright \pi_2 \triangleright g) \quad .$$

The pair product notation can shorten calculations with functors that involve product types (tuples). For example, the lifting for the functor $F^A \triangleq A \times A \times Z$ can be shortened to

$$f^{\uparrow F} \triangleq a_1^A \times a_2^A \times z^Z \rightarrow f(a_1) \times f(a_2) \times z = f \boxtimes f \boxtimes \text{id} \quad .$$

The last formula is often more convenient in symbolic derivations.

7.2.2 Deriving laws for functions with known implementations

The task is to prove a given law (an equation) for a function whose code is known. An example of such an equation is the naturality law of Δ , which states that for any function $f^{A \rightarrow B}$ we have

$$f \circ \Delta = \Delta \circ (f \boxtimes f) \quad . \quad (7.4)$$

Laws for fully parametric functions are often written without type annotations. However, it is important to check that types match. So we begin by finding suitable type parameters for Eq. (7.4).

Since it is given that f has type $A \rightarrow B$, the function Δ in the left-hand side of Eq. (7.4) must take arguments of type B and thus returns a value of type $B \times B$. We see that the left-hand side must be a function of type $A \rightarrow B \times B$. So, the Δ in the right-hand side must take arguments of type A . It then returns a value of type $A \times A$, which is consumed by $f \boxtimes f$. In this way, we see that all types match. We can put the resulting types into a type diagram and write the law with type annotations:

$$\begin{array}{ccc}
 A & \xrightarrow{\Delta^A} & A \times A \\
 f \downarrow & & \downarrow f \boxtimes f \\
 B & \xrightarrow{\Delta^B} & B \times B
 \end{array}
 \quad f : A \rightarrow B ; \Delta : B \rightarrow B \times B = \Delta : A \rightarrow A \times A ; (f \boxtimes f) \quad .$$

To prove the law, we need to use the known code of the function Δ .

We substitute that code into the left-hand side of the law and into the right-hand side of the law, hoping to transform these two expressions until they are the same.

We will now perform this computation in the Scala syntax and in the code notation.

```

x.pipe(f andThen delta)
  == (f(x)).pipe { a => (a, a) }
  == (f(x), f(x)) // Left-hand side.
x.pipe(delta andThen { case (a, b) => (f(a), f(b)) })
  == (x, x).pipe { case (a, b) => (f(a), f(b)) }
  == (f(x), f(x)) // Right-hand side.
  
```

$$\begin{aligned}
 x \triangleright f ; \Delta &= f(x) \triangleright (b \rightarrow b \times b) \\
 &= f(x) \times f(x) \quad . \\
 x \triangleright \Delta ; (f \boxtimes f) &= (x \times x) \triangleright (a \times b \rightarrow f(a) \times f(b)) \\
 &= f(x) \times f(x) \quad .
 \end{aligned}$$

At each step of the derivation, typically there is only one symbolic transformation we can perform. In the example above, each step either substitutes a definition of a known function or applies some function to an argument and computes the result. To help us retrace the steps later, we use a green underline as a hint indicating a sub-expression that will change at the next step.

We will prefer to derive laws in the code notation rather than in Scala syntax. The code notation covers all purely functional code, i.e., all programs that use only the nine basic code constructions.

7.2.3 Working with disjunctive functions in matrix notation

Although the matrix notation is unusual, it provides a general way of performing symbolic computations with disjunctive types in point-free style (the matrix elements are *functions*). Writing all matrices with type annotations makes it easier to translate between matrices and Scala code.

In many cases, the rules of matrix multiplication and function composition are sufficient for calculating with disjunctive functions. For example, consider the following functions `swap[A]` and `merge[A]`,

```

def swap[A]: Either[A, A] => Either[A, A] = {
  case Left(a)    => Right(a)
  case Right(a)   => Left(a)
}
def merge[A]: Either[A, A] => A = {
  case Left(a)    => a
  case Right(a)   => a
}
  
```

$$\text{swap}^A \triangleq \begin{vmatrix} & \text{A} & \text{A} \\ \text{A} & 0 & \text{id} \\ \text{A} & \text{id} & 0 \end{vmatrix}, \quad \text{merge}^A \triangleq \begin{vmatrix} & \text{A} \\ \text{A} & \text{id} \\ \text{A} & \text{id} \end{vmatrix} .$$

We can quickly prove by matrix multiplication that $\text{swap} ; \text{swap} = \text{id}$ and $\text{swap} ; \text{merge} = \text{merge}$:

$$\begin{aligned}
 \text{swap} ; \text{swap} &= \begin{vmatrix} 0 & \text{id} \\ \text{id} & 0 \end{vmatrix} ; \begin{vmatrix} 0 & \text{id} \\ \text{id} & 0 \end{vmatrix} = \begin{vmatrix} \text{id} ; \text{id} & 0 \\ 0 & \text{id} ; \text{id} \end{vmatrix} = \begin{vmatrix} \text{id} & 0 \\ 0 & \text{id} \end{vmatrix} = \text{id} \quad , \\
 \text{swap} ; \text{merge} &= \begin{vmatrix} 0 & \text{id} \\ \text{id} & 0 \end{vmatrix} ; \begin{vmatrix} \text{id} \\ \text{id} \end{vmatrix} = \begin{vmatrix} \text{id} ; \text{id} \\ \text{id} ; \text{id} \end{vmatrix} = \begin{vmatrix} \text{id} \\ \text{id} \end{vmatrix} = \text{merge} \quad .
 \end{aligned}$$

The identity function for any disjunctive type, e.g., $A + B + C$, is the “identity diagonal” matrix,

$$\text{id}^{A+B+C \rightarrow A+B+C} = \begin{array}{c|ccc} & \begin{array}{c|ccc} & A & B & C \end{array} \\ \hline A & \begin{array}{c|ccc} & \text{id} & 0 & 0 \end{array} \\ B & \begin{array}{c|ccc} & 0 & \text{id} & 0 \end{array} \\ C & \begin{array}{c|ccc} & 0 & 0 & \text{id} \end{array} \end{array} .$$

The type constructor $E^A \triangleq A + A$ is a functor whose lifting is defined by

```
def fmap[A, B](f: A => B): Either[A, A] => Either[B, B] = {
  case Left(a) => Left(f(a))
  case Right(a) => Right(f(a))
}
```

$$(f^{A \rightarrow B})^{\uparrow E} \triangleq \begin{array}{c|cc} & B & B \\ \hline A & \begin{array}{c|cc} & f & 0 \end{array} \\ A & \begin{array}{c|cc} & 0 & f \end{array} \end{array} .$$

With this definition, we can formulate a law of `merge`, called the “naturality law”:

$$\begin{array}{ccc} A + A & \xrightarrow{\text{merge}^A} & A \\ f^{\uparrow E} \downarrow & & \downarrow f \\ B + B & \xrightarrow{\text{merge}^B} & B \end{array}$$

$$(f^{A \rightarrow B})^{\uparrow E} ; \text{merge}^B = \text{merge}^A ; f^{A \rightarrow B} .$$

Proving this law is a simple matrix calculation:

$$\begin{array}{l} \text{left-hand side : } f^{\uparrow E} ; \text{merge} = \begin{array}{c|cc} f & 0 \\ 0 & f \end{array} ; \begin{array}{c|cc} \text{id} & \text{id} \\ \text{id} & \text{id} \end{array} = \begin{array}{c|cc} f ; \text{id} & f ; \text{id} \\ f ; \text{id} & f \end{array} = \begin{array}{c|c} f & f \end{array} , \\ \text{right-hand side : } \text{merge} ; f = \begin{array}{c|cc} \text{id} & f \\ \text{id} & \text{id} \end{array} ; \begin{array}{c|cc} \text{id} & f \\ \text{id} & \text{id} \end{array} = \begin{array}{c|cc} \text{id} ; f & f \\ \text{id} ; f & \text{id} \end{array} = \begin{array}{c|c} f & f \end{array} . \end{array}$$

In the last line we replaced f by a 1×1 matrix, $\begin{array}{c|c} f \end{array}$, in order to apply matrix multiplication.

Matrix rows and columns can be split or merged when necessary to accommodate various disjunctive types. As an example, let us verify the “associativity law” of `merge`,

$$\begin{array}{ccc} E^{A+A} & \xrightarrow{\text{merge}^{A+A}} & A + A \\ \downarrow \text{merge}^{\uparrow E} & & \downarrow \text{merge}^A \\ E^A & \xrightarrow{\text{merge}^A} & A \end{array}$$

$$(\text{merge}^A)^{\uparrow E} ; \text{merge}^A = \text{merge}^{A+A} ; \text{merge}^A .$$

Both sides of this law are functions of type $A + A + A + A \rightarrow A$. To transform the left-hand side, we use the definition of $^{\uparrow E}$ and write

$$\text{merge}^{\uparrow E} ; \text{merge} = \begin{array}{c|cc|c|cc|c|cc} & \begin{array}{c|cc} A & A \end{array} & & \begin{array}{c|cc} A & A \end{array} & & \begin{array}{c|cc} A & A \end{array} \\ \hline A + A & \text{merge} & 0 & A & \text{id} & & A + A & \text{merge} \\ A + A & 0 & \text{merge} & A & \text{id} & & A + A & \text{merge} \end{array} .$$

However, we have not yet substituted the definition of `merge` into the matrix. To do that, add more rows to the matrix in order to accommodate the full disjunctive type $A + A + A + A$:

$$\text{merge}^{\uparrow E} ; \text{merge} = \begin{array}{c|cc} & A \\ \hline A + A & \text{merge} \\ A + A & \text{merge} \end{array} = \begin{array}{c|c} A \\ \hline A & \text{id} \\ A & \text{id} \\ A & \text{id} \\ A & \text{id} \end{array} .$$

Now we compute the right-hand side of the law by substituting the code of `merge`:

$$\text{merge}^{A+A} ; \text{merge}^A = \begin{array}{c|cc|c|cc|c|cc} & \begin{array}{c|cc} A + A & A \end{array} & & \begin{array}{c|cc} A & A \end{array} \\ \hline A + A & \text{id} & ; & A & \text{id} \\ A + A & \text{id} & ; & A & \text{id} \end{array} .$$

We cannot proceed with matrix multiplication because the dimensions of the matrices do not match. We need to expand the rows and the columns of the first matrix; then we compute

$$\begin{array}{c|c|c} & | A + A | & | A | \\ \hline A + A & id & ; & A & id \\ A + A & id & ; & A & id \end{array} = \begin{array}{c|c|c} & | A & A | & | A | \\ \hline A & id & 0 & ; & A & id \\ A & 0 & id & ; & A & id \\ A & id & 0 & ; & A & id \\ A & 0 & id & ; & A & id \end{array} = \begin{array}{c|c} & | A | \\ \hline A & id \\ A & id \\ A & id \end{array} .$$

This proves the law (and also helps visualize how the transformations work with various types).

In some cases, we cannot fully split the rows or the columns of a matrix. For instance, if we are calculating with an arbitrary function $f:1+A \rightarrow 1+B$, we cannot write this function in a form of a 2×2 matrix because we do not know which parts of the disjunction will be returned (the code of the function f is arbitrary and unknown). At most, we could split the *rows* by writing the function f as a product of arbitrary functions $g:1 \rightarrow 1+B$ and $h:A \rightarrow 1+B$:

$$f = \begin{array}{c|c} & | 1 + B | \\ \hline 1 & g \\ A & h \end{array} .$$

The single column of this matrix remains unsplit. Either that column will remain unsplit throughout the derivation, or additional information about f , g , or h will allow us to split the column.

Finally, there are two tricks that complement the matrix intuition and may sometimes simplify a disjunctive function.¹

Ignored arguments If all rows of the disjunctive function ignore their arguments and always return the same results, we may collapse all rows into one, as shown in this example:

```
def same[A]: Either[A, Option[A]] => Option[A] =
  {
    case Left(a)      => None
    case Right(None)  => None
    case Right(Some(a)) => None
  }
```

$$\begin{aligned} \text{same}^{A+1+A \rightarrow 1+A} &= \begin{array}{c|c|c} & | 1 & A | \\ \hline A & - \rightarrow 1 & 0 \\ 1 & - \rightarrow 1 & 0 \\ A & - \rightarrow 1 & 0 \end{array} \\ &= \begin{array}{c|c|c} & | 1 & A | \\ \hline A + 1 + A & - \rightarrow 1 & 0 \end{array} . \end{aligned}$$

A more general formula for arbitrary functions $f:X \rightarrow C$ is

$$x:X \rightarrow p^{A+B} \triangleright \begin{array}{c|c} & | C | \\ \hline A & - \rightarrow f(x) \\ B & - \rightarrow f(x) \end{array} = x:X \rightarrow f(x) = f .$$

In this case, we can completely collapse the matrix, getting an ordinary (non-disjunctive) function.

Simplification of diagonal pair products Consider the pair product of two disjunctive functions such as $f:A+B \rightarrow R$ and $g:P+Q \rightarrow S$. Computing $f \boxtimes g$ in the matrix notation requires, in general, to split the rows and the columns of the matrices because the type of $f \boxtimes g$ is

$$\begin{aligned} f \boxtimes g : (A+B) \times (P+Q) &\rightarrow R \times S \\ &\cong A \times P + A \times Q + B \times P + B \times Q \rightarrow R \times S . \end{aligned}$$

¹These tricks are adapted from Section 2.8 of the book “Program design by calculation” (draft version of October 2019), see <http://www4.di.uminho.pt/~jno/ps/pdbc.pdf>

So, the pair product of two 2×1 matrices must be written *in general* as a 4×1 matrix:

$$\text{for any } f \triangleq \begin{vmatrix} & \parallel R \\ A & \parallel f_1 \\ B & \parallel f_2 \end{vmatrix} \text{ and } g \triangleq \begin{vmatrix} & \parallel S \\ P & \parallel g_1 \\ Q & \parallel g_2 \end{vmatrix}, \text{ we have } f \boxtimes g = \begin{vmatrix} & \parallel R \times S \\ A \times P & \parallel f_1 \boxtimes g_1 \\ A \times Q & \parallel f_1 \boxtimes g_2 \\ B \times P & \parallel f_2 \boxtimes g_1 \\ B \times Q & \parallel f_2 \boxtimes g_2 \end{vmatrix}.$$

A simplification trick exists when the pair product is composed with the diagonal function Δ :

$$\Delta ; (f \boxtimes g) = \Delta^{A+B \rightarrow (A+B) \times (A+B)} ; (f^{A+B \rightarrow R} \boxtimes g^{A+B \rightarrow S}) = p \rightarrow f(p) \times g(p).$$

This “diagonal pair product” is well-typed only if f and g have the same argument types (so, $A = P$ and $B = Q$). It turns out that the function $\Delta ; (f \boxtimes g)$ can be written as a 2×1 matrix, i.e., we do not need to split the rows:

$$\text{for any } f \triangleq \begin{vmatrix} & \parallel R \\ A & \parallel f_1 \\ B & \parallel f_2 \end{vmatrix} \text{ and } g \triangleq \begin{vmatrix} & \parallel S \\ A & \parallel g_1 \\ B & \parallel g_2 \end{vmatrix}, \text{ we have } \Delta ; (f \boxtimes g) = \begin{vmatrix} & \parallel R \times S \\ A & \parallel \Delta ; (f_1 \boxtimes g_1) \\ B & \parallel \Delta ; (f_2 \boxtimes g_2) \end{vmatrix}.$$

The rules of matrix multiplication do not help in deriving this law. So, we use a more basic approach: show that both sides are equal when applied to arbitrary values p of type $A + B$,

$$p^{A+B} \triangleright \Delta ; (f \boxtimes g) = f(p) \times g(p) \stackrel{?}{=} p \triangleright \begin{vmatrix} & \parallel R \times S \\ A & \parallel \Delta ; (f_1 \boxtimes g_1) \\ B & \parallel \Delta ; (f_2 \boxtimes g_2) \end{vmatrix}.$$

The type $A + B$ has two cases. Applying the left-hand side to $p \triangleq a^{A} + \mathbb{0}^{B}$, we get

$$\begin{aligned} f(p) \times g(p) &= ((a^{A} + \mathbb{0}^{B}) \triangleright f) \times ((a^{A} + \mathbb{0}^{B}) \triangleright g) \\ &= (\begin{vmatrix} a & \mathbb{0} \end{vmatrix} \triangleright \begin{vmatrix} f_1 \\ f_2 \end{vmatrix}) \times (\begin{vmatrix} a & \mathbb{0} \end{vmatrix} \triangleright \begin{vmatrix} g_1 \\ g_2 \end{vmatrix}) = (a \triangleright f_1) \times (a \triangleright g_1) = f_1(a) \times g_1(a). \end{aligned}$$

Applying the right-hand side to the same p , we find

$$\begin{aligned} \text{expect to equal } f_1(a) \times g_1(a) : \quad p \triangleright \begin{vmatrix} & \parallel R \times S \\ A & \parallel \Delta ; (f_1 \boxtimes g_1) \\ B & \parallel \Delta ; (f_2 \boxtimes g_2) \end{vmatrix} &= \begin{vmatrix} a & \mathbb{0} \end{vmatrix} \triangleright \begin{vmatrix} & \parallel R \times S \\ A & \parallel \Delta ; (f_1 \boxtimes g_1) \\ B & \parallel \Delta ; (f_2 \boxtimes g_2) \end{vmatrix} = a \triangleright \Delta ; (f_1 \boxtimes g_1) \\ \text{definition of } \Delta : \quad &= (a \times a) \triangleright (f_1 \boxtimes g_1) = f_1(a) \times g_1(a). \end{aligned}$$

A similar calculation with $p \triangleq \mathbb{0}^{A} + b^{B}$ shows that both sides of the law are equal to $f_2(b) \times g_2(b)$.

7.2.4 Derivations involving unknown functions with laws

A more challenging task is to derive an equation that uses arbitrary functions about which we only know that they satisfy certain given laws. Such derivations usually proceed by trying to transform the code until the given laws can be applied.

As an example, let us derive the property that $L^A \triangleq A \times F^A$ is a functor if F^\bullet is known to be a functor. We are in the situation where we only know that the function fmap_F exists and satisfies the functor law, but we do not know the code of fmap_F . Let us discover the derivation step by step.

First, we need to define fmap_L . We use the lifting notation $\uparrow F$ and write, for any $f^{A \rightarrow B}$,

```
def fmap_L[A, B](f: A => B): ((A, F[A])) => (B, F[B]) = { case (a, p) => (f(a), p.map(f)) }
```

$$f^{\uparrow L} \triangleq a^A \times p^{F^A} \rightarrow f(a) \times (p \triangleright f^{\uparrow F}) \quad .$$

To verify the identity law of L :

$$\text{expect to equal id : } \text{id}^{\uparrow L} = a^A \times p^{F^A} \rightarrow \text{id}(a) \times (p \triangleright \text{id}^{\uparrow F}) \quad .$$

At this point, the only things we can simplify are the identity functions applied to arguments. We know that F is a lawful functor; therefore, $\text{id}^{\uparrow F} = \text{id}$. So we continue the derivation, omitting types:

$$\text{expect to equal id : } \text{id}^{\uparrow L} = a \times p \rightarrow \text{id}(a) \times (p \triangleright \text{id}^{\uparrow F})$$

$$\text{identity law of } F : = a \times p \rightarrow a \times (p \triangleright \text{id})$$

$$\text{apply function : } = a \times p \rightarrow a \times p = \text{id} \quad .$$

To verify the composition law of L , we assume two arbitrary functions $f:A \rightarrow B$ and $g:B \rightarrow C$:

$$\text{expect to equal } (f \circ g)^{\uparrow L} : f^{\uparrow L} \circ g^{\uparrow L} = (a \times p \rightarrow f(a) \times f^{\uparrow F}(p)) \circ (b \times q \rightarrow g(b) \times g^{\uparrow F}(q)) \quad .$$

At this point, we pause and try to see how we might proceed. We do not know anything about f and g , so we cannot evaluate $f(a)$ or $f^{\uparrow F}(p)$. We also do not have the code of $\uparrow F$ (i.e., of fmap_F). The only information we have about these functions is that F 's composition law holds,

$$f^{\uparrow F} \circ g^{\uparrow F} = (f \circ g)^{\uparrow F} \quad . \quad (7.5)$$

We could use this law only if we somehow bring $f^{\uparrow F}$ and $g^{\uparrow F}$ together in the formula. The only way forward is to compute the function composition of the two functions whose code we *do* have:

$$\begin{aligned} & (a \times p \rightarrow f(a) \times f^{\uparrow F}(p)) \circ (b \times q \rightarrow g(b) \times g^{\uparrow F}(q)) \\ &= a \times p \rightarrow g(f(a)) \times g^{\uparrow F}(f^{\uparrow F}(p)) \quad . \end{aligned}$$

In order to use the law (7.5), we need to rewrite this code via the composition $f \circ g$. We notice that the formula contains exactly those function compositions:

$$g(f(a)) \times g^{\uparrow F}(f^{\uparrow F}(p)) = (a \triangleright f \circ g) \times (p \triangleright f^{\uparrow F} \circ g^{\uparrow F}) \quad .$$

So, we can now apply the composition law of F and write up the complete derivation, adding hints:

$$\text{expect to equal } (f \circ g)^{\uparrow L} : f^{\uparrow L} \circ g^{\uparrow L} = (a \times p \rightarrow f(a) \times f^{\uparrow F}(p)) \circ (b \times q \rightarrow g(b) \times g^{\uparrow F}(q))$$

$$\text{compute composition : } = a \times p \rightarrow g(f(a)) \times g^{\uparrow F}(f^{\uparrow F}(p))$$

$$\triangleright\text{-notation : } = a \times p \rightarrow (a \triangleright f \circ g) \times (p \triangleright f^{\uparrow F} \circ g^{\uparrow F})$$

$$\text{composition law of } F : = a \times p \rightarrow (a \triangleright f \circ g) \times (p \triangleright (f \circ g)^{\uparrow F})$$

$$\text{definition of } \uparrow L : = (f \circ g)^{\uparrow L} \quad .$$

The derivation becomes significantly shorter if we use the pair product (\boxtimes) to define $\uparrow L$:

$$f^{\uparrow L} \triangleq \text{id} \boxtimes f^{\uparrow F} \quad .$$

For instance, verifying the identity law looks like this:

$$\text{id}^{\uparrow L} = \text{id} \boxtimes \text{id}^{\uparrow F} = \text{id} \boxtimes \text{id} = \text{id} \quad .$$

This technique was used in the proof of Statement 6.2.3.3. The cost of having a shorter proof is the need to remember the properties of the pair product (\boxtimes), which is not often used in derivations.

7.2.5 Exercises

Exercise 7.2.5.1 Show using matrix calculations that $\text{swap} ; \text{swap} = \text{id}$, where swap is the function defined in Section 7.2.3.

Exercise 7.2.5.2 Now consider a different function $\text{swap}[\mathbf{A}, \mathbf{B}]$ with the type signature

```
def swap[A, B]: ((A, B)) => (B, A) = { case (a, b) => (b, a) } swapA,B ≡ a:A × b:B → b × a .
```

Show that $\Delta ; \text{swap} = \Delta$. Write out all types in this law and draw a type diagram.

Exercise 7.2.5.3 Given an arbitrary functor F , define the functor $L^A \doteq F^A \times F^A$ and prove, for an arbitrary function $f^{:A \rightarrow B}$, the “lifted naturality” law

$$f^{\uparrow F} ; \Delta = \Delta ; f^{\uparrow L} .$$

Write out all types in this law and draw a type diagram.

Exercise 7.2.5.4 Show that the types $(\mathbb{1} + \mathbb{1}) \times A$ and $A + A$ are equivalent. One direction of this equivalence is given by a function $\text{two}[\mathbf{A}]$ with the type signature

```
def two[A]: ((Either[Unit, Unit], A)) => Either[A, A] = ??? twoA : (\mathbb{1} + \mathbb{1}) \times A \rightarrow A + A .
```

Implement that function and prove that it satisfies the “naturality law”: for any $f^{:A \rightarrow B}$,

$$(\text{id} \boxtimes f) ; \text{two} = \text{two} ; f^{\uparrow E} ,$$

where $E^A \doteq A + A$ is the functor whose lifting $^{\uparrow E}$ was defined in Section 7.2.3. Write out the types in this law and draw a type diagram.

Exercise 7.2.5.5 Prove that the following laws hold for arbitrary $f^{:A \rightarrow B}$ and $g^{:C \rightarrow D}$:

$$\text{left projection law : } (f \boxtimes g) ; \pi_1 = \pi_1 ; f ,$$

$$\text{right projection law : } (f \boxtimes g) ; \pi_2 = \pi_2 ; g .$$

Exercise 7.2.5.6 Given arbitrary functors F and G , define the functor $L^A \doteq F^A \times G^A$ and prove that for arbitrary $f^{:A \rightarrow B}$,

$$f^{\uparrow L} ; \pi_1 = \pi_1 ; f^{\uparrow F} .$$

Write out the types in this law and draw a type diagram.

Exercise 7.2.5.7 Consider the functor L^A defined as

$$L^A \doteq \text{Int} \times \text{Int} + A .$$

Implement the functions fmap and flatten (denoted ftn_L) and write their code in matrix notation:

$$(f^{:A \rightarrow B})^{\uparrow L} : \text{Int} \times \text{Int} + A \rightarrow \text{Int} \times \text{Int} + B ,$$

$$\text{ftn}_L : \text{Int} \times \text{Int} + \text{Int} \times \text{Int} + A \rightarrow \text{Int} \times \text{Int} + A .$$

Exercise 7.2.5.8* Show that flatten (denoted ftn_L) from Exercise 7.2.5.7 satisfies the naturality law: for any $f^{:A \rightarrow B}$,

$$f^{\uparrow L} ; \text{ftn}_L = \text{ftn}_L ; f^{\uparrow L} .$$

8 Typeclasses and functions of types

8.1 Motivation and first examples

8.1.1 Constraining type parameters

The summation method, `sum`, works for any collection of numeric values:

```
scala> Seq(1, 2, 3).sum
res0: Int = 6

scala> Seq(1.0, 2.0, 3.0).sum
res1: Double = 6.0
```

We can use `sum` to compute the average of a sequence of numbers,

```
def avg(s: Seq[Double]): Double = s.sum / s.length
```

Can we generalize the averaging function `avg` from `Double` to other numeric types, e.g.

```
def avg[T](s: Seq[T]): T = ???
```

This code is impossible because averaging works only for certain types `T`, not for arbitrary types `T` as implied by the type signature above. We will be able to define `avg[T]` only if we constrain the type parameter `T` to be a type representing a suitable numeric value (e.g., `Float`, `Double`, or `BigDecimal`).

Another example of a similar situation is a function with type signature $A \times F^B \rightarrow F^{A \times B}$,

```
def inject[F[_], A, B](a: A, f: F[B]): F[(A, B)] = f.map(b => (a, b)) // Must have 'f.map'.
```

This function requires the type constructor `F[_]` to have a `map` method, i.e., to be a functor. We can implement `inject` only if we constrain the parameter `F` to be a functor.

What would that constraint be like? Consider an ordinary function with no type parameters, e.g.:

```
def f(x: Int): Int = x + 1
```

In this code, the syntax `x: Int` constrains the value of the argument `x` to be integer. It is a type error to apply `f` to a non-integer argument.

Using a similar syntax for *type parameters*, we would write the type signatures for `avg` and `inject` as

```
def avg[T: Fractional](s: Seq[T]): T
def inject[F[_]: Functor, A, B](a: A, f: F[B]): F[(A, B)]
```

Scala uses the syntax `[T: Fractional]` to constrain the type parameter `T` to “fractional numeric” types. Similarly, `[F[_]: Functor]` requires the type constructor `F[_]` to be a functor. Applying `avg` or `inject` to types that do not obey those constraints will be a type error detected at compile time.

In these examples, we are restricting a type parameter to a subset of possible types, because only types from that subset have certain properties that we need. A subset of types, together with the required properties that those types must satisfy, is called a **typeclass**. The syntax `[T: Fractional]` is a **typeclass constraint** that forces the type `T` to belong to the typeclass `Fractional`.

This chapter focuses on defining and using typeclasses and on understanding their properties. We will see in detail how the syntax such as `[T: Fractional]` is implemented and used.

8.1.2 Functions of types and values

The similarity between the type parameter `T` and the argument `s` is clear in this type signature,

```
def avg[T: Fractional](s: Seq[T]): T
```

8 Typeclasses and functions of types

We can view `avg` as a function that takes *two* parameters (a type τ and a value s) and returns a value. We can also view `avg` as a function from a *type* τ to a *value* of type $\text{Seq}[\tau] \Rightarrow \tau$. We may call functions of this kind **type-to-value** functions (TVF). The syntax for TVFs supported in a future version of Scala 3 will show this more clearly,

```
val avg: [T] => Seq[T] => T = ... // Scala 3 only.
```

To emphasize that `avg` is a TVF, we may write the type signature of `avg` as

```
def avg[T: Fractional]: Seq[T] => T // Inconvenient in practice! Use avg[T: Fractional](s: Seq[T]): T
```

A type constructor such as `Seq[_]` can be seen as a **type-to-type** function (TTF) because it takes a type τ and returns a new type $\text{Seq}[\tau]$.

Functions can map from values or from types and to values or to types, as this table shows:

functions...	from value	from type
to value	(VVF) <code>def f(x:Int):Int</code>	(TVF) <code>def pure[A]: A => List[A]</code>
to type	(VTF) dependent type	(TTF) <code>type MyData[A] = Either[Int, A]</code>

We have already seen examples of VVFs, TVFs, and TTFs. Value-to-type functions (VTFs) are known as **dependent types** (or, more verbosely, as “value-dependent types”). An example in Scala:

```
val x = new { type T = Int }
val y: x.T = 123
```

In this example, `x.T` is a dependent type because it is a type that depends on the *value* `x`. For the value `x` defined in this code, the expression `x.T` evaluates to the type `Int`.

We will not consider dependent types (VTFs) in this chapter because typeclasses only require a combination of a TTF and a TVF.

8.1.3 Partial functions of types and values

We would like to define the function `avg[T: Fractional]` as a TVF that can be applied only to a subset of possible types τ . This is similar to a **partial function**, i.e., a function defined only for a subset of possible values of its argument’s type. We may call such functions partial type-to-value functions (PTVFs), to distinguish them from partial value-to-value functions (PVVFs) we saw before.

In some situations, partial functions are safe to use. For instance, the following partial function `p`,

```
def p: Either[Int, String] => Int = { case Left(x) => x - 1 }
```

can be applied only to values of the form `Left(...)`. Applying `p` to a value `Right(...)` will cause a run-time error. However, consider this code:

```
val x = Seq(Left(1), Right("abc"), Left(2))
scala> x.filter(_.isLeft).map(p)
res0: Seq[Int] = List(0, 1)
```

Although `x.filter(_.isLeft)` has type `Seq[Either[Int, String]]`, all values in that sequence are guaranteed to be of type `Left`. So we know it is safe to apply the partial function `p` in `.map(p)`.

Although safe, this code is brittle: if the `filter` operation were moved to another place, we might by mistake write code equivalent to `x.map(p)`, causing a run-time exception. It is better to refactor the code so that the compile-time type-checking guarantees the safety of all operations at run time. For the example shown above, the `collect` method would make a partial function, such as `p`, safe to use:

```
scala> x.collect { case Left(y) => y - 1 }
res1: Seq[Int] = List(0, 1)
```

The `collect` method is a **total** function because it is defined for all values of its arguments and does not throw exceptions.

Total functions are safer to use than partial functions. The partial function `p` can be converted into a total function by changing its type to `Left[Int, String] => Int`. Another example: applying `head` to a `List` is unsafe, but the non-empty list type guarantees at compile time that the first element exists:

```
val xs: NonEmptyList[Int] = ...
val h = xs.head // .head is a total function for a NonEmptyList.
```

In these cases, we achieve safety by making types more strictly constrained. Similarly, partial type-to-value functions (PTVFs) become safe to use if we impose suitable typeclass constraints on the type parameters. Typeclasses can be viewed as a systematic way of safely managing PTVFs.

8.2 Implementing typeclasses

A typeclass constraint `[T: Fractional]` will generate a compile-time error when a function such as `avg[T]` is applied to an incorrectly chosen type parameter `T`. If the Scala library did not already implement the `Fractional` typeclass, how could we reproduce that functionality?

8.2.1 Creating a partial function at type level

The code needs to specify that the type parameter must belong to a certain subset of allowed types. To simplify the task, assume that the allowed types are `BigDecimal` and `Double`. One example of a type constraint is shown by unfunctors (see Section 6.1.6), which are type constructors whose type parameters are restricted to specific types. In this code:

```
sealed trait Frac[A]           // Unfunctor.
final case class FracBD() extends Frac[BigDecimal]
final case class FracD() extends Frac[Double]
```

values of type `Frac[A]` can be created only if `A = BigDecimal` OR `A = Double`. The keywords `sealed` and `final` guarantee that no further code could extend this definition and allow us to create a value of type, say, `Frac[String]` OR `Frac[Boolean]`. Although the Scala compiler will not detect any errors in this code,

```
type T = Frac[String]
type U = Frac[Boolean]
```

we will never be able to create and use any values of types `T` or `U`. In other words, the types `Frac[String]` and `Frac[Boolean]` are *void* types. Trying to create and use values of these types will result in type errors, as the following code shows:

```
1 def f[A]: Frac[A] = FracD()    // Type error.
2 val x: U = FracD()             // Type error.
3 val y: U = FracD().asInstanceOf[U]
4 y match { case FracD() => }    // Type error.
```

In line 3, we disabled the type checker and forced the Scala compiler to ignore the type error in the definition of `y`. However, line 4 shows that we are unable to use that value `y` in further computations.

The type `Frac[A]` is non-void (i.e., has values) only for `A` belonging to the set `{BigDecimal, Double}` of types. This set is called the **type domain** of the type function `Frac`. We now need to define the function `avg[T]` with a type parameter `T` constrained to that type domain.

The type constraint $T \in \{\text{BigDecimal, Double}\}$ is equivalent to the requirement that a value of type Frac^T should exist. So, we will implement the type constraint if we include an *additional argument* of type Frac^T into the type signature of `avg`:

```
def avg[T](s: Seq[T], frac: Frac[T]): T
```

The value `frac: Frac[T]` is called a **typeclass instance** value. Because that value needs to be passed to every call of `avg[T]`, we will be unable to use types `T` for which `Frac[T]` is void (i.e., has no values).

In this way, we implemented the typeclass constraint for the PTVF `avg[T]`. The main steps were:

1. Define a type constructor `Frac[_]`.
2. Make sure values of type `Frac[A]` exist only when `A = BigDecimal` OR `A = Double`.
3. Pass a value of type `Frac[T]` to the function `avg[T]` as an additional argument.

It is not necessary to define the type constructor `Frac` via an unfunctor. The type constructor `Frac` is only needed to define the type domain `{BigDecimal, Double}`. We can use a simple case class instead:

```
final case class Frac[T]()
val fracBD: Frac[BigDecimal] = Frac()
val fracD: Frac[Double] = Frac()
```

This code creates a type constructor `Frac` and makes values of type `Frac[T]` available for chosen type parameters `T`. In this way, we implemented the required type domain.

To write the code for `avg[T]`, we need to be able to add numeric values and to divide by an integer value. More precisely, the body of `avg[T]` needs access to two PTVFs that we may call `add` and `intdiv`,

```
def add[T](x: T, y: T): T
def intdiv[T](x: T, n: Int): T
```

Since `avg[T]` now has an additional argument `frac`, we may use that argument to hold the required functions. So, we redefine `Frac` as a named tuple (case class) containing the functions `add` and `intdiv`:

```
final case class Frac[T](add: (T, T) => T, intdiv: (T, Int) => T)
```

Typeclass instances for `BigDecimal` and `Double` are then created by the following code:

```
val fracBD = Frac[BigDecimal]( (x, y) => x + y, (x, n) => x / n )
val fracD = Frac[Double]( (x, y) => x + y, (x, n) => x / n )
```

With these definitions, implementing `avg[T]` becomes straightforward:

```
def avg[T](s: Seq[T], frac: Frac[T]): T = {
  val sum = s.reduce(frac.add)           // Assuming 's' is a non-empty sequence.
  frac.intdiv(sum, s.length)           // Here, 'reduce' would fail on an empty sequence 's'.
}
```

To use this function, we need to pass a typeclass instance corresponding to the type `T`:

```
scala> avg(Seq(1.0, 2.0, 3.0), fracD) // It will be a type error to use fracBD instead of fracD here.
res0: Double = 2.0

scala> avg(Seq(BigDecimal(1.0), BigDecimal(2.0)), fracBD)
res1: BigDecimal = 1.5
```

This is a fully working implementation of the `avg` function with a `Frac` typeclass constraint. We have achieved compile-time safety since `avg[T]` cannot be applied to values of unsupported types `T`. We have also achieved easy extensibility: To implement another function as a PTVF with the same type domain, we need to add an extra argument of type `Frac[T]` to the function. To add another supported type `T` to the type domain, we just write one more line of code similar to `val fracD = ...`

An equivalent implementation of the `Frac` typeclass via a `trait` with methods requires this code:

```
trait Frac[T] {                                // Trait is not 'sealed'.
  def add(x: T, y: T): T
  def intdiv(x: T, n: Int): T
}

val fracBD = new Frac[BigDecimal] {
  def add(x: BigDecimal, y: BigDecimal): BigDecimal = x + y
  def intdiv(x: BigDecimal, n: Int): BigDecimal = x / n
}

val fracD = new Frac[Double] {
  def add(x: Double, y: Double): Double = x + y
  def intdiv(x: Double, n: Int): Double = x / n
}
```

The function `avg[T]` will work unchanged with this implementation of the `Frac` typeclass.

The implementation via a `trait` is significantly longer than the code using a case class as shown previously. One advantage of the longer code is the ability to combine different typeclasses by `trait` mixing. We will look at that in more detail below. For now, we note that both implementations

will require the programmer to add a significant amount of new code:

- Calls to `func[T](args)` need to be changed to `func[T](args, ti)` with typeclass instances `ti`.
- For each supported type `T`, a corresponding typeclass instance value needs to be created.
- All those values need to be passed to all places in the code where PTVFs are used.

The extra work can be reduced (and sometimes avoided) by using Scala's "implicit value" feature.

8.2.2 Scala's implicit values

An **implicit** declaration is a feature of Scala that makes values automatically available to any function that declares an “implicit argument” of the same type. Scala’s syntax for implicit values is

```
implicit val x: Int = 123
```

This declaration introduces an implicit value of type `Int` into the current scope. That value will be automatically passed as an argument to any function declaring an argument of type `Int` as `implicit`:

```
def f(a: String)(implicit n: Int) = s"$a with $n"  
  
scala> f("xyz")  
res0: String = xyz with 123
```

We need to declare the arguments as `implicit` in the function’s type signature, and the implicit arguments must be in a *separate* argument list.

The simplest useful function with an implicit argument is the identity function. In the Scala library, this function is called `implicitly`. Compare its code with the code of the ordinary identity function:

```
def implicitly[T](implicit t: T): T = t  
def identity[T](t: T): T = t
```

What does `implicitly[T]` do? Since its only argument is declared as `implicit`, we can simply write `implicitly[T]` with no arguments to apply that function. (The type parameter usually needs to be specified.) If no implicit value of type `T` is available, a compile-time error will occur. If an implicit value of type `T` is available in the current scope, `implicitly[T]` will return that value:

```
implicit val s: String = "qqq"  
  
scala> implicitly[String]  
res1: String = qqq
```

It is an error to declare more than one implicit value of the same type in the same scope, because implicit arguments are specified by type alone. The Scala compiler will not be able to set implicit arguments of functions automatically when the function’s outer scope contains more than one implicit value of a required type, as in this code:

```
implicit val x: Int = 1  
implicit val y: Int = 2  
  
scala> implicitly[Int]  
<console>:14: error: ambiguous implicit values:  
  both value x of type => Int  
  and value y of type => Int  
  match expected type Int  
    implicitly[Int]  
    ^
```

But it is not an error to declare several implicit arguments of the *same* type, e.g.

```
def f(a: String)(implicit x: MyType, y: MyType)  
implicit val z: MyType = ???  
  
f("abc") // Same as f("abc")(z, z) since z is the unique implicit value of type MyType.
```

In the example above, the arguments `x` and `y` will be set to the same value, `z`. A compile-time error will occur if no `implicit` value of type `MyType` is visible in the current scope:

```
scala> implicitly[MyType]  
<console>:12: error: could not find implicit value for parameter e: MyType  
  implicitly[MyType]  
  ^
```

8.2.3 Implementing typeclasses by making instances `implicit`

The idea is to declare typeclass instance values as `implicit`. Typeclass instance arguments of functions are also declared as `implicit`. As a result, typeclass instances will be passed to all PTVFs automatically (as long as the appropriate implicits are visible in the scope of the PTVFs). This makes typeclasses easier to use because instance values need to be written out much less often.

The example with the `Frac` typeclass is implemented using implicit values like this:

```
final case class Frac[T](add: (T, T) => T, intdiv: (T, Int) => T)
implicit val fracBD = Frac[BigDecimal]( (x, y) => x + y, (x, n) => x / n )
implicit val fracD = Frac[Double]( (x, y) => x + y, (x, n) => x / n )
```

To define the function `avg[T]`, we declare an implicit argument as a `Frac` typeclass instance for `T`:

```
def avg[T](s: Seq[T])(implicit frac: Frac[T]): T = { // Assuming 's' is a non-empty sequence.
  val sum = s.reduce(frac.add) // Here, 'reduce' would fail on an empty sequence 's'.
  frac.intdiv(sum, s.length) // Compute 'sum/length'.
}
```

It is now easier to use the function `avg` because the typeclass instances are inserted automatically:

```
scala> avg(Seq(1.0, 2.0, 3.0))
res0: Double = 2.0

scala> avg(Seq(BigDecimal(1.0), BigDecimal(2.0)))
res1: BigDecimal = 1.5
```

Scala supports the “type constraint” syntax for implicit typeclass instance arguments: the code

```
def f[A, B](args...)(implicit t1: Typeclass1[A], t2: Typeclass2[B])
```

is equivalent to the shorter code

```
def f[A: Typeclass1, B: Typeclass2](args...)
```

The shorter code omits the names (`t1`, `t2`) of the typeclass instances. These values can be extracted via the standard function `implicitly` because all implicit arguments are automatically made available as implicit values in the scope of a function’s body. The code of `avg[T]` can then be written as

```
def avg[T: Frac](s: Seq[T]): T = {
  val frac = implicitly[Frac[T]]
  val sum = s.reduce(frac.add)
  frac.intdiv(sum, s.length)
}
```

When an implicit argument is required, the Scala compiler will search for implicit values of the required type in different places of the code. If implicit values are declared in another module, they can be made available by using an `import` declaration. In many cases, explicit `import` declarations can be avoided. One way to avoid them is to declare the required implicit values within the **companion object** of the typeclass (i.e., the Scala `object` with the same name as the type constructor):

```
final case class Frac[T](add: (T, T) => T, intdiv: (T, Int) => T)

object Frac { // The companion object of 'Frac[T]' creates some typeclass instances as 'implicit'.
  implicit val fracBD = Frac[BigDecimal]( (x, y) => x + y, (x, n) => x / n )
  implicit val fracD = Frac[Double]( (x, y) => x + y, (x, n) => x / n )
}
```

Whenever a function needs an implicit value of type `Frac[T]` for a specific type `T`, the Scala compiler will automatically look within the companion object of `Frac` (as well as within the companion object of the type `T`) for any instances declared there. So, the programmer’s code will not need to `import` those typeclass instances explicitly even if the companion object is in a different module:

```
scala> avg(Seq(1.0, 2.0, 3.0))
res0: Double = 2.0
```

8.2.4 Extension methods

In Scala, function applications can use three kinds of syntax:

1. The “function” syntax: arguments are to the right of the function as in `plus(x, y)` or `plus(x)(y)`.
2. The “method” syntax: the first argument is to the left of the function, and all other arguments (if any) are placed to the right of the function, as in `x.plus(y)` or `xs.foldLeft(0)(updater)`.
3. The “infix method” syntax (only applies to functions with two *explicit* arguments): no dot character is used, as in `x plus y`, or in `xs map {x => x + 1}`, or in `Set(1,2,3) contains 1`.

The last two syntax features are often used when writing chains of function applications, such as `xs.map(f).filter(g)`, because that code is easier to read than `filter(map(xs, f), g)`.

The method syntax is available only for methods defined in a class. A special feature of Scala allows us to add new functions with method syntax to previously defined types. New functions added in this way are called **extension methods**.

Suppose we would like to convert a previously defined function, say

```
def func(x: X, y: Y): Z = { ... }
```

into an extension method on the type `x` with the syntax `x.func(y)`. To do that, we define a new helper `class` that has a method named `func`. The class’s constructor must be declared as an `implicit` function having a *single* argument of type `x`:

```
implicit class FuncSyntax(x: X) { def func(y: Y): Z = ... }
```

After this code, we can write the method syntax `x.func(y)` as well as the infix method syntax `x func y`. The new syntax will work because the compiler automatically rewrites `x.func(y)` into `new FuncSyntax(x).func(y)`, creating a new temporary value `FuncSyntax(x)`. The method `func` will be available since it is defined in the class `FuncSyntax`.

As an example, let us define an extension method `average` for the type `Seq[T]`. Both the type parameter `T` and its typeclass constraint can be written in the constructor of the helper class:

```
implicit class AvgSyntax[T: Frac](xs: Seq[T]) {
  def average: T = avg(xs) // Use a different name, 'average', to avoid name clash with 'avg'.
}
```

We can now use the method `average` on numeric sequences:

```
scala> Seq(1.0, 2.0, 3.0).average
res0: Double = 2.0
```

The Scala compiler automatically rewrites the syntax `Seq(1.0, 2.0, 3.0).average` as the expression

```
new AvgSyntax(Seq(1.0, 2.0, 3.0))(implicitly[Frac[Double]]).average
```

In this way, the method `average` is actually invoked on a temporarily created value of type `AvgSyntax`. These values will be created automatically because the class constructor of `AvgSyntax` is declared as `implicit`. Since the constructor of `AvgSyntax` includes the typeclass constraint `[T: Frac]`, we will not be able to create values of type `AvgSyntax[T]` for types `T` not in the type domain of `Frac`.

This example illustrates the convenience of implementing PTVFs as extension methods. An extension method is defined only once but automatically becomes available for all types in the domain of the typeclass. Because of the typeclass constraint, the new method will be available *only* on values of supported types.

This convenience comes at a cost: helper classes such as `AvgSyntax` need to be explicitly imported into every scope where extension methods are used, with a declaration like this:

```
import some.library.having.a.longwinded.packagename.AvgSyntax
```

If the helper class is defined in some library, the programmer will have to look at the library’s source code to determine the full name of the helper class that needs to be imported.

8.2.5 Solved examples: Implementing typeclasses in practice

We will now look at some practical examples of programming tasks implemented via typeclasses.

Example 8.2.5.1 (metadata extractors) An application needs to work with data structures implemented in various external libraries. All those data structures are case classes containing certain metadata: “name” (a `String`) and “count” (a `Long` integer). However, the specific data structures define the metadata differently, each in its own way:

```
final case class Data1(p: String, q: String, r: Long) // "name" = p, "count" = r
final case class Data2(s: String, c: Long, x: Int)    // "name" = s, "count" = c
final case class Data3(x: Int, y: Long, z: String)   // "name" = z, "count" = x * y
```

The task is to implement two functions, `getName[T]` and `getCount[T]`, for extracting the metadata out of the data structures of type `T`, where `T` is one of `Data1`, `Data2`, `Data3`. Type signatures and sample tests:

```
def getName[T: HasMetadata](t: T): String = ???
def getCount[T: HasMetadata](t: T): Long = ???

scala> getName(Data2("abc", 123, 0))
res0: String = abc

scala> getCount(Data3(10, 20, "x"))
res1: Long = 200
```

Solution We will implement a typeclass `HasMetadata` and declare instances only for `Data1`, `Data2`, and `Data3`. The code for extracting the metadata will be contained within the typeclass instances. Since the metadata extractors have types `T => String` and `T => Long`, a simple solution is to define the typeclass as a `case class` containing these two functions:

```
final case class HasMetadata[T](getName: T => String, getCount: T => Long)
```

The required typeclass instances are declared as implicit values within the companion object:

```
object HasMetadata {    // Extract metadata from each type as appropriate.
  implicit val data1 = HasMetadata[Data1](_.p, _.r)
  implicit val data2 = HasMetadata[Data2](_.s, _.c)
  implicit val data3 = HasMetadata[Data3](_.z, data3 => data3.x * data3.y)
}
```

Now we can define `getName` and `getCount` as PTVFs with typeclass constraints. First, let us write the code using an implicit argument to pass the typeclass instance:

```
def getName[T](t: T)(implicit ti: HasMetadata[T]): String = ti.getName(t)
def getCount[T](t: T)(implicit ti: HasMetadata[T]): Long = ti.getCount(t)
```

Or we could use the typeclass constraint syntax, replacing `ti` by calls to `implicitly[...]`:

```
def getName[T: HasMetadata](t: T): String = implicitly[HasMetadata[T]].getName(t)
def getCount[T: HasMetadata](t: T): Long = implicitly[HasMetadata[T]].getCount(t)
```

This code defines PTVFs `getName` and `getCount` with the type domain that contains the three types `Data1`, `Data2`, `Data3`. In order to add a new type, say `Data4`, to the type domain, we will need to declare a new typeclass instance as an implicit value of type `HasMetadata[Data4]`. New implicit values can be defined anywhere in the code, not necessarily within the companion object `HasMetadata`. To avoid extra `import` statements, the implicit value may be defined within the companion object of `Data4`:

```
final case class Data4(x: Int, message: String)
object Data4 {    // Add Data4 to the type domain of HasMetadata:
  implicit val data4 = HasMetadata[Data4](_.message, _.x.toLong)
}

scala> getName(Data4(1, "abc"))
res2: String = abc
```

For convenience, let us declare the metadata extractors as extension methods:

```
implicit class ExtractorsSyntax[T: HasMetadata](t: T) {
  def name: String = getName(t)
  def count: Long = getCount(t)
}
```

With this definition, we can write:

```
scala> Data2("abc", 123, 0).name
res3: String = "abc"

scala> Data3(10, 20, "x").count
res4: Long = 200
```

need to modify. The typeclass pattern enables us to add externally defined types to a type domain whenever necessary, and to implement new PTVFs for them.

Example 8.2.5.2 (counters) A certain application needs to count the number of files processed and to send this number to external observability services. The functionality of a counter is provided by an external library as a special class `Counter` with an `inc` method that increments the counter. To test the code, we want to be able to pass a test-only counter provided by another library as a type `TestCounter` with an `incr` method. The task is to implement a function `bump[C]()`, where `c` is a type constrained to be one of the supported types of counters. The type signature and sample code:

```
def bump[C](): ... = ???

val counter = Counter(...)
val testCounter = TestCounter(...)

bump(counter)      // Should call counter.inc()
bump(testCounter) // Should call testCounter.incr()
```

Solution We will implement a typeclass `Bumpable` whose type domain contains the types `Counter` and `TestCounter`. Typeclass instances should allow us to increment a counter of any supported type. So, a typeclass instance value of type `Bumpable[C]` needs to contain a function of type `c => Unit` that will increment a counter of type `c` appropriately:

```
final case class Bumpable[C](bump: C => Unit)
```

We can now create the typeclass instances and implement the function `bump[C]`:

```
object Bumpable {
  implicit val b1 = Bumpable[Counter](c => c.inc())
  implicit val b2 = Bumpable[TestCounter](c => c.incr())

  def bump[C](counter: C)(implicit ti: Bumpable[C]): Unit = ti.bump(counter)
}
```

An equivalent implementation with the type constraint syntax looks like this:

```
def bump[C: Bumpable](counter: C): Unit = implicitly[Bumpable[C]].bump(counter)
```

Example 8.2.5.3 (default values) Certain types have naturally chosen “default” values (e.g., integer zero, empty string, empty array, etc.). The task is to implement a function `default[T]` restricted to types `T` for which default values are available. The required type signature and sample tests:

```
def default[T: HasDefault]: T = ???

scala> default[Int]
res0: Int = 0

scala> default[Double]
res1: Double = 0.0
```

Solution We need to define a PTVF `default[T]` with a type domain that contains (at least) the types `Int` and `Double`. For every supported type τ , we need to store the known default value of that type. So, the typeclass instance can be defined as a simple wrapper for values of type τ :

```
final case class HasDefault[T](value: T)
```

Typeclass instances are declared straightforwardly:

```
object HasDefault {
  implicit val defaultInt = HasDefault[Int](0)
  implicit val defaultDouble = HasDefault[Double](0.0)
  implicit val defaultString = HasDefault[String]("")
  implicit val defaultUnit = HasDefault[Unit](())
}
```

The implementation of `default[T]` is written as

```
def default[T](implicit ti: HasDefault[T]) = ti.value
```

When using the typeclass constraint syntax, the code is

```
def default[T: HasDefault]: T = implicitly[HasDefault[T]].value
```

How to define an empty list as a default value for lists of *any* chosen type? We cannot write

```
implicit val defaultList = HasDefault[List[A]](List()) // Error: 'A' is undefined.
```

The type parameter `A` needs to be defined in the left-hand side. Since Scala 2 does not support `val` declarations with type parameters (Scala 3 will), we need to write the typeclass instance as a `def`:

```
implicit def defaultList[A] = HasDefault[List[A]](List())
```

Another example of a `HasDefault` instance with a type parameter is for functions of type $A \rightarrow A$:

```
implicit def defaultFunc[A]: HasDefault[A => A] = HasDefault[A => A](identity)
```

Types that have default values are also called **pointed** types. This book defines the typeclass `Pointed` for pointed *functors* (Section 8.3.5) because they are more widely used than pointed types.

Example 8.2.5.4 (semigroups) In many cases, data items can be combined or merged into a larger data item of the same type. For instance, two numbers can be added, two sets combined into one set, two strings concatenated into one string, and two lists into one list. The “merge” operation can be defined as a function `combine[T]` taking two arguments of type τ and returning a new value of type τ . We will denote that operation by \oplus , e.g., we write $x \oplus y$. In all the examples just mentioned (integers, strings, lists, etc.), that operation is associative:

$$\forall(x, y, z). (x \oplus y) \oplus z = x \oplus (y \oplus z) \quad . \quad (8.1)$$

This associativity law makes parentheses in the expression $x \oplus y \oplus z$ unnecessary.

A type τ with an associative binary operation is called a **semigroup**. The task in this example is to define the semigroup operation for the types `Int`, `String`, and `List[A]`.

Solution For every supported type T , the required data is a function of type $T \times T \rightarrow T$. So, we define the typeclass as a wrapper over that type:

```
final case class Semigroup[T](combine: (T, T) => T)
```

The typeclass instances for the supported types are defined using a short syntax as

```
object Semigroup {
  implicit val semigroupInt = Semigroup[Int](_ + _)
  implicit val semigroupString = Semigroup[String](_ + _)
  implicit def semigroupList[A] = Semigroup[List[A]](_ ++ _)
}
```

The function `combine[T]` is implemented as

```
def combine[T](x: T, y: T)(implicit ti: Semigroup[T]): T = ti.combine(x, y)
```

Since `combine` is a binary operation, it is convenient to define infix method syntax for it:

```
implicit class SemigroupSyntax[T: Semigroup](x: T) { def |+|(y: T): T = combine(x, y) }
```

After this definition, we may use the infix operation `|+|` like this,

```
scala> List(1, 2, 3) |+| List(4)
res0: List[Int] = List(1, 2, 3, 4)
```

Due to the associativity law (8.1), the result of $x \mid+| y \mid+| z$ does not depend on the choice of parentheses: $(x \mid+| y) \mid+| z == x \mid+| (y \mid+| z)$. This makes programs written using the semigroup operation `|+|` easier to understand and reason about.

Semigroup types represent data that can be pairwise “merged” in a certain well-defined way. Using the `Semigroup` typeclass, we can write code that is parameterized by the type of “mergeable” data. As an example, given a `Seq[T]` where the type `T` is a semigroup, we can “merge” all elements to compute a result of type `T`. This computation can be implemented as a function parameterized by `T`,

```
def merge[T: Semigroup](ts: Seq[T]): T = ts.reduce(combine[T])
```

This function assumes a non-empty input sequence `ts` whose elements are of a semigroup type `T`. We can also implement the same function as an extension method,

```
implicit class SumSyntax[T: Semigroup](ts: Seq[T]) { def merge: T = ts.reduce(combine[T]) }
```

With the previous definitions, we can now evaluate expressions such as

```
scala> Seq(1, 2, 3).merge
res1: Int = 6

scala> Seq(List(), List(true), List(), List(true, false)).merge
res2: List[Boolean] = List(true, true, false)
```

It is important that the associativity law (8.1) should hold for each of the supported types. If that is not so, programs written with `merge` will not work as expected. (For instance, programmers would certainly expect that `xs.merge |+| ys.merge == (xs ++ ys).merge` for any sequences `xs` and `ys`.) However, the code of the typeclass *does not* check the associativity law. It is the responsibility of the programmer to verify that the implementation of each typeclass instance is lawful.

The associativity law of integer addition is a standard arithmetic identity

$$(x + y) + z = x + (y + z) \quad .$$

Verifying associativity for lists and strings (which are lists of characters) is intuitively simple because concatenation preserves the order of elements. If `x`, `y`, and `z` are lists, the concatenation `(x ++ y) ++ z` is a list containing all elements from `x`, `y`, and `z` in their original order. It is evident that the concatenation `x ++ (y ++ z)` is a list with the same elements in the same order. However, a rigorous proof of the associativity law for lists, starting from the code of the `concat` function, requires significant work (see Section 8.5.2).

Example 8.2.5.5 (alternative semigroup implementations) The definitions of the semigroup operation \oplus as concatenation for strings and as addition for integers may appear to be “natural”. However, alternative implementations are useful in certain applications. As long as the associativity law holds, *any* function of type $T \times T \rightarrow T$ may be used as the semigroup operation. The task of this example is to show that the following implementations of the semigroup operation are lawful and to implement the corresponding typeclass instances in Scala.

- (a) For any type T , define $x \oplus y \triangleq x$ (ignoring the value of y).
- (b) For pair types $T \triangleq A \times B$, define the operation \oplus by $(a_1 \times b_1) \oplus (a_2 \times b_2) \triangleq a_1 \times b_2$.
- (c) For $T \triangleq \text{String}$, define $x \oplus y$ as the longer of the strings x and y .
- (d) For $T \triangleq S \rightarrow S$ (the type S is fixed), define $x \oplus y \triangleq x \circ y$ (the forward function composition) or $x \oplus y \triangleq x \circ y$ (the backward function composition).

Solution (a) To verify the associativity law, use the definition $x \oplus y \triangleq x$ to compute

$$(x \oplus y) \oplus z = x \oplus z = x \quad , \quad x \oplus (y \oplus z) = x \quad .$$

So the associativity law holds: $(x \oplus y) \oplus z = x \oplus (y \oplus z)$ for any x, y, z .

It is clear that $x \oplus y \oplus z \oplus \dots = x$ for any number of values; the binary operation keeps the first value and ignores all further values. We can implement this semigroup instance at once for all types T :

```
implicit def semigroup1[T] = Semigroup[T]{ (x, y) => x }
```

Similarly, the definition $x \oplus y \triangleq y$ gives an associative binary operation for a (different) semigroup.

(b) To verify the associativity law:

$$\begin{aligned} ((a_1 \times b_1) \oplus (a_2 \times b_2)) \oplus (a_3 \times b_3) &= (a_1 \times b_1) \oplus (a_3 \times b_3) = a_1 \times b_3 \quad , \\ (a_1 \times b_1) \oplus ((a_2 \times b_2) \oplus (a_3 \times b_3)) &= (a_1 \times b_1) \oplus (a_2 \times b_3) = a_1 \times b_3 \quad . \end{aligned}$$

The implementation is possible for any types A, B :

```
implicit def semigroup2[A, B] = Semigroup[(A, B)]{ case ((a1, b1), (a2, b2)) => (a1, b2) }
```

One use case for this semigroup is to maintain a pair of timestamps for the first and the last events in a temporally ordered series. Merging two such pairs for consecutive events means to keep the first value from the first pair and the second value from the second pair.

(c) It is clear that $x \oplus y \oplus z$ is the longest of the strings x, y , and z . Since the definition of “longest” does not depend on the order in which we select pairs for comparison, the operation is associative. (For the same reason, any “maximum” or “minimum” operation is associative.) Implementation:

```
implicit val semigroup3 = Semigroup[String]{ (x, y) => if (x.length > y.length) x else y }
```

(d) The composition of functions is associative (see proofs in Section 4.2.2). Whether we choose to define $x \oplus y \oplus z = x \circ y \circ z$ or $x \oplus y \oplus z = x \circ y \circ z = z \circ y \circ x$, the results do not depend on inserting parentheses. The code for these two typeclass instances is

```
implicit def semigroup4[S] = Semigroup[S => S]{ (x, y) => x andThen y }
implicit def semigroup5[S] = Semigroup[S => S]{ (x, y) => x compose y }
```

Example 8.2.5.6 (monoids) When a data type is a semigroup, it is often possible to find a special value that acts as an “default” value with respect to the semigroup operation. Merging with the default value will not change any other value. For instance, concatenating with an empty list does not change any other list; so the empty list plays the role of the default value for lists. Merging an empty set with any other set does not change the other set; so the empty set is the default for sets.

A semigroup with such a default value is called a **monoid**. Formally, a type T is a monoid when it has an associative binary operation \oplus_T and a chosen default or “empty” value e^T such that

$$\text{for all } x^T : \quad e \oplus_T x = x \quad , \quad x \oplus_T e = x \quad . \quad (8.2)$$

The laws (8.2) are called the **identity laws** of monoid.

The task in this example is to define a typeclass describing monoids.

Solution The typeclass instances should contain the same information as semigroups and additionally the default value of type T . So, we write

```
final case class Monoid[T](combine: (T, T) => T, empty: T)
```

Let us define some typeclass instances for illustration:

```
object Monoid {
  implicit val monoidInt = Monoid[Int](_ + _, 0)
  implicit val monoidString = Monoid[String](_ + _, "")
  implicit def monoidList[A] = Monoid[List[A]](_ ++ _, List())
  implicit def monoidFunc[A] = Monoid[A => A](_ andThen _, identity)
}
```

Monoids formalize the general properties of data aggregation. Section 8.3.4 will study monoids in more detail and show further examples of their use. At this point, we look at one more example that defines a `Monoid` instance using two previously defined typeclasses, `Semigroup` and `HasDefault`.

Example 8.2.5.7 (monoids as semigroups with default) A monoid combines the properties of a semigroup and a type with a default value. If a type T is a semigroup and has a default value, it is likely that T is a monoid. The task is to define a `Monoid` typeclass instance given `Semigroup` and `HasDefault` instances for a type T .

Solution We need to define a function `monoidOf[T]` that returns the required monoid typeclass instance for T . The typeclass constraints for this function are `Semigroup` and `HasDefault`. So the type signature must be

```
def monoidOf[T](implicit ti1: Semigroup[T], ti2: HasDefault[T]): Monoid[T] = ???
```

To implement a value of type `Monoid[T]`, we need to provide a function of type $T \times T \rightarrow T$ and a value of type T . Precisely that data is available in the typeclass instances of `Semigroup` and `HasDefault`, and so it is natural to use them,

```
def monoidOf[T](implicit ti1: Semigroup[T], ti2: HasDefault[T]): Monoid[T] =
  Monoid(ti1.combine, ti2.value)
```

Using the type constraint syntax, the equivalent code is

```
def monoidOf[T: Semigroup : HasDefault]: Monoid[T] =
  Monoid(implicitly[Semigroup[T]].combine, implicitly[HasDefault[T]].value)
```

We can also define this function as an `implicit`, so that every type T with a `Semigroup` and `HasDefault` instances will automatically receive a `Monoid` typeclass instance as well.

Writing the types of the `Semigroup`, `HasDefault`, and `Monoid` instances in the type notation, we get

$$\text{Semigroup}^T \triangleq T \times T \rightarrow T \quad , \quad \text{HasDefault}^T \triangleq T \quad , \quad \text{Monoid}^T \triangleq (T \times T \rightarrow T) \times T \quad .$$

It is clear that

$$\text{Monoid}^T \cong \text{Semigroup}^T \times \text{HasDefault}^T \quad .$$

Indeed, the code for `monoidOf` computes a pair of values from the `Semigroup` and `HasDefault` instances.

Is this implementation lawful with respect to the monoid laws? The associativity law will hold for the monoid if the `Semigroup` typeclass instance was already lawful. However, the value stored in the `HasDefault` instance is not guaranteed to satisfy the identity laws (8.2) with respect to the `combine` operation stored in the `Semigroup` instance. The programmer must verify that the identity laws hold. It can happen that, for some type T , typeclass instances of `Semigroup` and `HasDefault` already exist but are not “compatible”, so that the monoid identity laws do not hold. In that case, a simple combination of `Semigroup` and `HasDefault` instances will not work, and a different `Monoid` instance must be defined.

Are there alternative implementations of the `Monoid` typeclass instance given `Semigroup` and `HasDefault` instances? The function `monoidOf` needs to produce a value of type $(T \times T \rightarrow T) \times T$ given values of type $T \times T \rightarrow T$ and a value of type T :

$$\text{monoidOf} : (T \times T \rightarrow T) \times T \rightarrow (T \times T \rightarrow T) \times T \quad .$$

When the type signature of `monoidOf` is written in this notation, it is clear that `monoidOf` should be the identity function; indeed, that is what our code translates to. Although there are many other implementations of the same type signature, only the code shown above will satisfy the monoid laws. An example of an unlawful implementation is

```
def badMonoidOf[T](implicit ti1: Semigroup[T], ti2: HasDefault[T]): Monoid[T] =
  Monoid((x, y) => ti1.combine(x, ti1.combine(x, y)), ti2.value)
```

This implementation defines the monoid operation as $x \oplus x \oplus y$ instead of the correct definition $x \oplus y$. If we set $y = e_T$, we will get $x \oplus x$ instead of x , violating one of the identity laws.

8.2.6 Typeclasses for type constructors

An example of a function parameterized by a type constructor is

```
def inject[F[_]: Functor, A, B](a: A, f: F[B]): F[(A, B)] = ???
```

We can use a suitable typeclass to implement this constraint. Since the type parameter `F` is itself a type constructor, the typeclass constructor will use the syntax `Functor[F[_]]`.

What information needs to be wrapped by a typeclass instance? A functor `F` must have a `map` function with the standard type signature

```
def map[A, B](fa: F[A])(f: A => B): F[B]
```

In the type notation, this type signature is written as

$$\text{map} : \forall(A, B). F^A \rightarrow (A \rightarrow B) \rightarrow F^B .$$

So, a typeclass instance of the `Functor` typeclass must contain this function as a value. But defining the typeclass as before via a `case class` does not work with Scala 2,

```
final case class Functor[F[_]](map: ∀(A, B). F[A] => (A => B) => F[B]) // Not possible in Scala 2.
```

Scala 3 will directly support an argument type that *itself* contains type quantifiers such as $\forall(A, B)$. In Scala 2, we need to represent such “nested” type quantifiers by writing a `trait` with a `def` method:

```
trait Functor[F[_]] {
  def map[A, B](fa: F[A])(f: A => B): F[B]
}
```

The type constructor `Functor` has the type parameter `F[_]`, which must be itself a type constructor. For any type constructor `F`, a value of type `Functor[F]` is a wrapper for a value of type $\forall(A, B). F^A \rightarrow (A \rightarrow B) \rightarrow F^B$.

Values of type `Functor` (i.e., typeclass instances) are implemented with the “`new { ... }`” syntax:

```
implicit val functorSeq = new Functor[Seq] {
  def map[A, B](fa: Seq[A])(f: A => B): Seq[B] = fa.map(f)
}
```

This is currently the most common way of defining typeclasses in Scala.

It is convenient to declare `map` as an extension method on the `Functor` type constructors,

```
implicit class FunctorSyntax[F[_]: Functor, A](fa: F[A]) { // Syntax helper class.
  def map[B](f: A => B): F[B] = implicitly[Functor[F]].map(fa)(f)
}
```

If this class definition is in scope, the `map` method becomes available for values of functor types.

Using the `Functor` typeclass and the syntax helper, we can now implement the function `inject`:

```
def inject[F[_]: Functor, A, B](a: A, f: F[B]): F[(A, B)] = f.map { b => (a, b) }

scala> inject("abc", Seq(1, 2, 3)) // Assuming that implicit Functor[Seq] is in scope.
res0: Seq[(String, Int)] = List(("abc", 1), ("abc", 2), ("abc", 3))
```

Just like the `Monoid` typeclass, the code of the `Functor` typeclass does not enforce the functor laws on the implementation. It is the programmer responsibility to verify that the laws hold.

One way of checking the laws is to use the `scalacheck` library¹ that automatically runs random tests for the given assertions, trying to discover a set of values for which some assertion fails. Using the `Functor` typeclass constraint, we can implement a function (in our terminology, a PTVF) that checks the functor laws for *any* given type constructor `F[_]`:

```
import org.scalacheck.Arbitrary // Necessary imports and definitions.
import org.scalatest.prop.GeneratorDrivenPropertyChecks
class FunctorTest extends Matchers with GeneratorDrivenPropertyChecks {
  def checkFunctorLaws[F[_], A, B, C]()(implicit ff: Functor[F], // Use the 'Arbitrary' typeclass
    fa: Arbitrary[F[A]], ab: Arbitrary[A => B], bc: Arbitrary[B => C]) = { // from 'scalacheck'.
    forAll { (fa: F[A]) => fa.map(identity[A]) shouldEqual fa } // Identity law. "For all 'fa', ..."
    forAll { (f: A => B, g: B => C, fa: F[A]) => // Composition law. The assertion must hold
      fa.map(f).map(g) shouldEqual fa.map(f andThen g) // for all 'f', 'g', 'fa'.
    }
  }
}
```

¹<https://www.scalacheck.org>

```

    }
    // Check the laws for F = Seq using specific types instead of type parameters A, B, C.
    checkFunctorLaws[Seq, Int, String, Double]()
}

```

The `scalacheck` library will substitute a large number of random values into the given assertions. Note that the laws are being tested only with a finite number of values and with type parameters set to specific types. While it is useful to test laws with `scalacheck` (we might find a bug), only a symbolic derivation provides a rigorous proof that the laws hold. One of the main themes of this book is to show how to perform symbolic derivations efficiently.

8.3 Deriving typeclass instances via structural analysis of types

In Chapter 6 we analyzed the structure of functors by checking which of the six standard type constructions can make new functors out of previous ones. We will now apply the same **structural analysis** to various typeclasses. Is a product of two monoids a monoid? Is a co-product of two semigroups a semigroup? Answers to such questions will enable us to:

- Quickly decide whether a given type can have a typeclass instance of `Monoid`, `Semigroup`, etc.
- If so, derive the code for the new typeclass instance without guessing.
- Have assurance that the required typeclass laws will hold for newly constructed instances.

In the following sections, we will show how to use this approach for some simple typeclasses.

8.3.1 Extractors

Our first typeclass is a generalization of Example 8.2.5.1. The functionality was to extract metadata of fixed types from a value of type τ . The typeclass instance contained a pair of functions,

```
final case class HasMetadata[T](getName: T => String, getCount: T => Long)
```

In the type notation, this type constructor is written as

$$\text{HasMetadata}^T \triangleq (T \rightarrow \text{String}) \times (T \rightarrow \text{Long}) \quad .$$

A standard type equivalence (see Table 5.6) shows that this type is equivalent to $T \rightarrow \text{String} \times \text{Long}$. This motivates us to denote $\text{String} \times \text{Long}$ by Z and to consider a typeclass whose instances are values of type $T \rightarrow Z$. We may call this typeclass a “ Z -extractor” since types T from its type domain permit us somehow to extract values of type Z . With a fixed type Z , we denote the typeclass by

$$\text{Extractor}^T \triangleq T \rightarrow Z \quad .$$

```
final case class Extractor[Z, T](extract: T => Z)
```

What kind of types can have an `Extractor` typeclass instance? To answer that question, we apply structural analysis and check whether any of the standard type constructions produce new typeclass instances. The results are summarized in Table 8.1. Let us show the required calculations.

Fixed types We check whether an `Extractor` typeclass instance can be computed for the `Unit` type or for another fixed type C . To compute $\text{Extractor}^{\mathbb{1}} = \mathbb{1} \rightarrow Z$ requires creating a value of type Z from scratch, which we cannot do in a fully parametric function. For a fixed type C , a value of type $\text{Extractor}^C = C \rightarrow Z$ can be computed only if we can compute a value of type Z given a value of type C . This is possible only if we choose C as $C = Z$. The typeclass instance for Extractor^Z is implemented as an identity function of type $Z \rightarrow Z$:

```
implicit def extractorZ[Z] = Extractor[Z, Z](identity)
```

Type parameters Creating a typeclass instance Extractor^A for an arbitrary type A means to compute $\forall A. A \rightarrow Z$; this is not possible since we cannot create values of type Z using only a value of an unknown type A . So, there is no `Extractor` for arbitrary types A .

Products If the types A and B are known to be within the type domain of `Extractor`, can we add the pair $A \times B$ to that domain? If we can extract a value of type Z from each of two values $a:A$ and $b:B$, we can certainly extract a value of type Z from the product $a \times b$ by choosing to extract only from a or only from b . So, it appears that we have two possibilities for implementing the typeclass for $A \times B$. Reasoning more rigorously, we see that computing a new typeclass instance from two previous ones requires implementing a conversion function with type signature

$$\forall(A, B, Z). \text{Extractor}^A \times \text{Extractor}^B \rightarrow \text{Extractor}^{A \times B} = (A \rightarrow Z) \times (B \rightarrow Z) \rightarrow A \times B \rightarrow Z .$$

We can derive only two fully parametric implementations of this type signature:

$$f:A \rightarrow Z \times g:B \rightarrow Z \rightarrow a \times b \rightarrow f(a) \quad \text{and} \quad f:A \rightarrow Z \times g:B \rightarrow Z \rightarrow a \times b \rightarrow g(b) .$$

Both implementations give a valid `Extractor` instance (since there are no laws to check). However, every choice will use one of the two `Extractor` instances and ignore the other. So, we can simplify this construction by keeping the typeclass constraint only for A and allowing *any* type B ,

$$\text{extractorPair} : \forall(A, B, Z). \text{Extractor}^A \rightarrow \text{Extractor}^{A \times B} .$$

```
def extractorPair[Z, A, B](implicit ti: Extractor[Z, A]) =
  Extractor[Z, (A, B)] { case (a, b) => ti.extract(a) }
```

If A has an `Extractor` instance, the product of A with any type B also has an `Extractor` instance. Examples of this construction are $Z \times B$ and $P \times Q \times Z$ (since the type Z itself has an `Extractor` instance).

Co-products Given typeclass instances Extractor^A and Extractor^B , can we compute a value of type Extractor^{A+B} ? Writing out the types, we get

$$\text{Extractor}^A \times \text{Extractor}^B \rightarrow \text{Extractor}^{A+B} = (A \rightarrow Z) \times (B \rightarrow Z) \rightarrow A + B \rightarrow Z .$$

Due to a known type equivalence (Table 5.6), we have a unique implementation of this function:

$$\text{extractorEither} \triangleq f:A \rightarrow Z \times g:B \rightarrow Z \rightarrow \left| \begin{array}{c|c} & Z \\ \hline A & a \rightarrow f(a) \\ B & b \rightarrow g(b) \end{array} \right| .$$

```
def extractorEither[Z, A, B](implicit ti1: Extractor[Z, A], ti2: Extractor[Z, B]) =
  Extractor[Z, Either[A, B]] {
    case Left(a)    => ti1.extract(a)
    case Right(b)   => ti2.extract(b)
  }
```

So, the co-product of A and B can be given a unique `Extractor` instance.

Since the product and the co-product constructions preserve `Extractor` instances, we conclude that any polynomial type expression has an `Extractor` instance as long as every product type contains at least one Z or another `Extractor` type. For example, the type expression

$$A \times Z + Z \times (P + Z \times Q) + B \times C \times Z$$

is of that form and therefore has an `Extractor` instance. Generally, a polynomial in Z will have an `Extractor` instance only if the polynomial is of the form $Z \times P^Z$ for some functor P .

Function types We need to investigate whether $C \rightarrow A$ or $A \rightarrow C$ can have an `Extractor` instance for some choice of C , assuming that we have an instance for A . The required conversion functions must have type signatures

$$\text{Extractor}^A \rightarrow \text{Extractor}^{A \rightarrow C} \quad \text{or} \quad \text{Extractor}^A \rightarrow \text{Extractor}^{C \rightarrow A} \quad .$$

Writing out the types, we find

$$(A \rightarrow Z) \rightarrow (A \rightarrow C) \rightarrow Z \quad \text{or} \quad (A \rightarrow Z) \rightarrow (C \rightarrow A) \rightarrow Z \quad .$$

None of these type signatures can be implemented. The first one is hopeless since we do not have values of type A ; the second one is missing values of type C . However, since the type C is fixed, we may store a value of type C as part of the newly constructed type. So, we consider the pair type $C \times (C \rightarrow A)$ and find that its `Extractor` instance, i.e., a value of type $C \times (C \rightarrow A) \rightarrow Z$, can be derived from a value of type $A \rightarrow Z$ as

$$f^{A \rightarrow Z} \rightarrow c^C \times g^{C \rightarrow A} \rightarrow f(g(c)) \quad .$$

```
def extractorFunc[Z, A, C](implicit ti: Extractor[Z, A]) =
  Extractor[Z, (C, C => A)] { case (c, g) => ti.extract(g(c)) }
```

Examples of this construction are the type expressions $C \times (C \rightarrow Z)$ and $D \times (D \rightarrow Z \times P)$.

Another situation where an `Extractor` instance exists for the type $C \rightarrow A$ is when the type C has a known “default value” e_C (as in the `HasDefault` typeclass). In that case, we may omit the first C in $C \times (C \rightarrow A)$ and instead substitute the default value when necessary.

Recursive types We can use type recursion with any of the non-recursive constructions that create a new type with an `Extractor` instance out of a previous such type. For clarity, let us use type constructors F_1, F_2 , etc., for describing the new types produced out of previous extractor types. For the product construction, we define $F_1^{B,A} \triangleq A \times B$. For the co-product construction, $F_2^{B,A} \triangleq B + A$ (where B must be also an extractor type). For the function construction, $F_3^{C,A} \triangleq C \times (C \rightarrow A)$.

Any recursive type equation that uses F_1, F_2 , and/or F_3 will define a new recursive type with an `Extractor` instance. An example of such a recursive type is a composition of F_2 and F_1 defined by

$$T \triangleq F_2^{Z \times P, F_1^{T,Q}} = Z \times P + Q \times T \quad . \quad (8.3)$$

We can visualize this recursive type as an “infinite disjunction”

$$\begin{aligned} T &\cong Z \times P + Q \times T \\ &= Z \times P + Q \times (Z \times P + Q \times (Z \times P + \dots)) \\ &= Z \times P \times (1 + Q + Q \times Q + \dots) \cong Z \times P \times \text{List}^Q \quad . \end{aligned}$$

Since the resulting type is equivalent to $Z \times C$ where $C \triangleq P \times \text{List}^Q$, we find that the recursive equation (8.3) is equivalent to the product construction with a different type.

This will happen with any recursive equation containing F_1 and F_2 (but no F_3): since F_1 and F_2 are polynomial functors, the resulting type T will be a recursive polynomial in Z . A polynomial in Z will have an `Extractor` instance only if the polynomial is of the form $Z \times P^Z$ for some functor P .

Recursive equations involving F_3 will produce new examples of `Extractor` types, such as

$$T \triangleq F_3^{C, F_2^{Z, F_1^{T,P}}} = C \times (C \rightarrow Z + P \times T) \quad . \quad (8.4)$$

Heuristically, this type can be seen as an “infinite” exponential-polynomial type expression

$$T = C \times (C \rightarrow Z + P \times C \times (C \rightarrow Z + P \times C \times (C \rightarrow Z + \dots))) \quad .$$

Types of this form are useful in some applications involving lazy streams.

We will now show how to define an `Extractor` instance for any recursive type defined using F_1 , F_2 , and/or F_3 . A recursive type equation defining a type T can be written generally as $T \triangleq S^T$, where S is a type constructor built up by composing F_1 , F_2 , and/or F_3 in some way. (The type constructor S may use Z or other fixed types.) For each of F_1 , F_2 , and/or F_3 , we implemented a function with type

$$\text{extractorF} : \text{Extractor}^A \rightarrow \text{Extractor}^{F^A} .$$

Since S is a composition of F_1 , F_2 , and/or F_3 , we are able to implement a function

$$\text{extractorS} : \text{Extractor}^A \rightarrow \text{Extractor}^{S^A} .$$

The `Extractor` instance for the recursive type T is then defined recursively as

$$x^{T \rightarrow Z} \triangleq \text{extractorS}(x) .$$

The types match because the type T is equivalent to the type S^T . As long as the definition of the recursive type T is valid (i.e., the type recursion terminates), the extractor function will also terminate.

To illustrate this construction, let us derive an `Extractor` instance for the type T defined by Eq. (8.4). That equation has the form $T = S^T$ with the type constructor S defined by $S^A \triangleq C \times (C \rightarrow Z + P \times A)$.

```
type S[A] = (C, C => Either[Z, (P, A)]) // The fixed types 'C' and 'P' must be defined previously.
final case class TypeT(t: S[TypeT]) // Define the recursive type 'TypeT'.
```

To implement the function of type `Extractor[T] => Extractor[S[T]]`, which is $(T \rightarrow Z) \rightarrow S^T \rightarrow Z$, we begin with a typed hole

$$f^{T \rightarrow Z} \rightarrow s^{C \times (C \rightarrow Z + P \times T)} \rightarrow ???^Z .$$

To fill $???^Z$, we could apply $f^{T \rightarrow Z}$ to some value of type T ; but the only value of type T can be obtained if we apply the function of type $C \rightarrow Z + P \times T$ to the given value of type C . So we write

$$f^{T \rightarrow Z} \rightarrow c^C \times g^{C \rightarrow Z + P \times T} \rightarrow g(c) \triangleright ???^{Z + P \times T \rightarrow Z} .$$

The new typed hole has a function type. We can write the code in matrix notation as

$$\text{extractorS} \triangleq f^{T \rightarrow Z} \rightarrow c^C \times g^{C \rightarrow Z + P \times T} \rightarrow g(c) \triangleright \left| \begin{array}{c|c} & Z \\ \hline Z & \text{id} \\ P \times T & _^{P \times T} \rightarrow f(t) \end{array} \right| .$$

```
def extractorS[A](f: Extractor[A]): Extractor[S[A]] = Extractor[S[A]] { case (c, g) =>
  g(c) match {
    case Left(z)      => z
    case Right((_, t)) => f.extract(t)
  }
}
```

The recursive construction defines an `Extractor` instance for T by a recursive equation,

$$\text{extractorT} \triangleq \text{extractorS}(\text{extractorT}) . \quad (8.5)$$

```
def extractorT: Extractor[TypeT] = Extractor[TypeT] { case TypeT(t) =>
  extractorS(extractorT).extract(t)
}
```

To test this code, we define a value of type T while setting `C = Int`, `P = Boolean`, and `z = String`:

```
val t = TypeT((10, x => Right((true, TypeT((x * 2, y => Left("abc")))))))

scala> extractorT.extract(t) // The recursive definition of 'extractorT' terminates.
res0: String = abc
```

Why the recursion terminates The above code shows that the recursive definition (8.5) terminates. But why? A recursive definition of the form $x \triangleq f(x)$ can easily create an infinite loop, as in this code:

```
def f(x: Int): Int = x + 1
def x: Int = f(x)

scala> x          // Infinite loop: f(f(f(f(...))))
java.lang.StackOverflowError
```

The code for `extractorT` works because `extractorT` is a value of a *function* type, and because the presence of the case class `Extractor` forces us to rewrite Eq. (8.5) in the form of an “expanded function”,

$$\text{extractorT} \triangleq t \rightarrow \text{extractorS}(\text{extractorT})(t) \quad .$$

```
def extractorT: Extractor[TypeT] = Extractor { case TypeT(t) => extractorS(extractorT).extract(t) }
```

Although a function f is equivalent to its **expanded form** $t \rightarrow f(t)$, there is an important difference: using expanded forms in the code will make recursive definitions terminate.

To see an example of that, consider a recursive equation $f \triangleq k(f)$, where k is some function. The Scala code `val f = k(f)` creates an infinite loop when we compute anything that involves f :

```
def k(f: Int => Int): Int => Int = { x => if (x <= 0) 1 else 2 * f(x - 1) }
def f: Int => Int = k(f)    // This definition is invalid!

scala> f          // Infinite loop: k(k(k(k(...))))
java.lang.StackOverflowError

scala> f(4)        // Infinite loop: k(k(k(k(...)))(4)
java.lang.StackOverflowError
```

This code is clearly invalid. But if we expand the right-hand side of the recursive equation to

$$f \triangleq t \rightarrow k(f)(t)$$

instead of $f \triangleq k(f)$, the code will become valid, and the infinite loop disappears:

```
def f: Int => Int = { x => k(f)(x) }    // This defines f(n) = 2^n for n ≥ 0.

scala> f          // We can compute f without an infinite loop.
res0: Int => Int = <function1>

scala> f(4)        // We can compute f(4) without an infinite loop.
res1: Int = 16
```

The recursive use of f now occurs *within* a function body, and so $k(f)$ is evaluated only when f is applied to an argument. This allows the recursive definition of f to terminate.

Summary We derived the constructions that create new types with `Extractor` typeclass instances from previous ones. Any number of these constructions can be combined to create a new type expression that will always have an `Extractor` instance. An example is the type expression

$$K^{Z,P,Q,R,S} \triangleq Z \times P + Q \times (Q \rightarrow Z + R \times (R \rightarrow Z \times S)) \quad .$$

Since the type K is built up step by step from fixed types via the product, co-product, and function constructions, an `Extractor` instance for K can be derived systematically with no guessing:

```
type K[Z,P,Q,R,S] = Either[(Z, P), (Q, Q => Either[Z, (R, R => (Z, S))])]

implicit def extractorK[Z,P,Q,R,S]: Extractor[Z, K[Z,P,Q,R,S]] = {           // Extractor values for types:
  implicit val e1 = extractorPair[Z, Z, S]                                // Z × S. Needs Extractor[Z, Z].
  implicit val e2 = extractorFunc[Z, (Z, S), R]                            // R × (R ⇒ Z × S).
  implicit val e3 = extractorEither[Z, Z, (R, R => (Z, S))]              // Z + R × (R ⇒ Z × S).
  implicit val e4 = extractorFunc[Z, Either[Z, (R, R => (Z, S))], Q]    // Q × (Q ⇒ Z + R × (R ⇒ Z × S)).
  implicit val e5 = extractorPair[Z, Z, P]                                // Z × P.
  extractorEither[Z, (Z,P), (Q, Q => Either[Z, (R, R => (Z, S))])]  // Extractor for type K.
}
```

Construction	Type signature to implement	Results
The Unit type, or other fixed type C	Extractor^1 or Extractor^C	Extractor^Z
Product of extractor type A and any B	$\text{Extractor}^A \rightarrow \text{Extractor}^{A \times B}$	one possibility
Co-product of extractor types A and B	$\text{Extractor}^A \times \text{Extractor}^B \rightarrow \text{Extractor}^{A+B}$	one possibility
Function from or to another type C	$\text{Extractor}^A \rightarrow \text{Extractor}^{A \rightarrow C}$ or $\text{Extractor}^{C \rightarrow A}$	$\text{Extractor}^{C \times (C \rightarrow A)}$
Recursive type	$\text{Extractor}^A \rightarrow \text{Extractor}^{S^A}$ where $T \triangleq S^T$	Extractor^T

Table 8.1: Type constructions producing the `Extractor` typeclass.

The code computes each implicit value e_1, e_2, \dots , using constructions that require previously computed values to be present as implicits.

The `Extractor` typeclass is often used with $Z = \text{String}$ as a way to “print” values of different types, and it is then called `Show`. When $Z = \text{Array[Byte]}$, the typeclass is often called a “serializer”.

8.3.2 Equality comparison — the `Eq` typeclass

In Scala, the built-in operation `==` is not type-safe because the code `x == y` will compile regardless of the types of `x` and `y`. We can replace `==` by a new operation `==` constrained to a typeclass called `Eq`, ensuring that types can be meaningfully compared for equality. The equality comparison for values of type A is a function of type $A \times A \rightarrow 2$ (where 2 denotes the `Boolean` type). Typeclass instances of `Eq` need to wrap a function of that type. We also define `==` as an extension method:

```
final case class Eq[A](equal: (A, A) => Boolean)
object Eq {
  implicit class EqOps[A: Eq](a: A) {
    def ===(b: A): Boolean = implicitly[Eq[A]].equal(a, b)
  } // Use type-specific comparisons to define some typeclass instances here:
  implicit val eqInt: Eq[Int] = Eq[Int](_ == _)
  implicit val eqString: Eq[String] = Eq[String](_ == _)
}
```

If we define an `Eq` typeclass instance for all types via the built-in `==` operation, we will find that this operation does not work as expected when comparing values of a function type, e.g., `Int => Int`:

```
import Eq._
implicit def eqTypeA[A] = Eq[A](_ == _)

scala> ((n: Int) => n) === ((n: Int) => n)
res1: Boolean = false
```

Our expectations about equality comparisons are formalized via the laws of identity, symmetry, reflexivity, and transitivity (see also a discussion on page 183 in Section 6.1.6). The **reflexivity** law states that $x = x$ for any x ; so the comparison $x == x$ should always return `true`. The example shown above violates that law when we choose $x \triangleq n: \text{Int} \rightarrow n$.

Let us perform structural analysis for the `Eq` typeclass, defining $\text{Eq}^A \triangleq A \times A \rightarrow 2$. The results (see Table 8.2 below) will show which types can be usefully compared for equality.

Fixed types All primitive types have `Eq` instances that use type-specific equality comparisons.

Products If A and B have equality comparisons, we can compare pairs of type $A \times B$ by comparing each part of the pair separately:

```
def eqPair[A: Eq, B: Eq] = Eq[(A, B)]{ case ((a1, b1), (a2, b2)) => a1 === a2 && b1 === b2 }
```

It is easy to check that the identity, symmetry, reflexivity, and transitivity laws hold for the new comparison operation if they hold for comparisons of A and B separately.

It is important that the code for this construction assumes that both types A and B have lawful `Eq` instances and performs the comparisons $a_1 \equiv a_2$ and $b_1 \equiv b_2$. If the above code performed only, say, the comparison $a_1 \equiv a_2$, the resulting comparison operation would have violated the **identity law** of the equality operation (if $x = y$ then $f(x) = f(y)$ for any function f): we would have pairs such as $a \times b_1$ and $a \times b_2$ that would be “equal” according to this definition, and yet many functions $f: A \times B \rightarrow C$ exist such that $f(a \times b_1) \neq f(a \times b_2)$.

Co-products If A and B have equality comparisons, we can compare values of type $A + B$ while ensuring that a value of type $A + \emptyset$ is never equal to a value of type $\emptyset + B$:

```
def eqEither[A: Eq, B: Eq] = Eq[Either[A, B]] {
  case (Left(a1), Left(a2))  => a1 === a2    // Compare a1 + 0 and a2 + 0.
  case (Right(b1), Right(b2)) => b1 === b2    // Compare 0 + b1 and 0 + b2.
  case _                      => false      // a + 0 is never equal to 0 + b.
}
```

The laws hold for the new operation because the code defines values of type `Either[A, B]` as equal only when the types are both `Left` or both `Right`. If the comparisons of types A and B satisfy the laws separately, the laws for $A + B$ will be satisfied separately for values of type `Left` and of type `Right`.

Defining the comparison operation in any other way (e.g., setting $a_1 + \emptyset \neq a_2 + \emptyset$ for all a_1^A and a_2^A , or $\emptyset + b_1 = \emptyset + b_2$ for all b_1^B and b_2^B) would have violated the reflexivity law or the identity law.

Since the `Eq` typeclass has both the product and the co-product constructions, any polynomial type expression made of primitive types will also have an `Eq` typeclass instance.

Functions If A has an `Eq` instance, can we create an `Eq` instance for $R \rightarrow A$ where R is some other type? This would be possible with a function of type

$$\forall A. (A \times A \rightarrow 2) \rightarrow (R \rightarrow A) \times (R \rightarrow A) \rightarrow 2 \quad . \quad (8.6)$$

Here we assume that the type R is an arbitrary chosen type, so no values of type R can be computed from scratch. (This would not be the case when $R = \mathbb{1}$ or $R = \mathbb{2}$, say. But in those cases the type $R \rightarrow A$ can be simplified to a polynomial type, e.g., $\mathbb{1} \rightarrow A \cong A$ and $\mathbb{2} \rightarrow A \cong A \times A$, etc.) Without values of type R , we cannot compute any values of type A and cannot apply comparisons to them. Then the only possible implementations of the type signature (8.6) are constant functions returning `true` or `false`. However, the implementation that always returns `false` will violate the reflexivity law $x = x$. The implementation that always returns `true` is not useful.

We will be able to evaluate functions of type $R \rightarrow A$ if some chosen values of type R are available. Examples of that situation are types R of the form $R \cong \mathbb{1} + S$ for some type S ; we then always have a chosen value $r_1 \triangleq \mathbb{1} + 0^S$ of type R . What if we compare functions only via their values at r_1 ?

```
def equals[R, A: Eq](f: R => A, g: R => A): Boolean = f(r1) === g(r1) // Violates the identity law.
```

The above code defines a comparison operation that violates the identity law: there are many functions f and g that will give different results for arguments not equal to r_1 .

Another way to see the problem is to write the type equivalence

$$R \rightarrow A \cong \mathbb{1} + S \rightarrow A \cong A \times (S \rightarrow A) \quad ,$$

which reduces this case to the product construction we saw above. It follows that we need to have an `Eq` instance for $S \rightarrow A$ to define the equality operation for $R \rightarrow A$. If we have a chosen value of type S , e.g., if $S \cong \mathbb{1} + T$, we will again reduce the situation to the product construction with the function type $T \rightarrow A$. This process will end only if the type R has the form

$$R \cong \mathbb{1} + \mathbb{1} + \dots + \mathbb{1} \quad , \quad R \rightarrow A \cong \mathbb{1} + \mathbb{1} + \dots + \mathbb{1} \rightarrow A \cong A \times A \times \dots \times A \quad ,$$

i.e., if R has a *known* finite number of distinct values. Then we can write code that applies functions of type $R \rightarrow A$ to every possible value of the argument of type R and compares all the resulting values of type A . However, we also see that the type $R \rightarrow A$ is equivalent to a polynomial type, which is a product of a fixed number of A ’s. The product construction already covers that type.

So, we can compare functions $f_1:R \rightarrow A$ and $f_2:R \rightarrow A$ only if we are able to check whether $f_1(r) = f_2(r)$ for *every* possible value of type R . We cannot implement such a comparison for a general type R .

We conclude that functions of type $R \rightarrow A$ cannot have an `Eq` instance for a general type R . A similar argument shows that functions of type $A \rightarrow R$ also cannot have a useful `Eq` instance. The only exceptions are types $R \rightarrow A$ or $A \rightarrow R$ that are equivalent to some polynomial types.

Recursive types Since all polynomial type expressions preserve `Eq` instances, the same logic can be applied to recursive polynomial types. For instance, lists and trees with `Eq`-comparable values are also `Eq`-comparable. Let us prove this rigorously.

Consider a recursive polynomial type T defined using a polynomial functor S^\bullet ,

$$T \triangleq S^T \quad .$$

The functor S^\bullet may use other fixed types that have `Eq` instances. To construct the typeclass instance for T , we first implement a function `eqS` of type

$$\text{eqS} : \text{Eq}^A \rightarrow \text{Eq}^{S^A} \quad .$$

This function produces an `Eq` instance for S^A using `Eq` instances of A and of all other types that S^A depends on. The product and co-product constructions guarantee that it is always possible to implement this function for a polynomial functor S^\bullet . Then we define an `Eq` instance for T recursively:

$$\text{eqT} : \text{Eq}^T \quad , \quad \text{eqT} \triangleq \text{eqS}(\text{eqT}) \quad .$$

The recursive equation for `eqT` needs to be implemented as an expanded function,

$$\text{eqT} \triangleq t^T \times t^T \rightarrow \text{eqS}(\text{eqT})(t \times t) \quad ,$$

and then, as we have seen in the previous section, the recursion will terminate.

As an example, let us define an `Eq` instance for the type T defined by $T \triangleq \text{Int} + T + \text{Int} \times T \times T$.

```
type S[A] = Either[Either[Int, A], (Int, (A, A))]
final case class T(s: S[T]) // Recursive type equation T \triangleq \text{Int} + T + \text{Int} \times T \times T.
def eqS[A](implicit ti: Eq[A]): Eq[S[A]] = { // Function of type Eq[A] \Rightarrow Eq[S[A]].
  implicit val e1 = eqEither[Int, A] // Instance for Int + A.
  implicit val e2 = eqPair[A, A] // Instance for A \times A.
  implicit val e3 = eqPair[Int, (A, A)] // Instance for Int \times A \times A.
  eqEither[Either[Int, A], (Int, (A, A))] // Instance for Int + A + Int \times A \times A.
}
implicit def eqT: Eq[T] = Eq { case (T(s1), T(s2)) => eqS(eqT).equal(s1, s2) }
```

To test that the recursion terminates, define a value of type T and run a comparison:

```
val t = T(Left(Right(T(Left(Left(10)))))) // t : 0 + (10 : \text{Int} + 0 + 0) : T + 0
scala> t === t
res0: Boolean = true
```

Checking laws for recursive instances We use induction to verify that the laws of identity, symmetry, reflexivity, and transitivity will hold for the `Eq` instance just defined. The `Eq` instance for T was defined as a recursive function `eqT = eqS(eqT)`. We know that `eqS(x)` satisfies all the required laws if x satisfies them (“`eqS` preserves the laws”). Let us visualize what happens when the function `eqT` is applied to some values of type T . Are the laws satisfied in that case (e.g., is `eqT(t, t) == true` for all $t:T$)? The function `eqT` will call itself recursively whenever the body of `eqS` calls `eqT`. *Assuming* that all the recursive calls to `eqT` satisfy the laws, we find that the outer call to `eqT` also satisfies the laws because `eqS` preserves them. This corresponds to the inductive step.

Eventually the recursion will terminate; so *some* calls of `eqT(t1, t2)` with certain values t_1 and t_2 do not cause any more recursive self-calls to `eqT`. For *those* values t_1 and t_2 , the function `eqT` will satisfy the required laws with no additional assumptions, because `eqS(eqT)(t1, t2)`, which satisfies the laws, does not call `eqT`. This corresponds to the base case of an inductive proof.

So, the recursively defined `Eq` instance is lawful. This argument is general and works in all cases when a typeclass instance t is defined by recursion, $t \triangleq s(t)$, via a function s that preserves the laws.

Construction	Type signature to implement	Results
The <code>Unit</code> type, or other primitive type	$\text{Eq}^T \triangleq T \times T \rightarrow 2$	the method <code>==</code>
Product of <code>Eq</code> types A and B	$\text{Eq}^A \times \text{Eq}^B \rightarrow \text{Eq}^{A \times B}$	one possibility
Co-product of <code>Eq</code> types A and B	$\text{Eq}^A \times \text{Eq}^B \rightarrow \text{Eq}^{A+B}$	one possibility
Recursive types	$\text{Eq}^A \rightarrow \text{Eq}^{S^A}$ where $T \triangleq S^T$	Eq^T

Table 8.2: Type constructions producing the `Eq` typeclass.

Summary Instances of the `Eq` typeclass can be derived for any polynomial or recursive polynomial type expressions containing primitive types or type parameters constrained to be `Eq`-comparable. The derivation of the `Eq` instance is unambiguous and can be automated with libraries such as the `kittens`², `magnolia`³, or `scalaz-deriving`⁴.

8.3.3 Semigroups

A type T has an instance of `Semigroup` when an associative binary operation of type $T \times T \rightarrow T$ is available. We will now apply structural analysis to this typeclass. The results are shown in Table 8.3.

Fixed types Each of the primitive types (`Boolean`, `Int`, `Double`, `String`, etc.) has at least one well-known associative binary operation that can be used to define a semigroup instance. Booleans have the conjunction and the disjunction operations; numbers can be added or multiplied, or the maximum or the minimum number chosen; strings can be concatenated or chosen in the alphabetical order. Examples 8.2.5.4 and 8.2.5.5 show several implementations of such binary operations. The `Unit` type has a trivially defined binary operation, which is also associative (since it always returns the same value). The same is true for any fixed type that has a chosen “default” value: the binary operation that always returns the default value is associative (although not likely to be useful).

Type parameters A semigroup instance parametric in type T means a value of type $\forall T. T \times T \rightarrow T$. There are two implementations of this type signature: $a^T \times b^T \rightarrow a$ and $a^T \times b^T \rightarrow b$. Both provide an associative binary operation, as Example 8.2.5.5(a) shows. So, any type T can be made into a “trivial” semigroup in one of these two ways. (“Trivial” semigroups are occasionally useful.)

Products If types A and B are semigroups, the product $A \times B$ can be also given a `Semigroup` instance. To compute that instance means, in the general case, to implement a function with type

$$\text{semigroupPair} : \forall(A, B). \text{Semigroup}^A \times \text{Semigroup}^B \rightarrow \text{Semigroup}^{A \times B} .$$

Writing out the type expressions, we get the type signature

$$\text{semigroupPair} : \forall(A, B). (A \times A \rightarrow A) \times (B \times B \rightarrow B) \rightarrow (A \times B \times A \times B \rightarrow A \times B) .$$

While this type signature can be implemented in a number of ways, we look for code that preserves information, in hopes of satisfying the associativity law. The code should be a function of the form

$$\text{semigroupPair} \triangleq f^{A \times A \rightarrow A} \times g^{B \times B \rightarrow B} \rightarrow a_1^A \times b_1^B \times a_2^A \times b_2^B \rightarrow ???^A \times ???^B .$$

Since we are trying to define the new semigroup operation through the previously given operations f and g , it is natural to apply f and g to the given data a_1, a_2, b_1, b_2 and write

$$f^{A \times A \rightarrow A} \times g^{B \times B \rightarrow B} \rightarrow a_1^A \times b_1^B \times a_2^A \times b_2^B \rightarrow f(a_1, a_2) \times g(b_1, b_2) .$$

²<https://github.com/typelevel/kittens>

³<https://github.com/propensive/magnolia>

⁴<https://github.com/scalaz/scalaz-deriving>

This code defines a new binary operation $\oplus_{A \times B}$ via the previously given \oplus_A and \oplus_B as

$$(a_1^A \times b_1^B) \oplus_{A \times B} (a_2^A \times b_2^B) = (a_1 \oplus_A a_2) \times (b_1 \oplus_B b_2) \quad . \quad (8.7)$$

```
def semigroupPair[A: Semigroup, B: Semigroup] =  
  Semigroup[(A, B)]{ case ((a1, b1), (a2, b2)) => (a1 |+| a2, b1 |+| b2) }
```

This implementation satisfies the associativity law if the operations \oplus_A , \oplus_B already do, i.e., if the results of computing $a_1 \oplus_A a_2 \oplus_A a_3$ and $b_1 \oplus_B b_2 \oplus_B b_3$ do not depend on the order of parentheses:

$$\begin{aligned} ((a_1 \times b_1) \oplus_{A \times B} (a_2 \times b_2)) \oplus_{A \times B} (a_3 \times b_3) &= ((a_1 \oplus_A a_2) \times (b_1 \oplus_B b_2)) \oplus_{A \times B} (a_3 \times b_3) \\ &= (a_1 \oplus_A a_2 \oplus_A a_3) \times (b_1 \oplus_B b_2 \oplus_B b_3) \quad , \\ (a_1 \times b_1) \oplus_{A \times B} ((a_2 \times b_2) \oplus_{A \times B} (a_3 \times b_3)) &= (a_1 \times b_1) \oplus_{A \times B} ((a_2 \oplus_A a_3) \times (b_2 \oplus_B b_3)) \\ &= (a_1 \oplus_A a_2 \oplus_A a_3) \times (b_1 \oplus_B b_2 \oplus_B b_3) \quad . \end{aligned}$$

Co-products To compute a `semigroup` instance for the co-product $A + B$ of two semigroups, we need

$$\text{semigroupEither} : \forall(A, B). \text{Semigroup}^A \times \text{Semigroup}^B \rightarrow \text{Semigroup}^{A+B} \quad .$$

Writing out the type expressions, we get the type signature

$$\text{semigroupEither} : \forall(A, B). (A \times A \rightarrow A) \times (B \times B \rightarrow B) \rightarrow (A + B) \times (A + B) \rightarrow A + B \quad .$$

Begin by writing a function with a typed hole:

$$\text{semigroupEither} \triangleq f^{A \times A \rightarrow A} \times g^{B \times B \rightarrow B} \rightarrow c^{(A+B) \times (A+B) \rightarrow ???^{A+B}} \quad .$$

Transforming the type expression $(A + B) \times (A + B)$ into an equivalent disjunctive type,

$$(A + B) \times (A + B) \cong A \times A + A \times B + B \times A + B \times B \quad ,$$

we can continue to write the function's code in matrix notation,

$$f^{A \times A \rightarrow A} \times g^{B \times B \rightarrow B} \rightarrow \begin{array}{c|cc} & A & B \\ \hline A \times A & ???^{A \times A \rightarrow A} & ???^{A \times A \rightarrow B} \\ A \times B & ???^{A \times B \rightarrow A} & ???^{A \times B \rightarrow B} \\ B \times A & ???^{B \times A \rightarrow A} & ???^{B \times A \rightarrow B} \\ B \times B & ???^{B \times B \rightarrow A} & ???^{B \times B \rightarrow B} \end{array} \quad .$$

The matrix is 4×2 because the input type, $A \times A + A \times B + B \times A + B \times B$, is a disjunction with 4 parts, while the result type $A + B$ is a disjunction with 2 parts. In each row, we need to fill only one of the two typed holes because only one part of the disjunction $A + B$ can have a value.

To save space, we will omit the types in the matrices. The first and the last rows of the matrix must contain functions of types $A \times A \rightarrow A$ and $B \times B \rightarrow B$, and so it is natural to fill them with f and g :

$$f^{A \times A \rightarrow A} \times g^{B \times B \rightarrow B} \rightarrow \begin{array}{c|cc} & f^{A \times A \rightarrow A} & \emptyset \\ \hline ???^{A \times B \rightarrow A} & ???^{A \times B \rightarrow B} & \\ ???^{B \times A \rightarrow A} & ???^{B \times A \rightarrow B} & \\ \hline \emptyset & & g^{B \times B \rightarrow B} \end{array} \quad .$$

The remaining two rows can be filled in four different ways:

$$\begin{array}{c}
 f \times g \rightarrow \left| \begin{array}{cc} f:A \times A \rightarrow A & \emptyset \\ a:A \times b:B \rightarrow a & \emptyset \\ b:B \times a:A \rightarrow a & \emptyset \\ \emptyset & g:B \times B \rightarrow B \end{array} \right|, \quad f \times g \rightarrow \left| \begin{array}{cc} f & \emptyset \\ \emptyset & a:A \times b:B \rightarrow b \\ \emptyset & b:B \times a:A \rightarrow b \\ \emptyset & g \end{array} \right|, \\
 f \times g \rightarrow \left| \begin{array}{cc} f:A \times A \rightarrow A & \emptyset \\ \emptyset & a:A \times b:B \rightarrow b \\ b:B \times a:A \rightarrow a & \emptyset \\ \emptyset & g:B \times B \rightarrow B \end{array} \right|, \quad f \times g \rightarrow \left| \begin{array}{cc} f & \emptyset \\ a:A \times b:B \rightarrow a & \emptyset \\ \emptyset & b:B \times a:A \rightarrow b \\ \emptyset & g \end{array} \right|
 \end{array}.$$

The Scala code corresponding to the four possible definitions of \oplus is

```

def semigroupEither1[A: Semigroup, B: Semigroup] = Semigroup[Either[A, B]] {
  case (Left(a1), Left(a2)) => a1 |+| a2 // Here a1 |+| a2 is a1 ⊕_A a2.
  case (Right(b1), Right(b2)) => b1 |+| b2 // Here b1 |+| b2 is b1 ⊕_B b2.
  case (Left(a), Right(b)) => a // "Take A" - discard all data of type B.
  case (Right(b), Left(a)) => a
}

def semigroupEither2[A: Semigroup, B: Semigroup] = Semigroup[Either[A, B]] {
  case (Left(a1), Left(a2)) => a1 |+| a2
  case (Right(b1), Right(b2)) => b1 |+| b2
  case (Left(a), Right(b)) => b // "Take B" - discard all data of type A.
  case (Right(b), Left(a)) => b
}

def semigroupEither3[A: Semigroup, B: Semigroup] = Semigroup[Either[A, B]] {
  case (Left(a1), Left(a2)) => a1 |+| a2
  case (Right(b1), Right(b2)) => b1 |+| b2
  case (Left(a), Right(b)) => a // "Take first" - discard y in x ⊕ y.
  case (Right(b), Left(a)) => b
}

def semigroupEither4[A: Semigroup, B: Semigroup] = Semigroup[Either[A, B]] {
  case (Left(a1), Left(a2)) => a1 |+| a2
  case (Right(b1), Right(b2)) => b1 |+| b2
  case (Left(a), Right(b)) => b // "Take last" - discard x in x ⊕ y.
  case (Right(b), Left(a)) => a
}

```

The four different choices of the binary operation can be described as:

1. ("Take A") Discard all data of type B : $(a^A + \emptyset) \oplus (\emptyset + b^B) = a$ and $(\emptyset + b^B) \oplus (a^A + \emptyset) = a$.
2. ("Take B ") Discard all data of type A : $(a^A + \emptyset) \oplus (\emptyset + b^B) = b$ and $(\emptyset + b^B) \oplus (a^A + \emptyset) = b$.
3. ("Take first") $x \oplus y$ discards y : $(a^A + \emptyset) \oplus (\emptyset + b^B) = a$ and $(\emptyset + b^B) \oplus (a^A + \emptyset) = b$.
4. ("Take last") $x \oplus y$ discards x : $(a^A + \emptyset) \oplus (\emptyset + b^B) = b$ and $(\emptyset + b^B) \oplus (a^A + \emptyset) = a$.

Does the semigroup law (8.1) hold for the new typeclass instance with any of these implementations?

It turns out that *all four* implementations are lawful. To verify the associativity law, we need to show that values such as `Left(x) |+| Left(y) |+| Right(z)` or `Left(x) |+| Right(y) |+| Right(z)` do not depend on the order of inserted parentheses. Expressions of the form $x \oplus y \oplus z$ can have 8 possible combinations of `Left` and `Right` types. Each of them needs to be checked against each of the 4 implementations of \oplus . Instead of doing 32 separate derivations, we will reason case by case.

First consider the case when all three values are of type `Left(x)`. In all four implementations, the binary operation reduces to the binary operation of the semigroup A , which is associative since we assume that A is a lawful semigroup:

```
Left(x) |+| Left(y) |+| Left(z) == Left(x |+| y |+| z) == Left(x) |+| (Left(y) |+| Left(z))
```

The same argument applies to three values of type `Right`. It remains to consider the "mixed" cases.

In the first implementation (“take A”), we discard all data of type `Right` from the expression $x \oplus y \oplus z$, keeping only data of type `Left`. It is clear that discarding data of type `Right` will yield the same result regardless of the order of parentheses. If more than one item of type `Left` remains, the data is aggregated with the operation \oplus_A . So, the results do not depend on the order of parentheses:

```
Right(x) |+| Right(y) |+| Left(z) == Left(z) == Right(x) |+| (Right(y) |+| Left(z))  
Left(x) |+| Right(y) |+| Left(z) == Left(x |+| z) == Left(x) |+| (Right(y) |+| Left(z))
```

A similar argument shows that the second implementation (“take B”) is also associative.

The implementation “take first” will select the first value whenever the types are mixed. Therefore

```
Left(x) |+| Left(y) |+| Right(z) == Left(x |+| y) == Left(x) |+| (Left(y) |+| Right(z))  
Left(x) |+| Right(y) |+| p == Left(x) == Left(x) |+| (Right(y) |+| p) // Regardless of p.  
Right(x) |+| Left(y) |+| p == Right(x) == Right(x) |+| (Left(y) |+| p) // Regardless of p.  
Right(x) |+| Right(y) |+| Left(z) == Right(x |+| y) == Right(x) |+| (Right(y) |+| Left(z))
```

The results are independent of the parentheses. The same applies to the “take last” implementation.

Functions If A is a semigroup and E is any fixed type, are the types $A \rightarrow E$ and/or $E \rightarrow A$ semigroups? To create a `Semigroup` instance for $E \rightarrow A$ means to implement the type signature

$$\text{Semigroup}^A \rightarrow \text{Semigroup}^{E \rightarrow A} = (A \times A \rightarrow A) \rightarrow (E \rightarrow A) \times (E \rightarrow A) \rightarrow E \rightarrow A .$$

An implementation that preserves information is

$$\text{semigroupFunc} \triangleq f^{A \times A \rightarrow A} \rightarrow g_1^{E \rightarrow A} \times g_2^{E \rightarrow A} \rightarrow e^E \rightarrow f(g_1(e), g_2(e)) .$$

This defines the new \oplus operation by $g_1 \oplus g_2 \triangleq e \rightarrow g_1(e) \oplus_A g_2(e)$.

```
def semigroupFunc[E, A: Semigroup] = Semigroup[E => A] { case (g1, g2) => e => g1(e) |+| g2(e) }
```

In the pipe notation, $e \triangleright (f \oplus g) = f(e) \oplus_A g(e)$. The associativity law holds for this operation:

$$\text{left-hand side : } e \triangleright ((f \oplus g) \oplus h) = (e \triangleright (f \oplus g)) \oplus_A h(e) = f(e) \oplus_A g(e) \oplus_A h(e) .$$

$$\text{right-hand side : } e \triangleright (f \oplus (g \oplus h)) = f(e) \oplus_A (e \triangleright (g \oplus h)) = f(e) \oplus_A g(e) \oplus_A h(e) .$$

The type $A \rightarrow E$ only allows semigroup operations that discard the left or the right element: the type signature $f^{A \times A \rightarrow A} \rightarrow h_1^{A \rightarrow E} \times h_2^{A \rightarrow E} \rightarrow ???^{A \rightarrow E}$ can be implemented only by discarding f and one of h_1 or h_2 . Either choice makes $A \rightarrow E$ into a trivial semigroup.

We have seen constructions that create new semigroups via products, co-products, and functions. Thus, any exponential-polynomial type expression built up from primitive types and/or existing semigroups is again a semigroup.

Recursive types A type T defined by a recursive type equation $T \triangleq S^T$ can have a semigroup instance when S^A is any exponential-polynomial type expression built up from primitive types, products, co-products, and the type parameter A . The known semigroup constructions guarantee that a typeclass instance `Semigroup[S[A]]` can be created out of `Semigroup[A]`. This gives us a function

$$\text{semigroupS} : \text{Semigroup}^A \rightarrow \text{Semigroup}^{S^A} .$$

Then a semigroup instance for T is defined recursively as

$$\text{semigroupT} \triangleq \text{semigroupS}(\text{semigroupT}) .$$

The recursive definition will terminate as long as we implement it in code as an expanded function. The associativity law holds for the semigroup T by induction, as shown at the end of Section 8.3.2.

Summary Any type expression whatsoever can have a `Semigroup` instance. Since the instances have several inequivalent implementations, automatic derivation of `Semigroup` instances is not often useful.

Construction	Type signature to implement	Results
The <code>Unit</code> type, or other fixed type C	Semigroup^C	two trivial semigroups
Product of semigroups A and B	$\text{Semigroup}^A \times \text{Semigroup}^B \rightarrow \text{Semigroup}^{A \times B}$	one possibility
Co-product of semigroups A and B	$\text{Semigroup}^A \times \text{Semigroup}^B \rightarrow \text{Semigroup}^{A+B}$	four possibilities
Function from another type E	$\text{Semigroup}^A \rightarrow \text{Semigroup}^{E \rightarrow A}$	one possibility
Recursive type	$\text{Semigroup}^A \rightarrow \text{Semigroup}^{S^A}$ where $T \triangleq S^T$	Semigroup^T

Table 8.3: Type constructions producing the `Semigroup` typeclass.

8.3.4 Monoids

Since a monoid is a semigroup with a default value, a `Monoid` instance is a value of type

$$\text{Monoid}^A \triangleq (A \times A \rightarrow A) \times A \quad .$$

For the binary operation $A \times A \rightarrow A$, we can re-use the results of structural analysis for semigroups. Additionally, we will need to verify that the default value satisfies monoid's identity laws. The results are shown in Table 8.4.

Fixed types Each of the primitive types (`Boolean`, `Int`, `Double`, `String`, etc.) has a well-defined monoidal operation (addition or multiplication for numbers, concatenation for strings, and so on). The unit type, `Unit`, also has a monoid instance where all methods return the unit value `1`.

Type parameters This construction works for semigroups but *not* for monoids: the “trivial” semigroup operations $x \oplus y = x$ and $x \oplus y = y$ are not compatible with monoid's identity laws. (e.g., with the definition $x \oplus y = x$, no default value e could possibly satisfy the left identity law $e \oplus y = y$ because $e \oplus y = e$ for all y).

Products For two monoids A and B , a monoid instance for the product $A \times B$ is computed by

$$\text{monoidPair} : \forall(A, B). \text{Monoid}^A \times \text{Monoid}^B \rightarrow \text{Monoid}^{A \times B} \quad .$$

The empty value for the monoid $A \times B$ is $e_{A \times B} \triangleq e_A \times e_B$, the pair of empty values from the monoids A and B . The new binary operation is defined by Eq. (8.7) as in the pair semigroup construction. We can now verify the new monoid's identity laws, assuming that they hold for the monoids A and B :

$$\begin{aligned} (a_1 \times b_1) \oplus (e_A \times e_B) &= (a_1 \oplus_A e_A) \times (b_1 \oplus_B e_B) = a_1 \times b_1 \quad , \\ (e_A \times e_B) \oplus (a_2 \times b_2) &= (e_A \oplus_A a_2) \times (e_B \oplus_B b_2) = a_2 \times b_2 \quad . \end{aligned}$$

An implementation in Scala is

```
def monoidPair [A: Monoid, B: Monoid]: Monoid[(A, B)] = Monoid[(A, B)](
  { case ((a1, b1), (a2, b2)) => (a1 |+| a2, b1 |+| b2) },
  (implicitly[Monoid[A]].empty, implicitly[Monoid[B]].empty)
)
```

Co-products For two monoids A and B , how can we implement a `Monoid` instance for $A + B$? We have seen four versions of the semigroup operation \oplus for the type $A + B$. Independently of those, we need to define the empty element e_{A+B} , which must have type $A + B$. There are two possibilities:

$$e_{A+B} \triangleq e_A + \emptyset^B \quad \text{or} \quad e_{A+B} \triangleq \emptyset^A + e_B \quad .$$

It remains to see which of the eight combinations will satisfy the monoid identity laws,

$$\begin{aligned} (a + \emptyset) \oplus e_{A+B} &= a + \emptyset \quad , \quad e_{A+B} \oplus (a + \emptyset) = a + \emptyset \quad , \\ (\emptyset + b) \oplus e_{A+B} &= \emptyset + b \quad , \quad e_{A+B} \oplus (\emptyset + b) = \emptyset + b \quad . \end{aligned}$$

First choose $e_{A+B} = e_A + \mathbb{0}$; the reasoning for the other case will be quite similar. The first line above,

$$(a + \mathbb{0}) \oplus (e_A + \mathbb{0}) = a + \mathbb{0} \quad , \quad (e_A + \mathbb{0}) \oplus (a + \mathbb{0}) = a + \mathbb{0} \quad ,$$

will hold because all four versions of the operation \oplus will reduce to \oplus_A on values of type $A + \mathbb{0}$. The second line, however, is compatible only with one version of the \oplus operation, namely with “take B”:

$$(a^A + \mathbb{0}) \oplus (\mathbb{0} + b^B) = b \quad \text{and} \quad (\mathbb{0} + b^B) \oplus (a^A + \mathbb{0}) = b \quad .$$

So, the co-product construction must choose one of the monoids, say B , as “preferred”. The code is

```
def monoidEitherPreferB[A: Monoid, B: Monoid] = Monoid[Either[A, B]]( {
  case (Left(a1), Left(a2))    => Left(a1 |+| a2)
  case (Left(a), Right(b))     => Right(b) // "Take B".
  case (Right(b), Left(a))    => Right(b)
  case (Right(b1), Right(b2))  => Right(b1 |+| b2)
}, Left(implicitly[Monoid[A]].empty) )
```

Similarly, the choice $e_{A+B} \triangleq \mathbb{0} + e_B$ forces us to choose the version “take A ” of the \oplus operation.

Functions The semigroup construction for function types works also for monoids. Exercise 8.4.2.3 will show that the function type $R \rightarrow A$ is a lawful monoid for any type R and any monoid A .

Additionally, the function type $R \rightarrow R$ is a monoid for any type R (even if R is not a monoid). The operation \oplus and the empty value are defined as $f:R \rightarrow R \oplus g:R \rightarrow R \triangleq f \circ g$ and $e_{R \rightarrow R} \triangleq \text{id}^R$. The code is

```
def monoidFunc1[R]: Monoid[R => R] = Monoid( (f, g) => f andThen g, identity )
```

The monoidal operation \oplus is the forward function composition $f \circ g$, so the monoid laws for this operation are the identity and the associativity laws of function composition (see Section 4.2.2).

We could also define the monoidal operation as the backward function composition, $f \circ g$.

Recursive types Can we define a `Monoid` instance for a type T defined by $T \triangleq S^T$, where S^\bullet is some type constructor? As we have seen, products, co-products, and function type constructions preserve monoids. For any type built up via these constructions from monoids, a `Monoid` instance can be derived. These constructions cover all exponential-polynomial types. So, let us consider an exponential-polynomial type constructor S^A that contains a type parameter A , primitive types, and other known monoid types. For such type constructors S^\bullet , we will always be able to implement a function `monoids` that derives a `Monoid` instance for S^A from a monoid instance for A :

$$\text{monoidS} : \text{Monoid}^A \rightarrow \text{Monoid}^{S^A} \quad .$$

A monoid instance for T is then defined recursively by

$$\text{monoidT} \triangleq \text{monoidS}(\text{monoidT}) \quad .$$

As we saw before, the code for this definition will terminate only if we implement it as a recursive function. However, the type Monoid^A is not a function type: it is a pair $(A \times A \rightarrow A) \times A$. To obtain a working implementation of `monoidT`, we need to rewrite that type into an equivalent function type,

$$\text{Monoid}^A = (A \times A \rightarrow A) \times A \cong (A \times A \rightarrow A) \times (\mathbb{1} \rightarrow A) \cong (\mathbb{1} + A \times A \rightarrow A) \quad ,$$

where we used the known type equivalences $A \cong \mathbb{1} \rightarrow A$ and $(A \rightarrow C) \times (B \rightarrow C) \cong A + B \rightarrow C$.

```
final case class Monoid[A](methods: Option[(A, A)] => A)
```

With this new definition of the `Monoid` typeclass (and with the appropriate changes to the code of `monoidPair`, `monoidEitherPreferB`, and `monoidFunc`), we can now implement the recursive construction.

To illustrate how that works, consider the exponential-polynomial type constructor S^\bullet defined as

$$S^A \triangleq (\text{Int} + A) \times \text{Int} + \text{String} \times (A \rightarrow (A \rightarrow \text{Int}) \rightarrow A) \quad .$$

Construction	Type signature to implement	Results
The Unit type, or primitive types	Monoid ¹ , Monoid ^{Int} , etc.	custom code
Product of monoids A and B	Monoid $^A \times$ Monoid $^B \rightarrow$ Monoid $^{A \times B}$	one possibility
Co-product of monoid A and semigroup B	Monoid $^A \times$ Semigroup $^B \rightarrow$ Monoid $^{A+B}$	one possibility
Function from another type E	Monoid $^A \rightarrow$ Monoid $^{E \rightarrow A}$	Monoid $^{E \rightarrow A}$
Recursive type	Monoid $^A \rightarrow$ Monoid $^{S^A}$ where $T \triangleq S^T$	Monoid T

Table 8.4: Type constructions producing the `Monoid` typeclass.

```
type S[A] = Either[(Either[Int, A], Int), (String, A => (A => Int) => A)]
```

It is clear that S^A is built up from type constructions that preserve monoids at each step. So, we expect that the recursive type $T \triangleq S^T$ is a monoid. We first implement the function `monoidS`,

```
def monoidS[A](implicit ti: Monoid[A]): Monoid[S[A]] = {
  implicit val m0 = monoidEitherPreferB[Int, A]
  implicit val m1 = monoidPair[Either[Int, A], Int]
  implicit val m2 = monoidFunc[A, A => Int]
  implicit val m3 = monoidFunc[(A => Int) => A, A]
  implicit val m4 = monoidPair[String, A => (A => Int) => A]
  monoidEitherPreferB[(Either[Int, A], Int), (String, A => (A => Int) => A)]
}
```

We can now use this function to define the recursive type T and a `Monoid` instance for it,

```
final case class T(s: S[T])
def monoidT: Monoid[T] = Monoid[T] {
  case None          => T(monoidS[T](monoidT).methods(None))
  case Some((t1, t2)) => T(monoidS[T](monoidT).methods(Some(t1.s, t2.s)))
}
```

To test this code, create a value of type T and perform a computation:

```
val t = T(Right(("a", t => f => T(Left((Left(f(t)), 10))))))

scala> t |+| t
res0: T = T(Right((aa,<function1>)))
```

Another way of implementing the recursive construction is to write the `Monoid` typeclass using a `trait`. Although the code is longer, it is easier to read. The recursive instance is implemented by

```
def monoidT: Monoid[T] = new Monoid[T] {
  def empty: T = T(monoidS[T](monoidT).empty) // This must be a 'def empty', not a 'val empty'.
  def combine: (T, T) => T = (x, y) => T(monoidS[T](monoidT).combine(x.s, y.s))
}
```

The recursive definition of `monoidT` terminates because the methods of the `trait` are declared as `def` (not as `val`). The code of `monoids` remains the same; we need to rewrite `monoidPair`, `monoidEitherPreferB`, and `monoidFunc` to accommodate the new definition of the `Monoid` typeclass. The full code is shown in Figure 8.1.

The monoid laws hold for the recursive instances by induction (see Section 8.3.2).

Summary A `Monoid` instance can be implemented, in at least one way, for *any* exponential-polynomial type expression (including recursive types) built from primitive types and other monoids.

```

trait Monoid[T] {
  def empty: T
  def combine: (T, T) => T
}
implicit val monoidInt: Monoid[Int] = new Monoid[Int] {
  def empty: Int = 0
  def combine: (Int, Int) => Int = _ + _
}
implicit val monoidString: Monoid[String] = new Monoid[String] {
  def empty: String = ""
  def combine: (String, String) => String = _ + _
}
implicit class MonoidOps[T: Monoid](t: T) {
  def |+|(a: T): T = implicitly[Monoid[T]].combine(t, a)
}
def monoidPair[A: Monoid, B: Monoid]: Monoid[(A, B)] = new Monoid[(A, B)] {
  def empty: (A, B) = (implicitly[Monoid[A]].empty, implicitly[Monoid[B]].empty)
  def combine: ((A, B), (A, B)) => (A, B) = {
    case ((a1, b1), (a2, b2)) => (a1 |+| a2, b1 |+| b2)
  }
}
def monoidEitherPreferB[A: Monoid, B: Monoid] = new Monoid[Either[A, B]] {
  def empty: Either[A, B] = Left(implicitly[Monoid[A]].empty)
  def combine: (Either[A, B], Either[A, B]) => Either[A, B] = {
    case (Left(a1), Left(a2)) => Left(a1 |+| a2)
    case (Left(a), Right(b)) => Right(b) // "Take B".
    case (Right(b), Left(a)) => Right(b)
    case (Right(b1), Right(b2)) => Right(b1 |+| b2)
  }
}
def monoidFunc[A: Monoid, E] = new Monoid[E => A] {
  def empty: E => A = _ => implicitly[Monoid[A]].empty
  def combine: (E => A, E => A) => E => A = {
    case (f, g) => e => f(e) |+| g(e)
  }
}
// This type constructor will be used below to define a recursive type T.
type S[A] = Either[(Either[Int, A], Int), (String, A => (A => Int) => A)]
// If we have a Monoid instance for A, we can get a Monoid instance for S[A].
def monoidsS[A](implicit ti: Monoid[A]): Monoid[S[A]] = {
  implicit val m0 = monoidEitherPreferB[Int, A]
  implicit val m1 = monoidPair[Either[Int, A], Int]
  implicit val m2 = monoidFunc[A, A => Int]
  implicit val m3 = monoidFunc[(A => Int) => A, A]
  implicit val m4 = monoidPair[String, A => (A => Int) => A]
  monoidEitherPreferB[(Either[Int, A], Int), (String, A => (A => Int) => A)]
}
// Define a recursive type T and a Monoid instance for it.
final case class T(s: S[T])
implicit def monoidT: Monoid[T] = new Monoid[T] {
  def empty: T = T(monoidsS[T](monoidT).empty) // Here, 'val empty' will cause a StackOverflowError.
  def combine: (T, T) => T = (x, y) => T(monoidsS[T](monoidT).combine(x.s, y.s))
}
val t = T(Right(("abc", t => f => T(Left((Left(f(t)), 10))))))
val e = implicitly[Monoid[T]].empty

scala> t |+| t |+| e    // Expect to see the string "abcabc".
res0: T = T(Right((abcabc,<function1>)))

```

Figure 8.1: Implementing a recursive instance of the `Monoid` typeclass via a `trait`.

8.3.5 Pointed functors: motivation and laws

Section 8.2.6 showed how to implement typeclasses for type *constructors*, e.g., the `Functor` typeclass. Typeclass instances in such cases often contain a nested type quantifier such as $\forall A. (...)$, so the implementation needs to use Scala's `trait` with `def` methods inside. We will now look at some examples of typeclasses that add further methods to `Functor`. Chapter 6 performed a structural analysis of functors, which we will extend to the new typeclasses.

The first typeclass is a “pointed” functor. A functor type F^T represents, in a generalized sense, “wrapped” values of type T . A frequently used operation is to create a “wrapped” value of type F^T out of a single given value of type T . This operation, usually called `pure` in Scala libraries, is implemented as a function with a type signature

```
def pure[A]: A => F[A]
```

The code notation for this function is pu_F , and the type signature is written as $\text{pu}_F : \forall A. A \rightarrow F^A$.

Some examples of pointed functors in Scala are `Option`, `List`, `Try`, and `Future`. Each of these type constructors has a method that “wraps” a given single value:

```
val x: Option[Int] = Some(10)           // A non-empty option that holds a value.
val y: List[String] = List("abc")       // A list that holds a single value.
val z: Try[Int] = Success(200)          // A value computed without errors.
val f: Future[String] = Future.successful("OK") // A 'Future' value that is already available.
```

As we can see, “wrapping a single value” means a different thing for each of the type constructors. Although the relevant methods of these type constructors are not called “pure”, we can create a PTVF `pure[F]` that would be defined only for F that can “wrap” a single value. Such type constructors F are called “pointed”. We may define the typeclass `Pointed` via this code:

```
trait Pointed[F[_]] { def pure[A]: A => F[A] }
```

Now we can implement instances of the `Pointed` typeclass for some functors:

```
implicit val pointedOption = new Pointed[Option] { def pure[A]: A => Option[A] = x => Some(x) }
implicit val pointedList = new Pointed[List] { def pure[A]: A => List[A] = x => List(x) }
implicit val pointedTry = new Pointed[Try] { def pure[A]: A => Try[A] = x => Success(x) }
```

The PTVF `pure` can be defined and used like this,

```
def pure[F[_]: Pointed, A](x: A): F[A] =
  implicitly[Pointed[F]].pure(x)

scala> pure[Option, Int](123)
res0: Option[Int] = Some(123)
```

When a pointed type constructor F is a functor, we may use both the functor's `map` method and the `pure` method. Do these two methods need to be compatible in some way? If we “wrap” a value 123 in a `List` and then apply $.map(x \Rightarrow x + 1)$, we expect to obtain a list containing 124;

any other result would break our intuition about “wrapping”. We can generalize this situation to an arbitrary value $x:A$ wrapped using `pure` and a function $f:A \rightarrow B$ applied to the wrapped value via `map`:

```
pure(x).map(f) // pu_F(x) ▷ f↑F
```

We expect that the result should be the same as a wrapped $f(x)$. This expectation can be formulated as a law, called the **naturality law** of `pure`; it must hold for any $f:A \rightarrow B$:

```
pure(x).map(f) == pure(f(x))
```

$$\text{pu}_F(x) \triangleright f^{\uparrow F} = \text{pu}_F(f(x)) .$$

In the \triangleright -notation, this law is $x \triangleright \text{pu}_F \triangleright f^{\uparrow F} = x \triangleright f \triangleright \text{pu}_F$ or equivalently $x \triangleright \text{pu}_F \circ f^{\uparrow F} = x \triangleright f \circ \text{pu}_F$. Since both sides of the law are functions applied to an arbitrary value $x:A$, we can omit x and write

$$\begin{array}{ccc} A & \xrightarrow{\text{pu}_F} & F^A \\ f \downarrow & & \downarrow f^{\uparrow F} \\ B & \xrightarrow{\text{pu}_F} & F^B \end{array}$$

$$\text{pu}_F \circ f^{\uparrow F} = f \circ \text{pu}_F . \quad (8.8)$$

This motivates the following definition: A functor F^\bullet is **pointed** if there exists a fully parametric function $\text{pu}_F : \forall A. A \rightarrow F^A$ satisfying the naturality law (8.8) for any function $f:A \rightarrow B$.

It turns out that we can avoid checking the naturality law of pointed functors if we use a trick: reduce `pure` to a simpler but equivalent form for which the law is satisfied automatically.

Both sides of the naturality law (8.8) are functions of type $A \rightarrow F^B$. The trick is to choose $A = \mathbb{1}$ (the `Unit` type) and $f: \mathbb{1} \rightarrow B \triangleq (_ \rightarrow b)$, a constant function returning some fixed value b^B . Both sides of the naturality law may then be applied to the unit value 1 and must evaluate to the same result:

$$1 \triangleright \text{pu}_F \triangleright (_ \rightarrow b)^{\uparrow F} = 1 \triangleright f \triangleright \text{pu}_F \quad .$$

Since $1 \triangleright f = f(1) = b$, we find

$$\text{pu}_F(1) \triangleright (_ \rightarrow b)^{\uparrow F} = \text{pu}_F(b) \quad . \quad (8.9)$$

The naturality law (8.8) applies to all types A, B and to any function $f: A \rightarrow B$. Thus, Eq. (8.9) must apply to an arbitrary value b^B for any type B . That formula expresses the function pu_F through one value $\text{pu}_F(1)$ of type $F^{\mathbb{1}}$. This value can be viewed as a “wrapped unit” value.

To perform the same derivation in Scala syntax, we may write

```
val one: Unit = ()  
val f: Unit => B = { _ => b }  
pure(one).map(f) == pure(f(one)) == pure(b) // Because f(one) == b.
```

It follows that if pu_F satisfies the naturality law then a single “wrapped unit” value of that function, $\text{pu}_F(1)$, is sufficient to recover the entire function $\text{pu}_F(b)$ by using the code

$$\text{pu}_F(b) \triangleq \text{pu}_F(1) \triangleright (_ \rightarrow b)^{\uparrow F}$$

So, given just a “wrapped unit” value (denoted wu_F) of type $F^{\mathbb{1}}$, we can define a new function pu_F :

```
def pure[A](x: A): F[A] = wu.map { _ => x }  $\text{pu}_F^{A \rightarrow F^A} \triangleq x^A \rightarrow \text{wu}_F \triangleright (\_ \rightarrow x)^{\uparrow F} \quad . \quad (8.10)$ 
```

Does this function satisfy the naturality law with respect to an arbitrary $f: A \rightarrow B$? It does:

$$\begin{aligned} \text{expect to equal } x \triangleright f \circ \text{pu}_F : & x \triangleright \text{pu}_F \circ f^{\uparrow F} \\ \text{definition of } \text{pu}_F : & = \text{wu}_F \triangleright (_ \rightarrow x)^{\uparrow F} \circ f^{\uparrow F} \\ \text{functor composition law of } F : & = \text{wu}_F \triangleright ((_ \rightarrow x) \circ f)^{\uparrow F} \\ \text{compute function composition} : & = \text{wu}_F \triangleright (_ \rightarrow f(x))^{\uparrow F} \\ \text{definition of } \text{pu}_F : & = \text{pu}_F(f(x)) \\ \triangleright\text{-notation} : & = x \triangleright f \triangleright \text{pu}_F = x \triangleright f \circ \text{pu}_F \quad . \end{aligned}$$

Applied to the unit value, this new function gives $\text{pu}_F(1) = \text{wu}_F$ because

$$\begin{aligned} \triangleright\text{-notation} : \text{pu}_F(1) &= 1 \triangleright \text{pu}_F \\ \text{definition of } \text{pu}_F \text{ via } \text{wu}_F : & = \text{wu}_F \triangleright (_ \rightarrow 1)^{\uparrow F} \\ \text{the function } (_ \rightarrow 1) \text{ is the identity function } \text{id}^{\mathbb{1} \rightarrow \mathbb{1}} : & = \text{wu}_F \triangleright \text{id}^{\uparrow F} \\ \text{functor identity law of } F : & = \text{wu}_F \triangleright \text{id} = \text{wu}_F \quad . \end{aligned}$$

To summarize our results: for any functor F ,

- If F is a lawful pointed functor, its `pure` method satisfies Eq. (8.10) where $\text{wu}_F \triangleq \text{pu}_F(1)$ is a fixed “wrapped unit” value of type $F^{\mathbb{1}}$.
- If any “wrapped unit” value $\text{wu}_F : F^{\mathbb{1}}$ is given, we may define a `pure` method by Eq. (8.10) and make the functor F into a lawful pointed functor; the naturality law will be satisfied automatically. The value $\text{pu}_F(1)$ will be equal to the originally given value wu_F .

So, the function pu_F and the value wu_F are **computationally equivalent**: each one can be converted into the other and back, with no loss of information. We may define a pointed functor equivalently as a functor with a chosen value wu_F of type $F^{\mathbb{1}}$. When reasoning about pointed functors, it is simpler to use the definition via the “wrapped unit” wu_F because it has a simpler type and no laws. When writing code, the `pure` method is more convenient.

8.3.6 Pointed functors: structural analysis

To perform structural analysis, we begin with the known functor-building constructions from Chapter 6 and impose an additional requirement that a “wrapped unit” value $wu_F : F^1$ should exist. As we have seen in the previous section, no additional laws need to be checked. The results are shown in Table 8.5. The typeclass can be defined by the simpler code,

```
final case class Pointed[F[_]](wu: F[Unit])
def pure[F[_]: Pointed : Functor, A](a: A): F[A] = implicitly[Pointed[F]].wu.map(_ => a)
```

When the type F^1 has several distinct values, the choice of wu_F is application-dependent. E.g., if $F^A \triangleq \text{List}^A$, the type List^1 has values such as an empty list $\text{List}[\text{Unit}]()$, the list of length 1, i.e., $\text{List}((\text{Unit}))$, the list of length 2, i.e., $\text{List}((\text{Unit}), (\text{Unit}))$, etc. Each of these choices gives a valid `Pointed` instance for the `List` functor. It is up to the programmer to choose the `Pointed` instance that will be useful for the application at hand. In the case of `List`, the standard choice `wu == List()` and correspondingly `pure(x) = List(x)` is motivated by the usage of the `List` type constructor to represent several possibilities, e.g., in a search problem; then the “pure” list represents the situation with only one possibility.

Nameless type-to-type functions To implement the `Pointed` typeclass instances for the following constructions, we need to use some advanced syntax features of Scala. In the previous sections, we wrote PTVFs parameterized by a type built out of other types, for example

```
def monoidPair[A: Monoid, B: Monoid]: Monoid[(A, B)] = ???
```

The function `monoidPair` creates a typeclass instance for the type $A \times B$, which is built out of the types A and B . In Scala, the typeclass instance is a value of type `Monoid[(A, B)]`. The analog for type constructors is a value of type `Pointed[L]` where the type constructor `L` is defined by the type alias

```
type L[A] = (F[A], G[A])
```

However, the following declaration of the analogous function `pointedPair` is invalid in Scala:

```
def pointedPair[F[_]: Pointed, G[_]: Pointed]: Pointed[L] // Does not work in Scala.
```

It is not possible to use the type alias `L` within this function declaration, because the type alias needs to use the type parameters `F` and `G` that are defined only within the type signature of the function. To achieve that, we would need somehow to insert a new type alias declaration within the type signature of `pointedPair`, but the syntax of Scala does not support that:

```
def pointedPair[F[_]: Pointed, G[_]: Pointed]: Pointed[L] // Not a valid Scala syntax.
  type L[A] = (F[A], G[A]) // Temporarily define a type constructor L, and now use it:
  Pointed[L] }
```

The return type is required to be `Pointed[L]`, where `L` needs to be a type expression that defines a type constructor, i.e., a type-to-type function with a *single* type parameter. Writing a type alias with parameters `F` and `G`,

```
type L[F[_], G[_], A] = (F[A], G[A])
def pointedPair[F[_]: Pointed, G[_]: Pointed]: Pointed[L[F, G]] // Still incorrect.
```

will not help because the type expression `L[F, G]` in `Pointed[L[F, G]]` is not a valid type constructor with one type parameter. We cannot define a new type name (such as `L`) within the type signature; what we need is a *nameless* type-to-type function (i.e., a nameless type constructor).

The special Scala plugin called the “kind projector”⁵ adds syntax for nameless type constructors. The syntax is similar to defining a nameless function: for instance, the pair functor $F^* \times G^*$ is defined as `Lambda[X => (F[X], G[X])]`. Such type expressions can be understood as nameless type-to-type functions. When using the “kind projector” plugin, the syntax for defining `pointedPair` is

```
def pointedPair[F[_]: Pointed, G[_]: Pointed]: Pointed[Lambda[X => (F[X], G[X])]] = ???
```

⁵<https://github.com/typelevel/kind-projector>

Scala 3 will support a shorter syntax for nameless type-to-type functions, e.g., `[X] => (F[X], G[X])`. We will use the “kind projector” syntax in this book’s code examples.

Fixed types A constant functor $\text{Const}^{Z,\bullet}$ is defined as $\text{Const}^{Z,A} \triangleq Z$, where Z is a fixed type. A “wrapped unit” value is thus also a value of type Z . Since we cannot produce values of an arbitrary type Z from scratch, the constant functor is not pointed in general. The constant functor *will* be pointed when there exists a known value of type Z . Examples are $Z = \mathbb{1}$ or $Z = \mathbb{1} + U$ (where U is an arbitrary type). If we know that Z is equivalent to $\mathbb{1} + U$, we will be able to produce a value of type Z as $\mathbb{1} + \mathbb{0}^U$. In that case, we set $\text{wu}_{\text{Const}^{Z,\bullet}} = \mathbb{1} + \mathbb{0}^U$.

```
type Const[Z, A] = Z
def pointedOpt[U]: Pointed[Lambda[X => Const[Option[U], X]]] = Pointed(None: Const[Option[U], Unit])
```

Other cases are types such as $Z = \text{Int}$, $Z = \text{String}$, etc., that have well-known “default” values.

Type parameters The identity functor $\text{Id}^A \triangleq A$ is pointed since $\text{Id}^{\mathbb{1}} = \mathbb{1}$, and we can set $\text{wu}_{\text{Id}} = \mathbb{1}$.

```
type Id[A] = A
def pointedId: Pointed[Id] = Pointed(Id(()))
```

The other functor constructions that work by setting type parameters are functor compositions. If F and G are two functors or two contrafunctors then $F \circ G$ is a functor. The functor $F \circ G$ is pointed when we can create a value of type F^G . If both F and G are pointed, we can apply F ’s `pure` method to $\text{wu}_G : G^{\mathbb{1}}$ and obtain a value of type F^G .

```
def pointedFoG[F[_]: Pointed : Functor, G[_]: Pointed]: Pointed[Lambda[X => F[G[X]]]] =
  Pointed[Lambda[X => F[G[X]]]](pure[F, G[Unit]](implicitly[Pointed[G]].wu))
```

The case when F and G are contrafunctors requires us to assume that F belongs to the “pointed contrafunctor” typeclass (see Section 8.3.8 below). A pointed contrafunctor has a “wrapped unit” value of type $F^{\mathbb{1}}$, which can be transformed into F^A for any type A by using `contramap` with a constant function $A \rightarrow \mathbb{1}$:

```
def cpure[F[_]: Pointed : Contrafunctor, A]: F[A] = implicitly[Pointed[F]].wu.cmap(_ => ())
```

In this way, we can create a value of type F^G . The contrafunctor G does not need to be pointed.

```
def pointedCFoG[F[_]: Pointed : Contrafunctor, G[_]: Pointed]: Pointed[Lambda[X => F[G[X]]]] =
  Pointed[Lambda[X => F[G[X]]]](cpure[F, G[Unit]])
```

Products If F and G are two pointed functors, is the functor product $L^A \triangleq F^A \times G^A$ a pointed functor? We need to produce a value $\text{wu}_L : F^{\mathbb{1}} \times G^{\mathbb{1}}$, and we have values $\text{wu}_F : F^{\mathbb{1}}$ and $\text{wu}_G : G^{\mathbb{1}}$. It is clear that we must set $\text{wu}_L = \text{wu}_F \times \text{wu}_G$.

```
def pointedFxG[F[_]: Pointed, G[_]: Pointed]: Pointed[Lambda[X => (F[X], G[X])]] =
  Pointed[Lambda[X => (F[X], G[X])]]((implicitly[Pointed[F]].wu, implicitly[Pointed[G]].wu))
```

Co-products If F and G are two pointed functors, is the functor co-product $L^A \triangleq F^A + G^A$ a pointed functor? We need to produce a value $\text{wu}_L : F^{\mathbb{1}} + G^{\mathbb{1}}$, and we have values $\text{wu}_F : F^{\mathbb{1}}$ and $\text{wu}_G : G^{\mathbb{1}}$. There are two choices, $\text{wu}_L = \text{wu}_F + \mathbb{0}^{G^{\mathbb{1}}}$ and $\text{wu}_L = \mathbb{0}^{F^{\mathbb{1}}} + \text{wu}_G$, both making L^{\bullet} a pointed functor.

It is sufficient if just F^{\bullet} is a pointed functor: $\text{wu}_L \triangleq \text{wu}_F + \mathbb{0}^{G^{\mathbb{1}}}$ is a `Pointed` typeclass instance for $F^{\bullet} + G^{\bullet}$, even if G^{\bullet} is not pointed.

```
def pointedEitherFG[F[_]: Pointed, G[_]: Pointed]: Pointed[Lambda[X => Either[F[X], G[X]]]] =
  Pointed[Lambda[X => Either[F[X], G[X]]]](Left(implicitly[Pointed[F]].wu))
```

Functions If C is any contrafunctor and F is a pointed functor, the exponential functor $L^A \triangleq C^A \rightarrow F^A$ will be pointed if we are able to produce a value $\text{wu}_L : C^{\mathbb{1}} \rightarrow F^{\mathbb{1}}$. We already have a value $\text{wu}_F : F^{\mathbb{1}}$, and we cannot use a value of type $C^{\mathbb{1}}$ with a general contrafunctor C . So, we have to set $\text{wu}_L \triangleq (_ \rightarrow \text{wu}_F)$. This makes L into a pointed functor.

```
def pointedFuncFG[F[_]: Pointed, C[_]: Pointed]: Pointed[Lambda[X => C[X] => F[X]]] =
```

```
Pointed[Lambda[X => C[X] => F[X]]](_ => implicitly[Pointed[F]].wu)
```

Recursive types The recursive construction for functors (see Statement 6.2.3.7) assumes a bifunctor $S^{\bullet,\bullet}$ and defines a recursive functor F^\bullet via the type equation $F^\bullet \triangleq S^{\bullet,F^\bullet}$. The functor F^\bullet will be pointed if we can compute a value wu_F of type F^\bullet . The type F^\bullet is a recursive type defined via the type equation $F^\bullet \triangleq S^{\bullet,F^\bullet}$. If that type is not void, i.e., if there exists some value of that type, we will be able to define wu_F as that value.

How can we construct wu_F for a given bifunctor S ? The procedure can be derived by structural analysis of S (see Section 8.5.1 below). For *polynomial* bifunctors S (which is the most often used kind of bifunctors), the necessary and sufficient condition is that the type $S^{1,0}$ should be non-void. If we can create a value of type $S^{1,0}$, arguments shown in Section 8.5.1 will guarantee that we can also create a value of type F^\bullet , and so the recursive functor F will be pointed.

As an example, consider the polynomial bifunctor $S^{A,R} \triangleq A + A \times R$. The corresponding recursive functor $F^A \triangleq S^{A,F^A} = A + A \times F^A$ is the non-empty list (see Example 3.3.2.1, Table 6.3, and Statement 6.2.3.7). The type F^A can be (non-rigorously) viewed as an “infinite disjunction”

$$F^A = A + A \times (A + A \times (A + \dots)) = A + A \times A + A \times A \times A + \dots$$

Since the type $S^{1,0} = 1 + 1 \times 0 \cong 1$ is non-void, the necessary and sufficient condition holds, so we expect that the recursive construction will work. The type F^\bullet is defined by

$$F^\bullet \triangleq S^{\bullet,F^\bullet} = 1 + 1 \times F^\bullet \cong 1 + F^\bullet \quad .$$

This type can be (non-rigorously) viewed as an “infinite disjunction”

$$F^\bullet \cong 1 + F^\bullet \cong 1 + 1 + F^\bullet = 1 + 1 + 1 + \dots$$

It is clear that a value of that type can be computed, for example, as

$$wu_F = 1 + 0 + 0 + \dots \cong 1 + 0^{F^\bullet} \quad .$$

In Scala, this is `Left()`. So, a `Pointed` typeclass instance for F is implemented by the following code:

```
type S[A, R] = Either[A, (A, R)]
final case class F[A](s: S[A, F[A]])
implicit val pointedF: Pointed[F] =
  Pointed(F(Left(())))
```

The corresponding `pure` method will use F ’s `map` to transform $wu_F = 1 + 0 + 0 + \dots$ into
 $a + 0 + 0 + \dots : A + A \times A + A \times A \times A + \dots \quad .$

The `pure` method of F creates a non-empty list with a single element a^A .

Summary Can we recognize a pointed functor F by looking at its type expression, e.g.

$$F^{A,B} \triangleq ((1 + A \rightarrow \text{Int}) \rightarrow A \times B) + \text{String} \times A \times A \quad ?$$

This type constructor is a functor in both A and B , and we ask whether F is pointed with respect to A and/or with respect to B .

To answer this question with respect to A , we set $A = 1$ in $F^{A,B}$ and obtain the type expression

$$F^{1,B} = ((1 + 1 \rightarrow \text{Int}) \rightarrow 1 \times B) + \text{String} \times 1 \times 1 \cong ((2 \rightarrow \text{Int}) \rightarrow B) + \text{String} \quad .$$

The functor F will be pointed with respect to A if we can compute a value of this type from scratch. At the outer level, this type expression is a disjunctive type with two parts; it is sufficient to compute one of the parts. Can we compute a value of type $(2 \rightarrow \text{Int}) \rightarrow B$? Since the parameter B is an arbitrary, unknown type, we cannot construct values of type B using a given value of type $2 \rightarrow \text{Int}$. The remaining possibility is to compute the second part of the co-product, which is $\text{String} \times 1 \times 1$. We are able to compute a value of this type because `String` is a fixed type with a known default value (an empty string). So, we conclude that $F^{A,B}$ is pointed with respect to A .

Construction	Type signature to implement	Results
Constant functor returning a fixed type Z	value of type Z	Z has a default
Identity functor	$\mathbb{1}$	one possibility
Composition of pointed functors/contrafunctors	$\text{Pointed}^{F^\bullet} \times \text{Pointed}^{G^\bullet} \rightarrow \text{Pointed}^{F^G}$	one possibility
Product of pointed functors F and G	$\text{Pointed}^{F^\bullet} \times \text{Pointed}^{G^\bullet} \rightarrow \text{Pointed}^{F^\bullet \times G^\bullet}$	one possibility
Co-product of a pointed functor F and any G	$\text{Pointed}^{F^\bullet} \times \text{Functor}^{G^\bullet} \rightarrow \text{Pointed}^{F^\bullet + G^\bullet}$	one possibility
Function from any C to a pointed F	$\text{Pointed}^{F^\bullet} \times \text{Contrafunctor}^{C^\bullet} \rightarrow \text{Pointed}^{C^\bullet \rightarrow F^\bullet}$	one possibility
Recursive type	$\text{Pointed}^{F^\bullet} \rightarrow \text{Pointed}^{S^\bullet F^\bullet}$ where $F^A \triangleq S^{A, F^A}$	$\text{Pointed}^{F^\bullet}$

Table 8.5: Type constructions producing the `Pointed` functor typeclass.

Considering now the type parameter B , we set $B = \mathbb{1}$ and obtain

$$F^{A, \mathbb{1}} = ((\mathbb{1} + A \rightarrow \text{Int}) \rightarrow A \times \mathbb{1}) + \text{String} \times A \times A \quad .$$

The type A is now an arbitrary and unknown type, so we cannot compute any values of A or $\text{String} \times A \times A$ from scratch. The function type $(\mathbb{1} + A \rightarrow \text{Int}) \rightarrow A \times \mathbb{1}$ cannot be implemented because a value of type A cannot be computed from a function $\mathbb{1} + A \rightarrow \text{Int}$ that *consumes* values of type A . So, $F^{A, B}$ is not pointed with respect to B .

8.3.7 Co-pointed functors

Pointed functors provide the functionality of wrapping a given value in a “pure wrapper”. Another useful operation is *extracting* a value from a given “wrapper”:

```
def extract[F[_], A]: F[A] => A
```

$\text{ex} : \forall A. F^A \rightarrow A \quad .$

Functors having this operation are called **co-pointed**. We may define the `Copointed` typeclass as

```
trait Copointed[F[_]] { def ex[A]: F[A] => A }
def extract[F[_]: Copointed, A](f: F[A]): A = implicitly[Copointed[F]].ex(f)
```

The `extract` function must be fully parametric and obey the **naturality law** (compare to Eq. (8.8)),

$$\begin{array}{ccc}
 F^A & \xrightarrow{\text{ex}_F} & A \\
 f^F \downarrow & & \downarrow f \\
 F^B & \xrightarrow{\text{ex}_F} & B
 \end{array}
 \quad \text{ex}_F ; f = f^F ; \text{ex}_F \quad . \quad (8.11)$$

The naturality law formulates our expectation that the extractor function somehow “selects” a value of type A among all the values wrapped by F^A , and the “selection” works independently of the values. If all wrapped

values are transformed by a function f into wrapped values of type B , the extractor function will still select a value of type B in the same way as it did for values of type A . So, the result will be the same as if we first extracted a value of type A and then transformed that value with f .

Both sides of the law (8.11) are functions of type $F^A \rightarrow B$. We saw in the previous section that the `pure` method of the `Pointed` typeclass is computationally equivalent to a single chosen value of type F^1 when F is a functor. That provides a simpler form of the `Pointed` typeclass. For co-pointed functors, there is no simpler form of the `extract` method. If we set $A = \mathbb{1}$ and $f: \mathbb{1} \rightarrow B \triangleq (1 \rightarrow b)$ in the naturality law, both sides will become functions of type $F^1 \rightarrow B$. But the type F^1 might be void, or a value of type F^1 may not be computable via fully parametric code. So, we cannot deduce any further information from the naturality law of co-pointed functors.

However, if F is a pointed functor, we *will* have a chosen value $\text{wu}_F : F^1$ to which we may then apply both sides of the naturality law (8.11) and obtain⁶

$$\text{wu}_F \triangleright \text{ex}_F \circ f \stackrel{!}{=} \text{wu}_F \triangleright f^{\uparrow F} \circ \text{ex}_F \quad .$$

Calculating both sides separately, we find

$$\begin{aligned} \text{wu}_F \triangleright \text{ex}_F \circ f &= \text{wu}_F \triangleright \text{ex}_F \triangleright f = 1 \triangleright f = b \quad . \\ \text{wu}_F \triangleright f^{\uparrow F} \circ \text{ex}_F &= b \triangleright \text{pu}_F \circ \text{ex}_F \quad . \end{aligned}$$

So, we know that $b \triangleright \text{pu}_F \circ \text{ex}_F \stackrel{!}{=} b$. This can hold for all $b : B$ only if

$$\text{pu}_F \circ \text{ex}_F = \text{id} \quad .$$

This additional **compatibility law** is a consequence of naturality laws if the functor F is pointed and co-pointed at the same time.

Let us perform structural analysis for co-pointed functors; the results are shown in Table 8.6.

Fixed types A constant functor $\text{Const}^{Z,A} \triangleq Z$ is *not* co-pointed because we cannot implement $\forall A. Z \rightarrow A$ (a value of an arbitrary type A cannot be computed from a value of a fixed type Z).

Type parameters The identity functor $\text{Id}^A \triangleq A$ is co-pointed with $\text{ex} \triangleq \text{id}^{A \rightarrow A}$. An identity function will always satisfy any naturality law.

```
type Id[A] = A
def copointedId: Copointed[Id] = new Copointed[Id] { def ex[A]: Id[A] => A = identity }
```

Composition of two co-pointed functors F, G is co-pointed:

$$\text{ex}_{F \circ G} \triangleq h^{\circ F^G} \rightarrow \text{ex}_G(\text{ex}_F(h)) \quad \text{or equivalently} \quad \text{ex}_{F \circ G} = \text{ex}_F \circ \text{ex}_G \quad .$$

```
def copointedFoG[F[_]: Copointed, G[_]: Copointed]: Copointed[Lambda[X => F[G[X]]]] =
  new Copointed[Lambda[X => F[G[X]]]] { def ex[A]: F[G[A]] => A = extract[F,G[A]] _ andThen extract[G, A] }
```

The naturality law holds for $\text{ex}_{F \circ G}$ because

$$\begin{aligned} \text{expect to equal } \text{ex}_{F \circ G} \circ f &: f^{\uparrow F \circ G} \circ \text{ex}_{F \circ G} \\ \text{definition of } f^{\uparrow F \circ G}, \text{ see Eq. (6.18)} &: = (f^{\uparrow G})^{\uparrow F} \circ \text{ex}_F \circ \text{ex}_G \\ \text{naturality law of } \text{ex}_F &: = \text{ex}_F \circ f^{\uparrow G} \circ \text{ex}_G \\ \text{naturality law of } \text{ex}_G &: = \text{ex}_F \circ \text{ex}_G \circ f = \text{ex}_{F \circ G} \circ f \quad . \end{aligned}$$

Products If functors F and G are co-pointed, we can implement a function of type $F^A \times G^A \rightarrow A$ in two different ways: by discarding F^A or by discarding G^A . With either choice, the functor product $F^* \times G^*$ is made into a co-pointed functor. For instance, if we choose to discard G^A then the functor G will not need to be co-pointed, and the code for the `extract` method will be

$$\text{ex}_{F \times G} \triangleq f^{\circ F^A} \times g^{\circ G^A} \rightarrow \text{ex}_F(f) = \pi_1 \circ \text{ex}_F \quad ,$$

where we used the pair projection function $\pi_1 \triangleq (a \times b \rightarrow a)$.

```
def copointedFxF[G[_]: Copointed]: Copointed[Lambda[X => (F[X], G[X])]] =
  new Copointed[Lambda[X => (F[X], G[X])]] { def ex[A]: ((F[A], G[A])) => A = { case (f, g) => extract(f) } }
```

⁶The symbol $\stackrel{!}{=}$ means “must be equal, according to what we know”.

The following calculation verifies the naturality law (8.11) for this definition of $\text{ex}_{F \times G}$:

$$\begin{aligned}
 & \text{expect to equal } \text{ex}_{F \times G} \circ f : \quad \underline{f^{\uparrow F \times G}} \circ \text{ex}_{F \times G} \\
 & \text{definition of } f^{\uparrow F \times G}, \text{ see Eq. (6.13)} : \quad = (f^{\uparrow F} \boxtimes f^{\uparrow G}) \circ \underline{\text{ex}_{F \times G}} \\
 & \text{definition of } \text{ex}_{F \times G} : \quad = \underline{(f^{\uparrow F} \boxtimes f^{\uparrow G})} \circ \underline{\pi_1} \circ \text{ex}_F \\
 & \text{use the property } (p \boxtimes q) \circ \pi_1 = \pi_1 \circ p : \quad = \underline{\pi_1} \circ \underline{f^{\uparrow F}} \circ \text{ex}_F \\
 & \text{naturality law of } \text{ex}_F : \quad = \underline{\pi_1} \circ \text{ex}_F \circ f \\
 & \text{definition of } \text{ex}_{F \times G} : \quad = \text{ex}_{F \times G} \circ f \quad .
 \end{aligned}$$

To demonstrate the property $(p \boxtimes q) \circ \pi_1 = \pi_1 \circ p$ used in this proof, apply both sides to $a \times b$:

$$\begin{aligned}
 (a \times b) \triangleright (p \boxtimes q) \circ \pi_1 &= \underline{(a \times b) \triangleright (p \boxtimes q)} \circ \pi_1 = (p(a) \times q(b)) \triangleright \pi_1 = p(a) \quad , \\
 (a \times b) \triangleright \pi_1 \circ p &= \underline{(a \times b) \triangleright \pi_1} \circ p = a \triangleright p = p(a) \quad .
 \end{aligned}$$

Co-products For co-pointed functors F and G , there is only one possible implementation of the type signature $\text{ex}_{F+G} : F^A + G^A \rightarrow A$, given that we have functions ex_F and ex_G :

```

def copointedEitherFG[F[_], G[_]: Copointed, G[_]: Copointed]:
    Copointed[Lambda[X => Either[F[X], G[X]]]] =
  new Copointed[Lambda[X => Either[F[X], G[X]]]] {
    def ex[A]: Either[F[A], G[A]] => A = {
      case Left(f) => extract(f)
      case Right(g) => extract(g)
    }
  }
}
  
```

$$\text{ex}_{F+G} \triangleq \begin{array}{|c|c|} \hline & A \\ \hline F^A & \text{ex}_F \\ \hline G^A & \text{ex}_G \\ \hline \end{array} \quad .$$

To verify that ex_{F+G} satisfies the naturality law, we compute:

$$\begin{aligned}
 & \text{expect to equal } \text{ex}_{F+G} \circ f : \quad \underline{f^{\uparrow F+G}} \circ \text{ex}_{F+G} \\
 & \text{definition of } f^{\uparrow F+G}, \text{ see Eq. (6.14)} : \quad = \begin{array}{|c|c|} \hline f^{\uparrow F} & \emptyset \\ \hline \emptyset & f^{\uparrow G} \\ \hline \end{array} \circ \begin{array}{|c|} \hline \text{ex}_F \\ \hline \text{ex}_G \\ \hline \end{array} \\
 & \text{matrix function composition} : \quad = \begin{array}{|c|} \hline \underline{f^{\uparrow F} \circ \text{ex}_F} \\ \hline \underline{f^{\uparrow G} \circ \text{ex}_G} \\ \hline \end{array} \\
 & \text{naturality laws of } \text{ex}_F \text{ and } \text{ex}_G : \quad = \begin{array}{|c|} \hline \text{ex}_F \circ f \\ \hline \text{ex}_G \circ f \\ \hline \end{array} = \begin{array}{|c|} \hline \text{ex}_F \\ \hline \text{ex}_G \\ \hline \end{array} \circ f = \text{ex}_{F+G} \circ f \quad .
 \end{aligned}$$

Functions An exponential functor of the form $L^A \triangleq C^A \rightarrow P^A$ (where C is a contrafunctor and P is a functor) will be co-pointed if we can implement a function of type $\forall A. (C^A \rightarrow P^A) \rightarrow A$,

$$\text{ex}_L \triangleq h : C^A \rightarrow P^A \rightarrow ???^A \quad .$$

Since the type A is arbitrary, the only way of computing a value of type A is somehow to use the function h . The only way of using h is to apply it to a value of type C^A , which will yield a value of type P^A . So, we need to assume that we can somehow create a value of type C^A , for any type A . We may call a contrafunctor C with a method `cpure` of type $\forall A. C^A$ a **pointed contrafunctor**. Assuming that C is pointed and denoting its `cpure` method by cpu_C , we can thus compute a value of type P^A as $h(\text{cpu}_C)$. To extract A from P^A , we need to assume additionally that P is co-pointed and use its method $\text{ex}_P : P^A \rightarrow A$. Finally we have

$$\text{ex}_L \triangleq h : C^A \rightarrow P^A \rightarrow \text{ex}_P(h(\text{cpu}_C)) \quad \text{or equivalently} \quad h : C^A \rightarrow P^A \triangleright \text{ex}_L = \text{cpu}_C \triangleright h \triangleright \text{ex}_P \quad . \quad (8.12)$$

To verify the naturality law, we apply both sides to an arbitrary $h:C^A \rightarrow P^A$ and compute

$$\begin{aligned}
 \text{expect to equal } h \triangleright \text{ex}_L \circ f &: \triangleright \text{ex}_L h \triangleright f^{\uparrow L} \circ \text{ex}_L = (h \triangleright f^{\uparrow L}) \\
 \text{use Eq. (8.12)} &: = \text{cpu}_C \triangleright (h \triangleright f^{\uparrow L}) \triangleright \text{ex}_P = \text{cpu}_C \triangleright \underline{(h \triangleright f^{\uparrow L}) \circ \text{ex}_P} \\
 \text{definition of } f^{\uparrow L}, \text{ see Eq. (6.16)} &: = \text{cpu}_C \triangleright f^{\downarrow C} \circ h \circ \underline{f^{\uparrow P} \circ \text{ex}_P} \\
 \text{naturality law of } \text{ex}_P &: = \text{cpu}_C \triangleright f^{\downarrow C} \circ h \circ \text{ex}_P \circ f \quad .
 \end{aligned}$$

We expect the last expression to equal

$$h \triangleright \text{ex}_L \circ f = \text{cpu}_C \triangleright h \triangleright \text{ex}_P \triangleright f = \text{cpu}_C \triangleright h \circ \text{ex}_P \circ f \quad .$$

This is possible only if $\text{cpu}_C \triangleright f^{\downarrow C} = \text{pu}_C$ for all f . This motivates us to assume that law as the **naturality law** of cpu_C for pointed contrafunctors. With that last assumption, we have finished proving the naturality law of ex_L . The code for ex_L is

```
def copointedFunc[C[_]: Pointed, P[_]: Copointed]: Copointed[Lambda[X => C[X] => P[X]]] =  
  new Copointed[Lambda[X => C[X] => P[X]]] {  
    def ex[A]: (C[A] => P[A]) => A = h => extract[P, A](h(cpure[C, A]))  
  }      // In Scala 2.13: h => cpure[C, A] pipe h pipe extract[P, A] as in h → pu_C ∘ h ∘ exp
```

We will analyze pointed contrafunctors and their naturality law in Section 8.3.8.

Recursive types Consider a functor F defined by a recursive equation $F^A \triangleq S^{A,F^A}$ where S is a bifunctor (see Section 6.2.2). The functor F is co-pointed if a method $\text{ex}_F : (F^A \rightarrow A) \cong (S^{A,F^A} \rightarrow A)$ can be defined. Since the recursive definition of F uses F^A as a type argument in S^{A,F^A} , we may assume (by induction) that an extractor function $F^A \rightarrow A$ is already available when applied to the recursively used F^A . Then we can use the `bimap` method of S to map $S^{A,F^A} \rightarrow S^{A,A}$. It remains to extract a value of type A out of a bifunctor value $S^{A,A}$. We call a bifunctor S **co-pointed** if a fully parametric function $\text{ex}_S : S^{A,A} \rightarrow A$ exists satisfying the corresponding **naturality law**

$$\text{ex}_S \circ f = \text{bimap}_S(f)(f) \circ \text{ex}_S \quad . \quad (8.13)$$

Assuming that S is co-pointed, we can finally define ex_F by recursion,

$$\text{ex}_F \triangleq s : S^{A,F^A} \rightarrow s \triangleright (\text{bimap}_S(\text{id})(\text{ex}_F)) \triangleright \text{ex}_S \quad \text{or equivalently} \quad \text{ex}_F \triangleq \text{bimap}_S(\text{id})(\text{ex}_F) \circ \text{ex}_S \quad .$$

To verify the naturality law of ex_F , we denote recursive uses by an overline and compute:

$$\begin{aligned}
 \text{expect to equal } \text{ex}_F \circ f &: \underline{f^{\uparrow F}} \circ \text{ex}_F \\
 \text{definition of } f^{\uparrow F}, \text{ see Eq. (6.20)} &: = \text{bimap}_S(f)(\overline{f^{\uparrow F}}) \circ \text{ex}_F \\
 \text{definition of } \text{ex}_F &: = \underline{\text{bimap}_S(f)(\overline{f^{\uparrow F}})} \circ \text{bimap}_S(\text{id})(\overline{\text{ex}_F}) \circ \text{ex}_S \\
 \text{bifunctor composition law (6.10)} &: = \text{bimap}_S(f \circ \text{id})(\overline{f^{\uparrow F} \circ \text{ex}_F}) \circ \text{ex}_S \\
 \text{naturality law of } \text{ex}_F &: = \underline{\text{bimap}_S(f)(\overline{\text{ex}_F \circ f})} \circ \text{ex}_S \\
 \text{bifunctor composition law in reverse} &: = \text{bimap}_S(\text{id})(\overline{\text{ex}_F}) \circ \text{bimap}_S(f)(\overline{f}) \circ \text{ex}_S \\
 \text{naturality law (8.13) of } \text{ex}_S &: = \underline{\text{bimap}_S(\text{id})(\overline{\text{ex}_F})} \circ \text{ex}_S \circ f \\
 \text{definition of } \text{ex}_F &: = \text{ex}_F \circ f \quad .
 \end{aligned}$$

An example illustrating the recursive construction is the bifunctor $S^{A,R} \triangleq A + R \times R$ that defines F^A to be the binary tree functor,

$$F^A \triangleq S^{A,F^A} = A + F^A \times F^A \quad .$$

The bifunctor S is co-pointed because there exists a suitable function $\text{ex}_S : S^{A,A} \rightarrow A$.

Construction	Type signature to implement	Results
Identity functor	$\text{id} : A \rightarrow A$	one possibility
Composition of co-pointed functors	$\text{Copointed}^{F^\bullet} \times \text{Copointed}^{G^\bullet} \rightarrow \text{Copointed}^{F^G}$	one possibility
Product of co-pointed functor F and any G	$\text{Copointed}^{F^\bullet} \times \text{Functor}^{G^\bullet} \rightarrow \text{Copointed}^{F^\bullet \times G^\bullet}$	one possibility
Co-product of co-pointed functors F and G	$\text{Copointed}^{F^\bullet} \times \text{Copointed}^{G^\bullet} \rightarrow \text{Copointed}^{F^\bullet + G^\bullet}$	one possibility
Function from pointed C to co-pointed F	$\text{Pointed}^C \times \text{Copointed}^F \rightarrow \text{Copointed}^{C^\bullet \rightarrow F^\bullet}$	one possibility
Recursive type	$\text{Copointed}^F \rightarrow \text{Copointed}^{S^\bullet, F^\bullet}$ where $F^A \triangleq S^{A, F^A}$	Copointed^F

Table 8.6: Type constructions producing the `Copointed` functor typeclass.

```
type S[A, R] = Either[A, (R, R)]
def exS[A]: S[A, A] => A = {
  case Left(a)      => a
  case Right((a1, a2)) => a1 // Could be
    'a2'.
}
```

In the code notation, the function `exS` is written as

$$\text{ex}_S \triangleq \begin{array}{c|c|c} & & A \\ \hline & A & \text{id} \\ \hline A \times A & & \pi_1 \end{array}.$$

This function extracts the left-most leaf of a binary tree because it is using the projection π_1 (not π_2).To implement the co-pointed instance `exF` for the binary tree functor, we need to use the implementation of `bimapS`,

```
def bimap_S[A,B,P,Q](f: A => B)(g: P => Q): S[A, P] => S[B, Q] = {
  case Left(a)      => Left(f(a))
  case Right((x, y)) => Right((g(x), g(y)))
}
```

Now we can define the recursive type constructor F^\bullet and a co-pointed instance for it,

```
final case class F[A](s: S[A, F[A]])
val copointedF: Copointed[F] = new Copointed[F1] {
  def ex[A]: F1[A] => A = { case F1(s) => exS(bimap_S(identity[A], ex[A])(s)) }
}
```

The naturality law holds for the recursive instance by induction (see Section 8.3.2).

Summary Can we recognize a co-pointed functor F by looking at its type expression, e.g.

$$F^{A,B} \triangleq ((\mathbb{1} + A \rightarrow \text{Int}) \rightarrow A \times B) + \text{String} \times A \times A \quad ?$$

The constructions shown in this section tell us that a co-product of two co-pointed functors is again co-pointed. Let us first check whether $\text{String} \times A \times A$ is co-pointed. We can certainly extract a value of type A out of $\text{String} \times A \times A$, but not a value of type B . So, there is no hope for $F^{A,B}$ to be co-pointed with respect to the parameter B .It remains to consider A as the type parameter. The type constructor $\text{String} \times A \times A$ is co-pointed, but we still need to check $(\mathbb{1} + A \rightarrow \text{Int}) \rightarrow A \times B$, which is a function type. The function construction requires $\mathbb{1} + A \rightarrow \text{Int}$ to be a pointed contrafunctor and $A \times B$ to be a co-pointed functor (with respect to A). It is clear that $A \times B$ is co-pointed with respect to A since we have $\pi_1 : A \times B \rightarrow A$. It remains to check that the contrafunctor $C^A \triangleq \mathbb{1} + A \rightarrow \text{Int}$ is pointed. A contrafunctor C^\bullet is pointed if values of type C^1 can be computed (see Section 8.3.8); this requires us to compute a value of type $\mathbb{1} + \mathbb{1} \rightarrow \text{Int}$. One such value is `{ _ => 0 }`, a constant function that always returns the integer 0.We conclude that $F^{A,B}$ is co-pointed with respect to A but not co-pointed with respect to B .

8.3.8 Pointed contrafunctors

In the previous section, the function-type construction required a contrafunctor C^\bullet to have a method `cpuC` of type $\forall A. C^A$; we called such contrafunctors **pointed**. We also needed to assume that the

naturality law holds for all functions $f: A \rightarrow B$,

$$\text{cpu}_C \triangleright f^{\downarrow C} = \text{cpu}_C \quad \text{or equivalently} \quad \text{cmap}_C(f: A \rightarrow B)(\text{cpu}_C^{\downarrow C}) = \text{cpu}_C^{\downarrow C} \quad . \quad (8.14)$$

$$\begin{array}{ccc} \text{cpu}_C : C^B & & B \\ \downarrow \text{cmap}_C(f) & & \uparrow f \\ \text{cpu}_C : C^A & & A \end{array}$$

We may simplify the formulation of the typeclass by setting $B = \mathbb{1}$ in the naturality law (8.14) and denoting $\text{wu}_C \triangleq \text{cpu}_C^{\downarrow \mathbb{1}}$. The law (8.14) then gives

$$\text{cpu}_C^A = \text{wu}_C \triangleright (\underline{\text{:_}^A \rightarrow 1})^{\downarrow C} \quad . \quad (8.15)$$

In this way, we express the `cpure` method through a chosen value $\text{wu}_C : C^{\mathbb{1}}$. For the same reasons as in the case of pointed functors, cpu_C and wu_C are computationally equivalent. The law (8.14) for cpu_C will be satisfied automatically if cpu_C is defined via Eq. (8.15). To verify that, compute

$$\begin{aligned} \text{expect to equal } \text{cpu}_C^A : & \quad \text{cpu}_C^B \triangleright f^{\downarrow C} \\ \text{use definition (8.15)} : & = \text{wu}_C \triangleright (\underline{\text{:_}^B \rightarrow 1})^{\downarrow C} \triangleright f^{\downarrow C} \\ \text{composition law of contrafunctor } C : & = \text{wu}_C \triangleright (\underline{f \circ (\text{:_}^B \rightarrow 1)})^{\downarrow C} \\ \text{compute function composition} : & = \text{wu}_C \triangleright (\underline{\text{:_}^A \rightarrow 1})^{\downarrow C} = \text{cpu}_C^A \quad . \end{aligned}$$

So, a pointed contrafunctor instance for C^{\bullet} is equivalent to a chosen value of type $C^{\mathbb{1}}$.

```
final case class Pointed[F[_]](wu: F[Unit])
def cpure[F[_]: Pointed : Contrafunctor, A]: F[A] = implicitly[Pointed[F]].wu.cmap(_ => ())
```

Here we used a `Contrafunctor` typeclass (see Example 8.4.1.7 below). We can now apply structural analysis to pointed contrafunctors, similarly to Section 6.2.4. The results are shown in Table 8.7.

Fixed types This construction is the same as for pointed functors (Section 8.3.6). A fixed type Z gives a constant contrafunctor $C^A \triangleq Z$. Since $C^{\mathbb{1}} = Z$, the constant contrafunctor is pointed if we have a chosen value of type Z ; this will be the case, for instance, if $Z = \mathbb{1} + U$ for some type U .

Type parameters Since the identity functor $\text{Id}^A \triangleq A$ is not a contrafunctor, it remains to consider the functor compositions C^{F^A} and F^{C^A} where C^{\bullet} is a contrafunctor and F^{\bullet} is a functor.

If C is pointed, we can always obtain a value cpu_C of type C^A for any type A , in particular for $A = F^{\mathbb{1}}$ (whether or not a value of type $F^{\mathbb{1}}$ can be computed). So, $C^{F^{\bullet}}$ is a pointed contrafunctor whenever C^{\bullet} is one, for any (not necessarily pointed) functor F .

```
def pointedCoF[C[_]: Pointed: Contrafunctor, F[_]: Pointed[Lambda[X => C[F[X]]]] = 
  Pointed[Lambda[X => C[F[X]]]](cpure[C, F[Unit]])
```

Creating a value of type $F^{C^{\mathbb{1}}}$ requires F to have a `pure` method that could be applied to a value of type $C^{\mathbb{1}}$ to compute a value of type $F^{C^{\mathbb{1}}}$. So, $F^{C^{\bullet}}$ is pointed whenever both C^{\bullet} and F^{\bullet} are pointed.

```
def pointedFoC[C[_]: Pointed, F[_]: Pointed : Functor]: Pointed[Lambda[X => F[C[X]]]] = 
  Pointed[Lambda[X => F[C[X]]]](pure[F, C[Unit]](implicitly[Pointed[C]].wu))
```

Products The construction is the same as for pointed functors: If we have values of type $C^{\mathbb{1}}$ and $D^{\mathbb{1}}$, we can compute the pair $C^{\mathbb{1}} \times D^{\mathbb{1}}$. This makes the product contrafunctor $L^A \triangleq C^A \times D^A$ pointed if both C^{\bullet} and D^{\bullet} are pointed contrafunctors.

Co-products The construction is the same as for pointed functors: If at least one of the contrafunctors C^{\bullet} and D^{\bullet} is pointed, we can create a `Pointed` instance for the co-product contrafunctor $L^A \triangleq C^A + D^A$ as either $\text{wu}_L = \text{wu}_C + \mathbb{0}^{\downarrow D^{\mathbb{1}}}$ or $\text{wu}_L = \mathbb{0}^{\downarrow C^{\mathbb{1}}} + \text{wu}_D$.

Construction	Type signature to implement	Results
Constant functor returning a fixed type Z	value of type Z	Z has a default
Composition of pointed functors/contrafunctors	$\text{Pointed}^{F^\bullet} \times \text{Pointed}^{G^\bullet} \rightarrow \text{Pointed}^{F^\bullet G^\bullet}$	one possibility
Product of pointed contrafunctors F and G	$\text{Pointed}^{F^\bullet} \times \text{Pointed}^{G^\bullet} \rightarrow \text{Pointed}^{F^\bullet \times G^\bullet}$	one possibility
Co-product of a pointed F and any G	$\text{Pointed}^{F^\bullet} \times \text{Contrafunctor}^{G^\bullet} \rightarrow \text{Pointed}^{F^\bullet + G^\bullet}$	one possibility
Function from a functor F to a pointed C	$\text{Pointed}^{C^\bullet} \times \text{Functor}^{F^\bullet} \rightarrow \text{Pointed}^{F^\bullet \rightarrow C^\bullet}$	one possibility
Recursive type	$\text{Pointed}^{C^\bullet} \rightarrow \text{Pointed}^{S^\bullet C^\bullet}$ where $C^A \triangleq S^{A, C^A}$	$\text{Pointed}^{C^\bullet}$

Table 8.7: Type constructions producing the `Pointed` contrafunctor typeclass.

Functions The exponential contrafunctor construction is $L^A \triangleq F^A \rightarrow C^A$, where C^\bullet is a contrafunctor and F^\bullet is a functor. To create a value $\text{wu}_L : L^1$ means to create a function of type $F^1 \rightarrow C^1$. That function cannot use its argument of type F^1 for computing a value C^1 since F is an arbitrary functor. So, wu_L must be a constant function ($_ : F^1 \rightarrow \text{wu}_C$), where we assumed that a value $\text{wu}_C : C^1$ is available. Thus, $F^A \rightarrow C^A$ is pointed when C is a pointed contrafunctor and F is any functor.

```
def pointedFuncFC[C[_]: Pointed, F[_]: Pointed[Lambda[X => F[X] => C[X]]] =  
  Pointed[Lambda[X => F[X] => C[X]]](_ => implicitly[Pointed[C]].wu)
```

Recursive types The recursive construction for contrafunctors (see Statement 6.2.4.3) is $C^A \triangleq S^{A, C^A}$ where $S^{A, R}$ is a contrafunctor in A and a functor in R . Values of type C^1 will exist when the recursive type equation $T \triangleq S^{1, T}$ defines a non-void type T . This condition is similar to that for pointed functors, and the resulting construction is the same.

Summary Can we recognize a pointed contrafunctor C by looking at its type expression, e.g.

$$C^{A, B} \triangleq (\mathbb{1} + A \rightarrow B) + (\text{String} \times A \times B \rightarrow \text{String}) \text{ with respect to type parameter } A?$$

We need to set $A = \mathbb{1}$ and try to create a value $\text{wu} : C^{1, B}$. In this example, $C^{1, B} = (\mathbb{1} + \mathbb{1} \rightarrow B) + (\text{String} \times \mathbb{1} \times B \rightarrow \text{String})$. A value of this type is $\text{wu}_C \triangleq \mathbb{0} + (s : \text{String} \times \mathbb{1} \times b : B \rightarrow s)$. So, the contrafunctor $C^{A, B}$ is pointed with respect to A .

8.4 Summary

What problems can we solve now?

- Define arbitrary PTVFs using typeclass constraints on type parameters.
- Define typeclasses and typeclass instances for types and for type constructors.
- Implement `Monoid`, `Functor`, and other standard typeclasses, and prove their laws.
- Use known constructions to derive typeclass instances from previous ones.

What problems cannot be solved with these tools?

- *Automatically* derive type class instances for given data types or type constructors.
- Combine typeclasses and express dependencies, e.g., typeclass `tc1` requires `tc2` and `tc3`.

We may want to write code such as

```

type F[A] = (A => Int) => A           // Define a type constructor.
implicit val functorF: Functor[F] = implement // Automatically implement typeclass instance for F.
implicit val pointedF: Pointed[F] = implement // Automatically use the function-type construction.

```

However, no currently available library provides such functionality. Also, typeclass instances are not always derived uniquely, as we have seen in several cases (e.g., the co-product construction of monoids or pointed functors).

We will discuss how to combine typeclasses in Section 8.5.6 below.

8.4.1 Solved examples

Example 8.4.1.1 Define a PTVF with type signature `def bitsize[T]: Int` such that `bitsize[Short]` returns 16, `bitsize[Int]` returns 32, and `bitsize[Long]` returns 64. For all other types `T`, the expression `bitsize[T]` should remain undefined.

Solution The function `bitsize[T]` needs to take an additional implicit argument that will be available only for `T = Short`, `T = Int`, and `T = Long`. To implement that, we need to define a new type constructor, say `HasBitsize[T]`, and create implicit values of the corresponding types. The new type constructor defines a typeclass whose instances need to carry the information about the bit size:

```

final case class HasBitsize[T](size: Int)
object HasBitsize {
  implicit val bitsizeShort = HasBitsize[Short](16)
  implicit val bitsizeInt   = HasBitsize[Int] (32)
  implicit val bitsizeLong  = HasBitsize[Long] (64)
}

```

Now we can define the function `bitsize` as a PTVF,

```
def bitsize[T](implicit ti: HasBitsize[T]): Int = ti.size
```

The instance argument such as `ti: HasBitsize[T]` is sometimes called the “evidence” argument because its presence provides “evidence” that the type `T` belongs to the type domain of the typeclass.

We can check that the function `bitsize` is defined only for supported types:

```

scala> bitsize[Long]
res0: Int = 64

scala> bitsize[String]
<console>:15: error: could not find implicit value for evidence parameter of type HasBitsize[String]
          bitsize[String]
          ^

```

The current implementation of `HasBitsize` allows the programmer to add new types to its type domain whenever necessary. For example, the following code will add support for the `Boolean` type so that `bitsize[Boolean]` will evaluate to 1:

```
implicit val bitsizeBoolean = HasBitsize[Boolean](1)
```

In some applications, it is important that the type domain of a PTVF should remain fixed (e.g., as defined in a library). To prevent the programmer from creating any further values of type `HasBitsize`, we could make it a non-case class whose constructor is declared as a `private` function like this:

```

final class HasBitsize[T] private (val size: Int) // Not a case class; the constructor is private.
object HasBitsize {                           // The companion object is allowed to call the private constructor.
  implicit val bitsizeShort = new HasBitsize[Short](16)
  implicit val bitsizeInt   = new HasBitsize[Int] (32)
  implicit val bitsizeLong  = new HasBitsize[Long] (64)
}

```

The code of `bitsize[T]` remains unchanged. With these definitions, no further typeclass instances can be created by any code outside of the companion object `HasBitsize`:

```
scala> implicit val bitsizeBoolean = new HasBitsize[Boolean](1)
<console>:16: error: constructor HasBitsize in class HasBitsize cannot be accessed in object $iw
      implicit val bitsizeBoolean = new HasBitsize[Boolean](1)

```

An implementation via a `trait` requires longer code but brings no significant advantages:

```
trait HasBitsize[T] { def size: Int } // Declare the trait as 'sealed' to prohibit further instances.
object HasBitsize {
  implicit val bitsizeShort = new HasBitsize[Short]{ def size: Int = 16 }
  implicit val bitsizeInt = new HasBitsize[Int] { def size: Int = 32 }
  implicit val bitsizeLong = new HasBitsize[Long] { def size: Int = 64 }
}
```

Example 8.4.1.2 Define a `Monoid` instance for the type $1 + (\text{String} \rightarrow \text{String})$.

Solution We look for suitable monoid constructions (Section 8.3.4) that build up the given type expression from simpler parts. Since the type expression $1 + (\text{String} \rightarrow \text{String})$ is a co-product at the outer level, we must start with the co-product construction, which requires us to choose one of the parts of the disjunction, say $\text{String} \rightarrow \text{String}$, as the “preferred” monoid. Next, we need to produce `Monoid` instances for 1 and for $\text{String} \rightarrow \text{String}$. While the `Unit` type has a unique `Monoid` instance, there are several for $\text{String} \rightarrow \text{String}$. The function-type construction gives two possible monoid instances: the monoid $R \rightarrow A$ with $R = A = \text{String}$, and the function composition monoid. Let us choose the latter. The code using the monoid constructions from Section 8.3.4 can then be written as

```
val monoidX: Monoid[Either[Unit, String => String]] = {
  implicit val m1 = Monoid[Unit]( (x, y) => (), () )
  implicit val m2: Monoid[String => String] = monoidFunc1[String]
  monoidEitherPreferB[Unit, String => String]
}
```

We can translate the constructions into code for a `Monoid` instance for the type `Option[String => String]`:

```
val monoidX: Monoid[Option[String => String]] = Monoid( {
  case (None, None)      => None
  case (None, Some(f))    => Some(f)
  case (Some(f), None)    => Some(f)
  case (Some(f), Some(g)) => Some(f andThen g)
}, None )
```

Example 8.4.1.3 Show that if A is a monoid and B is a semigroup then $A + B$ is a monoid.

Solution The co-product construction $A + B$ (where both A, B are monoids) has two implementations: one of the empty elements e_A or e_B must be chosen as the empty element for the monoid $A + B$. If B is not a monoid, the only choice is to set $e_{A+B} \triangleq e_A + \emptyset$. This is the implementation in the function `monoidEitherPreferB` (Section 8.3.4). We just need to replace the `Monoid` typeclass constraint for B by `Semigroup`:

```
def monoidEitherSemigroup[A: Monoid, B: Semigroup] = Monoid[Either[A, B]]( {
  case (Left(a1), Left(a2))      => Left(a1 |+| a2)
  case (Left(a), Right(b))       => Right(b) // "Take B".
  case (Right(b), Left(a))       => Right(b)
  case (Right(b1), Right(b2))    => Right(b1 |+| b2)
}, Left(implicitly[Monoid[A]].empty) ) // The type B does not need an empty element.
```

The monoid laws hold here because the proofs of the laws do not depend on the existence of e_B .

Example 8.4.1.4 (a routing monoid) A (much simplified) web server is implemented as a number of “routes”. Each route may respond to one or more URL paths by evaluating a custom function. The task is to implement a `combine` operation for routes. The combined route should respond to all paths that at least one of the previous routes responds to:

```
type Path = String          // Here, the types 'Path' and 'Response' are defined only as an
type Response = (Int, String) // illustration. The code will use these types as type parameters.
```

```

type Route = Path => Option[Response]

val r1: Route = { case "/get_users" => (200, "user1, user2, user3") }
val r2: Route = { case "/get_names" => (200, "name1, name2, name3") }
    // The task is to implement an extension method |+| such that this works correctly:
val route: Route = r1 |+| r2      // Should respond to both '/get_users' and '/get_names'.

```

Use the cats library for implementing a `Monoid` instance for routes; verify that the monoid laws hold.

Solution We will first figure out how to implement the required functionality, and then adapt the code to the cats library's definition of the `Monoid` typeclass.

A “route” is a function of type `Path => Option[Response]` that returns a non-empty option if the route responds to a given path value. A combination of two routes `r1` and `r2` needs to be a new route, i.e., a function `Path => Option[Response]`. The new function will first check whether `r1` responds to a given path. If so, it will evaluate the result of applying `r1`. Otherwise, it will try applying `r2` to the given path. We implement the required business logic in a function called `combineRoutes`:

```

def combineRoutes(r1: Route, r2: Route): Route = { path =>
  r1(path) match {
    case Some(response) => Some(response)
    case None => r2(path)
  }
}

```

A monoid also needs to have an empty element. An “empty route” can be combined with any other route and will not change the behavior of that route. If the empty route responds to any path, it would prevent another route from also responding to the same path. So, the only solution is to define the “empty route” as a function that *never* responds to any path:

```
val emptyRoute: Route = { _ => None }
```

The cats library defines the `Monoid` typeclass via a `trait` with methods `empty` and `combine`. We can define that typeclass using our existing code:

```

import $ivy.`org.typelevel::cats-core:1.5.0`, cats.Monoid      // Using 'ammonite' for convenience.
implicit val catsMonoidRoute: Monoid[Route] = new Monoid[Route] {
  def empty: Route = emptyRoute
  def combine(x: Route, y: Route): Route = combineRoutes(x, y)
}

```

We can now check that the routes can be combined as we intended:

```

import cats.syntax.monoid._
val route: Route = r1 |+| r2

ammonite@ route("/get_users")
res0: Response = (200, "user1, user2, user3")

ammonite@ route("/get_names")
res1: Response = (200, "name1, name2, name3")

```

To verify that the monoid laws hold, we could look for direct proofs using the code of `emptyRoute` and `combineRoutes`. However, it is easier to figure out how to reduce the definition of the `Route` monoid to a number of constructions listed in Section 8.3.4.

For more convenient reasoning, we replace the types `Path` and `Response` by type parameters P and R . We note that these types are used fully parametrically by the code of our functions. So, we will write the type `Route` as $\text{Route} \triangleq P \rightarrow \mathbb{1} + R$. Since `Route` is a function type at the outer level, we start with the function-type construction, which requires $\mathbb{1} + R$ to be a monoid. This suggests using the co-product construction; however, R is not necessarily a monoid. There are several possible monoid instances for $\mathbb{1} + R$, so let us look at how the code of `combineRoutes` handles values of that type. The value $\mathbb{1} + \mathbb{0}^R$ corresponds to a route that is not responding to a given path; in that case, `combineRoutes` will switch to the other route. So, the empty value of the monoid $\mathbb{1} + R$ must be $\mathbb{1} + \mathbb{0}^R$. This is indeed

the value returned by `emptyRoute` when applied to any path.

A non-empty value $0 + r$ corresponds to a route that responds to a given path; given two such values, `combineRoutes` will take the first one. This corresponds to the binary operation \oplus_R defined by $r_1 \oplus_R r_2 = r_1$. This operation makes R into a trivial semigroup (see Section 8.3.3). As Example 8.4.1.3 showed, $\mathbb{1} + R$ is a monoid if R is a semigroup (since the unit type $\mathbb{1}$ is a monoid).

So, we have reduced the monoid instance defined by `emptyRoute` and `combineRoutes` to a trivial semigroup, a co-product construction, and a function-type construction. Since all those constructions are guaranteed to produce lawful monoids, we do not need to prove the monoid laws by hand.

Example 8.4.1.5 Using the `cats` library, define a `Functor` instance for the type `Seq[Try[T]]`.

Solution The `cats` library defines the `Functor` typeclass as a `trait` with a `map` method,

```
trait Functor[F[_]] { def map[A, B](fa: F[A])(f: A => B): F[B] }
```

The functor `Seq[Try[T]]` is a composition of functors `Seq` and `Try`, so we can define the `map` method for it using the functor composition (see Statement 6.2.3.6). We wrap `Seq[Try[T]]` into a case class:

```
import cats.Functor
final case class F[T](s: Seq[Try[T]]) // 'type F[T] = Seq[Try[T]]' does not work with 'map' method.
implicit val functorF: Functor[F] = new Functor[F] {
  def map[A, B](fa: F[A])(f: A => B): F[B] = F(fa.s.map(_.map(f)))
}
```

One more `cats`-specific import is necessary to enable the `map` extension method for functors:

```
import cats.syntax.functor._          // Enable the 'map' method.
val s = F(Seq(Try(1), Try(2), Try(3)))

ammonite@ s.map(_ * 10)
res0: F[Int] = F(List(Success(10), Success(20), Success(30)))
```

For the `map` method to work, the type constructor `F` must be defined as a `class` or a `trait`. Defining `F` as a type alias would make the Scala compiler confused: if the type `F[T]` were *the same* as `Seq[Try[T]]`, the expression `Seq(Try(1)).map(_ * 10)` would mean to apply `Seq`'s built-in `map` method rather than the extension method `map` defined for the type constructor `F` via the `Functor` typeclass.

Example 8.4.1.6 Using the `cats` library, implement a `Bifunctor` instance for $Q^{X,Y} \triangleq X + X \times Y$.

Solution The `cats` library defines the `Bifunctor` typeclass as a `trait` with the `bimap` method. Implementing that method as an information-preserving, fully parametric function is straightforward:

```
final case class Q[X, Y](q: Either[X, (X, Y)])
implicit val bifunctorQ = new Bifunctor[Q] {
  def bimap[A, B, C, D](fab: Q[A, B])(f: A => C, g: B => D): Q[C, D] = fab.q match {
    case Left(a)      => Q(Left(f(a)))
    case Right((a, b)) => Q(Right((f(a), g(b))))
  }
}
```

Example 8.4.1.7 Define a `Contrafunctor` typeclass having the method `contramap` (see Section 6.1.7):

```
def contramap[A, B](c: C[A])(f: B => A): C[B]
```

Implement a `Contrafunctor` instance for the type constructor $C^A \triangleq A \rightarrow \text{Int}$.

Solution Since the typeclass method has type parameters, the instance value will have type

$$\text{contramap} : \forall(A, B). C^A \rightarrow (B \rightarrow A) \rightarrow C^B \quad .$$

So, the typeclass needs to be implemented as a `trait`:

```
trait Contrafunctor[C[_]] { def contramap[A, B](c: C[A])(f: B => A): C[B] }
```

A typeclass instance for the type constructor $C^A \triangleq A \rightarrow \text{Int}$ is created by

```
type C[A] = A => Int
```

```
implicit val contrafunctorC = new Contrafunctor[C] {
  def contramap[A, B](c: A => Int)(f: B => A): B => Int = f andThen c
}
```

The cats library defines an equivalent typeclass named `Contravariant` with the method `contramap`.

Example 8.4.1.8 Define a `Functor` instance for recursive type constructor $Q^A \triangleq (\text{Int} \rightarrow A) + \text{Int} + Q^A$.

Solution Begin by defining Q^A as a recursive disjunctive type:

```
sealed trait Q[A]
final case class C1[A](i: Int => A) extends Q[A]
final case class C2[A](x: Int) extends Q[A]
final case class C3[A](q: Q[A]) extends Q[A]
```

The methods of Section 6.2.3 show how to implement the `map` function for `Q[A]`. Without repeating those steps, we will write code for a `Functor` instance directly:

```
implicit val functorQ: Functor[Q] = new Functor[Q] { // The function 'map' is recursive.
  def map[A, B](qa: Q[A])(f: A => B): Q[B] = qa match {
    case C1(i) => C1(i andThen f)
    case C2(x) => C2(x)
    case C3(q) => C3(map(q)(f)) // Recursive case.
  }
}
```

Example 8.4.1.9 Using a function parameterized by the type constructors F and G (required to be functors), implement a `Functor` instance for $F^A + G^A$.

Solution The co-product construction (Statement 6.2.3.4) shows how to implement the `map` function for the functor $F^A + G^A$. We begin by writing code for a `Functor` instance assuming that the type constructors `F` and `G` are given:

```
type L[A] = Either[F[A], G[A]]
implicit val functorEither = new Functor[L] {
  def map[A, B](e: L[A])(f: A => B): L[B] = e match {
    case Left(fa) => Left(fa.map(f))
    case Right(ga) => Right(ga.map(f))
  }
}
```

We will now rewrite this code by making `F` and `G` into type parameters. To achieve that, we need to use the “kind projector” plugin and replace the type constructor `L` by a nameless type function `Lambda[X => Either[F[X], G[X]]]`. The code becomes

```
implicit def functorEither[F[_], G[_]] = new Functor[Lambda[X => Either[F[X], G[X]]]] {
  type L[A] = Either[F[A], G[A]] // We may use F and G to define a type alias in this scope.
  def map[A, B](e: L[A])(f: A => B): L[B] = e match {
    case Left(fa) => Left(fa.map(f))
    case Right(ga) => Right(ga.map(f))
  }
}
```

Example 8.4.1.10* (a) Implement a function with type signature $C^A + C^B \rightarrow C^{A \times B}$ parameterized by a type constructor C (required to be a contrafunctor) and by arbitrary types A, B . Show that the inverse type signature $C^{A \times B} \rightarrow C^A + C^B$ is not implementable for some contrafunctors C .

(b) Implement a function with type signature $F^{A \times B} \rightarrow F^A \times F^B$ parameterized by a type constructor F (required to be a functor) and by arbitrary types A, B . Show that the inverse type signature $F^A \times F^B \rightarrow F^{A \times B}$ is not implementable for some functors F .

Solution (a) We need to implement a function with type signature

$$\forall(A, B). C^A + C^B \rightarrow C^{A \times B} .$$

Begin by looking at the types involved. We need to relate values $C^{A \times B}$, C^A , and C^B ; can we relate $A \times B$, A , and B ? There exist unique fully parametric functions π_1 and π_2 of types $A \times B \rightarrow A$ and

$A \times B \rightarrow B$. If we lift these functions to the contrafunctor C , we will get $\pi_1^{\downarrow C} : C^A \rightarrow C^{A \times B}$ and $\pi_2^{\downarrow C} : C^B \rightarrow C^{A \times B}$. The required type signature is then implemented via a `match` expression like this:

The code notation for this function is

```
def f[C[_]: Contrafunctor, A, B]
  : Either[C[A], C[B]] => C[(A, B)] = {
  case Left(ca)  => ca.contramap { case (a, b) => a }
  case Right(cb)  => cb.contramap { case (a, b) => b }
}
```

$$f : C^A + C^B \rightarrow C^{A \times B} \triangleq \begin{array}{c|c|c} & & C^{A \times B} \\ \hline C^A & \pi_1^{\downarrow C} & \\ \hline C^B & \pi_2^{\downarrow C} & \end{array} .$$

To show that it is not possible to implement a function g with the inverse type signature,

$$g : \forall(A, B). C^{A \times B} \rightarrow C^A + C^B ,$$

we choose the contrafunctor $C^A \triangleq A \rightarrow R$, where R is a fixed type. The type signature becomes

$$g : \forall(A, B). (A \times B \rightarrow R) \rightarrow (A \rightarrow R) + (B \rightarrow R) .$$

To implement this function, we need to decide whether to return values of type $A \rightarrow R$ or $B \rightarrow R$. Can we compute a value of type $A \rightarrow R$ given a value of type $A \times B \rightarrow R$?

$$g : (A \times B \rightarrow R) \rightarrow A \rightarrow R \triangleq q : A \times B \rightarrow R \rightarrow a : A \rightarrow ??? : R$$

We cannot compute a value of type R because that requires us to apply the function q to a pair $A \times B$, while we only have a value of type A . So, the typed hole $??? : R$ cannot be filled.

Similarly, we are not able to compute a value of type $B \rightarrow R$ from a value of type $A \times B \rightarrow R$. Whatever choice we make, $A \rightarrow R$ or $B \rightarrow R$, we cannot implement the required type signature.

(b) We need to implement a function with type signature

$$\forall(A, B). F^{A \times B} \rightarrow F^A \times F^B .$$

The types $A \times B$, A , and B are related by the functions $\pi_1 : A \times B \rightarrow A$ and $\pi_2 : A \times B \rightarrow B$. Lifting these functions to the functor F , we obtain $\pi_1^{\uparrow F} : F^{A \times B} \rightarrow F^A$ and $\pi_2^{\uparrow F} : F^{A \times B} \rightarrow F^B$. It remains to take the product of the resulting values:

$$f : F^{A \times B} \rightarrow F^A \times F^B \triangleq p : F^{A \times B} \rightarrow (p \triangleright \pi_1^{\uparrow F}) \times (p \triangleright \pi_2^{\uparrow F}) .$$

```
def f[F[_]: Functor, A, B](p: F[(A, B)]): (F[A], F[B]) =
  (p.map { case (a, b) => a }, p.map { case (a, b) => b }) // Or (p.map(_._1), p.map(_._2))
```

A shorter code for f via the “diagonal” function $\Delta \triangleq (q \circ \Delta \rightarrow q \times q)$ and the pair product \boxtimes is

$$f : F^{A \times B} \rightarrow F^A \times F^B \triangleq \Delta \circ (\pi_1^{\uparrow F} \boxtimes \pi_2^{\uparrow F}) .$$

This notation is sometimes easier to reason about when deriving properties of functions.

If we try implementing a function g with the inverse type signature,

$$g : \forall(A, B). F^A \times F^B \rightarrow F^{A \times B} , \tag{8.16}$$

we will find that g *can* be implemented for functors such as $F^A \triangleq A \times A$, $F^A \triangleq \mathbb{1} + A$, and $F^A \triangleq P \rightarrow A$. It is not obvious how to find a functor F for which the function g has no implementation. By looking through the known functor constructions (Table 6.2) and trying various combinations, we eventually find a suitable functor: $F^A \triangleq (P \rightarrow A) + (Q \rightarrow A)$. The type signature of g becomes

$$g : \forall(A, B). ((P \rightarrow A) + (Q \rightarrow A)) \times ((P \rightarrow B) + (Q \rightarrow B)) \rightarrow (P \rightarrow A \times B) + (Q \rightarrow A \times B) .$$

The argument of this function is of type

$$((P \rightarrow A) + (Q \rightarrow A)) \times ((P \rightarrow B) + (Q \rightarrow B)) ,$$

which can be transformed equivalently into a disjunction of four cases,

$$(P \rightarrow A) \times (P \rightarrow B) + (P \rightarrow A) \times (Q \rightarrow B) + (Q \rightarrow A) \times (P \rightarrow B) + (Q \rightarrow A) \times (Q \rightarrow B) .$$

Implementing the function g requires, in particular, to handle the case when we are given values of types $P \rightarrow A$ and $Q \rightarrow B$, and we are required to produce a value of type $(P \rightarrow A \times B) + (Q \rightarrow A \times B)$. The resulting type signature

$$(P \rightarrow A) \times (Q \rightarrow B) \rightarrow (P \rightarrow A \times B) + (Q \rightarrow A \times B)$$

cannot be implemented: If we choose to return a value of type $P \rightarrow A \times B$, we would need to produce a pair of type $A \times B$ from a value of type P . However, producing a pair $A \times B$ requires, in this case, to have values of *both* types P and Q , since the given arguments have types $P \rightarrow A$ and $Q \rightarrow B$. Similarly, we cannot return a value of type $Q \rightarrow A \times B$.

We find that the function g cannot be implemented for the functor $F^A \triangleq (P \rightarrow A) + (Q \rightarrow A)$. Function F for which the type signature (8.16) *can* be implemented are called “applicative” (see Chapter 11 for precise conditions). The functor F is an example of a functor that is not applicative.

Example 8.4.1.11* (R. O’Connor⁷) Assume that a functor F admits a function p with type signature

$$\text{def } p[A, B, F[_]: \text{Functor}] : \text{Either}[A, F[B]] \Rightarrow F[\text{Either}[A, B]] \quad p^{A,B} : A + F^B \rightarrow F^{A+B} ,$$

additionally satisfying the special laws of identity and associativity,

$$p^{0,B} = (b^B \rightarrow 0 + b)^{\uparrow F} , \quad p^{A+B,C} = \begin{array}{c|c|c} & A & F^{B+C} \\ \hline A & \text{id} & 0 \\ B + F^C & 0 & p^{B,C} \end{array} ; p^{A,B+C} .$$

Show that the functor F is pointed if such a function p exists. Conversely, show that any pointed functor F admits a function p with these properties.

Solution To show that F is pointed, it is sufficient to find a value wu_F of type F^1 . We note that the given function p can create values of type F^{A+B} from input values of type $A + 0$. So, we set the type parameters $A = 1$ and $B = 0$ and apply p to the value $1 + 0^{F^0}$,

$$\text{wu}_F \triangleq (1 + 0^{F^0}) \triangleright p^{1,0} \triangleright (1 + 0 \rightarrow 1)^{\uparrow F} .$$

$$\text{val wu}[F[_]: \text{Functor}]: F[\text{Unit}] = p[\text{Unit}, \text{Nothing}](\text{Left}(())) \mapsto \text{Left}(\text{wu}_F)$$

Conversely, assuming that F is pointed, we use its pu_F function to define p as

$$\text{def } p[F[_]: \text{Functor} : \text{Pointed}, A, B] \\ : \text{Either}[A, F[B]] \Rightarrow F[\text{Either}[A, B]] = \{ \\ \text{case Left}(a) \Rightarrow \text{pure}[F, \text{Either}[A, B]](\text{Left}(a)) \\ \text{case Right}(fb) \Rightarrow fb \mapsto \text{Right}[A, B](b) \} \\ \}$$

$$p^{A,B} \triangleq \begin{array}{c|c|c} & & F^{A+B} \\ \hline A & (a^A \rightarrow a + 0^B) ; \text{pu}_F \\ F^B & (b^B \rightarrow 0^A + b)^{\uparrow F} \end{array} .$$

It remains to show that p satisfies the required laws. The identity law holds because

$$\text{expect to equal } (b^B \rightarrow 0 + b)^{\uparrow F} : \quad p^{0,B} = \begin{array}{c|c|c} & & F^{0+B} \\ \hline 0 & (\text{we may delete this line}) \\ F^B & (b^B \rightarrow 0 + b)^{\uparrow F} \end{array} = (b^B \rightarrow 0 + b)^{\uparrow F} .$$

⁷This example is based on the post <https://mail.haskell.org/pipermail/haskell-cafe/2015-November/122357.html>

To verify the associativity law, we begin with its right-hand side since it is more complicated:

$$\begin{array}{c}
 \begin{array}{c|c|c}
 & A & F^{B+C} \\ \hline
 A & \text{id} & \emptyset \\ \hline
 B+F^C & \emptyset & p^{B,C}
 \end{array} ; p^{A,B+C} = \begin{array}{c|c|c}
 & A & F^{B+C} \\ \hline
 A & \text{id} & \emptyset \\ \hline
 B+F^C & \emptyset & p^{B,C}
 \end{array} ; \begin{array}{c|c}
 & F^{A+B+C} \\ \hline
 A & (a:A \rightarrow a + \emptyset^{B+C}) ; \text{pu}_F \\ \hline
 F^{B+C} & (x^{B+C} \rightarrow \emptyset^A + x) \uparrow F
 \end{array} \\ \\
 = \begin{array}{c|c}
 & F^{A+B+C} \\ \hline
 A & (a:A \rightarrow a + \emptyset^{B+C}) ; \text{pu}_F \\ \hline
 B+F^C & p^{B,C} ; (x^{B+C} \rightarrow \emptyset^A + x) \uparrow F
 \end{array} = \begin{array}{c|c}
 & F^{A+B+C} \\ \hline
 A & (a:A \rightarrow a + \emptyset^{B+C}) ; \text{pu}_F \\ \hline
 B & (b:B \rightarrow b + \emptyset^C) ; \text{pu}_F ; (x^{B+C} \rightarrow \emptyset^A + x) \uparrow F \\ \hline
 F^C & (c:C \rightarrow \emptyset^{A+B} + c) \uparrow F
 \end{array} .
 \end{array}$$

In the last line, we have expanded the type matrix to three rows corresponding to the disjunctive type $A + B + F^C$. We need to show that the last matrix equals $p^{A+B,C}$; so let us rewrite $p^{A+B,C}$ as a similarly expanded type matrix, using the type isomorphisms such as $\emptyset^A + \emptyset^B \cong \emptyset^{A+B}$:

$$p^{A+B,C} = \begin{array}{c|c}
 & F^{A+B+C} \\ \hline
 A & (a:A \rightarrow a + \emptyset^B + \emptyset^C) ; \text{pu}_F \\ \hline
 B & (b:B \rightarrow \emptyset^A + b + \emptyset^C) ; \text{pu}_F \\ \hline
 F^C & (c:C \rightarrow \emptyset^A + \emptyset^B + c) \uparrow F
 \end{array} = \begin{array}{c|c}
 & F^{A+B+C} \\ \hline
 A & (a:A \rightarrow a + \emptyset^{B+C}) ; \text{pu}_F \\ \hline
 B & (b:B \rightarrow \emptyset^A + b + \emptyset^C) ; \text{pu}_F \\ \hline
 F^C & (c:C \rightarrow \emptyset^{A+B} + c) \uparrow F
 \end{array} .$$

The only remaining difference is in the second lines of the matrices. We write those lines separately:

$$\begin{array}{l}
 \text{expect to equal } (b:B \rightarrow \emptyset^A + b + \emptyset^C) ; \text{pu}_F : \quad (b:B \rightarrow b + \emptyset^C) ; \text{pu}_F ; (x^{B+C} \rightarrow \emptyset^A + x) \uparrow F \\
 \text{naturality law of } \text{pu}_F : \quad = (b:B \rightarrow b + \emptyset^C) ; (x^{B+C} \rightarrow \emptyset^A + x) ; \text{pu}_F \\
 \text{compute function composition :} \quad = (b:B \rightarrow \emptyset^A + b + \emptyset^C) ; \text{pu}_F .
 \end{array}$$

This completes the proof of the required laws.

8.4.2 Exercises

Exercise 8.4.2.1 Define a PTVF `def isLong[T]: Boolean` that returns `true` for `T = Long` or `Double` and returns `false` for `T = Int, Short, or Float`. The function should remain undefined for other types `T`.

Exercise 8.4.2.2 Implement a `Monoid` instance for the type `String × (1 + Int)`.

Exercise 8.4.2.3 (a) If A is a monoid and R any type, implement a `Monoid` instance for $R \rightarrow A$:

```
def monoidFunc[A: Monoid, R]: Monoid[R => A] = ???
```

Prove that the monoid laws hold for that instance.

(b) With the choice `R = Boolean`, use the type equivalence $(R \rightarrow A) = (2 \rightarrow A) \cong A \times A$ and verify that the monoid instance `monoidFunc[A, Boolean]` is the same as the monoid instance for $A \times A$ computed by `monoidPair[A, A]` in Section 8.3.4.

Exercise 8.4.2.4 Show that if s is a semigroup then `Option[S]` and `Option[(S, S)]` are monoids.

Exercise 8.4.2.5 A framework implements a “route” type R as $R \triangleq Q \rightarrow (E + S)$, where Q is a query, E is an error response, and S is a success response. A server is defined as a combination of several routes. For a given query Q , the response is the first route (if it exists) that yields a success response. Implement the route combination operation and show that it makes R into a semigroup. What would be necessary to make R into a monoid?

Exercise 8.4.2.6 Using the `cats` library, implement a `Functor` instance for `type F[T] = Try[Seq[T]]`.

Exercise 8.4.2.7 Using the `cats` library, implement a `Bifunctor` instance for $B^{X,Y} \triangleq (\text{Int} \rightarrow X) + Y \times Y$.

Exercise 8.4.2.8 Define a `Profunctor` typeclass having the method `xmap`:

```
def xmap[A, B](f: A => B, g: B => A): F[A] => F[B]
```

Implement a `Profunctor` instance for $P^A \triangleq A \rightarrow (\text{Int} \times A)$.

Exercise 8.4.2.9 Implement a `Functor` instance for the recursive type $Q^A \triangleq \text{String} + A \times A \times Q^A$.

Exercise 8.4.2.10 Show explicitly that a value $wu_C : C^1$ is computationally equivalent to a value $pu_C : \forall A. C^A$ that satisfies the naturality law (8.14).

Exercise 8.4.2.11 Using a function parameterized by the type constructors F and G (required to be functors), implement a `Functor` instance for $F^A \times G^A$.

Exercise 8.4.2.12 Implement a `Functor` instance for $F^A \rightarrow G^A$ as a function parameterized by type constructors F and G , where F^A is required to be a contrafunctor and G^A is required to be a functor. For the contrafunctor F , use either the `Contrafunctor` typeclass from Example 8.4.1.7 or the `cats` library's typeclass `Contravariant`.

Exercise 8.4.2.13* (a) Implement a function with type signature $F^A + F^B \rightarrow F^{A+B}$ parameterized by a type constructor F (required to be a functor) and by arbitrary types A, B . Show that the inverse type signature $F^{A+B} \rightarrow F^A + F^B$ is not implementable for some functors F .

(b) Implement a function with type signature $C^{A+B} \rightarrow C^A \times C^B$ parameterized by a type constructor C (required to be a contrafunctor) and by arbitrary types A, B . Show that the inverse type signature $C^A \times C^B \rightarrow C^{A+B}$ is not implementable for some contrafunctors C .

Exercise 8.4.2.14* Implement a function with type signature $F^{A \rightarrow B} \rightarrow A \rightarrow F^B$ parameterized by a functor F and arbitrary types A, B . Show that the inverse type signature, $(A \rightarrow F^B) \rightarrow F^{A \rightarrow B}$, cannot be implemented for some functors F . (Functors admitting a function with that type signature are called “rigid”; see Section 14.5.5.)

Exercise 8.4.2.15* (dual O’Connor) Assume that a functor F admits a function q with type signature

```
def q[A, B, F[_]: Functor]: F[(A, B)] => (A, F[B])
```

$q^{A,B} : F^{A \times B} \rightarrow A \times F^B$, additionally satisfying the special laws of identity and associativity,

$$q^{1,B} = f^{F^{1 \times B}} \rightarrow 1 \times (f \triangleright (1 \times b^B \rightarrow b)^{\uparrow F}) \quad , \quad q^{A,B \times C} \circ (id^A \boxtimes q^{B,C}) = q^{A \times B,C} \quad .$$

This Scala code illustrates the required laws in full detail:

```
// For any value f: F[(Unit, B)], we first compute
val (a1, b1) = q[Unit, B, F](f)._1
// Then we must have: a1 == () and b1 == f.map { case (_, b) => b }

// For any value g: F[(A, B, C)], we first compute
val ((a1: A, b1: B), f1: F[C]) = q[(A, B), C, F](g.map { case (a, b, c) => ((a, b), c) })
val (a2: A, f2: F[(B, C)]) = q[A, (B, C), F](g.map { case (a, b, c) => (a, (b, c)) })
// Then we must have: a1 == a2 and (b1, f1) == g[B, C, F](f2)
```

The following naturality law should also hold for q ,

$$(f^{A \rightarrow C} \boxtimes g^{B \rightarrow D})^{\uparrow F} \circ q^{C,D} = q^{A,B} \circ f \boxtimes (g^{\uparrow F}) \quad .$$

```
// For any value k: F[(A, B)] and any functions f: A => C, g: B => D, first compute
val (a: A, fb: F[B]) = q[A, B, F](k) // Then we must have:
(f(a), fb.map(g)) == q[C, D, F](k.map { case (a, b) => (f(a), g(b)) })
```

Show that the functor F is co-pointed if a function q with these properties exists. Conversely, show that any lawful co-pointed functor F admits a function q with these properties.

8.5 Further developments

8.5.1 The existence of values for recursive types

A recursive type T is defined by a type equation such as $T \triangleq 1 + \text{Int} + T \times T$. Can we decide whether such a type equation has a “solution”, i.e., a well-defined type T with values that we can create and manipulate in a program?

In all examples seen so far, the recursive type equations had the form $T \triangleq S^T$ where the type constructor S is a *functor*. Type equations with non-functor S (e.g., the equation $T \triangleq T \rightarrow \text{Int}$) do not seem to be useful in practice, and we will not consider them in this book.

In a rigorous approach, showing that T is a “solution” (called a **fixpoint**) of the type equation $T \triangleq S^T$ means proving that the types T and S^T are equivalent (isomorphic). We must implement this type isomorphism as two functions, named e.g., `fix` and `unfix`, satisfying the conditions

$$\text{fix} : S^T \rightarrow T \quad , \quad \text{unfix} : T \rightarrow S^T \quad , \quad \text{fix} ; \text{unfix} = \text{id} \quad , \quad \text{unfix} ; \text{fix} = \text{id} \quad .$$

Given a type constructor S , we can define the recursive type T with this Scala code,

```
final case class T(s: S[T]) // Type constructor S[_] must be already defined.
def fix: S[T] => T = { s => T(s) }
def unfix: T => S[T] = { t => t.s }
```

We can generalize this code to a “fixpoint” constructor Fix^S that is *parameterized* by S :

```
final case class Fix[S[_]](s: S[Fix[S]]) // Type constructor S[_] must be already defined.
def fix[S[_]]: S[Fix[S]] => Fix[S] = { s => Fix(s) }
def unfix[S[_]]: Fix[S] => S[Fix[S]] = { t => t.s }
```

In both implementations, the functions `fix` and `unfix` are inverses of each other because they merely wrap and unwrap values within a case class. So, we are always able to write code that *defines* the recursive type T . The remaining question is whether we will be able to create values of type T .

Applying structural analysis to the functor S , we begin with the constant functor $S^A \triangleq Z$, where Z is a fixed type. The type equation $T \triangleq S^T$ is just $T \triangleq Z$; so T is not a recursive type.

The next case is the identity functor $S^A \triangleq A$. The type equation $T \triangleq S^T$ has the form $T \triangleq T$; since all types T satisfy that equation trivially, we find that the identity functor does not define any specific type T . If we translate this type equation into Scala code, we will run into a problem with recursion:

```
type Id[A] = A
final case class T(s: Id[T])
// Equivalent to 'final case class T(s: T)'.
val x = T(x) // Infinite loop!
```

A value of `case class T` can be created only if we supply a value of type T . Writing `def x: T = T(x)` instead of `val` does not help: the evaluation of the code for x will not terminate. It is clear that this recursive definition is invalid: we are unable to create values of the resulting type T , i.e., the type T is void.

Next, consider a product functor such as $S^A \triangleq A \times A \times \text{Int}$. Can we create values of type $T \triangleq S^T$?

```
final case class T(s: (T, T, Int))
val x = T(s = (x, x, 123)) // Infinite loop!
```

We again have an infinite loop when creating values of type T . As in the example with the identity functor, the case class T requires us to compute some values of type T before we can create a value of type T . That requirement is impossible to satisfy.

A disjunctive type, e.g., $S^A \triangleq \text{Int} + A \times A$, allows us to create values of type Fix^S with no difficulty:

```
final case class T(s: Either[Int, (T, T)])
val x: T = T(Left(123)) // OK
val y: T = T(Right(x, x)) // OK
```

We are able to create x of type $\text{Int} + 0:T \times T$ without need for any previous values of T . We can then use x to create a value y of type T . This resembles defining a value by induction: the base case is the type $\text{Int} + 0:T \times T$, which is a disjunctive part of S^A that does *not* contain any values of type A . The inductive step is the type $0:\text{Int} + T \times T$, which creates a new value of type T from two previous values. Type recursion terminates when the base case exists.

The examples we saw previously, $S^A \triangleq A$ and $S^A \triangleq \text{Int} \times A \times A$, do not have a base case where a value of type S^A could be computed without need for any previous values of type A .

Given a functor S^A , how can we determine whether the type recursion $T \triangleq S^T$ terminates? If S^A is a *polynomial* functor, we can view S^A as a polynomial function of A and reduce it to the form

$$S^A \cong C_0 + C_1 \times A + C_2 \times A \times A + \dots \quad (8.17)$$

It is clear from this formula that the type C_0 is found by setting $A = \mathbb{0}$ in S^A , i.e., $C_0 \cong S^0$. If we can compute a value c_0 of type C_0 (i.e., if C_0 is not void, $C_0 \not\cong \mathbb{0}$), a base-case value t_0 of type T is found as

$$t_0^T \triangleq c_0^{C_0} + \mathbb{0}^{C_1 \times A} + \mathbb{0}^{C_2 \times A \times A} + \dots$$

Further values of type T could be computed as $c_1 \times t_0$ or $c_2 \times t_0 \times t_0$, etc.

We conclude that the recursive type $T \triangleq S^T$ defined via a *polynomial* functor S will be non-void (i.e., type recursion will terminate) when the type S^0 is non-void.

In the examples $S^A \triangleq A$ and $S^A \triangleq \text{Int} \times A \times A$, we find $S^0 \cong \mathbb{0}$, so type recursion is invalid.

It remains to consider exponential functors S that cannot be reduced to the polynomial form (8.17). As an example, take $S^A \triangleq \text{String} \times (\text{Int} \rightarrow A)$. For this functor S , the condition $S^0 \not\cong \mathbb{0}$ does not hold:

$$S^0 \cong \text{String} \times (\text{Int} \rightarrow \mathbb{0})$$

use the type equivalence $(\text{Int} \rightarrow \mathbb{0}) \cong \mathbb{0}$: $\cong \text{String} \times \mathbb{0} \cong \mathbb{0}$.

Nevertheless, we can write Scala code implementing values of the type T defined by $T \triangleq S^T$:

```
final case class T(message: String, next: Int => T)
val t: T = T("start", n => T(s"have $n", _ => t)) // A recursive 'val'.
```

The value t refers to itself recursively within a nested function $_ \Rightarrow t$, and there is no “base case” in the type S^T . However, the recursion does not lead to an infinite loop; we can use t safely:

```
scala> t.next(10)
res0: T = T(have 10,<function1>)

scala> t.next(10).next(10)
res1: T = T(start,<function1>)
```

A value of type T contains a function `next` that, when applied, returns a new value of the same type T . The new value of type T does not need to be computed in advance; its evaluation is *delayed* until some code decides to call `next`. For this reason, an infinite loop is avoided even though the structure functor S has no “base case”. Values of type T can be viewed as an infinite stream of `String` values computed on demand; an `Int` value is required in order to produce the next element of the stream. A symbolic (and non-rigorous) representation of that type is

$$T = \text{String} \times (\text{Int} \rightarrow \text{String} \times (\text{Int} \rightarrow \text{String} \times (\text{Int} \rightarrow \text{String} \times \dots))) .$$

As another example, consider $S^A \triangleq \mathbb{1} \rightarrow \text{String} + \text{Int} \times A$. Using the type equivalence $P \cong (\mathbb{1} \rightarrow P)$, we could transform S^A into an equivalent functor \tilde{S}^A in the polynomial form (8.17),

$$\tilde{S}^A \triangleq \text{String} + \text{Int} \times A .$$

Although the types S^A and \tilde{S}^A are equivalent, the recursive types Fix^S and $\text{Fix}^{\tilde{S}}$ are different. While

$$\begin{aligned} \text{Fix}^{\tilde{S}} &= \text{String} + \text{Int} \times (\text{String} + \text{Int} \times (\text{String} + \text{Int} \times \dots)) \\ &\cong (\mathbb{1} + \text{Int} + \text{Int} \times \text{Int} + \text{Int} \times \text{Int} \times \text{Int} + \dots) \times \text{String} \cong \text{String} \times \text{List}^{\text{Int}} , \end{aligned}$$

the type Fix^S admits recursively defined values representing *unbounded* streams of integers, e.g.

```
final case class T(e: () => Either[String, (Int, T)]) // Define the type T \triangleq \text{String} + \text{Int} \times T .
val t1: T = T(() => Right((1, t1))) // Stream [1, 1, 1, ...].
def t2(n: Int): T = T(() => Right((n, t2(n+1)))) // Stream [n, n+1, n+2, ...].
```

The type $T \triangleq \text{Fix}^S$ also admits finite streams that may be defined as *non-recursive* values of type T :

```
val t0 = T(() => Right((10, T(() => Left("stop"))))) // Finite stream [10, "stop"].
```

We can recognize that Fix^S has non-recursive values by checking that the type S^0 is not void:

$$S^0 = \mathbb{1} \rightarrow \text{Int} \times \emptyset + \text{String} \cong \mathbb{1} \rightarrow \text{String} \cong \text{String} \not\cong \emptyset .$$

How can we recognize functors S that admit valid recursive values of type Fix^S ? The exponential functor construction (Statement 6.2.3.5) shows that $C^A \rightarrow P^A$ is a functor when C is a contrafunctor and P is a functor. If the type expression for S^A contains a sub-expression of the form $C^A \rightarrow P^A$, how can we implement a value of type $C^T \rightarrow P^T$? We can use recursion to implement a value of T , as long as we can then somehow produce a value of type P^T out of a value of type T . This is precisely the condition for P^\bullet to be a pointed functor. The contrafunctor C^A is an argument of the function of type $C^A \rightarrow P^A$, so we are not required to produce values of type C^A —we *consume* those values. It follows that we can implement a value of type $C^T \rightarrow P^T$ (with $T = \text{Fix}^S$) as long as P^\bullet is a pointed functor. As we saw in Section 8.3.5, a functor P^\bullet is pointed if we can compute a value of type $P^\mathbb{1}$.

In the example $S^A \triangleq \text{String} \times (\text{Int} \rightarrow A)$, the functor S^A is pointed since

$$S^\mathbb{1} \cong \text{String} \times (\text{Int} \rightarrow \mathbb{1}) \cong \text{String} \not\cong \emptyset .$$

This consideration applies to any sub-expression of the form $C^A \rightarrow P^A$ within the type constructor S^A . The condition for values of Fix^S to exist is that every functor P^A involved in such sub-expressions should be pointed. To check that, we can set the type parameter $A = \mathbb{1}$ in all return types of *functions* within S^A . If the resulting type is not void, we will be able to implement a recursively defined value of type Fix^S .

If the functor S^A has the property $S^0 \not\cong \emptyset$ (i.e., we have a base case for the inductive definition of Fix^S), we will also be able to implement non-recursive values of type Fix^S .

We conclude that $S^0 \not\cong \emptyset$ is a *sufficient* condition for the type Fix^S to be non-void. If S is a polynomial functor, this condition is also a necessary condition. For exponential-polynomial functors S , recursive values of type Fix^S can be implemented if every function-type sub-expression $C^A \rightarrow P^A$ in S^A involves a pointed functor P (i.e., if $P^\mathbb{1} \not\cong \emptyset$).

8.5.2 Proofs of associativity of concat for lists and arrays

The `concat` function is defined for both lists and arrays, and works similarly for these data types:

```
scala> Array.concat(Array(1, 2), Array(3, 4), Array(5, 6))
res0: Array[Int] = Array(1, 2, 3, 4, 5, 6)

scala> List.concat(List(1, 2), List(3, 4), List(5, 6))
res1: List[Int] = List(1, 2, 3, 4, 5, 6)
```

In this section, we will show rigorously that concatenation is an associative operation.

In Scala, `Array[A]` is a sequence whose elements can be accessed by index. The array access function (the `apply` method defined on the `Array` class) is a partial function of type `Int => A` whose integer argument must be between 0 and $n - 1$, where n is the array length:

```
val x = Array("a", "b", "c")

scala> x(2) // The syntax 'x(2)' is the same as 'x.apply(2)'.
res2: String = c

scala> x(3) // Applying the partial function 'x.apply' to the value 3 will fail:
java.lang.ArrayIndexOutOfBoundsException: 3
```

We can denote the type of this function by $\text{Int}_{[0, n-1]} \rightarrow A$ to indicate the bounds of the index.

The `List` type constructor is a recursive disjunctive type defined by the type equation

$$\text{List}^A \triangleq \mathbb{1} + A \times \text{List}^A = \mathbb{1} + A + A \times A + A \times A \times A + \dots$$

Although the definitions of types are different, lists and arrays are conceptually similar — they are both sequences of values of type A . However, proving properties is easier for non-recursive types than for recursive types. We begin by proving the associativity property of the array concatenation.

Statement 8.5.2.1 For arrays, $\text{Array}_n^A \triangleq \text{Int}_{[0,n-1]} \rightarrow A$, the `concat` function (denoted `++`) defined by

$$a_1 : \text{Array}_{n_1}^A \quad a_2 : \text{Array}_{n_2}^A \triangleq i : \text{Int}_{[0,n_1+n_2-1]} \rightarrow \begin{cases} 0 \leq i < n_1 : & a_1(i) \\ n_1 \leq i < n_1 + n_2 : & a_2(i - n_1) \end{cases}$$

satisfies the associativity law

$$(a_1 \text{++} a_2) \text{++} a_3 = a_1 \text{++} (a_2 \text{++} a_3) .$$

Proof Both sides of the law evaluate to the same partial function of type $\text{Int} \rightarrow A$:

$$(a_1 \text{++} a_2) \text{++} a_3 = a_1 \text{++} (a_2 \text{++} a_3) = i : \text{Int}_{[0,n_1+n_2+n_3-1]} \rightarrow \begin{cases} 0 \leq i < n_1 : & a_1(i) \\ n_1 \leq i < n_1 + n_2 : & a_2(i - n_1) \\ n_1 + n_2 \leq i < n_1 + n_2 + n_3 : & a_3(i - n_1 - n_2) \end{cases} .$$

Now we establish a rigorous equivalence between the `List` and `Array` types.

Statement 8.5.2.2 The type $\text{List}^A \triangleq \mathbb{1} + A \times \text{List}^A$ is equivalent to $\text{Array}_n^A \triangleq \text{Int}_{[0,n-1]} \rightarrow A$, where $n \geq 0$ and $\text{Int}_{[0,n-1]}$ is the (possibly empty) subset of integers i within the range $0 \leq i \leq n - 1$.

Proof We need to implement two isomorphism maps f_1, f_2 and show that

$$f_1 : \text{Array}_n^A \rightarrow \text{List}^A , \quad f_2 : \text{List}^A \rightarrow \text{Array}_n^A , \quad f_1 \circ f_2 = \text{id} , \quad f_2 \circ f_1 = \text{id} .$$

To implement f_1 , we proceed by induction in n . The base case is $n = 0$, and we map Array_0 into an empty list. The inductive step assumes that f_1 is already defined on arrays of length n , and we now need to define f_1 for arrays of length $n + 1$. An array of length $n + 1$ is a partial function $g : \text{Int}_{[0,n]} \rightarrow A$ defined on the integer interval $[0, n]$. We now split that array into its first element, $g(0)$, and the rest of the array, which needs to be represented by another partial function, say $g' \triangleq i \rightarrow g(i + 1)$, defined on the integer interval $[0, n - 1]$. The function g represents an array of length n . By the inductive assumption, f_1 is already defined for arrays of length n . So, we can compute $f_1(g') : \text{List}^A$ and thus create a value of type $\mathbb{0} + A \times \text{List}^A$, which is equivalent to a value of type List^A .

```
def f1[A](arr: Array[A]): List[A] =
  if (arr.length == 0) List()
  else arr(0) :: f1(arr.tail)
```

$$f_1(\text{Array}_0^A) \triangleq \mathbb{1} + \mathbb{0} : A \times \text{List}^A ,$$

$$f_1(g : \text{Int}_{[0,n]} \rightarrow A) \triangleq \mathbb{0} + g(0) : A \times \overline{f_1}(i \rightarrow g(i + 1)) .$$

To implement f_2 , we use induction in the structure of the list. The length n of the array is not known in advance and needs to be computed as we perform pattern-matching on the given `List[A]` value. The base case is an empty list, which yields an empty array (i.e., an array of length $n = 0$). In the inductive step, we assume that we already defined f_2 on lists of length n , and we now need to define f_2 for lists of length $n + 1$. Such a list must have the form $\mathbb{0} + x : A \times s : \text{List}^A$, where s is a list of length n . By the inductive assumption, we are allowed to apply f_2 to s and obtain an array of length n , i.e., a partial function $g : \text{Int}_{[0,n-1]} \rightarrow A$. So we define $f_2(\mathbb{0} + x \times s)$ as a new array whose 0^{th} element is x and the i^{th} element is computed by applying the function $g \triangleq f_2(s)$ to $i - 1$:

```
def f2[A: ClassTag]: List[A] => Array[A] =
  { case List() => Array()
    case x :: s => Array(x) ++ f2.apply(s)
  } // This code is for illustration only.
```

$$f_2(\mathbb{1} + \mathbb{0} : A \times \text{List}^A) \triangleq \text{Array}_0^A ,$$

$$f_2(\mathbb{0} + x : A \times s : \text{List}^A) \triangleq i : \text{Int}_{[0,n]} \rightarrow \begin{cases} i = 0 : & x \\ i \geq 1 : & \overline{f_2}(s)(i - 1) \end{cases} .$$

To show that f_1 and f_2 are inverse functions for each other, we again need to use induction. The base case is the empty list and the empty array, which are indeed mapped isomorphically to each

other. The inductive step for $f_1 \circ f_2$ is an array of length $n + 1$ with the inductive assumption that $\overline{f_1 \circ f_2} = \text{id}$ for arrays of length n . Writing out the code of $f_1 \circ f_2$, we find that $g \triangleright f_1 \circ f_2 = g$:

$$\begin{aligned} g^{\text{Int}_{[0,n]} \rightarrow A} \triangleright f_1 \circ f_2 &= g \triangleright f_1 \triangleright f_2 = (\mathbb{0} + g(0) \times \overline{f_1}(i \rightarrow g(i + 1))) \triangleright f_2 \\ &= i \rightarrow \begin{cases} i = 0 : & g(0) \\ i \geq 1 : & \overline{f_2}(\overline{f_1}(i \rightarrow g(i + 1))(i - 1)) \end{cases} = i \rightarrow \begin{cases} i = 0 : & g(0) \\ i \geq 1 : & \text{id}(i \rightarrow g(i + 1))(i - 1) \end{cases} \\ &= i \rightarrow \begin{cases} i = 0 : & g(0) \\ i \geq 1 : & g(\overline{(i - 1) + 1}) \end{cases} = (i \rightarrow g(i)) = g. \end{aligned}$$

Similarly, we find that $(\mathbb{0} + x^A \times s^{\text{List}^A}) \triangleright f_2 \circ f_1 = (\mathbb{0} + x^A \times s^{\text{List}^A})$ via this calculation:

$$\begin{aligned} (\mathbb{0} + x^A \times s^{\text{List}^A}) \triangleright f_2 \circ f_1 &= f_1 \left(i^{\text{Int}_{[0,n]}} \rightarrow \begin{cases} i = 0 : & x \\ i \geq 1 : & \overline{f_2}(s)(i - 1) \end{cases} \right) \\ &= \mathbb{0} + x \times \overline{f_1}(i \rightarrow \overline{f_2}(s)(i + 1 - 1)) = \mathbb{0} + x \times \overline{f_1}(\overline{f_2}(s)) = \mathbb{0} + x \times s. \end{aligned}$$

This concludes the proof of the isomorphism between `Array` and `List`.

Since arrays and lists are isomorphic as types, the concatenation for lists is associative as long as we show that the concatenation operation for lists is isomorphic to that we defined for arrays.

Statement 8.5.2.3 The `concat` function for lists is defined recursively as

```
def concat[A](p: List[A], q: List[A])
  : List[A] = p match {
  case List() => q
  case a :: t => a :: concat(t, q)
}
```

$$p^{\text{List}^A} ++ q^{\text{List}^A} \triangleq p \triangleright \begin{array}{c|c} & \text{List}^A \\ \hline 1 & q \\ \hline A \times \text{List}^A & a \times t \rightarrow \mathbb{0} + a \times (t \overline{+} q) \end{array},$$

and is equivalent to the `concat` function on arrays defined in Statement 8.5.2.1:

$$\begin{aligned} \forall a^{\text{Array}_{n_1}^A}, b^{\text{Array}_{n_2}^A}. \quad f_1(a) ++ f_1(b) &= f_1(a ++ b), \\ \forall p^{\text{List}^A}, q^{\text{List}^A}. \quad f_2(p) ++ f_2(q) &= f_2(p ++ q), \end{aligned}$$

where f_1, f_2 are the isomorphism maps defined in Statement 8.5.2.2.

Proof If we show the property for f_2 , we can apply f_1 to both sides and obtain the other property.

The base case for lists is $[] ++ q = q$. This is clearly isomorphic to concatenating an empty array with another array since $f_2([])$ is an empty array.

The inductive step is $p ++ q$ where p is a non-empty list, $p = \mathbb{0} + a^A \times t^{\text{List}^A}$. We need to show that

$$f_2(\mathbb{0} + a \times t) ++ f_2(q) = f_2((\mathbb{0} + a \times t) ++ q) = f_2(\mathbb{0} + a \times (t ++ q)).$$

By definition of f_2 , we have

$$\begin{aligned} f_2(\mathbb{0} + a \times t) &= i \rightarrow \begin{cases} i = 0 : & a \\ i \geq 1 : & \overline{f_2}(t)(i - 1) \end{cases}, \\ f_2(\mathbb{0} + a \times (t ++ q)) &= i \rightarrow \begin{cases} i = 0 : & a \\ i \geq 1 : & \overline{f_2}(t ++ q)(i - 1) \end{cases}. \end{aligned}$$

The inductive assumption guarantees that $\overline{f_2}(t ++ q) = f_2(t) ++ f_2(q)$. Using the definition of array concatenation and assuming that the length of t is n_1 and the length q is n_2 , we get

$$f_2(\mathbb{0} + a \times (t ++ q)) = i \rightarrow \begin{cases} i = 0 : & a \\ 1 \leq i < n_1 + 1 : & f_2(t)(i - 1) \\ n_1 + 1 \leq i < n_1 + 1 + n_2 : & f_2(q)(i - n_1 - 1) \end{cases}.$$

This is the same array as the concatenation $f_2(\mathbb{0} + a \times t) ++ f_2(q)$. This concludes the proof.

8.5.3 “Kinds” and higher-order type functions

Type constructors are types parameterized by other types. We have also seen types parameterized by type constructors, e.g.

```
trait Functor[F[_]] { ... }
```

It is important to distinguish the ways in which types can be parameterized. A type constructor such as `List` can be understood as a type-to-type function (TTF): given a type, e.g., `Int`, it produces another type, `List[Int]`. It is a type error to apply `List` to a type parameter that is not a simple type:

```
scala> val x: List[List] = ???
<console>:11: error: type List takes type parameters
      val x: List[List] = ???
      ^
```

To describe the restriction on possible type parameters of `List`, we say that `List` has **kind signature** $* \rightarrow *$. The symbol $*$ means an ordinary type (not a type function), i.e., a type that can have values. In the type notation, we can write “kind signatures” as

$$\text{Int} : * , \quad \text{List} : * \rightarrow * , \quad \text{List}^{\text{Int}} : * .$$

So, the concept of “kind” can be understood as the “type signature of a type”. Types of kind $*$ (such as `Int` or `String`) can have values, but types of kind $* \rightarrow *$ cannot:

```
scala> val x: Seq = ???
<console>:11: error: type Seq takes type parameters
      val x: Seq = ???
      ^
```

Although Scala will check that all kinds match, there is no syntax in Scala to declare kinds.

The `Functor` typeclass is another example of a non-trivial “kind”: a type whose type parameter is itself a type constructor,

$$\text{Functor} : (* \rightarrow *) \rightarrow * .$$

The `Functor` type can be seen as a higher-order type-to-type function (TTF) since it takes a type parameter that is itself a TTF. Such types are often called “higher-kinded types”.

For higher-order TTFs, Scala requires syntax such as `Functor[F[_]]` or `F[_]`, for example:

```
// Ap1 and Ap2 will simply substitute type arguments into type constructors of various kinds.
type Ap1[F[_], A] = F[A]    // Ap1 : ((* \rightarrow *) \times *) \rightarrow *
type Ap2[P[_[_], _], Q[_], R] = P[Q, R] // Ap2: (((* \rightarrow *) \times *) \times (* \rightarrow *) \times *) \rightarrow *
type G[A] = Either[(A, A, String), A] // G: * \rightarrow *
type X = Ap2[Ap, G, Int] // OK; X is now Either[(Int, Int, String), Int]
type Y = Ap2[Ap, Ap, Int] // Type error: the second argument of Ap2 has wrong kind.
type Z = Ap2[G, G, Int] // Type error: the first argument of Ap2 has wrong kind.
```

The “kind projector” plugin is often needed when writing code with higher-order TTFs:

```
// 'Twice' will apply the type constructor Q twice to its type argument, and substitute into P.
type Twice[P[_[_], _], Q[_], R] = P[Lambda[X => Q[Q[X]]], R] // Twice: (((* \rightarrow *) \times * \rightarrow *) \times (* \rightarrow *) \times *) \rightarrow *
type O2[A] = Option[(A, A)] // O2: * \rightarrow *
type X2 = Twice[Ap1, O2, Int] // X2 is now Option[(Option[(Int, Int)], Option[(Int, Int)])].
val x2: X2 = Some((Some((1, 2)), Some((3, 4)))) // Types match for 'x2'.
```

8.5.4 Inductive typeclasses and their properties

We have seen many examples of typeclasses that have similar structural properties. For instance, a product of semigroups is a semigroup, a product of monoids is a monoid, a product of functors is a functor, a product of pointed functors is a pointed functor, and so on. It turns out that all these typeclasses have a common structure that we can recognize and reason about. The common

Typeclass	Instance type as a function	Inductive form	Structure functor
default value	$\mathbb{1} \rightarrow A$	$P^A \rightarrow A$	$P^A \triangleq \mathbb{1}$
semigroup	$A \times A \rightarrow A$	$P^A \rightarrow A$	$P^A \triangleq A \times A$
monoid	$\mathbb{1} + A \times A \rightarrow A$	$P^A \rightarrow A$	$P^A \triangleq \mathbb{1} + A \times A$
functor	$F^A \times (A \rightarrow B) \rightarrow F^B$	$S^{\bullet, F} \rightarrow F^{\bullet}$	$S^{B, F^{\bullet}} \triangleq \forall A. F^A \times (A \rightarrow B)$
pointed functor	$B + F^A \times (A \rightarrow B) \rightarrow F^B$	$S^{\bullet, F} \rightarrow F^{\bullet}$	$S^{B, F^{\bullet}} \triangleq \forall A. B + F^A \times (A \rightarrow B)$
contrafunctor	$F^A \times (B \rightarrow A) \rightarrow F^B$	$S^{\bullet, F} \rightarrow F^{\bullet}$	$S^{B, F^{\bullet}} \triangleq \forall A. F^A \times (B \rightarrow A)$
pointed contrafunctor	$\mathbb{1} + F^A \times (B \rightarrow A) \rightarrow F^B$	$S^{\bullet, F} \rightarrow F^{\bullet}$	$S^{B, F^{\bullet}} \triangleq \forall A. \mathbb{1} + F^A \times (B \rightarrow A)$

Table 8.8: Structure of typeclass instance values for various inductive typeclasses.

structure is that all typeclass methods can be expressed as a *single* uncurried function of a specific form $P^A \rightarrow A$, as Table 8.8 shows.

Functions of type $P^A \rightarrow A$ compute new values of type A from previous values (of type A or other types) wrapped by the functor P . This superficially resembles defining values of type A by induction and so motivates the following definition: A typeclass is **inductive** if a type A 's typeclass instance is a value of type $P^A \rightarrow A$ with some functor P called the **structure functor** of the typeclass.

A value of type $P^A \rightarrow A$ represents all the methods of the an inductive typeclass in a single function. We can implement this definition by the Scala code

```
final case class InductiveTypeclass[P[_]: Functor, A](methods: P[A] => A)
```

For example, the `Monoid` typeclass is inductive because its instances have type

$$\text{Monoid}^A = (A \times A \rightarrow A) \times A \cong (\mathbb{1} + A \times A \rightarrow A) = P^A \rightarrow A \quad \text{where } P^A \triangleq \mathbb{1} + A \times A \quad .$$

So, the `Monoid` typeclass can be declared as an inductive typeclass by code like this,

```
final case class MonoidStructure[A](s: Option[(A, A)])
val monoidStructureFunctor: Functor[MonoidStructure] = ... /* implementation */
type InductiveMonoid[A] = InductiveTypeclass[MonoidStructure, A]
```

Implementing the `Monoid` typeclass via a structure functor P and a function $P^A \rightarrow A$ is inconvenient for practical programming. The inductive form $P^A \rightarrow A$ is mainly useful for reasoning about general properties of typeclasses. To illustrate that kind of reasoning, we will show that the product-type, the function-type, and the recursive-type constructions work for all inductive typeclasses.

Consider an arbitrary inductive typeclass `TC` defined via a structure functor P ,

```
type P[A] = ...
type TC[A] = P[A] => A
```

If two types A, B have an instance of the typeclass `TC`, an instance for the product $A \times B$ can be derived automatically using the function `productTC`:

```
def productTC[A, B](f: P[A] => A, g: P[B] => B): P[(A, B)] => (A, B) =
  p => (f(p.map(_._1)), g(p.map(_._2)))
```

$$\begin{aligned} \text{productTC} : (P^A \rightarrow A) \times (P^B \rightarrow B) \rightarrow P^{A \times B} \rightarrow A \times B \quad , \\ \text{productTC} \triangleq (f: P^A \rightarrow A \times B) \rightarrow p: P^{A \times B} \rightarrow (p \triangleright \pi_1^A \triangleright f) \times (p \triangleright \pi_2^B \triangleright g) \quad . \end{aligned}$$

This explains why product types have instances for all typeclasses listed in Table 8.8: those typeclasses are inductive. (However, this argument does not prove that the *laws* of a given typeclass will also hold for the product types.)

It is interesting to note that the co-product construction cannot be derived in general for arbitrary inductive typeclasses: given $P^A \rightarrow A$ and $P^B \rightarrow B$, it is not guaranteed that we can compute $P^{A+B} \rightarrow A + B$. Not all inductive typeclasses support the co-product construction (although many do).

The function-type construction promises a typeclass instance for $E \rightarrow A$ if the type A has a typeclass instance. This construction works for any inductive typeclass because a value of type $P^{E \rightarrow A} \rightarrow E \rightarrow A$ can be computed from a value of type $P^A \rightarrow A$ when P is a functor:

$$q : (P^A \rightarrow A) \rightarrow P^{E \rightarrow A} \rightarrow E \rightarrow A \quad , \quad q \triangleq h : P^A \rightarrow A \rightarrow p : P^{E \rightarrow A} \rightarrow E \rightarrow A \rightarrow e : E \rightarrow p \triangleright (x : E \rightarrow A \rightarrow e \triangleright x) \uparrow^P \triangleright h \quad .$$

As in the case of the product construction, the laws still need to be checked for the new instances.

Finally, let us show that the recursive type construction works for inductive typeclasses. Consider a recursive type T defined by a type equation $T \triangleq S^T$, where the functor S^\bullet “preserves” typeclass instances: if A has an instance then S^A also does, as we saw in all our examples in this chapter. In other words, we have a function $\text{tcS} : (P^A \rightarrow A) \rightarrow P^{S^A} \rightarrow S^A$ that creates typeclass instances of type $P^{S^A} \rightarrow S^A$ out of instances of type $P^A \rightarrow A$. Then we define an instance tcT for T as

$$\text{tcT} : P^T \rightarrow T \triangleq p : P^T \rightarrow \text{tcS}(\overline{\text{tcT}})(p) \quad .$$

This recursive definition terminates because it is implemented as an expanded function. The types match since we can convert between the equivalent types T and S^T whenever necessary, so $p : P^T$ can be converted to a value of type P^{S^T} , while the value $\text{tcS}(\text{tcT})(p)$ can be converted from type S^T back to type T . In Scala code, the conversions between T and S^T are implemented by the constructor $\text{T}(\dots)$ and accessor $(_.s)$ methods of the case class that wraps the type T .

```
type S[A] = ... // Define a functor S as required.
final case class T(s: S[T]) // Define the recursive type T as T ≈ S^T.
def tcS: TC[S[A]] => TC[S[A]] = ... // Compute instances for S[A] from instances of A.
def tcT: P[T] => T = p => T(tcS(tcT)(p.map(_.s))) // Define the recursive instance.
```

In this way, we can implement the recursive-type construction with any inductive typeclass. (The typeclass laws still need to be checked for the recursively defined instances.)

The `Functor` and `Pointed` typeclasses are also inductive. Instead of a structure functor, we need to use a higher-order type function denoted by S^{\bullet, F^\bullet} and parameterized by a type constructor F^\bullet as well as by a type parameter A . (The type S has kind $* \times (* \rightarrow *) \rightarrow *$.) The methods of these typeclasses are expressed as $S^{\bullet, F} \rightarrow F^\bullet$, which is analogous to $P^A \rightarrow A$ except for additional type parameters. Similar arguments can be made for these typeclasses, although it is more difficult to reason about type constructors (and the laws will not hold without additional assumptions). As we have seen, the product-type, the function-type, and the recursive-type constructions work for functors, contrafunctors, pointed functors, and pointed contrafunctors.

Important inductive typeclasses are functors and contrafunctors (Chapter 6), filterables (Chapter 9), monads (Chapter 10), and applicatives (Chapter 11). We will see in Chapter 13 that another general construction also works for all inductive typeclasses, — the “free” type construction.

We have also seen examples of typeclasses that are *not* inductive. The `Extractor` typeclass has instances of type $A \rightarrow Z$, which is not of the form $P^A \rightarrow A$ for any functor P . Another such typeclass is `Copointed`, whose method $F^A \rightarrow A$ is not of the form $S^{\bullet, F} \rightarrow F^\bullet$. However, the methods of these typeclasses can be written in the inverted form $A \rightarrow P^A$ with some functor P . We call them **co-inductive typeclasses**. A motivation for this name is that the *co-product* construction (rather than the product construction) works with co-inductive typeclasses: given values of types $A \rightarrow P^A$ and $B \rightarrow P^B$, we can produce a value of type $A + B \rightarrow P^{A+B}$, i.e., a co-product instance, but not necessarily a value of type $A \times B \rightarrow P^{A \times B}$, which would be a product-type instance. The function-type construction is not guaranteed to work with co-inductive typeclasses, but the recursive-type construction and the “free” type construction can be implemented after appropriate modifications.

The `Eq` typeclass is “uninductive” (neither inductive nor co-inductive) because its instance type, $A \times A \rightarrow \mathbb{2}$, is neither of the form $P^A \rightarrow A$ nor $A \rightarrow P^A$. The traversable functor (Chapter 12) is

another example of an uninductive typeclass. Uninductive typeclasses usually support fewer type constructions. For example, only *polynomial* types can have instances of `Eq` or `Traversable` typeclasses.

8.5.5 Typeclasses with more than one type parameter (type relations)

A typeclass constraint in a function, such as `func[A: Monoid]`, restricts a type parameter to a certain type domain. Sometimes it is necessary to restrict *several* type parameters to satisfy some conditions together. Let us look at two simple examples of this.

The first example is converting integer numbers to floating point. The ranges of the available types allow us to convert a `Short` to a `Float` and an `Int` to a `Double`. Can we implement the type signature

```
def convertNumber[M, N](x: M): N
```

where the type parameters `M, N` are constrained to be either `M = Short` and `N = Float`, or `M = Int` and `N = Double`? (Of course, we will want to be able to add further supported pairs of types later.)

A condition that constrains several type parameters at once is called a **type relation**.

The second example is converting mutable data structures into the corresponding immutable ones. The Scala library contains data structures such as sequences, sets, and dictionaries, each having a mutable and an immutable version. Can we implement a function with type signature

```
def convertData[Mut[_], Immut[_], A](data: Mut[A]): Immut[A]
```

where the type parameters `Mut` and `Immut` are constrained to represent the mutable / immutable versions of a supported data structure (for example, `Mut = mutable.Set` and `Immut = immutable.Set`, etc.)?

A typeclass constraint for a single type parameter, such as `func[A: Monoid]`, is implemented by requiring an “evidence value” of type `Monoid[A]` as an additional argument of the function `func`. Implementing a type relation for *two* type parameters `A, B` is similar: We define a type constructor, say `Rel[_ _, _]`, and create some evidence values of type `Rel[A, B]` with chosen type parameters `A, B`. Any function, say `f[A, B]`, that needs a relation constraint on its type parameters will take an extra evidence argument of type `Rel[A, B]`. This will prevent our code from using `f[A, B]` with types `A, B` that are not in the required relation.

Using this technique, we can define a relation `MayConvert` and write the code for `convertNumber`:

```
final case class MayConvert[A, B](convert: A => B) // Evidence value contains a conversion function.
implicit val ev1 = MayConvert[Short, Float](_.toFloat) // Evidence value for [Short, Float].
implicit val ev2 = MayConvert[Int, Double](_.toDouble) // Evidence value for [Int, Double].
def convertNumber[M, N](x: M)(implicit ev: MayConvert[M, N]): N = ev.convert(x)
```

With these definitions, it will be a compile-time error to use `convertNumber` with unsupported types:

```
scala> convertNumber(123)
res0: Double = 123.0

scala> convertNumber(123:Short)
res1: Float = 123.0

scala> convertNumber("abc")
<console>:17: error: could not find implicit value for parameter ev: MayConvert[String,N]
      convertNumber("abc")
           ^
```

As we have just seen, a type relation is defined by creating a set of evidence values. The code above defines `MayConvert` as a one-to-one type relation because the evidence values (or “relation instances”) `ev1` and `ev2` do not have any types in common. So, `MayConvert` is equivalent to a type-to-type *function* that maps `Short` to `Float` and `Int` to `Double`. However, type relations are not limited to one-to-one relations or to type functions. By creating suitable implicit evidence values, we can implement many-to-one or many-to-many relations when needed.

A practical example of a many-to-many type relation is the compatibility between physical units. Miles can be converted into kilometers; pounds can be converted into ounces or kilograms; but

kilograms cannot be converted into miles. Type relations allow us to implement type-safe operations for quantities with units. Adding kilograms to pounds will automatically convert the quantity to a common unit, while adding kilograms to miles will raise a compile-time error.

Begin by declaring a type for “quantity with units” and some type names for the supported units:

```
trait KG; trait LB; trait OZ; trait KM; trait MI; trait FT // Declare names for some supported units.
final case class Quantity[U](value: Double)      // Constant functor: No values of type 'U' are stored.
def add[U1, U2](x: Quantity[U1], y: Quantity[U2]): Quantity[U2] = ???
```

The parameter `U` in `Quantity[U]` is called a **phantom** type parameter because `Quantity[U]` contains no values that use `U`. The parameter `U` would not be phantom if `Quantity[U]` were not a constant functor but instead contained values e.g., of type `U` or of type `U => Double` or of some other type that uses `U`.

The function `add[U1, U2](x, y)` must impose a type relation constraint on `U1` and `U2`, so that `x` and `y` may be added only when they have compatible units. The type relation is implemented as a type constructor `Convertible` with two type parameters. A “relation instance” (i.e., a value of type `Convertible[U1, U2]`) will contain a multiplier for converting any quantity from units `U1` to units `U2`:

```
final case class Convertible[U1, U2](multiplier: Double)
implicit val c1 = Convertible[KG, KG](1.0)
implicit val c2 = Convertible[LB, KG](0.453592) // Pounds in kilograms.
implicit val c3 = Convertible[KM, KM](1.0)
implicit val c4 = Convertible[MI, KM](1.60934) // Miles in kilometers. Add more definitions here.
```

Now we can implement the `add` function and verify that the type relation works as we intended:

```
def add[U1, U2](x: Quantity[U1], y: Quantity[U2])(implicit ev: Convertible[U1, U2]): Quantity[U2] =
  Quantity(x.value * ev.multiplier + y.value)

scala> add(Quantity[LB](1), Quantity[KG](1)) // 1 pound + 1 kg = 1.453592 kg.
res0: Quantity[KG] = Quantity(1.453592)

scala> add(Quantity[MI](1), Quantity[KG](1)) // Compile-time error: cannot add miles to kilograms.
<console>:25: error: could not find implicit value for parameter ev: Convertible[MI,KG]
```

To make this code more convenient for practical use, we can add extension methods that shorten the syntax from `Quantity[KG](1)` to `1.kg` and from `add(2.lb, 2.kg)` to a more readable `2.lb + 2.kg`:

```
implicit class QuantitySyntax(x: Double) {
  def kg = Quantity[KG](x)
  def lb = Quantity[LB](x)           // Add more definitions as needed.
}
implicit class QuantityAdd[U1](x: Quantity[U1]) {
  def +[U2](y: Quantity[U2])(implicit ev: Convertible[U1, U2]): Quantity[U2] = add(x, y)
}

scala> 2.lb + 2.kg      // Compute 2 pounds + 2 kilograms; the result is in kilograms.
res1: Quantity[KG] = Quantity(2.907184)
```

Another necessary improvement is reducing the number of implicit values. The current code uses n^2 implicit values for every group of n compatible units. Adding support for a new unit (say, inches) requires adding implicit values for converting between inches and all previously defined units of length. To avoid this problem, the code must be reorganized to convert all compatible quantities to chosen units (e.g., all length to kilometers and all mass to kilograms). This makes the `Convertible` relation many-to-one instead of many-to-many. The resulting code is shown in Figure 8.2.

8.5.6 Inheritance and automatic conversions of typeclasses

It often happens that one typeclass is a subset of another; for example, a `Monoid` instance for a type `T` means that `T` already has the properties of both `Semigroup` and `HasDefault` typeclasses. One says that the `Monoid` typeclass **inherits** from the `Semigroup` and `HasDefault` typeclasses. We may want to express inheritance relations between typeclasses, so that e.g., a `Monoid` instance should automatically imply the presence of a `Semigroup` instance, without extra code.

```

trait KG; trait LB; trait OZ; trait KM; trait MI; trait FT           // Some supported units.

// This type relation is many-to-1: it relates all mass units to KG and all length units to KM.
final case class Convertible[U1, U2](multiplier: Double) extends AnyVal
// Units of mass.
implicit val cKG = Convertible[KG, KG](1.0)
implicit val cLB = Convertible[LB, KG](0.453592)
implicit val cOZ = Convertible[OZ, KG](0.0283495)
// Units of distance.
implicit val cKM = Convertible[KM, KM](1.0)
implicit val cMI = Convertible[MI, KM](1.60934)
implicit val cFT = Convertible[FT, KM](0.0003048)
// We can add support for new units whenever necessary.

final case class Quantity[U](value: Double) extends AnyVal { // Use 'AnyVal' to reduce run-time cost.
  def +(U2, C)(q: Quantity[U2])(implicit ev1: Convertible[U, C], ev2: Convertible[U2, C]) =
    Quantity[U2](value * ev1.multiplier / ev2.multiplier + q.value)

  def ==(U2, C)(q: Quantity[U2])(implicit ev1: Convertible[U, C], ev2: Convertible[U2, C]) =
    value * ev1.multiplier == q.value * ev2.multiplier
}

implicit class QuantitySyntax(x: Double) { // Extension methods defined for convenience.
  def kg = Quantity[KG](x)
  def lb = Quantity[LB](x)
  def oz = Quantity[OZ](x)
  def km = Quantity[KM](x)
  def mi = Quantity[MI](x)
  def ft = Quantity[FT](x)
  // This general extension method, e.g., '1.in[KM]', will also work for any units defined later.
  def in[U](implicit ev: Convertible[U, _]): Quantity[U] = Quantity[U](x)
}

scala> 1.in[KM]           // Use the general method '.in' with a type parameter.
res1: Quantity[KM] = Quantity(1.0)

scala> 10000.ft + 1.km // Use the '.ft' and '.km' extension methods.
res2: Quantity[KM] = Quantity(4.048)

scala> 1.kg + 2.lb == 32.oz + 1.kg // Compare values safely.
res2: Boolean = true

scala> 1.km + 2.lb           // Compile-time error: cannot add kilometers to pounds.
<console>:29: error: could not find implicit value for parameter ev2: Convertible[LB,C]
      1.km + 2.lb
      ^

scala> trait YD; implicit val cYD = Convertible[YD, KM](0.0009144) // Add support for yards.
defined trait YD
cIN: Convertible[YD,KM] = Convertible(9.144E-4)

scala> 1.in[YD] + 1.ft // Use .in[YD] to compute 1 YD + 1 FT = 4 FT.
res3: Quantity[FT] = Quantity(4.0)

```

Figure 8.2: Implementing type-safe computations with units of length and mass.

One way of inheriting a typeclass is to add a constraint to the new typeclass constructor:

```
final case class Semigroup[T](combine: (T, T) => T)
final case class Monoid[T: Semigroup](empty: T) // Inherit 'Semigroup'.
```

The case class `Monoid` combines a previous `Semigroup` instance with the new method `empty`. Creating an instance of `Monoid` for a type `T` will then require having an instance of `Semigroup` for `T`:

```
implicit val c1 = Semigroup[Int](_ + _)
implicit val c2 = Monoid[Int](0) // Works only if a 'Semigroup[Int]' is available.
```

Given a `Monoid[T]` instance, how can we recover the inherited implicit value of type `Semigroup[T]`? This is achieved by redefining the `Monoid` class like this,

```
final case class Monoid[T](empty: T)(implicit val semigroup: Semigroup[T]) // Use 'implicit val'.
implicit def semigroupFromMonoid[T](implicit ti: Monoid[T]): Semigroup[T] = ti.semigroup
```

Another possibility of implementing typeclass inheritance is to use object-oriented inheritance of traits. This approach is often used when defining typeclasses as traits with methods:

```
trait Semigroup[T] { def combine: (T, T) => T }
trait Monoid[T] extends Semigroup[T] { def empty: T } // 'def combine' is inherited from 'Semigroup'.
```

Creating an instance of `Monoid` for a type `T` no longer requires having an instance of `Semigroup` for `T`:

```
implicit val c: Monoid[Int] = new Monoid[Int] {
  def empty = 0
  def combine: (Int, Int) => Int = _ + _
```

With this approach, we cannot avoid repeating the code for `def combine...` even if we *do* already have a `Semigroup` instance. However, conversion from `Monoid[T]` to `Semigroup[T]` is now automatic

due to **object-oriented inheritance**: Scala considers `Monoid[T]` a subtype of `Semigroup[T]` because the `Monoid` class is declared as `extends Semigroup`.

A problem with the object-oriented inheritance is that automatic conversions to parent typeclasses cannot be *disabled*. When several typeclasses inherit from the same parent typeclass, duplicate implicit instances of the parent typeclass will be present, which is a compile-time error:

```
trait TC[A] // Parent typeclass.
trait TC1[A] extends TC[A] // Two different typeclasses TC1 and TC2 inherit from TC.
trait TC2[A] extends TC[A]
// The function f requires A to have both TC1 and TC2 instances and then wants to access TC instance.
def f[A: TC1 : TC2]() = {
  implicitly[TC[A]] // Compilation fails because two implicit values of type TC[A] are found.
}
```

When typeclass inheritance is implemented by combining instances without object-oriented inheritance, conversions to parent typeclasses are not automatic: they are implemented by implicit functions that need to be imported into the current scope. The programmer's code can avoid producing duplicate instances by choosing which implicit conversions to import:

```
final case class TC[A]()
final case class TC1[A]()(implicit val tc: TC[A]) // TC1 inherits TC.
object TC1 { implicit def toTC[A](implicit x: TC1[A]): TC[A] = x.tc }
final case class TC2[A]()(implicit val tc: TC[A]) // TC2 inherits TC.
object TC2 { implicit def toTC[A](implicit x: TC2[A]): TC[A] = x.tc }

// The function f requires A to have both TC1 and TC2 instances and then wants to access TC instance.
def f[A: TC1 : TC2]() = {
  import TC1._ // Can import TC1._ or TC2._ but not both. If the next line is uncommented,
  // import TC2._ // compilation will fail because two implicits of type TC[A] will be found!
  implicitly[TC[A]] // This compiles successfully. One implicit instance of TC[A] can be found.
}
```

The problem of duplicate inherited instances can be solved with less work for the programmer if the typeclasses are implemented using a special, more complicated encoding.⁸

⁸See slides 5–14 of the talk by John de Goes: <https://www.slideshare.net/jdegoes/scalaz-8-a-whole-new-game>

9 Computations in functor blocks. I.

Filterable functors and contrafunctors

Chapter 6 studied the mathematical properties of `map` and derived the concept of “functor” as a general type constructor that permits a `map` method with useful properties. In this chapter, we will study the `filter` method in a similar way. We will start from practical examples and then derive the relevant mathematical properties. This will give us a precise definition of “filterable” type constructors — those that permit useful implementations of `filter`.

9.1 Practical uses of filtering

The Scala standard library defines the `filter` method on sequences, sets, and other data structures. An example of using `filter` is the following calculation,

$$\sum_{x \in \mathbb{Z}; 0 \leq x \leq 100; \cos x > 0} \sqrt{\cos(x)} \approx 38.71 \quad .$$

```
scala> (0 to 100).map(x => math.cos(x)).filter(_ > 0).map(math.sqrt).sum
res0: Double = 38.71218949848382
```

The role of `filter` in this computation is to select only the positive values of $\cos(x)$. It is safe to apply the square root function to positive values, so the code will work correctly.

The code above is a chain of methods, but the same code can be written using Scala’s `for/yield` syntax, which is called a **functor block** in this book. The `filter` operation is represented by an `if` keyword embedded in a functor block. Compare each line of the functor block code with the corresponding line of the code written via method chains:

```
(for { x <- 0 to 100
      y = math.cos(x)
      if y > 0
    } yield math.sqrt(y)
).sum
```

```
(0 to 100).map { x =>
  math.cos(x) }.filter { y =>
  y > 0 }.map { y =>
  math.sqrt(y)
}.sum
```

Functor blocks require the first line to have a left arrow placed before a “source”, i.e., a value of a functor type (e.g., a sequence or a `Try`). Each line of a functor block can be viewed as a computation that creates an intermediate value of the same “source” functor type. The result value of the entire functor block is again a value of the same functor type. So, a functor block manipulates data wrapped by a functor without changing the type of the wrapper.

The name “functor block” is chosen because the code is a single expression (a “block”) whose type is required to be a functor. The functor’s type is selected by the source’s type in the first line.

We have seen in Chapter 6 that the functor block syntax becomes available for a type constructor that has a `map` method (i.e., for a functor). To support the embedded `if` keyword, the type constructor must support a method called `withFilter` with the same type signature as the `filter` method. The type signatures of `map` and `withFilter` methods for a type constructor `F[_]` can be written as

```
class F[A] { // F[A] is some type constructor that has .map and .withFilter methods.
  def map[B](f: A => B): F[B] = ...
  def withFilter(p: A => Boolean): F[A] = ...
}
```

A type constructor that supports the filtering operation is called **filterable**.

The main focus of this chapter is to explore filterable functors in detail. Programmers would intuitively expect the filtering operation to have certain properties. We will now examine the expected properties and translate them into mathematical laws for the function `withFilter`.

9.1.1 Examples and intuitions for the filtering operation

Examples of often used filterable functors defined in the Scala library are `List` and `Option`:

```
scala> List(64, 128).filter(_ > 100).map(_ * 2)
res0: List[Int] = List(256)
```

```
scala> Some(128).filter(_ > 100).map(_ * 2)
res1: Option[Int] = Some(256)
```

```
scala> Some(64).filter(_ > 100).map(_ * 2)
res2: Option[Int] = None
```

In an intuitive view, a functor wraps one or more data values, and the filtering operation may decrease the number of values wrapped. e.g., the length of the sequence `List(64, 128)` is decreased from 2 to 1 after filtering with the condition `_ > 100`. The `Option` functor can wrap at most one value, so filtering with a predicate returning `false` will return an empty `Option` value (i.e., `None`). In all cases, the resulting collection will not contain values that fail the filtering predicate.

Note that `Option[T]` is written in the type notation as a disjunctive type $1 + T$, while `List[T]` can be viewed as an “infinite disjunction”,

$$\text{List}^T = 1 + T + T \times T + T \times T \times T + \dots \quad (9.1)$$

Disjunctive type constructors are able to hold a different number of values of type T in different parts of the disjunction (e.g., $T \times T$ for 2 values and $T \times T \times T$ for 3 values of T).

So, we expect that a filterable functor should contain a disjunctive type supporting a different number of values of T , including *zero* values. When the filtering operation `.filter(p)` is applied, some values of type T will fail the predicate `p` and will be removed from the collection. The example

```
scala> List(64, 128).filter(_ > 100)
res0: List[Int] = List(128)
```

corresponds to mapping a disjunctive part of type $T \times T$ to a part of type T within List^T in Eq. (9.1).

Consider now a custom data type that implements a given application’s business requirements:

Example 9.1.1.1 On a given week, an order (data type A) can be placed on Tuesday and/or on Friday. An order is approved under certain conditions given by a predicate `p: A => Boolean`. Can we represent the order approval by a filtering operation on a suitable data type?

The data type describing a week’s orders must describe a possible order on Tuesday (`Option[A]`) and a possible order on Friday (again `Option[A]`). So, we can represent a week’s orders as a product $F^A \triangleq (1 + A) \times (1 + A)$ and implement it as a case class having methods `map` and `withFilter`:

```
final case class Orders[A](tue: Option[A], fri: Option[A]) {
  def map[B](f: A => B): Orders[B] = Orders(tue.map(f), fri.map(f)) // Functor.
  def withFilter(p: A => Boolean): Orders[A] = Orders(tue.filter(p), fri.filter(p)) // Filterable.
}

scala> Orders(Some(500), Some(2000)).withFilter(_ < 1000) // Approved if the amount is below $1000.
res0: Orders[Int] = Orders(Some(500), None)
```

This code applies filtering independently to both parts of the product. With this definition, we will be able to use `Orders` as a “source” of data in functor blocks:

```
scala> for {
  x <- Orders(Some(500), Some(2000)) // "Source" of type Orders[Int].
  y = x - 200 // Apply discount of $200 to each order.
```

```

1  if y < 500      // Orders are approved if the amount is below $500 after discount.
2  } yield y * 1.10 // Add 10% tax. Result is of type Orders[Double].
3  res1: Orders[Double] = Orders(Some(330.0), None)

```

Suppose we are considering an additional business rule, such as:

- (a) Both orders must be approved, or else no orders can be placed that week.
- (b) Both orders can be placed that week if at least one of them is approved.

We could modify the code of `withFilter` to implement one of the rules (a) or (b). Will the resulting function still be a “filtering” operation?

We cannot decide this without knowing the mathematical laws that a filtering operation must satisfy. Let us now consider what intuitive expectations we have for the concept of filtering.

9.1.2 Motivation for and derivation of the laws of filtering

Computations in a functor block will “make sense” if we easily understand what the program does when we look at the code. Consider this schematic example of a functor block program that uses a filterable functor `List`:

```

1  val result = for { // Some computations in the context of the 'List' functor.
2    x <- List(...) // For each x in the given list...
3    y = f(x)        // ... compute y
4    if p1(y)         // ... impose condition p1: discard all x, y for which p1(y) == false
5    if p2(y)         // ... same for condition p2
6    z = g(x, y)    // ... compute z
7    if q(x, y, z)  // ... and impose another condition.
8  } yield          // For all x in the given list, such that all the conditions hold,
9    k(x, y, z)    // compute the values k, put them into a list, and return as the list 'result'.

```

There are several properties that we intuitively expect such programs to have. One property is that computing $y = f(x)$ in line 3 and then checking a condition for y , such as “`if p1(y)`” in line 4, should be the same as checking the condition $p1(f(x))$ and then computing $y = f(x)$: since the code says that $y = f(x)$, we expect the conditions $p1(y)$ and $p1(f(x))$ to be equivalent.

Translating this equivalence into code, we obtain the requirement that the following two expressions (`result1` and `result2`) should always be equal to each other:

```

val result1 = for {
  x <- xs
  y = f(x)
  if p(y)
} yield y
// Translating the functor block into methods:
val result1 = xs.map(f).filter(p)

```

```

val result2 = for {
  x <- xs
  if p(f(x))
  y = f(x)
} yield y
// Translating the functor block into methods:
val result2 = xs.filter(x => p(f(x))).map(f)

```

Lines 4–5 of the listing above show two filtering operations, “`if p1(y)`” and “`if p2(y)`”, applied one after another. We expect that the first filtering operation keeps only values that satisfy the condition $p1$, and the second filtering operation is applied to the results of the first one, additionally imposing the condition $p2$. So, we expect that applying these two filtering operations is equivalent to filtering by the condition “`if p1(y) && p2(y)`”.

We can translate this expectation into equality of the following two code expressions:

```

val result1 = for {
  x <- xs
  if p1(x)
  if p2(x)
} yield x
// Translating the functor block into methods:
val result1 = xs.filter(p1).filter(p2)

```

```

val result2 = for {
  x <- xs
  if (p1(x) && p2(x))
} yield x
// Translating the functor block into methods:
val result2 = xs.filter(x => p1(x) && p2(x))

```

When a filter predicate $p(x)$ returns `true` for all x , the filtering operation `xs.filter(p)` will never discard any values. So we expect the result to remain the same if we *delete* the line “`if true`” from a functor block program. The corresponding code equivalence can be written as

```
xs.filter(_ => true) == xs
```

Now, suppose a predicate $p(x)$ returns `false` for certain values x . Then we expect those values x to be excluded from any computations performed *after* the line “`if p(x)`”. In particular, we should be able to use a partial function safely as long as that function is defined for x such that $p(x) == true$. To express this in code, first define a general “factory” for partial functions,

```
def if_p[A, B](p: A => Boolean)(f: A => B): A => B = x => p(x) match { case true => f(x) }
```

This “factory” takes a predicate $p:A \rightarrow 2$ and a function $f:A \rightarrow B$, and returns a partial function with the same signature, $A \rightarrow B$, but defined only for values x for which $p(x) = \text{true}$. Let us denote that function for brevity by $f|_p$. Since the `Boolean` type is equivalent to a disjunction of two “named unit types”, $2 \cong 1 + 1$ (meaning “`false`” + “`true`”), we can write the code notation for the function $f|_p$ as

$$(f:A \rightarrow B)|_p \triangleq x:A \rightarrow p(x) \triangleright \begin{array}{c|c|c} & & B \\ \hline & 1 (\text{false}) & 0 \\ \hline 1 (\text{true}) & & 1 \rightarrow f(x) \end{array} . \quad (9.2)$$

The top row contains the void type 0 , indicating that the partial function $f|_p$ will crash if applied to a value x for which $p(x) = \text{false}$.

Using the partial function $f|_p$, we write the last property of the filtering operation as the equality of the expressions `result1` and `result2` defined like this:

<pre>val result1 = for { x <- xs if p(x) y = f(x) } yield y // Translating the functor block into methods: val result1 = xs.filter(p).map(f)</pre>	<pre>val result2 = for { x <- xs if p(x) y = fp(x) // Defined as val fp = if_p(p)(f) } yield y // Translating the functor block into methods: val result2 = xs.filter(p).map(fp)</pre>
---	---

We found 4 requirements for the `filter` function, written in terms of equal code fragments. These requirements are the four “laws” (i.e., equations) that any reasonable `filter` must satisfy. In the code notation, `filter` is filt_F ,

$$\text{filt}_F : (A \rightarrow 2) \rightarrow F^A \rightarrow F^A ,$$

and its 4 laws (called the naturality, identity, composition, and partial function laws of `filter`) are

$$\text{naturality law : } f \uparrow^F ; \text{filt}_F(p) = \text{filt}_F(f \circ p) ; f \uparrow^F \quad \text{for } \forall (f:A \rightarrow B, p:B \rightarrow 2) . \quad (9.3)$$

$$\text{identity law : } \text{filt}_F(_ \rightarrow \text{true}) = \text{id}^{F^A \rightarrow F^A} . \quad (9.4)$$

$$\text{composition law : } \text{filt}_F(p_1) ; \text{filt}_F(p_2) = \text{filt}_F(x \rightarrow p_1(x) \wedge p_2(x)) \quad \text{for } \forall (p_1:A \rightarrow 2, p_2:A \rightarrow 2) . \quad (9.5)$$

$$\text{partial function law : } \text{filt}_F(p) ; f \uparrow^F = \text{filt}_F(p) ; f|_p \uparrow^F \quad \text{for } \forall (f:A \rightarrow B, p:A \rightarrow 2) . \quad (9.6)$$

The following type diagram illustrates the naturality law of `filter`:

$$\begin{array}{ccc} F^A & \xrightarrow{\text{filt}_F(f:A \rightarrow B ; p:B \rightarrow 2)} & F^A \\ (f:A \rightarrow B) \uparrow^F \downarrow & & \downarrow (f:A \rightarrow B) \uparrow^F \\ F^B & \xrightarrow{\text{filt}_F(p:B \rightarrow 2)} & F^B \end{array}$$

A functor F is called **filterable** if there exists a function filt_F satisfying these four laws.

We may define a typeclass `Filterable` and extension methods `filter` and `withFilter` like this,

```
trait Filterable[F[_]] { def filt[A](p: A => Boolean)(fa: F[A]): F[A] }
implicit class FilterableSyntax[F[_], A](fa: F[A])(implicit ev: Filterable[F]) {
  def filter(p: A => Boolean): F[A] = ev.filt(p)(fa)
  def withFilter(p: A => Boolean): F[A] = filter(p)
}
```

It is intuitively clear why functors such as `Option` and `List` obey the filtering laws: those types can be viewed as “containers” holding zero or more items of data, and the `filter` operation removes all data that fails the filtering condition. What about the custom data type `Orders` from Example 9.1.1? In principle, we would need to verify all four laws symbolically, using the code of `withFilter` as we implemented it for `Orders`. Later in this chapter we will see that the four laws can be simplified, reduced to just two laws, and proved more quickly. For now, we can use the `scalacheck` library to implement randomized tests for the four filtering laws:

```
def checkFilteringLaws[F[_] : Filterable : Functor, A, B](implicit
  faEv: Arbitrary[F[A]], fbEv: Arbitrary[F[B]], abEv: Arbitrary[A => B],
  aEv: Arbitrary[A => Boolean], bEv: Arbitrary[B => Boolean]): Assertion = {
  forAll { (f: A => B, p: B => Boolean, fa: F[A]) =>           // Naturality law.
    fa.map(f).filter(p) shouldEqual fa.filter(f andThen p).map(f)
  }
  forAll { (p1: B => Boolean, p2: B => Boolean, fa: F[B]) =>      // Composition law.
    fa.filter(p1).filter(p2) shouldEqual fa.filter(b => p1(b) && p2(b))
  }
  forAll { (fb: F[B]) => fb.filter(_ => true) shouldEqual fb }      // Identity law.

  forAll { (f: A => B, p: A => Boolean, fa: F[A]) =>           // Partial function law.
    fa.filter(p).map(f) shouldEqual fa.filter(p).map[B](x => p(x) match { case true => f(x) })
  }
}
```

Creating a `Filterable` typeclass instance for `Orders` and running the tests will show no errors:

```
implicit val filterableOrders = new Filterable[Orders] {
  def filt[A](p: A => Boolean)(fa: F[A]): F[A] = fa.filter(p)
}
checkFilteringLaws[Orders, Int, String]           // Need to set type parameters to specific types.
```

9.1.3 Examples of non-filterable functors

As usual with typeclasses, the code of the `Filterable` typeclass fixes the type signature of the `filter` function but does not enforce its laws. It is up to the programmer to verify that the implementation of `filter` satisfies the laws.

If we define the filtering operation for the `Orders` data type (see Example 9.1.1) with the extra business rule (a), we get

```
implicit val filterableOrdersRuleA = new Filterable[Orders] {
  def filt[A](p: A => Boolean)(fa: F[A]): F[A] =
    if (fa.tue.forall(p) && fa.fri.forall(p)) fa.filter(p)
    else Orders(None, None)           // Rule (a): No orders are approved unless both are approved.
}
checkFilteringLaws[Orders, Int, String]           // Tests pass.
```

However, implementing business rule (b) will violate some laws:

```
implicit val filterableOrdersRuleB = new Filterable[Orders] {
  def filt[A](p: A => Boolean)(fa: F[A]): F[A] =
    if (fa.tue.exists(p) || fa.fri.exists(p)) fa      // Here, the value 'fa' remains unchanged.
    else Orders(None, None)           // Rule (b): Both orders are approved if at least one is approved.
}
checkFilteringLaws[Orders, Boolean, Boolean]()      // Tests will fail! A specific failing case:

scala> Orders(Some(500), Some(2000)).filter(x => x < 1000).filter(x => x > 1000)
res0: Orders[Int] = Orders(Some(500),Some(2000))

scala> Orders(Some(500), Some(2000)).filter(x => x < 1000 && x > 1000) // Composition law fails:
res1: Orders[Int] = Orders(None, None)
```

If we implement rule (b), the filtering operation will not correspond to an intuitive understanding of computations in functor blocks:

```
scala> for { x <- Orders(Some(500), Some(2000))
  if x < 1000           // Intuition says that values of x must be below 1000 from now on.
  y = s"Amount: $x"      // So, the value x = 2000 should never appear in this line.
} yield y                // But the final result does not correspond to this intuition:
res2: Orders[String] = Orders(Some("Amount: 500"), Some("Amount: 2000"))
```

This computation violates the partial function law because the value $x = 2000$ is not excluded from further computations despite filtering with the predicate " $x < 1000$ ". This happened because the code of `filter` does not remove the value $x = 2000$ from the data structure in that case.

The four laws of filtering are a rigorous formulation of our intuitions about what it means to "filter data". The type `Orders` with business rule **(b)** is an example of a filtering operation that does not correspond to our intuitions and, as a consequence, violates the filtering laws. This does not mean that business rule **(b)** cannot be used in real-world programs; it only means that order approval according to rule **(b)** is not a filtering operation. For instance, applying two order approvals one after another will not give the intuitively expected results. Nevertheless, this may be acceptable in applications where only one order approval is ever applied.

Violations of the filtering laws by business rule **(b)** also does not mean that the functor `Orders` is not filterable. The same functor with business rule **(a)** has a lawful implementation of `filter`.

What are examples of functors that are not filterable? One way of implementing the type signature of the `filter` function is to use the identity function,

```
def filt[A](p: A => Boolean)(fa: F[A]): F[A] = fa // Ignore the predicate 'p'; always return 'fa'.
```

This implementation never removes any data and so will violate the partial function law if the functor F^A wraps a value of type A that does not pass the filter. However, the identity function is the *only* possible implementation of `filter` for certain functors F , e.g., the identity functor $F^A \triangleq A$ or the exponential functor $F^A \triangleq Z \rightarrow A$. So, functors of the form $F^A \triangleq Z \rightarrow A$ (where Z is a fixed type) are not filterable.

The functor $F^A \triangleq A \times (1 + A)$ is not filterable because the filtering operation $\text{filt}_F(p)$ cannot remove the first value of type A when it does not pass the filter predicate p .

Functors such as $F^A \triangleq 1 + A$ and $F^A \triangleq 1 + A \times A$ are filterable because the function filt_F can be defined correctly; but incorrect implementations are also possible. For example, one could imagine defining `filter` for an `Option` type as

```
implicit val wrongFilterableOption = new Filterable[Option] {
  def filt[A](p: A => Boolean)(fa: Option[A]): Option[A] = None // Discard input, always return None.
}
```

This code discards information and violates the identity law: the result of filtering with an identically `true` predicate is not the identity function of type `Option[A] => Option[A]`.

Finally, one could violate the naturality law by defining `filter` in a special way when the type parameter A is set to, say, `Int`. To obey the naturality law, the `filter` function must be fully parametric and must not use hard-coded values of specific types or make decisions based on specific types.

Note that the `Boolean` type is equivalent to $2 \cong 1 + 1$; in other words, this type can be expressed via the basic type constructions (disjunction and the `Unit` type) without using any externally defined values. For this reason, it is allowed to use the `Boolean` type in fully parametric functions.

9.1.4 Solved examples: Programming with filterable functors

Example 9.1.4.1 A cluster has 2 servers; each server needs to have valid credentials, which expire periodically. If credentials expire for one server, it may copy valid credentials from the other server. If no server has valid credentials, the cluster is down. Is this setup described by a filterable functor?

Solution Assuming that credentials have type A , we can have two possibilities: both servers have valid credentials, or the cluster is down. The corresponding data type is the functor $F^A \triangleq 1 + A \times A$. This functor is filterable if we can implement a lawful `filter` function. Begin writing code for `filter`:

```

1 type F[A] = Option[(A, A)]
2 def filter[A](p: A => Boolean): F[A] => F[A] = {
3   case None          => None
4   case Some((a1, a2)) => ???
5 }

```

In line 4, we need to compute a value of type F^A using the given values a_1 and a_2 . We need to check whether the predicate p holds for a_1 and a_2 . What if $p(a_1) == \text{false}$ but $p(a_2) == \text{true}$? We need to remove a_1 from the result, or else the filtering laws will not hold. But the functor $F^A \triangleq 1 + A \times A$ does not allow us to keep just one of the values of type A ; it requires two such values or none. So, we may return `Some((a2, a2))` or `None`.

Looking at the business requirements, we see that $p(a_1) == \text{false}$ means the first server's credentials expired. In that case, if $p(a_2) == \text{true}$, the first server copies the second server's valid credentials, a_2 . So, we must return `Some((a2, a2))`. Other cases are handled similarly. The full code is

```

def filter[A](p: A => Boolean): F[A] => F[A] = {
  case None          => None          // Cluster is down, no valid credentials.
  case Some((a1, a2)) => (p(a1), p(a2)) match {
    case (true, true)  => Some((a1, a2)) // Both credentials are still valid.
    case (true, false)  => Some((a1, a1)) // Server 2 copies credentials from server 1.
    case (false, true)  => Some((a2, a2)) // Server 1 copies credentials from server 2.
    case (false, false) => None          // Both credentials expired, the cluster is down.
  }
}

```

It remains to check that the filtering laws hold. The naturality law requires that the equation $f^{\uparrow F} \circ \text{filt}_F(p) = \text{filt}_F(f \circ p) \circ f^{\uparrow F}$ must hold for any values $s: F^A$, $f: A \rightarrow B$, and $p: B \rightarrow 2$. The code for `fmap` is

```

def fmap[A, B](f: A => B): F[A] => F[B] = {
  case None          => None
  case Some((a1, a2)) => Some((f(a1), f(a2)))
}

```

The code for the left-hand side of the law, $f^{\uparrow F} \circ \text{filt}_F(p)$, is written by composing `fmap` and `filter`:

```

fmap(f) andThen filter(p) == {
  case None          => None
  case Some((a1, a2)) => (p(f(a1)), p(f(a2))) match {
    case (true, true)  => Some((f(a1), f(a2)))
    case (true, false)  => Some((f(a1), f(a1)))
    case (false, true)  => Some((f(a2), f(a2)))
    case (false, false) => None
  }
}

```

The code for the right-hand side of the law, $\text{filt}_F(f \circ p) \circ f^{\uparrow F}$, is

```

filter(f andThen p) = {
  case None          => None
  case Some((a1, a2)) => (p(f(a1)), p(f(a2))) match {
    case (true, true)  => Some((a1, a2))
    case (true, false)  => Some((a1, a1))
    case (false, true)  => Some((a2, a2))
    case (false, false) => None
  }
} andThen fmap(f) = {
  case None          => None
  case Some((a1, a2)) => (p(f(a1)), p(f(a2))) match {
    case (true, true)  => Some((f(a1), f(a2)))
    case (true, false)  => Some((f(a1), f(a1)))
    case (false, true)  => Some((f(a2), f(a2)))
    case (false, false) => None
  }
}

```

Since the code is exactly the same, the law holds.

This computation illustrates why fully parametric functions such as `filter` obey the naturality law: such functions manipulate their arguments purely as symbols of unknown types, without referring

to any specific types or values. Applying a lifted function $f^{\uparrow F}$ before `filter` is the same as inserting `f(...)` around values of type `A` at every place in the code of `filter(p)` where a value of type `A` is used. Applying a lifted function $f^{\uparrow F}$ after `filter` is the same as inserting `f(...)` at every place where a value of type `A` is *returned*; but this does not put `f(...)` around values of type `A` used by the predicate `p`. To compensate, the right-hand side of the naturality law replaces the predicate `p` by the new predicate `f andThen p`, which inserts `f(...)` into the remaining places in the code.

This turns out to be the rule: fully parametric functions always satisfy a suitably formulated naturality law. We will see many more examples of naturality laws in this book, and we will find that the naturality law of a method `q` often has the form $f^{\uparrow}; q = \langle \text{modified } q \rangle; f^{\uparrow}$, i.e., the composition of a lifted function f^{\uparrow} and a method `q` can be interchanged, possibly after some modifications.

The identity law holds because the code of `filter(p)` is the identity function when `p = (_ \rightarrow \text{true})`,

```
filter[A](_ => true) == {
  case None      => None
  case Some((a1, a2))  => Some((a1, a2))
} == identity[F[A]]
```

It takes a bit more work to show that the composition law holds. We need to consider two cases: the cluster is down, or both servers have valid credentials.

In the first case, the value of F^A is `None` and remains `None` after any filtering operation. (If the cluster is down, a check of credentials will not bring it up.) So, the composition law holds for that case.

In the second case, we have two credentials a_1, a_2 in a value $s \triangleq \mathbb{0} + a_1 \times a_2$. The filtering operation $s \triangleright \text{filt}_F(p_1) \triangleright \text{filt}_F(p_2)$ will produce different results according to the values of the predicates p_1 and p_2 applied to a_1 and a_2 , which we can summarize as a table:

$p_1(a_1)$	$p_1(a_2)$	$p_2(a_1)$	$p_2(a_2)$	$p_{12}(a_1)$	$p_{12}(a_2)$	$s \triangleright \text{filt}_F(p_1)$	$s \triangleright \text{filt}_F(p_1) \triangleright \text{filt}_F(p_2)$	$s \triangleright \text{filt}_F(p_{12})$
<code>true</code>	<code>true</code>	<code>true</code>	<code>false</code>	<code>true</code>	<code>false</code>	$\mathbb{0} + a_1 \times a_2$	$\mathbb{0} + a_1 \times a_1$	$\mathbb{0} + a_1 \times a_1$
<code>true</code>	<code>false</code>	<code>true</code>	<code>true</code>	<code>true</code>	<code>false</code>	$\mathbb{0} + a_1 \times a_1$	$\mathbb{0} + a_1 \times a_1$	$\mathbb{0} + a_1 \times a_1$
<code>???</code>	<code>???</code>	<code>false</code>	<code>false</code>	<code>false</code>	<code>false</code>	<code>???</code>	$\mathbb{1} + \mathbb{0}^{\mathbb{A} \times \mathbb{A}}$	$\mathbb{1} + \mathbb{0}^{\mathbb{A} \times \mathbb{A}}$
<code>true</code>	<code>false</code>	<code>false</code>	<code>true</code>	<code>false</code>	<code>false</code>	$\mathbb{0} + a_1 \times a_1$	$\mathbb{1} + \mathbb{0}^{\mathbb{A} \times \mathbb{A}}$	$\mathbb{1} + \mathbb{0}^{\mathbb{A} \times \mathbb{A}}$
<code>false</code>	<code>false</code>	<code>???</code>	<code>???</code>	<code>false</code>	<code>false</code>	$\mathbb{1} + \mathbb{0}^{\mathbb{A} \times \mathbb{A}}$	$\mathbb{1} + \mathbb{0}^{\mathbb{A} \times \mathbb{A}}$	$\mathbb{1} + \mathbb{0}^{\mathbb{A} \times \mathbb{A}}$

Here we denoted $p_{12} \triangleq x \rightarrow p_1(x) \wedge p_2(x)$ for brevity. We see that the last two columns are always equal, which verifies the composition law. We omitted some rows from the table because the filtering code is completely symmetric with respect to interchanging a_1 and a_2 , and because the composition law is trivial when $p_1 = p_2$.

The partial function law holds because the code of `filter` will always remove the value `a1`, or `a2`, or both of them when the filtering predicate `p` returns `false` for any of those values.

So, we have seen that all filtering laws hold. If the business logic requirements change, the `filter` function will need to be implemented differently. For instance, the first server might be the only source of credentials: the second server may copy the first server's credentials if needed, but the cluster will go down whenever the first server's credentials expire. This corresponds to the code

```
def filter[A](p: A => Boolean): F[A] => F[A] = {
  case None      => None
  case Some((a1, a2))  => (p(a1), p(a2)) match {
    case (true, true)    => Some((a1, a2))
    case (true, false)   => Some((a1, a1))
    case (false, _)      => None    // The cluster is down if credentials expired for server 1.
  }
}
```

Alternatively, we may have a requirement that copying credentials between servers is not possible:

```
def filter[A](p: A => Boolean): F[A] => F[A] = {
  case None      => None
```

```

  case Some((a1, a2))  => (p(a1), p(a2)) match {
    case (true, true)    => Some((a1, a2))           // Both credentials are valid.
    case _                => None                  // The cluster is down if any credentials expired.
  }
}

```

The filtering laws will still hold with these alternative implementations. We omit the proofs.

Example 9.1.4.2 John can have up to 3 coupons, and Jill up to 2. All of John's coupons must be valid on purchase day, while each of Jill's coupons is checked independently. Implement a filterable functor describing this situation.

Solution We use a type parameter A for the type of "coupons". A data structure holding "up to 3" values of type A is written as $1 + A + A \times A + A \times A \times A$ and can be implemented in Scala by

```

sealed trait JohnsCoupons[A]           // This represents the type 1 + A + A × A + A × A × A.
final case class John0[A]()           extends JohnsCoupons[A]
final case class John1[A](c1: A)       extends JohnsCoupons[A]
final case class John2[A](c1: A, c2: A) extends JohnsCoupons[A]
final case class John3[A](c1: A, c2: A, c3: A) extends JohnsCoupons[A]

```

This code models John's coupons. Jill's coupons are implemented similarly,

```

sealed trait JillsCoupons[A]           // This represents the type 1 + A + A × A.
final case class Jill0[A]()           extends JillsCoupons[A]
final case class Jill1[A](c1: A)       extends JillsCoupons[A]
final case class Jill2[A](c1: A, c2: A) extends JillsCoupons[A]

```

The full data type is the product of John's and Jill's coupons.

```
final case class Coupons[A](johns: JohnsCoupons[A], jills: JillsCoupons[A])
```

It is convenient to define the filtering functions separately for `JohnsCoupons` and for `JillsCoupons`:

```

def filterJohn[A](p: A => Boolean): JohnsCoupons[A] => JohnsCoupons[A] = {
  case John0()           => John0() // We return John0() unless all coupons are valid.
  case John1(c1)         => if (p(c1)) John1(c1) else John0()
  case John2(c1, c2)     => if (p(c1) && p(c2)) John2(c1, c2) else John0()
  case John3(c1, c2, c3) => if (p(c1) && p(c2) && p(c3)) John3(c1, c2, c3) else John0()
}
def filterJill[A](p: A => Boolean): JillsCoupons[A] => JillsCoupons[A] = {
  case Jill0() => Jill0() // We must remove each invalid coupon but keep the rest.
  case Jill1(c1) => if (p(c1)) Jill1(c1) else Jill0()
  case Jill2(c1, c2) => (p(c1), p(c2)) match {
    case (true, true)  => Jill2(c1, c2)
    case (true, false) => Jill1(c1)
    case (false, true) => Jill1(c2)
    case (false, false) => Jill0()
  }
}

```

Now we can define the `filter` function for `Coupons`:

```
def filter[A](p: A => Boolean)(fa: Coupons[A]): Coupons[A] =
  Coupons(filterJohn(p)(fa.johns), filterJill(p)(fa.jills))
```

We will not prove the laws for the `filter` function because its code follows from general constructions derived later in this chapter. Running the `scalacheck` tests shows no failures:

```

implicit val filterableCoupons = new Filterable[Coupons] {
  def filt[A](p: A => Boolean)(fa: Coupons[A]): Coupons[A] = filter(p)(coupons)
}
checkFilteringLaws[Coupons, Int, String]           // Tests pass.

```

Example 9.1.4.3 A server receives a sequence of requests. Each request must be authenticated. Once a non-authenticated request is found, no further requests are accepted. Is this situation described by a filterable functor?

Solution We represent the requests by a `Seq[A]` wrapped in a `Server` type:

```
final case class Server[A](requests: Seq[A])
```

The filtering operation truncates the sequence when the predicate `p` first returns `false`:

```
def filter[A](p: A => Boolean)(s: Server[A]): Server[A] = Server(s.requests.takeWhile(p))
```

We will not prove the laws because this implementation reduces to general constructions of filterable functors (see page 317 below). Intuitively, we expect laws to hold because the `filter` function always removes values that fail the predicate `p`. The filtering function also removes other values that may or may not fail the predicate, but the filtering laws allow removing *more* values.

Example 9.1.4.4 If possible, implement a `Filterable` typeclass instance for:

(a) The functor F^T defined by the Scala code

```
final case class F[T](x: Option[T], yy: Option[(T, T)])
```

(b) $F^A \triangleq \text{Int} + \text{Int} \times A + \text{Int} \times A \times A + \text{Int} \times A \times A \times A$.

(c) A non-empty list functor, $F^A = \text{NEList}^A$, defined recursively as $F^A \triangleq A + A \times F^A$.

(d) $F^{Z,A} \triangleq Z + \text{Int} \times Z \times A \times A$ (with respect to the type parameter A).

(e) $F^{Z,A} \triangleq Z + \text{Int} \times A \times \text{List}^A$ (with respect to the type parameter A).

Solution (a) The functor F is written in the code notation as $F^T \triangleq (\mathbb{1} + T) \times (\mathbb{1} + T \times T)$. This is a product of `Option[T]`, which is filterable, and the functor $\mathbb{1} + T \times T$, which was shown to be filterable in Example 9.1.4.1. We can apply the corresponding `filter` operation to each part of the product:

```
def filter[T](p: T => Boolean): F[T] => F[T] = {
  case F(t, None)          => F(t.filter(p), None)
  case F(t, Some((t1, t2))) => F(t.filter(p), if (p(t1) && p(t2)) Some((t1, t2)) else None)
}
```

Each part of the product type satisfies the filtering laws separately (see Statement 9.2.4.2 below).

(b) The functor F is equivalent to $F^A \cong \text{Int} \times (\mathbb{1} + A + A \times A + A \times A \times A) = \text{Int} \times \text{JohnsCoupons}^A$, where we used the functor `JohnsCoupons` from Example 9.1.4.2. So we use the same filtering operation for `JohnsCoupons` as in Example 9.1.4.2, while keeping the `Int` value unchanged:

```
final case class F[A](n: Int, johns: JohnsCoupons[A])
def filter[A](p: A => Boolean)(fa: F[A]): F[A] = fa.copy(johns = filterJohn(p)(johns))
```

An interesting alternative implementation of `filter` uses the integer value `n` for tracking the total number of data items *removed* by filtering:

```
def filter[A](p: A => Boolean)(fa: F[A]): F[A] = {
  val (removed, newJohnsCoupons) = fa.johns match {
    case John0()          => (0, John0())
    case John1(c1)         => if (p(c1)) (0, John1(c1)) else (1, John0())
    case John2(c1, c2)      => if (p(c1) && p(c2)) (0, John2(c1, c2)) else (2, John0())
    case John3(c1, c2, c3)  => if (p(c1) && p(c2) && p(c3)) (0, John3(c1, c2, c3)) else (3, John0())
  }
  F(fa.n + removed, newJohnsCoupons)
}
```

The new code still satisfies the filtering laws (we omit the proof).

(c) The type definition of `NEList` is

```
sealed trait NEList[A]
final case class Last[A](x: A)           extends NEList[A]
final case class More[A](x: A, tail: NEList[A]) extends NEList[A]
```

We find that we *cannot* implement the type signature of `filter`:

```
1 def filter[A](p: A => Boolean): NEList[A] => NEList[A] = {
2   case Last(x)      => if (p(x)) ??? else ??? // Need to compute a value of type NEList[A] here.
3   case More(x, tail) => ???
4 }
```

The problem is in line 2 above when `p(x) == false`: we need to remove the value `x`, making the list empty, but the type `NEList[A]` disallows empty lists. So, line 2 must always return a list containing the value `x`, which violates the partial function law of filtering. We conclude that `NEList` is not filterable.

(d) Looking at the type expression $F^{Z,A} \triangleq Z + \text{Int} \times Z \times A \times A$, we need to check whether we could remove the values of type A that do not pass the filter. If none of the two values of type A within $\text{Int} \times Z \times A \times A$ pass the filter, we will need to remove both of them and to return a value of type Z . Luckily, we have a value of type Z within $\text{Int} \times Z \times A \times A$. So we can implement `filter` e.g., like this,

```
type F[A] = Either[Z, (Int, Z, A, A)]           // The type 'Z' must be already defined.
def filter[A](p: A => Boolean): F[A] = {
  case Left(z)          => Left(z)
  case Right((n, z, a1, a2)) => if (p(a1) && p(a2)) Right((n, z, a1, a2))
                                else Left(z) // If anything fails the filter, use 'z'.
}
```

The filtering laws will hold similarly to Example 9.1.4.1.

(e) The `filter` function must be able to remove values of type A from $F^{Z,A} \triangleq Z + \text{Int} \times A \times \text{List}^A$. What if *no* values of type A pass the filter? If we do not remove all those values of type A from the data structure, we will violate the partial function law. So, the only choice in that case is to return a value of type Z . But Z is a type parameter, and we cannot create values of type Z from scratch. We conclude that $F^{Z,A}$ is not filterable with respect to A .

```
type F[A] = Either[Z, (Int, A, List[A])]
def filter[A](p: A => Boolean): F[A] = {
  case Left(z)          => Left(z)
  case Right((n, a, as)) => // What to compute in case p(a) == false and as.filter(p) == List() ?
                                ??? // In that case, we will have neither values of type A nor of type Z.
}
```

9.1.5 Exercises: Programming with filterable functors

Exercise 9.1.5.1 Confucius gave wisdom on each of the 7 days of a week. Sometimes the wise proverbs were hard to remember. If Confucius forgets what he said on a given day, he also forgets what he said on all the *previous* days of the week. Is this situation described by a filterable functor?

Exercise 9.1.5.2 Define an extension method `evenFilter(p)` on the type `IndexedSeq[T]` such that a value `x: T` is kept in the sequence if `p(x) == true` and only if the initial sequence has an *even* total number of elements `y` for which `p(y) == false`. Does `evenFilter` define a lawful filterable functor?

Exercise 9.1.5.3 If possible, implement the `filter` function or a `Filterable` typeclass instance (law checking is optional) for:

(a) The functor $Q^{A,Z}$ with respect to the type parameter A , where $Q^{\bullet,\bullet}$ is defined by this Scala code:

```
final case class Q[A, Z](id: Long, user1: Option[(A, Z)], user2: Option[(A, Z)])
```

(b) The functor R^A defined by the Scala code

```
final case class R[A](x: Int, y: Int, z: A, data: List[A])
```

where the functor `List` already has the method `filter` defined in the standard library.

(c) $F^P, Q^A \triangleq (P \rightarrow P) + (Q \rightarrow Q) \times A \times A \times A$.

(d) The functor MyTree^A defined recursively as $\text{MyTree}^A \triangleq \mathbb{1} + A \times A \times \text{MyTree}^A \times \text{MyTree}^A$.

Exercise 9.1.5.4 Three times each week (on Mondays, Wednesdays, and Fridays) a data set is collected and a test is run. If a test fails, the corresponding data set may be discarded. Is this situation described by a filterable functor if we impose one of the following additional requirements:

(a) All data sets for a given week are discarded if at least *two* of the tests failed that week?

(b) All data sets for a given week are retained if at least *two* of the tests passed that week?

Exercise 9.1.5.5 Implement a `filter` function for a regular-shaped tree R^A defined by:

- (a) $R^A \triangleq \mathbb{1} + A + R^{A \times A}$.
 (b) $R^A \triangleq A + R^{(\mathbb{1} + A) \times (\mathbb{1} + A)}$.

Is the simplest regular-shaped tree R^A defined by $R^A \triangleq A + R^{A \times A}$ filterable?

9.2 Laws and structure

9.2.1 Simplifying the filtering laws: Motivation for `deflate`

The four laws of `filter` (Section 9.1.2) require considerable work to verify directly. Is there a shorter reformulation of the laws that is easier to remember and to verify?

To motivate such a reformulation, begin by considering a heuristic picture of a filter operation $\text{filt}_F(p) : F^A \rightarrow F^A$ that may remove some values of type A from a wrapper F^A . For example, if $F^A \triangleq \mathbb{1} + A$, the filtering operation may remove the value of type A and return the result $\mathbb{1} + \mathbb{0}$. It appears that the filtering operation, at the type level, replaces the type A by the unit type $\mathbb{1}$ whenever a value does not pass the filtering condition. Indeed, in examples of non-filterable functors F , such as $F^A \triangleq A \times (\mathbb{1} + A)$, the type expression for F^A cannot accommodate replacing the type A by $\mathbb{1}$.

If we had the functor $F^{\mathbb{1}+A}$ instead of F^A , we would be able to replace A by $\mathbb{1}$ whenever necessary:

```
def filter[F[_]: Functor, A](p: A => Boolean): F[Option[A]] => F[Option[A]] =
  _.map(_.filter(p)) // Using the standard .filter method on Option.
```

We can always convert a value of type F^A into a value of type $F^{\mathbb{1}+A}$:

```
def inflate[F[_]: Functor, A]: F[A] => F[Option[A]] =
  _.map(x => Some(x))
```

The code notation for the function `inflate` is
 $\text{inflate}^{F,A} \triangleq (x^A \rightarrow \mathbb{0}^{\mathbb{1}} + x)^{\uparrow F}$.

It remains to convert $F^{\mathbb{1}+A}$ to F^A . If we could *somewhat* do that, say, via a function `deflate`,

```
def deflate[F[_], A]: F[Option[A]] => F[A] = ???
```

$\text{deflate}^{F,A} : F^{\mathbb{1}+A} \rightarrow F^A$,

we would then express `filter` through `map` and `deflate` like this,

$$(\text{filt}_F(p))^{\mathbb{1}+A \rightarrow F^A} = \text{inflate} \circ (\text{filt}_{\text{Opt}}(p))^{\uparrow F} \circ \text{deflate} \quad .$$

$$F^A \xrightarrow{\text{inflate}} F^{\mathbb{1}+A} \xrightarrow{(\text{filt}_{\text{Opt}}(p))^{\uparrow F}} F^{\mathbb{1}+A} \xrightarrow{\text{deflate}} F^A$$

Here filt_{Opt} is the standard `filter` method defined for the `Option[_]` types.

We notice that both functions in the composition $\text{inflate} \circ (\text{filt}_{\text{Opt}}(p))^{\uparrow F}$ are some lifted functions in the functor F , and so we can simplify that composition to a single lifted function,

$$\begin{aligned} & \text{inflate} \circ (\text{filt}_{\text{Opt}}(p))^{\uparrow F} \\ \text{definition of inflate : } &= (x^A \rightarrow \mathbb{0}^{\mathbb{1}} + x)^{\uparrow F} \circ (\text{filt}_{\text{Opt}}(p))^{\uparrow F} \\ \text{functor composition law of } F : &= (x \rightarrow \text{filt}_{\text{Opt}}(p)(\mathbb{0} + x))^{\uparrow F} \quad . \end{aligned}$$

We will need to use this function often, so let us call it $\psi(p)$ or even shorter, ψ_p , for convenience:

```
def psi[A](p: A => Boolean): A => Option[A] =
  x => Some(x).filter(p)
```

$$\begin{aligned} \psi^A : (A \rightarrow \mathbb{2}) \rightarrow A \rightarrow \mathbb{1} + A \quad , \\ \psi_p \triangleq \psi(p^{\mathbb{A} \rightarrow \mathbb{2}}) \triangleq x^A \rightarrow \text{filt}_{\text{Opt}}(p)(\mathbb{0} + x) \quad . \end{aligned}$$

Using the function ψ , we can express the `filter` operation as

$$F^A \xrightarrow{\psi_p^{\uparrow F}} F^{\mathbb{1}+A} \xrightarrow{\text{deflate}} F^A \quad \text{filt}_F(p) = \psi_p^{\uparrow F} \circ \text{deflate} \quad . \quad (9.7)$$

```
def filter[A](p: A => Boolean)(fa: F[A]): F[A] = deflate(fa.map(psi(p)))
```

We derived the code for `filter` assuming that a suitable function `deflate` exists. Can we derive the code for `deflate` for a given filterable functor F , e.g., for $F = \text{Seq}$? The required type signature is `deflate: Seq[Option[T]] => Seq[T]`. The Scala library has a `flatten` method for `Seq` with exactly that type signature; it removes all empty `Option` values from a sequence. This shows how to derive the `deflate` method for any filterable F : use F 's `filter` to remove the empty `Option` values from F^{1+A} .

```
def deflate[F[_]: Filterable : Functor, A]: F[Option[A]] => F[A] =  
  _.filter(_.nonEmpty).map(_.get)
```

$$\text{deflate} : F^{1+A} \xrightarrow{\text{filt}_F(\text{nonEmpty})} F^{1+A} \xrightarrow{\text{get}^{1F}} F^A$$

$$\text{deflate}^{F^{1+A} \rightarrow F^A} = \text{filt}_F(\text{nonEmpty}) \circ \text{get}^{1F} \quad . \quad (9.8)$$

The method `get` is a partial function and will fail on empty `Option` values. However, it is safe to use `get1F` here, because the partial function law of filtering guarantees that `.filter(_.nonEmpty)` will prevent `get` from being applied to empty `Option` values. Because of this, `deflate` is a *total* function when defined by Eq. (9.8), as long as F 's `filter` obeys its partial function law.

The ability to express `filter` via `deflate` means that a functor's filtering logic is fully described as long as we know how to exclude empty `Option` values from F^{1+A} and convert the result to F^A . We can define the `Filterable` typeclass through `deflate`, providing `filter` and `deflate` as extension methods:

```
abstract class Filterable[F[_]: Functor] { // Need a 'Functor' instance to use '.map'.  
  def deflate[A]: F[Option[A]] => F[A] // Typeclass instances will implement 'deflate'.  
  def filt[A](p: A => Boolean)(fa: F[A]): F[A] = deflate(fa.map(x => Some(x).filter(p)))  
} // Typeclass instances don't have to implement 'filt', but may override it for performance reasons.  
  
implicit class FilterableSyntax[F[_]: Filterable, A](fa: F[Option[A]]) {  
  def deflate: F[A] = implicitly[Filterable[F]].deflate(fa)  
  def withFilter(p: A => Boolean): F[A] = implicitly[Filterable[F]].filt(p)(fa)  
}
```

Example 9.2.1.1 Use `deflate` to implement a `Filterable` instance for the functor $F^A \triangleq Z \rightarrow \text{List}^A$.

Solution The type signature of `deflate` is implemented as

```
type F[A] = Z => List[A] // The type Z should have been defined before.  
def deflateF[A](fa: F[Option[A]]): F[A] = { z => fa(z).flatten }
```

We used Scala's library method `flatten` with the type signature `Seq[Option[A]] => Seq[A]`. This method is defined for all sequences, similarly to the `flatten` method that has type `Seq[Seq[A]] => Seq[A]`.

Using this function, we implement the typeclass instance as

```
implicit val filterableF = new Filterable[F] {  
  def deflate[A]: F[Option[A]] => F[A] = deflateF  
}
```

Example 9.2.1.2 Implement `deflate` for the functor $F^A \triangleq A \times A + (Z \rightarrow Z)$.

Solution The type signature of `deflate` is $F^{1+A} \rightarrow F^A$ and can be implemented e.g., as

```
type F[A] = Either[(A, A), Z => Z] // The type Z should have been defined before.  
def deflateF[A]: F[Option[A]] => F[A] = { // Pattern-match on Either[Option[A], Option[A]], Z => Z].  
  case Left((Some(a1), Some(a2))) => Left((a1, a2)) // Both values pass the filter.  
  case Left(_) => Right(identity) // We can use a fixed value of type Z => Z.  
  case Right(zz) => Right(zz)  
}
```

This code implements a “greedy” filter that requires both values in the pair $A \times A$ to satisfy the predicate. Otherwise, the filter returns the special “empty” value $\emptyset^{A \times A} + \text{id}^{Z \rightarrow Z}$ of type F^A .

Example 9.2.1.3 Use `deflate` to verify that the functor $F^A \triangleq A + A \times A \times \text{String}$ is not filterable.

Solution We begin by checking whether the type signature of `deflate` can be implemented:

$$\text{deflate}_F : \mathbb{1} + A + (\mathbb{1} + A) \times (\mathbb{1} + A) \times \text{String} \rightarrow A + A \times A \times \text{String} \quad .$$

An immediate problem is that we need to map all disjunctive cases, including $\mathbb{1} + 0 + 0$, into a value of type F^A , which contains values of type A in every disjunctive case. So, implementing `deflate`

requires us to produce a value of type A from scratch, $\forall A. \mathbb{1} \rightarrow A$, which is impossible in a fully parametric function. Since `deflate` is not implementable, the functor F is not filterable.

These examples show that `deflate` is easier to implement and to reason about than `filter`.

9.2.2 Equivalence of filter and deflate

We have expressed `filter` through `deflate` by Eq. (9.7) and `deflate` through `filter` by Eq. (9.8). Are `deflate` and `filter` computationally equivalent? It turns out that Eqs. (9.7)–(9.8) are inverses of each other only if we assume certain laws.

Statement 9.2.2.1 Begin with a filterable functor F 's function `filter`, define `deflate` through Eq. (9.8), and then define a new function `filter'` via Eq. (9.7):

$$\text{deflate} = \text{filt}_F(\text{nonEmpty}) \circ \text{get}^{\uparrow F} , \quad \text{filter}'(p) = \psi_p^{\uparrow F} \circ \text{deflate} .$$

Then `filter'` is equal to `filter`, assuming that the partial function law holds.

Proof We need to show that $\text{filt}_F(p)$ is the same as $\text{filter}'(p)$ for any predicate $p : A \rightarrow \mathbb{2}$:

$$\begin{aligned} \text{expect to equal } \text{filt}_F(p) : & \quad \text{filter}'(p) = \psi_p^{\uparrow F} \circ \text{filt}_F(\text{nonEmpty}) \circ \text{get}^{\uparrow F} \\ \text{naturality law of } \text{filt}_F : & \quad = \text{filt}_F(\psi_p \circ \text{nonEmpty}) \circ \psi_p^{\uparrow F} \circ \text{get}^{\uparrow F} \\ \text{composition law of } F : & \quad = \text{filt}_F(\psi_p \circ \text{nonEmpty}) \circ (\psi_p \circ \text{get})^{\uparrow F} . \end{aligned} \quad (9.9)$$

To proceed with the calculation, we need to simplify the two expressions $\psi_p \circ \text{nonEmpty}$ and $\psi_p \circ \text{get}$. Begin with writing the code for the standard methods `nonEmpty` and `get`, using the equivalent type `Option[Unit]` (i.e., $\mathbb{1} + \mathbb{1}$) instead of `Boolean` (i.e., $\mathbb{2}$):

```
def nonEmpty[A]: Option[A] => Option[Unit] = _.map(_ => ()) // Option[Unit] ≈ Boolean
def get[A]: Option[A] => A = { case Some(a) => a }
```

$$\text{nonEmpty} \triangleq \begin{array}{c|cc} & \mathbb{1}(\text{false}) & \mathbb{1}(\text{true}) \\ \hline \mathbb{1} & \text{id} & 0 \\ A & 0 & (_ : A \rightarrow \mathbb{1}) \end{array} = \begin{array}{c|c} & 2 \\ \hline \mathbb{1} & _ \rightarrow \text{false} \\ A & _ \rightarrow \text{true} \end{array} , \quad \text{get} : \mathbb{1} + A \rightarrow A \triangleq \begin{array}{c|cc} & A \\ \hline \mathbb{1} & 0 \\ A & \text{id} \end{array} . \quad (9.10)$$

These methods are fully parametric since their code is defined purely through the eight standard code constructions of functional programming (see Section 5.2.3), with no externally defined types such as `Int` or `String`. The function ψ is also fully parametric since we can write it as

```
def psi[A](p: A => Option[Unit]): A => Option[A] = x => p(x).map(_ => x) // Option[Unit] ≈ Boolean
```

$$\text{view } p : A \rightarrow \mathbb{2} \text{ as having type } \mathbb{1} + \mathbb{1} : \quad x : A \triangleright \psi_p \triangleq p(x) \triangleright \begin{array}{c|cc} & \mathbb{1} & A \\ \hline \mathbb{1}(\text{false}) & \text{id} & 0 \\ \mathbb{1}(\text{true}) & 0 & \mathbb{1} \rightarrow x \end{array} .$$

We may write that code equivalently as $\text{nonEmpty} \triangleq (_ : A \rightarrow \mathbb{1})^{\uparrow \text{Opt}}$ and $\psi_p(x) \triangleq x \triangleright p \circ (1 \rightarrow x)^{\uparrow \text{Opt}}$.

Now we compute the function compositions we need. First, we show that $\psi_p \circ \text{nonEmpty} = p$:

$$\begin{aligned} \text{expect to equal } x \triangleright p : & \quad x : A \triangleright \psi_p \circ \text{nonEmpty} \\ & \quad = x \triangleright p \circ (1 \rightarrow x)^{\uparrow \text{Opt}} \circ (_ : A \rightarrow \mathbb{1})^{\uparrow \text{Opt}} \\ \text{composition under } \uparrow \text{Opt} : & \quad = x \triangleright p \circ (1 \rightarrow 1)^{\uparrow \text{Opt}} \\ \text{identity function of type } \mathbb{1} \rightarrow \mathbb{1} : & \quad = x \triangleright p \circ (\text{id} : \mathbb{1} \rightarrow \mathbb{1})^{\uparrow \text{Opt}} \\ \text{identity law of Opt} : & \quad = x \triangleright p \circ \text{id} = x \triangleright p . \end{aligned} \quad (9.11)$$

The expression $\psi_p \circ \text{get}$ is simplified to the partial function $\text{id}_{|p}$ by matrix composition:

$$\begin{aligned}
 \text{definitions of } \psi_p \text{ and get : } \underline{x : A \triangleright \psi_p \circ \text{get}} &= p(x) \triangleright \begin{array}{c|c|c|c} & & 1 & A \\ \hline 1 & \text{(false)} & \text{id} & 0 \\ \hline 1 & \text{(true)} & 0 & 1 \rightarrow x \end{array} ; \begin{array}{c|c|c|c} & & 1 & 0 \\ \hline 1 & & A & \text{id} \\ \hline 0 & & & \end{array} \\
 \text{matrix composition : } &= p(x) \triangleright \begin{array}{c|c|c} & & A \\ \hline 1 & \text{(false)} & 0 \\ \hline 1 & \text{(true)} & 1 \rightarrow x \end{array} \\
 \text{use Eq. (9.2) as definition of } |_p : &= x \triangleright \text{id}_{|p} .
 \end{aligned} \tag{9.12}$$

The same derivation performed in Scala syntax looks like this:

```

psi(p)(x).get      // Expand the code for 'psi' and 'get':
== p(x) match {
    case false  => None
    case true   => Some(x)
} match {
    case Some(x) => x
}                      // Compute function composition:
== p(x) match { case true => x }      // Rewrite this code equivalently as
== x match { case x if p(x) => x }      // x \triangleright \text{id}_{|p}
    
```

We can now finish the calculation in Eq. (9.9):

$$\begin{aligned}
 \text{expect to equal } \text{filt}_F(p) : & \text{filter}'(p) = \text{filt}_F(\underline{\psi_p \circ \text{nonEmpty}}) \circ (\underline{\psi_p \circ \text{get}})^{\uparrow F} \\
 \text{simplify using Eqs. (9.11)–(9.12) : } &= \text{filt}_F(p) \circ \underline{\text{id}_{|p}}^{\uparrow F} \\
 \text{partial function law (9.6) of } \text{filt}_F : &= \text{filt}_F(p) \circ \underline{\text{id}}^{\uparrow F} \\
 \text{identity law of } F : &= \text{filt}_F(p) \circ \text{id} = \text{filt}_F(p) .
 \end{aligned}$$

So, the new function `filter'` equals the original `filter` function.

Statement 9.2.2.2 Beginning with a given `deflate` function, define the corresponding `filter` function via Eq. (9.7) and then use Eq. (9.8) to define a new function `deflate'`:

$$\text{filt}_F(p) = \psi_p^{\uparrow F} \circ \text{deflate} , \quad \text{deflate}' = \text{filt}_F(\text{nonEmpty}) \circ \text{get}^{\uparrow F} .$$

Then `deflate'` is equal to `deflate`, assuming that a suitable naturality law holds (Eq. (9.14) below).

Proof Try showing that the new function `deflate'` is equal to `deflate`:

$$\text{deflate}' = \text{filt}_F(\text{nonEmpty}) \circ \text{get}^{\uparrow F} = \psi_{\text{nonEmpty}}^{\uparrow F} \circ \text{deflate} \circ \text{get}^{\uparrow F} \stackrel{?}{=} \text{deflate} . \tag{9.13}$$

The derivation is stuck here: we cannot simplify the last expression unless we can somehow switch the order of function compositions so that $\psi_{\text{nonEmpty}}^{\uparrow F}$ and $\text{get}^{\uparrow F}$ are placed together and the functor composition law of F can be applied. To achieve that, we need a law that switches the order of lifted function compositions around `deflate`. The naturality law (9.3) of `filter` has that form, so we can try

$$\begin{array}{ccc}
 F^{1+A} & \xrightarrow{\text{deflate}} & F^A \\
 & \downarrow f^{\uparrow F} & \\
 ??? & \xrightarrow{\text{deflate}} & F^B
 \end{array}
 \quad \begin{array}{l}
 \text{deriving a similar naturality law for } \text{deflate}. \text{ To switch the order of composition of } \text{deflate} \text{ with a lifted function, the law must have the form} \\
 \text{deflate} \circ f^{\uparrow F} = (???)^{\uparrow F} \circ \text{deflate} ,
 \end{array}$$

as illustrated by the type diagram at left. The types match in the right-hand side only if the argument of `deflate` is of type F^{1+B} . So, the law must have the form $\text{deflate} \circ f^{\uparrow F} = (???:^{1+A \rightarrow 1+B})^{\uparrow F} \circ \text{deflate}$. The typed hole $???:^{1+A \rightarrow 1+B}$ must be filled with a value, say, $g^{1+A \rightarrow 1+B}$, which is somehow related to f . The only way to obtain g is to lift the function f to the `Option` functor, i.e., to define $g \triangleq f^{\uparrow \text{Opt}}$.

So, the **naturality law** of `deflate` is

$$\begin{array}{ccc} F^{\mathbb{1}+A} & \xrightarrow{\text{deflate}} & F^A \\ \downarrow (f^{\text{Opt}})^F & & \downarrow f^F \\ F^{\mathbb{1}+B} & \xrightarrow{\text{deflate}} & F^B \end{array} \quad \text{deflate} ; f^{\text{Opt}} = f^{\text{Opt}} ; \text{deflate} \quad , \quad (9.14)$$

where $f : A \rightarrow B$ is an arbitrary function.

Assuming that this naturality law holds for `deflate`, we can continue the derivation in Eq. (9.13) towards showing that `deflate'` equals `deflate`:

$$\begin{aligned} \text{expect to equal } \text{deflate} : \quad & \text{deflate}' = \psi_{\text{nonEmpty}}^F ; \text{deflate} ; \text{get}^F \\ \text{naturality law (9.14) of } \text{deflate} : \quad & = \psi_{\text{nonEmpty}}^F ; \text{get}^{\text{Opt}} ; \text{deflate} \\ \text{composition law of } F : \quad & = (\psi_{\text{nonEmpty}} ; \text{get}^{\text{Opt}})^F ; \text{deflate} \quad . \end{aligned} \quad (9.15)$$

We now need to perform a separate calculation in order to simplify the function $\psi_{\text{nonEmpty}} ; \text{get}^{\text{Opt}}$, which must be a function of type $\mathbb{1} + A \rightarrow \mathbb{1} + A$ for the types to match. Writing this calculation in Scala syntax, we obtain (skipping some steps for brevity)

```
psi[Option[A]](nonEmpty)(x).map(get) == (nonEmpty(x) match {
  case false  => None
  case true   => Some(x)
}) .map { case Some(y) => y } == // Expand code for 'nonEmpty'.
( ( x match {
  case None   => false
  case Some(_) => true
}) match {
  case false   => None
  case true    => Some(x) // This will be of the form Some(Some(y)).
}) .map { case Some(y) => y } == // Compute all function compositions.
  x match {
  case None   => None
  case Some(y) => Some(y)
} == x // Identity function applied to 'x'.
```

To perform the same calculation in the code notation, we first prepare a formula for the lifting operation ${}^{\text{Opt}}$ of the `Option` functor:

$$(h : C \rightarrow D)^{\text{Opt}} = \begin{vmatrix} & \mathbb{1} & D \\ \mathbb{1} & \text{id} & \mathbf{0} \\ C & \mathbf{0} & h \end{vmatrix} \quad , \text{ so} \quad \text{get}^{\text{Opt}} = \begin{vmatrix} & \mathbb{1} & A \\ \mathbb{1} & \text{id} & \mathbf{0} \\ \mathbb{1} + A & \mathbf{0} & \text{get} \end{vmatrix} = \begin{vmatrix} & \mathbb{1} & A \\ \mathbb{1} & \text{id} & \mathbf{0} \\ \mathbb{1} & \mathbf{0} & \mathbf{0} \\ A & \mathbf{0} & \text{id} \end{vmatrix} \quad , \quad (9.16)$$

where we expanded the matrix to accommodate the disjunctive type $\mathbb{1} + A$. Then we compute

$$\begin{aligned} x : \mathbb{1} + A \triangleright \psi_{\text{nonEmpty}} &= \text{nonEmpty}(x) \triangleright \begin{vmatrix} & \mathbb{1} & \mathbb{1} + A \\ \mathbb{1} & \text{id} & \mathbf{0} \\ \mathbb{1} & \mathbf{0} & 1 \rightarrow x \end{vmatrix} \\ \text{definition (9.10) of } \text{nonEmpty} : \quad &= x \triangleright \begin{vmatrix} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \text{id} & \mathbf{0} \\ A & \mathbf{0} & _ \rightarrow 1 \end{vmatrix} ; \begin{vmatrix} & \mathbb{1} & \mathbb{1} & A \\ \mathbb{1} & \text{id} & \mathbf{0} & \mathbf{0} \\ \mathbb{1} & \mathbf{0} & 1 \rightarrow x & \mathbf{0} \\ A & \mathbf{0} & _ \rightarrow x & \text{id} \end{vmatrix} \\ \text{matrix composition} : \quad &= x \triangleright \begin{vmatrix} & \mathbb{1} & \mathbb{1} + A \\ \mathbb{1} & \text{id} & \mathbf{0} \\ A & \mathbf{0} & _ \rightarrow x \end{vmatrix} = x \triangleright \begin{vmatrix} & \mathbb{1} & \mathbb{1} & A \\ \mathbb{1} & \text{id} & \mathbf{0} & \mathbf{0} \\ A & \mathbf{0} & \mathbf{0} & \text{id} \end{vmatrix} \quad , \end{aligned}$$

where the last matrix uses the fact that x matches the type $0 + A$ in the bottom row. Finally,

$$\begin{aligned}
 x^{\mathbb{1}+A} \triangleright \psi_{\text{nonEmpty}} \circ \text{get}^{\text{Opt}} &= x \triangleright \left| \begin{array}{c|cc|c} & 1 & 1 & A \\ \hline 1 & \text{id} & 0 & 0 \\ A & 0 & 0 & \text{id} \end{array} \right| \circ \text{get}^{\text{Opt}} \\
 &= x \triangleright \left| \begin{array}{c|cc|c} & 1 & A \\ \hline 1 & \text{id} & 0 \\ 1 & 0 & 0 \\ A & 0 & \text{id} \end{array} \right| = x \triangleright \left| \begin{array}{c|cc|c} & 1 & A \\ \hline 1 & \text{id} & 0 \\ A & 0 & \text{id} \end{array} \right| = x \triangleright \text{id} = x \quad . \quad (9.17)
 \end{aligned}$$

So, we find $\psi_{\text{nonEmpty}} \circ \text{get}^{\text{Opt}} = \text{id}$ and thus Eq. (9.15) gives $\text{deflate}' = \text{deflate}$.

We conclude that `filter` and `deflate` are computationally equivalent as long as the partial function law (9.6) holds for `filter` and the naturality law (9.14) holds for `deflate`.

Statement 9.2.2.3 The partial function law always holds for the `filter` function defined via `deflate`.

Proof Assume that a `deflate` function is given, and define `filter` through Eq. (9.7). Then the partial function law (9.6) is transformed into an equation illustrated by the following diagram:

$$\begin{array}{ccccc}
& \psi_p^F & \nearrow & & \\
F^A & \swarrow & F^{\mathbb{1}+A} & \xrightarrow{\text{deflate}} & F^A \xrightarrow{(f:A \rightarrow B)^F} F^B \\
& \psi_p^F & \nearrow & & \\
& \swarrow & F^{\mathbb{1}+A} & \xrightarrow{\text{deflate}} & F^A \xrightarrow{(f_p^{A \rightarrow B})^F} F^B
\end{array}
\quad \quad \quad
\psi_p^F ; \text{deflate} ; f^F = \psi_p^F ; \text{deflate} ; f|_p^F$$

To show that the law holds, we transform both sides using the naturality law (9.14):

$$\begin{aligned} \text{left-hand side : } & \psi_p^F ; \underline{\text{deflate}} ; f^{\uparrow F} = \underline{\psi_p^F ; f^{\uparrow \text{Opt}} ; \uparrow F} ; \text{deflate} = (\psi_p ; f^{\uparrow \text{Opt}})^{\uparrow F} ; \text{deflate} \\ \text{right-hand side : } & \psi_p^F ; \text{deflate} ; f_p^{\uparrow F} = \psi_p^F ; f_p^{\uparrow \text{Opt}} ; \uparrow F ; \text{deflate} = (\psi_p ; f_p^{\uparrow \text{Opt}})^{\uparrow F} ; \text{deflate} \end{aligned} .$$

It remains to show that $\psi_p; f \uparrow^{\text{Opt}}_{p_1} = \psi_p; f \uparrow^{\text{Opt}}$. Apply the function $\psi_p; f \uparrow^{\text{Opt}}$ to an x^A ,

```

psi(p)(x).map(f) == (p(x) match {
  case false    => None
  case true     => Some(x)
}) match {
  case None     => None
  case Some(y)  => Some(f(y))
} == p(x) match {
  case false    => None
  case true     => Some(f(x))
}

```

The same calculation in the code notation looks like this:

$$x^A \triangleright \psi_p \circ f^{\text{Opt}} = p(x) \triangleright \begin{array}{c|cc} & 1 & A \\ \hline 1 (\text{false}) & \text{id} & 0 \\ 1 (\text{true}) & 0 & 1 \rightarrow x \end{array} \circ \begin{array}{c|cc} & 1 & A \\ \hline 1 & \text{id} & 0 \\ A & 0 & f \end{array} = p(x) \triangleright \begin{array}{c|cc} & 1 & A \\ \hline 1 (\text{false}) & \text{id} & 0 \\ 1 (\text{true}) & 0 & 1 \rightarrow f(x) \end{array} .$$

In the last expression, the function f is applied to x only when $p(x) = \text{true}$. So, the result will be the same if we replace $f(x)$ by the partial function $f|_p(x)$, which was defined by Eq. (9.2) to be equal to $f(x)$ when $p(x)$ holds. It follows that $\psi_p \circ f|_p^{\text{Opt}} = \psi_p \circ f^{\text{Opt}}$, concluding the proof.

The equivalence between `filter` and `deflate` extends to their respective naturality laws:

Statement 9.2.2.4 If the naturality law (9.3) holds for `filter`, the law (9.14) will also hold for `deflate` if it is defined through `filter` by Eq. (9.8).

Proof Compare the two sides of the law (9.14) if `deflate` is defined by Eq. (9.8):

$$\text{left-hand side of Eq. (9.14)} : f^{\uparrow \text{Opt} \uparrow F} ; \underline{\text{deflate}} = f^{\uparrow \text{Opt} \uparrow F} ; \underline{\text{filt}_F(\text{nonEmpty})} ; \text{get}^{\uparrow F}$$

$$\text{naturality law (9.3) of } \text{filt}_F : = \text{filt}_F(f^{\uparrow \text{Opt}} ; \text{nonEmpty}) ; f^{\uparrow \text{Opt} \uparrow F} ; \text{get}^{\uparrow F} .$$

$$\text{right-hand side of Eq. (9.14)} : \underline{\text{deflate}} ; f^{\uparrow F} = \text{filt}_F(\text{nonEmpty}) ; \text{get}^{\uparrow F} ; f^{\uparrow F} .$$

The two sides will be equal if we prove that $f^{\uparrow \text{Opt}} ; \text{nonEmpty} = \text{nonEmpty}$ and that $f^{\uparrow \text{Opt}} ; \text{get} = \text{get} ; f$, which can be viewed as the two naturality laws specific to these functions. Use the definitions (9.10) of `get` and `nonEmpty`, set the type parameters as needed to match the types, and compute:

$$\begin{aligned} f^{\uparrow \text{Opt}} ; \text{nonEmpty} &= \begin{array}{c|c|c|c|c} & 1 & B & & 2 \\ \hline 1 & \text{id} & 0 & ; & 1 \\ \hline A & 0 & f & ; & \begin{array}{c|c} & \text{--} \rightarrow \text{false} \\ B & \text{--} \rightarrow \text{true} \end{array} \end{array} = \begin{array}{c|c|c|c|c} & & & 2 & \\ \hline 1 & & & \text{--} \rightarrow \text{false} & \\ \hline A & & & \text{--} \rightarrow \text{true} & \end{array} = \text{nonEmpty} . \\ f^{\uparrow \text{Opt}} ; \text{get} &= \begin{array}{c|c|c|c|c} & 1 & B & & B \\ \hline 1 & \text{id} & 0 & ; & 1 \\ \hline A & 0 & f & ; & \begin{array}{c|c} & 0 \\ B & \text{id} \end{array} \end{array} = \begin{array}{c|c|c|c|c} & & & B & \\ \hline 1 & & & 0 & \\ \hline A & & & f & \end{array} , \\ \text{get} ; f &= \begin{array}{c|c|c|c|c} & & A & & B \\ \hline & & 1 & 0 & \\ \hline & & A & \text{id} & \end{array} ; f^{\uparrow A \rightarrow B} = \begin{array}{c|c|c|c|c} & & B & & B \\ \hline & & 1 & 0 & \\ \hline & & A & \text{id} ; f & \end{array} = \begin{array}{c|c|c|c|c} & & & B & \\ \hline & & & 0 & \\ \hline & & & f & \end{array} = f^{\uparrow \text{Opt}} ; \text{get} . \end{aligned}$$

This concludes the proof.

The converse statement is also true:

Statement 9.2.2.5 If the naturality law (9.14) holds for `deflate` and the function `filter` is defined via `deflate` by Eq. (9.7) then the naturality law (9.3) holds for `filter`.

Proof Begin by writing the two sides of the naturality law (9.3):

$$\text{left-hand side of Eq. (9.3)} : f^{\uparrow F} ; \underline{\text{filt}(p)} = f^{\uparrow F} ; \underline{\psi_p^{\uparrow F}} ; \text{deflate} = (f ; \psi_p)^{\uparrow F} ; \text{deflate} .$$

$$\text{right-hand side of Eq. (9.3)} : \text{filt}(f ; p) ; f^{\uparrow F} = \psi_{f ; p}^{\uparrow F} ; \underline{\text{deflate}} ; f^{\uparrow F}$$

$$\text{naturality law (9.14) of } \text{deflate} : = \underline{\psi_{f ; p}^{\uparrow F}} ; f^{\uparrow \text{Opt} \uparrow F} ; \text{deflate} = (\psi_{f ; p} ; f^{\uparrow \text{Opt}})^{\uparrow F} ; \text{deflate} .$$

The remaining difference is in the order of composition of f and ψ . The proof will be complete if we show that, for any $f^{\uparrow A \rightarrow B}$ and $p^{\uparrow B \rightarrow 2}$,

$$f ; \psi_p = \psi_{f ; p} ; f^{\uparrow \text{Opt}} . \quad (9.18)$$

This equation can be viewed as a naturality law specific to the function ψ . To prove Eq. (9.18), it is convenient to use a definition of ψ_p that represents p as a function of type $A \rightarrow \mathbb{1} + \mathbb{1} \cong A \rightarrow \text{Opt}^1$,

$$x^{\uparrow A} ; \psi_p \triangleq x^{\uparrow A} ; p^{\uparrow A \rightarrow \text{Opt}^1} ; (1 \rightarrow x)^{\uparrow \text{Opt}} . \quad (9.19)$$

Using this definition of ψ_p , we can derive Eq. (9.18) by applying both sides to an x^A ,

$$\begin{aligned}
 \text{left-hand side : } & x \triangleright f \circ \psi_p = x \triangleright f \triangleright \psi_p \\
 \text{use Eq. (9.19) : } & = x \triangleright f \triangleright p \circ (1 \rightarrow x \triangleright f)^{\uparrow \text{Opt}} = x \triangleright f \circ p \circ (1 \rightarrow x \triangleright f)^{\uparrow \text{Opt}} \quad . \\
 \text{right-hand side : } & x \triangleright \psi_{f \circ p} \circ f^{\uparrow \text{Opt}} \\
 \text{use Eq. (9.19) : } & = x \triangleright f \circ p \circ (1 \rightarrow x)^{\uparrow \text{Opt}} \circ f^{\uparrow \text{Opt}} \\
 \text{composition law of Opt : } & = x \triangleright f \circ p \circ ((1 \rightarrow x) \circ f)^{\uparrow \text{Opt}} \\
 \text{compute composition : } & = x \triangleright f \circ p \circ (1 \rightarrow x \triangleright f)^{\uparrow \text{Opt}} \quad .
 \end{aligned}$$

9.2.3 Motivation and laws for `liftOpt`

In several derivations we just saw, the function `deflate` was composed with $\psi_p^{\uparrow F}$ or another lifted function that mapped types as $A \rightarrow \mathbb{1} + A$. This suggests considering a more general type signature, $f^{A \rightarrow \mathbb{1} + B}$, and composing $f^{\uparrow F}$ with `deflate` into a function that maps $F^A \rightarrow F^B$. It turns out that the resulting function, which we will call `liftOpt` and denote by liftOpt_F or simply by `liftOpt`,

```
def liftOpt_F[A, B](f: A => Option[B]): F[A] => F[B] = _ .map(f).deflate
```

$$\text{liftOpt}_F^{A,B}(f^{A \rightarrow \mathbb{1} + B}) \triangleq f^{\uparrow F} \circ \text{deflate}_F \quad , \quad (9.20)$$

has simpler laws and is particularly convenient for symbolic computations.

The name `liftOpt` ("lifting from `Option`") is motivated by the type signature,

$$\text{liftOpt}_F^{A,B} : (A \rightarrow \mathbb{1} + B) \rightarrow F^A \rightarrow F^B \quad ,$$

This lifts a function of type `A => Option[B]` into a function of type `F[A] => F[B]`. Except for using a "twisted" type $A \rightarrow \mathbb{1} + B$ instead of $A \rightarrow B$, this is similar to lifting via the `fmap` function,

$$\text{fmap}_F^{A,B} : (A \rightarrow B) \rightarrow F^A \rightarrow F^B \quad .$$

As this chapter will show, similarities between `liftOpt` and `fmap` go well beyond type signatures.

We will now derive some properties of `liftOpt`. The definition of `liftOpt` through `map` and `deflate` is illustrated in the diagram at left. Since f is arbitrary in Eq. (9.20), we may set $f = \text{id}^{A \rightarrow A}$ and the type parameter $A = \mathbb{1} + B$ to find

$$\begin{array}{ccc}
 F^A & \xrightarrow{(f^{A \rightarrow \mathbb{1} + B})^{\uparrow F}} & F^{\mathbb{1} + B} \\
 \text{liftOpt}(f) \triangleq & \searrow & \downarrow \text{deflate} \\
 & & F^B
 \end{array}$$

$$\text{liftOpt}^{\mathbb{1} + B, B}(\text{id}^{\mathbb{1} + B \rightarrow \mathbb{1} + B}) = \text{id}^{\uparrow F} \circ \text{deflate} = \text{deflate} \quad . \quad (9.21)$$

This expresses `deflate` through `liftOpt`. Are these two functions computationally equivalent?

Statement 9.2.3.1 The functions `liftOpt` and `deflate` are computationally equivalent as expressed by Eqs. (9.20)–(9.21), assuming that a suitable naturality law, Eq. (9.23) below, holds for `liftOpt`.

Proof We need to show that the equivalence of `liftOpt` and `deflate` holds in both directions:

(a) Given a `deflate` function, compute a `liftOpt` function via Eq. (9.20) and then a new `deflate'` function via Eq. (9.21). Then the new `deflate'` will be the same function as the initial `deflate`.

(b) Given a `liftOpt` function, compute a `deflate` function via Eq. (9.21) and then a new `liftOpt'` function via Eq. (9.21). The new `liftOpt'` will be the same as the initial `liftOpt`, assuming Eq. (9.23).

Proof for (a) directly derives the formula $\text{deflate}' = \text{deflate}$ by this calculation:

use Eq. (9.21) to define `deflate'` : $\text{deflate}' = \text{liftOpt}(\text{id})$

use Eq. (9.20) to define `liftOpt` : $= \text{id}^{\uparrow F} \circ \text{deflate} = \text{deflate}$.

Proof for (b) begins by expressing `liftOpt'` through the initial `liftOpt`:

$$\text{liftOpt}'(f) = f^{\uparrow F} \circ \text{deflate} = f^{\uparrow F} \circ \text{liftOpt}(\text{id}) \quad . \quad (9.22)$$

If nothing is known about `liftOpt`, the calculation will get stuck at this point. To proceed, we need to assume that `liftOpt` obeys a law that switches the order of function compositions around `liftOpt` and allows us to pull f inside of `liftOpt` in the equation above. Laws that switch the order of lifted function compositions are often naturality laws, as we have already seen. So, let us derive a suitable naturality law for `liftOpt`, beginning with a `liftOpt` function defined via `deflate`:

$$(h^{A \rightarrow B})^{\uparrow F} ; \text{liftOpt}^{B,C}(f^{B \rightarrow \mathbb{1} + C}) = h^{\uparrow F} ; f^{\uparrow F} ; \text{deflate} = (h ; f)^{\uparrow F} ; \text{deflate} = \text{liftOpt}^{A,C}(h ; f) .$$

$$\begin{array}{ccc} F^A & & \\ \downarrow h^{\uparrow F} & \searrow \text{liftOpt}(h ; f) & \\ F^B & \xrightarrow{\text{liftOpt}(f)} & F^C \end{array}$$

It follows that if `liftOpt` were defined via `deflate` then `liftOpt` would automatically satisfy the naturality law (see diagram at left)

$$(h^{A \rightarrow B})^{\uparrow F} ; \text{liftOpt}_F^{B,C}(f^{B \rightarrow \mathbb{1} + C}) = \text{liftOpt}_F^{A,C}(h^{A \rightarrow B} ; f^{B \rightarrow \mathbb{1} + C}) . \quad (9.23)$$

This motivates *imposing* that law on `liftOpt`. We can then finish the proof, resuming from Eq. (9.22),

$$\text{expect to equal } \text{liftOpt}(f) : \quad \text{liftOpt}'(f) = f^{\uparrow F} ; \text{liftOpt}(\text{id}) = \text{liftOpt}(f ; \text{id}) = \text{liftOpt}(f) .$$

Since `deflate` is computationally equivalent to `filter` (Statement 9.2.2.2), it follows that `filter` is computationally equivalent to `liftOpt`. The next step is to translate the laws of `filter` into the corresponding laws for `liftOpt`. Do the laws become simpler when formulated for `liftOpt`?

We have already seen that the partial function law of `filter` is satisfied automatically when `filter` is defined via `deflate` (Statement 9.2.2.3). It appears that `deflate` has only 3 laws while `filter` has 4. We will now show that `liftOpt` has only 2 laws, and yet they are *equivalent* to the 4 laws of `filter`.

To see that, we first express `filter` through `deflate` and then finally through `liftOpt`:

$$\text{filt}(p^{A \rightarrow \mathbb{2}}) = \psi_p^{\uparrow F} ; \text{deflate} = \text{liftOpt}(\psi_p) . \quad (9.24)$$

Conversely, `liftOpt` is expressed via `filter` like this,

$$\text{liftOpt}^{A,B}(f^{A \rightarrow \mathbb{1} + B}) \triangleq f^{\uparrow F} ; \text{deflate} = f^{\uparrow F} ; \text{filt}(\psi_{\text{nonEmpty}}) , \quad (9.25)$$

Identity law Let us now translate `filter`'s identity law (9.4) into the corresponding law for `liftOpt`. Begin by expressing `filt` ($_ \rightarrow \text{true}$) via `liftOpt` using Eq. (9.24):

$$\text{filt}(_ \rightarrow \text{true}) = \text{liftOpt}(\psi_{_ \rightarrow \text{true}}) = \text{liftOpt}(x^{A \rightarrow \mathbb{0} + x}) .$$

The function $\psi_{_ \rightarrow \text{true}}$ is equivalent to a simpler function $x^{A \rightarrow \mathbb{0} + x}$ (i.e., $x \Rightarrow \text{Some}(x)$ in Scala),

$$\text{use Eq. (9.19)} : \quad x^{A \rightarrow \psi_{_ \rightarrow \text{true}}} = x^{A \rightarrow (_^{A \rightarrow \text{true}^{\text{Opt}^1}}) ; (1 \rightarrow x)^{\uparrow \text{Opt}}}$$

$$\text{use equivalence } \mathbb{0} + 1 \cong \text{true}^{\text{Opt}^1} : \quad = x^{A \rightarrow (_^{A \rightarrow \mathbb{0} + 1}) ; (1 \rightarrow x)^{\uparrow \text{Opt}}}$$

$$\text{apply function to } x : \quad = (\mathbb{0} + 1) \triangleright (1 \rightarrow x)^{\uparrow \text{Opt}} = \begin{array}{c|cc} & \mathbb{0} & x \\ \hline \mathbb{0} & 1 & A \\ 1 & \text{id} & \mathbb{0} \\ \hline 1 & \mathbb{0} & 1 \rightarrow x \end{array}$$

$$\text{substitute the row into the matrix} : \quad = \begin{array}{c|cc} & \mathbb{0} & x \\ \hline \mathbb{0} & \mathbb{0} & x \end{array} = \mathbb{0} + x^{A \rightarrow \mathbb{0} + x} . \quad (9.26)$$

So, we expect the **identity law** of `liftOpt` to be

$$\text{liftOpt}^{A,A}(x^{A \rightarrow \mathbb{0} + x}) = \text{id}^{F^A \rightarrow F^A} . \quad (9.27)$$

Statement 9.2.3.2 (a) If `filter` obeys its naturality law (9.3) and identity law (9.4) and `liftOpt` is defined via Eq. (9.25) then `liftOpt` obeys its identity law (9.27).

(b) If `liftOpt` obeys its identity law (9.27) then `filter`, defined via Eq. (9.24), obeys its law (9.4).

Proof (a) Compute the identity law of `liftOpt`:

$$\begin{aligned}
 \text{expect to equal } \text{id}^A : \quad & \text{liftOpt}(x:A \rightarrow \mathbb{0} + x) = (x:A \rightarrow \mathbb{0} + x)^{\uparrow F} \circ \text{filt}^{\mathbb{1}+A}(\text{nonEmpty}) \circ \text{get}^{\uparrow F} \\
 \text{naturality law (9.3)} : \quad & = \text{filt}^A((x:A \rightarrow \mathbb{0} + x) \circ \text{nonEmpty}) \circ (x:A \rightarrow \mathbb{0} + x)^{\uparrow F} \circ \text{get}^{\uparrow F} \\
 \text{compute composition} : \quad & = \text{filt}^A(x:A \rightarrow \text{true}) \circ (x:A \rightarrow \mathbb{0} + x)^{\uparrow F} \circ \text{get}^{\uparrow F} \\
 \text{identity law (9.4)} : \quad & = \text{id}^A \circ (x:A \rightarrow \mathbb{0} + x)^{\uparrow F} \circ \text{get}^{\uparrow F} = ((x:A \rightarrow \mathbb{0} + x) \circ \text{get})^{\uparrow F} \\
 \text{compute composition} : \quad & = (\text{id}^A)^{\uparrow F} = \text{id}^A \quad .
 \end{aligned}$$

(b) Compute the identity law of `filter` using Eq. (9.26):

$$\begin{aligned}
 \text{use Eq. (9.24)} : \quad & \text{filt}(_ \rightarrow \text{true}) = \text{liftOpt}(\psi(_ \rightarrow \text{true})) \\
 \text{use Eq. (9.26)} : \quad & = \text{liftOpt}(x \rightarrow \mathbb{0} + x) \\
 \text{use Eq. (9.27)} : \quad & = \text{id} \quad .
 \end{aligned}$$

This completes the proof.

The function $x:A \rightarrow \mathbb{0} + x$ plays the role of the `pure` method for the `Option` type, if we view `Option[_]` as a pointed functor (see Section 8.3.5). Denote that `pure` method for brevity by `puOpt`:

$$\text{pu}_{\text{Opt}}^{\mathbb{1}+A} \triangleq x:A \rightarrow \mathbb{0} + x \quad .$$

Then `liftOpt`'s identity law (9.27) is written more concisely as

$$\text{liftOpt}_F(\text{pu}_{\text{Opt}}) = \text{id} \quad . \quad (9.28)$$

We can combine Eq. (9.28) and naturality law (9.23) into a single law if we compose the identity law with a lifted arbitrary function $f:A \rightarrow B$:

$$f^{\uparrow F} \circ \text{liftOpt}_F(\text{pu}_{\text{Opt}}) = \text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}}) \quad .$$

So the “combined” **naturality-identity law** of `liftOpt` is

$$\text{liftOpt}_F(f:A \rightarrow B \circ \text{pu}_{\text{Opt}}^B) = f^{\uparrow F} \quad . \quad (9.29)$$

Composition law Next, we translate Eq. (9.5) into a corresponding law of `liftOpt`. That law needs to combine two predicates p_1 and p_2 into a new predicate $p(x) \triangleq p_1(x) \wedge p_2(x)$. As Eq. (9.24) shows, `liftOpt` uses a predicate p only through the function ψ_p . So, we will derive the composition law of `liftOpt` if we somehow express ψ_p as a combination of ψ_{p_1} and ψ_{p_2} . Begin by writing

```
psi(p) == { x => Some(x).filter(p) } == { x => Some(x).filter(p1).filter(p2) }
```

$$\psi_p = x:A \rightarrow (\mathbb{0} + x) \triangleright \text{filt}_{\text{Opt}}(p) = x:A \rightarrow (\mathbb{0} + x) \triangleright \text{filt}_{\text{Opt}}(p1) \triangleright \text{filt}_{\text{Opt}}(p2) \quad .$$

We need to transform this code into some sort of combination of the functions ψ_{p_1} and ψ_{p_2} :

```
psi(p1) == { x => Some(x).filter(p1) }
psi(p2) == { x => Some(x).filter(p2) }
```

Since the value `Some(x).filter(p1)` is still of type `Option[A]`, we may apply `.map(psi(p2))` to that value:

```
x => Some(x).filter(p1).map(y => Some(y).filter(p2)) // Type is Option[Option[A]].
```

Except for the type `Option[Option[A]]`, the result is correct: a value $x:A$ will be present within the `Option[Option[A]]` wrapper only if both $p_1(x)$ and $p_2(x)$ return `true`. To convert the result to the required type `Option[A]`, we apply `Option`'s method `flatten`:

```

psi(p) == x => Some(x).filter(p1).map { y => Some(y).filter(p2) }.flatten // Use flatMap instead.
== x => Some(x).filter(p1).flatMap { y => Some(y).filter(p2) }
== psi(p1) andThen (_.flatMap(psi(p2))) // Using standard flatten and flatMap for Option.

```

Denote this combination of the functions ψ_{p_1} and ψ_{p_2} by the symbol \diamond_{Opt} , so that we may write

$$\psi_p = \psi_{p_1} \diamond_{\text{Opt}} \psi_{p_2} \triangleq x^A \rightarrow x \triangleright \psi_{p_1} \triangleright \text{flm}_{\text{Opt}}(\psi_{p_2}) = \psi_{p_1} \circ (y \rightarrow y \triangleright \text{flm}_{\text{Opt}}(\psi_{p_2})) \quad .$$

We use the symbol flm_{Opt} for `Option`'s `flatMap`; the Scala code `x.flatMap(f)` is denoted by $x \triangleright \text{flm}_{\text{Opt}}(f)$ if we view flm_{Opt} as a curried function with the type signature

$$\text{flm}_{\text{Opt}} : (A \rightarrow \text{Opt}^B) \rightarrow \text{Opt}^A \rightarrow \text{Opt}^B \quad .$$

The operation \diamond_{Opt} is a special kind of composition called **Kleisli¹ composition**. It applies to functions such as ψ_p that have type $A \rightarrow \mathbb{1} + A$; such functions cannot be composed with the ordinary function composition ($\psi_{p_1} \circ \psi_{p_2}$ does not type-check). It is straightforward to extend the operation \diamond_{Opt} from functions of type $A \rightarrow \mathbb{1} + A$ to functions with more general types, $A \rightarrow \mathbb{1} + B$ and $B \rightarrow \mathbb{1} + C$:

```

def kleisliOpt[A, B](f: A => Option[B], g: B => Option[C]): A => Option[C] =
  { x: A => f(x).flatMap(g) } // Using the standard flatMap for Option.

```

$$f^{A \rightarrow \mathbb{1} + B} \diamond_{\text{Opt}} g^{B \rightarrow \mathbb{1} + C} \triangleq x^A \rightarrow f(x) \triangleright \text{flm}_{\text{Opt}}(g) \quad .$$

The Kleisli composition $f \diamond_{\text{Opt}} g$ yields a function of type $A \rightarrow \mathbb{1} + C$ and is similar to the ordinary composition $f \circ g$ except for using “twisted” types, e.g., $A \rightarrow \text{Opt}^B$ instead of $A \rightarrow B$. (The “twisted” functions cannot be composed as $f \circ g$ because their types do not match.)

We can now derive the composition law of `liftOpt` starting from Eq. (9.5):

$$\begin{aligned} \text{left-hand side: } & \text{filt}(p_1) \circ \text{filt}(p_2) = \text{liftOpt}(\psi_{p_1}) \circ \text{liftOpt}(\psi_{p_2}) \quad . \\ \text{right-hand side: } & \text{filt}(p) = \text{liftOpt}(\psi_p) = \text{liftOpt}(\psi_{p_1} \diamond_{\text{Opt}} \psi_{p_2}) \quad . \end{aligned}$$

If `filter`'s composition law holds, we obtain (without any additional assumptions) the equation

$$\text{liftOpt}(\psi_{p_1}) \circ \text{liftOpt}(\psi_{p_2}) = \text{liftOpt}(\psi_{p_1} \diamond_{\text{Opt}} \psi_{p_2}) \quad . \quad (9.30)$$

This looks like a law typical for a “lifting”: the composition of lifted functions is equal to the lifted Kleisli composition. So, it appears useful to formulate the composition law of `liftOpt` in a more general way by allowing arbitrary $f^{A \rightarrow \mathbb{1} + B}$ and $g^{B \rightarrow \mathbb{1} + C}$ instead of specific functions ψ_{p_1} and ψ_{p_2} .

The composition law of `liftOpt` is then written as

$$\begin{array}{ccc} & F^B & \\ \text{liftOpt}(f) & \nearrow & \text{liftOpt}(g) \\ F^A & \xrightarrow{\text{liftOpt}(f \diamond_{\text{Opt}} g)} & F^C \end{array} \quad \text{liftOpt}_F(f^{A \rightarrow \mathbb{1} + B}) \circ \text{liftOpt}_F(g^{B \rightarrow \mathbb{1} + C}) = \text{liftOpt}_F(f \diamond_{\text{Opt}} g) \quad . \quad (9.31)$$

Because this law holds for f and g involving arbitrary types A, B, C , it is stronger than `filter`'s composition law. We will now show that `deflate`'s naturality law (9.14) and hence `filter`'s naturality law (9.3) can be derived from Eqs. (9.29)–(9.31) if we choose f and g in a special way, such that one of f or g always returns a non-empty `Option` value:

Statement 9.2.3.3 (a) The naturality law (9.23) of `liftOpt` follows from Eqs. (9.29)–(9.31).

(b) If `deflate` is defined via `liftOpt`, the naturality law (9.14) follows from Eqs. (9.29)–(9.31).

Proof (a) Choose functions $f^{A \rightarrow \mathbb{1} + B}$ to be of the form

$$f^{A \rightarrow \mathbb{1} + B} \triangleq h^{A \rightarrow B} \circ \text{pu}_{\text{Opt}}^{B \rightarrow \mathbb{1} + B}$$

or, written out in more detail : $= h^{A \rightarrow B} \circ (x^B \rightarrow \mathbb{0} + x) = x^A \rightarrow \mathbb{0} + h(x) \quad ,$

¹The Swiss-German name Kleisli is pronounced “cli-slee” (“cli” as in “climb”).

where $h:A \rightarrow B$ is an arbitrary function, and use those f in Eq. (9.31):

$$\text{liftOpt}(f \diamond_{\text{Opt}} g) = \text{liftOpt}(h \circ \text{pu}_{\text{Opt}}) \circ \text{liftOpt}(g)$$

use Eq. (9.29) : $= h^{\uparrow F} \circ \text{liftOpt}(g)$.

To simplify this formula, we need to compute $f \diamond_{\text{Opt}} g$. The Kleisli composition \diamond_{Opt} is defined via the standard `flatMap` method of the `Option` type. With the notation flm_{Opt} for the curried `flatMap` function, the definition of \diamond_{Opt} is simplified to

$$f:A \rightarrow \mathbb{1} + B \quad \diamond_{\text{Opt}} \quad g:B \rightarrow \mathbb{1} + C \triangleq f \circ \text{flm}_{\text{Opt}}(g) \quad . \quad (9.32)$$

Then we compute $f \diamond_{\text{Opt}} g$ using this definition of \diamond_{Opt} as

$$\text{definition of } f : \quad f \diamond_{\text{Opt}} g = (h \circ \text{pu}_{\text{Opt}}) \diamond_{\text{Opt}} g = h \circ \text{pu}_{\text{Opt}} \circ \text{flm}_{\text{Opt}}(g)$$

compute composition (see below) : $= h \circ g$.

The result of composing `pure` and `flatMap` for `Option` is not obvious, but it turns out that pu_{Opt} followed by $\text{flm}_{\text{Opt}}(g)$ is equal to just g . To verify that, let us first use the syntax of Scala:

```
pure(x) == Some(x)           // By definition of 'pure' for 'Option'.
p.flatMap(g) == p match {      // By definition of 'flatMap' for 'Option'.
  case None      => None
  case Some(x)   => g(x)
}
pure(x).flatMap(g) == Some(x).flatMap(g) == g(x)
```

The same symbolic computation is written in the code notation like this:

$$\text{pu}_{\text{Opt}} = \begin{array}{|c|c|} \hline & \mathbb{1} & A \\ \hline A & \mathbb{0} & \text{id} \\ \hline \end{array} \quad , \quad \text{flm}_{\text{Opt}}(g:A \rightarrow \mathbb{1} + B) = \begin{array}{|c|c|} \hline & \mathbb{1} + B \\ \hline \mathbb{1} & \mathbb{1} \rightarrow \mathbb{1} + \mathbb{0}^B \\ \hline A & g \\ \hline \end{array} \quad , \quad (9.33)$$

$$\text{pu}_{\text{Opt}} \circ \text{flm}_{\text{Opt}}(g) = \begin{array}{|c|c|} \hline & \mathbb{1} & A \\ \hline A & \mathbb{0} & \text{id} \\ \hline \end{array} \circ \begin{array}{|c|c|} \hline & \mathbb{1} + B \\ \hline \mathbb{1} & \mathbb{1} \rightarrow \mathbb{1} + \mathbb{0}^B \\ \hline A & g \\ \hline \end{array}$$

$$\text{matrix composition} : \quad = \begin{array}{|c|c|} \hline & \mathbb{1} + B \\ \hline A & \text{id} \circ g \\ \hline \end{array} = g \quad . \quad (9.34)$$

Since we have now shown that $f \diamond_{\text{Opt}} g = h \circ g$, the naturality law (9.23) follows:

$$\text{liftOpt}(f \diamond_{\text{Opt}} g) = \text{liftOpt}(h \circ g) = h^{\uparrow F} \circ \text{liftOpt}(g) \quad .$$

This derivation is possible because the function f always returns a non-empty `Option`, which corresponds to filtering with a predicate that always returns `true`. This reduces filtering to an identity function, which simplifies the composition law (9.31) by eliminating one of the `liftOpt` functions.

(b) We keep $f:A \rightarrow \mathbb{1} + B$ arbitrary but choose $g:B \rightarrow \mathbb{1} + C$ to be of the form

$$g:B \rightarrow \mathbb{1} + C \triangleq h:B \rightarrow C \circ \text{pu}_{\text{Opt}}^C \quad ,$$

where $h:B \rightarrow C$ is an arbitrary function, and substitute into `liftOpt`'s composition law (9.31):

$$\text{liftOpt}(f \diamond_{\text{Opt}} g) = \text{liftOpt}(f) \circ \text{liftOpt}(h \circ \text{pu}_{\text{Opt}})$$

use Eq. (9.29) : $= \text{liftOpt}(f) \circ h^{\uparrow F} \quad .$

To proceed, we need to compute the Kleisli composition $f \diamond_{\text{Opt}} g$ for the chosen form of g ,

$$f \diamond_{\text{Opt}} g = f \diamond_{\text{Opt}} (h \circ \text{pu}_{\text{Opt}})$$

$$\text{definition (9.32) of } \diamond_{\text{Opt}} : \quad = f \circ \text{flm}_{\text{Opt}}(h \circ \text{pu}_{\text{Opt}}) = f \circ \text{flm}_{\text{Opt}}(y \rightarrow 0 + h(y))$$

$$\text{definition (9.33) of } \text{flm}_{\text{Opt}} : \quad = f \circ \begin{array}{c|c|c|c} & & & 1 + C \\ \hline 1 & & 1 \rightarrow 1 + 0 : C \\ \hline B & & y \rightarrow 0 + h(y) \\ \hline & & 1 & C \\ \hline 1 & id & 0 \\ \hline B & 0 & h \end{array}$$

split matrix into 2 columns : $= f \circ \begin{array}{c|c|c} & & 1 + C \\ \hline 1 & id & 0 \\ \hline B & 0 & h \end{array}$

$$\text{definition of lifting, } \text{h}^{\text{Opt}} : \quad = f \circ h^{\text{Opt}} \quad .$$

This yields the **right naturality** law of liftOpt ,

$$\begin{array}{ccc} F^A & & \\ \text{liftOpt}(f) \downarrow & \searrow \text{liftOpt}(f \circ h^{\text{Opt}}) & \\ F^B & \xrightarrow{h^F} & F^C \end{array}$$

$$\text{liftOpt}_F(f^{A \rightarrow 1+B}) \circ h^F = \text{liftOpt}_F(f \circ h^{\text{Opt}}) \quad . \quad (9.35)$$

Assuming that deflate is defined via liftOpt by Eq. (9.21), we express the two sides of deflate 's naturality law (9.14) through liftOpt :

$$\text{left-hand side of Eq. (9.14)} : \quad \text{deflate} \circ f^{\text{Opt}} = \text{liftOpt}(\text{id}) \circ f^{\text{Opt}}$$

$$\text{use Eq. (9.35)} : \quad = \text{liftOpt}(f^{\text{Opt}}) \quad .$$

$$\text{right-hand side of Eq. (9.14)} : \quad f^{\text{Opt}} \circ \text{deflate} = f^{\text{Opt}} \circ \text{liftOpt}(\text{id})$$

$$\text{naturality law (9.23)} : \quad = \text{liftOpt}(f^{\text{Opt}} \circ \text{id}) = \text{liftOpt}(f^{\text{Opt}}) \quad .$$

We are justified to use the naturality law (9.23) because we proved in part (a) that it holds.

If filter obeys the naturality law then so does deflate , and vice versa (Statements 9.2.2.4–9.2.2.5).

To summarize, we have proved that if filter is defined via liftOpt and the two laws (9.29), (9.31) hold for liftOpt then all four laws of filter will hold.

The converse statement also holds: if we define liftOpt through filter by Eq. (9.25), the four laws of filter will imply the two laws of liftOpt .

Exercise 9.2.3.4 Show that any function $f^{A \rightarrow 1+B}$ can be expressed as $f = \psi_p \circ h^{\text{Opt}}$ with a suitable choice of a predicate $p^{A \rightarrow 2}$ and a partial function $h^{A \rightarrow B}$. Derive explicit formulas for h and ψ in terms of f and implement them in Scala.

Exercise 9.2.3.5* Show that the law (9.31) holds if liftOpt is defined via filter using Eq. (9.25).

Hint: first derive the two naturality laws for liftOpt ; then use Exercise 9.2.3.4 to extend the restricted composition law (9.30) to the full law (9.31).

9.2.4 Constructions of filterable functors

How can we recognize a filterable functor F^A by its type expression, without having to prove laws? One intuition is that the type F^A must be able to accommodate replacing values of A by unit values; this replacement is performed by the function deflate . To make this intuition more precise, it helps to perform structural analysis that systematically looks for type constructions creating new filterable functors out of existing ones while preserving the laws.

To begin, we note that `Option[_]`, `Either[L, _]`, `Try[_]`, `Seq[_]`, and `Map[K, _]` are filterable. Let us now go through all constructions available for exponential-polynomial types. To check whether a functor is filterable, it is convenient to use the `liftOpt` function and its two laws (9.29)–(9.31).

Type parameters There are three constructions that work solely by manipulating type parameters: the identity functor $\text{Id}^A \triangleq A$, the constant functor $\text{Const}^{Z,A} \triangleq Z$ (where Z is a fixed type), and the functor composition, $F^A \triangleq G^{H^A}$.

The identity functor is *not* filterable because `deflate` of type $\mathbb{1} + A \rightarrow A$ cannot be implemented.

The constant functor $\text{Const}^{Z,A} \triangleq Z$ can be viewed as a “wrapper” that never wraps any values of type A . This functor is filterable because we can define $\text{liftOpt}(_) \triangleq \text{id}^{Z \rightarrow Z}$ (filtering is a no-op for a “wrapper” that is always empty). All laws usually hold for an identity function. To verify the laws, note that the lifting to the `Const` functor is also an identity function: $f^{\uparrow \text{Const}} = \text{id}^{Z \rightarrow Z}$ for any $f^{A \rightarrow B}$.

check law (9.29) : $\text{liftOpt}_{\text{Const}}(f \circ \text{pu}_{\text{Opt}}) = \text{id} = f^{\uparrow \text{Const}}$,

check law (9.31) : $\text{liftOpt}_{\text{Const}}(f) \circ \text{liftOpt}_{\text{Const}}(g) = \text{id} \circ \text{id} = \text{id} = \text{liftOpt}_{\text{Const}}(f \diamond_{\text{Opt}} g)$.

The functor composition $F^A \triangleq G^{H^A}$ requires only H to be a filterable functor:

Statement 9.2.4.1 The functor $F^A \triangleq G^{H^A}$ is filterable when H^A is filterable and G^A is any functor.

Proof Assuming that liftOpt_H is available and lawful, we define liftOpt_F as

```
def liftOpt_F[A, B](f: A => Option[B]): G[H[A]] => G[H[B]] = liftOpt_F(f) ≡ (liftOpt_H(f))↑G .  
{ g: G[H[A]] => g.map(liftOpt_H(f)) }
```

To verify the identity-naturality law of `liftOpt_F`, note that $f^{\uparrow F} = f^{\uparrow H \uparrow G}$ by definition of F :

expect to equal $f^{\uparrow F}$: $\text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}}) = (\text{liftOpt}_H(f \circ \text{pu}_{\text{Opt}}))↑G$
law (9.29) of liftOpt_H : $= (f^{\uparrow H})↑G = f^{\uparrow F}$.

To verify the composition law of liftOpt_F , compute:

expect to equal $\text{liftOpt}_F(f \diamond_{\text{Opt}} f')$: $\text{liftOpt}_F(f) \circ \text{liftOpt}_F(f') = (\text{liftOpt}_H(f) \circ \text{liftOpt}_H(f'))↑G$
composition law of liftOpt_H : $= (\text{liftOpt}_H(f \diamond_{\text{Opt}} f'))↑G = \text{liftOpt}_F(f \diamond_{\text{Opt}} f')$.

Products To show that the product of two filterable functors is filterable, we will use a definition of $\text{liftOpt}_{G^\bullet \times H^\bullet}$ and a proof quite similar to what we did for the product of functors (Statement 6.2.3.3).

Statement 9.2.4.2 The functor $F^A \triangleq G^A \times H^A$ is filterable if G^\bullet and H^\bullet are filterable functors.

Proof To define liftOpt_F , we use the pair product operation \boxtimes similarly to Eq. (6.13),

$\text{liftOpt}_F(p) \triangleq \text{liftOpt}_G(p) \boxtimes \text{liftOpt}_H(p)$.

The lifting to F is defined by Eq. (6.13) as $f^{\uparrow F} = f^{\uparrow G} \boxtimes f^{\uparrow H}$. To verify the naturality-identity law:

expect to equal $f^{\uparrow F}$: $\text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}}) = \text{liftOpt}_G(f \circ \text{pu}_{\text{Opt}}) \boxtimes \text{liftOpt}_H(f \circ \text{pu}_{\text{Opt}})$
law (9.29) for G and H : $= f^{\uparrow G} \boxtimes f^{\uparrow H} = f^{\uparrow F}$.

To verify the composition law:

expect to equal $\text{liftOpt}_F(f \diamond_{\text{Opt}} g)$: $\text{liftOpt}_F(f) \circ \text{liftOpt}_F(g)$
definition of liftOpt_F : $= (\text{liftOpt}_G(f) \boxtimes \text{liftOpt}_H(f)) \circ (\text{liftOpt}_G(g) \boxtimes \text{liftOpt}_H(g))$
composition property (9.36) : $= (\text{liftOpt}_G(f) \circ \text{liftOpt}_G(g)) \boxtimes (\text{liftOpt}_H(f) \circ \text{liftOpt}_H(g))$
composition laws of G and H : $= \text{liftOpt}_G(f \diamond_{\text{Opt}} g) \boxtimes \text{liftOpt}_H(f \diamond_{\text{Opt}} g)$
definition of liftOpt_F : $= \text{liftOpt}_F(f \diamond_{\text{Opt}} g)$.

In this calculation, we used the composition property

$$(f \boxtimes g) \circ (p \boxtimes q) = (f \circ p) \boxtimes (g \circ q) \quad , \quad (9.36)$$

which follows from the definition of the pair product operation \boxtimes ,

$$(f \boxtimes g) \boxtimes (p \boxtimes q) = (a \times b \rightarrow f(a) \times g(b)) \boxtimes (c \times d \rightarrow p(c) \times q(d))$$

$$\begin{aligned} \text{compute composition : } &= (a \times b \rightarrow p(f(a)) \times q(g(b))) \\ \text{definition of } \boxtimes : &= (a \rightarrow p(f(a))) \boxtimes (b \rightarrow q(g(b))) \\ \text{definition of the } \boxtimes \text{ operation : } &= (f \boxtimes p) \boxtimes (g \boxtimes q) \quad . \end{aligned}$$

Co-products There are two constructions that produce new filterable functors involving disjunctive types (co-product types). The first construction is the filterable co-product $F^A \triangleq G^A + H^A$, where G^A and H^A are filterable functors. This is similar to the functor co-product (Statement 6.2.3.4). The second construction is $F^A \triangleq \mathbb{1} + A \times G^A$ where G^A is a filterable functor. This cannot be reduced to the first construction because $A \times G^A$ is not filterable.

Statement 9.2.4.3 The functor $F^A \triangleq G^A + H^A$ is filterable if G and H are filterable functors.

Proof Assuming that liftOpt_G and liftOpt_H are available, we define liftOpt_F as

```
def liftOpt_F[A, B](f: A => Option[B]): Either[G[A], H[A]] => Either[G[B], H[B]] = {
  case Left(ga) => liftOpt_G(f)(ga)
  case Right(ha) => liftOpt_H(f)(ha)
}
```

$$\text{liftOpt}_F(f: A \rightarrow \mathbb{1} + B) \triangleq \begin{vmatrix} & G^A & H^A \\ G^A & \text{liftOpt}_G(f) & \mathbb{0} \\ H^A & \mathbb{0} & \text{liftOpt}_H(f) \end{vmatrix} .$$

Lifting to the functor F^A is defined as in Statement 6.2.3.4,

$$f^{\uparrow F} \triangleq \begin{vmatrix} & G^A & H^A \\ G^A & f^{\uparrow G} & \mathbb{0} \\ H^A & \mathbb{0} & f^{\uparrow H} \end{vmatrix} .$$

In the matrix calculations, we will always have the rows and columns of type $G^A + H^A$, so we will omit the type annotations for brevity. To verify the naturality-identity law (9.29):

$$\begin{aligned} \text{expect to equal } f^{\uparrow F} : \quad & \text{liftOpt}_F(f \boxtimes \text{pu}_{\text{Opt}}) = \begin{vmatrix} \text{liftOpt}_G(f \boxtimes \text{pu}_{\text{Opt}}) & \mathbb{0} \\ \mathbb{0} & \text{liftOpt}_H(f \boxtimes \text{pu}_{\text{Opt}}) \end{vmatrix} \\ \text{law (9.29) for } G \text{ and } H : \quad & = \begin{vmatrix} f^{\uparrow G} & \mathbb{0} \\ \mathbb{0} & f^{\uparrow H} \end{vmatrix} = f^{\uparrow F} \quad . \end{aligned}$$

To verify the composition law (9.31):

$$\begin{aligned} \text{expect to equal } \text{liftOpt}_F(f \diamond_{\text{Opt}} g) : \quad & \text{liftOpt}_F(f) \boxtimes \text{liftOpt}_F(g) \\ \text{definition of } \text{liftOpt}_F : \quad & = \begin{vmatrix} \text{liftOpt}_G(f) & \mathbb{0} \\ \mathbb{0} & \text{liftOpt}_H(f) \end{vmatrix} \boxtimes \begin{vmatrix} \text{liftOpt}_G(g) & \mathbb{0} \\ \mathbb{0} & \text{liftOpt}_H(g) \end{vmatrix} \\ \text{matrix composition : } & = \begin{vmatrix} \text{liftOpt}_G(f) \boxtimes \text{liftOpt}_G(g) & \mathbb{0} \\ \mathbb{0} & \text{liftOpt}_H(f) \boxtimes \text{liftOpt}_H(g) \end{vmatrix} \\ \text{law (9.31) for } G \text{ and } H : & = \begin{vmatrix} \text{liftOpt}_G(f \diamond_{\text{Opt}} g) & \mathbb{0} \\ \mathbb{0} & \text{liftOpt}_H(f \diamond_{\text{Opt}} g) \end{vmatrix} \\ \text{definition of } \text{liftOpt}_F : & = \text{liftOpt}_F(f \diamond_{\text{Opt}} g) \quad . \end{aligned}$$

Statement 9.2.4.4 The functor $F^A \triangleq 1 + A \times G^A$ is filterable if G is a filterable functor.

Proof Assuming that liftOpt_G is available, we define liftOpt_F as

```

def liftOpt_F[A, B](f: A => Option[B]): Option[(A, G[A])] => Option[(B, G[B])] = {
  case None          => None                                // An empty wrapper remains empty.
  case Some((a, ga)) => f(a) match {                         // Does 'a' pass the filtering predicate?
    case None         => None                                // No. Drop all data, return an empty wrapper.
    case Some(b)      => Some((b, liftOpt_G(f)(ga)))        // Yes. Keep 'b' and filter 'ga' using 'liftOpt_G'.
  }
}

```

		$1 + B \times G^B$	
		$1 \rightarrow 1 + \mathbb{0}^{B \times G^B}$	
$\text{liftOpt}_F(f) \triangleq$	1	1	$B \times G$
	$A \times G^A$	$a \times g \rightarrow a \triangleright f \triangleright$	1
			$\mathbb{0}$
		B	$b \rightarrow b \times \text{liftOpt}_G(f)(g)$

The matrix for liftOpt_F has a non-split output column (representing the disjunctive type $\mathbb{1} + B \times G^B$). This is because the code must pattern-match on $f(a)$ in order to determine whether the result is of type $\mathbb{1}$ or of type $B \times G^B$. It is inconvenient to use such matrices in calculations since the conventions of matrix products do not work unless all parts of a disjunctive type are represented by separate columns. To be able to do symbolic calculations, we need to rewrite the code in a different way.

Note that the code pattern when destructuring an `Option` value looks like this function,

```
def p[A, B, C](q: A => Option[B], r: B => C): Option[A] => Option[C] = {
  case None    => None
  case Some(a) => q(a) match {    // Destructure the result of applying the function 'q'.
    case None    => None
    case Some(b) => Some(r(b)) // Apply a final transformation 'r'.
  }
}
```

The code of `p` can be rewritten using the standard `flatMap` method of the `Option` type:

```
def p[A, B, C](q: A => Option[B], r: B => C): Option[A] => Option[C] = _.flatMap { a => q(a).map(r) }
```

So we may write the code of `liftOpt_F` as

```
def liftOpt_F[A, B](f: A => Option[B]): Option[(A, G[A])] => Option[(B, G[B])] = _.flatMap {
  case (a, ga)  => f(a).map { b => (b, liftOpt_G(f)(ga)) }
}
```

$$\text{liftOpt}_E(f) \triangleq \text{flm}_{\text{Opt}}(a^{\cdot A} \times g^{\cdot G^A} \rightarrow a \triangleright f \circ (b \rightarrow b \times \text{liftOpt}_G(f)(g))^{\text{Opt}}) \quad . \quad (9.37)$$

To verify the laws, we first need to define the lifting to the functor F :

```
def fmap_F[A, B](f: A => B): Option[(A, G[A])] => Option[(B, G[B])] = {
  case None          => None
  case Some((a, ga)) => Some((f(a), ga.map(f)))
}
```

Again, it is convenient to rewrite the code using the standard `map` method of the `Option` type:

```
def fmap_F[A, B](f: A => B): Option[(A, G[A])] => Option[(B, G[B])] = _ .map {  
  case (a, ga) => (f(a), ga.map(f))  
}
```

$$f^{\uparrow F} \equiv (a^{:A} \times g^{:G^A} \rightarrow f(a) \times (g \triangleright f^{\uparrow G}))^{\uparrow \text{Opt}} \equiv (f \boxtimes f^{\uparrow G})^{\uparrow \text{Opt}}$$

For brevity, we omit type annotations. The naturality-identity law (9.29) for F is verified by

$$\begin{aligned}
 \text{expect to equal } f^{\uparrow F} : & \text{ liftOpt}_F(f \circ \text{pu}_{\text{Opt}}) \\
 \text{definition of liftOpt}_F : & = \text{flm}_{\text{Opt}}(a \times g \rightarrow a \triangleright f \circ \text{pu}_{\text{Opt}} \circ (b \rightarrow b \times \text{liftOpt}_G(f \circ \text{pu}_{\text{Opt}})(g))^{\uparrow \text{Opt}}) \\
 \text{law (9.29) for } G : & = \text{flm}_{\text{Opt}}(a \times g \rightarrow a \triangleright f \circ \text{pu}_{\text{Opt}} \circ (b \rightarrow b \times f^{\uparrow G}(g))^{\uparrow \text{Opt}}) \\
 \text{naturality (8.8) of pu}_{\text{Opt}} : & = \text{flm}_{\text{Opt}}(a \times g \rightarrow a \triangleright f \circ (b \rightarrow b \times f^{\uparrow G}(g)) \circ \text{pu}_{\text{Opt}}) \\
 \text{function composition} : & = \text{flm}_{\text{Opt}}(a \times g \rightarrow (f(a) \times f^{\uparrow G}(g)) \triangleright \text{pu}_{\text{Opt}}) = \text{flm}_{\text{Opt}}((f \boxtimes f^{\uparrow G}) \circ \text{pu}_{\text{Opt}}) \\
 \text{use Eq. (9.38)} : & = (f \boxtimes f^{\uparrow G})^{\uparrow \text{Opt}} = f^{\uparrow F} .
 \end{aligned}$$

Here we used a property that applies to a composition of any $q : A \rightarrow B$ with pu_{Opt} under `flatMap`:

$$\text{flm}_{\text{Opt}}(q : A \rightarrow B \circ \text{pu}_{\text{Opt}} : B \rightarrow \mathbb{1} + B) = q^{\uparrow \text{Opt}} . \quad (9.38)$$

This property is derived by applying the code for `Option`'s `flatMap` and `map` to suitable values:

<code>None.flatMap(x => Some(q(x))) == None</code>	<code>None.map(q) == None</code> // <code>pure_Opt(x) == Some(x)</code>
<code>Some(x).flatMap(x => Some(q(x))) == Some(q(x))</code>	<code>Some(x).map(q) == Some(q(x))</code>

Additional work is necessary to check `liftOpt`'s composition law,

$$\text{liftOpt}_F(f) \circ \text{liftOpt}_F(f') = \text{liftOpt}_F(f \circ_{\text{Opt}} f') . \quad (9.39)$$

Since $\text{liftOpt}_F(f)$ is equal to $\text{flm}_{\text{Opt}}(\dots)$, so we need somehow to transform $\text{flm}_{\text{Opt}}(\dots) \circ \text{flm}_{\text{Opt}}(\dots)$ into an expression of the form $\text{flm}_{\text{Opt}}(\dots)$. To obtain such a transformation, we use a trick: consider the composition law for the $\text{liftOpt}_{\text{Opt}}$ operation of the `Option` functor (not the liftOpt_F of the functor F). Since we already know that `Option` is filterable, the composition law must hold for $\text{liftOpt}_{\text{Opt}}$:

$$\text{liftOpt}_{\text{Opt}}(f : A \rightarrow \text{Opt}^B) \circ \text{liftOpt}_{\text{Opt}}(g : B \rightarrow \text{Opt}^C) = \text{liftOpt}_{\text{Opt}}(f \circ_{\text{Opt}} g) .$$

To express $\text{liftOpt}_{\text{Opt}}$ through known functions, recall that `Option`'s `deflate` function, with the type signature $\text{Opt}^{\text{Opt}^A} \rightarrow \text{Opt}^A$, is equal to the standard method `flatten` (denoted by ftn_{Opt} for brevity):

$$\text{liftOpt}_{\text{Opt}}(f) = f^{\uparrow \text{Opt}} \circ \text{deflate}_{\text{Opt}} = f^{\uparrow \text{Opt}} \circ \text{ftn}_{\text{Opt}} = \text{flm}_{\text{Opt}}(f) .$$

So, `Option`'s `flatMap` method equals `liftOpt` and obeys a law similar to `liftOpt`'s composition law,

$$\text{flm}_{\text{Opt}}(f) \circ \text{flm}_{\text{Opt}}(f') = \text{flm}_{\text{Opt}}(f \circ_{\text{Opt}} f') = \text{flm}_{\text{Opt}}(f \circ \text{flm}_{\text{Opt}}(f')) . \quad (9.40)$$

We call Eq. (9.40) the **associativity law** of `flatMap` for reasons explained in Statement 9.4.2.1 below.

To make the calculation quicker, denote by $r_{f,g}$ a sub-expression in Eq. (9.37), so that we can write

$$r_{f,g} \triangleq b \rightarrow b \times \text{liftOpt}_G(f)(g) , \quad \text{liftOpt}_F(f) = \text{flm}_{\text{Opt}}(a \times g \rightarrow a \triangleright f \circ r_{f,g}^{\uparrow \text{Opt}}) . \quad (9.41)$$

We can now start with the left-hand side of Eq. (9.39):

$$\begin{aligned}
 & \text{liftOpt}_F(f) \circ \text{liftOpt}_F(f') \\
 \text{use Eq. (9.41)} : & = \text{flm}_{\text{Opt}}(a \times g \rightarrow a \triangleright f \circ r_{f,g}^{\uparrow \text{Opt}}) \circ \text{flm}_{\text{Opt}}(a' \times g' \rightarrow a' \triangleright f' \circ r_{f',g'}^{\uparrow \text{Opt}}) \\
 \text{use Eq. (9.40)} : & = \text{flm}_{\text{Opt}}((a \times g \rightarrow a \triangleright f \circ r_{f,g}^{\uparrow \text{Opt}}) \circ \text{flm}_{\text{Opt}}(a' \times g' \rightarrow a' \triangleright f' \circ r_{f',g'}^{\uparrow \text{Opt}})) \\
 \text{»-notation} : & = \text{flm}_{\text{Opt}}(a \times g \rightarrow a \triangleright f \circ r_{f,g}^{\uparrow \text{Opt}} \circ \text{flm}_{\text{Opt}}(a' \times g' \rightarrow a' \triangleright f' \circ r_{f',g'}^{\uparrow \text{Opt}})) .
 \end{aligned} \quad (9.42)$$

Here we used a convenient “parenthesis-canceling” property of the \triangleright -notation:

$$(x \rightarrow x \triangleright f) ; g = (x \rightarrow \underline{g(f(x))}) = (x \rightarrow x \triangleright \underline{f \triangleright g}) = (x \rightarrow x \triangleright f ; g) \quad .$$

It is not immediately clear how to proceed, so let us transform the right-hand side of Eq. (9.39):

$$\begin{aligned} & \text{liftOpt}_F(f \diamond_{\text{Opt}} f') \\ \text{definition (9.32) of } \diamond_{\text{Opt}} : &= \text{liftOpt}_F(f ; \text{flmOpt}(f')) \\ \text{use Eq. (9.41) : } &= \text{flmOpt}(a \times g \rightarrow a \triangleright f ; \text{flmOpt}(f') ; r_{f ; \text{flmOpt}(f'), g}^{\text{Opt}}) \quad . \end{aligned} \quad (9.43)$$

How can we show that the last expressions in Eqs. (9.42)–(9.43) are equal? To find a way forward, let us compare these two equations and find the sub-expressions that remain different:

$$\text{sub-expression from the left-hand side : } r_{f, g}^{\text{Opt}} ; \text{flmOpt}(a' \times g' \rightarrow a' \triangleright f' ; r_{f', g'}^{\text{Opt}}) \quad (9.44)$$

$$\text{we would like that to become equal to : } \text{flmOpt}(f') ; r_{f ; \text{flmOpt}(f'), g}^{\text{Opt}} \quad . \quad (9.45)$$

The difference is in the lifted functions to the left and to the right of flmOpt . If we somehow bring those functions inside $\text{flmOpt}(...)$, we may be able to simplify them further. So, we look for properties of Option ’s flatMap that should have the form

$$(p : A \rightarrow B)^{\text{Opt}} ; \text{flmOpt}(q : B \rightarrow \text{Opt}^C) = \text{flmOpt}(\text{???} : A \rightarrow \text{Opt}^C) \quad ,$$

$$\text{flmOpt}(p : A \rightarrow \text{Opt}^B) ; (q : B \rightarrow C)^{\text{Opt}} = \text{flmOpt}(\text{???} : A \rightarrow \text{Opt}^C) \quad .$$

The typed holes must be filled using the only available data (the functions p and q):

$$(p : A \rightarrow B)^{\text{Opt}} ; \text{flmOpt}(q : B \rightarrow \text{Opt}^C) = \text{flmOpt}(p ; q) \quad , \quad (9.46)$$

$$\text{flmOpt}(p : A \rightarrow \text{Opt}^B) ; (q : B \rightarrow C)^{\text{Opt}} = \text{flmOpt}(p ; q^{\text{Opt}}) \quad . \quad (9.47)$$

We omit the proofs for these **naturality laws** of flatMap . With them, we transform Eqs. (9.44)–(9.45):

$$\begin{aligned} & r_{f, g}^{\text{Opt}} ; \text{flmOpt}(a' \times g' \rightarrow a' \triangleright f' ; r_{f', g'}^{\text{Opt}}) \\ \text{use Eq. (9.46) : } &= \text{flmOpt}(r_{f, g} ; (a' \times g' \rightarrow a' \triangleright f' ; r_{f', g'}^{\text{Opt}})) \\ \text{expand } r_{f, g} : &= \text{flmOpt}((a \rightarrow a \times \text{liftOpt}_G(f)(g)) ; (a' \times g' \rightarrow a' \triangleright f' ; r_{f', g'}^{\text{Opt}})) \\ \text{compute composition : } &= \text{flmOpt}(a \rightarrow a \triangleright f' ; r_{f', \text{liftOpt}_G(f)(g)}^{\text{Opt}}) \quad , \end{aligned}$$

and

$$\begin{aligned} & \text{flmOpt}(f') ; r_{f ; \text{flmOpt}(f'), g}^{\text{Opt}} = \text{flmOpt}(f' ; r_{f ; \text{flmOpt}(f'), g}^{\text{Opt}}) \\ \text{expand function } f' : &= \text{flmOpt}(a \rightarrow a \triangleright f' ; r_{f ; \text{flmOpt}(f'), g}^{\text{Opt}}) \quad . \end{aligned}$$

The difference between sub-expressions has become smaller: it just remains to show that

$$r_{f', \text{liftOpt}_G(f)(g)}^{\text{Opt}} \stackrel{?}{=} r_{f ; \text{flmOpt}(f'), g}^{\text{Opt}} \quad .$$

Expand the definition of $r_{f, g}$ in both sides:

$$b \rightarrow b \times \text{liftOpt}_G(f')(\text{liftOpt}_G(f)(g)) \stackrel{?}{=} b \rightarrow b \times \text{liftOpt}_G(f ; \text{flmOpt}(f'))(g) \quad .$$

Omitting the common sub-expressions, we find the remaining difference:

$$\text{liftOpt}_G(f')(\text{liftOpt}_G(f)(g)) \stackrel{?}{=} \text{liftOpt}_G(f \circ \text{flmOpt}(f'))(g) .$$

This is equivalent to liftOpt_G 's composition law applied to the function g ,

$$g \triangleright \text{liftOpt}_G(f) \circ \text{liftOpt}_G(f') = g \triangleright \text{liftOpt}_G(f \diamond_{\text{opt}} f') = g \triangleright \text{liftOpt}_G(f \circ \text{flmOpt}(f')) .$$

Since the composition law of liftOpt_G is assumed to hold, we have finished the proof of Eq. (9.39).

The construction in Statement 9.2.4.4 implements a special kind of filtering where the value $a^{\wedge A}$ in the pair of type $A \times G^A$ needs to pass the filter for any data to remain in the functor after filtering. We can use the same construction repeatedly with $G^\bullet \triangleq \mathbb{1}$ and obtain the type

$$L_n^A \triangleq \underbrace{\mathbb{1} + A \times (\mathbb{1} + A \times (\mathbb{1} + \dots \times (\mathbb{1} + A \times \mathbb{1}))))}_{\text{parameter } A \text{ is used } n \text{ times}} ,$$

which is equivalent to a list of up to n elements. The construction defines a filtering operation for L_n^\bullet that will delete any data beyond the first value of type A that does not pass the predicate. It is clear that this filtering operation implements the standard `takeWhile` method defined on sequences. So, `takeWhile` is a lawful filtering operation (see Example 9.1.4.3 where it was used).

We can also generalize the construction of Statement 9.2.4.4 to the functor

$$F^A \triangleq \mathbb{1} + \underbrace{A \times A \times \dots \times A \times G^A}_{n \text{ times}} .$$

We implement the filtering operation with the requirement that *all* n values of type A in the tuple $A \times A \times \dots \times A \times G^A$ must pass the filtering predicate, or else F^A becomes empty. Example 9.1.4.2 shows how such filtering operations may be used in practice.

Function types As we have seen in Chapter 6 (Statement 6.2.3.5), functors involving a function type, such as $F^A \triangleq G^A \rightarrow H^A$, require G^\bullet to be a *contrafunctor* rather than a functor. It turns out that the functor $G^A \rightarrow H^A$ is filterable only if the contrafunctor G^\bullet has certain properties (Eqs. (9.50)–(9.51) below) similar to properties of filterable functors. We will call such contrafunctors **filterable**.

To motivate the definition of filterable contrafunctors, consider the operation `liftOpt` for F :

$$\text{liftOpt}_F(f^{\wedge A \rightarrow \mathbb{1} + B}) : (G^A \rightarrow H^A) \rightarrow G^B \rightarrow H^B , \quad \text{liftOpt}_F(f) = p^{\wedge G^A \rightarrow H^A} \rightarrow g^{\wedge G^B} \rightarrow ???^{\wedge H^B} .$$

Assume that H is filterable, so that we have the function $\text{liftOpt}_H(f) : H^A \rightarrow H^B$. We will fill the typed hole $???^{\wedge H^B}$ if we somehow get a value of type H^A ; that is only possible if we apply $p^{\wedge G^A \rightarrow H^A}$,

$$\text{liftOpt}_F(f) = p^{\wedge G^A \rightarrow H^A} \rightarrow g^{\wedge G^B} \rightarrow \text{liftOpt}_H(f)(p(???^{\wedge G^A})) .$$

The only way to proceed is to have a function $G^B \rightarrow G^A$. We cannot obtain such a function by lifting f to the contrafunctor G : that gives $f^{\downarrow G} : G^{\mathbb{1} + B} \rightarrow G^A$. So, we need to require having a function

$$\text{liftOpt}_G(f^{\wedge A \rightarrow \mathbb{1} + B}) : G^B \rightarrow G^A . \quad (9.48)$$

This function is analogous to `liftOpt` for functors, except for the reverse direction of transformation ($G^B \rightarrow G^A$ instead of $G^A \rightarrow G^B$). We can now complete the implementation of `liftOpt`:

$$\begin{aligned} \text{liftOpt}_F(f^{\wedge A \rightarrow \mathbb{1} + B}) &\triangleq p^{\wedge G^A \rightarrow H^A} \rightarrow g^{\wedge G^B} \rightarrow \text{liftOpt}_H(f)(p(\text{liftOpt}_G(f)(g))) \\ \triangleright\text{-notation : } &= p^{\wedge G^A \rightarrow H^A} \rightarrow g^{\wedge G^B} \rightarrow g \triangleright \text{liftOpt}_G(f) \triangleright p \triangleright \text{liftOpt}_H(f) \\ \text{omit } (g \rightarrow g \triangleright) : &= p \rightarrow \text{liftOpt}_G(f) \circ p \circ \text{liftOpt}_H(f) . \end{aligned} \quad (9.49)$$

Note that the last line is similar to Eq. (6.15) but with `liftOpt` instead of `map`:

$$(f^{\wedge A \rightarrow B})^{\uparrow F} = p^{\wedge G^A \rightarrow H^A} \rightarrow f^{\downarrow G} \circ p \circ f^{\uparrow F} = p \rightarrow \text{cmap}_G(f) \circ p \circ \text{fmap}_F(f) .$$

The laws for filterable contrafunctors are chosen such that $F^A \triangleq G^A \rightarrow H^A$ can be shown to obey filtering laws when H^\bullet is a filterable functor and G^\bullet is a filterable contrafunctor.

Statement 9.2.4.5 Assume that H^\bullet is a lawful filterable functor and G^\bullet is a contrafunctor with a function liftOpt_G having type signature (9.48) and obeying the laws (9.50)–(9.51) shown below. Then the functor $F^A \triangleq G^A \rightarrow H^A$ is filterable.

Proof We will arrive at the required laws for G by trying to prove the laws for F .

Because of the function type of F^\bullet , it is convenient for derivations to apply both sides of the laws to an arbitrary value $p:G^A \rightarrow H^A$. Consider the naturality-identity law of F :

$$\begin{aligned} \text{expect to equal } p \triangleright f^{\uparrow F} &= f^{\downarrow G} ; p ; f^{\uparrow H} : p \triangleright \text{liftOpt}_F(f ; \text{pu}_{\text{Opt}}) \\ \text{definition (9.49) of liftOpt}_F &: = \text{liftOpt}_G(f ; \text{pu}_{\text{Opt}}) ; p ; \text{liftOpt}_H(f ; \text{pu}_{\text{Opt}}) \\ \text{naturality-identity law of liftOpt}_H &: = \text{liftOpt}_G(f ; \text{pu}_{\text{Opt}}) ; p ; f^{\uparrow H} . \end{aligned}$$

The only sub-expression that remains different is $\text{liftOpt}_G(f ; \text{pu}_{\text{Opt}})$, and the derivation will be finished if we assume the **naturality-identity law** of filterable contrafunctors to be

$$\text{liftOpt}_G(f ; \text{pu}_{\text{Opt}}) = f^{\downarrow G} . \quad (9.50)$$

The composition law of F applied to a value to a value $p:G^A \rightarrow H^A$ is

$$\begin{aligned} \text{left-hand side of Eq. (9.31) for } F &: p \triangleright \text{liftOpt}_F(f) ; \text{liftOpt}_F(g) = p \triangleright \text{liftOpt}_F(f) \triangleright \text{liftOpt}_F(g) \\ \text{definition (9.49) of liftOpt}_F &: = (\text{liftOpt}_G(f) ; p ; \text{liftOpt}_H(f)) \triangleright \text{liftOpt}_F(g) \\ \text{definition (9.49) :} &= \text{liftOpt}_G(g) ; (\text{liftOpt}_G(f) ; p ; \text{liftOpt}_H(f)) ; \text{liftOpt}_H(g) \\ \text{composition law (9.31) of liftOpt}_H &: = \text{liftOpt}_G(g) ; \text{liftOpt}_G(f) ; p ; \text{liftOpt}_H(f \diamond_{\text{Opt}} g) . \end{aligned}$$

The right-hand side of Eq. (9.31) for F is

$$p \triangleright \text{liftOpt}_F(f \diamond_{\text{Opt}} g) = \text{liftOpt}_G(f \diamond_{\text{Opt}} g) ; p ; \text{liftOpt}_H(f \diamond_{\text{Opt}} g) .$$

Clearly, we need to require the **composition law** of filterable contrafunctor G to be

$$\text{liftOpt}_G(g) ; \text{liftOpt}_G(f) = \text{liftOpt}_G(f \diamond_{\text{Opt}} g) . \quad (9.51)$$

Assuming that liftOpt_G satisfies this law, we obtain Eq. (9.31) for F and so conclude the proof.

Recursive types How to generalize the filtering operation from sequences to other recursive types? For motivation, we look at two examples: the filterable `List` functor defined by

$$\text{List}^A \triangleq \mathbb{1} + A \times \text{List}^A ,$$

and the recursive construction for ordinary functors (Statement 6.2.3.7) that requires a bifunctor $S^{\bullet,\bullet}$.

Statement 9.2.4.6 If G^\bullet is a filterable functor, the recursive functor F^\bullet defined by

$$F^A \triangleq G^A + A \times F^A$$

is filterable. With $G^A \triangleq \mathbb{1}$, this construction reproduces the standard filtering operation of `List`.

Proof We first need to implement the type constructor F and the function liftOpt_F :

```
sealed trait F[A] // Assume that the functor G was defined previously.
final case class FG[A](g: G[A]) extends F[A]
final case class FAF[A](a: A, rf: F[A]) extends F[A]
// Assume that liftOpt_G is available and define liftOpt_F:
def liftOpt_F[A, B](f: A => Option[B]): F[A] => F[B] = {
  case FG(g) => FG(liftOpt_G(f)(g))
  case FAF(a, rf) => f(a) match {
    case None => liftOpt_F(f)(rf) // Does 'a' pass the filtering predicate?
    case Some(b) => FAF[B](b, liftOpt_F(f)(rf)) // No. Drop 'a' and filter 'rf' recursively.
  }
}
```

		F^B
G^A	$g:G^A \rightarrow \text{liftOpt}_G(f)(g) + \mathbb{0}^{B \times F^B}$	
$A \times F^A$	$a:A \times r:F^A \rightarrow f(a) \triangleright$	$\begin{array}{c c} F^B & \\ \hline 1 & 1 \rightarrow \overline{\text{liftOpt}_F(f)(r)} \\ B & b:B \rightarrow \mathbb{0}^{G^B} + b \times \overline{\text{liftOpt}_F(f)(r)} \end{array}$

The overline denotes recursive uses of liftOpt_F within its definition.

With this definition of liftOpt_F , it is inconvenient to use matrix composition because the matrix shown above has a single column instead of columns split by the disjunctive type $F^A = G^A + A \times F^A$. We have seen a similar problem in the proof of Statement 9.2.4.4, where we rewrote the code via flmOpt and avoided using matrices with non-split columns. But it is not clear how to rewrite the code for liftOpt_F in that way. Instead, we use a more straightforward approach: apply both sides of the laws to an arbitrary value of type F^A .

The disjunctive type F^A has two cases, $G^A + \mathbb{0}^{A \times F^A}$ and $\mathbb{0}^{G^A} + A \times F^A$. When applied to a value $g:G^A + \mathbb{0}$, the function liftOpt_F is exactly the same as liftOpt_G , so both laws are satisfied since G^A is a lawful filterable functor. It remains to verify the laws when applied to a value $\mathbb{0}^{G^A} + a:A \times r:F^A$.

To prepare for the calculations, write the result of applying liftOpt_F to a value $\mathbb{0} + a \times r$:

		F^B
$(\mathbb{0}^{G^A} + a:A \times r:F^A) \triangleright \text{liftOpt}_F(f:A \rightarrow \mathbb{1} + B) = f(a) \triangleright$		$\begin{array}{c c} F^B & \\ \hline 1 & 1 \rightarrow r \triangleright \overline{\text{liftOpt}_F(f)(r)} \\ B & b:B \rightarrow \mathbb{0}^{G^B} + (b \times r) \triangleright \overline{\text{liftOpt}_F(f)(r)} \end{array}$

To check the naturality-identity law (9.29) of liftOpt_F , begin with the left-hand side:

$$\begin{aligned}
 (\mathbb{0} + a \times r) \triangleright \text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}}) &= (\mathbb{0} + a \times r) \triangleright \text{liftOpt}_F(x \rightarrow \mathbb{0} + f(x)) \\
 \text{use Eq. (9.52)} : &= a \triangleright (f \circ \text{pu}_{\text{Opt}}) \circ \begin{array}{c|c} 1 \rightarrow r \triangleright \overline{\text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}})(r)} \\ b:B \rightarrow \mathbb{0} + (b \times r) \triangleright \overline{\text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}})(r)} \end{array} \\
 \text{evaluate function of } a : &= (\mathbb{0} + f(a)) \triangleright \begin{array}{c|c} 1 \rightarrow r \triangleright \overline{\text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}})} \\ b:B \rightarrow \mathbb{0} + (b \times r) \triangleright \overline{\text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}})} \end{array} \\
 \text{substitute into matrix} : &= \mathbb{0} + (f(a) \times r) \triangleright \overline{\text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}})} \\
 \text{inductive assumption} : &= \mathbb{0} + (f(a) \times r) \triangleright \underline{f^{\uparrow F}} \quad .
 \end{aligned} \tag{9.52}$$

The same expression is found by applying the right-hand side of Eq. (9.29) to $\mathbb{0} + a \times r$:

$$(\mathbb{0}^{G^A} + a:A \times r:F^A) \triangleright (f:A \rightarrow B)^{\uparrow F} = \mathbb{0}^{G^A} + f(a) \times f^{\uparrow F}(r) = \mathbb{0} + f(a) \times r \triangleright f^{\uparrow F} \quad .$$

To verify the composition law (9.31), apply its left-hand side to $\mathbb{0} + a \times r$:

$$\begin{aligned}
 (\mathbb{0} + a \times r) \triangleright \text{liftOpt}_F(f) \circ \text{liftOpt}_F(g) \\
 \text{use Eq. (9.52)} : &= f(a) \triangleright \begin{array}{c|c} 1 \rightarrow r \triangleright \overline{\text{liftOpt}_F(f)} \\ b:B \rightarrow \mathbb{0} + (b \times r) \triangleright \overline{\text{liftOpt}_F(f)} \end{array} \circ \underline{\text{liftOpt}_F(g)}
 \end{aligned}$$

$$\begin{array}{ll}
 \text{apply liftOpt}_F(g) : & = f(a) \triangleright \left\| \begin{array}{l} 1 \rightarrow r \triangleright \overline{\text{liftOpt}_F}(f) \triangleright \text{liftOpt}_F(g) \\ b^B \rightarrow (0 + (b \times r) \triangleright \overline{\text{liftOpt}_F}(f)) \triangleright \text{liftOpt}_F(g) \end{array} \right\| \\
 \text{use Eq. (9.52)} : & = f(a) \triangleright \left\| \begin{array}{l} 1 \rightarrow r \triangleright \overline{\text{liftOpt}_F}(f) ; \text{liftOpt}_F(g) \\ b^B \rightarrow g(b) \triangleright \left\| \begin{array}{l} 1 \rightarrow r \triangleright \overline{\text{liftOpt}_F}(f) ; \overline{\text{liftOpt}_F}(g) \\ c^C \rightarrow 0 + (c \times r) \triangleright \overline{\text{liftOpt}_F}(f) ; \overline{\text{liftOpt}_F}(g) \end{array} \right\| \end{array} \right\| \quad (9.53)
 \end{array}$$

$$\text{inductive assumption :} \quad = a \triangleright f ; \left\| \begin{array}{l} 1 \rightarrow r \triangleright \overline{\text{liftOpt}_F}(f \diamond_{\text{Opt}} g) \\ b^B \rightarrow g(b) \triangleright \left\| \begin{array}{l} 1 \rightarrow r \triangleright \overline{\text{liftOpt}_F}(f \diamond_{\text{Opt}} g) \\ c^C \rightarrow 0 + (c \times r) \triangleright \overline{\text{liftOpt}_F}(f \diamond_{\text{Opt}} g) \end{array} \right\| \end{array} \right\| \quad . \quad (9.54)$$

We are justified in using the inductive assumption on the composition $\overline{\text{liftOpt}_F}(f) ; \text{liftOpt}_F(g)$ even though $\text{liftOpt}_F(g)$ is not a recursive call. It is sufficient that at least one term in the function composition is a recursive call to $\overline{\text{liftOpt}_F}$.

The right-hand side of Eq. (9.31) applied to $0 + a \times r$ gives

$$\begin{array}{ll}
 & (0 + a \times r) \triangleright \text{liftOpt}_F(f \diamond_{\text{Opt}} g) \\
 \text{use Eq. (9.52)} : & = a \triangleright (f \diamond_{\text{Opt}} g) \triangleright \left\| \begin{array}{l} 1 \rightarrow r \triangleright \overline{\text{liftOpt}_F}(f \diamond_{\text{Opt}} g) \\ c^C \rightarrow 0 + (c \times r) \triangleright \overline{\text{liftOpt}_F}(f \diamond_{\text{Opt}} g) \end{array} \right\| \quad .
 \end{array}$$

The only remaining difference compared with the last expression in Eq. (9.54) is

$$f ; \left\| \begin{array}{l} p \\ b^B \rightarrow g(b) \triangleright \left\| \begin{array}{l} p \\ q \end{array} \right\| \end{array} \right\| \stackrel{?}{=} (f \diamond_{\text{Opt}} g) ; \left\| \begin{array}{l} p \\ q \end{array} \right\| \quad , \quad (9.55)$$

where $p \triangleq 1 \rightarrow r \triangleright \text{liftOpt}_F(f \diamond_{\text{Opt}} g)$ and $q \triangleq c \rightarrow 0 + c \times r \triangleright \text{liftOpt}_F(f \diamond_{\text{Opt}} g)$ are some fixed functions. We can show that Eq. (9.55) holds for arbitrary functions p and q . Start from the right-hand side:

$$\begin{array}{ll}
 \text{expect the l.h.s. of Eq. (9.55)} : & (f \diamond_{\text{Opt}} g) ; \left\| \begin{array}{l} p \\ q \end{array} \right\| = f ; \overline{\text{flmOpt}}(g) ; \left\| \begin{array}{l} p \\ q \end{array} \right\| = f ; \left\| \begin{array}{l} 1 \rightarrow 1 + 0 \\ b \rightarrow g(b) \end{array} \right\| ; \left\| \begin{array}{l} p \\ q \end{array} \right\| \\
 \text{apply } \left\| \begin{array}{l} p \\ q \end{array} \right\| : & = f ; \left\| \begin{array}{l} 1 \rightarrow (1 + 0) \triangleright \left\| \begin{array}{l} p \\ q \end{array} \right\| \\ b \rightarrow g(b) \triangleright \left\| \begin{array}{l} p \\ q \end{array} \right\| \end{array} \right\| = f ; \left\| \begin{array}{l} 1 \rightarrow p \\ b \rightarrow g(b) \triangleright \left\| \begin{array}{l} p \\ q \end{array} \right\| \end{array} \right\| \quad .
 \end{array}$$

This proves Eq. (9.55) and concludes the proof of Statement 9.2.4.6.

This implementation preserves all values x^A except those that fail the filtering predicate (i.e., when $f(x) = 1 + 0$). So, it has the same filtering logic as the standard `filter` method for sequences.

Statement 9.2.4.7 If $S^{A,R}$ is a bifunctor that is filterable with respect to A , the recursive functor F^A defined by the type equation $F^A \triangleq S^{A,F^A}$ is filterable.

Proof We follow the proof of Statement 6.2.3.7. Implement liftOpt_F recursively as

$$\text{liftOpt}_F(f : A \rightarrow 1 + B) \triangleq \text{liftOpt}_S(f) ; \text{bimap}_S(\text{id})(\overline{\text{liftOpt}_F}(f)) = \text{liftOpt}_S(f) ; (\overline{\text{liftOpt}_F}(f))^{\uparrow S^{B,*}} \quad .$$

where liftOpt_S is assumed to obey the laws. The lifting to F is defined (also recursively) by

$$(f:A \rightarrow B)^{\uparrow F} \triangleq \text{bimap}_S(f)(\overline{f^{\uparrow F}}) = f^{\uparrow S^{\bullet,R}} \circ (\overline{f^{\uparrow F}})^{\uparrow S^{B,\bullet}} \quad .$$

To verify the naturality-identity law (9.29), compute

$$\begin{aligned} \text{expect to equal } f^{\uparrow F} : \quad & \text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}}) = \text{liftOpt}_S(f \circ \text{pu}_{\text{Opt}}) \circ (\overline{\text{liftOpt}_F}(f \circ \text{pu}_{\text{Opt}}))^{\uparrow S^{B,\bullet}} \\ \text{law (9.29) of liftOpt}_S : \quad & = f^{\uparrow S^{\bullet,R}} \circ (\overline{\text{liftOpt}_F}(f \circ \text{pu}_{\text{Opt}}))^{\uparrow S^{B,\bullet}} \\ \text{inductive assumption :} \quad & = f^{\uparrow S^{\bullet,R}} \circ (\overline{f^{\uparrow F}})^{\uparrow S^{B,\bullet}} = f^{\uparrow F} \quad . \end{aligned}$$

To verify the composition law (9.31), compute

$$\begin{aligned} & \text{liftOpt}_F(f) \circ \text{liftOpt}_F(g) \\ & = \text{liftOpt}_S(f) \circ (\overline{\text{liftOpt}_F}(f))^{\uparrow S^{B,\bullet}} \circ \text{liftOpt}_S(g) \circ (\overline{\text{liftOpt}_F}(g))^{\uparrow S^{C,\bullet}} \\ \text{naturality of liftOpt}_S : \quad & = \text{liftOpt}_S(f) \circ \text{liftOpt}_S(g) \circ (\overline{\text{liftOpt}_F}(f))^{\uparrow S^{C,\bullet}} \circ (\overline{\text{liftOpt}_F}(g))^{\uparrow S^{C,\bullet}} \\ \text{law (9.31) of liftOpt}_S : \quad & = \text{liftOpt}_S(f \diamond_{\text{Opt}} g) \circ (\overline{\text{liftOpt}_F}(f) \circ \overline{\text{liftOpt}_F}(g))^{\uparrow S^{C,\bullet}} \\ \text{inductive assumption :} \quad & = \text{liftOpt}_S(f \diamond_{\text{Opt}} g) \circ \overline{\text{liftOpt}_F}(f \diamond_{\text{Opt}} g) = \text{liftOpt}_F(f \diamond_{\text{Opt}} g) \quad . \end{aligned}$$

We have assumed a **naturality law** of liftOpt_S with respect to the type parameter R of $S^{A,R}$:

$$(p:R \rightarrow R')^{\uparrow S^{B,\bullet}} \circ \text{liftOpt}_S(g:R \rightarrow \mathbb{1} + C) = \text{liftOpt}_S(g:R \rightarrow \mathbb{1} + C) \circ (p:R \rightarrow R')^{\uparrow S^{C,\bullet}} \quad .$$

$$\begin{array}{ccc} S^{B,R} & \xrightarrow{\text{liftOpt}_S(g:R \rightarrow \mathbb{1} + C)} & S^{C,R} \\ (p:R \rightarrow R')^{\uparrow S^{B,\bullet}} \downarrow & & \downarrow (p:R \rightarrow R')^{\uparrow S^{C,\bullet}} \\ S^{B,R'} & \xrightarrow{\text{liftOpt}_S(g:R \rightarrow \mathbb{1} + C)} & S^{C,R'} \end{array}$$

Because this law interchanges liftOpt_S with a lifting that affects the *other* type parameter of $S^{\bullet,\bullet}$, all fully parametric functions will obey this law (see the discussion at the end of Section 6.2.2).

9.2.5 Filterable contrafunctors: motivation and examples

An intuitive view is that functors are “wrappers” of data, while contrafunctors “consume” data. Filterable functors permit us to exclude certain data from a wrapper; filterable contrafunctors permit us to exclude certain data from being *consumed*. Let us now make this intuition precise.

A simple contrafunctor is $C^A \triangleq A \rightarrow Z$, where Z is a constant type. This is a general form of an “extractor”—for example, a function that extracts logging information of type Z from data of an arbitrary type A . It is sometimes necessary to exclude particular kinds of information (e.g., private personal data) from logging. We can implement this by providing a predicate of type $A \Rightarrow \text{Boolean}$ that decides, depending on the given value $x:A$, whether x should be passed to the extractor. That predicate will be attached to a given extractor $c:C^A$ by the `filter` operation:

```
def filter[A](p: A => Boolean): C[A] => C[A]
val extractor: C[Payload] = ??? // Original code for extracting metadata from payloads.
val noPrivateData: Payload => Boolean = ??? // Returns true only if payload has no private data.
val filtered: C[Payload] = filter(noPrivateData)(extractor) // Will not extract private data.
```

How could filtering work if the predicate returns `false`? Even if the data (of type A) is excluded from the filtered extractor, the function of type $A \rightarrow Z$ must still return a value of type Z . A solution is to have a *default value* of type Z that the extractor will return when data is excluded from it.

There are two simple ways of implementing filtering for $C^A = A \rightarrow Z$ via a default value of type Z : First, by storing that default value in the type C^A and considering the contrafunctor $C^A \triangleq Z \times (A \rightarrow Z)$ instead of $A \rightarrow Z$. Second, by using a type $\mathbb{1} + Z$ instead of Z , since the type $\mathbb{1} + Z$ has a default value $1 + \mathbb{0}^Z$. The following two examples show how this works.

Example 9.2.5.1 Implement `filter` for the contrafunctor $C^A \triangleq A \rightarrow \mathbb{1} + Z$, where Z is a fixed type.

Solution This contrafunctor “consumes” data of type A and computes values of type $\mathbb{1} + Z$. Given a value c of type C^A , which is a function that we might write like this,

```
val c: A => Option[Z] = ???
```

we need somehow to impose a filter predicate $p: A \rightarrow \mathbb{2}$ ensuring that the function c is applied only to values that pass the predicate. The result will be a new function $d: A \rightarrow \mathbb{1} + Z$ that will use its argument only if it passes the predicate. The function d could return `None` for all arguments, but that implementation would lose information. If $p(a) == \text{true}$, we may compute $c(a)$, getting a value of type $\mathbb{1} + Z$ that d should return. If $p(a) == \text{false}$, the function d must return `None`. So, the code of d must be

```
val d: A => Option[Z] = { a => if (p(a)) c(a) else None }
```

The transformation from c to d is a filtering operation for the contrafunctor C^A , implemented as

```
def filter[A](p: A => Boolean)(c: A => Option[Z]): A => Option[Z] = { a =>
  if (p(a)) c(a) else None
}
```

// Equivalent code is { a => Some(a).filter(p).flatMap(c) }

$$\text{filt}_C(p: A \rightarrow \mathbb{2}) \triangleq c: A \rightarrow \mathbb{1} + Z \rightarrow \text{pu}_{\text{Opt}} \circ \text{filtOpt}(p) \circ \text{flm}_{\text{Opt}}(c) = c \rightarrow \psi_p \circ \text{flm}_{\text{Opt}}(c) .$$

Example 9.2.5.2 Implement `filter` for the contrafunctor $C^A \triangleq Z \times (A \rightarrow Z)$.

Solution The code for the `filter` function is

```
def filter[A](p: A => Boolean): ((Z, A => Z)) => (Z, A => Z) = {
  case (z, f) => (z, a => if (p(a)) f(a) else z)
}
```

$$\text{filt}_C(p: A \rightarrow \mathbb{2}) \triangleq z: Z \times f: A \rightarrow Z \rightarrow z \times (a: A \rightarrow p(a) \triangleright \begin{array}{c|c} & Z \\ \hline 1(\text{false}) & z \\ 1(\text{true}) & f(a) \end{array}) .$$

Note that the type $Z \times (A \rightarrow Z)$ is equivalent to $\mathbb{1} + A \rightarrow Z$:

$$Z \times (A \rightarrow Z) \cong (\mathbb{1} \rightarrow Z) \times (A \rightarrow Z) \cong \mathbb{1} + A \rightarrow Z .$$

For the contrafunctor $C^A \triangleq \mathbb{1} + A \rightarrow Z$, the `filter` function is implemented by

```
def filter[A](p: A => Boolean)(c: Option[A] => Z): Option[A] => Z = {
  case Some(a) if p(a) => c(Some(a)) // Only apply 'c' to 'a' if 'p(a) == \text{true}'.
  case _ => c(None) // Return c(None) otherwise, or for empty Option.
}
```

// Equivalent code is `_.filter(p).pipe(c)` (Scala 2.13).

$$\begin{aligned} \text{filt}_C(p: A \rightarrow \mathbb{2}) &\triangleq c: \mathbb{1} + A \rightarrow Z \rightarrow \underline{x: \mathbb{1} + A \rightarrow x} \triangleright \text{filtOpt}(p) \triangleright c \\ &= c: \mathbb{1} + A \rightarrow Z \rightarrow \text{filtOpt}(p) \circ c . \end{aligned}$$

Another motivation for filterable contrafunctors comes from the construction $F^A \triangleq G^A \rightarrow H^A$ (Statement 9.2.4.5): In order to assure the properties of a filterable functor for F^\bullet , the contrafunctor G^\bullet must have the `liftOpt` function as shown in Eq. (9.48). The existence of the `liftOpt` function for a contrafunctor turns out to be equivalent to the existence of the `filter` function, if suitable laws are

assumed to hold. To verify that equivalence, begin by defining the function `inflate`, whose role is similar to that of `deflate` for filterable functors. The type signature of `inflate` is

$$\text{inflate}_C : C^A \rightarrow C^{\mathbb{1}+A} .$$

We can relate `inflate` to `filter` and `liftOpt` by the following equations (to be derived below):

$$\text{filt}_C(p) = \text{inflate}_C \circ (\psi_p^{A \rightarrow \mathbb{1}+A})^{\downarrow C} , \quad \text{inflate}_C = (\text{get}^{\mathbb{1}+A \rightarrow A})^{\downarrow C} \circ \text{filt}_C(\text{nonEmpty}) , \quad (9.56)$$

$$\text{inflate}_C = \text{liftOpt}_C(\text{id}^{\mathbb{1}+A \rightarrow \mathbb{1}+A}) , \quad \text{liftOpt}_C(f^{A \rightarrow \mathbb{1}+B}) = \text{inflate}_C \circ f^{\downarrow C} . \quad (9.57)$$

$$\begin{array}{ccc} C^{\mathbb{1}+A} & \xleftarrow{\text{get}^{\downarrow C}} & C^A \xrightarrow{\text{filt}_C(p^{A \rightarrow \mathbb{2}})} C^A \\ \text{filt}_C^{\mathbb{1}+A}(\text{nonEmpty}^{\mathbb{1}+A \rightarrow \mathbb{2}}) \searrow & \downarrow \text{inflate}_C & \swarrow \psi_p^{\downarrow C} \\ C^{\mathbb{1}+A} & & \end{array} \quad \begin{array}{ccc} C^B & \xrightarrow{\text{inflate}_C} & C^{\mathbb{1}+B} \\ \text{liftOpt}_C(f) \triangleq \searrow & \downarrow (f^{A \rightarrow \mathbb{1}+B})^{\downarrow C} & \swarrow \\ C^A & & \end{array}$$

These functions have different but equivalent laws: filt_C has 4 laws, inflate_C has 3, and liftOpt_C has just 2 laws. So, liftOpt_C is the most convenient function for proving laws, while inflate_C is the easiest to implement in code (and to check whether a given contrafunctor is filterable). The laws of liftOpt_C are similar to the laws of `liftOpt` for filterable functors (we omit the derivations):

naturality law of liftOpt_C : $\text{liftOpt}_C(f^{A \rightarrow B} \circ g^{B \rightarrow \mathbb{1}+E}) = \text{liftOpt}_C(g) \circ f^{\downarrow C} ,$

naturality-identity law of liftOpt_C : $\text{liftOpt}_C(f^{A \rightarrow B} \circ \text{pu}_{\text{Opt}}^{B \rightarrow \mathbb{1}+B}) = f^{\downarrow C} ,$

composition law of liftOpt_C : $\text{liftOpt}_C(g^{B \rightarrow \mathbb{1}+E}) \circ \text{liftOpt}_C(f^{A \rightarrow \mathbb{1}+B}) = \text{liftOpt}_C(f \circ_{\text{Opt}} g) .$

As an illustration, let us implement these functions for the contrafunctor $C^A \triangleq A \rightarrow \mathbb{1} + Z$.

Example 9.2.5.3 For $C^A \triangleq A \rightarrow \mathbb{1} + Z$, implement `inflate` and `liftOpt`, and directly verify Eq. (9.56).

Solution We implement the type signatures of `inflate` and `liftOpt`, preserving information:

```
def inflate[A](c: A => Option[Z]): Option[A] => Option[Z] = _.flatMap(c)
def liftOpt[A, B](f: A => Option[B])(c: B => Option[Z]): A => Option[Z] = { a => f(a).flatMap(c) }
```

$\text{inflate}_C \triangleq c^{A \rightarrow \mathbb{1}+Z} \rightarrow \text{flmOpt}(c) , \quad \text{liftOpt}_C(f^{A \rightarrow \mathbb{1}+B}) \triangleq c^{B \rightarrow \mathbb{1}+Z} \rightarrow f \circ \text{flmOpt}(c) = c \rightarrow f \circ_{\text{Opt}} c .$

To verify Eq. (9.56), we need the code for lifting a function $g^{A \rightarrow B}$ to the contrafunctor C :

$$(g^{A \rightarrow B})^{\downarrow C} = c^{B \rightarrow \mathbb{1}+Z} \rightarrow a^A \rightarrow c(g(a)) = c^{B \rightarrow \mathbb{1}+Z} \rightarrow g \circ c .$$

Now we use the code for filt_C from Example 9.2.5.1 and compute $\text{get}^{\downarrow C} \circ \text{filt}_C(\text{nonEmpty})$ as

expect to equal inflate_C : $\text{get}^{\downarrow C} \circ \text{filt}_C(\text{nonEmpty})$

definitions of $\downarrow C$ and filt_C : $= (c \rightarrow \text{get} \circ c) \circ (c \rightarrow \psi_{\text{nonEmpty}} \circ \text{flmOpt}(c))$

compute composition : $= c \rightarrow \psi_{\text{nonEmpty}} \circ \text{flmOpt}(\text{get} \circ c)$

use naturality law (9.46) of flmOpt : $= c \rightarrow \psi_{\text{nonEmpty}} \circ \text{get}^{\uparrow \text{Opt}} \circ \text{flmOpt}(c)$

use Eq. (9.17) to simplify : $= c \rightarrow \text{flmOpt}(c) = \text{inflate}_C .$

In the formula (9.56), the lifted partial function $\text{get}^{\downarrow C}$ is applied *before* the filtering operation (rather than after filtering, as would be the case for filterable functors). The derivation shows how the partial function `get` is moved around due to the reverse order of contrafunctor-lifted function composition, until we find the expression $\psi_{\text{nonEmpty}} \circ \text{get}^{\uparrow \text{Opt}}$ where the partial function `get` is applied *after* a filter ψ (which makes it a total function due to `filter`'s partial function law). For this reason, it is correct to apply lifted partial functions before filtering in contrafunctors.

Example 9.2.5.4 Verify Eq. (9.56) for an arbitrary filterable contrafunctor C^\bullet , assuming needed laws.

Solution We need to check the two directions of the isomorphism in Eq. (9.56).

First part: starting from a given filt_C , we compute inflate_C and then use that to define a new filt'_C ; we must show that $\text{filt}_C = \text{filt}'_C$.

$$\text{expect to equal } \text{filt}_C(p) : \text{filt}'_C(p) = \text{inflate}_C \circ \psi_p^{\downarrow C} = \text{get}^{\downarrow C} \circ \text{filt}_C(\text{nonEmpty}) \circ \psi_p^{\downarrow C} .$$

The computation gets stuck here: We could simplify the composition of ψ_p and get , if only we could move these functions next to each other. It is clear that we need a law that exchanges the order of compositions of filt_C with lifted functions. Typically, that is done by naturality laws. By analogy with Eq. (9.3) and making sure types match, we write a **naturality law** of filt_C as

$$\text{filt}_C(p^{\cdot A \rightarrow 2}) \circ (f^{\cdot B \rightarrow A})^{\downarrow C} = f^{\downarrow C} \circ \text{filt}_C(f \circ p) . \quad (9.58)$$

Assuming this law, we find that $\text{filt}'_C(p) = \text{filt}_C(p)$:

$$\text{use naturality law of } \text{filt}_C : \text{get}^{\downarrow C} \circ \text{filt}_C(\text{nonEmpty}) \circ \psi_p^{\downarrow C} = \text{get}^{\downarrow C} \circ \psi_p^{\downarrow C} \circ \text{filt}_C(\psi_p \circ \text{nonEmpty})$$

$$\text{use Eqs. (9.11)–(9.12)} : = \text{id}_{|p}^{\downarrow C} \circ \text{filt}_C(p)$$

$$\text{partial function law of } \text{filt}_C : = \text{id}^{\downarrow C} \circ \text{filt}_C(p) = \text{filt}_C(p) .$$

Here we assumed the partial function law in the form similar to that for functors,

$$f_{|p}^{\downarrow C} \circ \text{filt}_C(p) = f^{\downarrow C} \circ \text{filt}_C(p) .$$

The lifted partial function $f_{|p}^{\downarrow C}$ is applied *before* filtering, as appropriate for filterable contrafunctors.

Second part: starting from a given inflate_C , we compute filt_C and then use that to define a new $\text{inflate}'_C$; we must show that $\text{inflate}_C = \text{inflate}'_C$.

$$\text{expect to equal } \text{inflate}_C : \text{inflate}'_C = \text{get}^{\downarrow C} \circ \text{filt}_C(\text{nonEmpty}) = \text{get}^{\downarrow C} \circ \text{inflate}_C \circ \psi_{\text{nonEmpty}}^{\downarrow C} .$$

The calculation cannot proceed unless we can exchange lifted functions around inflate . By analogy with Eq. (9.14) and making changes suitable for contrafunctors, we obtain the **naturality law** for the inflate_C function:

$$\begin{array}{ccc} C^A & \xrightarrow{\text{inflate}_C} & C^{\mathbb{1}+A} \\ (f^{\cdot B \rightarrow A})^{\downarrow C} \downarrow & & \downarrow (f^{\cdot B \rightarrow A})^{\uparrow \text{Opt} \downarrow C} \\ C^B & \xrightarrow{\text{inflate}_C} & C^{\mathbb{1}+B} \end{array} \quad \text{With help of this law, we can finish the derivation:} \quad (9.59)$$

$$\text{use the naturality law} : \text{get}^{\downarrow C} \circ \text{inflate}_C \circ \psi_{\text{nonEmpty}}^{\downarrow C} = \text{inflate}_C \circ \text{get}^{\uparrow \text{Opt} \downarrow C} \circ \psi_{\text{nonEmpty}}^{\downarrow C}$$

$$\text{composition lifted to } C : = \text{inflate}_C \circ (\psi_{\text{nonEmpty}} \circ \text{get}^{\uparrow \text{Opt}})^{\downarrow C}$$

$$\text{use Eq. (9.17)} : = \text{inflate}_C \circ \text{id}^{\downarrow C} = \text{inflate}_C .$$

Exercise 9.2.5.5 Implement inflate and liftOpt for the filterable contrafunctor $C^A \triangleq \mathbb{1} + A \rightarrow \mathbb{Z}$.

Exercise 9.2.5.6* Verify Eq. (9.57) for an arbitrary filterable contrafunctor C^\bullet , assuming needed laws.

9.2.6 Constructions of filterable contrafunctors

How to build up a filterable contrafunctor from parts? Structural analysis produces a number of type constructions guaranteed to create filterable contrafunctors.

The `Filterable` typeclass is inductive (see Section 8.5.4) if formulated via `filter` or via `liftOpt`, because those methods return the type C^A itself. So, we expect that the product, the exponential, and the recursive constructions will apply to filterable contrafunctors (as they do to filterable functors).

Type parameters Constant contrafunctors $C^A \triangleq Z$ are “trivially” filterable: all methods are identity functions, so all laws hold.

Further constructions that work with type parameters are functor compositions. The composition $P \circ Q \triangleq P^{Q^\bullet}$ is a contrafunctor when P is a functor and Q is a contrafunctor, or vice versa. The contrafunctor P^{Q^\bullet} is filterable if Q^\bullet (whether it is a functor or a contrafunctor) is filterable:

Statement 9.2.6.1 (a) If P^\bullet is any contrafunctor and Q^\bullet is a filterable functor then P^{Q^\bullet} is filterable.

(b) If P^\bullet is any functor and Q^\bullet is a filterable contrafunctor then P^{Q^\bullet} is filterable.

Proof We follow the proof of Statement 9.2.4.1 with necessary modifications.

(a) We define the `liftOpt` operation for $P \circ Q$ as $\text{liftOpt}_{P \circ Q}(f: A \rightarrow \text{Option}[B]) \triangleq (\text{liftOpt}_Q(f)) \downarrow^P$.

```
def liftOpt_PQ[A, B](f: A => Option[B]): P[Q[B]] => P[Q[A]] = _.contramap(liftOpt_Q(f))
```

To verify the naturality-identity law (9.50):

$$\begin{aligned} \text{expect to equal } f \downarrow^{(P \circ Q)} : \text{liftOpt}_{P \circ Q}(f \circ \text{pu}_{\text{Opt}}) &= (\text{liftOpt}_Q(f \circ \text{pu}_{\text{Opt}})) \downarrow^P \\ \text{naturality-identity law (9.29) of } Q : &= f \uparrow^Q \downarrow^P = f \downarrow^{(P \circ Q)} . \end{aligned}$$

To verify the composition law (9.51), we show that its left-hand side equals $\text{liftOpt}_{P \circ Q}(f \diamond_{\text{Opt}} g)$:

$$\begin{aligned} \text{definition of liftOpt}_{P \circ Q} : \text{liftOpt}_{P \circ Q}(g) \circ \text{liftOpt}_{P \circ Q}(f) &= (\text{liftOpt}_Q(g)) \downarrow^P \circ (\text{liftOpt}_Q(f)) \downarrow^P \\ \text{composition law of } P : &= (\text{liftOpt}_Q(f) \circ \text{liftOpt}_Q(g)) \downarrow^P \\ \text{composition law (9.31) of } Q : &= (\text{liftOpt}_Q(f \diamond_{\text{Opt}} g)) \downarrow^P = \text{liftOpt}_{P \circ Q}(f \diamond_{\text{Opt}} g) . \end{aligned}$$

(b) We define the `liftOpt` operation for $P \circ Q$ as $\text{liftOpt}_{P \circ Q}(f: A \rightarrow \text{Option}[B]) \triangleq (\text{liftOpt}_Q(f)) \uparrow^P$.

```
def liftOpt_PQ[A, B](f: A => Option[B]): P[Q[B]] => P[Q[A]] = _.map(liftOpt_Q(f))
```

To verify the naturality-identity law (9.50):

$$\begin{aligned} \text{expect to equal } f \downarrow^{(P \circ Q)} : \text{liftOpt}_{P \circ Q}(f \circ \text{pu}_{\text{Opt}}) &= (\text{liftOpt}_Q(f \circ \text{pu}_{\text{Opt}})) \uparrow^P \\ \text{naturality-identity law (9.29) of } Q : &= f \downarrow^Q \uparrow^P = f \downarrow^{(P \circ Q)} . \end{aligned}$$

To verify the composition law (9.51), we show that its left-hand side equals $\text{liftOpt}_{P \circ Q}(f \diamond_{\text{Opt}} g)$:

$$\begin{aligned} \text{definition of liftOpt}_{P \circ Q} : \text{liftOpt}_{P \circ Q}(g) \circ \text{liftOpt}_{P \circ Q}(f) &= (\text{liftOpt}_Q(g)) \uparrow^P \circ (\text{liftOpt}_Q(f)) \uparrow^P \\ \text{composition law of } P : &= (\text{liftOpt}_Q(g) \circ \text{liftOpt}_Q(f)) \uparrow^P \\ \text{composition law (9.51) of } Q : &= (\text{liftOpt}_Q(f \diamond_{\text{Opt}} g)) \uparrow^P = \text{liftOpt}_{P \circ Q}(f \diamond_{\text{Opt}} g) . \end{aligned}$$

Composition of two filterable contrafunctors is a filterable *functor* (Exercise 9.3.2.1).

Products and co-products If G^\bullet and H^\bullet are filterable contrafunctors, the product $G^A \times H^A$ and the co-product $G^A + H^A$ will also be filterable contrafunctors. Proofs are analogous to the case of filterable functors and are delegated to Exercise 9.3.2.11.

Functions We have a construction similar to that of Statement 9.2.4.5 for filterable functors:

Statement 9.2.6.2 The contrafunctor $F^A \triangleq G^A \rightarrow H^A$ is filterable for any filterable functor G^A and any filterable contrafunctor H^A .

Proof We define the `liftOpt` operation for F^\bullet by

```
def liftOpt_F[A, B](f: A => Option[B])(p: G[B] => H[B]): G[A] => H[A] =
  { ga => liftOpt_H(f)(p(liftOpt_G(f)(ga))) }
```

To obtain a clearer code formula, rewrite the Scala code using the \triangleright -notation and then simplify:

$$\begin{aligned} \text{liftOpt}_F(f) &\triangleq p^{G^B \rightarrow H^B} \rightarrow g^{G^A} \rightarrow g \triangleright \text{liftOpt}_G(f) \triangleright p \triangleright \text{liftOpt}_H(f) \\ \text{simplify } (x \rightarrow x \triangleright y) = y : &= p^{G^B \rightarrow H^B} \rightarrow \text{liftOpt}_G(f) \circ p \circ \text{liftOpt}_H(f) . \end{aligned}$$

To verify the naturality-identity law (9.50), apply both its sides to an arbitrary $p^{G^B \rightarrow H^B}$:

$$\begin{aligned} \text{expect to equal } p \triangleright f^{\downarrow F} : & p \triangleright \text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}}) = \text{liftOpt}_G(f \circ \text{pu}_{\text{Opt}}) \circ p \circ \text{liftOpt}_H(f \circ \text{pu}_{\text{Opt}}) \\ \text{use laws (9.50) and (9.29)} : &= f^{\uparrow G} \circ p \circ f^{\downarrow H} = p \triangleright f^{\downarrow F} . \end{aligned}$$

To verify the composition law (9.51), show that its left-hand side equals $p \triangleright \text{liftOpt}_F(f \diamond_{\text{Opt}} g)$:

$$\begin{aligned} \triangleright\text{-notation} : & p \triangleright \text{liftOpt}_F(g) \circ \text{liftOpt}_F(f) = (p \triangleright \text{liftOpt}_F(g)) \triangleright \text{liftOpt}_F(f) \\ \text{definition of liftOpt}_F(f) : &= \text{liftOpt}_G(f) \circ (p \triangleright \text{liftOpt}_F(g)) \circ \text{liftOpt}_H(f) \\ \text{definition of liftOpt}_F(g) : &= \text{liftOpt}_G(f) \circ \text{liftOpt}_G(g) \circ p \circ \text{liftOpt}_H(g) \circ \text{liftOpt}_H(f) \\ \text{composition laws of } G \text{ and } H : &= \text{liftOpt}_G(f \diamond_{\text{Opt}} g) \circ p \circ \text{liftOpt}_H(f \diamond_{\text{Opt}} g) \\ \text{definition of liftOpt}_F(f \diamond_{\text{Opt}} g) : &= p \triangleright \text{liftOpt}_F(f \diamond_{\text{Opt}} g) . \end{aligned}$$

Example 9.2.6.3 (the search functor) A non-trivial application of Statement 9.2.6.2 is the **search functor** S_Z^A defined by $S_Z^A \triangleq (A \rightarrow \mathbb{1} + Z) \rightarrow \mathbb{1} + A$, where Z is a fixed type. This functor is filterable because it is a function from the filterable contrafunctor $A \rightarrow \mathbb{1} + Z$ to the filterable functor $\mathbb{1} + A$, and so it is filterable. The simplest case of the search functor is found by setting $Z \triangleq \mathbb{1}$, which gives the type constructor

$$S_{\mathbb{1}}^A \triangleq (A \rightarrow \mathbb{2}) \rightarrow \mathbb{1} + A .$$

Values of type $S_{\mathbb{1}}^A$ may be viewed as “searchers” taking a predicate $q^{A \rightarrow \mathbb{2}}$ and looking for a value of type A that satisfies the predicate. A searcher will return either a suitable value of type $\mathbb{0} + A$, or an empty value $\mathbb{1} + \mathbb{0}$ (“not found”). Applying a filter with an a predicate $p^{A \rightarrow \mathbb{2}}$ to a searcher will exclude values of type A from the search unless they satisfy p .

Another function-type construction is a generalization of the filterable contrafunctor $A \rightarrow \mathbb{1} + Z$.

Statement 9.2.6.4 If a contrafunctor H^{\bullet} is filterable, so is the contrafunctor $F^A \triangleq A \rightarrow \mathbb{1} + H^A$.

Proof We extend the implementation of `liftOpt` from Example 9.2.5.3:

```
def liftOpt_F[A, B](f: A => Option[B])(c: B => Option[H[B]]): A => Option[H[A]] =
  { a => f(a).flatMap(c).map(liftOpt_H(f)) }
```

$$\begin{aligned} \text{liftOpt}_F(f^{A \rightarrow \mathbb{1} + B}) &\triangleq p^{B \rightarrow \mathbb{1} + H^B} \rightarrow a^{A \rightarrow a \triangleright f} \triangleright \text{flm}_{\text{Opt}}(p) \triangleright (\text{liftOpt}_H(f))^{\uparrow \text{Opt}} \\ &= p \rightarrow f \circ \text{flm}_{\text{Opt}}(p) \circ (\text{liftOpt}_H(f))^{\uparrow \text{Opt}} . \end{aligned}$$

To verify the naturality-identity law (9.50), apply both sides to an arbitrary $p^{B \rightarrow \mathbb{1} + H^B}$:

$$\begin{aligned} \text{expect to equal } p \triangleright f^{\downarrow F} : & p \triangleright \text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}}) = f \circ \text{pu}_{\text{Opt}} \circ \text{flm}_{\text{Opt}}(p) \circ (\text{liftOpt}_H(f \circ \text{pu}_{\text{Opt}}))^{\uparrow \text{Opt}} \\ \text{use Eq. (9.34)} : &= f \circ p \circ (\text{liftOpt}_H(f \circ \text{pu}_{\text{Opt}}))^{\uparrow \text{Opt}} \\ \text{law (9.50) of } H : &= f \circ p \circ (f^{\downarrow H})^{\uparrow \text{Opt}} = p \triangleright f^{\downarrow F} . \end{aligned}$$

To verify the composition law (9.51), we apply its right-hand side to an arbitrary $p^{C \rightarrow \mathbb{1} + H^C}$ and

transform the result until it is equal to $p \triangleright \text{liftOpt}_F(f \diamond_{\text{Opt}} g)$:

$$\begin{aligned}
 \triangleright\text{-notation} : \quad & p \triangleright \text{liftOpt}_F(g) \circ \text{liftOpt}_F(f) = (p \triangleright \text{liftOpt}_F(g)) \triangleright \text{liftOpt}_F(f) \\
 \text{definition of } \text{liftOpt}_F(f) : \quad & = f \circ \text{flmOpt}(p \triangleright \text{liftOpt}_F(g)) \circ (\text{liftOpt}_H(f))^{\text{Opt}} \\
 \text{definition of } \text{liftOpt}_F(g) : \quad & = f \circ \text{flmOpt}(g \circ \text{flmOpt}(p) \circ (\text{liftOpt}_H(g))^{\text{Opt}}) \circ (\text{liftOpt}_H(f))^{\text{Opt}} \\
 \text{Eqs. (9.40) and (9.47)} : \quad & f \circ \text{flmOpt}(g) \circ \text{flmOpt}(p) \circ (\text{liftOpt}_H(g) \circ \text{liftOpt}_H(f))^{\text{Opt}} \\
 \text{definition (9.32) of } f \diamond_{\text{Opt}} g : \quad & = (f \diamond_{\text{Opt}} g) \circ \text{flmOpt}(p) \circ (\text{liftOpt}_H(f \diamond_{\text{Opt}} g))^{\text{Opt}} \\
 \text{definition of } \text{liftOpt}_F : \quad & = p \triangleright \text{liftOpt}_F(f \diamond_{\text{Opt}} g) \quad .
 \end{aligned}$$

Recursive types To define a contrafunctor via type recursion, we need to use a profunctor $S^{A,R}$ (see Section 6.4.1) that is contravariant in A and covariant in R .

Statement 9.2.6.5 If $S^{A,R}$ is contravariant in A and covariant in R , and additionally $S^{\bullet,R}$ is filterable (with the type parameter R fixed), then the recursive contrafunctor $F^A \triangleq S^{A,F^A}$ is filterable.

Proof The recursive contrafunctor F^{\bullet} is implemented by wrapping S in a case class:

```
type S[A, R] = ...
final case class F[A](s: S[A, F[A]])
```

The code of the function liftOpt for F^{\bullet} is recursive and uses the `xmap` method of the profunctor S .

```
def liftOpt_F[A, B](f: A => Option[B]): F[B] => F[A] = { case F(sfb) => F(
  liftOpt_S(f)(sfb).xmap_S(identity)(liftOpt_F(f))
)}
```

$$\begin{array}{ccc}
 S^{B,F^B} & \xrightarrow{\text{liftOpt}_S(f:A \rightarrow \overline{1+B})} & S^{A,F^B} \\
 & \searrow \text{liftOpt}_F(f) \triangleq & \downarrow (\text{liftOpt}_F(f))^{\text{S}^A, \bullet} \\
 & & S^{A,F^A}
 \end{array}$$

Note that $F^B \cong S^{B,F^B}$. As before, we use an overline to mark recursive calls to the same function:

$$\begin{aligned}
 \text{liftOpt}_F(f:A \rightarrow \overline{1+B}) & \triangleq \text{liftOpt}_S(f) \circ (\overline{\text{liftOpt}_F(f)})^{\text{S}^A, \bullet} \\
 & = \text{liftOpt}_S(f) \circ \text{xmap}_S(\text{id})(\overline{\text{liftOpt}_F(f)}) \quad .
 \end{aligned}$$

To verify the laws, we need the code for lifting to the contrafunctor F :

```
def cmap_F[A, B](f: A => B): F[B] => F[A] = { case F(sfb) => F(sfb.xmap_S(f)(cmap_F(f))) }
```

$$f \downarrow F \triangleq \text{xmap}_S(f)(\overline{f \downarrow F}) = f \downarrow S^{\bullet,F^B} \circ (\overline{f \downarrow F})^{\text{S}^A, \bullet} \quad .$$

The naturality-identity law (9.50) is verified by

$$\begin{aligned}
 \text{expect to equal } f \downarrow F : \quad & \text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}}) = \text{liftOpt}_S(f \circ \text{pu}_{\text{Opt}}) \circ (\overline{\text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}})})^{\text{S}^A, \bullet} \\
 \text{law (9.50) for } S^{\bullet,B} : \quad & = f \downarrow S^{\bullet,B} \circ (\overline{\text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}})})^{\text{S}^A, \bullet} \\
 \text{inductive assumption} : \quad & = f \downarrow S^{\bullet,B} \circ (\overline{f \downarrow F})^{\text{S}^A, \bullet} = f \downarrow F \quad .
 \end{aligned}$$

To verify the composition law (9.51):

$$\begin{aligned}
 \text{expect } \text{liftOpt}_F(f \diamond_{\text{Opt}} g) : \quad & \text{liftOpt}_F(g \circ B \rightarrow \overline{1+C}) \circ \text{liftOpt}_F(f:A \rightarrow \overline{1+B}) \\
 \text{definition of } \text{liftOpt}_F : \quad & = \text{liftOpt}_S(g) \circ (\overline{\text{liftOpt}_F(g)})^{\text{S}^B, \bullet} \circ \text{liftOpt}_S(f) \circ (\overline{\text{liftOpt}_F(f)})^{\text{S}^A, \bullet} \\
 \text{law (9.60) of } \text{liftOpt}_S : \quad & = \text{liftOpt}_S(g) \circ \text{liftOpt}_S(f) \circ (\overline{\text{liftOpt}_F(g)})^{\text{S}^A, \bullet} \circ (\overline{\text{liftOpt}_F(f)})^{\text{S}^A, \bullet} \\
 \text{law (9.51) of } \text{liftOpt}_S : \quad & = \text{liftOpt}_S(f \diamond_{\text{Opt}} g) \circ (\overline{\text{liftOpt}_F(g)} \circ \overline{\text{liftOpt}_F(f)})^{\text{S}^A, \bullet} \\
 \text{inductive assumption} : \quad & = \text{liftOpt}_S(f \diamond_{\text{Opt}} g) \circ (\overline{\text{liftOpt}_F(f \diamond_{\text{Opt}} g)})^{\text{S}^A, \bullet} = \text{liftOpt}_F(f \diamond_{\text{Opt}} g) \quad .
 \end{aligned}$$

In this derivation, we have used the naturality law of liftOpt_S with respect to lifting in the other type parameter of $S^{A,R}$:

$$\begin{array}{ccc}
 S^{B,R} & \xrightarrow{\text{liftOpt}_S(f: A \rightarrow \mathbb{1} + B)} & S^{A,R} \\
 \downarrow (h: R \rightarrow R') \uparrow S^{B,\bullet} & & \downarrow h \uparrow S^{A,\bullet} \\
 S^{B,R'} & \xrightarrow{\text{liftOpt}_S(f)} & S^{A,R'}
 \end{array} \quad \text{liftOpt}_S(f: A \rightarrow \mathbb{1} + B) ; (h: R \rightarrow R') \uparrow S^{A,\bullet} = h \uparrow S^{B,\bullet} ; \text{liftOpt}_S(f) \quad . \quad (9.60)$$

We expect this naturality law to hold for fully parametric functions, as discussed in Section 6.2.2.

9.3 Summary

What can we do with the techniques of this chapter?

- Use functor blocks to manipulate data wrapped in filterable functors.
- Decide whether a given filtering behavior satisfies the laws of `filter`.
- Decide whether a given type constructor (functor or contrafunctor) is filterable; if so, implement a `filter` or `liftOpt` function that satisfies the appropriate laws.
- Use constructions to derive the code of `filter` without trial and error.

What cannot be done with these techniques?

- Given a custom type constructor, generate the code for `filter` or `liftOpt` automatically.

Most non-trivial type constructors have many lawful but inequivalent implementations of `filter`. It is not possible to choose a “preferred” implementation automatically, since different applications may need different filtering behavior. While in most cases the standard library provides sufficient implementations of filtering (e.g., the `filter` or `takeWhile` methods on sequences), in some situations the programmer will need to provide a custom implementation of `filter` for a custom data type. The programmer must examine the given business requirements and decide whether they can be implemented as a lawful `filter` function.

9.3.1 Solved examples

Example 9.3.1.1 Show that the functor $F^A \triangleq G^A \rightarrow A$ is not filterable (for any contrafunctor G^A).

Solution Try to implement `deflate` : $F^{\mathbb{1}+A} \rightarrow F^A$, writing out its full type signature:

$$\text{deflate}_F : (G^{\mathbb{1}+A} \rightarrow \mathbb{1} + A) \rightarrow G^A \rightarrow A \quad , \quad \text{deflate}_F = p: G^{\mathbb{1}+A} \rightarrow \mathbb{1} + A \rightarrow g: G^A \rightarrow ???^A \quad .$$

We cannot extract a value of type A from $g: G^A$ since the contrafunctor G^A does not wrap any values of A . So, the only hope of filling the typed hole $???^A$ is to apply the function p to an argument of type $G^{\mathbb{1}+A}$. Even if we are able to map $G^A \rightarrow G^{\mathbb{1}+A}$ (e.g., if G^\bullet is filterable), the result of applying p will be a value of type $\mathbb{1} + A$. We cannot compute a value of type A out of that. So, the type signature of `deflate` for F^\bullet is not implementable. We conclude that F^\bullet is not filterable.

Example 9.3.1.2 Use known filterable constructions to show that

$$F^A \triangleq \text{Int} \times \text{String} \rightarrow \mathbb{1} + \text{Int} \times A + A \times (\mathbb{1} + A) + (\text{Int} \rightarrow \mathbb{1} + A + A \times A \times \text{String})$$

is a filterable functor. (Using the constructions avoids the need for proofs.)

Solution We need to analyze the structure of the functor F^\bullet to decide which constructions we may use. Define some auxiliary functors that represents sub-expressions in F^A ,

$$\begin{aligned} R_1^A &\triangleq \text{Int} \times \text{String} \rightarrow A \quad , \quad R_2^A \triangleq \text{Int} \rightarrow A \quad , \\ G^A &\triangleq \mathbb{1} + \text{Int} \times A + A \times (\mathbb{1} + A) \quad , \quad H^A \triangleq \mathbb{1} + A + A \times A \times \text{String} \quad . \end{aligned}$$

Now we can rewrite the type F^A as

$$F^A = R_1^{L^A} \quad , \quad L^A \triangleq G^A + R_2^{H^A} \quad .$$

The type of G^\bullet is a co-product, so we need to check which of the two co-product constructions (Statements 9.2.4.3 or 9.2.4.4) might apply. The first of them does not apply because the functors $\text{Int} \times A$ and $A \times (\mathbb{1} + A)$ are not filterable. But the second construction applies if we write G^A in the form $G^A = \mathbb{1} + A \times K^A$ where $K^A \triangleq \mathbb{1} + \text{Int} + A$. Since K^A is the co-product of the `Option` functor and a constant functor (the fixed type `Int`), K^\bullet is filterable by Statement 9.2.4.3. So, G^A is filterable.

Similarly, we find that H^\bullet is filterable by Statement 9.2.4.4 if we write $H^A = \mathbb{1} + A \times (\mathbb{1} + A \times \text{String})$, where the functor $\mathbb{1} + A \times \text{String}$ is filterable by the same construction.

The functor $R_2^{H^\bullet}$ is filterable since it is a functor composition (Statement 9.2.4.1) and H^\bullet is filterable. The co-product $L^A \triangleq G^A + R_2^{H^A}$ is filterable by Statement 9.2.4.3, and $R_1^{L^A}$ by Statement 9.2.4.1.

Each construction gives a specific code for the corresponding `liftOpt` function, and so we could derive the code for liftOpt_F that is guaranteed to obey the filter laws. However, keep in mind that there are several inequivalent ways of implementing a lawful `liftOpt` for this functor. For instance, the filtering operation for H^\bullet could be defined similarly to that for `JillsCoupons` in Example 9.1.4.2 and not through a co-product construction. The constructions give one possibility out of many. The programmer needs to choose the implementation according to the business requirements at hand.

Example 9.3.1.3 (identity law of deflate) The function `deflate` : $F^{\mathbb{1}+A} \rightarrow F^A$ is available only for filterable functors; but the function with the inverse type signature, `inflate` : $F^A \rightarrow F^{\mathbb{1}+A}$, can be implemented for any functor F . (See definition of `inflate` in Section 9.2.1.) Assuming that a given functor F is filterable, show that `deflate` is a **left inverse** of `inflate`:

$$\text{inflate}_F \circ \text{deflate}_F = \text{id} \quad . \quad (9.61)$$

Also show that it is not a right inverse: $\text{deflate}_F \circ \text{inflate}_F \neq \text{id}$ for some functors F .

Solution We may assume that F satisfies the filtering laws. The function `inflate` can be equivalently written as

$$\text{inflate}_F = (x:A \rightarrow \mathbb{0} + x)^{\uparrow F} = \text{pu}_{\text{Opt}}^F = (\psi_{(_ \rightarrow \text{true})})^{\uparrow F} \quad .$$

Now we can use the identity law (9.4) to derive

$$\text{filt}_F(_ \rightarrow \text{true}) = \psi_{(_ \rightarrow \text{true})}^{\uparrow F} \circ \text{deflate}_F = \text{inflate}_F \circ \text{deflate}_F \quad .$$

Since the identity law says $\text{filt}_F(_ \rightarrow \text{true}) = \text{id}$, we obtain $\text{inflate}_F \circ \text{deflate}_F = \text{id}$.

To show that the inverse equation does not always hold, we need to find an explicit example: a specific functor F and a value $x:F^{\mathbb{1}+A}$ such that $x \triangleright \text{deflate}_F \circ \text{inflate}_F \neq x$. Looking at the simplest filterable functors, we find that the constant functor $F^A \triangleq \mathbb{Z}$ is not a suitable example because all its methods are identity functions. The next nontrivial example is the `Option` functor, $F^A \triangleq \mathbb{1} + A = \text{Opt}^A$. With the equivalence $\text{Opt}^{\text{Opt}^A} \cong \mathbb{1} + \mathbb{1} + A$, we write the methods `inflate` and `deflate` as

```
def inflate[A]: Option[A] => Option[Option[A]] = _.map(x => Some(x))
def deflate[A]: Option[Option[A]] => Option[A] = _.flatten
```

$$\text{inflate}_{\text{Opt}} \triangleq \text{pu}_{\text{Opt}}^{\uparrow \text{Opt}} = \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & 1 & 1 & A \\ \hline 1 & id & 0 & 0 \\ \hline A & 0 & 0 & id \\ \hline \end{array} \end{vmatrix}, \quad \text{deflate}_{\text{Opt}} \triangleq \text{ftn}_{\text{Opt}} = \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & 1 & A \\ \hline 1 & id & 0 \\ \hline 1 & id & 0 \\ \hline A & 0 & id \\ \hline \end{array} \end{vmatrix}.$$

The composition $\text{deflate}_{\text{Opt}} \circ \text{inflate}_{\text{Opt}}$ is computed as

$$\text{deflate}_{\text{Opt}} \circ \text{inflate}_{\text{Opt}} = \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & 1 & A \\ \hline 1 & id & 0 \\ \hline 1 & id & 0 \\ \hline A & 0 & id \\ \hline \end{array} \end{vmatrix} \circ \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & 1 & 1 & A \\ \hline 1 & id & 0 & 0 \\ \hline A & 0 & 0 & id \\ \hline \end{array} \end{vmatrix}$$

matrix composition : $= \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & 1 & 1 & A \\ \hline 1 & id & 0 & 0 \\ \hline 1 & id & 0 & 0 \\ \hline A & 0 & 0 & id \\ \hline \end{array} \end{vmatrix} \neq \text{id} = \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & 1 & 1 & A \\ \hline 1 & id & 0 & 0 \\ \hline 1 & 0 & id & 0 \\ \hline A & 0 & 0 & id \\ \hline \end{array} \end{vmatrix}.$

The result differs from an identity matrix in the second row. A value $x^{1+1+A} \triangleq 0^1 + 1 + 0^A$ will give a non-void result in the second row of the matrix, showing the difference:

$$(0 + 1 + 0) \triangleright \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & 1 & 1 & A \\ \hline 1 & id & 0 & 0 \\ \hline 1 & id & 0 & 0 \\ \hline A & 0 & 0 & id \\ \hline \end{array} \end{vmatrix} = \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & id & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline id & 0 & 0 \\ \hline 0 & 0 & id \\ \hline \end{array} \end{vmatrix} \triangleright \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & 1 & 0 & 0 \\ \hline 1 & 0 & 0 \\ \hline 0 & 0 & id \\ \hline \end{array} \end{vmatrix} = \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & 1 & 0 & 0 \\ \hline 1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline \end{array} \end{vmatrix} \neq \begin{vmatrix} & \begin{array}{|c|c|c|} \hline & 1 & 0 & 0 \\ \hline 1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline \end{array} \end{vmatrix}.$$

In Scala code, this value x is `Some(None)`, so the calculation corresponds to

```
inflate(deflate(Some(None))) == None // Would have been Some(None) if the function were an identity.
```

Example 9.3.1.4 Assume that a given functor $H^A \triangleq 1 + K^A$ is filterable (but K^{\bullet} is not necessarily filterable). The functor H^{\bullet} is a “data wrapper” with a fixed empty value, $1 + 0^{K^A}$. Show that an empty wrapper remains empty after any filtering: the function filt_H satisfies, for any $p^{A \rightarrow 2}$,

$$(1 + 0^{K^A}) \triangleright \text{filt}_H(p) = 1 + 0^{K^A} . \quad (9.62)$$

Solution We know nothing about H^{\bullet} and K^{\bullet} other than the fact that filt_H obeys the filtering laws. Rewrite Eq. (9.62) via the simpler function deflate_H , which is computationally equivalent to filt_H :

$$(1 + 0^{K^A}) \triangleright \psi_p^{\uparrow H} \circ \text{deflate}_H \stackrel{?}{=} 1 + 0^{K^A} .$$

Any function lifted to H^{\bullet} works separately for the two parts of the disjunctive type $H^A = 1 + K^A$. So

$$(1 + 0^{K^A}) \triangleright f^{\uparrow H} = 1 + 0^{K^B} ,$$

regardless of the choice of $f^{A \rightarrow B}$. Setting $f \triangleq \psi_p$, we obtain

$$(1 + 0^{K^A}) \triangleright \psi_p^{\uparrow H} = 1 + 0^{K^{1+A}} . \quad (9.63)$$

It remains to show that

$$(1 + 0^{K^{1+A}}) \triangleright \text{deflate}_H \stackrel{?}{=} 1 + 0^{K^A} . \quad (9.64)$$

We can proceed only if we use some law of `deflate`; Example 9.3.1.3 shows a suitable law. Since Eq. (9.63) holds for all predicates p , we can choose the predicate p to be identically `true` and get

$$(1 + \mathbb{0}^{:K^A}) \triangleright \psi_{(_) \rightarrow \text{true}}^{\uparrow H} = (1 + \mathbb{0}^{:K^A}) \triangleright \text{inflate}_H = 1 + \mathbb{0}^{:K^{1+A}} . \quad (9.65)$$

Substituting this into Eq. (9.64), we find

$$\begin{aligned} \text{expect to equal } 1 + \mathbb{0}^{:K^A} : & \quad (1 + \mathbb{0}^{:K^{1+A}}) \triangleright \text{deflate}_H \\ \text{use Eq. (9.65)} : & \quad = (1 + \mathbb{0}^{:K^A}) \triangleright \underline{\text{inflate}_H \circ \text{deflate}_H} \\ \text{identity law (9.61) of } \text{deflate}_H : & \quad = (1 + \mathbb{0}^{:K^A}) \triangleright \underline{\text{id}} = 1 + \mathbb{0}^{:K^A} . \end{aligned}$$

This completes the proof of Eq. (9.64) and of the property (9.62). The same property can be expressed in terms of liftOpt_H as

$$(1 + \mathbb{0}^{:K^B}) \triangleright \text{liftOpt}_H(f^{:A \rightarrow 1+B}) = 1 + \mathbb{0}^{:K^B} . \quad (9.66)$$

Example 9.3.1.5 Assuming that G^\bullet and H^\bullet are filterable functors and H^\bullet is of the form $H^A \triangleq \mathbb{1} + K^A$ (where K^\bullet is not necessarily filterable), prove that the functor $F^\bullet \triangleq G^{K^\bullet}$ is filterable.

Solution We need to define liftOpt_F and verify its laws, assuming that liftOpt_G and liftOpt_H ,

$$\text{liftOpt}_G(f^{:A \rightarrow 1+B}) : G^A \rightarrow G^B , \quad \text{liftOpt}_H(f^{:A \rightarrow 1+B}) : \mathbb{1} + K^A \rightarrow \mathbb{1} + K^B ,$$

are already available and obey the same laws. We need to implement the type signature

$$\text{liftOpt}_F(f^{:A \rightarrow 1+B}) : G^{K^A} \rightarrow G^{K^B} .$$

We can map G^{K^A} to G^{K^B} using $\text{liftOpt}_G^{K^A, K^B}(k)$ if we supply k of type $K^A \rightarrow \mathbb{1} + K^B$ and use the type parameters K^A, K^B instead of A, B . We can compute the function k as $\text{liftOpt}_H(f^{:A \rightarrow 1+B})$ if we extend the argument to $\mathbb{1} + K^A$ instead of K^A using the function pu_{Opt} with type parameter K^A :

$$\text{pu}_{\text{Opt}}^{K^A} : K^A \rightarrow \mathbb{1} + K^A , \quad \text{pu}_{\text{Opt}}^{K^A} \circ \text{liftOpt}_H^{K^B, K^B}(f^{:A \rightarrow 1+B}) : K^A \rightarrow \mathbb{1} + K^B .$$

Now we are ready to write the code for liftOpt_F as

$$\text{liftOpt}_F(f^{:A \rightarrow 1+B}) \triangleq \text{liftOpt}_G^{K^A, K^B}(\text{pu}_{\text{Opt}}^{K^A} \circ \text{liftOpt}_H(f)) .$$

It remains to verify the laws. The naturality-identity law (9.29) for liftOpt_F :

$$\begin{aligned} \text{expect to equal } f^{\uparrow F} : & \quad \text{liftOpt}_F(f \circ \text{pu}_{\text{Opt}}) = \text{liftOpt}_G^{K^A, K^B}(\text{pu}_{\text{Opt}}^{K^A} \circ \underline{\text{liftOpt}_H(f \circ \text{pu}_{\text{Opt}})}) \\ \text{law (9.29) of } \text{liftOpt}_H : & \quad = \text{liftOpt}_G^{K^A, K^B}(\text{pu}_{\text{Opt}}^{K^A} \circ \underline{f^{\uparrow H}}) \\ \text{lifting } \uparrow^H \text{ expressed via } \uparrow^K : & \quad = \text{liftOpt}_G^{K^A, K^B}(\text{pu}_{\text{Opt}}^{K^A} \circ \underline{f^{\uparrow K} \uparrow^{\text{Opt}}}) \\ \text{naturality (8.8) of } \text{pu}_{\text{Opt}} : & \quad = \text{liftOpt}_G^{K^A, K^B}(f^{\uparrow K} \circ \text{pu}_{\text{Opt}}^{K^B}) \\ \text{law (9.29) of } \text{liftOpt}_G : & \quad = \underline{(f^{\uparrow K})^{\uparrow G}} = f^{\uparrow F} . \end{aligned}$$

To verify the composition law (9.31) for liftOpt_F :

$$\begin{aligned} \text{expect to equal } \text{liftOpt}_F(f \diamond_{\text{Opt}} g) : & \quad \text{liftOpt}_F(f) \circ \text{liftOpt}_F(g) \\ \text{definition of } \text{liftOpt}_F : & \quad = \underline{\text{liftOpt}_G(\text{pu}_{\text{Opt}} \circ \text{liftOpt}_H(f))} \circ \underline{\text{liftOpt}_G(\text{pu}_{\text{Opt}} \circ \text{liftOpt}_H(g))} \\ \text{law (9.31) for } \text{liftOpt}_G : & \quad = \text{liftOpt}_G((\text{pu}_{\text{Opt}} \circ \text{liftOpt}_H(f)) \diamond_{\text{Opt}} (\text{pu}_{\text{Opt}} \circ \text{liftOpt}_H(g))) \\ \text{simplify using Eq. (9.67)} : & \quad = \text{liftOpt}_G(\text{pu}_{\text{Opt}} \circ \underline{\text{liftOpt}_H(f) \circ \text{liftOpt}_H(g)}) \\ \text{law (9.31) for } \text{liftOpt}_H : & \quad = \text{liftOpt}_G(\text{pu}_{\text{Opt}} \circ \text{liftOpt}_H(f \diamond_{\text{Opt}} g)) = \text{liftOpt}_F(f \diamond_{\text{Opt}} g) . \end{aligned}$$

In this derivation, we used a property that simplifies the Kleisli composition with pu_{Opt} :

$$f^{:K^A \rightarrow \mathbb{1} + K^B} \diamond_{\text{Opt}} (\text{pu}_{\text{Opt}}^{K^B} ; \text{liftOpt}_H^{\mathbb{1} + K^B, \mathbb{1} + K^C}(g^{:B \rightarrow \mathbb{1} + C})) = f ; \text{liftOpt}_H(g) . \quad (9.67)$$

This simplification depends on the property (9.66) of liftOpt_H shown in Example 9.3.1.4 and does not work for arbitrary functions p, q having the same type signatures as in Eq. (9.67),

$$p^{:A \rightarrow \mathbb{1} + B} \diamond_{\text{Opt}} (\text{pu}_{\text{Opt}}^{B \rightarrow \mathbb{1} + B} ; q^{:\mathbb{1} + B \rightarrow \mathbb{1} + C}) \neq p ; q .$$

To prove Eq. (9.67), use the code for \diamond_{Opt} and pu_{Opt} from Eq. (9.33):

$$\begin{aligned} \text{left-hand side of Eq. (9.67)} : & f \diamond_{\text{Opt}} (\text{pu}_{\text{Opt}}^{K^B} ; \text{liftOpt}_H(g)) \\ \text{definition (9.32) of } \diamond_{\text{Opt}} : & = f ; \text{flm}_{\text{Opt}}(\text{pu}_{\text{Opt}}^{K^B} ; \text{liftOpt}_H(g)) \\ \text{definition of } \text{pu}_{\text{Opt}} : & = f ; \text{flm}_{\text{Opt}}(x^{:K^B} \rightarrow (\mathbb{0} + x) \triangleright \text{liftOpt}_H(g)) \\ \text{use Eq. (9.33)} : & = f^{:K^A \rightarrow \mathbb{1} + K^B} ; \begin{array}{c|c} & \mathbb{1} + K^C \\ \hline \mathbb{1} & 1 \rightarrow 1 + \mathbb{0}^{:K^C} \\ \hline K^B & x \rightarrow (\mathbb{0} + x) \triangleright \text{liftOpt}_H(g) \end{array} . \end{aligned}$$

We expect the last expression to equal the right-hand side of Eq. (9.67), which is $f ; \text{liftOpt}_H(g)$. To be able to compare these expressions, we rewrite $\text{liftOpt}_H(g)$ equivalently as a matrix:

$$\text{liftOpt}_H(g^{:B \rightarrow \mathbb{1} + C}) = \begin{array}{c|c} & \mathbb{1} + K^C \\ \hline \mathbb{1} & 1 \rightarrow (1 + \mathbb{0}^{:K^C}) \triangleright \text{liftOpt}_H(g) \\ \hline K^B & x \rightarrow (\mathbb{0} + x) \triangleright \text{liftOpt}_H(g) \end{array} .$$

Usually, such an equivalent rewriting gives no advantages; but in this case, Eq. (9.66) simplifies the first row of the matrix to $1 \rightarrow 1 + \mathbb{0}^{:K^C}$. This makes both sides of Eq. (9.66) equal.

Example 9.3.1.6 Prove that the recursive functor $F^A \triangleq \mathbb{1} + A \times A + A \times (\mathbb{1} + A) \times F^A$ is filterable.

Solution Rather than proving the laws by hand (as we did in a similar case in Statement 9.2.4.6), we will use a trick that will make calculations shorter. The trick is to “unroll” the recursive equation and to reduce F^A to the `List` functor, which has a standard filtering operation.

The **unrolling trick** works for any recursive definition of the form $F^A \triangleq P^A + Q^A \times F^A$:

$$\text{if } F^A \triangleq P^A + Q^A \times F^A \text{ then } F^A \cong P^A \times \text{List}^{Q^A} , \quad (9.68)$$

where P^\bullet and Q^\bullet are arbitrary functors. The functor F^A given in this example is of this form with the functors $P^A \triangleq \mathbb{1} + A \times A$ and $Q^A \triangleq A \times (\mathbb{1} + A)$. Comparing with the definition of the `List` functor,

$$\begin{aligned} \text{List}^A &\triangleq \mathbb{1} + A \times \text{List}^A \cong \mathbb{1} + A \times (\mathbb{1} + A \times (\mathbb{1} + \dots (\mathbb{1} + A \times \text{List}^A))) \\ &\cong \mathbb{1} + A + A \times A + A \times A \times A + \dots + \underbrace{A \times \dots \times A}_{n \text{ times}} \times \text{List}^A , \end{aligned}$$

we find that the definition (9.68) of F^A can be “unrolled” n times as

$$\begin{aligned} F^A &\triangleq P^A + Q^A \times F^A \cong P^A + Q^A \times (P^A + Q^A \times (P^A + \dots (P^A + Q^A \times F^A))) \\ &\cong P^A + Q^A \times P^A + Q^A \times Q^A \times P^A + \dots + \underbrace{Q^A \times \dots \times Q^A}_{n \text{ times}} \times F^A \\ &\cong P^A \times (\mathbb{1} + Q^A + Q^A \times Q^A + \dots \underbrace{Q^A \times \dots \times Q^A}_{n-1 \text{ times}}) + \underbrace{Q^A \times \dots \times Q^A}_{n \text{ times}} \times F^A . \end{aligned}$$

The type equivalence (9.68), $F^A \cong P^A \times \text{List}^{Q^A}$, follows by induction on the number of “unrolled” functors F . We can now use the functor product construction for F^\bullet if we show that List^{Q^\bullet} is filterable. This does not follow by functor composition because $Q^A \triangleq A \times (\mathbb{1} + A)$ is *not* filterable (we saw that in Section 9.1.3). However, we can derive from Statement 9.2.4.4, setting $G^A \triangleq \mathbb{1} + A$, that

$$H^A \triangleq \mathbb{1} + A \times (\mathbb{1} + A) = \mathbb{1} + Q^A$$

is filterable. So, we can use the result of Example 9.3.1.5 and conclude that List^{Q^\bullet} is filterable.

Example 9.3.1.7 Show that the contrafunctor $C^A \triangleq A \rightarrow Z$ is not filterable (where the fixed type Z does not have a known default value).

Solution Try to implement the function $\text{inflate}_C : C^A \rightarrow C^{\mathbb{1}+A}$,

$$\text{inflate}_C : (A \rightarrow Z) \rightarrow \mathbb{1} + A \rightarrow Z \quad , \quad \text{inflate}_C = c^{A \rightarrow Z} \rightarrow p^{\mathbb{1}+A} \rightarrow ???^Z \quad .$$

The only way to fill the type hole $???^Z$ is to apply c to an argument of type A . However, we do not have values of type A ; we only have p of type $\mathbb{1} + A$, and p might be the “empty” value, $1 + 0$. So, it is impossible to implement inflate_C , and we conclude that C is not filterable.

Example 9.3.1.8 Given a filterable functor F^\bullet , show that the type $F^\mathbb{1} \rightarrow F^0$ is not void.

Solution A filterable functor must have a `deflate` function with type signature $F^{\mathbb{1}+A} \rightarrow F^A$. Set the type parameter $A = 0$ in the code of `deflate` to obtain the code for a function of type $F^\mathbb{1} \rightarrow F^0$.

9.3.2 Exercises

Exercise 9.3.2.1 Implement a `Filterable` instance for the functor $F[T] = G[H[T]]$ assuming that the contrafunctor $H[_]$ already has a `Filterable` instance and $G[_]$ is an arbitrary contrafunctor. Verify the laws of filterable functor rigorously (i.e. by calculations, not tests).

Exercise 9.3.2.2 Implement a `Filterable` instance for $F[T] = \text{Option[Int} \Rightarrow \text{Option}[(T, T)]]$. Show that the laws hold by using known constructions (avoiding explicit proofs or tests).

Exercise 9.3.2.3 Prove rigorously (not by tests) that $\text{flm}_{\text{Opt}}(\text{pu}_{\text{Opt}}) = \text{id}^{\mathbb{1}+A \rightarrow \mathbb{1}+A}$.

Exercise 9.3.2.4 Show that one can define $\text{deflate} : C^{\mathbb{1}+A} \rightarrow C^A$ for any contrafunctor C^A (not necessarily filterable). Prove that *in case* C^\bullet is filterable, the **identity law** will hold:

$$\text{inflate}_C ; \text{deflate}_C = \text{id}^{C^A \rightarrow C^A} \quad .$$

Show that the inverse equation does not hold in general: $\text{deflate}_C ; \text{inflate}_C \neq \text{id}$.

Exercise 9.3.2.5 Assuming that G^\bullet is a filterable functor, prove rigorously that the recursive functor $F^A \triangleq G^A + \text{Int} \times A \times A \times A \times F^A$ is filterable. Implement a `Filterable` instance for F^\bullet .

Exercise 9.3.2.6 Show that the functor $F^A \triangleq A + (\text{Int} \rightarrow A)$ is not filterable.

Exercise 9.3.2.7 Prove that $F^A \triangleq \mathbb{1} + A \times G^A$ is in general not filterable if G^A is an arbitrary (non-filterable) functor; give an example of a suitable G^A . Since $F^\mathbb{1} \rightarrow F^0 \cong \mathbb{1} + G^\mathbb{1} \rightarrow \mathbb{1} \cong \mathbb{1}$, this will demonstrate that Example 9.3.1.8 gives a necessary but not a sufficient condition for a functor F^\bullet to be filterable.

Exercise 9.3.2.8 Show that $F^A \triangleq \mathbb{1} + G^A + H^A$ is filterable if $\mathbb{1} + G^A$ and $\mathbb{1} + H^A$ are filterable (even when G^\bullet and H^\bullet are not filterable).

Exercise 9.3.2.9 Show that $C^A \triangleq A + A \times A \rightarrow \mathbb{1} + Z$ is a filterable contrafunctor (no law checking).

Exercise 9.3.2.10 Verify Eqs. (9.46)–(9.47) in Scala syntax and in the code notation.

Exercise 9.3.2.11 If G^A and H^A are filterable contrafunctors, prove that the contrafunctors $G^A \times H^A$ and $G^A + H^A$ are also filterable.

Exercise 9.3.2.12 Show that the contrafunctor $C^A \triangleq A \times F^A \rightarrow Z$ is not filterable for any functor F^\bullet and any fixed type Z that does not have a known default value.

Exercise 9.3.2.13 Show that a *necessary* condition for a contrafunctor C^\bullet to be filterable is that a function of type $C^0 \rightarrow C^1$ can be implemented (i.e., the type $C^0 \rightarrow C^1$ is not void).

Exercise 9.3.2.14* Show that a *polynomial* functor F^\bullet is filterable (in some way) if and only if the type $F^1 \rightarrow F^0$ is not void. Find an example of a non-filterable polynomial functor F^\bullet that violates the condition $F^1 \rightarrow F^0 \not\equiv \emptyset$.

9.4 Further developments

9.4.1 Naturality laws and natural transformations

While deriving various laws, we often need to interchange the order of compositions that involve lifted functions. For instance, in the derivation of Example 9.2.5.4, we needed to move $\psi_p^{\downarrow C}$ to the left of filt_C in the expression $\text{get}^{\downarrow C} ; \text{filt}_C(\text{nonEmpty}) ; \psi_p^{\downarrow C}$, or else we could make no progress with the calculations. The required interchange was possible by using the law (9.58),

$$\text{filt}_C(p) ; f^{\downarrow C} = f^{\downarrow C} ; \text{filt}_C(f ; p) \quad .$$

In this and previous chapters, we discovered a number of laws of that form, e.g.:

Eq. (8.8) :	$\text{pu}_F ; f^{\uparrow F} = f ; \text{pu}_F$	for $\text{pu}_F : A \rightarrow F^A$
Eq. (8.11) :	$\text{ex}_F ; f = f^{\uparrow F} ; \text{ex}_F$	for $\text{ex}_F : F^A \rightarrow A$
Eq. (8.13) :	$\text{ex}_S ; f = \text{bimap}_S(f)(f) ; \text{ex}_S$	for $\text{ex}_S : S^{A,A} \rightarrow A$
Eq. (9.3) :	$f^{\uparrow F} ; \text{filt}_F(p) = \text{filt}_F(f ; p) ; f^{\uparrow F}$	for $\text{filt}_F : (A \rightarrow 2) \rightarrow F^A \rightarrow F^A$
Eq. (9.14) :	$\text{deflate} ; f^{\uparrow F} = f^{\text{Opt}^{\uparrow F}} ; \text{deflate}$	for $\text{deflate}_F : F^{1+A} \rightarrow F^A$
Eq. (9.18) :	$f ; \psi_p = \psi_{f ; p} ; f^{\text{Opt}}$	for $\psi : (A \rightarrow 2) \rightarrow A \rightarrow 1 + A$
Eq. (9.23) :	$h^{\uparrow F} ; \text{liftOpt}_F(f) = \text{liftOpt}_F(h ; f)$	for $\text{liftOpt}_F : (A \rightarrow \text{Opt}^B) \rightarrow F^A \rightarrow F^B$
Eq. (9.35) :	$\text{liftOpt}_F(f) ; h^{\uparrow F} = \text{liftOpt}_F(f ; h^{\text{Opt}})$	for $\text{liftOpt}_F : (A \rightarrow \text{Opt}^B) \rightarrow F^A \rightarrow F^B$
Eq. (9.58) :	$\text{filt}_C(p) ; f^{\downarrow C} = f^{\downarrow C} ; \text{filt}_C(f ; p)$	for $\text{filt}_C : (A \rightarrow 2) \rightarrow C^A \rightarrow C^A$
Eq. (9.59) :	$f^{\downarrow C} ; \text{inflate}_C = \text{inflate}_C ; f^{\text{Opt}^{\downarrow C}}$	for $\text{inflate}_C : C^A \rightarrow C^{1+A}$

We called all these laws “naturality laws”, although they were derived from different premises and do not look similar at first sight. Is there a common pattern for these laws? For a given function, can we guess the form of its naturality law?

Looking at the examples shown above, we find that each law follows one of the four patterns shown in the following table, where we use arbitrary functors F, G, H and contrafunctors C :

Pattern	Type signature	Naturality law
natural transformation	$t^A : F^A \rightarrow G^A$	$(f : A \rightarrow B)^{\uparrow F} ; t^B = t^A ; f^{\uparrow G}$
parameterized transformation	$t^A : C^A \rightarrow F^A \rightarrow G^A$	$f^{\uparrow F} ; t^B(c) = t(c \triangleright f^{\downarrow C}) ; f^{\uparrow G}$
generalized lifting; A -naturality	$t^{A,B} : (A \rightarrow G^B) \rightarrow F^A \rightarrow F^B$	$(f : A \rightarrow B)^{\uparrow F} ; t^{B,C}(p) = t^{A,C}(f ; p)$
generalized lifting; B -naturality	$t^{A,B} : (A \rightarrow G^B) \rightarrow F^A \rightarrow F^B$	$t^{A,B}(p) ; (g : B \rightarrow C)^{\uparrow F} = t^{A,C}(p ; g^{\uparrow G})$

Let us now look at each of these patterns in detail.

Natural transformations This pattern covers functions with type signatures of the form $F^A \rightarrow G^A$, where F and G must be both functors or both contrafunctors. Functions of that kind, i.e., fully parametric functions with type signatures of the form $t^A : F^A \rightarrow G^A$, are called **natural transformations** between F and G . Heuristically, we may view t^A as a function that copies data of type A from one “wrapper” to another and may rearrange that data in a way that does not depend on the type A . An example is the `headOption` method defined in Scala on various sequence types such as `List`,

$$\text{headOpt} : \text{List}^A \rightarrow \text{Opt}^A .$$

We can write an equivalent Scala code for this function as

```
def headOption[A]: List[A] => Option[A] = {
  case List()  => None
  case x :: _   => Some(x)
}
```

It is clear that this code works in the same way for all types A . This property is formulated mathematically as the requirement that we may first transform a list with a lifted function $(f:A \rightarrow B)^{\uparrow \text{List}}$ and then apply `headOption`; or we may first apply `headOption` and then transform the data with a lifted function f (and we will then need to lift f to the `Option` functor rather than to the `List` functor); the results will be equal. We write this requirement as an equation called the naturality law of `headOption`:

$$\begin{array}{ccc} \text{List}^A & \xrightarrow{\text{headOpt}^A} & \text{Opt}^A \\ \downarrow (f:A \rightarrow B)^{\uparrow \text{List}} & & \downarrow f^{\uparrow \text{Opt}} \\ \text{List}^B & \xrightarrow{\text{headOpt}^B} & \text{Opt}^B \end{array} \quad (f:A \rightarrow B)^{\uparrow \text{List}} ; \text{headOpt}^B = \text{headOpt}^A ; f^{\uparrow \text{Opt}} .$$

It is important to keep in mind that this law does not depend on the fact that `headOption` extracts the *first* element of a list. The same naturality law will hold for any fully parametric function of type $\text{List}^A \rightarrow \text{Opt}^A$, e.g., the functions `_.lastOption` or `_.drop(2).headOption`. The naturality law only expresses the property that `headOption` works in the same way for all types without examining any specific values in the list.

Other examples of natural transformations are the functions `pure`, `deflate`, and `inflate`, whose naturality laws we have already seen. All these naturality laws are captured by the “natural transformation” pattern, which we can formulate for arbitrary functors F^{\bullet} and G^{\bullet} as the following law:

$$\begin{array}{ccc} F^A & \xrightarrow{t^A} & G^A \\ \downarrow (f:A \rightarrow B)^{\uparrow F} & & \downarrow f^{\uparrow G} \\ F^B & \xrightarrow{t^B} & G^B \end{array} \quad (f:A \rightarrow B)^{\uparrow F} ; t^B = t^A ; f^{\uparrow G} . \quad (9.69)$$

Once we recognize that a given function $t^A : F^A \rightarrow G^A$ has the type signature of a natural transformation, how can we remember its naturality law? A naturality law always involves an arbitrary function $f:A \rightarrow B$ with two type parameters. Types will match only if the function f is lifted to the functor F when f is applied before t , and to the functor G when f is applied after t . So, the naturality law is $f^{\uparrow F} ; t = t ; f^{\uparrow G}$ with appropriate type parameters.

The analogous naturality law for natural transformations $t : C^A \rightarrow D^A$ between contrafunctors C , D has exactly the same form, but the order of type parameters must be swapped:

$$\begin{array}{ccc} C^A & \xrightarrow{t^A} & D^A \\ \downarrow (f:B \rightarrow A)^{\downarrow C} & & \downarrow f^{\downarrow D} \\ C^B & \xrightarrow{t^B} & D^B \end{array} \quad (f:B \rightarrow A)^{\downarrow C} ; t^B = t^A ; f^{\downarrow D} . \quad (9.70)$$

Examples of this pattern are the naturality law (8.14) for `pu_D` (where we need to set $C^A \triangleq \mathbb{1}$) and the naturality law (9.59) for `inflate_C`.

Parameterized transformations The second pattern, which may be called “parameterized transformation”, is a curried function that takes a first argument of type C^A and returns a natural transformation of type $F^A \rightarrow G^A$. This pattern is recognized by a type signature $C^A \rightarrow F^A \rightarrow G^A$ where C is a contrafunctor while F and G are functors (or vice versa, C is a functor while F and G are contrafunctors). Examples of parameterized transformations are the methods `filter`, `takeWhile`, and `find`, defined in the Scala library for sequence-like type constructors (e.g., `List`, `Array`).

For functions of this kind, naturality laws must modify the argument of type C^A when changing the order of lifted functions. In order to formulate the naturality law for a parameterized transformation $t^A : C^A \rightarrow F^A \rightarrow G^A$, use an arbitrary function $f:A \rightarrow B$ and first transform the argument of

type F^A into F^B using $f^{\uparrow F}$ before applying $t^B(c)$:

$$f^{\uparrow F} ; t^B(c : C^B) = ???^{F^A \rightarrow G^B} .$$

It is clear that we must choose an arbitrary value c of type C^B (rather than C^A) for all types to match. It remains to fill the typed hole, which must be of the form $t(\dots) ; f^{\uparrow G}$,

$$f^{\uparrow F} ; t^B(c : C^B) = t^A(???^{C^A}) ; f^{\uparrow G} .$$

The argument of t of type C^A is obtained by applying $f^{\downarrow C}$ to c . So, the law is

$$f^{\uparrow F} ; t(c) = t(c \triangleright f^{\downarrow C}) ; f^{\uparrow G} . \quad (9.71)$$

A similar law can be derived for the case when C^\bullet is a functor and F^\bullet, G^\bullet are contrafunctors.

Parameterized transformations $t : C^A \rightarrow F^A \rightarrow G^A$ are reduced to natural transformations if we swap the order of curried arguments to $F^A \rightarrow C^A \rightarrow G^A$ and define the functor $H^A \triangleq C^A \rightarrow G^A$. Then $F^A \rightarrow C^A \rightarrow G^A = F^A \rightarrow H^A$, which is a type signature of a natural transformation. Denoting by $\tilde{t} : F^A \rightarrow H^A$ the function t with its arguments swapped, we can write the naturality law of \tilde{t} as

$$f^{\uparrow F} ; \tilde{t} = \tilde{t} ; f^{\uparrow H} , \quad \text{where } \tilde{t} \triangleq p^{F^A} \rightarrow c^{C^A} \rightarrow p \triangleright t(c) .$$

To show that this law is equivalent to Eq. (9.71), we use the definition of H ,

$$f^{\uparrow H} \triangleq h^{H^A} \rightarrow c^{C^B} \rightarrow c \triangleright f^{\downarrow C} ; h ; f^{\uparrow G} .$$

Substituting this into the naturality law of \tilde{t} , we obtain the naturality law (9.71):

$$\begin{aligned} \text{left-hand side applied to } p^{F^A} : & \quad p \triangleright f^{\uparrow F} ; \tilde{t} = p \triangleright f^{\uparrow F} ; (p \rightarrow c \rightarrow p \triangleright t(c)) \\ \text{apply function to } p \triangleright f^{\uparrow F} : & \quad = c \rightarrow p \triangleright f^{\uparrow F} \triangleright t(c) = c \rightarrow p \triangleright f^{\uparrow F} ; t(c) , \\ \text{left-hand side applied to } p^{F^A} : & \quad p \triangleright (p \rightarrow c \rightarrow p \triangleright t(c)) ; f^{\uparrow H} = (c \rightarrow p \triangleright t(c)) \triangleright f^{\uparrow H} \\ \text{definition of } f^{\uparrow H} : & \quad = c \rightarrow c \triangleright f^{\downarrow C} ; (c \rightarrow p \triangleright t(c)) ; f^{\uparrow G} \\ \text{apply function to } c \triangleright f^{\downarrow C} : & \quad = c \rightarrow p \triangleright (c \triangleright f^{\downarrow C}) ; f^{\uparrow G} . \end{aligned}$$

So, we have reduced both sides of \tilde{t} 's law to the two sides of the law (9.71) applied to an arbitrary p^{F^A} and considered as functions of c^{C^B} .

Reduction to natural transformations works similarly when C^\bullet is a functor and F^\bullet, G^\bullet are contrafunctors. A naturality law of $t : C^A \rightarrow F^A \rightarrow G^A$ can then be derived from t 's type signature.

Liftings A “lifted” function $f^{\uparrow F}$ is the result of applying a functor F 's `map` operation to a function f . The type signatures of `flatMap` and `liftOpt` are similar to that of `map` except for using the function type $A \rightarrow \text{Opt}^B$ instead of $A \rightarrow B$:

$$\begin{aligned} \text{fmap}_F : (A \rightarrow B) & \rightarrow F^A \rightarrow F^B , \\ \text{flm}_{\text{Opt}} : (A \rightarrow \text{Opt}^B) & \rightarrow \text{Opt}^A \rightarrow \text{Opt}^B , \\ \text{liftOpt}_F : (A \rightarrow \text{Opt}^B) & \rightarrow F^A \rightarrow F^B . \end{aligned}$$

Replacing `Opt` by an arbitrary functor G , we obtain the type signature of a “generalized lifting”,

$$\text{lift}_{G,F}^{A,B} : (A \rightarrow G^B) \rightarrow F^A \rightarrow F^B ,$$

which we can view as a lifting of functions with a “twisted” type $A \rightarrow G^B$ (which we call **Kleisli functions**) to functions of type $F^A \rightarrow F^B$. We will look at properties of generalized liftings in the next subsection. Here we focus on the naturality laws for generalized liftings.

A generalized lifting has two type parameters and two naturality laws. Looking at the two naturality laws (9.23), (9.35) for liftOpt or at the two laws (9.46), (9.47) for fmapOpt , we notice that each naturality law replaces one of the type parameters but keeps the other type parameter unchanged. This motivates us to fix one of the type parameters in the type signature of $\text{lift}_{G,F}^{A,B}$. For fixed A , the function $\text{lift}_{G,F}^{A,B}$ is a natural transformation between functors $A \rightarrow G^\bullet$ and $F^A \rightarrow F^\bullet$. With a fixed B , the function $\text{lift}_{G,F}^{A,B}$ is a natural transformation between contrafunctors $\bullet \rightarrow G^B$ and $F^\bullet \rightarrow F^B$. One can show that the corresponding naturality laws for these two natural transformations are equivalent to the two naturality laws of a generalized lifting.

We have reduced the four patterns of naturality laws to the laws of natural transformations, Eq. (9.69) for functors and Eq. (9.70) for contrafunctors, which are easier to understand.

Parametricity theorem It turns out that the naturality law of a natural transformation $t : F^A \rightarrow G^A$ will *always hold* if the code of the function t is fully parametric. More precisely, naturality holds if the code of t is a combination of the eight standard code constructions (shown in Section 5.2.3) together with recursion. This is a consequence of the “parametricity theorem”, which is beyond the scope of this chapter.² So, we do *not* need to verify naturality laws for functions whose code is known to be fully parametric. This saves a significant amount of work, since every method of every typeclass will have one naturality law per type parameter. Until now, we have been systematically deriving and checking all naturality laws; but we will not check those laws in the rest of the book.

Even if naturality laws hold automatically, it is important to be able to recognize their form and to use them in derivations where they are frequently needed. The mnemonic recipe for naturality laws (9.69)–(9.70) is that an arbitrary function f is lifted to the functor F at the left side of $t : F^A \rightarrow G^A$ and to the functor G at the right side of t , matching the two sides of the type signature $F^A \rightarrow G^A$.

All methods of typeclasses considered in this book are covered by the natural transformation recipe. However, not all type signatures of fully parametric functions can be reduced to natural transformations. For example, $t^A : (A \rightarrow A) \rightarrow A$ is not of the form $F^A \rightarrow G^A$ where F, G are either functors or contrafunctors. The parametricity theorem will still produce naturality laws for such functions; but we will not be able to write those laws via the natural transformation recipe. (A general procedure that works for all type signatures is given in Section D.2.)

9.4.2 Generalizing the laws of liftings. Kleisli functions

As we have seen in this chapter, the laws of filtering may be formulated equivalently via the `filter`, `deflate`, or `liftOpt` methods. These methods and their laws are equivalent but play different roles: `filter` is the most convenient to use in program code; `deflate` is the easiest type signature to implement and to reason about, especially in order to demonstrate that a functor is not filterable; `liftOpt` has the fewest laws and is most convenient for proofs of general type constructions.

If we put the naturality laws aside, `liftOpt` has the laws of identity (9.27) and composition (9.31). It is notable how these two laws are similar to the functor laws (6.2)–(6.3):

$$\begin{aligned} \text{liftOpt}_F(\text{pu}_F) &= \text{id} \quad , \quad \text{liftOpt}_F(f^{:A \rightarrow \mathbb{1}+B}) ; \text{liftOpt}_F(g^{:B \rightarrow \mathbb{1}+C}) = \text{liftOpt}_F(f \diamond_{\text{Opt}} g) \quad . \\ \text{fmap}_F(\text{id}) &= \text{id} \quad , \quad \text{fmap}_F(f^{:A \rightarrow B}) ; \text{fmap}_F(g^{:B \rightarrow C}) = \text{fmap}_F(f ; g) \quad . \end{aligned}$$

The only difference between these laws is in replacing $\text{id}^{:A \rightarrow A}$ by $\text{pu}_F^{:A \rightarrow F^A}$ and the function composition $f ; g$ by the Kleisli composition $f \diamond_{\text{Opt}} g$. We will now focus on the analogy between these laws, which goes far beyond the superficial similarity of form.

Kleisli functions $f^{:A \rightarrow \mathbb{1}+B}$ and $g^{:B \rightarrow \mathbb{1}+C}$ cannot be composed as $f ; g$ with the ordinary function composition. If we instead use the Kleisli composition, $f \diamond_{\text{Opt}} g$, the properties of Kleisli functions with respect to composition become completely analogous to the properties of the ordinary functions, except that the $\text{pu}_{\text{Opt}} : A \rightarrow \mathbb{1} + A$ plays the role of the identity ($\text{id}^{:A \rightarrow A}$).

²Formulations and proofs sufficient for the scope of this book are given in Appendix D.

Statement 9.4.2.1 The Kleisli composition \diamond_{Opt} obeys the identity and the associativity laws:

$$\begin{aligned} \text{identity laws : } \text{pu}_{\text{Opt}}^{:A \rightarrow \text{Opt}^A} \diamond_{\text{Opt}} g^{:A \rightarrow \text{Opt}^B} &= g \quad , \quad f^{:A \rightarrow \text{Opt}^B} \diamond_{\text{Opt}} \text{pu}_{\text{Opt}}^{:B \rightarrow \text{Opt}^B} = f \, , \\ \text{associativity law : } (f^{:A \rightarrow \text{Opt}^B} \diamond_{\text{Opt}} g^{:B \rightarrow \text{Opt}^C}) \diamond_{\text{Opt}} h^{:C \rightarrow \text{Opt}^D} &= f \diamond_{\text{Opt}} (g \diamond_{\text{Opt}} h) \quad . \end{aligned}$$

Proof Use the definitions (9.32) and Eqs. (9.34), (9.67), (9.40) derived previously in this chapter:

$$\begin{aligned} \text{use Eq. (9.34) : } \text{pu}_{\text{Opt}} \diamond_{\text{Opt}} g &= \text{pu}_{\text{Opt}} \circ \text{flm}_{\text{Opt}}(g) = g \quad , \\ \text{use Exercise 9.3.2.3 : } f \diamond_{\text{Opt}} \text{pu}_{\text{Opt}} &= f \circ \text{flm}_{\text{Opt}}(\text{pu}_{\text{Opt}}) = f \circ \text{id} = f \quad , \\ \text{expect to equal } f \diamond_{\text{Opt}} (g \diamond_{\text{Opt}} h) : (f \diamond_{\text{Opt}} g) \diamond_{\text{Opt}} h &= f \circ \text{flm}_{\text{Opt}}(g) \circ \text{flm}_{\text{Opt}}(h) \\ \text{use Eq. (9.40) : } &= f \circ \text{flm}_{\text{Opt}}(g \diamond_{\text{Opt}} h) = f \diamond_{\text{Opt}} (g \diamond_{\text{Opt}} h) \quad . \end{aligned}$$

This calculation motivates the name “associativity law” for Eq. (9.40).

The function lift_{Opt} can be viewed as a “generalized lifting” from Kleisli functions $A \rightarrow \mathbb{1} + B$ to functions $F^A \rightarrow F^B$, just as fmap is a lifting from ordinary functions $A \rightarrow B$ to functions $F^A \rightarrow F^B$; the laws of composition and the laws of liftings are analogous. The close analogy between ordinary functions and Kleisli functions means that any proofs of properties of ordinary liftings can be mechanically translated into proofs of the corresponding properties of generalized liftings. Indeed, replacing fmap by lift_{Opt} and id by pu_{Opt} where appropriate, we can translate the proof of Statement 6.2.3.3 (functor product), written using the pair product operation \boxtimes , into the proof of Statement 9.2.4.2 (filterable functor product). The same holds for the proofs of functor co-product and functor composition constructions.

The similarity between these proofs means, in the mathematical sense, that we have been proving essentially the same statements twice but did not have the appropriate level of abstraction to see that. While functional programmers may accept the work of writing these proofs twice, a mathematician would not be satisfied without defining a “generalized lifting” that replaces Opt by a functor M ,

$$\begin{aligned} \text{lift}_{M,F} : (A \rightarrow M^B) &\rightarrow F^A \rightarrow F^B \quad , \\ \text{pu}_M : A \rightarrow M^A \quad , \quad \diamond_M : (A \rightarrow M^B) &\rightarrow (B \rightarrow M^C) \rightarrow (A \rightarrow M^C) \quad , \end{aligned}$$

and postulating the required properties as the set of identity, associativity, and composition laws,

$$\begin{aligned} \text{lift}_{M,F}(\text{pu}_M^{:A \rightarrow M^A}) &= \text{id}^{:F^A \rightarrow F^A} \quad , \quad \text{lift}_{M,F}(f) \circ \text{lift}_{M,F}(g) = \text{lift}_{M,F}(f \diamond_M g) \quad , \\ \text{pu}_M \diamond_M g &= g \quad , \quad f \diamond_M \text{pu}_M = f \quad , \quad (f \diamond_M g) \diamond_M h = f \diamond_M (g \diamond_M h) \quad . \end{aligned}$$

Now the two sets of proofs can be replaced by a single set of proofs formulated for an “ M -filterable” functor F , where M could be later set to the identity functor or the Opt functor.

Not all functors M support the Kleisli composition \diamond_M with the required laws. We will study such functors M , which are known as **monads**, in Chapter 10.

9.4.3 Motivation for using category theory

We have seen four examples of operations that have the form of a “lifting”:

$$\begin{aligned} \text{for functors } F : \text{fmap}_F : (A \rightarrow B) &\rightarrow (F^A \rightarrow F^B) \quad , \\ \text{for contrafunctors } C : \text{cmap}_C : (B \rightarrow A) &\rightarrow (C^A \rightarrow C^B) \quad , \\ \text{for } M\text{-filterable functors } F : \text{lift}_{M,F} : (A \rightarrow M^B) &\rightarrow (F^A \rightarrow F^B) \quad , \\ \text{for } M\text{-filterable contrafunctors } C : \text{lift}_{M,F} : (B \rightarrow M^A) &\rightarrow (C^A \rightarrow C^B) \quad . \end{aligned}$$

All these operations obey analogous laws of naturality, identity, and composition, and differ only in the type of functions being lifted: the ordinary function $A \rightarrow B$, the “reversed” type $B \rightarrow A$, the

Kleisli function $A \rightarrow M^B$, and the “reversed” Kleisli function $B \rightarrow M^A$. In turn, all these function types obey the laws of identity and composition. (For the types to match, composition of reversed functions needs to be performed in the reverse order.)

In order to avoid writing essentially the same proofs multiple times, we use a more abstract view of this situation: a new notion of a functor that lifts more general or “twisted” function types instead of the ordinary function types ($A \rightarrow B$). This notion (and its further generalizations) is provided by **category theory**, which turns out to be a convenient language for describing various laws and types of operations in functional programming.

Category theory generalizes functions of type $A \rightarrow B$ to **morphisms** $A \rightsquigarrow B$, which can be in any relation to A and B as long as the identity and composition laws hold. The morphism types, the “identity morphism” of type $A \rightsquigarrow A$, and the composition operation must be chosen appropriately. These choices, together with a definition of the admissible types A, B, \dots , define a **category**, e.g.:

Category	Types A, B, \dots	Morphisms $f : A \rightsquigarrow B$	Identity morphism	Composition
“plain”	<code>Int, String, ...</code>	$f : A \rightarrow B$	$\text{id}^{A \rightarrow A}$	$f \circ g$
“reversed”	<code>Int, String, ...</code>	$f : B \rightarrow A$	$\text{id}^{A \rightarrow A}$	$g \circ f$
“ M -Kleisli”	<code>Int, String, ...</code>	$f : A \rightarrow M^B$	$\text{pu}_M : A \rightarrow M^A$	$f \diamond_M g$
“reversed M -Kleisli”	<code>Int, String, ...</code>	$f : B \rightarrow M^A$	$\text{pu}_M : A \rightarrow M^A$	$g \diamond_M f$
“ F -lifted”	$F^{\text{Int}}, F^{\text{String}}, \dots$	$f : F^A \rightarrow F^B$	$\text{id}^{F^A \rightarrow F^A}$	$f \circ g$

Definition of category A category C is a collection of **objects**³ $\{A, B, \dots\}$ and a collection of morphisms $\{f_1, f_2, \dots\}$, where each morphism is labeled by two objects A, B as $f : A \rightsquigarrow B$ (the objects A and B do not need to be different). For any object A , the identity morphism $\text{id}_C^{A \rightsquigarrow A}$ must exist. For any two morphisms $f : A \rightsquigarrow B$ and $g : B \rightsquigarrow C$, the composition $f \circ g$ must exist as a morphism labeled $A \rightsquigarrow C$. Additionally, the identity and the associativity laws must hold:

$$f \circ_C \text{id}_C = f = f \circ_C \text{id}_C \quad , \quad (f \circ_C g) \circ_C h = f \circ_C (g \circ_C h) \quad .$$

Most applications of category theory to functional programming will use categories whose objects are the types (`Int, String, ...`) of the programming language (or, sometimes, type constructors), and whose morphisms are functions with type signatures defined in a specific way, e.g., $A \rightarrow M^B$ or $F^A \rightarrow F^B$ as we have seen. However, the definition of category is general and does not assume that morphisms are functions between objects.

In functional programming, a functor is a type constructor F^\bullet with a lawful lifting of functions $A \rightarrow B$ to functions $F^A \rightarrow F^B$. Category theory defines a **functor** much more generally — as a lawful lifting *from one category to another*. We will use the phrase “**categorical functor**” to distinguish the notion of functor in category theory from the programmer’s notion (a type constructor with `map`).⁴

Definition of categorical functor Given two categories C and D , a functor $\mathcal{F} : C \rightarrow D$ is a mapping of each type A in C to the corresponding type $\mathcal{F}(A)$ in D , as well as a mapping of each morphism $f : A \rightsquigarrow_C B$ from C to a corresponding morphism $\mathcal{F}(f) : \mathcal{F}(A) \rightsquigarrow_D \mathcal{F}(B)$ in D . Additionally, the laws of identity and composition must hold according to the rules of each category. That is, identity morphisms id_C from the category C must be mapped to those of the category D ; and the composition (\circ_C) of any two morphisms in the category C must be mapped to the composition (\circ_D) of morphisms in the category D .

$$\text{identity law of categorical functor : } \mathcal{F}(\text{id}_C^{A \rightsquigarrow C}) = \text{id}_D^{\mathcal{F}(A) \rightsquigarrow \mathcal{F}(C)} \quad ,$$

$$\text{composition law of categorical functor : } \mathcal{F}(f : A \rightsquigarrow_C B) \circ_D \mathcal{F}(g : B \rightsquigarrow_C C) = \mathcal{F}(f \circ_C g) \quad .$$

To clarify this definition, consider two examples: a contrafunctor C and a filterable functor F .

³The “objects” in category theory are not related to “object-oriented programming”.

⁴In category theory, programmer’s functors are called **endofunctors** — categorical functors from a category C to itself.

Example 9.4.3.1 A contrafunctor C is specified via a lawful function $\text{cmap} : (B \rightarrow A) \rightarrow C^A \rightarrow C^B$. To formulate this in category theory, we say that there must exist a categorical functor from the reversed category to the C -lifted category. To define that functor, we need to specify how types and morphisms are mapped from the first category to the second. Each type A of the reversed category is mapped to the corresponding type C^A of the C -lifted category. Each $(A \rightsquigarrow B)$ -morphism $f : B \rightarrow A$ of the reversed category is mapped to the morphism $f \downarrow C : C^A \rightarrow C^B$ of the C -lifted category.

To formulate the laws of identity and composition for the (categorical) functor, we look up the definitions of the identity morphisms and the composition operation in each category:

for the reversed category : $\text{id}^{A \rightarrow A}$ and $g ; f$,

for the C -lifted category : $\text{id}^{C^A \rightarrow C^A}$ and $f \circ g$.

Now we require that the first category's identity morphism is mapped to the second category's identity morphism, and that a composition of any two morphisms (as defined in the first category) is mapped to a composition as defined in the second category:

identity law : $(\text{id}^{A \rightarrow A}) \downarrow C = \text{id}^{C^A \rightarrow C^A}$,

composition law : $(g ; f) \downarrow C = f \downarrow C ; g \downarrow C$.

We derived these laws previously as the laws of contrafunctors.

Example 9.4.3.2 A filterable functor F is specified via a lawful function $\text{liftOpt}_F : (A \rightarrow \text{Opt}^B) \rightarrow F^A \rightarrow F^B$. To formulate the categorical definition for this situation, we say that there must exist a functor from the Opt-Kleisli category to the F -lifted category.

To define that (categorical) functor, we need to specify how types and morphisms are mapped from the first category to the second. Each type A of the Opt-Kleisli category is mapped to the corresponding type F^A of the F -lifted category. Each $(A \rightsquigarrow B)$ -morphism $f : A \rightarrow \text{Opt}^B$ of the Opt-Kleisli category is mapped to the morphism $\text{liftOpt}_F(f) : F^A \rightarrow F^B$ of the F -lifted category.

Formulate the laws of identity and composition for the (categorical) functor using the definitions of the identity morphisms and the composition operation in each category:

for the Opt-Kleisli category : $\text{pu}_{\text{Opt}}^{A \rightarrow \text{Opt}^A}$ and $f \diamond_{\text{Opt}} g$,

for the F -lifted category : $\text{id}^{F^A \rightarrow F^A}$ and $f ; g$.

Now we require that the first category's identity morphism is mapped to the second category's identity morphism, and that a composition of any two morphisms (as defined in the first category) is mapped to a composition as defined in the second category:

identity law : $\text{liftOpt}_F(\text{pu}_{\text{Opt}}) = \text{id}^{F^A \rightarrow F^A}$,

composition law : $\text{liftOpt}_F(f \diamond_{\text{Opt}} g) = \text{liftOpt}_F(f) ; \text{liftOpt}_F(g)$.

We derived these laws previously as the laws of `liftOpt`.

In both examples, laws of the appropriately formulated categorical functors turned out to be equivalent to laws we derived previously. This gives us assurance that we have correctly guessed the relevant laws. The choice of laws is not self-evident. For example, Section 9.1.2 derived the four laws of the `filter` function from heuristic ideas. The laws of `filter` have been formulated differently by different people, also starting from heuristic considerations.⁵ It is not obvious whether we correctly guessed the relevant laws of `filter` and did not assume more laws than necessary. In contrast, the two laws of the categorical functor are very general and appear time and again in different areas of mathematics. This gives us confidence that these laws are correctly chosen and will be useful in a

⁵E.g., with an additional “empty-value” law here: <https://github.com/fantasyland/fantasy-land#filterable>

wide range of contexts. Proving that the four laws of `filter` from Section 9.1.2 are equivalent to the two laws of a categorical functor gives assurance that our choice of filtering laws is mathematically consistent and is likely to prove useful in applications.

Having looked at the laws of `liftOpt`, we noticed that we can reduce the number of different proofs if we generalize the Opt-Kleisli category to an M -Kleisli category with a suitable functor M . It turns out that one can prove general theorems about products, co-products, and composition of (categorical) functors that map between any categories. In this way, we can replace four theorems (say, for the product of functors, contrafunctors, filterable functors, and filterable contrafunctors) by a single but more abstract theorem about the product of (categorical) functors being a functor between suitably defined categories. We will not look at these proofs here; Chapters 6–9 already worked through a few almost identical proofs that show the required techniques.

The categorical view also shows us two directions for developing the theory further, hoping to find useful applications. First, we can look for functors M (called “monads”) that admit the Kleisli composition with the properties (the identity and the associativity laws) required by an M -Kleisli category. Second, having found some new monads M , we can look for “ M -filterable” functors or contrafunctors F that admit an operation $\text{lift}_{M,F}$ similar to liftOpt_F but adapted to the monad M instead of `Option`. We will see some examples of M -filterable contrafunctors later in this book.

To summarize, using the category theory’s notion of functor brings important advantages:

- We have assurance that we formulated the laws correctly.
- We can find some promising directions for generalizing the constructions.
- Many proofs can be reduced to a single proof for properties of (categorical) functors.

In this way, we find that category theory is a useful tool for reasoning about abstract constructions that work with different typeclasses (functor, contrafunctor, filterable, etc.) in a similar way.

What does category theory *not* do for functional programming?

- It defines many abstract constructions but does not say which of these constructions will be useful in practical programming.
- It does not provide any proofs of specific laws for a given function and does not give any guidance towards finding such proofs or derivations.
- It does not help in determining whether a given type constructor is, say, a filterable functor, or a pointed functor, or a monad.
- It does not help in implementing typeclass instances that obey the required laws.
- It does not say whether there exists a natural transformation between two given functors, or how to implement one.

Answering these questions requires symbolic derivations using techniques specific to functional programming. Developing these techniques (while avoiding unnecessary theoretical material) is one of the main themes of this book.

10 Computations in functor blocks. II. Semimonads and monads

In a Scala programmer's view, functors are type constructors with a `map` method, filterable type constructors have a `filter` method, and **semimonads** are functors that have a `flatMap` method. This chapter begins by developing an intuition for the behavior of `flatMap`.

10.1 Practical use of monads

10.1.1 Motivation for semimonads: Nested iteration

How can we translate into code a computation that contains nested iterations, such as

$$\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \frac{1}{1+i+j+k} = ? \quad (10.1)$$

Recall that a "flat" (non-nested) iteration is translated into the `map` method applied to a sequence:

$$(1 \text{ to } n).map \{ i \Rightarrow 1.0 / (1 + i) \}.sum \quad \sum_{i=1}^n \frac{1}{1+i} .$$

The mathematical notation combines summation and iteration in one symbol, but the code separates these two steps: first, a sequence is computed as `(1 to n).map { i => 1.0 / (1 + i) }`, holding the values $\frac{1}{1+i}$ for $i = 1, \dots, n$, and only then the `sum` function is applied to the sequence. This separation is useful because it gives us full flexibility to transform or aggregate the sequence.

So, we will treat nested iterations in a similar way: first, compute a sequence of values that result from nested iterations, and then apply transformations or aggregations to that sequence.

If we use nested `map` operations, we will obtain a nested data structure, e.g., a vector of vectors:

```
scala> (1 to 5).map(i => (1 to i).map(j => i * j))
res0: IndexedSeq[IndexedSeq[Int]] = Vector(Vector(1), Vector(2, 4), Vector(3, 6, 9), Vector(4, 8, 12, 16), Vector(5, 10, 15, 20, 25))
```

We need to "flatten" this nested structure into a simple, non-nested sequence. The standard method for that is `flatten`, and its combination with `map` can be replaced by `flatMap`:

```
scala> (1 to 4).map(i => (1 to i).map(j => i * j)).flatten
res1: IndexedSeq[Int] = Vector(1, 2, 4, 3, 6, 9, 4, 8, 12, 16)

scala> (1 to 4).flatMap(i => (1 to i).map(j => i * j))      // Same result as above.
res2: IndexedSeq[Int] = Vector(1, 2, 4, 3, 6, 9, 4, 8, 12, 16)
```

To represent more nesting, we use more `flatMap` operations. For example, to implement Eq. (10.1):

```
def example(n: Int): Double = (1 to n).flatMap { i =>
  (1 to n).flatMap { j =>
    (1 to n).map { k =>
      1.0 / (1.0 + i + j + k)
    }
  }.sum
}

scala> example(10)
res3: Double = 63.20950497687006
```

These examples show that converting nested iterations into a simple iteration means replacing all `map` functions by `flatMap` except for the last `map` call (that by itself produces a non-nested sequence).

The `for/yield` syntax (or “functor block”) is an easier way of flattening nested iterations: just use a new source line for each level of nesting. Compare the two following code fragments line by line:

```
(for { i <- 1 to n
      j <- 1 to n
      k <- 1 to n
    } yield 1.0 / (1.0 + i + j + k)
).sum
```

```
(1 to n).flatMap { i =>
  (1 to n).flatMap { j =>
    (1 to n).map { k =>
      1.0 / (1.0 + i + j + k)
    }
  }.sum
}
```

The left arrows visually suggest that the variables `i`, `j`, `k` will iterate over the given sequences. All left arrows except the last one are replaced by `flatMap`s; the last left arrow is replaced by a `map`. These replacements are performed automatically by the Scala compiler.

A functor block with source lines and conditionals corresponds to the mathematical notation for creating sets of values. An example of using that notation is the formula

```
val t = for {
  x <- p
  y <- q
  z <- r
  if f(x, y, z) == 0
} yield x + y + z
```

$$T = \{x + y + z \mid x \in P, y \in Q, z \in R, f(x, y, z) = 0\} \quad .$$

Here, `P`, `Q`, `R` are given sets of numbers, and the result is a set `T` of numbers obtained by adding some `x` from `P`, some `y` from `Q`, and some `z` from `R` such that $f(x, y, z) = 0$. A direct implementation of this formula is the code shown at left. Here, `p`, `q`, `r` are given collections (say, arrays) and the result `t` is again an array. Just like the mathematical formula’s result is a collection of some $x + y + z$ values, the functor block’s result is a collection of values computed after the `yield` keyword.

To develop more intuition about using functor blocks with multiple left arrows, look at this code:

```
val result = for {
  i <- 1 to m
  j <- 1 to n
  x = f(i, j)
  k <- 1 to p
  y = g(i, j, k)
} yield h(x, y)
```

```
val result = {
  (1 to m).flatMap { i =>
    (1 to n).flatMap { j =>
      val x = f(i, j)
      (1 to p).map { k =>
        val y = g(i, j, k)
        h(x, y)
      }
    }
  }
}
```

One can imagine that each line (which we can read as “for all i in $[1, \dots, m]$ ”, “for all j in $[1, \dots, n]$ ”, etc.) will produce an intermediate sequence of the same type. Each next line continues the calculation from the previous intermediate sequence.

If this intuition is correct, we should be able to refactor the code by cutting the calculation at any place and continuing in another functor block, without changing the result value:

```
val result = for {
  i <- 1 to m
  j <- 1 to n
  // We will cut the block here, making i and j
  // available for further computations.
  x = f(i, j)
  k <- 1 to p
  y = g(i, j, k)
} yield h(x, y)
// The 'result' is equal to 'res2' at right.
```

```
val res1 = for {
  i <- 1 to m
  j <- 1 to n
} yield (i, j) // Intermediate sequence 'res1'.
val res2 = for {
  (i, j) <- res1 // Continue from 'res1'.
  x = f(i, j)
  k <- 1 to p
  y = g(i, j, k)
} yield h(x, y)
```

This example illustrates the two features of functor blocks that often cause confusion:

- Each “source” line computes an intermediate collection of the same type, so all values to the right of `<-` must use *the same* type constructor (or its subtypes).
- The entire functor block’s result is again a collection using the same type constructor. The result is *not* the expression under `yield`; instead, it is a collection of those expressions.

So far, we have been using sequences as the main type constructor. However, functor blocks with several left arrows will work with any other type constructor that has `map` and `flatMap` methods. In the next sections, we will see how to use functor blocks with different type constructors.

Functors having `flatMap` methods are called **semimonads** in this book.¹ In practice, most semimonads also have a `pure` method (i.e., belong to the `Pointed` typeclass, see Section 8.3.5). Semimonads with a `pure` method (and obeying the appropriate laws) are called **monads**. This chapter will study semimonads and monads in detail. For now, we note that the functor block syntax does not require functors to have a `pure` method; it works just as well with semimonads.

If a functor has a `withFilter` method, Scala's functor block will also support the `if` operation (see Section 9.1). So, the full functionality of functor blocks can be used with *filterable semimonads*.

10.1.2 List-like monads

List-like monads are types that model a collection of data values. Examples are `Seq` and its subtypes, `Stream`, `Array`, non-empty lists, sets, and dictionaries. These data types make different choices of lazy or eager evaluation and memory allocation, but their `flatMap` methods work similarly: they “flatten” nested collections.

All list-like monads are also filterable. A typical task for list-like monads is to find a set of solutions of a combinatorial problem by first taking all combinations and then filtering out unwanted ones.

Example 10.1.2.1 Compute all permutations of the three letters “a”, “b”, “c”.

Solution We will compute a *sequence* of all permutations by nested iteration. First attempt:

```
scala> for {
  x <- Seq("a", "b", "c")
  y <- Seq("a", "b", "c")
  z <- Seq("a", "b", "c")
} yield x + y + z
res0: Seq[String] = List(aaa, aab, aac, aba, abb, abc, aca, acb, acc, baa, bab, bac, bba, bbb, bbc,
  bca, bcb, bcc, caa, cab, cac, cba, cbb, cbc, cca, ccb, ccc)
```

To obtain all permutations and nothing else, we need to exclude all repeated subsequences such as “aab”. To achieve that, we must make `y` iterate over letters that do not include the current value of `x`:

```
val xs = Seq("a", "b", "c")

scala> for {
  x <- xs
  xsWithoutX = xs.filter(_ != x)
  y <- xsWithoutX
  xsWithoutXY = xsWithoutX.filter(_ != y)
  z <- xsWithoutXY
} yield x + y + z
res1: Seq[String] = List(abc, acb, bac, bca, cab, cba)
```

Example 10.1.2.2 Compute the set of all subsets of `xs = Set("a", "b", "c")`.

Solution We aim to write the code as a nested iteration. Begin by choosing one element, say “a”. Some subsets of `xs` will contain “a” and other subsets will not. So, let `xa` iterate over two possibilities: either an empty set, `Set()`, or a set containing just “a”. Then each subset of `xs` is the union of `xa` and some subset that does not contain “a”. We repeat the same logic for “b” and “c”. The code is:

```
val empty = Set[String]()

scala> for {
  xa <- Set(empty, Set("a"))
  xb <- Set(empty, Set("b"))
  xc <- Set(empty, Set("c"))
} yield xa ++ xb ++ xc      // p ++ q is the union of the sets p and q.
```

¹There is no single accepted name. The libraries `scalaz` and `cats` call the relevant typeclasses `Bind` and `FlatMap` respectively.

```
res0: Set[Set[String]] = Set(Set(), Set(a, b), Set(b, c), Set(a, c), Set(a, b, c), Set(b), Set(c), Set(a))
```

Example 10.1.2.3 Compute all sub-arrays of length 3 in a given array. Type signature and a test:

```
def subarrays3[A](input: IndexedSeq[A]): IndexedSeq[IndexedSeq[A]] = ???

scala> subarrays3(IndexedSeq("a", "b", "c", "d"))
res3: IndexedSeq[IndexedSeq[String]] = Vector(Vector(a, b, c), Vector(a, b, d), Vector(a, c, d),
Vector(b, c, d))
```

Solution What are the indices of the elements of the required sub-arrays? Suppose n is the length of the given array. In any sub-array of length 3, the first element must have an index i such that $0 \leq i < n$. The second element must have an index j such that $i < j < n$, and the third element's index k must satisfy $j < k < n$. So we can iterate over the indices like this,

```
def subarrays3[A](input: IndexedSeq[A]): IndexedSeq[IndexedSeq[A]] = {
  val n = input.length
  for {
    i <- 0 until n                         // Iterate over 0,1,...,n-1.
    j <- i + 1 until n
    k <- j + 1 until n
  } yield IndexedSeq(input(i), input(j), input(k))
}
```

Example 10.1.2.4 Generalize examples 10.1.2.1–10.1.2.3 to support an arbitrary size n instead of 3.

Solution (a) The task is to compute the set of all permutations of n letters. We note that the solution in Example 10.1.2.1 used three source lines (such as $x <- xs$), one for each letter. To generalize that code to n letters, we would need to write a functor block with n source lines. However, we cannot do that if n is a run-time parameter not fixed in advance. So, the functor block must use recursion in n . Begin implementing `permutations(xs)` as a recursive function:

```
def permutations(xs: Seq[String]): Seq[String] = for {
  x <- xs
  xsWithoutX = xs.filter(_ != x)
  ??? permutations(xsWithoutX) ??? // We need to use a recursive call somehow.
```

It is promising to use a recursive call of `permutations` with the sub-sequence `xsWithoutX` that does not contain a chosen letter `x`. It remains to formulate the code as nested iteration. Let us visualize the computation for `xs == Seq("a", "b", "c")`. While iterating over `xs`, we start with `x == "a"`, which gives `xsWithoutX == Seq("b", "c")`. Iterating over `permutations(xsWithoutX)`, we obtain the permutations `"bc"` and `"cb"`. These permutations need to be concatenated with `x == "a"`, yielding `"abc"` and `"acb"`, which is the correct part of the final answer. So, we write a nested iteration and concatenate the results:

```
def permutations(xs: Seq[String]): Seq[String] = for {
  x <- xs
  xsWithoutX = xs.filter(_ != x)
  rest <- permutations(xsWithoutX)
} yield x + rest

scala> permutations(Seq("a", "b", "c", "d"))
res0: Seq[String] = List()
```

The code is wrong: it always returns an empty list! The reason is that we provided no base case for the recursion. Eventually the intermediate value `xsWithoutX` becomes empty. A nested iteration with an empty list always makes the final result also an empty list. To fix this, add a base case:

```
def permutations(xs: Seq[String]): Seq[String] = if (xs.length == 1) xs else for {
  x <- xs
  xsWithoutX = xs.filter(_ != x)
  rest <- permutations(xsWithoutX)
} yield x + rest
```

```
scala> permutations(Seq("a", "b", "c", "d"))
res1: Seq[String] = List(abcd, abdc, acbd, acdb, adbc, adcb, bacd, badc, bcad, bcda, bdac, bdca,
    cabd, cadb, cbad, cbda, cdab, cdba, dabc, dacb, dbac, dbca, dcab, dcba)
```

(b) To find all subsets of a set via nested iteration, we cannot directly extend the code from Example 10.1.2.2 because we cannot write an unknown number of source lines in a functor block. As in part (a), we need to refactor the code and write a functor block that uses recursion.

```
def subsets[A](xs: Set[A]): Set[Set[A]] = for {
  x <- xs
  xsWithoutX = xs - x           // Use the difference operation for sets.
  ??? subsets(xsWithoutX) ???  // We need to use a recursive call somehow.
```

If $xs == \text{Set("a", "b", "c")}$ and $x == "a"$ during an iteration, we get $xsWithoutX == \text{Set("b", "c")}$. Once we compute $\text{subsets}(xsWithoutX)$, we need to use all those subsets together with x and also without adding x . We also should not forget to write the base case (an empty set xs). So, the code becomes

```
def subsets[A](xs: Set[A]): Set[Set[A]] = if (xs.isEmpty) Set(xs) else for {
  x <- xs
  xsWithoutX = xs - x
  rest <- subsets(xsWithoutX)    // Recursive call.
  maybeX <- Set(Set(x), Set())
} yield maybeX ++ rest

scala> subsets(Set("a", "b", "c", "d"))
res0: Set[Set[String]] = Set(Set(), Set(a, c, d), Set(a, b), Set(b, c), Set(a, d), Set(a, b, d),
  Set(b, c, d), Set(a, c), Set(c, d), Set(a, b, c), Set(d), Set(b), Set(b, d), Set(a, b, c, d),
  Set(c), Set(a))
```

(c) To compute all sub-arrays of length n from a given array, we rewrite the solution in Example 10.1.2.3 via a recursive function that computes the sequence of indices:

```
def subindices(begin: Int, end: Int, n: Int): IndexedSeq[IndexedSeq[Int]] =
  if (n == 0) IndexedSeq(IndexedSeq()) else for {
    i <- begin until end
    rest <- subindices(i + 1, end, n - 1)    // Recursive call.
  } yield IndexedSeq(i) ++ rest

scala> subindices(0, 4, 2)
res0: IndexedSeq[IndexedSeq[Int]] = Vector(Vector(0, 1), Vector(0, 2), Vector(0, 3), Vector(0, 4),
  Vector(1, 2), Vector(1, 3), Vector(1, 4), Vector(2, 3), Vector(2, 4), Vector(3, 4))
```

The sequence of subarrays is easy to compute once the sequence of subarray indices is known:

```
def subarrays[A](n: Int, input: IndexedSeq[A]): IndexedSeq[IndexedSeq[A]] =
  subindices(0, input.length, n).map(_.map(i => input(i)))

scala> subarrays(4, IndexedSeq("a", "b", "c", "d", "e"))
res1: IndexedSeq[IndexedSeq[String]] = Vector(Vector(a, b, c, d), Vector(a, b, c, e), Vector(a, b,
  d, e), Vector(a, c, d, e), Vector(b, c, d, e))
```

The solutions (a)–(c) are not tail recursive because recursive calls within a source line in a functor block are translated into recursive calls not in tail positions but *inside* `map` or `flatMap` methods. Achieving tail recursion in functor blocks requires techniques beyond the scope of this chapter.

Example 10.1.2.5 Find all solutions of the “8 queens” problem.

Solution The 8 queens must be placed on an 8×8 chess board so that no queen threatens any other queen. To make our work easier, we note that each queen must be placed in a different row. So, it is sufficient to find the column index for each queen. A solution is a sequence of 8 indices.

Begin by iterating over all possible combinations of column indices:

```
val solutions = for {
  x0 <- 0 until 8           // Queen 0 has coordinates (x0, 0).
  x1 <- 0 until 8           // Queen 1 has coordinates (x1, 1).
  x2 <- 0 until 8           // Queen 2 has coordinates (x2, 2).
```

It remains to filter out invalid positions. We should start filtering as early as possible, since the total number of combinations grows quickly during nested iterations:

```
val solutions = for {
  x0 <- 0 until 8           // Queen 0 has coordinates (x0, 0).
  x1 <- 0 until 8           // Queen 1 has coordinates (x1, 1).
  if noThreat(x1, Seq(x0)) // Queen 1 does not threaten queen 0.
  x2 <- 0 until 8           // Queen 2 has coordinates (x2, 2).
  if noThreat(x2, Seq(x0, x1)) // Queen 2 does not threaten queens 0 and 1.
  ...
} yield Seq(x0, x1, x2, ...)
```

Here, `noThreat` is a helper function that decides whether a new queen threatens previous ones:

```
def noThreat(otherX: Int, prev: Seq[Int]): Boolean = {
  val otherY = prev.length
  prev.zipWithIndex.forall { case (x, y) => // Check the vertical and the two diagonals.
    x != otherX && x - y != otherX - otherY && x + y != otherX + otherY
  }
}
```

We used a feature of Scala allowing us to pass a sequence of arguments via the syntax `Int*`, which means a variable number of arguments of type `Int`. We can now complete the code and test it:

```
val column = 0 until 8
val solutions = for {
  x0 <- column           // Queen 0 has coordinates (x0, 0).
  x1 <- column           // Queen 1 has coordinates (x1, 1).
  if noThreat(x1, Seq(x0)) // Queen 1 does not threaten queen 0.
  x2 <- column           // Queen 2 has coordinates (x2, 2).
  if noThreat(x2, Seq(x0, x1)) // Queen 2 does not threaten queens 0 and 1.
  x3 <- column
  if noThreat(x3, Seq(x0, x1, x2))
  x4 <- column
  if noThreat(x4, Seq(x0, x1, x2, x3))
  x5 <- column
  if noThreat(x5, Seq(x0, x1, x2, x3, x4))
  x6 <- column
  if noThreat(x6, Seq(x0, x1, x2, x3, x4, x5))
  x7 <- column
  if noThreat(x7, Seq(x0, x1, x2, x3, x4, x5, x6))
} yield Seq(x0, x1, x2, x3, x4, x5, x6, x7)

scala> solutions.take(5) // First 5 solutions.
res0: IndexedSeq[Seq[Int]] = Vector(List(0, 4, 7, 5, 2, 6, 1, 3), List(0, 5, 7, 2, 6, 3, 1, 4),
List(0, 6, 3, 5, 7, 1, 4, 2), List(0, 6, 4, 7, 1, 3, 5, 2), List(1, 3, 5, 7, 2, 0, 6, 4))
```

Example 10.1.2.6 Generalize Example 10.1.2.5 to solve the “ n queens” problem.

Solution In this problem, n queens must be placed on an $n \times n$ chess board. We need to enumerate and count all solutions.² As in the 8-queens problem, each queen must be placed in a different row, and so we represent a solution by a sequence of n column indices, each index between 0 and $n - 1$.

Begin by writing code that expects to use itself recursively:

```
def nQueens(n: Int): Seq[Seq[Int]] = for {
  x0 <- 0 until n
  ??? nQueens(n - 1) ???
```

Possible positions of new queens depend on the chosen positions for all previous queens. So, the recursive function must receive that information. Write an auxiliary recursive function `nQueensPartial` that computes all remaining positions given a sequence of (less than n) previously found queens:

```
def nQueensPartial(n: Int, prev: Seq[Seq[Int]]): Seq[Seq[Int]] = for {
```

²There is no known general formula for the number of solutions of the n -queens problem. See a discussion here, <https://math.stackexchange.com/questions/1872444>

```

x <- 0 until n
if noThreat(x, prev)           // The new queen does not threaten any previous queens.
  rest <- nQueensPartial(n - 1, prev +: x) // Find positions with n - 1 new queens.
} yield x +: rest             // Prepend the new queen to the other queen positions.

```

This code still has two problems: first, the base case ($n = 0$) is not covered; second, the recursive function must be initially called with correct arguments. The complete code is

```

def nQueens(n: Int): Seq[Seq[Int]] = {
  def nQueensPartial(m: Int, prev: Seq[Int]): Seq[Seq[Int]] = if (m == 0) Seq(Seq()) else for {
    x <- 0 until n
    if noThreat(x, prev)
    rest <- nQueensPartial(m - 1, prev +: x)
  } yield x +: rest
  nQueensPartial(n, Seq())
}

scala> (nQueens(8).length, nQueens(9).length, nQueens(10).length, nQueens(11).length)
res0: (Int, Int, Int, Int) = (92,352,724,2680)

```

Example 10.1.2.7* Formulas of Boolean logic may be transformed by “expanding brackets”,

expand brackets with a conjunction inside : $(a \wedge b) \vee c = (a \vee c) \wedge (b \vee c)$,

expand brackets with a disjunction inside : $(a \vee b) \wedge c = (a \wedge c) \vee (b \wedge c)$.

Implication is written as $(a \Rightarrow b) = ((\neg a) \vee b)$. We can write these transformations in Scala code as

```

(a && b) || c == (a || c) && (b || c)
(a || b) && c == (a && c) || (b && c)
(a implies b) == ((!a) || b)

```

So, we can rewrite any Boolean formula as a conjunction of disjunctions with no more nested conjunctions inside, e.g., $(a \vee b) \wedge ((\neg c) \vee d \vee e)$. This form is called the **conjunctive normal form**³ (CNF) of a Boolean formula.

We can also rewrite Boolean formula into a *disjunction* of conjunctions with no more nesting inside, e.g., $(p \wedge q \wedge \neg r) \vee (\neg x \wedge y)$. This is called the **disjunctive normal form**⁴ (DNF) of a Boolean formula. The task is to implement functions that convert formulas from CNF to DNF and back.

Solution We begin by designing a data type to represent CNFs. Let the type parameter A stand for the type of elementary Boolean formulas that we denoted by $a, b, !c, \dots$, i.e., Boolean formulas that contain no conjunctions or no disjunctions. Then we may represent a conjunction as a `Set[A]`, and a disjunction of conjunctions as a `Set[Set[A]]`. For instance, $(a \vee b) \wedge (c \vee d \vee e)$ is represented by `Set(Set(a, b), Set(c, d, e))`. Because we are using `Set`, disjunction of two disjunctions is easily implemented as a union of sets. Since `Set` eliminates repeated elements, the representation based on `Set` will also automatically perform simplification of formulas such as $(a \vee b \vee a) == (a \vee b)$.

Conjunction of two conjunctions may be implemented as a union of sets for the same reason.

We also need to figure out how to represent the Boolean constants `true` and `false` in the CNF. When we compute a conjunction such as $(a \vee b) \wedge c \wedge \text{true}$, the result must be just $(a \vee b) \wedge c$. Since $(a \vee b) \wedge c$ is represented as `Set(Set(a, b), Set(c))`, and since we compute conjunctions via unions of sets, we expect to have the equation

```
Set(Set(a, b), Set(c)) union Set(???) == Set(Set(a, b), Set(c))
```

This must hold for any formulas; so we must represent `true` as an *empty* set of disjunctions, `Set()`.

```

final case class CNF[A](s: Set[Set[A]])
def trueCNF[A] = CNF[A](Set())
def falseCNF[A] = CNF[A](Set(Set()))

```

The value `false` should have the property that for any group g of disjunctions, e.g., $g == (x \vee y)$, we must have $g \vee \text{false} == g$. So, `false` corresponds to an empty disjunction. However, CNF represents disjunctions by nested sets contained within conjunctions. Thus, `Set(Set())` must stand for `false`.

Similar arguments hold for representing DNFs. We again use nested sets, but this time the conjunctions and disjunctions are swapped, so `Set(Set(a, b), Set(c, d, e))` stands for the DNF formula

³https://en.wikipedia.org/wiki/Conjunctive_normal_form

⁴https://en.wikipedia.org/wiki/Disjunctive_normal_form

$(a \&& b) \mid\mid (c \&& d \&& e)$. The value `Set()` represents an empty disjunction and stands for the constant `false`, while the constant `true` is `Set(Set())`, a disjunction containing an empty conjunction.

```
final case class DNF[A] (s: Set[Set[A]])  
def trueDNF[A] = DNF[A] (Set(Set()))  
def falseDNF[A] = DNF[A] (Set())
```

It is easy to make mistakes when reasoning about the constants `true` and `false` in these representations. We need to test the code to make sure our implementations of `true` and `false` are correct.

Before writing the full code, let us implement the DNF-to-CNF conversion for a short example:

```
(a && b) \mid\mid (c && d)      ---->      (a \mid\mid c) && (a \mid\mid d) && (b \mid\mid c) && (b \mid\mid d)  
// The same formulas are written in the set-based representation as:  
DNF(Set(Set(a, b), Set(c, d)))  ---->  CNF(Set(Set(a, c), Set(a, d), Set(b, c), Set(b, d)))
```

A simple expression implementing this transformation for nested sets could look like this:

```
for {  
  x <- Set("a", "b")  
  y <- Set("c", "d")  
} yield Set(x) ++ Set(y)  // The result is Set(Set(a, c), Set(a, d), Set(b, c), Set(b, d)).
```

Generalizing this code to an arbitrary number of nested sets, we get a first draft of a solution,

```
def toCNF[A](dnf: Set[Set[A]]): Set[Set[A]] = for {  
  x <- dnf.head          // The 'head' of a set; corresponds to Set("a", "b") in the example above.  
  ys <- toCNF(dnf.tail)  // The 'tail' of a set; corresponds to Set(Set("c", "d")) in the example.  
  // Converted to CNF, this corresponds to Set(Set("c"), Set("d")).  
} yield Set(x) ++ ys
```

This code is not yet fully correct: we have no base case for the recursion, and the `head` method will crash for an empty set. To make the pattern matching on a `Set` easier, let us implement a function

```
def headTailOption[A](xs: Set[A]): Option[(A, Set[A])] =      // 'List'-like pattern matching for Set.  
  if (xs.isEmpty) None else Some((xs.head, xs.tail))
```

With that function, we can write the final code as

```
def toCNF[A](dnf: DNF[A]): CNF[A] = headTailOption(dnf.s) match {  
  case None => falseCNF[A]          // Base case: The empty set Set() means 'false' in DNF.  
  case Some((head, tail)) => CNF(for {  
    x <- head                      // The first group of conjunctions.  
    ys <- toCNF(DNF(tail)).s        // The remaining groups of conjunctions, recursively converted to CNF.  
  } yield ys + x)                  // For sets, 'ys + x' is the same as 'Set(x) ++ ys'.  
}  
val dnf1 = DNF(Set(Set("a", "b"), Set("c", "d", "e")))  
val cnf1 = CNF(Set(Set("a", "c"), Set("a", "d"), Set("a", "e"), Set("b", "c"), Set("b", "d"),  
  Set("b", "e")))  
  
scala> toCNF(dnf1) == cnf1  
res0: Boolean = true  
  
scala> (toCNF(trueDNF) == trueCNF, toCNF(falseDNF) == falseCNF)  
res1: (Boolean, Boolean) = (true,true)
```

Because the rules of Boolean logic are completely symmetric with respect to swapping the operations \vee and \wedge , the function `cnf2dnf` can be implemented via `dnf2cnf` like this,

```
def toDNF[A](cnf: CNF[A]): DNF[A] = DNF(toCNF(DNF(cnf.s)).s)  
  
scala> (toDNF(trueCNF) == trueDNF, toDNF(falseCNF) == falseDNF)  
res2: (Boolean, Boolean) = (true,true)
```

If we test some more, we will find that the functions `toCNF` and `toDNF` are not inverses:

```
scala> toDNF(cnf1)      // Expected this to equal DNF(Set(Set("a", "b"), Set("c", "d", "e")) == dnf1.  
res3: DNF[String] = DNF(Set(Set(b, a), Set(b, a, c, e), Set(d, a, b), Set(b, e, a), Set(b, a, c),  
  Set(d, e, c, b), Set(d, e, c), Set(d, a, c, b), Set(e, a, b, c, d), Set(d, a, e, c), Set(d, a, e, b)))
```

However, we already verified that `cnf1 == toCNF(dnf1)`. We expect the formula `dnf1` to remain the same when we convert it to CNF and back to DNF. Why the discrepancy? The reason is that the transformations `toDNF` and `toCNF` fail to perform certain simplifications allowed in Boolean logic. For example, both `(a) || (a && b)` and `(a) && (a || b)` are equivalent to just `(a)`. In the set-based representations, we must be able to simplify `Set(Set("a"), Set("a", "b"))` to just `Set(Set("a"))`. Generally, we may discard any conjunction group that happens to be a superset of another conjunction group, and similarly for disjunction groups. After these simplifications, the formula `toDNF(cnf1)` becomes equal to `dnf1` because, as we can see, all inner sets in `toDNF(cnf1)` are supersets of either `Set("a", "b")` or `Set("c", "d", "e"))` or both. We can implement the simplifications as a function `simplifyCNF`:

```
def simplifyCNF[A](cnf: CNF[A]): CNF[A] = CNF(
  cnf.s.toIndexedSeq.sortBy(_.size).foldLeft(Set[Set[A]]()) { case (prevGroups, group) =>
    if (prevGroups.exists(_ subsetOf group)) prevGroups else prevGroups + group
  } // Omit 'group' if it happens to be a superset of some previous group within 'prevGroups'.
) // Since 'prevGroups' is sorted by group size, we only need to check later groups for supersets.

scala> simplifyCNF(CNF(Set(Set("a", "b"), Set("a", "b", "c", "d"))))
res3: CNF[String] = CNF(Set(Set(a, b)))
```

With this function, we may define more practically useful transformations `dnf2cnf` and `cnf2dnf` as

```
def dnf2cnf[A](dnf: DNF[A]): CNF[A] = simplifyCNF(toCNF(dnf))
def cnf2dnf[A](cnf: CNF[A]): DNF[A] = DNF(dnf2cnf(DNF(cnf.s)).s)
// Verify that dnf2cnf and cnf2dnf are inverses:
scala> (dnf2cnf(cnf2dnf(cnf1)) == cnf1, cnf2dnf(dnf2cnf(dnf1)) == dnf1)
res4: (Boolean, Boolean) = (true,true)
```

The new conversion functions `dnf2cnf` and `cnf2dnf` are inverses of each other.

Example 10.1.2.8* (matrix operations) If matrices are represented by nested sequences of numbers, matrix products can be calculated via nested functor blocks. The present task is to implement: (a) matrix transposition, (b) the vector-matrix dot product, and (c) the matrix-matrix dot product.

Solution (a) If a matrix has type `Seq[Seq[A]]`, a transposed matrix has the same type. An example:

```
val matrix = Seq(
  Seq(1, 2, 3),
  Seq(10, 20, 30)
)
/* Both matrices have type Seq[Seq[Int]]. */
```

```
val matrix_T = Seq(
  Seq(1, 10),
  Seq(2, 20),
  Seq(3, 30)
)
```

To compute the transposed matrix, we have to iterate over the initial matrix. A functor block of type `Seq` will return a value of type `Seq[Seq[A]]` only if the `yield`

expression itself returns a value of type `Seq[A]`. The example above shows that the first iteration should return `Seq(1, 10)`, which contains the first elements of all inner sequences from `matrix`. The second iteration should return all second elements, and so on. We see that we need to iterate over the indices of the matrix columns. Those indices are returned by the library method `indices`:

```
scala> Seq("Wilhelm Roentgen", "Henri Beckerel", "Marie Curie").indices
res0: scala.collection.immutable.Range = Range(0, 1, 2)
```

The value `Seq(1, 10)` can be computed by the functor block

```
for { row <- matrix } yield row(0)
```

This functor block needs to be written *inside* the `yield` value of the iterations of the indices:

```
scala> val matrix_T = for { index <- matrix.head.indices } // [0, 1, 2] are the column indices.
           yield { for { row <- matrix } // Iterate over the rows of the matrix.
                     yield row(index)
           }
matrix_T: IndexedSeq[Seq[Int]] = Vector(List(1, 10), List(2, 20), List(3, 30))
```

The two nested `for/yield` blocks represent two nested loops whose result value is again a nested sequence. We have implemented nested loops over column and row indices by nested functor blocks that use the column indices but directly iterate over rows.

(b) To see how a vector-matrix dot product works, consider this example:

$$\begin{vmatrix} a_0 & a_1 & a_2 \end{vmatrix} \cdot \begin{vmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \\ b_{20} & b_{21} \end{vmatrix} = \begin{vmatrix} a_0 \cdot b_{00} + a_1 \cdot b_{10} + a_2 \cdot b_{20} & a_0 \cdot b_{01} + a_1 \cdot b_{11} + a_2 \cdot b_{21} \end{vmatrix} .$$

The 2×3 matrix is represented as a sequence containing three nested sequences. We need to iterate over the first elements of the nested sequences (b_{00}, b_{10}, b_{20}) , multiply them with the corresponding elements (a_0, a_1, a_2) of the vector, and compute the sum of all products. The code for that is

```
Seq(b00, b10, b20).zip(Seq(a0, a1, a2)).map { case (x, y) => x * y }.sum
```

Then we need to do the same with the second elements (b_{01}, b_{11}, b_{21}) . To obtain the full code, we iterate over the indices of the nested sequences, as we did in part **(a)** for the transposition:

```
import scala.math.Numeric.Implicits.infixNumericOps
def vectorMatrixProduct[N: Numeric](vector: Seq[N], matrix: Seq[Seq[N]]): Seq[N] =
  for { index <- matrix.head.indices } yield {
    val b = for { row <- matrix } yield row(index)
    val pairs = for { (x, y) <- b.zip(vector) } yield x * y
    pairs.sum
  }

scala> vectorMatrixProduct(Seq(3,4,5), matrix_T)
res1: Seq[Int] = Vector(26, 260)
```

(c) The matrix-matrix dot product is defined as a matrix containing the results of the vector-matrix dot products, for all row vectors of the first matrix. For example,

$$\begin{vmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \end{vmatrix} \cdot \begin{vmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \\ b_{20} & b_{21} \end{vmatrix} = \begin{vmatrix} a_{00} \cdot b_{00} + a_{01} \cdot b_{10} + a_{02} \cdot b_{20} & a_{00} \cdot b_{01} + a_{01} \cdot b_{11} + a_{02} \cdot b_{21} \\ a_{10} \cdot b_{00} + a_{11} \cdot b_{10} + a_{12} \cdot b_{20} & a_{10} \cdot b_{01} + a_{11} \cdot b_{11} + a_{12} \cdot b_{21} \end{vmatrix} .$$

We reuse the solution of part **(b)** and write the code as

```
def matrixProduct[N: Numeric](matrix1: Seq[Seq[N]], matrix2: Seq[Seq[N]]): Seq[Seq[N]] =
  for { row1 <- matrix1 } yield vectorMatrixProduct(row1, matrix2)
```

The code shown in this example is for illustration only; for higher performance, matrix operations must be implemented through flat arrays rather than nested sequences.

Exercise 10.1.2.9 Implement the matrix-vector dot product as a function

```
def matrixVectorProduct[N: Numeric](matrix: Seq[Seq[N]], vector: Seq[N]): Seq[N] = ???
```

The result of computing the matrix-vector product is a column vector, for example:

$$\begin{vmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \end{vmatrix} \cdot \begin{vmatrix} b_{00} \\ b_{10} \\ b_{20} \end{vmatrix} = \begin{vmatrix} a_{00} \cdot b_{00} + a_{01} \cdot b_{10} + a_{02} \cdot b_{20} \\ a_{10} \cdot b_{00} + a_{11} \cdot b_{10} + a_{12} \cdot b_{20} \end{vmatrix} .$$

Exercise 10.1.2.10 The **trace** of a square matrix is the sum of its diagonal elements, e.g.:

$$\text{Tr} \begin{vmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{vmatrix} \triangleq a_{00} + a_{11} + a_{22} .$$

Implement the trace of a matrix (assuming a square matrix) as a function

```
def trace[N: Numeric](matrix: Seq[Seq[N]]): N = ???
```

10.1.3 Pass/fail monads

The type `Option[A]` can be viewed as a collection that can either be empty or hold a single value of type `A`. An “iteration” over such a collection will perform a computation at most once:

```
scala> for { x <- Some(123) } yield x * 2      // The computation is performed once.
res0: Option[Int] = Some(246)
```

When an `Option` value is empty, the computation is not performed at all.

```
scala> for { x <- None: Option[Int] } yield x * 2      // The computation is not performed at all.
res1: Option[Int] = None
```

What would a *nested* “iteration” over several `Option` values do? When all of the `Option` values are non-empty, the “iteration” will perform some computations using the wrapped values. However, if even one of the `Option` values happens to be empty, the computed result will be an empty value:

```
scala> for {
  x <- Some(123)
  y <- None
  z <- Some(-1)
} yield x + y + z
res2: Option[String] = None
```

Computations with `Either` and `Try` values follow the same logic: nested “iteration” will perform no computations unless all values are non-empty. This logic is useful for implementing a series of computations that could produce failures, where any failure should stop all further processing. For this reason (and since they all support the `pure` method and are lawful monads, as this chapter will show), we call the type constructors `Option`, `Either`, and `Try` the **pass/fail monads**.

The following schematic example illustrates this logic:

```
val result: Try[A] = for { // Computations in the 'Try' monad.
  x <- Try(k())
  y = f(x)                // First computation 'k()', may fail.
  if p(y)                  // No possibility of failure in this line.
  z <- Try(g(x, y))        // The entire expression will fail if 'p(y) == false'.
  r <- Try(h(x, y, z))    // The computation may also fail here.
} yield r                  // If 'r' has type 'A' then 'result' has type 'Try[A]'.
```

The function `Try()` catches exceptions thrown by its argument. If one of `k()`, `g(x, y)`, or `h(x, y, z)` throws an exception, the corresponding `Try(...)` value will evaluate to a `Failure(...)` case class, and further computations will not be performed. The value `result` will indicate the *first* encountered failure. Only if all `Try(...)` values evaluate to a `Success(...)` case class, the entire expression evaluates to a `result` of type `Success` that wraps a value of type `A`.

Whenever this pattern of computation is found, a functor block gives concise and readable code that replaces a series of nested `if/else` or `match/case` expressions. A typical situation was shown in Example 3.2.2.4 (Chapter 3), where a “safe integer” computation continues only as long as every result is a success; the chain of operations stops at the first failure. The code of Example 3.2.2.4 introduced custom data type with hand-coded methods such as `add`, `mul`, and `div`. We can now implement equivalent functionality using functor blocks and a standard type `Either[String, Int]`:

```
type Result = Either[String, Int]
def div(x: Int, y: Int): Result = if (y == 0) Left(s"error: $x / $y") else Right(x / y)
def sqrt(x: Int): Result = if (x < 0) Left(s"error: sqrt($x)") else Right(math.sqrt(x).toInt)
val previous: Result = Right(20) // Start with some given 'previous' value of type 'Result'.

scala> val result: Result = for { // Safe computation: 'sqrt(1000 / previous - 100) + 20'.
  x <- previous
  y <- div(1000, x)
```

```

z <- sqrt(y - 100)
} yield z + 1
result: Result = Left("error: sqrt(-50)")

```

The concise and readable code of `val result` replaces more verbose implementations such as

```

val result: Result = previous match {
  case Left(error)  => Left(error)
  case Right(x)     => div(1000, x) match {
    case Left(error)  => Left(error)
    case Right(y)     => sqrt(y - 100) match {
      case Left(error)  => Left(error)
      case Right(z)     => ... // More repetitive code of this sort.
    }
  }
}

```

The following examples illustrate the typical tasks where pass/fail monads are used. These tasks perform a linear sequence of computations that may fail; the first failure is then returned as a value.

Example 10.1.3.1: chaining computations with Option Some clients have placed some orders with some companies. The information is made available via Java system properties, for example:

```

System.setProperty("client 0", "company 2")
System.setProperty("client 1", "company 3")
System.setProperty("company 2", "order 4")
System.setProperty("company 3", "order 5")
System.setProperty("order 4", "123")
System.setProperty("order 5", "456")

```

Given a client's name, obtain the corresponding order quantity if it exists.

Solution The Java method `System.getProperty` returns the property value as `String` if the property is present, and otherwise returns `null`. Wrapping that return value into an `Option()` call, we replace null values by empty `Option` values (i.e., `None`). This makes the result values safe: using a `null` value may throw an exception, which will not happen when using `map` and `flatMap` methods on `Option` values. It remains to chain the computations:

```

def getOrderAmount(client: String): Option[Int] = for {
  company    <- Option(System.getProperty(client))
  order      <- Option(System.getProperty(company))
  stringValue <- Option(System.getProperty("orders"))
  intValue   <- Try(stringValue.toInt).toOption           // Non-integer string values are invalid.
} yield intValue

scala> getOrderAmount("client 1")
res0: Option[Int] = Some(123)

scala> getOrderAmount("client 2")
res1: Option[Int] = Some(456)

scala> getOrderAmount("client 3")
res2: Option[Int] = None

```

In the example just shown, `Option` values are sufficient since the absence of a property is not an error situation. Now we consider a task where we need to keep track of error information.

Example 10.1.3.2: chaining computations with Try Three given functions f , g , h all have Scala type `Int => Int` but may throw exceptions. Given an integer x , compute $f(g(h(x)))$ safely, reporting the first encountered error.

Solution Wrap each function into a `Try()` and chain the resulting computations:

```

def f(x: Int): Int = 1 / x
def g(x: Int): Int = 2 - x
def h(x: Int): Int = 2 / x

import scala.util.Try
def result(x: Int): Try[Int] = for {
  p <- Try(h(x))

```

```

q <- Try(g(p))
r <- Try(f(q))
} yield r

scala> result(1)
res0: Try[Int] = Failure(java.lang.ArithmetiException: / by zero)

```

The result value shows information about the failure generated in this computation.

Example 10.1.3.3: chaining with Future Scala library's `Future` class can be seen as a pass/fail monad because `Future(x)` will encapsulate any exception thrown while computing `x`. However, in addition to the pass/fail features, a `Future` value has a concurrency effect: the encapsulated computation `x` is scheduled to be run in parallel on another CPU thread. For this reason, `Future`'s methods (such as `map` and `flatMap`) require an implicit `ExecutionContext` argument, which provides access to threads.

As soon as a `Future` value is created, its computation is scheduled immediately. So, several `Future` values may run their computations in parallel. Nevertheless, computations chained via `flatMap` (or in a functor block) will run sequentially if new values need to wait for previous values:

```

import scala.concurrent.ExecutionContext.Implicits.global
def longComputation(x: Double): Future[Double] = Future { ... } // A long computation.

val result1: Future[Double] = for {
  p <- longComputation(10.0)
  q <- longComputation(p + 20.0)
  r <- longComputation(q - 20.0)
} yield p + q + r // Three 'longComputation' calls are running sequentially.

```

This code waits for the first `Future` that computes `p`, then creates a `Future` value that will eventually compute `q`, and finally creates a `Future` that will eventually compute `r`; only then the sum `p + q + r` may be obtained (wrapped in a `Future` constructor). This computation cannot run the three `longComputation(...)` calls in parallel, since each call depends on the result of the previous one.

Another possibility is that each `longComputation(...)` is independent of the results of the other computations. Then the three `Future` values may be created up front, and the functor block code represents three "long computations" running in parallel:

```

val long1 = longComputation(10.0)
val long2 = longComputation(50.0)
val long3 = longComputation(100.0)

val result2: Future[Double] = for {
  p <- long1
  q <- long2
  r <- long3
} yield p + q + r // Three 'longComputation()' calls are running in parallel.

```

10.1.4 Tree-like semimonads and monads

Tree-like type constructors are recursive types such as those described in Section 3.3. A typical example of a tree-like type constructor is the binary tree defined by the type equation

$$BT^A \triangleq A + BT^A \times BT^A.$$

Statement 6.2.3.6 shows that BT^A is a functor by replacing the right-hand side $A + BT^A \times BT^A$ by an arbitrary "structure bifunctor" S^{A,BT^A} and then showing that the recursive type constructor L^A defined by $L^A \triangleq S^{A,L^A}$ is a functor. As we will see, the type constructor L^A will be a semimonad or a monad with certain choices of $S^{*,*}$.

For lists, nested iteration goes over inner lists contained in an outer list. How does nested iteration work for a tree-shaped collection? An iteration over a tree enumerates the values at the *leaves* of a tree. So, a tree analog of nested iteration implies that each leaf of an outer tree contains an inner tree. A `flatMap` function must concatenate all nested trees into a single "flattened" tree.

Let us implement the `flatMap` method for the binary tree BT^* in that way. It is convenient to define an equivalent curried function (denoted by “flm”) with type signature

$$\text{flm}^{A,B} : (A \rightarrow BT^B) \rightarrow BT^A \rightarrow BT^B \quad .$$

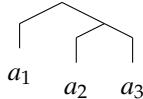
```

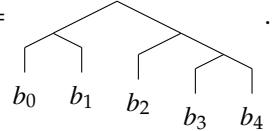
1  sealed trait BTTree[A]
2  final case class Leaf[A](x: A) extends BTTree[A]
3  final case class Branch[A](b1: BTTree[A], b2: BTTree[A]) extends BTTree[A]
4
5  def flm[A, B](f: A => BTTree[B]): BTTree[A] => BTTree[B] = {           // t.flatMap(f) == flm(f)(t)
6    case Leaf(x)          => f(x) // Here f(x) has type BTTree[B], which could be a Leaf or a Branch.
7    case Branch(b1, b2)   => Branch(flm(f)(b1), flm(f)(b2))           // Recursive calls of 'flm'.
8  }

```

The same implementation is written in the code notation as

$$\text{flm}^{A,B}(f: A \rightarrow BT^B) \triangleq \begin{array}{c|c|c} & B + BT^B \times BT^B & \\ \hline A & f & \\ \hline BT^A \times BT^A & b_1 \times b_2 \rightarrow \mathbb{0}^B + \overline{\text{flm}}(f)(b_1) \times \overline{\text{flm}}(f)(b_2) & \end{array} .$$

To visualize how the `flatMap` method operates on binary trees, let us compute `tree1.flatMap(f)`, where we take `tree1` =  and a function $f : A \rightarrow BT^B$ that has $f(a_1) = \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array}$, $f(a_2) = b_2$, $f(a_3) = b_3$

and $f(a_3) = \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array}$. (Here a_i for $i = 1, 2, 3$ are some values of type A and b_i for $i = 0, \dots, 4$ are some values of type B .) Evaluating the code of `flatMap`, we find `tree1.flatMap(f)` = .

So, we see that `flatMap` works by grafting a subtree into every `Leaf` of a given tree: A leaf is replaced by a new tree in line 6 in the code of `flatMap`. That code can be generalized to the recursive type PT^A (representing a “tree with P -shaped branches”) defined by

$$PT^A \triangleq A + P^{PT^A} \quad ,$$

for any given functor P . The disjunctive part $A + \mathbb{0}$ is replaced by a new tree:

```

sealed abstract class PTree[P[_]] : Functor, A] // Need an 'abstract class' due to implicits.
final case class Leaf[P[_] : Functor, A](x: A) extends PTree[P, A]
final case class Branch[P[_] : Functor, A](p: P[PTree[P, A]]) extends PTree[P, A]

def flm[P[_]: Functor, A, B](f: A => PTree[P, B]): PTree[P, A] => PTree[P, B] = {
  case Leaf(x)          => f(x) // Here f(x) has type PTree, which could be a Leaf or a Branch.
  case Branch(p)        => Branch(p.map(t => flm(f)(t))) // Conceptually, Branch(p.map(flm(f))).
}

```

The same function is written in the code notation as

$$\text{flm}^{A,B}(f: A \rightarrow P^{PT^B}) \triangleq \begin{array}{c|c|c} & B + P^{PT^B} & \\ \hline A & f & \\ \hline P^{PT^A} & (\overline{\text{flm}}(f))^{\uparrow P} & \end{array} .$$

We can also implement `flatMap` for more general type constructors L defined by $L^A \triangleq P^A + P^{L^A}$ for some functor P . Such L^A can be visualized as trees with P -shaped branches and P -shaped leaves.

```

sealed abstract class PLPT[P[_] : Functor, A]
final case class Leaf[P[_] : Functor, A](px: P[A]) extends PLPT[P, A]
final case class Branch[P[_] : Functor, A](pb: P[PLPT[P, A]]) extends PLPT[P, A]

def flm[P[_]: Functor, A, B](f: A => PLPT[P, B]): PLPT[P, A] => PLPT[P, B] = {
  case Leaf(px)      => Branch(px.map(f)) // Here px.map(f) has type P[PLPT[P, B]].
  case Branch(pb)    => Branch(pb.map(t => flm(f)(t)))
}

```

$$\text{flm}^{A,B} \left(f: A \rightarrow P^{\text{PLPT}^B} \right) \triangleq \begin{array}{|c|c|c|} \hline & P^B & P^{\text{PLPT}^B} \\ \hline P^A & \emptyset & f^{\uparrow P} \\ \hline P^{\text{PLPT}^A} & \emptyset & (\overline{\text{flm}}(f))^{\uparrow P} \\ \hline \end{array} .$$

The code of `flatMap` for the `PLPT` tree never creates any leaves, only branches. We will see later how this prevents `PLPT` from being a full monad (it is only a semimonad; no `pure` method can satisfy the required laws). Nevertheless, having a lawful `flatMap` is sufficient for using `PLPT` in functor blocks.

The following examples show some use cases for tree-like monads.

Example 10.1.4.1 Implement the `flatMap` operation for a tree of configuration properties of the form

```

url: http://server:8000
users:
  user:
    name: abcde
    pass: 12345
  user2:
    name: fghij
    pass: 67890

```

Write a function to convert trees into strings in this format.

Solution The code for this data structure must support any number of simple properties or branches with `String` labels. A suitable structure is a tree with P -shaped leaves and P -shaped branches, where the functor P is defined as $P^A \triangleq \text{List}^{\text{String} \times A}$. Implement the tree type:

```

sealed trait PropTree[A] // Introduce the type parameter A for the values of properties.
final case class Simple[A](props: List[(String, A)]) extends PropTree[A]
final case class Branches[A](branches: List[(String, PropTree[A])]) extends PropTree[A]

```

To pretty-print trees of this type, we need to keep track of the current level of indentation:

```

def prettyPrint[A](pt: PropTree[A], indent: String = "")(toString: A => String): String = (pt match {
  case Simple(props) => props.map { case (name, a) => indent + name + ": " + toString(a) }
  case Branches(brs) => brs.map { case (name, ps) => indent + name + "\n" +
    prettyPrint(ps, indent + "  ")(toString) }
}).mkString("\n") + "\n"

```

For convenience, the methods `flatMap` and `toString` can be defined directly on the trait,

```

sealed trait PropTree[A] {
  def flatMap[B](f: A => PropTree[B]): PropTree[B] = this match {
    case Simple(props) => Branches(props.map { case (name, a) => (name, f(a)) })
    case Branches(brs) => Branches(brs.map { case (name, ps) => (name, ps.flatMap(f)) })
  }
  override def toString: String = prettyPrint(this)_.toString
}

```

The following code illustrates the `flatMap` operation by replacing all integer leaf values below 100 by a simple property and other values by a branch of properties:

```

val pt1: PropTree[Int] = Simple(List("quantity1" -> 50, "quantity2" -> 250))

scala> println(pt1.toString)
quantity1: 50
quantity2: 250

```

```

val pt2 = pt1.flatMap { x => if (x < 100) Simple(List("value" -> x)) else Branches(List(
  "large" -> Simple(List("value" -> x / 100, "factor" -> 100)),
  "small" -> Simple(List("value" -> x % 100, "factor" -> 1)))
) }

scala> println(pt2.toString)
quantity1:
  value: 50
quantity2:
  large:
    value: 2
    factor: 100
  small:
    value: 50
    factor: 1

```

Example 10.1.4.2 Implement variable substitution for a simple arithmetic language. Example use:

```

val expr1: Term[String] = Const(123) * Var("a") + Const(456) * Var("b")
val expr2: Term[String] = Const(20) * Var("c")

scala> expr1.flatMap { x => if (x == "a") expr2 else Var(x) } // Substitute "a" = expr2 in expr1.
res0: Term[String] = Plus(Mult(Const(123), Mult(Const(20), Var("c"))), Mult(Const(456), Var("b")))

```

Solution Begin by implementing the basic functionality of the language: constants, variables, addition, and multiplication. The type parameter A in $\text{Term}[A]$ is the type of labels for variables.

```

sealed trait Term[A] {
  def +(other: Term[A]): Term[A] = Plus(this, other)
  def *(other: Term[A]): Term[A] = Mult(this, other)
  def map[B](f: A => B): Term[B] = ???
  def flatMap[B](f: A => Term[B]): Term[B] = ???
}
final case class Const[A](value: Int) extends Term[A]
final case class Var[A](name: A) extends Term[A]
final case class Plus[A](t1: Term[A], t2: Term[A]) extends Term[A]
final case class Mult[A](t1: Term[A], t2: Term[A]) extends Term[A]

```

The type constructor Term is an example of an abstract syntax tree and can be equivalently defined as

$$\text{Term}^A \triangleq L^A + S^{\text{Term}^A} , \quad L^A \triangleq \text{Int} + A , \quad S^B \triangleq B \times B + B \times B ,$$

where the functors L and S describe the structure of the leaves and the branches of the tree.

The code of the `map` method is

```

def map[B](f: A => B): Term[B] = this match { // This code must be within 'trait Term[A]'.
  case Const(value) => Const(value)
  case Var(name) => Var(f(name))
  case Plus(t1, t2) => Plus(t1 map f, t2 map f)
  case Mult(t1, t2) => Mult(t1 map f, t2 map f)
}

```

The code of `flatMap` replaces variables by new trees, leaving everything else unchanged:

```

def flatMap[B](f: A => Term[B]): Term[B] = this match { // This code must be within 'trait Term[A]'.
  case Const(value) => Const(value)
  case Var(name) => f(name)
  case Plus(t1, t2) => Plus(t1 flatMap f, t2 flatMap f)
  case Mult(t1, t2) => Mult(t1 flatMap f, t2 flatMap f)
}

```

Note that the `flatMap` functions for lists, pass/fail monads, and tree-like monads are information-preserving: no data is discarded from the original tree or from any computed results.

Example 10.1.4.3 Show that a regular-shaped tree *cannot* have an information-preserving `flatMap`.

Solution The type constructor R^A for a regular-shaped binary tree is defined by $R^A \triangleq A + R^{A \times A}$. We can (non-rigorously) view the type R^A as a disjunction with “infinitely many” parts,

$$R^A = A + A \times A + A \times A \times A \times A + \dots ,$$

where the n^{th} part ($n = 1, 2, \dots$) is a product of 2^n copies of A . The `flatMap` function must have the type

$$\text{flm} : (A \rightarrow R^B) \rightarrow R^A \rightarrow R^B .$$

A possible value r of type R^A is, say, a product of 4 copies of A . How could `flatMap` apply to that value? For any function $f : A \rightarrow R^B$, we must compute a result value $\text{flm}(f)(r)$ of type R^B . The only way of obtaining a value of type R^B is to apply f to some value of type A . But the given argument r contains 4 such values. If we apply f to each one of them, we will obtain 4 different values of type R^B . Some of these values may correspond to the part $B \times B$ in the disjunction, others to $B \times B \times B \times B$, etc. The total number of values of type B depends on the result of that computation and is not necessarily equal to a power of 2. Then it will be impossible to accommodate all of those values of type B within a *single* regular-shaped tree of type R^B . If we discard some values of type B so that the rest fits into the regular-shaped tree R^B , we will lose information. So, it is not possible to implement an information-preserving `flatMap` for a regular-shaped binary tree. (It is not a monad.)

10.1.5 The Reader monad

This chapter started with the list-like monads whose `flatMap` method is motivated by the requirements of nested iteration. It turns out that the `flatMap` method can be generalized to many other type constructors that are useful for various programming tasks not limited to nested iteration.

In general, a (semi)monad L^A is no longer a collection of data items of type A . Instead, L^A can be regarded (informally) as a value of type A wrapped in a special “computational effect”. We view “computations with an L -effect” as functions of type $A \rightarrow L^B$ (as in `flatMap`’s argument type). In this view, different monads — such as list-like, pass/fail, or tree-like — implement different kinds of effects. An ordinary function of type $A \rightarrow B$ is a computation with a “trivial effect”.

In this sense, monadic effects are *not* side effects. Functions of type $A \rightarrow L^B$ can be referentially transparent and have value semantics. Informally, an “ L -effect” describes everything that a function of type $A \rightarrow L^B$ computes other than a value of type B . To make the vague idea of “effect” concrete, we write the required type of the computation in the form $A \rightarrow L^B$ with a specific type constructor L . In this and the next subsections, we will look at some monads that can be derived in that approach: the `Reader`, `Writer`, `Eval`, `State`, and `Cont` (“continuation”) monads.

The first of those, the `Reader`, corresponds to a function that computes a result while using some additional data of a fixed type Z . A function consuming (“reading”) that additional data will have type $A \times Z \rightarrow B$ instead of $A \rightarrow B$. It remains to rewrite the type $A \times Z \rightarrow B$ in the form $A \rightarrow L^B$ with a suitable choice of a type constructor L^\bullet . By currying, we obtain an equivalent type

$$(A \times Z \rightarrow B) \cong (A \rightarrow Z \rightarrow B) ,$$

which has the form $A \rightarrow L^B$ if we define $L^A \triangleq Z \rightarrow A$. This type constructor is called the `Reader` monad and is denoted by $\text{Read}^{Z,A} \triangleq Z \rightarrow A$. The Scala definition is `type Reader[Z, A] = Z => A`.

Fully parametric implementations of `map` and `flatMap` directly follow from their type signatures:

```
def map[A, B](r: Z => A)(f: A => B): (Z => B) = r andThen f // Example 5.4.1.7.
def flatMap[A, B](r: Z => A)(f: A => Z => B): (Z => B) = { z => f(r(z))(z) } // Exercise 5.4.2.10(c).
```

What are the use cases for `Reader`? If we write a functor block with a value of type `Reader[Z, A]` as a “source” value, it follows that other source values must have type `Reader[Z, B]`, `Reader[Z, C]`, and so on. In other words, the type Z must be fixed for the entire functor block. So, data of type Z plays the role of a *common dependency* that may be used by all computations within the functor block.

As an example, imagine a program built up by composing several “procedures”. Each “procedure” takes an argument and returns a value, like an ordinary function, but also may need to run a Unix shell command. For simplicity, assume that the interface to shell commands is a function that runs a new command with given input, capturing the command’s entire output. The type of that function may be $\text{String} \times \text{String} \rightarrow \text{Int} \times \text{String}$, taking the command string and the input for the command, and returning the command’s exit code and the output.

```
type RunSh = (String, String) => (Int, String)
```

A simple implementation (that does not handle any run-time exceptions) is

```
import sys.process_
val runSh: RunSh = { (command, input) =>
  var result: Array[Char] = Array()
  val p: Process = command.run(new ProcessIO(
    { os => os.write(input.getBytes); os.close() },
    { is => result = scala.io.Source.fromInputStream(is).toArray; is.close() },
    _.close()
  ))
  val exitCode = p.exitValue()
  (exitCode, new String(result))
}
```

We can now use this “shell runner” to execute some standard Unix commands:

```
scala> runSh("echo -n abcd", "") // Use '-n' to avoid trailing newlines.
res0: (Int, String) = (0, "abcd")

scala> runSh("cat", "xyz")._2 // Equivalent to 'cat < $(echo -n xyz)'.
res1: String = "xyz"
```

Consider a program that determines the total line count of all files under a given directory, but only including files whose names match given patterns. We may split the computation into three “procedures”: 1) list all files under the given directory, 2) filter out files whose names do not match the given patterns, 3) determine the line count for each of the remaining files. For the purposes of this example, we will implement each stage as a function that uses `runSh` to run a shell command.

```
def listFiles(runSh: RunSh, dir: String): String = runSh(s"find $dir -type f", "")._2
def filterFiles(runSh: RunSh, files: String, patterns: String): String =
  runSh(s"grep -f $patterns", files)._2
def lineCounts(runSh: RunSh, files: String): Array[Int] = files.split("\n") // Array of file names.
  .map { file => runSh("wc -l $file", "")._2.replaceAll("^ +", "").split(" ")(0).toInt }
```

This code assumes that file names do not contain the newline character `\n`. Use this code only as an illustration of a use case for the `Reader` monad.

We can now write the program like this:

```
def getLineCount(runSh: RunSh, dir: String, patterns: String): Int = {
  val fileList = listFiles(runSh, dir)
  val filtered = filterFiles(runSh, fileList, patterns)
  val counts = lineCounts(runSh, filtered)
  counts.sum
}
```

The value `runSh` is a common dependency of all the “procedures” and is repeated throughout the code. This repetition becomes a problem if we have many “procedures” within different code modules; or when different “runners” of type `RunSh` are used within the program. For instance, one “runner” executes commands on a remote machine, while another “runner” is used for testing and returns fixed results without running any shell commands. How can we avoid writing repetitive code and at the same time assure that the correct “runners” are passed to all the “procedures”?

The `Reader` monad offers a solution: it allows us to combine smaller “procedures” into larger ones while passing the “runner” values automatically. We first need to convert all “procedures” into functions of type `A => Reader[RunSh, B]` with suitable choices of type parameters `A, B`:

```

type Reader[Z, A] = Z => A
def listFilesR(dir: String): Reader[RunSh, String] = runSh => runSh(s"find $dir -type f", "")._2
def filterFilesR(patterns: String): String => Reader[RunSh, String] = files => runSh =>
  runSh(s"grep -f $patterns", files)._2
def lineCountsR(files: String): Reader[RunSh, Array[Int]] = runSh => files.split("\n")
  .map { file => runSh("wc -l $file", "")._2.replaceAll("^\n", "").split(" ")_(0).toInt }

```

This allows us to express `getLineCount` as a combination of the three “procedures” by using the `Reader` monad’s `flatMap` function (for convenience, assume that we defined an extension method `flatMap`):

```

def getLineCount(dir: String, patterns: String): Reader[RunSh, Int] = listFilesR(dir)
  .flatMap(files => filterFilesR(patterns)(files))
  .flatMap(lineCountsR).map(_.sum) // Assuming an extension method 'map' is defined for 'Reader'.

```

For better readability, rewrite this code equivalently in the functor block syntax:

```

def getLineCountR(dir: String, patterns: String): Reader[RunSh, Int] = for {
  files <- listFilesR(dir)
  filtered <- filterFilesR(patterns)(files)
  lineCounts <- lineCountsR(filtered)
} yield lineCounts.sum

val program: Reader[RunSh, Int] = getLineCountR(".", "patterns.txt")

```

We obtained a value `program` of type `Reader[RunSh, Int]`, which is a function type. It is important to note that at this point no shell commands have been run yet. We merely packaged all the necessary actions into a function value (a **monadic program**). We now need to apply that function to a “runner” value of type `RunSh`. Only then we will obtain the actual line count:

```
val count: Int = program(runSh)
```

The `Reader` monad allows us to split the code into two stages: first, we build a monadic program (a value of type `Reader`) from smaller “procedures”. Monadic programs are ordinary values that can be passed as arguments to functions, stored in arrays, etc. We can compose smaller monadic programs into larger ones using `map`, `flatMap`, or `for/yield` blocks. We have the full flexibility of manipulating those monadic values and combining them into larger monadic programs in any order, since no shell commands are being run while we perform these manipulations. When we are done building the full monadic program, we can “run” it using a chosen runner.

Since the runner is the common dependency of all `Reader`-monadic programs, running a monadic program means performing **dependency injection**. At the point of running a `Reader` program, we have the full flexibility of choosing the value of the dependency. The code guarantees that the dependency will be passed to each individual part of the monadic program.

Implicit values for dependency injection Scala’s implicit argument feature allows us to solve the problem of dependency injection in a different way. Instead of converting all code to use the `Reader` monad, we convert all code to use an implicit argument for the common dependency:

```

def listFilesIm(dir: String)(implicit runSh: RunSh): String = ...
def filterFilesIm(patterns: String)(files: String)(implicit runSh: RunSh): String = ...
def lineCountsIm(files: String)(implicit runSh: RunSh): Array[Int] = ...

def getLineCountIm(dir: String, patterns: String)(implicit runSh: RunSh): Int = {
  val fileList = listFilesIm(dir)
  val filtered = filterFilesIm(fileList, patterns)
  val counts = lineCountsIm(filtered)
  counts.sum
}

```

Compare this code and the code that uses the `Reader` monad: all functions are curried, and their type signatures are exactly the same up to the `implicit` keyword in the last argument. Scala’s implicit arguments reproduce the code style of the `Reader` monad in a more idiomatic way (especially with

Scala 3's implicit function types⁵).

10.1.6 The Writer monad

The `Writer` monad represents a function of type $A \rightarrow B$ that returns its result (of type B) and additionally outputs some information (say, logging data) about the computation just performed. Let W be the type of the logging data. To model this situation, we then need a function $f : A \rightarrow B$ and additionally a function $g : A \rightarrow W$ that computes the logging output. So, we can define the type of a “computation with the `Writer` effect” rigorously as the product of the two function types,

$$(A \rightarrow B) \times (A \rightarrow W) \cong A \rightarrow B \times W \quad .$$

Since this type must be of the form $A \rightarrow L^B$, we must define the functor L as $L^A \triangleq A \times W$. This is the type constructor of the `Writer` monad, denoted by $\text{Writer}^{A,W} \triangleq A \times W$.

If several computations are performed one after another in the `Writer` monad, the logging information should be “accumulated” in some way. In the logging example, additional lines are simply appended to the log file. It means that we must be able somehow to combine several values of type W into one. A general way of doing that is to require W to be a semigroup (see Example 8.2.5.4) with a binary operation \oplus . We can then implement the `flatMap` method for `Writer` like this:

```
final case class Writer[A, W: Semigroup](a: A, log: W) {
  def flatMap[B](f: A => Writer[B, W]): Writer[B, W] = {
    val Writer(b, newLog) = f(a) // Pattern-match to destructure the value f(a).
    Writer(b, log |+| newLog) // Use the semigroup operation |+|.
  }
}
```

The logging type W is often a monoid (a semigroup with an “empty” value). If so, `Writer[A, W]` will be a full monad whose `pure` method is implemented as

```
def pure[A, W: Monoid]: A => (A, W) = a => (a, Monoid[W].empty)
```

When W is a semigroup but not a monoid, `Writer[A, W]` will be a semimonad but not a monad.

An example of using a `Writer` semimonad is logging with timestamps where we need to keep track of the earliest and the latest timestamp. Define the type $W \triangleq \text{Logs}$ and a semigroup operation $|+|$ by

```
final case class Logs(begin: LocalDateTime, end: LocalDateTime, message: String) {
  def |+|(other: Logs): Logs = Logs(begin, other.end, message + "\n" + other.message)
} // For simplicity, we assume that timestamps will be monotonically increasing.
```

The type `Logs` is not a monoid because its binary operation discards some of the input data, so we cannot define an “empty” value satisfying the identity laws (see Eq. (8.2) in Example 8.2.5.6).

We can now use the semimonad `Writer[A, Logs]`. Here are some example computations:

```
type Logged[A] = Writer[A, Logs]
def log[A](message: String)(x: A): Logged[A] = {           // Define this function for convenience.
  val timestamp = LocalDateTime.now
  new Logged(x, Logs(timestamp, timestamp, message))
}
def compute[A](x: => A): A = { Thread.sleep(100L); x }           // Simulate a long computation.

scala> val result: Logged[Double] = for {
  x <- log("begin with 3")(compute(3))           // The initial source type is 'Logged[Int]'.
  y <- log("add 1")(compute(x + 1))
  z <- log("multiply by 2.0")(compute(y * 2.0)) // The type of result becomes 'Logged[Double]'.
} yield z // The computation should take between 300 and 400 ms.
res0: Logged[Double] = Writer(8.0,Logs(2020-02-15T22:02:42.313,2020-02-15T22:02:42.484,begin with 3
add 1
multiply by 2.0))
```

Unlike the `Reader` monad, which delays all computations until a “runner” is called, a monadic value of type `Writer[A, W]` already contains the final computed values of types `A` and `W`.

⁵See <https://www.scala-lang.org/blog/2016/12/07/implicit-function-types.html>

10.1.7 The State monad

Heuristically, the `Reader` monad $\text{Read}^{Z,A}$ is able to “read” values of type Z , while the `Writer` monad $\text{Writer}^{A,W}$ may “write” values of type W , in addition to computing their result of type A . The `State` monad, denoted by $\text{State}^{S,A}$, combines the functionality of `Reader` and `Writer` in a special way: the extra value (of type S) is updated and automatically passed from one computation to the next.

To derive the required type constructor, consider a computation of type $A \rightarrow B$ that additionally needs to read and to write a value of type S . Since the total input is a pair of A and S , and the total output is a pair of B and S , this kind of computation is represented by a function of type $A \times S \rightarrow B \times S$. We now try to rewrite this type in the form $A \rightarrow L^B$ with a suitable type constructor L . It is clear that we need to curry the argument A . The result is

$$(A \times S \rightarrow B \times S) \cong (A \rightarrow S \rightarrow B \times S) = A \rightarrow L^B \quad , \quad \text{where} \quad L^A \triangleq S \rightarrow A \times S \quad .$$

So, the `State` monad must be defined by the type constructor $\text{State}^{S,A} \triangleq S \rightarrow A \times S$. This is a function that computes a value of type A while using and possibly updating the “state value” of type S .

The code of the `flatMap` method for this type constructor was derived in Example 5.4.1.11(c). It does indeed pass the updated state value to the next computation:

```
type State[S, A] = S => (A, S)
def flatMap[S, A, B](prev: State[S, A])(f: A => State[S, B]): State[S, B] = { s =>
  val (a, newState) = prev(s) // Compute result of type 'A', updating the state.
  f(a)(newState)           // Pass the updated state to the next computation.
}
```

An example of using the `State` monad is the task of implementing a random number generator. A simple generator is the **Lehmer algorithm**,⁶ which generates integer sequences x_n defined by

$$x_{n+1} \triangleq (48271 * x_n) \% (2^{31} - 1) \quad , \quad 1 \leq x_n \leq 2^{31} - 2 \quad , \quad n = 0, 1, 2, \dots$$

The “updating” function for this sequence, $x_{n+1} = \text{lehmer}(x_n)$, can be implemented as

```
def lehmer(x: Long): Long = x * 48271L \% ((1L << 31) - 1)
```

In many applications, one needs uniformly distributed floating-point numbers in the interval $[0, 1]$. To produce such numbers, let us define a helper function:

```
def uniform(x: Long): Double = (x - 1).toDouble / ((1L << 31) - 3) // Enforce the interval [0, 1].
```

To use the uniform generator, we need to provide an initial value x_0 (the “seed”) and then call the function `lehmer` repeatedly on successive values. The code would look like this:

```
val s0 = 123456789L // A "seed" value.
val s1 = lehmer(s0)
val r1 = uniform(s1)
... // Use pseudo-random value r1.
val s2 = lehmer(s1)
val r2 = uniform(s2)
... // Use pseudo-random value r2.
val s3 = lehmer(s2) // And so on.
```

We need to keep track of the generator’s state values s_1, s_2, \dots , that are not directly needed for other computations. This “bookkeeping” is error-prone since we might reuse a previous generator state by mistake. The `State` monad keeps track of the updated state values automatically and correctly. This comes at a cost: we need to convert all computations into `State`-typed monadic programs.

As a simple example, consider the task of generating uniformly distributed floating-point numbers in the interval $[0, 1]$. We need to maintain the generator state while computing the result. The floating-point generator is implemented as a monadic value of type `State[Long, Double]`:

```
val rngUniform: State[Long, Double] = { oldState =>
  val result = uniform(oldState) // Enforce the interval [0, 1].
  val newState = lehmer(oldState)
  (result, newState)
}
```

⁶See https://en.wikipedia.org/wiki/Lehmer_random_number_generator

Code using `rngUniform` will be of the monadic type `State[Long, A]` for some `A`:

```
val program: State[Long, String] = for { // Assume flatMap and map methods are defined for State.
  r1 <- rngUniform
  ... // Use pseudo-random value r1. The internal state of rngUniform is maintained automatically.
  r2 <- rngUniform
  ... // Use pseudo-random value r2.
} yield s"Pair is $r1, $r2" // Compute result of type String.
```

Monadic programs of this type can be composed in arbitrary ways. The “bookkeeping” of the state values is safely handled by the `State` monad and hidden from the programmer’s view. When the entire monadic program has been composed, it needs to be “run” to extract its result value. Since the type `State[S, A]` is a function with argument of type `S`, the runner is just an application of that function to an initial value of type `S`, specifying the initial state:

```
val seed = 123456789L      // Initial state of the generator.

scala> program(seed)      // Run this monadic program.
res0: (String, Long) = ("Pair is 0.028744523433557146, 0.5269012540999576", 216621204L)
```

10.1.8 The eager/lazy evaluation monad (Eval)

The monads `Reader`, `Writer`, and `State` manage extra information about computations. Those monads’ effects can be viewed as working with an extra value of a certain fixed type. We now turn to monads whose effects are not values but nonstandard evaluation strategies.

The first of these monads is called `Eval`, and its task is to encapsulate lazy and eager evaluations into a single type. A value of type `Eval[A]` can be eager (available now) or lazy (available later). Values of these sub-types can be combined with correct logic: for instance, a combination of eager and lazy values automatically becomes lazy.

To derive the type constructor, note that lazy values of type `A` are equivalent to functions of type $\mathbb{1} \rightarrow A$. So, the disjunctive type $\text{Eval}^A \triangleq A + (\mathbb{1} \rightarrow A)$ represents a value that is either lazy or eager:

```
sealed trait Eval[A]
final case class Eager[A](x: A)           extends Eval[A]
final case class Lazy[A](lazyX: Unit => A) extends Eval[A]
```

It is useful to have functions converting between eager and lazy values whenever needed:

```
def get: Eval[A] => A = {
  case Eager(x)    => x
  case Lazy(lazyX) => lazyX(())
}

def now(x: A): Eval[A] = Eager(x)
def later(e: => A): Eval[A] = Lazy(_ => e)
```

Now we can implement a `flatMap` method that correctly keeps track of the evaluation strategy:

```
def flatMap[A, B](f: A => Eval[B]): Eval[A] => Eval[B] = {
  case Eager(x)    => f(x)           // This value can be eager or lazy, according to f(x).
  case Lazy(lazyX) => Lazy(_ => get(f(lazyX(())))) // Call 'get' to avoid nested Lazy().
}
```

Assuming that `map` and `flatMap` are defined as extension methods for `Eval`, we may use functor blocks to combine eager and lazy computations freely:

```
val result: Eval[Int] = for {
  x <- later(longComputation1()) // Delay the long computation.
  y <- now(x + 2)              // Short computation, no need to delay.
  z <- later(longComputation2(y * 100))
} yield z
```

value of type `Int` out of `result`, we need to evaluate `get(result)`, which will take a longer time.

The value of `result` is a lazy computation because it involves lazy steps. It is quick to compute `result` because the long computations are not yet started. To extract the final

10.1.9 The continuation monad (Cont)

The continuation monad is another monadic design pattern that involves a non-standard evaluation strategy called the **continuation-passing** programming style. In that style, functions do not return their results directly but instead call an auxiliary function that consumes the result. The auxiliary function is called a “continuation” or a “callback”, and is passed as an additional argument.

To compare the direct style with the continuation-passing style, consider an arithmetic calculation:

```
def add3(x: Int): Int = x + 3
def mult4(x: Int): Int = x * 4
val result = add3(mult4(10)) // result = 43
```

Now we add callback arguments to each function and rewrite the code as

```
def add3(x: Int)(callback: Int => Unit): Unit = callback(x + 3)
def mult4(x: Int)(callback: Int => Unit): Unit = callback(x * 4)
def result(callback: Int => Unit): Unit = mult4(10)(r => add3(r)(callback))
```

To make the pattern more clear, replace the constant 10 by a function `pure` with a callback argument:

```
def pure(x: Int)(callback: Int => Unit): Unit =
  callback(x)

def result(callback: Int => Unit): Unit =
  pure(10) { x =>
    mult4(r1) { y =>
      add3(r2) { z =>
        callback(z)
      }
    }
  }
```

This code is a typical pattern of continuation-passing style. The final result of the calculation is only available as the bound variable `z` in a deeply nested function scope. This makes working with this code style more difficult: All subsequent code that needs to use the value `z` needs to be written either directly within that deeply nested scope, or must be contained within the given `callback`. This is the heuristic reason for the name “continuation”.

Another feature of the continuation-passing style is that callbacks could be called later, while the main thread of computation continues running concurrently. To show an example, redefine

```
def add3(x: Int)(callback: Int => Unit): Unit = { Future(callback(x + 3)); () }
def mult4(x: Int)(callback: Int => Unit): Unit = { Future(callback(x * 4)); () }
```

The new code schedules the calls to `callback` on separate threads. But the type signatures of `add3` and `mult4` hide this fact: they just return `Unit`. So, the code of `def result(...)` remains unchanged.

The type signature of `result` is $(\text{Int} \rightarrow \mathbb{1}) \rightarrow \mathbb{1}$, which shows that it consumes a callback (of type $\text{Int} \rightarrow \mathbb{1}$). A function that consumes a callback is at liberty to call the callback later, and to call it several times or not at all. The type $(\text{Int} \rightarrow \mathbb{1}) \rightarrow \mathbb{1}$ does not show whether the callback will be called; it merely “registers” the callback for possible later use. This gives us flexibility in the execution strategy, at the cost of making the code more complicated to write and to understand.

One complication is that it is difficult to maintain code that contains deeply nested function scopes. The continuation monad solves this problem by converting nested function scopes into more easily composable functor blocks. The code also becomes more readable.

To derive the required type constructor, consider a computation of type $A \rightarrow B$ that needs to use the continuation-passing style. Instead of returning a value of type B , it registers a callback function for possible later use and returns a `Unit` value. If the callback has type $B \rightarrow \mathbb{1}$, the total input of our computation is $A \times (B \rightarrow \mathbb{1})$ while the output is simply $\mathbb{1}$. So, the type of a continuation-passing computation is $A \times (B \rightarrow \mathbb{1}) \rightarrow \mathbb{1}$. Rewrite this type in the form $A \rightarrow L^B$ with a suitable functor L :

$$(A \times (B \rightarrow \mathbb{1}) \rightarrow \mathbb{1}) \cong (A \rightarrow (B \rightarrow \mathbb{1}) \rightarrow \mathbb{1}) = A \rightarrow L^B \quad , \quad \text{where} \quad L^A \triangleq (A \rightarrow \mathbb{1}) \rightarrow \mathbb{1} \quad .$$

It is sometimes helpful if the callback returns a more informative value than `Unit`. For instance, that value could show error information or give access to processes that were scheduled concurrently.

So, we generalize the type constructor L to $(A \rightarrow R) \rightarrow R$, where R is a fixed “result” type. This type constructor is called the **continuation monad** and is denoted by $\text{Cont}^{R,A} \triangleq (A \rightarrow R) \rightarrow R$.

How does the continuation monad make callback-based code composable? The answer is in the code of the `flatMap` method. Its implementation is (Exercise 5.4.2.11)

```
type Cont[R, A] = (A => R) => R
def flatMap[R, A, B](ca: Cont[R, A])(f: A => Cont[R, B]): Cont[R, B] = { br => ca(a => f(a)(br)) }
```

The code of `flatMap` substitutes the new callback, `br: B => R`, into the innermost scope of the computation `f`. In this way, we obtain easy access to the innermost callback scope. This trick makes deeply nested callback code easily composable.

After defining `map` and `flatMap` as extension methods on `Cont`, we can rewrite the code above as

```
def pure[A]: A => Cont[Unit, A] = a => ar => ar(a)
def add3(x: Int): Cont[Unit, Int] = callback => callback(x + 3)
def mult4(x: Int): Cont[Unit, Int] = callback => callback(x * 4)

val result: Cont[Unit, Int] = for {
  x <- pure(10)
  y <- mult4(x)
  z <- add3(y)
} yield z
```

This style of code is more readable and easier to modify.

The result of the computation has type `Cont[Int]` and is a function value. Like the `State` and the `Reader` monads, the continuation monad delays all computations until we apply a runner. One way of extracting values from `Cont`-monadic programs is to use a runner producing a `Future` value that resolves when the callback is called:

```
def runner[A](c: Cont[Unit, A]): Future[A] = {
  val pr = Promise[A]() // scala.concurrent.Promise
  c { a => pr.success(a) } // Resolve the promise.
  pr.future // Create a Future from Promise.
}
// Wait for the Future value.
Await.result(runner(result), Duration.Inf)
```

typically there will be only one place in the code where a monadic program is being “run”. In this way, we can isolate the high-level business logic in the `Cont` monad from the low-level code of the runners.

We conclude this subsection with some more examples of using the continuation monad.

Example 10.1.9.1 Each arithmetic computation (such as `add3` or `mult4`) has a certain arbitrarily specified cost (of a monoid type `W`). Use the monad `Cont[W, A]` to implement computations with specified cost. The total cost must automatically add up when computations are chained using `flatMap`.

Solution The functions `pure`, `add3`, and `mult4` need to be redefined with new types:

```
implicit val monoidW: Monoid[W] = ??? // Implement the "cost" monoid here.
def pure[A](a: A): Cont[W, A] = { ar => ar(a) }
def add3(x: Int, cost: W): Cont[W, Int] = { callback => callback(x + 3) |+| cost }
def mult4(x: Int, cost: W): Cont[W, Int] = { callback => callback(x * 4) |+| cost }
```

The computation of `result` can be now written as

```
val result: Cont[W, Int] = for {
  x <- pure(10)
  y <- mult4(x, cost1) // Here, cost1 and cost2 are some values of type W.
  z <- add3(y, cost2)
} yield z
```

The implementation of `runner` now needs to take an initial cost value. We can generalize the previous code of the runner to an arbitrary result type `R`:

```
def runCont[R, A](c: Cont[R, A], init: R): (R, Future[A]) = {
  val promise = Promise[A]()
  val res = c { a => promise.success(a); init } // Resolve the promise and return init.
  (res, promise.future) // Return the new result value r together with a Future[A].
```

This runner uses special low-level features of the `Future` class, such as a mutable value of type `Promise`. If these features are used in many places in the code, the programmer runs the risk of creating concurrency bugs, such as race conditions or deadlocks, that are difficult to fix. When using the continuation monad, typ-

```

    }
    val (totalCost, futureResult) = runCont(result, Monoid[W].empty)
    val resultInt = Await.result(futureResult, Duration.Inf) // Wait for the Future value.
  
```

If the code contains many other monadic operations such as `add3` and `mult4`, it is inconvenient to hard-code the cost each time. Instead, we can easily implement a function that adds a given cost to any given monadic operation:

```
def addCost[A](c: Cont[W, A], cost: W): Cont[W, A] = { callback => c(callback) |+| cost }
```

Example 10.1.9.2 Convert the callback-based API of `java.nio` to the continuation monad. Read a file into a string, write that string to another file, and finally read the new file again to verify that the string was written correctly. Do not implement any error handling.

Solution The `java.nio` package provides APIs for asynchronous input/output using buffers. For instance, the code for reading a file into a string works by creating a `fileChannel` and then calling `fileChannel.read` with a callback (encapsulated by the `CompletionHandler` class). The result of reading the file is available only when the callback is called, and only within the callback's scope:

```

import java.nio.ByteBuffer
import java.nio.channels.{AsynchronousFileChannel, CompletionHandler}
import java.nio.file.{Paths, StandardOpenOption => SOO}

val fileChannel = AsynchronousFileChannel.open(Paths.get("sample.txt"), SOO.READ)
val buffer = ByteBuffer.allocate(256)// In our simple example, the file is shorter than 256 bytes.

fileChannel.read(buffer, 0, null, new CompletionHandler[Integer, Object] {
  override def failed(e: Throwable, attachment: Object): Unit = println(s"Error reading file: $e")
  override def completed(byteCount: Integer, attachment: Object): Unit = {
    println(s"Read $byteCount bytes")
    fileChannel.close()
    buffer.rewind()
    buffer.limit(byteCount)
    val data = new String(buffer.array()) // Within this scope, we can work with the obtained data.
  }
}
  
```

Writing data to file is implemented similarly:

```

val outputFileChannel = AsynchronousFileChannel.open(Paths.get("sample2.txt"), SOO.CREATE, SOO.WRITE)
outputFileChannel.write(buffer, 0, null, new CompletionHandler[Integer, Object] {
  override def failed(e: Throwable, attachment: Object): Unit = println(s"Error writing file: $e")
  override def completed(byteCount: Integer, attachment: Object): Unit = {
    println(s"Wrote $byteCount bytes")
    outputFileChannel.close()
    ... // Continue the program within the scope of this callback.
  }
}
  
```

This API forces us to write the business logic of the program in deeply nested callbacks, since the results of input/output operations are only available within the callback scopes. The continuation monad solves this problem by converting all values into the function type `Cont[Unit, A]`. Begin by defining monadic-valued functions that encapsulate the `java.nio` APIs for reading and writing files:

```

type NioMonad[A] = Cont[Unit, A]
def nioRead(filename: String): NioMonad[ByteBuffer] = { callback =>
  val buffer = ByteBuffer.allocate(256)
  val channel = AsynchronousFileChannel.open(Paths.get(filename), SOO.READ)
  channel.read(buffer, 0, null, new CompletionHandler[Integer, Object] {
    override def failed(e: Throwable, attachment: Object): Unit = println(s"Error reading file: $e")
    override def completed(result: Integer, attachment: Object): Unit = {
      buffer.rewind()
      buffer.limit(result)
      channel.close()
      callback(buffer)
    }
  })
}
  
```

```

    }
  })
}

def nioWrite(buffer: ByteBuffer, filename: String): NioMonad[Int] = { callback =>
  val channel = AsynchronousFileChannel.open(Paths.get(filename), SOO.CREATE, SOO.WRITE)
  channel.write(buffer, 0, null, new CompletionHandler[Integer, Object] {
    override def failed(e: Throwable, attachment: Object): Unit = println(s"Error writing file: $e")
    override def completed(result: Integer, attachment: Object): Unit = {
      channel.close()
      callback(result.intValue)
    }
  })
}
}

```

Using these functions, we can implement the required code using a functor block in the `Cont` monad:

```

val filesHaveEqualContent: NioMonad[Boolean] = for {
  buffer1 <- nioRead("sample.txt")
  _           <- nioWrite("sample2.txt")
  buffer2 <- nioRead("sample2.txt")
} yield { new String(buffer1.array()) == new String(buffer2.array()) }

```

The code has become significantly easier to work with, as its high-level logic is clearly displayed.

Since the input/output operations are run concurrently, the value of type `NioMonad[Boolean]` is a function that will compute its `Boolean` result at some time in the future. We can use the runner shown above to wait for that value to become available:

```

val result: Future[Boolean] = runner(filesHaveEqualContent)

scala> Await.result(result, Duration.Inf)
res0: Boolean = true

```

10.1.10 Exercises

Exercise 10.1.10.1 For a given set of type `Set[Int]`, compute all subsets (w, x, y, z) of size 4 such that $w < x < y < z$ and $w + z = x + y$. (The values w, x, y, z must be all different.)

Exercise 10.1.10.2 Given 3 sequences xs, ys, zs of type `Seq[Int]`, compute all tuples (x, y, z) such that $x \in xs, y \in ys, z \in zs$ and $x < y < z$ and $x + y + z < 10$.

Exercise 10.1.10.3* Solve the n -queens problem on an $3 \times 3 \times 3$ cube.

Exercise 10.1.10.4 Read a file into a string and write it to another file using Java `Files` and `Paths` API. Use `Try` and `for/yield` to make that API composable and safe with respect to exceptions.

Exercise 10.1.10.5 Write a tiny library for arithmetic using `Futures`, implementing the functions

```

def const(implicit ec: ExecutionContext): Int => Future[Int] = ???
def add(x: Int)(implicit ec: ExecutionContext): Int => Future[Int] = ???
def isEqual(x: Int)(implicit ec: ExecutionContext): Int => Future[Boolean] = ???

```

Use these functions to write a functor block (`for/yield`) program that computes $1 + 2 + \dots + 100$ via a parallel computation and verifies that the result is correct.

Exercise 10.1.10.6 Given a semigroup W , make a semimonad out of the functor $F^A \triangleq E \rightarrow A \times W$.

Exercise 10.1.10.7 Implement `map` and `flatMap` for the tree-like functor $F^A \triangleq A + A \times A + F^A + F^A \times F^A$.

Exercise 10.1.10.8* Find the largest prime number below 1000 via a simple sieve of Eratosthenes.⁷ Use the `State[S, Int]` monad with `S = Array[Boolean]`.

⁷See https://en.wikipedia.org/wiki/Sieve_of_Eratosthenes

10.2 Laws of semimonads and monads

This chapter introduced semimonads to encode nested iteration as functor block programs with multiple source lines. What properties do we intuitively expect such programs to have?

When functor blocks describe iterations over data collections, a source line $x \leftarrow c$ means that the value of x iterates over items in the collection c . An assignment line $y = f(x)$ means that we define a local variable y to equal the expression $f(x)$. We expect to get the same result by iterating over a collection c whose values x were replaced by $f(x)$, i.e., by iterating over $c.map(f)$. It means that the following two code fragments should always give the same results:

```
val result1 = for {
  x <- c
  y = f(x)
  z <- g(y) // Same as z <- g(f(x)).
} yield z
// Translating the functor block into methods:
val result1 = c.flatMap(x => g(f(x)))
```

```
val result2 = for {
  y <- c.map(x =>
    f(x))
  z <- g(y) // This code is unchanged.
} yield z
// Translating the functor block into methods:
val result2 = c.map(f).flatMap(y => g(y))
```

In the code just shown, an assignment line $y = f(x)$ occurs before the last source line $z \leftarrow p(y)$. The other possibility is when the assignment line occurs *after* the last source line. A similar reasoning gives the requirement of equality between these two code fragments:

```
val result1 = for {
  x <- c
  z <- f(x)
  y = g(z)
} yield y
// Translating the functor block into methods:
val result1 = c.flatMap(f).map(g)
```

```
val result2 = for {
  x <- c
  y <- f(x).map(z =>
    g(z))
} yield y
// Translating the functor block into methods:
val result2 = c.flatMap { x => f(x).map(g) }
```

Now consider the case when there are more than two levels of nested iterations. Such programs are usually written as functor blocks with three or more source lines. However, any subgroup of source lines can be refactored into a separate functor block with unchanged result:

```
val result1 = for {
  x <- c
  y <- f(x)
  z <- g(y)
} yield z
// Translating the functor block into methods:
val result1 = c.flatMap{ x => f(x).flatMap(g) }
```

```
val result2 = for {
  yy <- for { x <- c
    y <- f(x) } yield y
  z <- g(yy)
} yield z
// Translating the functor block into methods:
val result2 = c.flatMap(f).flatMap(yy => g(yy))
```

We obtained three general requirements (or “laws”) for the `flatMap` method. Let us now formulate these requirements as rigorous equations.

For brevity, we will denote the `flatMap` method for a semimonad $S[_]$ by “ flm ”. We will write flm_S when we need to indicate explicitly the type constructor being used. The type signature is

```
def flm[A, B](f: A => S[B]): S[A] => S[B]
```

$$flm^{A,B} : (A \rightarrow S^B) \rightarrow S^A \rightarrow S^B .$$

The first law is written in Scala code as

```
c.flatMap(x => g(f(x))) ==
  c.map(f).flatMap(g)
```

$$S^A \xrightarrow{\begin{array}{c} f^S \\ \text{flm}(f:A \rightarrow B) \end{array}} S^B \xrightarrow{\begin{array}{c} \text{flm}(g:B \rightarrow S^C) \\ g^B \end{array}} S^C$$

and in the code notation, omitting the argument c (that is, using the point-free style), as

$$flm(f:A \rightarrow B) ; flm(g:B \rightarrow S^C) = f^S ; flm(g) . \quad (10.2)$$

This equation holds for arbitrary $f:A \rightarrow B$ and $g:B \rightarrow S^C$. This is a “**left naturality** law of `flatMap`” since it exchanges the order of lifted functions to the left of `flatMap`. More precisely, we may call this equation the naturality law of `flatMap[A, B]` “with respect to A ” since the application of f^S changes the type parameter A .

The second law holds for arbitrary $f:A \rightarrow S^B$ and $g:B \rightarrow C$:

```
c.flatMap(f).map(g) ==
  c.flatMap { x => f(x).map(g) }
```

$$\begin{array}{ccc} & S^B & \\ f\text{lm}(f:A \rightarrow S^B) \nearrow & \searrow (g:B \rightarrow C) \uparrow S \\ S^A & \xrightarrow{\quad} & S^C \\ & \text{f\lm}(f:A \rightarrow S^B \circ g \uparrow S) & \end{array}$$

```
c.flatMap { x => f(x).flatMap(g) } ==
  c.flatMap(f).flatMap(g)
```

$$\begin{array}{ccc} & S^B & \\ f\text{lm}(f:A \rightarrow S^B) \nearrow & \searrow \text{f\lm}(g:B \rightarrow S^C) \\ S^A & \xrightarrow{\quad} & S^C \\ & \text{f\lm}(f:A \rightarrow S^B \circ \text{f\lm}(g)) & \end{array}$$

Implementation does not obey one of these laws, programs written with that monad may give wrong results even though the code looks correct. To avoid those hard-to-debug errors, we must verify that the code for every monad's `flatMap` method obeys the laws.

At this point, the three laws of semimonads may appear complicated and hard to understand and to verify. In the next subsections, we will derive a shorter and clearer formulation of those laws. For now, let us define a `Semimonad` typeclass and test the laws using the `scalacheck` library:

```
abstract class Semimonad[F[_]]: Functor {
  def flatMap[A, B](fa: F[A])(f: A => F[B]): F[B]
}

implicit class SemimonadOps[F[_]](fa: F[A]): Semimonad[A] {
  def flatMap[B](f: A => F[B]): F[B] = implicitly[Semimonad[F]].flatMap(fa)(f)
}

def checkSemimonadLaws[F[_], A, B, C](): Semimonad[F], // Use the 'Arbitrary' typeclass
  fa: Arbitrary[F[A]], ab: Arbitrary[A => F[B]], bc: Arbitrary[B => F[C]] = { // from 'scalacheck'.
  forAll { (f: A => F[B], g: B => F[C], fa: F[A]) => // Associativity law of flatMap.
    fa.flatMap(x => f(x).flatMap(g)) shouldEqual fa.flatMap(f).flatMap(g)
  }
} // Assuming that a Semimonad instance was defined for Seq[_], check the laws with specific A, B, C.
checkSemimonadLaws[Seq, Int, String, Double]()
```

10.2.2 The laws of flatten

In Section 9.2.1 we simplified the laws of the `filter` operation by passing to a simpler `deflate` function. We then showed that these two functions are computationally equivalent if certain laws are assumed to hold for `filter`. We will now derive a similar relationship between the methods `flatMap` and `flatten`. We will see that `flatten` has fewer laws, and that its laws are simpler to verify.

Statement 10.2.2.1 The method `flatten` (denoted by $\text{ftn}^A : S^A \rightarrow S^A$) is computationally equivalent to `flatMap` as long as `flatMap` satisfies its left naturality law (10.2).

Proof By definition, `flatMap` is expressed as a composition of `map` and `flatten`,

```
c.flatMap(f) == c.map(f).flatten
```

$$\begin{array}{ccc} & S^B & \\ (f:A \rightarrow S^B) \uparrow S \nearrow & \searrow \text{ftn} \\ S^A & \xrightarrow{\quad} & S^B \\ & \text{f\lm}(f:A \rightarrow S^B) & \end{array}$$

$$\text{f\lm}_S(f) = f \uparrow S \circ \text{ftn}_S .$$

Substituting $f \triangleq \text{id}^{S^A \rightarrow S^A}$ into this equation, we get

$$\text{f\lm}(\text{id}) = \underline{\text{id}} \uparrow S \circ \text{ftn} = \text{ftn} .$$

This expresses `flatten` through `flatMap`. It remains to show that the relationship between `flatten` and `flatMap` is an isomorphism. For that, we need to prove two properties:

(1) Starting with a given function $\text{ftn} : S^A \rightarrow S^A$, define $\text{f\lm}(f) \triangleq f \uparrow S \circ \text{ftn}$ and then define a new function $\text{ftn}' \triangleq \text{f\lm}(\text{id})$. Prove that $\text{ftn}' = \text{ftn}$:

$$\text{ftn}' = \text{f\lm}(\text{id}) = \underline{\text{id}} \uparrow S \circ \text{ftn} = \text{id} \circ \text{ftn} = \text{ftn} .$$

If `flatMap` is defined via `flatten`, the left naturality law of `flatMap` is automatically satisfied:

$$\begin{aligned} \text{expect to equal } \text{flm}(f \circ g) : & \quad f^{\uparrow S} ; \text{flm}(g) = \underline{f^{\uparrow S} ; g^{\uparrow S}} ; \text{ftn} \\ \text{composition law of } S : & \quad = (f \circ g)^{\uparrow S} ; \text{ftn} = \text{flm}(f \circ g) \quad . \end{aligned}$$

(2) Starting with a given function `flm` that satisfies the left naturality law, define `ftn` \triangleq `flm(id)` and then define a new function `flm'(f) \triangleq f↑S ; ftn`. Prove that `flm' = flm`:

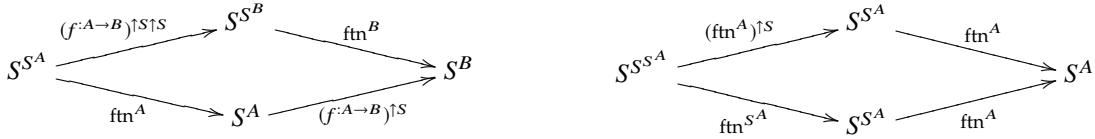
$$\begin{aligned} \text{expect to equal } \text{flm}(f) : & \quad \text{flm}'(f) = f^{\uparrow S} ; \text{ftn} = f^{\uparrow S} ; \text{flm(id)} \\ \text{left naturality law of } \text{flm} : & \quad = \text{flm}(f ; \text{id}) = \text{flm}(f) \quad . \end{aligned}$$

Statement 10.2.2.2 If a `flatMap` function satisfies the three laws (10.2)–(10.4), the corresponding `flatten` function defined as `ftn \triangleq flm(id)` satisfies its *two* laws, with an arbitrary $f : A \rightarrow B$:

$$\text{naturality law of } \text{ftn} : f^{\uparrow S \uparrow S} ; \text{ftn} = \text{ftn} ; f^{\uparrow S} \quad , \quad (10.5)$$

$$\text{associativity law of } \text{ftn} : \text{ftn}^{\uparrow S} ; \text{ftn} = \text{ftn} ; \text{ftn} \quad . \quad (10.6)$$

The following type diagrams illustrate these laws:



Note that the the associativity law involves two intermediate values of type S^{S^A} that are *not* necessarily equal. The associativity law requires only the final results of type S^A to be equal.

Proof By Statement 10.2.2.1, we have $\text{flm}(f) = f^{\uparrow S} ; \text{ftn}$ since it is given that the left naturality law of `flatMap` holds. Substituting that into the other two laws of `flatMap`, we get:

$$\text{right naturality law of } \text{flm} : \text{flm}(f ; g^{\uparrow S}) \stackrel{!}{=} \text{flm}(f) ; g^{\uparrow S} \quad ,$$

$$\text{substitute flm via ftn} : (f ; g^{\uparrow S})^{\uparrow S} ; \text{ftn} \stackrel{!}{=} f^{\uparrow S} ; \text{ftn} ; g^{\uparrow S} \quad ,$$

$$\text{set } f \triangleq \text{id} : g^{\uparrow S \uparrow S} ; \text{ftn} \stackrel{!}{=} \text{ftn} ; g^{\uparrow S} \quad ,$$

which shows that Eq. (10.5) holds; and finally:

$$\text{associativity law of } \text{flm} : \text{flm}(f ; \text{flm}(g)) \stackrel{!}{=} \text{flm}(f) ; \text{flm}(g) \quad ,$$

$$\text{substitute flm via ftn} : (f ; g^{\uparrow S} ; \text{flm}(g))^{\uparrow S} ; \text{ftn} = (f ; g^{\uparrow S} ; \text{ftn})^{\uparrow S} ; \text{ftn} \stackrel{!}{=} f^{\uparrow S} ; \text{ftn} ; g^{\uparrow S} ; \text{ftn} \quad ,$$

$$\text{set } f \triangleq \text{id} \text{ and } g \triangleq \text{id} : \text{ftn}^{\uparrow S} ; \text{ftn} \stackrel{!}{=} \text{ftn} ; \text{ftn} \quad .$$

This verifies Eq. (10.6).

Statement 10.2.2.3 If a `flatten` function satisfies the laws (10.5)–(10.6) then the corresponding `flatMap` function defined by $\text{flm}(f) \triangleq f^{\uparrow S} ; \text{ftn}$ will satisfy its three laws (10.2)–(10.4).

Proof By Statement 10.2.2.1, the left naturality law holds. To check the right naturality law:

$$\text{expect to equal } \text{flm}(f) ; g^{\uparrow S} : \text{flm}(f ; g^{\uparrow S}) = \underline{(f ; g^{\uparrow S})^{\uparrow S}} ; \text{ftn} = f^{\uparrow S} ; \underline{g^{\uparrow S \uparrow S} ; \text{ftn}}$$

$$\text{naturality law of } \text{ftn} : = \underline{f^{\uparrow S} ; \text{ftn}} ; g^{\uparrow S} = \text{flm}(f) ; g^{\uparrow S} \quad .$$

To check the associativity law:

$$\text{expect to equal } \text{flm}(f) ; \text{flm}(g) : \text{flm}(f ; \text{flm}(g)) = (f ; \text{flm}(g))^{\uparrow S} ; \text{ftn} = (f ; g^{\uparrow S} ; \text{ftn})^{\uparrow S} ; \text{ftn}$$

$$\text{associativity law of } \text{ftn} : = f^{\uparrow S} ; \underline{g^{\uparrow S \uparrow S} ; \text{ftn}} ; \text{ftn}$$

$$\text{naturality law of } \text{ftn} : = \underline{f^{\uparrow S} ; \text{ftn}} ; g^{\uparrow S} ; \text{ftn} = \text{flm}(f) ; \text{flm}(g) \quad .$$

We have proved that `flatten` is equivalent to `flatMap` but has fewer laws.

By the parametricity theorem (see Appendix D), any purely functional code will obey naturality laws. All semimonads and monads in this chapter (except `Future`) are purely functional, so we will not need to verify their naturality laws. However, verifying the associativity law involves complicated derivations. Using `flatten` instead of `flatMap` will often make those derivations shorter.

10.2.3 Verifying the associativity law via `flatten`

The following examples will verify the associativity law of `flatten` for some standard monads.

Example 10.2.3.1 The standard Scala types `Either` and `Try` are examples of the monad $F^A \triangleq Z + A$, where Z is a fixed type. Show that this monad satisfies the associativity law.

Solution The type signature of `flatten` is $\text{ftn} : Z + (Z + A) \rightarrow Z + A$, and its code is

```
def flatten[A]: Either[Z, Either[Z, A]] => Either[Z, A] = {
  case Left(z)          => Left(z)
  case Right(Left(z))   => Left(z)
  case Right(Right(a))  => Right(a)
}
```

$$\text{ftn}^{Z+Z+A \rightarrow Z+A} \triangleq \begin{array}{c|cc} & Z & A \\ \hline Z & \text{id} & \mathbb{0} \\ Z & \text{id} & \mathbb{0} \\ A & \mathbb{0} & \text{id} \end{array} .$$

Since `flatten` is fully parametric, both sides of the law are fully parametric functions with the type signature $Z + Z + Z + A \rightarrow Z + A$. This type signature has *only one* fully parametric implementation: since it is not possible to produce values of unknown types A and Z from scratch, an implementation of $Z + Z + Z + A \rightarrow Z + A$ must return $Z + \mathbb{0}$ when the input contains a value of type Z ; otherwise it must return $\mathbb{0} + A$. So, both sides of the law must have the same code.

To prove rigorously that only one implementation exists, we must use the Curry-Howard correspondence and a decision algorithm for constructive logic. Instead, we can verify the associativity law by an explicit derivation. First, we need to lift `flatten` to the functor F . The lifting code is

$$(f^{A \rightarrow B})^F \triangleq \begin{array}{c|cc} & Z & B \\ \hline Z & \text{id} & \mathbb{0} \\ A & \mathbb{0} & f \end{array} , \quad \text{ftn}^F = \begin{array}{c|ccc} & Z & Z & A \\ \hline Z & \text{id} & \mathbb{0} & \mathbb{0} \\ Z & \mathbb{0} & \text{id} & \mathbb{0} \\ Z & \mathbb{0} & \text{id} & \mathbb{0} \\ A & \mathbb{0} & \mathbb{0} & \text{id} \end{array} .$$

For comparison, the Scala code for ftn^F (had we needed to write it) would look like this,

```
def fmapFlatten[A]: Either[Z, Either[Z, Either[Z, A]]] => Either[Z, Either[Z, A]] = {
  case Left(z)          => Left(z)
  case Right(Left(z))   => Right(Left(z))
  case Right(Right(Left(z))) => Right(Left(z))
  case Right(Right(Right(a))) => Right(Right(a))
}
```

Now we can compute the two sides of the associativity law (10.6) via matrix composition:

$$\text{left-hand side : } \text{ftn}^F ; \text{ftn} = \begin{array}{c|ccc} & Z & Z & A \\ \hline Z & \text{id} & \mathbb{0} & \mathbb{0} \\ Z & \mathbb{0} & \text{id} & \mathbb{0} \\ Z & \mathbb{0} & \text{id} & \mathbb{0} \\ A & \mathbb{0} & \mathbb{0} & \text{id} \end{array} ; \begin{array}{c|cc} & Z & A \\ \hline Z & \text{id} & \mathbb{0} \\ Z & \text{id} & \mathbb{0} \\ A & \mathbb{0} & \text{id} \end{array} = \begin{array}{c|cc} & Z & A \\ \hline Z & \text{id} & \mathbb{0} \\ Z & \text{id} & \mathbb{0} \\ Z & \text{id} & \mathbb{0} \\ A & \mathbb{0} & \text{id} \end{array} ,$$

$$\begin{array}{l}
 \text{right-hand side : } \text{ftn} : Z + Z + (Z + A) \rightarrow Z + (Z + A) \circ \text{ftn} = \\
 \begin{array}{c|cc|c}
 & Z & Z + A & \\
 \hline
 Z & \text{id} & 0 & \\
 Z & \text{id} & 0 & \\
 \hline
 Z + A & 0 & \text{id} & \\
 \end{array} ; \begin{array}{c|cc}
 Z & Z & A \\
 \hline
 Z & \text{id} & 0 \\
 Z & \text{id} & 0 \\
 A & 0 & \text{id} \\
 \end{array} \\
 \\
 \text{expand id}^{Z+A} : \quad = \quad \begin{array}{c|ccc}
 & Z & Z & A \\
 \hline
 Z & \text{id} & 0 & 0 \\
 Z & \text{id} & 0 & 0 \\
 Z & 0 & \text{id} & 0 \\
 A & 0 & 0 & \text{id} \\
 \end{array} ; \begin{array}{c|cc}
 Z & A \\
 \hline
 Z & \text{id} & 0 \\
 Z & \text{id} & 0 \\
 A & 0 & \text{id} \\
 \end{array} = \begin{array}{c|cc}
 Z & A \\
 \hline
 Z & \text{id} & 0 \\
 Z & \text{id} & 0 \\
 A & 0 & \text{id} \\
 \end{array} .
 \end{array}$$

The two sides of the associativity law are equal.

When it works, the technique of Curry-Howard code inference gives much shorter proofs than explicit derivations:

Example 10.2.3.2 Verify that the Reader monad, $F^A \triangleq Z \rightarrow A$, satisfies the associativity law.

Solution The type signature of `flatten` is $(Z \rightarrow Z \rightarrow A) \rightarrow Z \rightarrow A$. Both sides of the law (10.6) are functions with the type signature $(Z \rightarrow Z \rightarrow Z \rightarrow A) \rightarrow Z \rightarrow A$. By code inference with typed holes, we find that there is only one fully parametric implementation of this type signature, namely

$$p : Z \rightarrow Z \rightarrow A \rightarrow z : Z \rightarrow p(z)(z)(z) .$$

So, both sides of the law must have the same code, and the law holds.

Example 10.2.3.3 Show that the List monad ($F^A \triangleq \text{List}^A$) satisfies the associativity law.

Solution The `flatten[A]` method has the type signature $\text{ftn}^A : \text{List}^{\text{List}^A} \rightarrow \text{List}^A$ and concatenates the nested lists in their order. Let us first show a more visually clear (but less formal) proof of the associativity law. Both sides of the law are functions of type $\text{List}^{\text{List}^{\text{List}^A}} \rightarrow \text{List}^A$. We can visualize how both sides of the law are applied to a triple-nested list value p defined by

$$p \triangleq [[[x_{11}, x_{12}, \dots], [x_{21}, x_{22}, \dots], \dots], [[y_{11}, y_{12}, \dots], [y_{21}, y_{22}, \dots], \dots], \dots] ,$$

where all x_{ij}, y_{ij}, \dots have type A . Applying ftn^{List} flattens the inner lists and produces

$$p \triangleright \text{ftn}^{\text{List}} = [[x_{11}, x_{12}, \dots, x_{21}, x_{22}, \dots], [y_{11}, y_{12}, \dots, y_{21}, y_{22}, \dots], \dots] .$$

Flattening that result gives a list of all values x_{ij}, y_{ij}, \dots , in the order they appear in p :

$$p \triangleright \text{ftn}^{\text{List}} \triangleright \text{ftn} = [x_{11}, x_{12}, \dots, x_{21}, x_{22}, \dots, y_{11}, y_{12}, \dots, y_{21}, y_{22}, \dots, \dots] .$$

Applying $\text{ftn}^{\text{List}^A}$ to p will flatten the outer lists,

$$p \triangleright \text{ftn}^{\text{List}^A} = [[x_{11}, x_{12}, \dots], [x_{21}, x_{22}, \dots], \dots, [y_{11}, y_{12}, \dots], [y_{21}, y_{22}, \dots], \dots] .$$

Flattening that value results in $p \triangleright \text{ftn}^{\text{List}^A} \triangleright \text{ftn} = [x_{11}, x_{12}, \dots, x_{21}, x_{22}, \dots, y_{11}, y_{12}, \dots, y_{21}, y_{22}, \dots, \dots]$. This is exactly the same as $p \triangleright \text{ftn}^{\text{List}} \triangleright \text{ftn}$: namely, the list of all values in the order they appear in p .

A formal proof of the associativity law is by an explicit derivation. Using the recursive type definition $\text{List}^A \triangleq \mathbb{1} + A \times \text{List}^A$, we can define `flatten` as a recursive function:

$$\text{ftn}^A \triangleq \begin{array}{c|c}
 & \mathbb{1} + \text{List}^A \times \text{List}^{\text{List}^A} \\
 \hline
 \mathbb{1} & 1 \rightarrow 1 + 0 \\
 \text{List}^A \times \text{List}^{\text{List}^A} & h : \text{List}^A \times t : \text{List}^{\text{List}^A} \rightarrow h \triangleright t \triangleright \text{ftn} \\
 \end{array} ,$$

where we have used the function `concat` (denoted `++`) whose associativity property was derived in Section 8.5.2. The operation of lifting to the `List` functor is defined for arbitrary functions f by

$$(f:A \rightarrow B)^{\uparrow \text{List}} \triangleq \left| \begin{array}{c|cc} & 1 & B \times \text{List}^B \\ \hline 1 & \text{id} & \emptyset \\ A \times \text{List}^A & \emptyset & h:A \times t:\text{List}^A \rightarrow f(h) \times t \triangleright f^{\uparrow \text{List}} \end{array} \right| .$$

The proof is by induction; the overline denotes recursive function calls, for which all laws hold by inductive assumption. We write the code for $\text{ftn}^{\uparrow \text{List}}$ and $\text{ftn}^{\text{List}^A}$, both of type $\text{List}^{\text{List}^{\text{List}^A}} \rightarrow \text{List}^{\text{List}^A}$:

$$\begin{aligned} \text{ftn}^{\uparrow \text{List}} &= \left| \begin{array}{c|cc} & 1 & \text{List}^A \times \text{List}^{\text{List}^A} \\ \hline 1 & \text{id} & \emptyset \\ \text{List}^{\text{List}^A} \times \text{List}^{\text{List}^{\text{List}^A}} & \emptyset & h:\text{List}^{\text{List}^A} \times t:\text{List}^{\text{List}^{\text{List}^A}} \rightarrow \text{ftn}(h) \times t \triangleright \text{ftn}^{\overline{\uparrow \text{List}}} \end{array} \right| , \\ \text{ftn}^{\text{List}^A} &= \left| \begin{array}{c|cc} & 1 + \text{List}^A \times \text{List}^{\text{List}^A} & \\ \hline 1 & 1 \rightarrow 1 + \emptyset & \\ \text{List}^{\text{List}^A} \times \text{List}^{\text{List}^{\text{List}^A}} & h:\text{List}^{\text{List}^A} \times t:\text{List}^{\text{List}^{\text{List}^A}} \rightarrow h \uparrow\uparrow t \triangleright \overline{\text{ftn}}^{\text{List}^A} & \end{array} \right| . \end{aligned}$$

It remains to compute $\text{ftn}^{\uparrow \text{List}} ; \text{ftn}^A$ and $\text{ftn}^{\text{List}^A} ; \text{ftn}^A$ via matrix composition. The first one is quick:

$$\begin{aligned} \text{ftn}^{\uparrow \text{List}} ; \text{ftn}^A &= \left| \begin{array}{c|cc} \text{id} & \emptyset & \\ \hline \emptyset & h \times t \rightarrow \text{ftn}(h) \times t \triangleright \text{ftn}^{\overline{\uparrow \text{List}}} & \end{array} \right| \circ \left| \begin{array}{c|cc} & 1 \rightarrow 1 + \emptyset & \\ \hline & h \times t \rightarrow h \uparrow\uparrow t \triangleright \overline{\text{ftn}} & \end{array} \right| \\ &= \left| \begin{array}{c|cc} & 1 + A \times \text{List}^A & \\ \hline 1 & 1 \rightarrow 1 + \emptyset & \\ \text{List}^{\text{List}^A} \times \text{List}^{\text{List}^{\text{List}^A}} & h \times t \rightarrow \text{ftn}(h) \uparrow\uparrow t \triangleright \text{ftn}^{\overline{\uparrow \text{List}}} \triangleright \overline{\text{ftn}} & \end{array} \right| . \end{aligned}$$

The second calculation gets stuck because the code matrix for $\text{ftn}^{\text{List}^A}$ has an unsplit column:

$$\text{ftn}^{\text{List}^A} ; \text{ftn}^A = \left| \begin{array}{c|cc} & 1 + \text{List}^A \times \text{List}^{\text{List}^A} & \\ \hline 1 & 1 \rightarrow 1 + \emptyset & \\ \text{List}^{\text{List}^A} \times \text{List}^{\text{List}^{\text{List}^A}} & h \times t \rightarrow h \uparrow\uparrow t \triangleright \overline{\text{ftn}}^{\text{List}^A} & \end{array} \right| \circ \left| \begin{array}{c|cc} & 1 \rightarrow 1 + \emptyset : A \times \text{List}^A & \\ \hline & h \times t \rightarrow h \uparrow\uparrow t \triangleright \overline{\text{ftn}} & \end{array} \right| .$$

We cannot split the column because the expression $h \uparrow\uparrow t \triangleright \overline{\text{ftn}}^{\text{List}^A}$ may evaluate to any part of the disjunction $1 + \text{List}^A \times \text{List}^{\text{List}^A}$ depending on the values of h and t . We are unable to compose the result with the second matrix unless we identify that part of the disjunction and then substitute into the first or the second column of that matrix. A general recipe in such situations is to perform additional pattern matching on the argument of the function and to substitute these values into the matrix. Consider the value of $h : \text{List}^{\text{List}^A}$, which can be an empty list, $h = 1 + \emptyset$, or a product, $h = \emptyset + g : \text{List}^A \times k : \text{List}^{\text{List}^A}$. Substitute these possibilities into the matrix expression for $\text{ftn}^{\text{List}^A} ; \text{ftn}^A$,

$$\text{with } h = 1 + \emptyset : (1 + \emptyset) \times t \triangleright \text{ftn}^{\text{List}^A} ; \text{ftn}^A = (1 + \emptyset) \uparrow\uparrow t \triangleright \overline{\text{ftn}}^{\text{List}^A} ; \text{ftn}^A$$

$$\text{concatenate with empty list} : = t \triangleright \overline{\text{ftn}}^{\text{List}^A} \triangleright \text{ftn}^A .$$

Substituting this value h into $\text{ftn}^{\uparrow\text{List}} \circ \text{ftn}^A$, we get

$$(\mathbb{0} + (1 + \mathbb{0}) \times t) \triangleright \text{ftn}^{\uparrow\text{List}} \circ \text{ftn}^A = t \triangleright \overline{\text{ftn}^{\uparrow\text{List}}} \triangleright \overline{\text{ftn}} \quad .$$

Now we just need to show that

$$t \triangleright \overline{\text{ftn}^{\text{List}^A}} \triangleright \text{ftn}^A \stackrel{?}{=} t \triangleright \overline{\text{ftn}^{\uparrow\text{List}}} \triangleright \overline{\text{ftn}} \quad .$$

This holds by the inductive assumption.

It remains to substitute the second possibility, $h = \mathbb{0} + g \times k$:

with $h = \mathbb{0} + g \times k$: $(\mathbb{0} + (\mathbb{0} + g \times k) \times t) \triangleright \text{ftn}^{\text{List}^A} \circ \text{ftn}^A = ((\mathbb{0} + g \times k) \uparrow\uparrow t \triangleright \overline{\text{ftn}^{\text{List}^A}}) \triangleright \text{ftn}^A$

code of $\uparrow\uparrow$: $= (\mathbb{0} + g \times (k \uparrow\uparrow t \triangleright \overline{\text{ftn}^{\text{List}^A}})) \triangleright \text{ftn}^A$

code of ftn^A : $= g \uparrow\uparrow (k \uparrow\uparrow t \triangleright \overline{\text{ftn}^{\text{List}^A}}) \triangleright \overline{\text{ftn}}$

Exercise 10.2.9.10 : $= g \uparrow\uparrow (k \triangleright \overline{\text{ftn}}) \uparrow\uparrow (t \triangleright \overline{\text{ftn}^{\text{List}^A}} \triangleright \overline{\text{ftn}}) = (\mathbb{0} + g \times k) \triangleright \overline{\text{ftn}} \uparrow\uparrow (t \triangleright \overline{\text{ftn}^{\text{List}^A}} \triangleright \overline{\text{ftn}})$

inductive assumption : $= \text{ftn}(h) \uparrow\uparrow (t \triangleright \overline{\text{ftn}^{\uparrow\text{List}}} \triangleright \overline{\text{ftn}}) \quad .$

This is the same as the result of substituting $\mathbb{0} + h \times t$ into $\text{ftn}^{\uparrow\text{List}} \circ \text{ftn}^A$.

Example 10.2.3.4 Consider the `List` type constructor with a `flatten` method that concatenates the nested lists in reverse order. Show that this implementation violates the associativity law of `flatten`.

Solution Apply both sides of the law to the nested list $p \triangleq [[a, b], [c, d], [e, f], [g, h]]$:

$$p \triangleright \text{ftn}^{\uparrow\text{List}} = [[c, d, a, b], [g, h, e, f]] \quad ,$$

left-hand side : $p \triangleright \text{ftn}^{\uparrow\text{List}} \triangleright \text{ftn} = [g, h, e, f, c, d, a, b] \quad ,$

$$p \triangleright \text{ftn}^{\text{List}^A} = [[e, f], [g, h], [a, b], [c, d]] \quad ,$$

right-hand side : $p \triangleright \text{ftn}^{\text{List}^A} \triangleright \text{ftn} = [c, d, a, b, g, h, e, f] \quad .$

The results are not the same. So, p is a counterexample that refutes the associativity law.

Example 10.2.3.5 Show that the `Writer` semimonad, $F^A \triangleq A \times W$, is lawful if W is a semigroup.

Solution The type signature of `flatten` is $\text{ftn} : (A \times W) \times W \rightarrow A \times W$, and the code is

$$\text{ftn} : (a \times w_1) \times w_2 \rightarrow a \times (w_1 \oplus w_2) \quad ,$$

where \oplus is the binary operation of the semigroup W . The lifting to F is $f^{\uparrow F} = a \times w \rightarrow (a \triangleright f) \times w$, so

$$\text{ftn}^{\uparrow F} = ((a \times w_1) \times w_2) \times w_3 \rightarrow (((a \times w_1) \times w_2) \triangleright \text{ftn}) \times w_3 = (a \times (w_1 \oplus w_2)) \times w_3 \quad .$$

To verify the associativity law, it is convenient to substitute a value $((a \times w_1) \times w_2) \times w_3$ of type FF^A into both sides of the law:

$$(((a \times w_1) \times w_2) \times w_3) \triangleright \text{ftn}^{\uparrow F} \triangleright \text{ftn} = ((a \times (w_1 \oplus w_2)) \times w_3) \triangleright \text{ftn} = a \times ((w_1 \oplus w_2) \oplus w_3) \quad ,$$

$$(((a \times w_1) \times w_2) \times w_3) \triangleright \text{ftn}^{F^A} \triangleright \text{ftn} = ((a \times w_1) \times (w_2 \oplus w_3)) \triangleright \text{ftn} = a \times (w_1 \oplus (w_2 \oplus w_3)) \quad .$$

The operation \oplus is associative since W is a semigroup. So, both sides of the law are equal.

Example 10.2.3.6 Consider the functor $F^A \triangleq A \times V \times V$ with the `flatten` method defined by

$$\text{ftn} \triangleq (a^A \times u_1^V \times u_2^V) \times v_1^V \times v_2^V \rightarrow a \times v_2 \times v_1 \quad .$$

Show that this definition violates the associativity law of `flatten`.

Solution Substitute a value of type FF^A into both sides of the law and get unequal results:

$$(((a \times u_1 \times u_2) \times v_1 \times v_2) \times w_1 \times w_2) \triangleright \text{ftn}^{\uparrow F} \triangleright \text{ftn} = ((a \times v_2 \times v_1) \times w_1 \times w_2) \triangleright \text{ftn} = a \times w_2 \times w_1 \quad ,$$

$$(((a \times u_1 \times u_2) \times v_1 \times v_2) \times w_1 \times w_2) \triangleright \text{ftn}^{F^A} \triangleright \text{ftn} = ((a \times u_1 \times u_2) \times w_2 \times w_1) \triangleright \text{ftn} = a \times w_1 \times w_2 \quad .$$

Had the code not exchanged the order of w_1 and w_2 , the law would have held.

10.2.4 Motivation for monad laws

Semimonads are heuristically viewed as values with a special “computational effect”. Semimonad-valued computations can be composed using the `flatMap` method, which will “merge” the effects associatively. It is generally useful to be able to create values with an “empty effect”, such that merging the empty effect leaves other effects unchanged. A full monad M is a semimonad that has a method for creating values with “empty effect”. That method is called `pure` (notation pu_M):

```
def pure[A](a: A): M[A]
```

$\text{pu}_M^A : A \rightarrow M^A$.

To get intuition about the properties of a vaguely defined “empty effect”, again consider nested iteration over arrays. The “empty effect” is an array containing *one* element, because an iteration of such an array goes over a single value, which is equivalent to no iteration. In a functor block, this intuition says that a source line with an “empty effect”, $y \leftarrow \text{pure}(x)$, should be equivalent to just $y = x$. This line must occur either before or after another source line, for instance:

```
result1 = for {
  ... // Some code, then:
  y <- pure(x) // "Empty effect" with x: A.
  z <- someArray(y) // someArray: A => M[B]
  // Same as z <- pure(x).flatMap(someArray)
```

```
result2 = for {
  ... // Some code, then:
  y = x // x: A
  z <- someArray(y) // someArray: A => M[B]
  // Same as z <- someArray(x)
```

The equality of `result1` and `result2` gives the following law: for all $g: A \rightarrow M^B$,

`pure(x).flatMap(g) == g(x)`

left identity law of M : $\text{pu}_M \circ \text{flm}_M(g: A \rightarrow M^B) = g$. (10.7)

The second possibility is that an empty effect comes *after* a source line:

```
result1 = for {
  x <- someArray // someArray: M[A]
  y <- pure(x) // Empty effect with x: A.
  // Same as y <- someArray.flatMap(x => pure(x))
```

```
result2 = for {
  x <- someArray // someArray: M[A]
  y = x
  // Same as y <- someArray
```

Then the equality of `result1` and `result2` gives the law

`g.flatMap(pure) == g`

right identity law of M : $\text{flm}_M(\text{pu}_M) = \text{id}^{M^A \rightarrow M^A}$. (10.8)

A typeclass for monads and a law-checking test function can be defined by

```
abstract class Monad[F[_]: Functor : Semimonad] {
  def pure[A](a: A): F[A]
}

def checkMonadIdentityLaws[F[_], A, B]()(implicit mf: Monad[F], sf: Semimonad[F],
  aa: Arbitrary[A], af: Arbitrary[F[A]], ab: Arbitrary[A => F[B]]) = {
  forAll { (x: A, g: A => F[B]) =>
    mf.pure(x).flatMap(g) shouldEqual g(x) // Left identity law.
  }
  forAll { (fa: F[A]) =>
    fa.flatMap(mf.pure[A]) shouldEqual fa // Right identity law.
  }
}
```

Note that `pure` is the same method as in the `Pointed` typeclass (Sections 8.3.5–8.3.6). So, we could say that a monad is a pointed semimonad whose `pure` method obeys the two identity laws (10.7)–(10.8).

Although the `pure` method can be replaced by a simpler “wrapped unit” value (wu_M), having no laws, derivations turn out to be easier when using pu_M .

The `Pointed` typeclass requires the `pure` method to satisfy the naturality law (8.8). A full monad’s `pure` method must satisfy that law, in addition to the identity laws.

Just as some useful semigroups are not monoids, there exist some useful semimonads that are not full monads. A simple example is the `Writer` semimonad $F^A \triangleq A \times W$ whose type W is a semigroup but not a monoid (see Exercise 10.2.9.1).

10.2.5 The monad identity laws in terms of `pure` and `flatten`

Since the laws of semimonads are simpler when formulated via the `flatten` method, let us convert the identity laws to that form. We use the code for `flatMap` in terms of `flatten`,

$$\text{flm}_M(f: A \rightarrow M^B) = f \uparrow^M \circ \text{ftn}_M \quad .$$

Begin with the left identity law of `flatMap`, written as

$$\begin{array}{ccc} & M^{M^A} & \\ \text{pu}^{M^A} \nearrow & \searrow \text{ftn}^A & \\ M^A & \xrightarrow{\text{id}} & M^A \end{array}$$

Since this law holds for arbitrary f , we can set $f \triangleq \text{id}$ and get

$$\text{pu}_M \circ \text{ftn}_M = \text{id}^{M^A \rightarrow M^A} \quad . \quad (10.9)$$

This is the **left identity law** of `flatten`. Conversely, if Eq. (10.9) holds, we can compose both sides with an arbitrary function $f: A \rightarrow M^B$ and recover the left identity law of `flatMap` (Exercise 10.2.9.2).

The **right identity law** of `flatten` is written as

$$\begin{array}{ccc} & M^{M^A} & \\ \text{(pu}^A\text{)} \uparrow^M \nearrow & \searrow \text{ftn}^A & \\ M^A & \xrightarrow{\text{id}} & M^A \end{array}$$

$$\text{flm}_M(\text{pu}_M) = \text{pu}_M \uparrow^M \circ \text{ftn}_M \stackrel{!}{=} \text{id} \quad . \quad (10.10)$$

In the next section, we will see a reason why these laws have their names.

10.2.6 Monad laws in terms of Kleisli functions

A **Kleisli function** is a function with type signature $A \rightarrow M^B$ where M is a monad. We first encountered Kleisli functions in Section 9.2.3 when deriving the laws of filterable functors using the `liftOpt` method. At that point, M was the simple `Option` monad. We found that functions of type $A \rightarrow \mathbb{1} + B$ can be composed using the Kleisli composition denoted by \diamond_{Opt} (see page 309). Later, Section 9.4.2 stated the general properties of Kleisli composition. We will now show that the Kleisli composition gives a useful way of formulating the laws of a monad.

The Kleisli composition operation for a monad M , denoted \diamond_M , is a function with type signature

$$\diamond_M : (A \rightarrow M^B) \rightarrow (B \rightarrow M^C) \rightarrow A \rightarrow M^C \quad .$$

This resembles the forward composition of ordinary functions, $(\circ) : (A \rightarrow B) \rightarrow (B \rightarrow C) \rightarrow A \rightarrow C$, except for different types of functions. If M is a monad, the implementation of \diamond_M is

```
def <>[M[_]]: Monad, A,B,C](f: A => M[B], g: B => M[C]): A => M[C] = { x => f(x).flatMap(g) } f <math>\diamond_M g \triangleq f \circ \text{flm}_M(g) \quad . \quad (10.11)
```

The Kleisli composition can be equivalently expressed by a functor block code as

```
def <>[M[_]]: Monad, A,B,C](f: A => M[B], g: B => M[C]): A => M[C] = { x =>
  for {
    y <- f(x)
    z <- g(y)
  } yield z
}
```

This example shows that Kleisli composition is a basic part of functor block code: it expresses the chaining of two consecutive “source” lines.

Let us now derive the laws of Kleisli composition \diamond_M , assuming that the monad laws hold for M .

Statement 10.2.6.1 For a lawful monad M , the Kleisli composition \diamond_M satisfies the identity laws

$$\text{left identity law of } \diamond_M : \text{pu}_M \diamond_M f = f \quad , \quad \forall f : A \rightarrow M^B \quad , \quad (10.12)$$

$$\text{right identity law of } \diamond_M : f \diamond_M \text{pu}_M = f \quad , \quad \forall f : A \rightarrow M^B \quad . \quad (10.13)$$

Proof We may assume that Eqs. (10.7)–(10.8) hold. Using the definition (10.11), we find

$$\text{left identity law of } \diamond_M, \text{ should equal } f : \text{pu}_M \diamond_M f = \underline{\text{pu}_M \circ \text{flm}_M(f)}$$

$$\text{use Eq. (10.7)} : = f \quad ,$$

$$\text{right identity law of } \diamond_M, \text{ should equal } f : f \diamond_M \text{pu}_M = f \circ \underline{\text{flm}_M(\text{pu}_M)}$$

$$\text{use Eq. (10.8)} : = f \circ \text{id} = f \quad .$$

The following statement and the identity law (10.8) show that `flatMap` can be viewed as a “lifting”,

$$\text{flm}_M : (A \rightarrow M^B) \rightarrow (M^A \rightarrow M^B) \quad ,$$

from Kleisli functions $A \rightarrow M^B$ to M -lifted functions $M^A \rightarrow M^B$, except that Kleisli functions must be composed using \diamond_M , while pu_M plays the role of the Kleisli-identity function.

Statement 10.2.6.2 For a lawful monad M , the `flatMap` method satisfies the composition law

$$\begin{array}{ccc} \text{flm}_M(f) & \nearrow M^B & \text{flm}_M(g) \\ M^A & \xrightarrow{\text{flm}_M(f \diamond_M g)} & M^C \end{array} \quad \text{flm}_M(f \diamond_M g) = \text{flm}_M(f) \circ \text{flm}_M(g) \quad .$$

Proof We may use Eq. (10.4) since M is a lawful monad. A direct calculation yields the law:

$$\text{expect to equal } \text{flm}_M(f) \circ \text{flm}_M(g) : \text{flm}_M(f \diamond_M g) = \text{flm}_M(f \circ \text{flm}_M(g))$$

$$\text{use Eq. (10.4)} : = \text{flm}_M(f) \circ \text{flm}_M(g) \quad .$$

The following statement motivates calling Eq. (10.4) an “associativity” law.

Statement 10.2.6.3 For a lawful monad M , the Kleisli composition \diamond_M satisfies the **associativity law**

$$(f \diamond_M g) \diamond_M h = f \diamond_M (g \diamond_M h) \quad , \quad \forall f : A \rightarrow M^B, g : B \rightarrow M^C, h : C \rightarrow M^D \quad . \quad (10.14)$$

So, we may write $f \diamond_M g \diamond_M h$ unambiguously with no parentheses.

Proof Substitute Eq. (10.11) into both sides of the law:

$$\text{left-hand side} : \underline{(f \diamond_M g)} \diamond_M h = (f \circ \underline{\text{flm}_M(g)}) \diamond_M h = f \circ \text{flm}_M(g) \circ \text{flm}_M(h) \quad ,$$

$$\text{right-hand side} : \underline{f \diamond_M (g \diamond_M h)} = f \circ \underline{\text{flm}_M(g \diamond_M h)}$$

$$\text{use Statement 10.2.6.2} : = f \circ \text{flm}_M(g) \circ \text{flm}_M(h) \quad .$$

Both sides of the law are now equal.

We find that the properties of the operation \diamond_M are similar to the identity and associativity properties of the function composition $f \circ g$ except for using pu_M instead of the identity function.⁸

Since the Kleisli composition describes the chaining of consecutive lines in functor blocks, its associativity means that multiple lines are chained unambiguously. For example, this code:

⁸It means that Kleisli functions satisfy the properties of morphisms of a category; see Section 9.4.3.

```

1 x => for {
2   y <- f(x)
3   z <- g(y)
4   t <- h(z)
5 } yield t

```

corresponds to the Kleisli composition

$$(x \rightarrow f(x)) \diamond_M (y \rightarrow g(y)) \diamond_M (z \rightarrow h(z))$$

and does not need to specify whether lines 2 and 3 are chained before appending line 4, or lines 3 and 4 are chained before prepending line 2.

We will now prove that the Kleisli composition with its laws is computationally equivalent to `flatMap` with *its* laws. In other words, we may equally well use the Kleisli composition when formulating the requirements for a functor M to be a monad.

Statement 10.2.6.4 The Kleisli composition \diamond_M and M 's `flatMap` are computationally equivalent,

$$f:A \rightarrow M^B \diamond_M g:B \rightarrow M^C = f \circ \text{flm}_M(g) \quad , \quad \text{flm}_M(f:A \rightarrow M^B) = \text{id}^{M^A \rightarrow M^A} \diamond_M f \quad , \quad (10.15)$$

provided that Eqs. (10.12)–(10.14) and the following additional law hold for \diamond_M :

$$\text{left naturality of } \diamond_M : (f:A \rightarrow B \circ g:B \rightarrow M^C) \diamond_M h:C \rightarrow M^D = f \circ (g \diamond_M h) \quad . \quad (10.16)$$

Note that this law makes parentheses unnecessary in the expression $f \circ g \diamond_M h$.

Proof Equations (10.15) map \diamond_M to `flmM` and back. We have to show that these mappings are isomorphisms when the given laws hold. We proceed in two steps:

(1) Given an operation \diamond_M , we define `flmM` and then a new operation \diamond'_M using Eq. (10.15). We then need to prove that $\diamond'_M = \diamond_M$. Calculate using arbitrary functions $f:A \rightarrow M^B$ and $g:B \rightarrow M^C$:

$$\text{use Eq. (10.15)} : f \diamond'_M g = f \circ \text{flm}_M(g) = f \circ (\text{id}^{M^B} \diamond_M g)$$

$$\text{left naturality (10.16) of } \diamond_M : = (f \circ \text{id}) \diamond_M g = f \diamond_M g \quad .$$

When \diamond_M is defined via `flmM`, the left naturality law (10.16) will hold because “ \circ ” is associative,

$$(f \circ g) \diamond_M h = (f \circ g) \circ \text{flm}_M(h) = f \circ (g \circ \text{flm}_M(h)) = f \circ (g \diamond_M h) \quad .$$

(2) Given a function `flmM`, we define \diamond_M and then a new function flm'_M using Eq. (10.15). We then need to prove that $\text{flm}'_M = \text{flm}_M$. Calculate using an arbitrary function $f:A \rightarrow M^B$:

$$\text{use Eq. (10.15)} : \text{flm}'_M(f) = \text{id}^{M^A} \diamond_M f = \text{id} \circ \text{flm}_M(f) = \text{flm}_M(f) \quad .$$

We have already derived the laws of Kleisli composition from the laws of `flatMap`. We will now derive the converse statement. In this way, we will show that the Kleisli composition laws and the `flatMap` laws are fully equivalent.

Statement 10.2.6.5 If the Kleisli composition \diamond_M obeys the laws (10.12)–(10.14), the corresponding `flatMap` method defined by Eq. (10.15) will satisfy the laws (10.2)–(10.4).

Proof To derive the identity laws of `flatMap`:

$$\text{left identity law} : \text{pu}_M \circ \text{flm}_M(f) = \text{pu}_M \circ \text{id} \diamond_M f = \text{pu}_M \diamond_M f = f \quad ,$$

$$\text{right identity law} : \text{flm}_M(\text{pu}_M) = \text{id} \diamond_M \text{pu}_M = \text{id} \quad .$$

To derive the associativity law (10.4) of `flatMap`, substitute the definition of `flmM` into both sides:

$$\text{left-hand side} : \text{flm}_M(f \circ \text{flm}_M(g)) = \text{id} \diamond_M (f \circ \text{id} \diamond_M g) = \text{id} \diamond_M (f \diamond_M g)$$

$$\text{associativity law (10.14)} : = (\text{id} \diamond_M f) \diamond_M g \quad ,$$

$$\text{right-hand side} : \text{flm}_M(f) \circ \text{flm}_M(g) = (\text{id} \diamond_M f) \circ (\text{id} \diamond_M g)$$

$$\text{left naturality (10.16) of } \diamond_M : = (\text{id} \diamond_M f) \circ \text{id} \diamond_M g = (\text{id} \diamond_M f) \diamond_M g \quad .$$

Both sides of the law are now equal.

The two naturality laws of `flatMap` are equivalent to the three naturality laws of \diamond_M , but we omit those derivations.

10.2.7 Verifying the monad laws using Kleisli functions

When the monad laws are formulated via Kleisli composition, the intuition behind the laws becomes clearer: they are analogous to the identity and associativity laws of the function composition (\circ). The price is that the type signatures become complicated. For instance, the associativity law (10.14) has four type parameters, while the corresponding law (10.6) for `flatten` has only one. For certain monads, however, a trick called **flipped Kleisli** makes direct proofs of laws much shorter. That trick applies to monads of a function type, such as the continuation and the state monads.

Statement 10.2.7.1 The continuation monad, $\text{Cont}^{R,A} \triangleq (A \rightarrow R) \rightarrow R$, satisfies all monad laws.

Proof Begin by writing the type of a Kleisli function corresponding to this monad,

$$A \rightarrow \text{Cont}^{R,B} = A \rightarrow (B \rightarrow R) \rightarrow R \quad .$$

This function type has two curried arguments. The first step of the flipped Kleisli technique is to change the types of the Kleisli functions by flipping their two curried arguments. We obtain

$$(B \rightarrow R) \rightarrow A \rightarrow R \quad .$$

This type looks like a function of the form $K^B \rightarrow K^A$, where we temporarily defined $K^A \triangleq A \rightarrow R$. The remaining steps are to flip the arguments of pu_{Cont} , obtaining a modified $\tilde{\text{pu}}_{\text{Cont}}$, and to modify the Kleisli composition \diamond_{Cont} into $\tilde{\diamond}_{\text{Cont}}$, so that we can compose the flipped Kleisli functions using $\tilde{\diamond}_{\text{Cont}}$. We will then prove the laws

$$\tilde{\text{pu}}_{\text{Cont}} \tilde{\diamond}_{\text{Cont}} f = f \quad , \quad f \tilde{\diamond}_{\text{Cont}} \tilde{\text{pu}}_{\text{Cont}} = f \quad , \quad (f \tilde{\diamond}_{\text{Cont}} g) \tilde{\diamond}_{\text{Cont}} h = f \tilde{\diamond}_{\text{Cont}} (g \tilde{\diamond}_{\text{Cont}} h) \quad .$$

Flipping the arguments is an operation that maps functions to computationally equivalent functions. So, if we prove the laws of identity and composition for flipped Kleisli functions, it will follow that the same laws hold for the original functions.

The original `pure` method is

$$\text{pu}_{\text{Cont}} \triangleq a:A \rightarrow f:A \rightarrow R \rightarrow f(a) \quad .$$

We find that the flipped `pure` method is just an identity function:

$$\tilde{\text{pu}}_{\text{Cont}} \triangleq f:A \rightarrow R \rightarrow a:A \rightarrow f(a) = f:A \rightarrow R \rightarrow f = \text{id}^{(A \rightarrow R) \rightarrow A \rightarrow R} = \text{id}^{K^A \rightarrow K^A} \quad .$$

This is a significant simplification. The flipped Kleisli composition $\tilde{\diamond}_{\text{Cont}}$ has the type signature

$$f:K^B \rightarrow K^A \tilde{\diamond}_{\text{Cont}} g:K^C \rightarrow K^B = ???^{K^C \rightarrow K^A} \quad .$$

There is only one implementation of that type signature, namely the backward composition,

$$f:K^B \rightarrow K^A \tilde{\diamond}_{\text{Cont}} g:K^C \rightarrow K^B \triangleq g \circ f = f \circ g \quad .$$

So, this must be the code of the flipped Kleisli composition. It is now quick to verify the laws:

$$\text{left identity law : } \tilde{\text{pu}}_{\text{Cont}} \tilde{\diamond}_{\text{Cont}} f = \text{id} \circ f = f \quad ,$$

$$\text{right identity law : } f \tilde{\diamond}_{\text{Cont}} \tilde{\text{pu}}_{\text{Cont}} = f \circ \text{id} = f \quad ,$$

$$\text{associativity law : } (f \tilde{\diamond}_{\text{Cont}} g) \tilde{\diamond}_{\text{Cont}} h = (f \circ g) \circ h = f \circ (g \circ h) = f \tilde{\diamond}_{\text{Cont}} (g \tilde{\diamond}_{\text{Cont}} h) \quad .$$

We could have avoided writing the last three lines by noticing that functions of types $K^B \rightarrow K^A$ automatically satisfy the laws of (backward) function composition (Section 4.2.2).

Statement 10.2.7.2 The state monad, $\text{State}^{S,A} \triangleq S \rightarrow A \times S$, satisfies all monad laws.

Proof Begin by writing the type of a Kleisli function corresponding to this monad,

$$A \rightarrow \text{State}^{S,B} = A \rightarrow S \rightarrow B \times S \quad .$$

An equivalent type is obtained by uncurrying the two arguments:

$$(A \rightarrow S \rightarrow B \times S) \cong (A \times S \rightarrow B \times S) = K^A \rightarrow K^B \quad ,$$

where we temporarily defined $K^A \triangleq A \times S$. This type looks like an ordinary function, which promises to simplify the proof. Uncurrying is an equivalence transformation. So, let us uncurry the arguments in all Kleisli functions and prove the laws in the “uncurried Kleisli” formulation. We need to uncurry the arguments in pu_{State} and \diamond_{State} as well. Denote the resulting functions $\tilde{\text{pu}}_{\text{State}}$ and $\tilde{\diamond}_{\text{State}}$:

$$\tilde{\text{pu}}_{\text{State}} \triangleq a^{:A} \times s^{:S} \rightarrow a \times s = \text{id}^{:K^A \rightarrow K^A} \quad , \quad f^{:K^A \rightarrow K^B} \tilde{\diamond}_{\text{State}} g^{:K^B \rightarrow K^C} \triangleq f \diamond g \quad .$$

We can see that the composition $f \diamond g$ implements the correct logic of the state monad: an initial state value $s^{:S}$ is updated by f and then passed to g .

We found that the uncurried Kleisli functions have simple types $K^A \rightarrow K^B$, and the operation $\tilde{\diamond}_{\text{State}}$ is just the ordinary composition of those functions. Since we already know that the laws of identity and associativity hold for ordinary functions (Section 4.2.2), the proof is finished.

For comparison, look at the type signatures of `flatten` for the state and continuation monads:

$$\begin{aligned} \text{ftn}_{\text{State}^{S,\bullet}} &: (S \rightarrow (S \rightarrow A \times S) \times S) \rightarrow S \rightarrow A \times S \quad , \\ \text{ftn}_{\text{Cont}^{R,\bullet}} &: (((A \rightarrow R) \rightarrow R) \rightarrow R) \rightarrow (A \rightarrow R) \rightarrow R \quad . \end{aligned}$$

These type signatures are complicated and confusing to read. Direct proofs of the monad laws for these functions are much longer than the proofs of Statements 10.2.7.2–10.2.7.1. When a monad M has a function type, the Kleisli function $A \rightarrow M^B$ has two curried arguments. Flipping or uncurrying these arguments often produces an equivalent function that is easier to work with.

10.2.8 Structural analysis of semimonads and monads

We have seen different examples of well-known monads that were discovered by programmers working on specific tasks. Hoping to find systematically all possible monads, we will now apply structural analysis to semimonads and monads. For each type construction, we will prove rigorously that the monad laws hold.

Type parameters Three type constructions are based on using just type parameters: a constant functor, $\text{Const}^{Z,A} \triangleq Z$, the identity functor $\text{Id}^A \triangleq A$, and the functor composition, $L^A \triangleq F^{G^A}$.

A constant functor $F^A \triangleq Z$ is a lawful semimonad because we can implement

$$\text{ftn}_F = \text{id}^{:Z \rightarrow Z} \quad .$$

An identity function will always satisfy the laws. To obtain a full monad, we need to implement

$$\text{pu}_F : A \rightarrow Z \quad .$$

This is possible only if we have a default value z_0 of type Z . Assuming that, we set $\text{pu}_F \triangleq - \rightarrow z_0$ and check the identity laws:

$$\text{left identity law of } F : \text{pu}_F^{:F^A \rightarrow F^{F^A}} \diamond \text{ftn}_F = \text{pu}_F \diamond \text{id} = \text{pu}_F \stackrel{?}{=} \text{id} \quad .$$

The function pu_F is a constant function that always returns z_0 , and yet it must be equal to the identity function. This law can be satisfied only if the identity function of type $Z \rightarrow Z$ always returns the same value z_0 . It follows that z_0 is the only available value of the type Z , which means $Z \cong \mathbb{1}$. In that

case, all functions become constants returning 1, and the laws are trivially satisfied. We conclude that the only case when a constant functor is a lawful monad is when $L^A \triangleq \mathbb{1}$ (the constant `Unit` type).

The identity functor $\text{Id}^A \triangleq A$ is a monad: its `pure` and `flatten` methods are identity functions.

Functor composition $F \circ G$ is not guaranteed to produce monads even if F and G are both monads. A counterexample can be found by taking $F^A \triangleq Z + A$ and $G^A \triangleq R \rightarrow A$; the composition $L^A \triangleq Z + (R \rightarrow A)$ is not even a semimonad.

Statement 10.2.8.1 The functor $L^A \triangleq Z + (R \rightarrow A)$, where R and Z are fixed but arbitrary types, cannot have a `flatten` method.

Proof The type signature of `flatten` is

$$\text{ftn}_L : Z + (R \rightarrow Z + (R \rightarrow A)) \rightarrow Z + (R \rightarrow A) .$$

A fully parametric implementation is impossible: we would need to compute either a value of type Z or of type $R \rightarrow A$ from an argument that may have either type Z or type $R \rightarrow Z + (R \rightarrow A)$. When the argument has the latter type, it is impossible to compute either a value of type Z or a value of type $R \rightarrow A$, because different values may be returned for different values of R , and we do not have any known values of type R available.

Products The product construction works for semimonads as well as for monads.

Statement 10.2.8.2 Given two semimonads F^A and G^A , the functor $L^A \triangleq F^A \times G^A$ is a semimonad. If both F^A and G^A are monads then L^A is also a monad.

Proof Begin by defining the `flatten` method for the semimonad L via the `flatten` methods of F and G . The `flatten` method needs to transform a value of type $L^{F^A \times G^A} = F^{F^A \times G^A} \times G^{F^A \times G^A}$ into a value of type $F^A \times G^A$. Since F and G are functors, we can extract F^{F^A} out of $F^{F^A \times G^A}$ by lifting the standard projection function π_1 ,

$$\pi_1^F : F^{F^A \times G^A} \rightarrow F^{F^A} .$$

Then we use F 's `flatten` method to obtain a value of type F^A . In a similar way, we transform a value of type $G^{F^A \times G^A}$ into G^A . The resulting code is

```
def flatten_L[A]: (F[(F[A], G[A])], G[(F[A], G[A])]) => (F[A], G[A]) = { case (ffa, gga) =>
  val ffa: F[F[A]] = ffa.map(_._1)
  val gga: G[G[A]] = gga.map(_._2)
  (flatten_F(ffa), flatten_G(gga))
}
```

$$\text{ftn}_L \triangleq f^{F^{F^A \times G^A}} \times g^{G^{F^A \times G^A}} \rightarrow (f \triangleright \pi_1^F \triangleright \text{ftn}_F) \times (g \triangleright \pi_2^G \triangleright \text{ftn}_G) = (\pi_1^F ; \text{ftn}_F) \boxtimes (\pi_2^G ; \text{ftn}_G) . \quad (10.17)$$

To verify the associativity law, we need to write the code of ftn_L^L :

$$(h^{A \rightarrow B})^{\uparrow L} = h^{\uparrow F} \boxtimes h^{\uparrow G} , \quad \text{ftn}_L^{\uparrow L} = \text{ftn}_L^{\uparrow F} \boxtimes \text{ftn}_L^{\uparrow G} .$$

Substitute these definitions into the associativity law:

$$\begin{aligned} \text{left-hand side : } & \text{ftn}_L^{\uparrow L} ; \text{ftn}_L = (\text{ftn}_L^{\uparrow F} \boxtimes \text{ftn}_L^{\uparrow G}) ; ((\pi_1^F ; \text{ftn}_F) \boxtimes (\pi_2^G ; \text{ftn}_G)) \\ \text{use Eq. (7.2) : } & = (\text{ftn}_L^{\uparrow F} ; \pi_1^F ; \text{ftn}_F) \boxtimes (\text{ftn}_L^{\uparrow G} ; \pi_2^G ; \text{ftn}_G) , \\ \text{right-hand side : } & \text{ftn}_L ; \text{ftn}_L = ((\pi_1^F ; \text{ftn}_F) \boxtimes (\pi_2^G ; \text{ftn}_G)) ; ((\pi_1^F ; \text{ftn}_F) \boxtimes (\pi_2^G ; \text{ftn}_G)) \\ \text{use Eq. (7.2) : } & = (\pi_1^F ; \text{ftn}_F ; \pi_1^F ; \text{ftn}_F) \boxtimes (\pi_2^G ; \text{ftn}_G ; \pi_2^G ; \text{ftn}_G) . \end{aligned}$$

The only given information about ftn_F and ftn_G is that they obey their associativity laws; e.g.,

$$\text{ftn}_F^{\uparrow F} ; \text{ftn}_F = \text{ftn}_F ; \text{ftn}_F .$$

In order to use this law, we need to move the two functions ftn_F next to each other in the expressions

$$(\text{ftn}_L^{\uparrow F} ; \pi_1^{\uparrow F} ; \text{ftn}_F) \quad \text{and} \quad (\pi_1^{\uparrow F} ; \text{ftn}_F ; \pi_1^{\uparrow F} ; \text{ftn}_F) \quad ,$$

hoping to show that these expressions are equal. We begin with the first of those expressions:

$$\begin{aligned} \text{ftn}_L^{\uparrow F} ; \pi_1^{\uparrow F} ; \text{ftn}_F &= (\text{ftn}_L ; \pi_1)^{\uparrow F} ; \text{ftn}_F \\ \text{left projection law (7.3)} : &= (\pi_1 ; \pi_1^{\uparrow F} ; \text{ftn}_F)^{\uparrow F} ; \text{ftn}_F = \pi_1^{\uparrow F} ; \pi_1^{\uparrow F \uparrow F} ; \text{ftn}_F^{\uparrow F} ; \text{ftn}_F \\ \text{associativity law of } F : &= \pi_1^{\uparrow F} ; \pi_1^{\uparrow F \uparrow F} ; \text{ftn}_F ; \text{ftn}_F \\ \text{naturality law of } \text{ftn}_F : &= \pi_1^{\uparrow F} ; \text{ftn}_F ; \pi_1^{\uparrow F} ; \text{ftn}_F \quad . \end{aligned}$$

Both expressions are now the same. An analogous derivation shows that

$$\text{ftn}_L^{\uparrow G} ; \pi_2^{\uparrow G} ; \text{ftn}_G = \pi_2^{\uparrow G} ; \text{ftn}_G ; \pi_2^{\uparrow G} ; \text{ftn}_G \quad .$$

So, both sides of the associativity law are equal.

Now we assume that F and G are monads with given `pure` methods pu_F and pu_G . We define

$$\text{pu}_L \triangleq a^A \rightarrow \text{pu}_F(a) \times \text{pu}_G(a) = \Delta ; (\text{pu}_F \boxtimes \text{pu}_G) \quad .$$

Assuming that identity laws hold for F and G , we can now verify the identity laws for L :

$$\begin{aligned} \text{left identity law of } L : \quad \text{pu}_L ; \text{ftn}_L &= \Delta ; (\text{pu}_F \boxtimes \text{pu}_G) ; ((\pi_1^{\uparrow F} ; \text{ftn}_F) \boxtimes (\pi_2^{\uparrow G} ; \text{ftn}_G)) \\ \text{use Eq. (7.2)} : &= \Delta ; ((\text{pu}_F ; \pi_1^{\uparrow F} ; \text{ftn}_F) \boxtimes (\text{pu}_G ; \pi_2^{\uparrow G} ; \text{ftn}_G)) \\ \text{naturality of } F, G : &= \Delta ; ((\pi_1 ; \text{pu}_F ; \text{ftn}_F) \boxtimes (\pi_2 ; \text{pu}_G ; \text{ftn}_G)) \\ \text{identity laws of } F, G : &= \Delta ; (\pi_1 \times \pi_2) = \text{id} \quad , \\ \text{right identity law of } L : \quad \text{pu}_L^{\uparrow L} ; \text{ftn}_L &= ((\Delta ; (\text{pu}_F \boxtimes \text{pu}_G))^{\uparrow F} \boxtimes (\Delta ; (\text{pu}_F \boxtimes \text{pu}_G))^{\uparrow G}) ; \text{ftn}_L \\ \text{use Eq. (7.2)} : &= ((\Delta ; (\text{pu}_F \boxtimes \text{pu}_G))^{\uparrow F} ; \pi_1^{\uparrow F} ; \text{ftn}_F) \boxtimes ((\Delta ; (\text{pu}_F \boxtimes \text{pu}_G))^{\uparrow G} ; \pi_2^{\uparrow G} ; \text{ftn}_G) \\ \text{projection laws (7.3)} : &= ((\Delta ; \pi_1 ; \text{pu}_F)^{\uparrow F} ; \text{ftn}_F) \boxtimes ((\Delta ; \pi_2 ; \text{pu}_G)^{\uparrow G} ; \text{ftn}_G) \\ \text{identity laws (7.1)} : &= (\text{pu}_F^{\uparrow F} ; \text{ftn}_F) \boxtimes (\text{pu}_G^{\uparrow G} ; \text{ftn}_G) = \text{id} \boxtimes \text{id} = \text{id} \quad . \end{aligned}$$

Let us build some intuition about how the product of two monads works in practice. A simple example is the product of two identity monads, $L^A \triangleq A \times A$. This type constructor is a monad whose `flatten` function is defined by

```
type Pair[A] = (A, A)
def flatten[A]: Pair[Pair[A]] => Pair[A] = { case ((a, b), (c, d)) => (a, d) }
```

A sample calculation shows that “nested iterations” apply functions element by element:

```
final case class P[A](x: A, y: A) {
  def map[B](f: A => B): P[B] = P(f(x), f(y))
  def flatMap[B](f: A => P[B]): P[B] = P(f(x).x, f(y).y)
}

scala> for {
  x <- P(1, 10)
  y <- P(2, 20)
  z <- P(3, 30)
} yield x + y + z      // The result is P(1 + 2 + 3, 10 + 20 + 30).
res0: P[Int] = P(6, 60)
```

Note that the pair type $A \times A$ is equivalent to the function type $2 \rightarrow A$ (in Scala, `Boolean => A`). We know that $2 \rightarrow A$ is a `Reader` monad. One can check that the implementation of `flatten` for the `Reader` monad $2 \rightarrow A$ is equivalent to the code of `flatten` for $L^A \triangleq A \times A$ as shown above.

Now consider the product $F^A \times G^A$ with arbitrary monads F and G that represent some effects. Then a Kleisli function of type $A \rightarrow F^B \times G^B$ contains both effects. How does the monad L combine two effects when we compose two such Kleisli functions? We see from the code of L 's `flatten` that the first effect in F is combined with the second effect in F , and the first effect in G is combined with the second effect in G . The part F^{G^A} from $F^{F^A \times G^A}$ is discarded by `flatten`, leaving only F^{F^A} , i.e., a combination of two F -effects. Also, G^{F^A} is discarded from $G^{F^A \times G^A}$, leaving only G^{G^A} . As an example, consider this code:

```
val result: (F[C], G[C]) = for {
  x <- (fa, ga)           // Assume fa: F[A], ga: G[A]
  y <- (fb, gb)           // Assume fb: F[B], gb: G[B]
} yield h(x, y)           // Assume h: (A, B) => C
```

The expression `result` is equivalent to the following code that works with F and G separately:

```
val result1: F[C] = for {
  x <- fa
  y <- fb
} yield h(x, y)
```

```
val result2: G[C] = for {
  x <- ga
  y <- gb
} yield h(x, y)
```

```
val result: (F[C], G[C]) =
  (result1, result2)
```

A composition of two L -effects is a pair consisting of F -effects and G -effects composed separately. Because of that, the composition of L -effects is associative (and so L is a lawful semimonad) as long as F - and G -effects are themselves composed associatively.

Statement 10.2.8.3 Given any functor F^A , the functor $L^A \triangleq A \times F^A$ is a semimonad.

Proof We begin by implementing the `flatten` method for L , with the type signature $A \times F^A \times F^{A \times F^A} \rightarrow A \times F^A$. Since we know nothing about the functor F , we cannot extract values of type A from F^A . We also do not have a `flatten` method for F . How can we get values of type A and F^A out of $A \times F^A \times F^{A \times F^A}$? One possibility is simply to discard the part of type $F^{A \times F^A}$:

```
def flatten1_L[A]: ((A, F[A]), F[(A, F[A])]) => (A, F[A]) = ...
```

The other possibility is to transform $F^{A \times F^A}$ to F^A within the functor F :

```
def flatten2_L[A]: ((A, F[A]), F[(A, F[A])]) => (A, F[A]) = {
  case (afa, fafa) =>
    (afa._1, fafa.map(_._1))
}
```

In the code notation, these alternative implementations may be written as

$$\text{ftn}_1 L \triangleq \pi_1 = p^{A \times F^A} \times q^{F^{A \times F^A}} \rightarrow p \quad , \quad \text{ftn}_2 L \triangleq p^{A \times F^A} \times q^{F^{A \times F^A}} \rightarrow (p \triangleright \pi_1) \times (q \triangleright \pi_1^F) = \pi_1 \boxtimes \pi_1^F \quad .$$

To check the associativity laws, we need to prepare the code for lifting to the functor L :

$$f^{\uparrow L} = f \boxtimes f^{\uparrow F} \quad , \quad \text{ftn}_1^{\uparrow L} = \pi_1 \boxtimes \pi_1^{\uparrow F} \quad , \quad \text{ftn}_2^{\uparrow L} = (\pi_1 \boxtimes \pi_1^{\uparrow F}) \boxtimes (\pi_1 \boxtimes \pi_1^{\uparrow F})^{\uparrow F} \quad .$$

Then we verify the associativity law for $\text{ftn}_1 L$ using the projection law (7.3):

$$\begin{aligned} \text{left-hand side : } & \text{ftn}_1^{\uparrow L} ; \text{ftn}_1 L = (\pi_1 \boxtimes \pi_1^{\uparrow F}) ; \pi_1 = \pi_1 ; \pi_1 \quad , \\ \text{right-hand side : } & \text{ftn}_1 L ; \text{ftn}_1 L = \pi_1 ; \pi_1 \quad . \end{aligned}$$

The law holds. For ftn_{2L} , we also find that the two sides of the law are equal:

$$\begin{aligned} \text{ftn}_{2L}^L \circ \text{ftn}_{2L} &= ((\pi_1 \boxtimes \pi_1^F) \boxtimes (\pi_1 \boxtimes \pi_1^F)^F) \circ (\pi_1 \boxtimes \pi_1^F) \\ \text{use Eq. (7.2)} : \quad &= ((\pi_1 \boxtimes \pi_1^F) \circ \pi_1) \boxtimes ((\pi_1 \boxtimes \pi_1^F)^F \circ \pi_1^F) = ((\pi_1 \boxtimes \pi_1^F) \circ \pi_1) \boxtimes ((\pi_1 \boxtimes \pi_1^F) \circ \pi_1)^F \\ \text{use Eq. (7.3)} : \quad &= (\pi_1 \circ \pi_1) \boxtimes (\pi_1 \circ \pi_1)^F = (\pi_1 \circ \pi_1) \boxtimes (\pi_1^F \circ \pi_1^F) \quad , \\ \text{ftn}_{2L} \circ \text{ftn}_{2L} &= (\pi_1 \boxtimes \pi_1^F) \circ (\pi_1 \boxtimes \pi_1^F) = (\pi_1 \circ \pi_1) \boxtimes (\pi_1^F \circ \pi_1^F) \quad . \end{aligned}$$

So, both implementations (ftn_{2L} and ftn_{2L}) are associative. These implementations of `flatten` discard either the first F -effect or the second one. Since all implementations of `flatten` discard effects, identity laws cannot hold, and so the semimonads of the form $L^A \triangleq A \times F^A$ are not monads. A semimonad L combines F -effects similarly to the “trivial” semigroup (see Section 8.3.3) whose binary operation simply discards the left or the right argument. Such semigroups cannot be made into monoids since the identity laws will not hold. This construction illustrates the connections between monoids, semigroups, monads, and semimonads.

Co-products As a rule, the co-product of two monads ($F^A + G^A$) is not a monad. For simple examples, see Exercise 10.2.9.7 for $1 + F^A$ (where $F^A \triangleq A \times A$) and Exercise 10.2.9.4 for $M^A + M^A$ with an arbitrary monad M . An exception to that rule is a co-product with the *identity* monad:

Statement 10.2.8.4 If F^A is any monad, the functor $L^A \triangleq A + F^A$ is a monad. The functor L is called the **free pointed functor on F** , for reasons explained in Chapter 13.

Proof We need to define the monad methods for L , for which we may use the `pure` and `flatten` methods of F . Begin with the `flatten` method, which needs to have the type signature

$$\text{ftn}_L : L^{L^A} \rightarrow L^A = A + F^A + F^{A+F^A} \rightarrow A + F^A \quad .$$

Since we know nothing about the specific monad F , we cannot extract a value of type A out of F^A . However, we can use F ’s `pure` method to create a value of type F^A out of A . This allows us to convert $A + F^A$ into F^A using the function we will denote γ :

```
type L[A] = Either[A, F[A]]
def gamma[A]: L[A] => F[A] = {
  case Left(a) => F.pure(a)
  case Right(fa) => fa
}
```

$$\gamma^A \triangleq \begin{array}{|c|c|} \hline & F^A \\ \hline A & \text{pu}_F \\ \hline F^A & \text{id} \\ \hline \end{array} \quad .$$

Lifting this function to F , we can convert F^{A+F^A} into F^{F^A} and finally into F^A via F ’s `flatten` method:

```
def flatten_L[A]: L[L[A]] => L[A] = {
  case Left(Left(a)) => Left(a)
  case Left(Right(fa)) => Right(fa)
  case Right(g) => Right(g.map(gamma).flatten)
} // The last line equals 'Right(g.flatMap(gamma))'.
```

$$\text{ftn}_L \triangleq \begin{array}{|c|c|c|} \hline & A & F^A \\ \hline A & \text{id} & \text{id} \\ \hline F^A & \text{id} & \text{id} \\ \hline F^{L^A} & \text{id} & \gamma^F \circ \text{ftn}_F \\ \hline \end{array} \quad .$$

Is there another implementation for ftn_L ? We could have replaced A by F^A using pu_F . However, that code would never return a result of type $A + \text{id}$, which makes it impossible to satisfy identity laws such as $\text{pu}_F \circ \text{ftn}_F = \text{id}$.

The `pure` method for L could be defined in two ways: $\text{pu}_L \triangleq a^{:A} \rightarrow a + \text{id}$ or $\text{pu}_L \triangleq a \rightarrow \text{id} + \text{pu}_F(a)$. It turns out that only the first definition satisfies the monad L ’s identity laws (Exercise 10.2.9.15).

To verify the identity laws for the definition $\text{pu}_L \triangleq a^{:A} \rightarrow a + \text{id}$, begin by writing the lifting code for the functor L and a fully split matrix for pu_L^L ,

$$(f^{:A \rightarrow B})^L = \begin{array}{|c|c|} \hline & B \quad F^B \\ \hline A & f \quad \text{id} \\ \hline F^A & \text{id} \quad f^F \\ \hline \end{array} \quad , \quad \text{pu}_L^L = \begin{array}{|c|c|c|} \hline & L^A & F^{L^A} \\ \hline A & \text{id} & \text{id} \\ \hline F^A & \text{id} & \text{id} \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline & A \quad F^A \quad F^{L^A} \\ \hline A & \text{id} & \text{id} & \text{id} \\ \hline F^A & \text{id} & \text{id} & \text{id} \\ \hline \end{array} \quad .$$

Then write the two identity laws and simplify using matrix compositions:

$$\mathrm{pu}_L^{L^A} \circ \mathrm{ftn}_L = \left| \begin{array}{c|cc|c} & A & F^A & F^{L^A} \\ \hline A & \mathrm{id} & 0 & 0 \\ F^A & 0 & \mathrm{id} & 0 \end{array} \right| \circ \left| \begin{array}{c|cc} A & F^A \\ \hline \mathrm{id} & 0 \\ 0 & \mathrm{id} \\ \hline 0 & \gamma^{\uparrow F} \circ \mathrm{ftn}_F \end{array} \right| = \left| \begin{array}{c|cc} A & F^A \\ \hline \mathrm{id} & 0 \\ 0 & \mathrm{id} \end{array} \right| = \mathrm{id} ,$$

$$\text{pu}_L^{\uparrow L} ; \text{ftn}_L = \left| \begin{array}{c|cc|c} & A & F^A & F^{L^A} \\ \hline A & \text{id} & \emptyset & \emptyset \\ F^A & \emptyset & \emptyset & \text{pu}_L^{\uparrow F} \end{array} \right| ; \left| \begin{array}{c|cc} A & F^A \\ \hline \text{id} & \emptyset \\ \emptyset & \text{id} \\ \gamma^{\uparrow F} ; \text{ftn}_F & \emptyset \end{array} \right| = \left| \begin{array}{c|cc} A & F^A \\ \hline A & \text{id} \\ F^A & \emptyset \\ \emptyset & \text{pu}_L^{\uparrow F} ; \gamma^{\uparrow F} ; \text{ftn}_F \end{array} \right| .$$

We will show that the last matrix equals identity by proving that

$$\text{id} \stackrel{?}{=} \text{pu}_L^{\uparrow F} ; \gamma^{\uparrow F} ; \text{ftn}_F = (\text{pu}_L ; \gamma)^{\uparrow F} ; \text{ftn}_F \quad .$$

Let us simplify the function $pu_L \circ \gamma$ in a separate calculation:

$$p u_L \circ \gamma = \left| \begin{array}{c|cc|c} & A & F^A \\ \hline A & \text{id} & \emptyset \end{array} \right| \circ \left| \begin{array}{c|cc} & F^A \\ \hline A & p u_F \\ F^A & \text{id} \end{array} \right| = p u_F$$

It remains to show that

$$\text{id} \stackrel{?}{=} (\text{pu}_L \circ \gamma)^{\uparrow F} \circ \text{ftn}_F = \text{pu}_F^{\uparrow F} \circ \text{ftn}_F$$

But this is the right identity law of the monad F , which (by assumption) is a lawful monad.

The next step is to verify the associativity law for L . Write the code for ftn_L^L and $\text{ftn}_L^{L^A}$:

$$\text{ftn}_L^{\uparrow L} = \left| \begin{array}{c|cc} & L^A & F^{L^A} \\ \hline L^{L^A} & \text{ftn}_L & 0 \\ F^{L^{L^A}} & 0 & \text{ftn}_L^{\uparrow F} \end{array} \right| = \left| \begin{array}{c|ccc} & A & F^A & F^{L^A} \\ \hline A & \text{id} & 0 & 0 \\ F^A & 0 & \text{id} & 0 \\ F^{L^A} & 0 & \gamma^F ; \text{ftn}_F & 0 \\ F^{L^{L^A}} & 0 & 0 & \text{ftn}_L^{\uparrow F} \end{array} \right|,$$

$$\text{ftn}_L^{L^A} = \begin{array}{c|cc} & L^A & F^{L^A} \\ \hline L^A & \text{id} & \emptyset \\ F^{L^A} & \emptyset & \text{id} \\ F^{L^{L^A}} & \emptyset & \gamma^{\uparrow F} \circ \text{ftn}_F \end{array} = \begin{array}{c|ccc} & A & F^A & F^{L^A} \\ \hline A & \text{id} & \emptyset & \emptyset \\ F^A & \emptyset & \text{id} & \emptyset \\ F^{L^A} & \emptyset & \emptyset & \text{id} \\ F^{L^{L^A}} & \emptyset & \emptyset & \gamma^{\uparrow F} \circ \text{ftn}_F \end{array}$$

We are ready to verify the associativity law. Simplify both sides using matrix compositions:

$$\text{ftn}_L^{\uparrow L} ; \text{ftn}_L = \left| \begin{array}{c|ccc} & A & F^A & F^{L^A} \\ \hline A & \text{id} & \emptyset & \emptyset \\ F^A & \emptyset & \text{id} & \emptyset \\ F^{L^A} & \emptyset & \gamma^{\uparrow F} ; \text{ftn}_F & \emptyset \\ F^{L^{L^A}} & \emptyset & \emptyset & \text{ftn}_L^{\uparrow F} \end{array} \right| ; \left| \begin{array}{c|cc} & A & F^A \\ \hline A & \text{id} & \emptyset \\ F^A & \emptyset & \text{id} \\ F^{L^A} & \emptyset & \gamma^{\uparrow F} ; \text{ftn}_F \end{array} \right| = \left| \begin{array}{c|cc} & A & F^A \\ \hline A & \text{id} & \emptyset \\ F^A & \emptyset & \text{id} \\ F^{L^A} & \emptyset & \gamma^{\uparrow F} ; \text{ftn}_F \\ F^{L^{L^A}} & \emptyset & \text{ftn}_L^{\uparrow F} ; \gamma^{\uparrow F} ; \text{ftn}_F \end{array} \right| ,$$

$$\text{ftn}_L^{L^A} \circ \text{ftn}_L = \begin{array}{c|ccc} & A & F^A & F^{L^A} \\ \hline A & \text{id} & 0 & 0 \\ F^A & 0 & \text{id} & 0 \\ F^{L^A} & 0 & 0 & \text{id} \\ F^{L^{L^A}} & 0 & 0 & \gamma^{\uparrow F} \circ \text{ftn}_F \end{array} \circ \begin{array}{c|cc} & A & F^A \\ \hline A & \text{id} & 0 \\ F^A & 0 & \text{id} \\ F^{L^A} & 0 & \gamma^{\uparrow F} \circ \text{ftn}_F \end{array} = \begin{array}{c|cc} & A & F^A \\ \hline A & \text{id} & 0 \\ F^A & 0 & \text{id} \\ F^{L^A} & 0 & \gamma^{\uparrow F} \circ \text{ftn}_F \\ F^{L^{L^A}} & 0 & \gamma^{\uparrow F} \circ \text{ftn}_F \circ \gamma^{\uparrow F} \circ \text{ftn}_F \end{array}.$$

We will show that the last matrices are equal if we can prove equality of the expressions

$$\mathbf{ftn}_L^{\uparrow F} ; \gamma^{\uparrow F} ; \mathbf{ftn}_F \stackrel{?}{=} \gamma^{\uparrow F} ; \mathbf{ftn}_F ; \gamma^{\uparrow F} ; \mathbf{ftn}_F \quad ,$$

which are functions of type $F^{L^A} \rightarrow F^A$. Simplify these expressions separately:

- left-hand side : $\text{ftn}_L^F ; \gamma^{\uparrow F} ; \text{ftn}_F = (\text{ftn}_L ; \gamma)^{\uparrow F} ; \text{ftn}_F$,
- right-hand side : $\gamma^{\uparrow F} ; \underline{\text{ftn}_F ; \gamma^{\uparrow F}} ; \text{ftn}_F$
- naturality of ftn_F : $= \gamma^{\uparrow F} ; \gamma^{\uparrow F \uparrow F} ; \underline{\text{ftn}_F ; \text{ftn}_F}$
- associativity law of F : $= \gamma^{\uparrow F} ; \gamma^{\uparrow F \uparrow F} ; \underline{\text{ftn}_F^{\uparrow F} ; \text{ftn}_F} = (\gamma ; \gamma^{\uparrow F} ; \text{ftn}_F)^{\uparrow F} ; \text{ftn}_F$.

It remains to show the equality of the functions under $(\dots)^{\uparrow F}$:

$$\text{ftn}_L \circ \gamma \stackrel{?}{=} \gamma \circ \gamma^F \circ \text{ftn}_F = \gamma^L \circ \gamma \circ \text{ftn}_F \quad , \quad (10.18)$$

where in the last step we used the naturality law of $\gamma: L^A \rightarrow F^A$, which is a natural transformation:

$$\gamma \circ f^{\uparrow F} = f^{\uparrow L} \circ \gamma \quad , \quad \text{for all } f^{A \rightarrow B} \quad .$$

The functions in Eq. (10.18) are equal due to F 's identity law ($\text{pu}_F \circ \text{ftn}_F = \text{id}$):

$$\begin{aligned}
\text{ftn}_L \circ \gamma &= \left| \begin{array}{c|cc} & A & F^A \\ \hline A & \text{id} & \emptyset \\ F^A & \emptyset & \text{id} \\ F^{L^A} & \emptyset & \gamma^{\uparrow F} \circ \text{ftn}_F \end{array} \right| \circ \left| \begin{array}{c|cc} & F^A \\ \hline A & \text{pu}_F \\ F^A & \text{id} \end{array} \right| = \left| \begin{array}{c|cc} & F^A \\ \hline A & \text{pu}_F \\ F^A & \text{id} \\ F^{L^A} & \gamma^{\uparrow F} \circ \text{ftn}_F \end{array} \right| = \left| \begin{array}{c|cc} & F^A \\ \hline L^A & \gamma \\ F^{L^A} & \gamma^{\uparrow F} \circ \text{ftn}_F \end{array} \right|, \\
\gamma^{\uparrow L} \circ \gamma \circ \text{ftn}_F &= \left| \begin{array}{c|cc} & F^A & F^{F^A} \\ \hline L^A & \gamma & \emptyset \\ F^{L^A} & \emptyset & \gamma^{\uparrow F} \end{array} \right| \circ \left| \begin{array}{c|cc} & F^{F^A} \\ \hline F^A & \text{pu}_F \\ F^{F^A} & \text{id} \end{array} \right| \circ \text{ftn}_F = \left| \begin{array}{c|cc} & F^A \\ \hline L^A & \gamma \circ \text{pu}_F \circ \text{ftn}_F \\ F^{L^A} & \gamma^{\uparrow F} \circ \text{id} \circ \text{ftn}_F \end{array} \right| = \left| \begin{array}{c|cc} & F^A \\ \hline L^A & \gamma \\ F^{L^A} & \gamma^{\uparrow F} \circ \text{ftn}_F \end{array} \right|.
\end{aligned}$$

This concludes the proof.

Function types If F^\bullet is a monad, can we find a type constructor H^\bullet such that $L^A \triangleq H^A \rightarrow F^A$ is a monad? In order for L to be a functor, H must be a contrafunctor. We have already seen that $Z \rightarrow A$ is a monad (a `Reader`) when Z is a fixed type. The type expression $Z \rightarrow A$ is indeed of the form $H^A \rightarrow F^A$ where the type constructors are set to $H^A \triangleq Z$ and $F^A \triangleq A$. The following statements show two generalizations of the `Reader` monad to other H and F .

Statement 10.2.8.5 For a fixed type Z , if a functor F is a (semi)monad then so is $L^A \triangleq Z \rightarrow F^A$.

Proof As with many monads of function type, derivations are made shorter by using the flipped Kleisli trick. We assume that F 's Kleisli composition \diamond_F is known and obeys the associativity law. When F is a monad, we also assume that its pure method pu_F is known and obeys the identity laws. For the monad L , we will now define the flipped Kleisli composition $\tilde{\diamond}_L$ and prove its laws.

The ordinary Kleisli functions have types $A \rightarrow Z \rightarrow F^B$, so the flipped Kleisli functions have types $Z \rightarrow A \rightarrow F^B$. We begin by defining the flipped Kleisli composition $\tilde{\diamond}_L$ and the flipped method $\tilde{\text{pu}}_L$ that has type $Z \rightarrow A \rightarrow F^A$ (note that we use the *non-flipped* methods \diamond_F and pu_F of the monad F):

$$\begin{aligned} f: Z \rightarrow A \rightarrow F^B \quad \tilde{\diamond}_L g: Z \rightarrow B \rightarrow F^C &\triangleq z: Z \rightarrow f(z) \diamond_F g(z) \quad , \\ \tilde{\text{pu}}_L &\triangleq \underline{_}: Z \rightarrow \text{pu}_F \quad . \end{aligned}$$

To verify the laws, it is convenient to substitute an arbitrary $z: Z$ and define $\tilde{\diamond}_L$ and $\tilde{\text{pu}}_L$ by

$$z \triangleright (f \tilde{\diamond}_L g) \triangleq (z \triangleright f) \diamond_F (z \triangleright g) \quad , \quad z \triangleright \tilde{\text{pu}}_L \triangleq \text{pu}_F \quad .$$

For the left-hand side of the associativity law, we write

$$z \triangleright ((f \tilde{\diamond}_L g) \tilde{\diamond}_L h) = (z \triangleright (f \tilde{\diamond}_L g)) \diamond_F (z \triangleright h) = (z \triangleright f) \diamond_F (z \triangleright g) \diamond_F (z \triangleright h) \quad ,$$

where we omitted parentheses around \diamond_F since its associativity is given. The right-hand side:

$$z \triangleright (f \tilde{\diamond}_L (g \tilde{\diamond}_L h)) = (z \triangleright f) \diamond_F (z \triangleright (g \tilde{\diamond}_L h)) = (z \triangleright f) \diamond_F (z \triangleright g) \diamond_F (z \triangleright h) \quad .$$

We obtained the same expression for both sides of the law.

Assuming now that F is a monad with a lawful `pure` method, we verify L 's identity laws:

$$\begin{aligned} \text{left identity law of } L: \quad z \triangleright (\tilde{\text{pu}}_L \tilde{\diamond}_L f) &= (z \triangleright \tilde{\text{pu}}_L) \diamond_F (z \triangleright f) = \underline{\text{pu}_F \diamond_F} (z \triangleright f) = z \triangleright f \quad , \\ \text{right identity law of } L: \quad z \triangleright (f \tilde{\diamond}_L \tilde{\text{pu}}_L) &= (z \triangleright f) \diamond_F (z \triangleright \tilde{\text{pu}}_L) = (z \triangleright f) \diamond_F \underline{\text{pu}_F} = z \triangleright f \quad . \end{aligned}$$

It remains to convert the flipped Kleisli methods $\tilde{\diamond}_L$ and $\tilde{\text{pu}}_L$ to the code of L 's `flatMap` and `pure`:

```
type L[A] = Z => F[A]           // The type Z and a semimonad F must be already defined.
def flatMap_L[A, B](la: L[A])(f: A => L[B]): L[B] = { z => la(z).flatMap(a => f(a)(z)) }
def pure_L[A](a: A): L[A] = { _ => implicitly[Monad[F]].pure(a) }
```

Statement 10.2.8.6 For any contrafunctor H^A , the functor $L^A \triangleq H^A \rightarrow A$ is a monad.

Proof We use the flipped Kleisli formulation for L . The flipped L -Kleisli functions have types $H^B \rightarrow A \rightarrow B$ instead of $A \rightarrow H^B \rightarrow B$. The flipped Kleisli composition $\tilde{\diamond}_L$ of an $f: H^B \rightarrow A \rightarrow B$ and a $g: H^C \rightarrow B \rightarrow C$ must have type $H^C \rightarrow A \rightarrow C$. To infer this function's code, begin with a typed hole:

$$f: H^B \rightarrow A \rightarrow B \quad \tilde{\diamond}_L g: H^C \rightarrow B \rightarrow C = k: H^C \rightarrow ???: A \rightarrow C \quad .$$

Looking at the available data, we notice that a value of type $A \rightarrow C$ will be found if we apply f and g to some arguments and then compose the resulting functions of types $A \rightarrow B$ and $B \rightarrow C$:

$$???: A \rightarrow C = f(???: H^B) \circ g(???: H^C) \quad .$$

We already have a value h of type H^C , so we can apply $g(h)$ and get a function of type $B \rightarrow C$. The remaining typed hole requires a value of type H^B ; but we only have the value $k: H^C$. We cannot create values of a contrafunctor type H^B from scratch. The only way of filling that type hole is to transform k into the type H^B . This is possible by lifting a function of type $B \rightarrow C$ to the contrafunctor H , and we already have such a function, namely $g(k)$. So we write

$$???: H^B = k: H^C \triangleright (???: B \rightarrow C) \downarrow H = k \triangleright (g(k)) \downarrow H \quad .$$

Putting the entire code together and substituting an arbitrary value k^{H^C} , we get

$$k^{H^C} \triangleright (f^{H^B \rightarrow A \rightarrow B} \tilde{\circ}_L g^{H^C \rightarrow B \rightarrow C}) \triangleq f(k \triangleright (g(k))^{H^B}) \circ g(k) \quad . \quad (10.19)$$

The flipped `pure` method ($\tilde{\text{pu}}_L$) is defined by

$$\tilde{\text{pu}}_L^{H^A \rightarrow A \rightarrow A} \triangleq \underline{\text{id}^{H^A}} \rightarrow \text{id}^{A \rightarrow A} \quad , \quad \tilde{\text{pu}}_L(k) = \text{id} \quad . \quad (10.20)$$

These code formulas can be converted to the following Scala definitions of `pure` and `flatMap`:

```
type L[A] = H[A] => A           // The contrafunctor H must be already defined.
def pure_L[A](a: A): L[A] = { _ => a }
def flatMap_L[B, C](lb: L[B])(g: B => L[C]): L[C] = { (k: H[C]) =>
  val bc: (B => C) = { b => g(b)(k) } // Corresponds to g(k): B → C.
  val hb: H[B] = k.contramap(bc)        // Fill the typed hole ???: H[B].
  g(lb(hb))(k)                         // Corresponds to the composition f(...): g(k).
}
```

Let us now verify the monad laws of L in the flipped Kleisli form. The identity laws:

$$\begin{aligned} \text{expect } g(k) : k \triangleright (\tilde{\text{pu}}_L \tilde{\circ}_L g) &= \tilde{\text{pu}}_L(k \triangleright (g(k))^{H^B}) \circ g(k) = \text{id} \circ g(k) = g(k) \quad , \\ \text{expect } f(k) : k \triangleright (f \tilde{\circ}_L \tilde{\text{pu}}_L) &= f(k \triangleright (\tilde{\text{pu}}_L(k))^{H^B}) \circ \tilde{\text{pu}}_L(k) = f(k \triangleright \text{id}^{H^B}) \circ \text{id} = f(k) \quad . \end{aligned}$$

To verify the associativity law, write the two sides separately and use Eq. (10.19) repeatedly:

$$\begin{aligned} k^{H^C} \triangleright ((f \tilde{\circ}_L g) \tilde{\circ}_L h) &= ((f \tilde{\circ}_L g)(k \triangleright (h(k))^{H^B})) \circ h(k) = (k \triangleright (h(k))^{H^B} \triangleright (f \tilde{\circ}_L g)) \circ h(k) \\ &= f(k \triangleright (h(k))^{H^B} \triangleright (g(k \triangleright (h(k))^{H^B}))^{H^B}) \circ g(k \triangleright (h(k))^{H^B}) \circ h(k) \quad , \\ k \triangleright (f \tilde{\circ}_L (g \tilde{\circ}_L h)) &= f(k \triangleright ((g \tilde{\circ}_L h)(k))^{H^B}) \circ (g \tilde{\circ}_L h)(k) = f(k \triangleright (k \triangleright (g \tilde{\circ}_L h))^{H^B}) \circ k \triangleright (g \tilde{\circ}_L h) \\ &= f(k \triangleright (g(k \triangleright (h(k))^{H^B})) \circ h(k))^{H^B}) \circ g(k \triangleright (h(k))^{H^B}) \circ h(k) \\ \text{lift to } H : &= f(k \triangleright (h(k))^{H^B} \triangleright (g(k \triangleright (h(k))^{H^B}))^{H^B}) \circ g(k \triangleright (h(k))^{H^B}) \circ h(k) \quad . \end{aligned}$$

Now both sides are rewritten into identical expressions. This confirms the associativity law for L .

Statement 10.2.8.6 applies to type constructors of the form $H^A \rightarrow A$ but not to $H^A \rightarrow G^A$ with an arbitrary monad G . It turns out that $H^A \rightarrow G^A$ is a monad when the contrafunctor H is suitably adapted to the monad G . The necessary properties of H are formulated by the notion of G -filterable contrafunctor, which we briefly introduced in Section 9.4.3. We repeat the definition here:

For a given monad G , a contrafunctor H is **G -filterable** if it has the `lift` method,

$$\text{lift}_{G,H} : (A \rightarrow G^B) \rightarrow H^B \rightarrow H^A \quad ,$$

satisfying the identity and (contravariant) composition laws

$$\text{lift}_{G,H}(\text{pu}_G) = \text{id}^{H^A \rightarrow H^A} \quad , \quad \text{lift}_{G,H}(f \diamond_G g) = \text{lift}_{G,H}(g) \circ \text{lift}_{G,H}(f) \quad .$$

Let us look at some simple examples of G -filterable contrafunctors:

Statement 10.2.8.7 For any given (semi)monad G and any fixed type Z , the following contrafunctors are G -filterable: (a) $H^A \triangleq Z$, (b) $H^A \triangleq G^A \rightarrow Z$, (c) $H^A \triangleq A \rightarrow G^Z$.

Proof In each case, we need to define the function $\text{lift}_{G,H}$ and verify its laws. If G is only a semimonad, we will derive the composition law of $\text{lift}_{G,H}$ assuming only the associativity law of G (but not assuming any identity laws). For a full monad G , we will derive the identity law of $\text{lift}_{G,H}$ by using the identity laws of G .

(a) Since $H^A \triangleq Z$ is a constant contrafunctor, we define $\text{lift}_{G,H}(f) \triangleq \text{id}^{Z \rightarrow Z}$. The identity function will always satisfy the identity and composition laws.

(b) Define the function $\text{lift}_{G,H}$ by

```
type H[A] = G[A] => Z      // The type Z and the monad G must be already defined.
def lift_G[A, B](k: H[B])(f: A => G[B]): H[A] = { (ga: G[A]) => k(ga.flatMap(f)) }
```

$$\text{lift}_{G,H}(f: A \rightarrow G^B) \triangleq k: G^B \rightarrow Z \rightarrow ???: G^A \rightarrow Z = k: G^B \rightarrow Z \rightarrow \text{flm}_G(f) \circ k \quad , \quad k \triangleright \text{lift}_{G,H}(f) \triangleq \text{flm}_G(f) \circ k \quad .$$

Verify the laws of $\text{lift}_{G,H}$ using the right identity and the associativity law of G 's `flatMap`:

$$\text{identity law, expect to equal } k : k \triangleright \text{lift}_{G,H}(\text{pu}_G) = \text{flm}_G(\text{pu}_G) \circ k = \text{id} \circ k = k \quad ,$$

$$\text{left-hand side of composition law : } k \triangleright \text{lift}_{G,H}(f \diamond_G g) = \text{flm}_G(f \diamond_G g) \circ k = \text{flm}_G(f \circ \text{flm}_G(g)) \circ k$$

$$\text{associativity law of } \text{flm}_G : = \text{flm}_G(f) \circ \text{flm}_G(g) \circ k \quad ,$$

$$\begin{aligned} \text{right-hand side of composition law : } k \triangleright \text{lift}_{G,H}(g) \circ \text{lift}_{G,H}(f) &= k \triangleright \text{lift}_{G,H}(g) \triangleright \text{lift}_{G,H}(f) \\ &= \text{flm}_G(f) \circ (k \triangleright \text{lift}_{G,H}(g)) = \text{flm}_G(f) \circ \text{flm}_G(g) \circ k \quad . \end{aligned}$$

Both sides of the composition law of $\text{lift}_{G,H}$ are equal.

(c) Define the function $\text{lift}_{G,H}$ by

```
type H[A] = A => G[Z] // The type Z and the monad G must be already defined.
def lift_G[A, B](k: H[B])(f: A => G[B]): H[A] = { (a: A) => f(a).flatMap(k) }
```

$$\text{lift}_{G,H}(f: A \rightarrow G^B) \triangleq k: B \rightarrow G^Z \rightarrow ???: A \rightarrow G^Z = k: G^B \rightarrow Z \rightarrow f \diamond_G k \quad , \quad k \triangleright \text{lift}_{G,H}(f) \triangleq f \diamond_G k \quad .$$

The laws of $\text{lift}_{G,H}$ are satisfied because the operation \diamond_G obeys its identity and associativity laws:

$$\text{identity law, expect to equal } k : k \triangleright \text{lift}_{G,H}(\text{pu}_G) = \text{pu}_G \diamond_G k = k \quad ,$$

$$\text{left-hand side of composition law : } k \triangleright \text{lift}_{G,H}(f \diamond_G g) = (f \diamond_G g) \diamond_G k \quad ,$$

$$\begin{aligned} \text{right-hand side of composition law : } k \triangleright \text{lift}_{G,H}(g) \circ \text{lift}_{G,H}(f) &= k \triangleright \text{lift}_{G,H}(g) \triangleright \text{lift}_{G,H}(f) \\ &= f \diamond_G (k \triangleright \text{lift}_{G,H}(g)) = f \diamond_G (g \diamond_G k) \quad . \end{aligned}$$

Both sides of the composition law of $\text{lift}_{G,H}$ are equal due to associativity of \diamond_G .

We will now use filterable contrafunctors to generalize Statements 10.2.8.5 and 10.2.8.6.

Statement 10.2.8.8 For any (semi)monad G and any G -filterable contrafunctor H , the functor L defined by $L^A \triangleq H^A \rightarrow G^A$ is a (semi)monad.

Proof We use the flipped Kleisli formulation for L and the ordinary Kleisli formulation for G . The flipped L -Kleisli functions have types $H^B \rightarrow A \rightarrow G^B$ instead of $A \rightarrow H^B \rightarrow G^B$. The flipped Kleisli composition \circ_L of an $f: H^B \rightarrow A \rightarrow G^B$ and a $g: H^C \rightarrow B \rightarrow G^C$ must have type $H^C \rightarrow A \rightarrow G^C$. Similarly to what we did in the proof of Statement 10.2.8.6, we will define \circ_L by applying f and g to typed holes, getting values of types $A \rightarrow G^B$ and $B \rightarrow G^C$, and using the given operation \diamond_G :

$$f: H^B \rightarrow A \rightarrow G^B \circ_L g: H^C \rightarrow B \rightarrow G^C = k: H^C \rightarrow ???: A \rightarrow G^C = k: H^C \rightarrow f(???: H^B) \diamond_G g(???: H^C) \quad .$$

The typed hole $???: H^C$ is filled by k , while $???: H^B$ is obtained from k using $\text{lift}_{G,H}$:

$$???: H^B = k \triangleright \text{lift}_{G,H}(???: B \rightarrow G^C) = k \triangleright \text{lift}_{G,H}(g(k)) \quad .$$

Putting the code together, we obtain

$$\begin{aligned} k: H^C \triangleright (f: H^B \rightarrow A \rightarrow G^B \circ_L g: H^C \rightarrow B \rightarrow G^C) &\triangleq f(k \triangleright \text{lift}_{G,H}(g(k))) \diamond_G g(k) \quad , \\ k: H^C \triangleright \tilde{\text{pu}}_L &\triangleq \text{pu}_G \quad . \end{aligned}$$

These definitions reduce to Eqs. (10.19)–(10.20) if we set G to the identity monad, $G^A = \text{Id}^A \triangleq A$. Any contrafunctor H is Id -filterable because we can define the required method $\text{lift}_{\text{Id},H}$ by

$$\text{lift}_{\text{Id},H}(f: B \rightarrow C) \triangleq f \downarrow H \quad .$$

The laws of $\text{lift}_{\text{Id},H}$ follow from the identity and composition laws of the contrafunctor H .

Translating the monad L 's methods into Scala code, we obtain:

```

type L[A] = H[A] => G[A] // The contrafunctor H and the monad G must be already defined.
def flatMap_L[A, B](la: L[A])(g: A => L[B]): L[B] = { (k: H[B]) =>
  val agb: (A => G[B]) = { a => g(a)(k) } // Corresponds to flipped g(k).
  val ha: H[A] = lift_G(k)(agb) // Corresponds to k > lift_{G,H}(g(k)).
  la(ha).flatMap(agb) // Corresponds to the Kleisli composition f(...) ◊_G g(k).
}

```

We now verify the laws via derivations quite similar to those in the proof of Statement 10.2.8.6.

We will need to use the laws of $\text{lift}_{G,H}$, which we will write simply as “lift” for brevity. The identity laws of L follow from the identity laws of $\text{lift}_{G,H}$ and G :

$$\begin{aligned} \text{expect } g(k) : k \triangleright (\tilde{p}u_L \tilde{\diamond}_L g) &= \tilde{p}u_L(k \triangleright \text{lift}(g(k))) \diamond_G g(k) = pu_G \diamond_G g(k) = g(k) , \\ \text{expect } f(k) : k \triangleright (f \tilde{\diamond}_L \tilde{p}u_L) &= f(k \triangleright \text{lift}(\tilde{p}u_L(k))) ; \tilde{p}u_L(k) = f(k \triangleright \text{id}) \diamond_G pu_G = f(k) . \end{aligned}$$

The associativity law of L follows from the associativity of \diamond_G and the composition law of $\text{lift}_{G,H}$:

$$\begin{aligned} k^{H^C} \triangleright ((f \tilde{\diamond}_L g) \tilde{\diamond}_L h) &= ((f \tilde{\diamond}_L g)(k \triangleright \text{lift}(h(k)))) \diamond_G h(k) = (k \triangleright \text{lift}(h(k)) \triangleright (f \tilde{\diamond}_L g)) \diamond_G h(k) \\ &= f(k \triangleright \text{lift}(h(k)) \triangleright \text{lift}(g(k \triangleright \text{lift}(h(k))))) \diamond_G g(k \triangleright \text{lift}(h(k))) \diamond_G h(k) , \\ k \triangleright (f \tilde{\diamond}_L (g \tilde{\diamond}_L h)) &= f(k \triangleright \text{lift}((g \tilde{\diamond}_L h)(k))) \diamond_G (g \tilde{\diamond}_L h)(k) = f(k \triangleright \text{lift}(k \triangleright (g \tilde{\diamond}_L h))) \diamond_G k \triangleright (g \tilde{\diamond}_L h) \\ &= f(k \triangleright \text{lift}(g(k \triangleright \text{lift}(h(k)) \diamond_G h(k))) \diamond_G g(k \triangleright \text{lift}(h(k))) \diamond_G h(k) \\ &= f(k \triangleright \text{lift}(h(k)) ; \text{lift}(g(k \triangleright \text{lift}(h(k))))) \diamond_G g(k \triangleright \text{lift}(h(k))) \diamond_G h(k) . \end{aligned}$$

In the \triangleright -notation, we have $k \triangleright p \triangleright q = k \triangleright p ; q$. Both sides are now rewritten into identical expressions. Some examples of monads obtained with these constructions for $G^A \triangleq \mathbb{1} + A$ are

$$\text{Sel}^{Z,A} \triangleq (A \rightarrow Z) \rightarrow A , \quad \text{Search}^{Z,A} \triangleq (A \rightarrow \mathbb{1} + Z) \rightarrow \mathbb{1} + A , \quad L^A \triangleq (\mathbb{1} + A \rightarrow Z) \rightarrow \mathbb{1} + A .$$

The first two of these monads are the selector monad and the search monad.⁹ Statement 10.2.8.8 shows how to implement the monad methods for these monads and guarantees that the laws hold.

Recursive types The tree-like monads can be generalized to arbitrary shape of leaves and branches. This gives two constructions that define a monad and a semimonad from an arbitrary functor.

Statement 10.2.8.9 For any given functor F , the recursive functor $L^A \triangleq A + F^{L^A}$ is a monad called the **free monad on F** (for reasons explained in Chapter 13).

Proof Begin by implementing the monad methods for L as usual for a tree-like monad:

```

type L[A] = Either[A, F[L[A]]] // The functor F must be already defined.
def pure_L[A](a: A): L[A] = Left(a)
def flatMap_L[A, B](la: L[A])(f: A => L[B]): L[B] = la match {
  case Left(a) => f(a)
  case Right(fla) => Right(fla.map { (x: L[A]) => flatMap_L(x)(f) }) // Recursive call of flatMap_L.
}
def flatten_L[A]: L[L[A]] => L[A] = { // Match on L[L[A]] = Either[Either[A, F[L[A]]], F[L[L[A]]]].
  case Left(a) => Left(a)
  case Left(Right(fla)) => Right(fla)
  case Right(Right(flla)) => Right(flla.map(x => flatten_L(x))) // Recursive call of flatten_L.
}

```

$$\begin{aligned} pu_L &\triangleq \begin{array}{c|cc} & A & F^{L^A} \\ \hline A & id & \mathbf{0} \end{array} , & ftn_L^A: L^{L^A} \rightarrow L^A &\triangleq \begin{array}{c|cc} & A & F^{L^A} \\ \hline A & id & \mathbf{0} \\ F^{L^A} & \mathbf{0} & id \\ F^{L^{L^A}} & \mathbf{0} & \overline{ftn_L^A} \end{array} . \end{aligned}$$

⁹See <http://math.andrej.com/2008/11/21/a-haskell-monad-for-infinite-search-in-finite-time/>

To verify the monad laws for L , we need the lift pu_L and ftn_L to the functor L . The lifting code is

$$(f:A \rightarrow B)^{\uparrow L} = \begin{vmatrix} & \parallel B & F^{L^B} \\ & \parallel f & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel (f^{\uparrow L})^{\uparrow F} \end{vmatrix}.$$

In order to compute function compositions such as $\text{pu}_L^{\uparrow L} ; \text{ftn}_L$ and $\text{ftn}_L^{\uparrow L} ; \text{ftn}_L$ in the matrix notation, we need to expand the matrices of pu_L and ftn_L , so that their output types are fully split:

$$\begin{aligned} \text{pu}_L^{\uparrow L} &= \begin{vmatrix} & \parallel L^A & F^{L^A} \\ & \parallel \text{pu}_L & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel (\text{pu}_L^{\uparrow L})^{\uparrow F} \end{vmatrix} = \begin{vmatrix} & \parallel A & F^{L^A} & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel (\text{pu}_L^{\uparrow L})^{\uparrow F} \end{vmatrix}, \\ \text{ftn}_L^{\uparrow L} &= \begin{vmatrix} & \parallel L^A & F^{L^A} \\ & \parallel \text{ftn}_L & \mathbf{0} \\ \parallel L^A & \parallel & \parallel \\ & \parallel F^{L^L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel (\text{ftn}_L^{\uparrow L})^{\uparrow F} \end{vmatrix} = \begin{vmatrix} & \parallel A & F^{L^A} & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} & \mathbf{0} \\ & \parallel \mathbf{0} & \parallel \text{id} & \mathbf{0} \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \overline{\text{ftn}}_L^{\uparrow F} & \mathbf{0} \\ & \parallel F^{L^L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \mathbf{0} & (\text{ftn}_L^{\uparrow L})^{\uparrow F} \end{vmatrix}. \end{aligned}$$

We can now verify the identity and associativity laws of pu_L and ftn_L . Identity laws:

$$\begin{aligned} \text{pu}_L^{L^A} ; \text{ftn}_L &= \begin{vmatrix} & \parallel A & F^{L^A} & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \text{id} & \mathbf{0} \end{vmatrix} ; \begin{vmatrix} & \parallel A & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \overline{\text{ftn}}_L^{\uparrow F} \end{vmatrix} = \begin{vmatrix} & \parallel A & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \text{id} \end{vmatrix} = \text{id}, \\ \text{pu}_L^{\uparrow L} ; \text{ftn}_L &= \begin{vmatrix} & \parallel A & F^{L^A} & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \mathbf{0} & (\text{pu}_L^{\uparrow L})^{\uparrow F} \end{vmatrix} ; \begin{vmatrix} & \parallel A & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \overline{\text{ftn}}_L^{\uparrow F} \end{vmatrix} \\ &= \begin{vmatrix} & \parallel A & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel (\text{pu}_L^{\uparrow L})^{\uparrow F} ; \overline{\text{ftn}}_L^{\uparrow F} \end{vmatrix} = \begin{vmatrix} & \parallel A & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \text{id} \end{vmatrix} = \text{id}, \end{aligned}$$

where in the last line we used the inductive assumption $(\text{pu}_L^{\uparrow L})^{\uparrow F} ; \overline{\text{ftn}}_L^{\uparrow F} = \text{id}$.

To verify the associativity law, simplify its both sides separately and use fully split matrices:

$$\text{ftn}_L^{L^A} ; \text{ftn}_L = \begin{vmatrix} & \parallel A & F^{L^A} & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \text{id} & \mathbf{0} \\ \parallel F^{L^A} & \parallel & \parallel \\ & \parallel \mathbf{0} & \parallel \mathbf{0} & \overline{\text{ftn}}_L^{\uparrow F} \end{vmatrix} ; \begin{vmatrix} & \parallel A & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \overline{\text{ftn}}_L^{\uparrow F} \end{vmatrix} = \begin{vmatrix} & \parallel A & F^{L^A} \\ & \parallel \text{id} & \mathbf{0} \\ \parallel A & \parallel & \parallel \\ & \parallel F^{L^A} & \parallel \\ & \parallel \mathbf{0} & \parallel \text{id} & \mathbf{0} \\ \parallel F^{L^A} & \parallel & \parallel \\ & \parallel \mathbf{0} & \parallel \overline{\text{ftn}}_L^{\uparrow F} & \overline{\text{ftn}}_L^{\uparrow F} \end{vmatrix},$$

$$\begin{array}{c}
 \text{ftn}_L^{\uparrow L} ; \text{ftn}_L = \\
 \left| \begin{array}{c|ccc}
 & A & F^{L^A} & F^{L^{L^A}} \\
 \hline
 A & \text{id} & 0 & 0 \\
 F^{L^A} & 0 & \text{id} & 0 \\
 F^{L^{L^A}} & 0 & \overline{\text{ftn}}_L^{\uparrow F} & 0 \\
 F^{L^{L^{L^A}}} & 0 & 0 & (\text{ftn}_L^{\uparrow L})^{\uparrow F}
 \end{array} \right| ; \left| \begin{array}{c|cc}
 & A & F^{L^A} \\
 \hline
 A & \text{id} & 0 \\
 F^{L^A} & 0 & \text{id} \\
 F^{L^{L^A}} & 0 & \overline{\text{ftn}}_L^{\uparrow F}
 \end{array} \right| = \\
 \left| \begin{array}{c|cc}
 & A & F^{L^A} \\
 \hline
 A & \text{id} & 0 \\
 F^{L^A} & 0 & \text{id} \\
 F^{L^{L^A}} & 0 & \overline{\text{ftn}}_L^{\uparrow F} \\
 F^{L^{L^{L^A}}} & 0 & (\text{ftn}_L^{\uparrow L})^{\uparrow F} ; \overline{\text{ftn}}_L^{\uparrow F}
 \end{array} \right|
 \end{array} .$$

The two resulting matrices are equal due to the inductive assumption,

$$\overline{\text{ftn}}_L^{\uparrow F} ; \overline{\text{ftn}}_L^{\uparrow F} = (\text{ftn}_L^{\uparrow L})^{\uparrow F} ; \overline{\text{ftn}}_L^{\uparrow F} .$$

Statement 10.2.8.10 For any functor F , the recursive functor $L^A \triangleq F^A + F^{L^A}$ (representing a tree with F -shaped leaves and branches) is a semimonad.

Proof We have already implemented the `flatMap` method for an equivalent functor in Section 10.1.4. The corresponding `flatten` method is

```

type L[A] = Either[F[A], F[L[A]]]      // The functor F must be already defined.
def flatten_L[A]: L[L[A]] => L[A] = { // Match on L[L[A]] = Either[F[L[A]], F[L[L[A]]]].
  case Left(fla)  => Right(fla)
  case Right(flla) => Right(flla.map(x => flatten_L(x)))
}
    
```

Prepare the code of `flatten` and its lifting to L in the matrix notation:

$$\begin{array}{c}
 \text{ftn}_L^A = \left| \begin{array}{c|cc}
 & F^A & F^{L^A} \\
 \hline
 F^{L^A} & 0 & \text{id} \\
 F^{L^{L^A}} & 0 & \overline{\text{ftn}}_L^{\uparrow F}
 \end{array} \right|, \quad \text{ftn}_L^{L^A} = \left| \begin{array}{c|cc}
 & F^{L^A} & F^{L^{L^A}} \\
 \hline
 F^{L^{L^A}} & 0 & \text{id} \\
 F^{L^{L^{L^A}}} & 0 & \overline{\text{ftn}}_L^{\uparrow F}
 \end{array} \right|, \\
 (f^{A \rightarrow B})^{\uparrow L} = \left| \begin{array}{c|cc}
 & F^B & F^{L^B} \\
 \hline
 F^A & f^{\uparrow F} & 0 \\
 F^{L^A} & 0 & (f^{\uparrow L})^{\uparrow F}
 \end{array} \right|, \quad \text{ftn}_L^{\uparrow L} = \left| \begin{array}{c|cc}
 & F^{L^A} & F^{L^{L^A}} \\
 \hline
 F^{L^{L^A}} & \overline{\text{ftn}}_L^{\uparrow F} & 0 \\
 F^{L^{L^{L^A}}} & 0 & (\text{ftn}_L^{\uparrow L})^{\uparrow F}
 \end{array} \right|
 \end{array} .$$

To verify the associativity law, write both its parts separately:

$$\begin{array}{c}
 \text{ftn}_L^{L^A} ; \text{ftn}_L = \left| \begin{array}{c|cc}
 & F^{L^A} & F^{L^{L^A}} \\
 \hline
 F^{L^{L^A}} & 0 & \text{id} \\
 F^{L^{L^{L^A}}} & 0 & \overline{\text{ftn}}_L^{\uparrow F}
 \end{array} \right| ; \left| \begin{array}{c|cc}
 & F^A & F^{L^A} \\
 \hline
 F^{L^A} & 0 & \text{id} \\
 F^{L^{L^A}} & 0 & \overline{\text{ftn}}_L^{\uparrow F}
 \end{array} \right| = \left| \begin{array}{c|cc}
 & F^A & F^{L^A} \\
 \hline
 F^{L^{L^A}} & 0 & \overline{\text{ftn}}_L^{\uparrow F} \\
 F^{L^{L^{L^A}}} & 0 & \overline{\text{ftn}}_L^{\uparrow F} ; \overline{\text{ftn}}_L^{\uparrow F}
 \end{array} \right|, \\
 \text{ftn}_L^{\uparrow L} ; \text{ftn}_L = \left| \begin{array}{c|cc}
 & F^{L^A} & F^{L^{L^A}} \\
 \hline
 F^{L^{L^A}} & \overline{\text{ftn}}_L^{\uparrow F} & 0 \\
 F^{L^{L^{L^A}}} & 0 & (\text{ftn}_L^{\uparrow L})^{\uparrow F}
 \end{array} \right| ; \left| \begin{array}{c|cc}
 & F^A & F^{L^A} \\
 \hline
 F^{L^A} & 0 & \text{id} \\
 F^{L^{L^A}} & 0 & \overline{\text{ftn}}_L^{\uparrow F}
 \end{array} \right| = \left| \begin{array}{c|cc}
 & F^A & F^{L^A} \\
 \hline
 F^{L^{L^A}} & 0 & \overline{\text{ftn}}_L^{\uparrow F} \\
 F^{L^{L^{L^A}}} & 0 & (\text{ftn}_L^{\uparrow L})^{\uparrow F} ; \overline{\text{ftn}}_L^{\uparrow F}
 \end{array} \right|
 \end{array} .$$

The two resulting matrices are equal due to the inductive assumption,

$$\overline{\text{ftn}}_L^{\uparrow F} ; \overline{\text{ftn}}_L^{\uparrow F} = (\text{ftn}_L^{\uparrow L})^{\uparrow F} ; \overline{\text{ftn}}_L^{\uparrow F} .$$

Could L be a full monad? Suppose we were given some code for the `pure` method (pu_L). If we substitute an arbitrary value k^{L^A} into the left identity law, we expect to obtain again k :

$$k^{L^A} \triangleright \text{pu}_L ; \text{ftn}_L = k \triangleright \text{pu}_L \triangleright \left| \begin{array}{c|cc}
 & F^A & F^{L^A} \\
 \hline
 F^{L^A} & 0 & \text{id} \\
 F^{L^{L^A}} & 0 & \overline{\text{ftn}}_L^{\uparrow F}
 \end{array} \right| \stackrel{?}{=} k .$$

We notice that the matrix for ftn_L has no elements in the F^A column. This suggests information loss: the value $k \triangleright \text{pu}_L \triangleright \text{ftn}_L$ will be always of the form $0 + q^{F^{L^A}}$ with some q , regardless of k and of the implementation of pu_L . So, if we set $k \triangleq p^{F^A} + 0$ with some p , the law cannot possibly hold. The same problem occurs with the right identity law. We conclude that L cannot be a full monad.

10.2.9 Exercises: laws and structure of monads

Exercise 10.2.9.1 For fixed types U and V , define `flatten` for the semimonad $F^A \triangleq A \times U \times V$ as

$$\text{ftn}_F \triangleq (a:A \times u_1^U \times v_1^V) \times u_2^U \times v_2^V \rightarrow a \times u_1 \times v_2 \quad .$$

A `pure` method could be defined for F if some fixed values u_0^U and v_0^V were given:

$$\text{pu}_F \triangleq a:A \rightarrow a \times u_0 \times v_0 \quad .$$

Show that F would still not be a full monad because identity laws would not hold.

Exercise 10.2.9.2 Suppose a monad M is defined using the `pure` and `flatten` methods that satisfy all laws, in particular, the left identity law $\text{pu}_M \circ \text{ftn}_M = \text{id}$. The `flatMap` function is then defined through `flatten` by $\text{flm}_M(f) \triangleq f^{\uparrow M} \circ \text{ftn}_M$. Show that this `flatMap` function will satisfy its left identity law, $\text{pu}_M \circ \text{flm}_M(f) = f$, for any $f:A \rightarrow M^B$.

Exercise 10.2.9.3 Verify the associativity law for the `Reader` monad by a fully explicit calculation.

Exercise 10.2.9.4 Given an arbitrary monad M , show that the functor $F^A \triangleq 2 \times M^A$ (which is equivalent to $M^A + M^A$) is a semimonad but not a full monad.

Exercise 10.2.9.5 Show that $P^A \triangleq Z + W \times A$ is a (full) monad if W is a monoid.

Exercise 10.2.9.6 If W and R are arbitrary fixed types, is any of the functors $F^A \triangleq W \times (R \rightarrow A)$ and $G^A \triangleq R \rightarrow (W \times A)$ a lawful semimonad? Hint: try to define `flatten` for these functors.

Exercise 10.2.9.7 Show that $F^A \triangleq (P \rightarrow A) + (Q \rightarrow A)$ is not a semimonad (cannot define `flatMap`) when P and Q are fixed types that are different (and not related by subtyping).

Exercise 10.2.9.8* Consider the functor $D^A \triangleq \mathbb{1} + A \times A$ (in Scala, `type D[A] = Option[(A, A)]`). Implement the `flatten` and `pure` methods for D in two or more different ways, and show that some of the monad laws fail to hold for each of those implementations.¹⁰

Exercise 10.2.9.9 For the type constructor $F^A \triangleq Z \times (A \rightarrow Z)$, a programmer defined `flatMap` by

```
type F[A] = (Z, A => Z) // The type Z must be already defined.
def flatMap_F[A, B](fa: F[A])(f: A => F[B]): F[B] = fa match { case (z, g) => (z, _ => z) }
```

$$\text{flm}_F(f:A \rightarrow Z \times (B \rightarrow Z)) \triangleq z:Z \times g:A \rightarrow Z \rightarrow z \times (_^B \rightarrow z) \quad .$$

Show that this implementation fails to satisfy the semimonad laws. (Here, Z is a fixed type.)

Exercise 10.2.9.10* Prove (by induction on p) that `flatten` and `concat` satisfy the distributive law,

$$(p \text{++} q) \triangleright \text{ftn} = (p \triangleright \text{ftn}) \text{++} (q \triangleright \text{ftn}) \quad , \quad \forall p : \text{List}^A, q : \text{List}^A \quad .$$

Exercise 10.2.9.11 Given a monad M , consider a function `toUnit` (denoted tu_M) with type signature

```
def toUnit[M[_]: Monad, A]: M[A] => M[Unit] = \_.map(_ => ()) \quad \text{tu}_M : M^A \rightarrow M^1 \triangleq (\_^A \rightarrow 1)^{\uparrow M} \quad .
```

Show that this function satisfies the equation

$$\text{pu}_M^{A \rightarrow M^A} \circ \text{tu}_M^{M^A \rightarrow M^1} = _^A \rightarrow \text{pu}_M(1) \quad .$$

¹⁰One can prove that $D^A \triangleq \mathbb{1} + A \times A$ cannot be a lawful monad; see <https://stackoverflow.com/questions/49742377/>

Exercise 10.2.9.12 Show that $M[W]$ is a monoid if $M[_]$ is a monad and W is a monoid.

Exercise 10.2.9.13 A monad is called **commutative** if swapping the order of its effects gives the same results. In terms of functor block code, this means that `result1` and `result2` are equal:

```
result1 = for { // Assume p: M[A], q: M[B],
  x <- p    // and q does not depend on x.
  y <- q
} yield f(x, y) // For any f: (A, B) => C.
```

```
result2 = for { // Assume p: M[A], q: M[B],
  y <- q    // and p does not depend on y.
  x <- p
} yield f(x, y) // For any f: (A, B) => C.
```

(a) Check whether the `Option`, `List`, `Reader`, `Writer`, and `State` monads are commutative.

(b) Consider the type M^1 (in Scala, this is `M[Unit]`). Since the `Unit` type is a (trivial) monoid, Exercise 10.2.9.12 shows that M^1 is also a monoid. Prove that if the monad M is commutative then the binary operation \oplus_M of the monoid M^1 is also commutative.

Exercise 10.2.9.14 Use monad constructions (so that law checking is unnecessary) to implement `flatten` and `pure` for $F^A \triangleq A + (R \rightarrow A)$, where R is a fixed type.

Exercise 10.2.9.15 Consider the alternative way of defining the `pure` method for the monad construction $L^A \triangleq A + G^A$ (Statement 10.2.8.4), where G is a monad. Show that the identity laws will not hold for L if `pure` is defined as `a => Right(Monad[G].pure(a))` (in the code notation, $\text{pu}_L \triangleq a \rightarrow \mathbb{0} + a \triangleright \text{pu}_G$).

Exercise 10.2.9.16 Implement the functor and monad methods for $F^A \triangleq (Z \rightarrow \mathbb{1} + A) \times \text{List}^A$ using the known constructions (so that law checking is unnecessary).

Exercise 10.2.9.17 Use monad constructions to show that the functors F defined below are monads:

(a) $F^A \triangleq A + A + \dots + A$ and $F^A \triangleq A \times A \times \dots \times A$ (with finitely many A 's).

(b) A polynomial functor $F^A \triangleq p(A)$ when $p(x)$ is a polynomial of the form $p(x) = x^{n_1} + x^{n_2} + \dots + x^{n_k}$ for some positive integers n_1, \dots, n_k ; for example, $F^A \triangleq A \times A \times A + A \times A \times A \times A \times A + A \times A$.

Exercise 10.2.9.18** If M is a monad, show that $L^A \triangleq M^{M^A}$ is a semimonad but not a full monad.

10.3 Discussion and further developments

10.3.1 Why monads must be functors

We defined monads as functors with extra properties. It turns out that monads cannot be contra-functors, and that the functor laws of a monad can be derived from the laws of `flatMap`.

A monad's code can be specified in three equivalent ways: via `flatMap` and `pure`, via `flatten` and `pure`, or via the Kleisli composition and `pure`. These methods have different laws and different uses. The `flatMap` method is used most frequently, while the Kleisli composition is rarely used. Since

$$\text{flm}_M(f) = f^{\uparrow M} \circ \text{ftn}_M ,$$

we could imagine that `flatMap` already has the capability of lifting a function ($f^{\uparrow M}$). Let us extract that capability from the code of `flatMap`: compose $f^{\uparrow M}$ with the right identity law,

$$\text{pu}_M^{\uparrow M} \circ \text{ftn}_M = \text{id} , \quad \text{so} \quad f^{\uparrow M} \circ \underline{\text{pu}_M^{\uparrow M} \circ \text{ftn}_M} = f^{\uparrow M} .$$

The left-hand side equals $\text{flm}_M(f \circ \text{pu}_M)$. We found a formula that expresses `map` through `flatMap`:

```
p.map(f) == p.flatMap(f andThen M.pure)
```

$$f^{\uparrow M} = \text{flm}_M(f \circ \text{pu}_M) .$$

The laws of `map` will then follow if we assume that `flatMap` is lawful. The identity law of `map`:

$$\text{id}^{\uparrow M} = \text{flm}_M(\text{id} \circ \text{pu}_M) = \underline{\text{flm}_M(\text{pu}_M)}$$

right identity law of flm_M : $= \text{id} .$

The composition law of `map`:

$$\begin{aligned} \text{expect to equal } (f \circ g)^{\uparrow M} : & f^{\uparrow M} \circ g^{\uparrow M} = \text{flm}_M(f \circ \text{pu}_M) \circ \text{flm}_M(g \circ \text{pu}_M) \\ \text{associativity law of } \text{flm}_M : & = \text{flm}_M(f \circ \text{pu}_M \circ \text{flm}_M(g \circ \text{pu}_M)) \\ \text{left identity law of } \text{flm}_M : & = \text{flm}_M(f \circ g \circ \text{pu}_M) = (f \circ g)^{\uparrow M} . \end{aligned}$$

If a monad M is specified via its Kleisli composition \diamond_M , we can first define `flatMap` by

$$\text{flm}_M(f : A \rightarrow M^B) \triangleq \text{id}^{M^A \rightarrow M^A} \diamond_M f ,$$

and then derive `map` from `flatMap` as before.

If a monad's code is defined via `pure` and `flatten`, we cannot derive the `map` method. So, we could indeed say heuristically that `flatten` has the functionality of `flatMap` without `map`.

In the previous chapter, we have seen that the laws of filterable functors are straightforwardly adapted to filterable contrafunctors. Could we do that with monads to obtain “contramonomads”? It turns out that nontrivial “contramonomads” do not exist.

The first problem is defining a method analogous to `flatMap` for contrafunctors. We turn to category theory (Section 9.4.3) for guidance. The type signature of a monad's `flatMap` is that of a lifting,

$$\text{flm}_M : (A \rightarrow M^B) \rightarrow M^A \rightarrow M^B ,$$

corresponding to a functor from the M -Kleisli category to the M -lifted category. To adapt this construction to a contrafunctor H^A , we need to define a lifting called, say, “`contraflatMap`”, with the type

$$\text{cflm}_H : (A \rightarrow H^B) \rightarrow H^A \rightarrow H^B .$$

We notice a curious feature of this type signature: since H^B is contravariant in B , the entire type of cflm_H is contravariant in B . So, cflm_H does not have the type of a natural transformation with respect to B , unlike all other liftings we have seen, such as `map`, `lift0pt`, and `flatMap`. If the code of cflm_H is fully parametric then cflm_H must satisfy a corresponding naturality law. It turns out that the naturality law gives a constraint so strong that it forces H to be a *constant* contrafunctor.

To see that, consider the left naturality law of cflm_H (with respect to the type parameter A):

$$\text{cflm}_H(f : A \rightarrow B \circ g : B \rightarrow H^C) = \text{cflm}_H(g) \circ f^{\downarrow H} . \quad (10.21)$$

If H were a “contramonomad”, it would have a method analogous to `pure` with type signature $A \rightarrow H^A$ satisfying the right identity law

$$\text{cflm}_H(\text{pu}_H) = \text{id} .$$

Substitute $g = \text{pu}_H$ into Eq. (10.21) and obtain

$$\text{cflm}_H(f \circ \text{pu}_H) \stackrel{!}{=} \text{id} \circ f^{\downarrow H} = f^{\downarrow H} . \quad (10.22)$$

The method `pure` for a contrafunctor H is equivalent to a value wu_H of type $H^{\mathbb{1}}$ (Section 8.3.8). In other words, the naturality law for $\text{pu}_H : A \rightarrow H^A$ forces it to be a function that ignores its argument and always returns a fixed value of type H^A , denoted by cpu_H in Sections 8.3.7–8.3.8:

$$\text{pu}_H = \underline{\text{ }}^A \rightarrow \text{cpu}_H^A = \underline{\text{ }}^A \rightarrow \text{wu}_H \triangleright (\underline{\text{ }}^A \rightarrow 1)^{\downarrow H} .$$

Since pu_H ignores its argument, we have $f \circ \text{pu}_H = \text{pu}_H$ for any function f . Then Eq. (10.22) gives

$$f^{\downarrow H} \stackrel{!}{=} \text{cflm}_H(f \circ \text{pu}_H) = \text{cflm}_H(\text{pu}_H) = \text{id} .$$

The equation $f^{\downarrow H} = \text{id}$ holds only for constant contrafunctors $H^A \triangleq Z$ (with some fixed type Z). Constant contrafunctors are functors at the same time. We know from Section 10.2.8 that constant functors $F^A = Z$ are monads only when $Z = \mathbb{1}$. So, nontrivial “contramonomads” are not possible.

10.3.2 Equivalence of a natural transformation and a “lifting”

In this and the previous chapters, we have derived the laws for three typeclasses: filterable functors, filterable contrafunctors, and monads. In each case, the laws were formulated in terms of lifting-like functions (`liftOpt`, `flatMap`), with type signatures:

for a filterable functor F : $\text{liftOpt}_F : (A \rightarrow \text{Opt}^B) \rightarrow F^A \rightarrow F^B$,
 for a filterable contrafunctor F : $\text{liftOpt}_F : (A \rightarrow \text{Opt}^B) \rightarrow F^B \rightarrow F^A$,
 for a monad F : $\text{flm}_F : (A \rightarrow F^B) \rightarrow F^A \rightarrow F^B$.

The laws were equivalently written in terms of natural transformations (`deflate`, `inflate`, `flatten`):

for a filterable functor F : $\text{deflate}_F : F^{\text{Opt}^A} \rightarrow F^A$,
 for a filterable contrafunctor F : $\text{inflate}_F : F^A \rightarrow F^{\text{Opt}^A}$,
 for a monad F : $\text{ftn}_F : F^{F^A} \rightarrow F^A$.

How can the liftings be *equivalent* to their corresponding natural transformations, which have fewer type parameters and simpler types? The equations relating these functions have a clear pattern:

$$\begin{aligned} \text{deflate}_F &= \text{liftOpt}_F(\text{id}) , & \text{liftOpt}_F(f) &= f^{\uparrow F} ; \text{deflate} , \\ \text{inflate}_F &= \text{liftOpt}_F(\text{id}) , & \text{liftOpt}_F(f) &= \text{inflate} ; f^{\downarrow F} , \\ \text{ftn}_F &= \text{flm}_F(\text{id}) , & \text{flm}_F(f) &= f^{\uparrow F} ; \text{ftn}_F . \end{aligned}$$

The pattern is that the code of the lifting-like function contains both the functionality of `map` (lifting the function f to $f^{\uparrow F}$ or $f^{\downarrow F}$) and the functionality of the corresponding natural transformation. Setting f to an identity function will cancel the lifting (since $\text{id}^{\uparrow F} = \text{id}$ and $\text{id}^{\downarrow F} = \text{id}$), so that only the functionality of the natural transformation remains.

We have proved the equivalence of lifting-like functions and their corresponding natural transformations for filterable functors in Statement 9.2.3.1 and for monads in Statement 10.2.2.1. Is it just a coincidence that these functions are equivalent for filterable functors and for monads, or are these properties special cases of a general construction? To generalize the type signatures of `deflate` and `flatten`, consider arbitrary functors F and G , and a natural transformation

$$\text{tr} : F^{G^A} \rightarrow F^A .$$

A lifting-like function corresponding to this natural transformation is defined by

$$\text{ftr} : (A \rightarrow G^B) \rightarrow F^A \rightarrow F^B , \quad \text{ftr}(f^{A \rightarrow G^B}) \triangleq f^{\uparrow F} ; \text{tr} , \quad (10.23)$$

which reproduces the definitions for filterable functors with $G = \text{Opt}$ and for monads with $G = F$.

By setting $f = \text{id}$ in Eq. (10.23), we find that $\text{tr} = \text{ftr}(\text{id})$. When are tr and ftr equivalent?

Statement 10.3.2.1 The natural transformation tr is equivalent to the lifting-like function ftr via:

$$\begin{array}{ccc} (f^{A \rightarrow G^B})^{\uparrow F} & \nearrow F^{G^B} & \text{ftr}(f) = f^{\uparrow F} ; \text{tr} , \quad \text{tr} = \text{ftr}(\text{id}) , \\ F^A & \xrightarrow{\text{tr}} & F^B & \text{if } \text{ftr} \text{ obeys the naturality law with respect to the type parameter } A: \\ & \text{ftr}(f^{A \rightarrow G^B}) & & g^{\uparrow F} ; \text{ftr}(f) = \text{ftr}(g ; f) . \end{array} \quad (10.24)$$

Proof We need to derive the equivalence between tr and ftr in both directions:

(a) Given a function tr , we first define ftr via Eq. (10.23) and then define a new function tr' via Eq. (10.24). We then show that tr' equals tr :

$$\text{tr}' = \text{ftr}(\text{id}) = \text{id}^{\uparrow F} ; \text{tr} = \text{tr} .$$

For the function `ftr` defined via `tr`, the naturality law (10.25) holds automatically:

$$g^{\uparrow F} ; ftr(f) = \underline{g^{\uparrow F} ; f^{\uparrow F}} ; tr = (g ; f)^{\uparrow F} ; tr = ftr(g ; f) \quad .$$

(b) Given a function `ftr` that satisfies the law (10.25), we define `tr` via Eq. (10.24) and then define a new function `ftr'` via Eq. (10.23). We then show that `ftr'` equals `ftr`: for an arbitrary $f: A \rightarrow G^B$, write

$$\begin{aligned} \text{expect to equal } ftr(f) : & ftr'(f) = f^{\uparrow F} ; tr = f^{\uparrow F} ; ftr(id) \\ \text{naturality law (10.25)} : & = ftr(f ; id) = ftr(f) \quad . \end{aligned}$$

The full equivalence between `tr` and `ftr` would not hold without assuming the naturality law (10.25). To build intuition for that fact, consider that `ftr` has a more complicated type signature with one more type parameter than `tr`. So, in general there are “many more” possible functions with the type of `ftr`. The imposed naturality law (10.25) constrains the code of `ftr` so much that all possible functions `ftr` satisfying Eq. (10.25) are in a 1-to-1 correspondence with all possible functions `tr`.

A similar statement can be proved for contrafunctors F , by considering the type signatures

$$tr : F^A \rightarrow F^{G^A} \quad , \quad ftr : (A \rightarrow G^B) \rightarrow F^B \rightarrow F^A \quad .$$

Finally, a generalization is also possible with functors F and the type signatures

$$tr : F^A \rightarrow F^{G^A} \quad , \quad ftr : (G^A \rightarrow B) \rightarrow F^A \rightarrow F^B \quad .$$

By referring to these constructions, we can omit detailed proofs of equivalence in all such cases.

10.3.3 Monads, effects, and runners

The main advantage of using monads is the ability to compose “monadic programs” (i.e., values of type M^A for some monad M). For list-like and tree-like monads M , a value of type M^A represents a collection of values of type A and is the desired result of nested iterations. The same applies to pass/fail monads: e.g., a value of type `Try[A]` holds either a result or information about a failure, which we also need to use. However, for most other monads M , a monadic program $m : M^A$ represents a single value of type A held by an “effectful wrapper” of some sort. To get a useful result, we usually need to extract that value of type A from the wrapper.

To see how this works in general, consider monads of function types (such as `Reader`, `State`, and `Cont`). These monads delay their computations until a runner is applied. A monadic program $m : M^A$ is a function whose body somehow wraps a value of type A , and a runner gets access to that value. The function-type monads are often defined as case classes with a single part called `run`, for example:

```
final case class Reader[Z, A](run: Z => A)
final case class State[S, A](run: S => (A, S))
final case class Cont[R, A](run: (A => R) => R)
```

In Scala syntax, this makes monadic programs appear to have a method called `run`:

```
val s: State[S, A] = for { ... } yield { ... } // A monadic program in the State monad.
val init: S = ??? // An initial state.
val result: A = s.run(init)._1 // Run the monadic program and extract the result.
```

We can package this code into a `runner` function:

```
def runner[S, A](init: S): State[S, A] => A = _.run(init)._1
```

The runner for the `State` monad performs all the state updates contained in the monadic program `s: State[S, A]`, in order to compute a final result value of type A . One could say that the runner “runs the effects” in order to extract the result value.

These examples motivate a general definition: A **runner** for a monad M is a function of type $M^A \rightarrow A$, which we will denote by $\theta^A : M^A \rightarrow A$.

To be useful in practice, a runner θ must satisfy certain laws. To motivate those laws, consider how we would use runners with a monadic program $m : M^B$ that is composed from two parts: the first part is a monadic program $m_1 : M^A$, and the second part is a function $m_2 : A \rightarrow M^B$ that depends on the result (of type A) of the first monadic program.

```
val m = for { // m == m1.flatMap(m2)
  x <- m1
  y <- m2(x)
} yield y
```

The composition of m_1 and m_2 can be written as a functor block or as an application of the `flatMap` method,

$$m = m_1 \triangleright \text{flm}_M(m_2) \quad . \quad (10.26)$$

We may imagine that $\theta(m)$ first runs the effects of m_1 obtaining a value x , and then runs the effects of $m_2(x)$ obtaining a value y . So, it is natural to require that a runner θ applied to m should give the same results as applying θ to m_1 , which extracts a value x , and then applying the runner to $m_2(x)$. We can formulate this requirement as a law called the runner's **composition law**,

```
runner(m) == runner(m2(runner(m1)))
```

$$m \triangleright \theta = m_1 \triangleright \theta \triangleright m_2 \triangleright \theta \quad .$$

Substituting m from Eq. (10.26), we get $m_1 \triangleright \text{flm}_M(m_2) \circ \theta = m_1 \triangleright \theta \circ m_2 \circ \theta$. Since this equation must hold for any m_1 , we can simplify it to

$$\text{flm}_M(m_2) \circ \theta = \theta \circ m_2 \circ \theta \quad .$$

Let us reformulate the composition law in terms of the monad M 's `flatten` method:

$$m_2^{\uparrow M} \circ \text{ftn}_M \circ \theta = \theta \circ m_2 \circ \theta \quad .$$

We would like to move m_2 to the left of θ in the right-hand side, so that m_2 could drop out of the equation. To interchange the order of composition $\theta \circ m_2$, we assume that a naturality law for θ ,

$$f^{\uparrow M} \circ \theta = \theta \circ f \quad . \quad (10.27)$$

It follows that $\theta \circ m_2 \circ \theta = m_2^{\uparrow M} \circ \theta \circ \theta$ and that $\theta \circ \theta = \theta^{\uparrow M} \circ \theta$. The runner's composition law becomes:

$$\begin{array}{ccc} M^{M^A} & \xrightarrow[\theta]{\theta^{\uparrow M}} & M^A \\ \text{ftn}_M \downarrow & \theta \downarrow & \\ M^A & \xrightarrow[\theta]{} & A \end{array} \quad \text{ftn}_M \circ \theta = \theta \circ \theta = \theta^{\uparrow M} \circ \theta \quad . \quad (10.28)$$

Another clearly motivated requirement for runners is that `pure(x)`, a monadic value with an "empty effect", should be correctly handled by the runner:

$$x \triangleright \text{pu}_M \triangleright \theta = x \quad , \quad \text{or equivalently: } \text{pu}_M \circ \theta = \text{id} \quad .$$

This is the **identity law** of monad runners.

If a runner has purely functional code (such as the runner for the `State` monad shown above), the naturality law (10.27) will hold automatically due to the parametricity theorem. However, monad runners are not always purely functional. For example, the runner for the continuation monad shown in Section 10.1.9 uses Scala's `Future` and `Promise` classes, which are not purely functional. (Nevertheless, the identity and composition laws hold for that runner.)

We have argued that list-like, tree-like, and pass/fail monads do not need runners; but in fact, those monads *cannot* have lawful runners. For instance, the `Option` monad cannot have a purely functional runner: a function $\theta^A : \mathbb{1} + A \rightarrow A$ must produce a value of type A even if the `Option` value is empty, but it is impossible to produce a value of an arbitrary type A from scratch using purely functional code. For the `Option` monad and other pass/fail monads, we could only use runners θ^A that work for a specific type A , for example:

```
def runner: Option[Int] => Int = _.getOrElse(0) // For empty Option values, return a default.
```

Even if we restrict all types to `Int`, this runner will fail to obey the composition law:

```

val m1: Option[Int] = None
val m2: Int => Option[Int] = { x => Some(x + 1) }
val m: Option[Int] = for { x <- m1; y <- m2(x) } yield y

scala> runner(m)    // Composition law: runner(m) == runner(m2(runner(m1))).
res0: Int = 0

scala> runner(m2(runner(m1)))
res1: Int = 1

```

Similar arguments apply to list-like and tree-like monads. These monads may have an empty value, which cannot be correctly handled by a runner θ^A whose result must be of parametric type A .

A solution is to generalize the notion of runner from a function of type $M^A \rightarrow A$ to a function of type $M^A \rightarrow N^A$ where N is another monad (the “target” monad of the runner). For example, if M is a pass/fail monad, we can choose the target monad as $N^A = E + A$ where a fixed type E represents error information. We can then define a generalized runner for the `Option` monad like this:

```

val error: E = ... // Describe the error.
def run[A]: Option[A] => Either[E, A] = {
  case None    => Left(error)
  case Some(a) => Right(a)
}

```

$$\theta: \mathbb{1}^{A \rightarrow E + A} \triangleq \begin{array}{c|c|c|c} & & E & A \\ \hline 1 & \rightarrow \text{error} & 0 & \\ \hline A & 0 & & \text{id} \end{array} .$$

In the next section, we will formulate the laws for such maps between monads.

10.3.4 Monads in category theory. Monad morphisms

For any monad M , one defines a category, called the M -Kleisli category, whose objects are all types (`Int`, `String`, etc.) and whose morphisms between types A and B are Kleisli functions of type $A \rightarrow M^B$.

One axiom of a category requires us to have an identity morphism $A \rightarrow M^A$ for every object A ; this is the monad M ’s `pure` method, $\text{pu}_M : A \rightarrow M^A$. Another axiom is the associativity of morphism composition operation, which must combine functions of types $A \rightarrow M^B$ and $B \rightarrow M^C$ into a function of type $A \rightarrow M^C$. The Kleisli composition \diamond_M is precisely that operation, and its associativity law, $f \diamond_M (g \diamond_M h) = (f \diamond_M g) \diamond_M h$, is equivalent to a law of `flatMap` (Statements 10.2.6.3–10.2.6.5).

So, a functor M is a monad if and only if the corresponding M -Kleisli category is lawful. This is an economical way of formulating the monad laws; the only information specific to monads is the type signature $A \rightarrow M^B$ for the M -Kleisli morphisms.

We have seen that, for some monads, proofs of the laws are easier when written in terms of Kleisli morphisms. It turns out that the properties of monad runners also have a concise formulation in the language of categories.

A monad M ’s `pure` method has type $A \rightarrow M^A$, while a runner θ_M has type $M^A \rightarrow A$. Since the type parameter A itself can be viewed as the identity monad $\text{Id}^A \triangleq A$, we can write the types as

$$\text{pu}_M : \text{Id}^A \rightarrow M^A , \quad \theta_M : M^A \rightarrow \text{Id}^A .$$

These two types can be generalized to a transformation between two monads M and N ,

$$\phi : M^A \rightarrow N^A .$$

As we have seen in the previous section, some monads M require runners of this more general type, rather than of type $M^A \rightarrow A$.

Another use case for the type $M^A \rightarrow N^A$ comes from our code for the continuation monad’s runner. The code first transforms a value of type `Cont[R, A]` into a `Future[A]` and then waits for the `Future` to complete. So, `Cont`’s runner can be seen as a composition of two transformations:

$$\text{Cont}^{R,A} \xrightarrow{\theta_{\text{Cont-Future}}} \text{Future}^A \xrightarrow{\theta_{\text{Future}}} A \quad \theta_{\text{Cont}} = \theta_{\text{Cont-Future}} \circ \theta_{\text{Future}} .$$

The intermediate “generalized runner” $\theta_{\text{Cont-Future}}$ converts continuation-based monadic programs into Future-based ones. For this conversion to be compatible with the way we write and refactor functor blocks, we need to require laws similar to the laws of runners shown in the previous section. The “generalized runners” are called monad morphisms. The following definition states their laws:

Definition 10.3.4.1 A natural transformation $\phi^A : M^A \rightarrow N^A$, also denoted by $\phi : M^\bullet \rightsquigarrow N^\bullet$ or $\phi : M \rightsquigarrow N$, is called a **monad morphism** between monads M and N if the following two laws hold:

$$\text{identity law for } \phi : \text{pu}_M \circ \phi = \text{pu}_N \quad , \quad (10.29)$$

$$\text{composition law for } \phi : \text{ftn}_M \circ \phi = \phi^M \circ \phi^{N^A} \circ \text{ftn}_N \quad . \quad (10.30)$$

$$\begin{array}{ccc} & \begin{array}{c} \text{pu}_N \\ \swarrow \\ \text{pu}_M \\ \downarrow \\ \text{M}^A \end{array} & \begin{array}{c} \text{M}^{M^A} \xrightarrow{\text{ftn}_M} \text{M}^A \\ \phi^M \downarrow \\ \text{M}^{N^A} \xrightarrow{\phi^{N^A}} \text{N}^{N^A} \xrightarrow{\text{ftn}_N} \text{N}^A \\ \phi \searrow \end{array} \end{array}$$

The composition law can be equivalently expressed using the `flatMap` method:

$$\text{flm}_M(f : A \rightarrow M^B) \circ \phi : M^B \rightarrow N^B = \phi \circ \text{flm}_N(f \circ \phi) \quad . \quad (10.31)$$

In terms of the Kleisli composition operations \diamond_M and \diamond_N , the composition law is

$$(f : A \rightarrow M^B \circ \phi : M^B \rightarrow N^B) \diamond_N (g : B \rightarrow M^C \circ \phi : M^C \rightarrow N^C) = (f \diamond_M g) \circ \phi \quad .$$

This formulation shows more visually that a monad morphism $\phi^A : M^A \rightarrow N^A$ replaces M -effects by N -effects while preserving the composition of effects.

The name “monad morphism” is motivated by considering the *category of monads*. The objects of that category are the different possible monads (type constructors such as `Option`, `Try`, etc.). The morphisms between objects M and N of that category are monad morphisms $M \rightsquigarrow N$ as defined above: natural transformations that preserve the structure of the monadic operations. At the same time, a monad morphism ϕ can be viewed as a correspondence $g = f \circ \phi$ between Kleisli functions $f : A \rightarrow M^B$ and $g : A \rightarrow N^B$. The laws of monad morphisms guarantee that the mapping ϕ is compatible with the Kleisli composition operations \diamond_M and \diamond_N . So, it can be seen also as expressing a (categorical) functor between the M -Kleisli and the N -Kleisli categories.

Monad morphisms between arbitrary monads will be also used in Chapters 13 and 14. To build up more intuition, let us look at some examples of monad morphisms.

Example 10.3.4.2 We would like to define a function $\phi^A : Z + A \rightarrow \mathbb{1} + A$ as a monad morphism between the `Either` and `Option` monads. The implementation of ϕ is:

```
def toOption[Z, A]: Either[Z, A] => Option[A] = {
  case Left(z) => None
  case Right(a) => Some(a)
}
```

$$\phi \triangleq \begin{array}{c|cc} & \mathbb{1} & A \\ \hline Z & _ & \rightarrow 1 & 0 \\ A & 0 & id \end{array} .$$

To verify the identity law (10.29):

$$\text{pu}_{\text{Either}} \circ \phi = \begin{array}{c|cc} & Z & A \\ \hline A & 0 & id \end{array} \circ \begin{array}{c|cc} & \mathbb{1} & A \\ \hline Z & _ & \rightarrow 1 & 0 \\ A & 0 & id \end{array} = \begin{array}{c|cc} & \mathbb{1} & A \\ \hline A & 0 & id \end{array} = \text{pu}_{\text{Opt}} \quad .$$

To verify the composition law (10.30), show that both sides are equal:

$$\text{left-hand side} : \text{ftn}_{\text{Either}} \circ \phi = \begin{array}{c|cc} & Z & A \\ \hline Z & id & 0 \\ Z & id & 0 \\ A & 0 & id \end{array} \circ \begin{array}{c|cc} & \mathbb{1} & A \\ \hline Z & _ & \rightarrow 1 & 0 \\ A & 0 & id \end{array} = \begin{array}{c|cc} & \mathbb{1} & A \\ \hline Z & _ & \rightarrow 1 & 0 \\ Z & _ & \rightarrow 1 & 0 \\ A & 0 & id \end{array} ,$$

$$\begin{aligned}
\text{right-hand side : } \phi^{\uparrow \text{Either}} \circ \phi \circ \text{ftn}_{\text{Opt}} &= \begin{array}{|c|c|c|c|} \hline & Z & 1 & A \\ \hline Z & \text{id} & 0 & 0 \\ \hline Z & 0 & _ \rightarrow 1 & 0 \\ \hline A & 0 & 0 & \text{id} \\ \hline \end{array} \circ \begin{array}{|c|c|c|c|} \hline & 1 & 1 & A \\ \hline 1 & _ \rightarrow 1 & 0 & 0 \\ \hline 1 & 0 & \text{id} & 0 \\ \hline A & 0 & 0 & \text{id} \\ \hline \end{array} \circ \begin{array}{|c|c|c|c|} \hline & 1 & A \\ \hline 1 & \text{id} & 0 \\ \hline 1 & \text{id} & 0 \\ \hline A & 0 & \text{id} \\ \hline \end{array} \\
&= \begin{array}{|c|c|c|c|} \hline & 1 & 1 & A \\ \hline Z & _ \rightarrow 1 & 0 & 0 \\ \hline Z & 0 & _ \rightarrow 1 & 0 \\ \hline A & 0 & 0 & \text{id} \\ \hline \end{array} \circ \begin{array}{|c|c|c|c|} \hline & 1 & A \\ \hline 1 & \text{id} & 0 \\ \hline 1 & \text{id} & 0 \\ \hline A & 0 & \text{id} \\ \hline \end{array} = \begin{array}{|c|c|c|c|} \hline & 1 & A \\ \hline Z & _ \rightarrow 1 & 0 \\ \hline Z & _ \rightarrow 1 & 0 \\ \hline A & 0 & \text{id} \\ \hline \end{array} .
\end{aligned}$$

This monad morphism ϕ maps the `Either`-effect (error information) to the `Option`-effect (absence of value) by discarding the error information. The laws of the monad morphism guarantee that composition of `Either`-effects is mapped into composition of `Option`-effects. For instance, if some computations caused an error encapsulated by `Either`, the corresponding value after ϕ will be an empty `Option` (i.e., `None`).

The function ϕ is so simple that the correspondence of effects appears to be automatic. The next example shows a monad morphism that translates effects in a nontrivial way.

Example 10.3.4.3 A monad morphism between the `Writer` and the `Cont` monads can be defined by

```
def toCont[W: Monoid, A]: ((A, W)) => (A => W) => W = { case (a, w) => k => k(a) ⊕ w }
```

$$\phi : A \times W \rightarrow (A \rightarrow W) \rightarrow W \quad , \quad \phi \triangleq a:A \times w:W \rightarrow k:A \rightarrow W \rightarrow k(a) \oplus w \quad ,$$

where the fixed type W is a monoid with a binary operation \oplus and empty value e . The continuation monad's callback returns a value of type W that depends on the output w from the `Writer` monad.

To verify the identity law, use the definitions of `pure` for the `Writer` and `Cont` monads:

$$\begin{aligned}
\text{pure}_{\text{Writer}} \circ \phi &= (a \rightarrow a \times e) \circ (a \times w \rightarrow k \rightarrow k(a) \oplus w) \\
\text{compute composition : } &= a \rightarrow k \rightarrow k(a) \oplus e = a \rightarrow k \rightarrow k(a) = \text{pure}_{\text{Cont}} \quad .
\end{aligned}$$

To verify the composition law, we need an implementation of `flatten` for the `Cont` monad:

```
def flatten[W, A]: (((A => W) => W) => W) => (A => W) => W = { g => k => g(c => c(k)) }
```

$$\text{ftn}_{\text{Cont}} = g:((A \rightarrow W) \rightarrow W) \rightarrow W \rightarrow k:A \rightarrow W \rightarrow g(c:(A \rightarrow W) \rightarrow W \rightarrow c(k)) \quad .$$

Now we show that the two sides of the law (10.30) reduce to the same function:

$$\begin{aligned}
\text{left-hand side : } \text{ftn}_{\text{Writer}} \circ \phi &= (a \times u \times w \rightarrow a \times (u \oplus w)) \circ (a \times w \rightarrow k \rightarrow k(a) \oplus w) \\
&= a \times u \times w \rightarrow k \rightarrow k(a) \oplus u \oplus w \quad , \\
\text{right-hand side : } \phi^{\uparrow \text{Writer}} \circ \phi^{\text{Cont}^{W,A}} \circ \text{ftn}_{\text{Cont}} &= (a \times u \times w \rightarrow (k \rightarrow k(a) \oplus u) \times w) \circ (c \times w \rightarrow h \rightarrow h(c) \oplus w) \circ \text{ftn}_{\text{Cont}} \\
&= (a \times u \times w \rightarrow h \rightarrow h(k \rightarrow k(a) \oplus u) \oplus w) \circ \text{ftn}_{\text{Cont}} \\
\text{compute composition : } &= (a \times u \times w \rightarrow h \rightarrow h(k \rightarrow k(a) \oplus u) \oplus w) \circ (g \rightarrow p \rightarrow g(c \rightarrow c(p))) \\
\text{compute composition : } &= a \times u \times w \rightarrow p \rightarrow (c \rightarrow c(p))(k \rightarrow k(a) \oplus u) \oplus w \\
&= a \times u \times w \rightarrow p \rightarrow p(a) \oplus u \oplus w \quad .
\end{aligned}$$

We conclude this section by proving some properties of monad morphisms.

Statement 10.3.4.4 If a function $\phi^A : M^A \rightarrow N^A$ satisfies Eqs. (10.29) and (10.31) then:

(a) ϕ is a natural transformation, and (b) ϕ also satisfies Eq. (10.30).

Proof (a) The naturality law of ϕ ,

$$(g^{:A \rightarrow B})^{\uparrow M} ; \phi = \phi ; g^{\uparrow N} , \quad (10.32)$$

is derived from Eq. (10.31) by setting $f^{:A \rightarrow M^B} \triangleq g^{:A \rightarrow B} ; \text{pu}_M^{:B \rightarrow M^B}$:

$$\text{left-hand side of Eq. (10.31)} : \text{flm}_M(f) ; \phi = \underline{\text{flm}_M(f ; \text{pu}_M)} ; \phi$$

$$\text{left naturality of } \text{flm}_M : = f^{\uparrow M} ; \underline{\text{flm}_M(\text{pu}_M)} ; \phi$$

$$\text{right identity law of } \text{flm}_M : = f^{\uparrow M} ; \phi ,$$

$$\text{right-hand side of Eq. (10.31)} : \phi ; \text{flm}_N(f ; \phi) = \phi ; \underline{\text{flm}_N(g ; \text{pu}_M)} ; \phi$$

$$\text{identity law (10.29)} : = \phi ; \text{flm}_N(g ; \text{pu}_N)$$

$$\text{left naturality of } \text{flm}_N : = \phi ; g^{\uparrow N} ; \underline{\text{flm}_N(\text{pu}_N)}$$

$$\text{right identity law of } \text{flm}_N : = \phi ; g^{\uparrow N} .$$

The two sides of Eq. (10.31) are equal to the two sides of Eq. (10.32).

(b) Substitute the definitions $\text{flm}_M(f) = f^{\uparrow M} ; \text{ftn}_M$ and $\text{flm}_N(f) = f^{\uparrow N} ; \text{ftn}_N$ into Eq. (10.31):

$$f^{\uparrow M} ; \text{ftn}_M ; \phi = \underline{\phi ; (f ; \phi)^{\uparrow N}} ; \text{ftn}_N .$$

$$\text{use Eq. (10.32)} : = (f ; \phi)^{\uparrow M} ; \phi ; \text{ftn}_N = f^{\uparrow M} ; \phi^{\uparrow M} ; \phi ; \text{ftn}_M .$$

This equality holds for any f , in particular with $f = \text{id}$, which directly gives Eq. (10.30).

If a function $\phi : M \rightsquigarrow N$ satisfies Eq. (10.31) then ϕ is a natural transformation.

Statement 10.3.4.5 If L, M, N are monads and $\phi : L \rightsquigarrow M$ and $\chi : M \rightsquigarrow N$ are monad morphisms then the composition $\phi ; \chi : L \rightsquigarrow N$ is also a monad morphism.

Proof The identity law for $\phi ; \chi$ is verified by:

$$\text{expect to equal } \text{pu}_N : \underline{\text{pu}_L ; (\phi ; \chi)}$$

$$\text{identity law for } \phi : = \text{pu}_M ; \chi$$

$$\text{identity law for } \chi : = \text{pu}_N .$$

The composition law for $\phi ; \chi$ is verified by:

$$\text{expect to equal } \text{ftn}_L ; \phi ; \chi : (\phi ; \chi)^{\uparrow L} ; (\phi ; \chi) ; \text{ftn}_N$$

$$\text{naturality of } \phi : = \phi^{\uparrow L} ; \phi ; \underline{\chi^{\uparrow M} ; \chi ; \text{ftn}_N}$$

$$\text{composition law for } \chi : = \underline{\phi^{\uparrow L} ; \phi ; \text{ftn}_M} ; \chi$$

$$\text{composition law for } \phi : = \text{ftn}_L ; \phi ; \chi .$$

Statement 10.3.4.6 For any monad M , the method $\text{pu}_M : A \rightarrow M^A$ is a monad morphism $\text{pu}_M : \text{Id} \rightsquigarrow M$ between the identity monad and M .

Proof The identity law requires $\text{pu}_{\text{Id}} ; \text{pu}_M = \text{pu}_M$. This holds because $\text{pu}_{\text{Id}} = \text{id}$. The composition law requires $\text{ftn}_{\text{Id}} ; \text{pu}_M = \text{pu}_M^{\uparrow \text{Id}} ; \text{pu}_M ; \text{ftn}_M$. Since $\text{ftn}_{\text{Id}} = \text{id}$, the left-hand side of the composition law simplifies to pu_M . Transform the right-hand side:

$$\text{expect to equal } \text{pu}_M : \underline{\text{pu}_M^{\uparrow \text{Id}} ; \text{pu}_M ; \text{ftn}_M}$$

$$\text{lifting to the identity functor} : = \text{pu}_M ; \underline{\text{pu}_M ; \text{ftn}_M}$$

$$\text{left identity law for } M : = \text{pu}_M .$$

Exercise 10.3.4.7 Suppose M is a given monad, Z is a fixed type, and a *fixed* value $m_0 : M^Z$ is given.

(a) Consider the function f defined as

$$f : (Z \rightarrow A) \rightarrow M^A \quad , \quad f(q^{Z \rightarrow A}) \triangleq m_0 \triangleright q^{\uparrow M} \quad .$$

Prove that f is *not* a monad morphism from the Reader monad $R^A \triangleq Z \rightarrow A$ to the monad M^A , despite having the correct type signature.

(b) Under the same assumptions, consider the function ϕ defined as

$$\phi : (Z \rightarrow M^A) \rightarrow M^A \quad , \quad \phi(q^{Z \rightarrow M^A}) \triangleq m_0 \triangleright \text{flm}_M(q) \quad .$$

Show that ϕ is *not* a monad morphism from the monad $Q^A \triangleq Z \rightarrow M^A$ to M^A .

10.3.5 Constructions of polynomial monads

“Polynomial monads” are polynomial functors that have lawful monad methods. The product and co-product constructions allow us to create polynomial monads via the following operations:

- Start with $F^A \triangleq Z + W \times A$, which is a monad (semimonad) when W is a monoid (semigroup).
- Given a polynomial monad F^A , create the monad $L^A \triangleq A + F^A$.
- Given two polynomial monads F^A and G^A , create the monad $L^A \triangleq F^A \times G^A$.
- Given a polynomial monad F^A , create the monad $L^A \triangleq F^{Z+W \times A}$ (see Section 14.4).

It is an open conjecture (no proof is known) that these are the only constructions available for polynomial monads. If so, we can create an algorithm that recognizes whether a given polynomial functor can be made into a monad by a suitable definition of `flatten` and `pure`.

As an example, consider the fact that the polynomial functor $F^A \triangleq 1 + A \times A$ cannot be made into a monad (Exercise 10.2.9.7). One can also show that F^A cannot be obtained through the monad constructions listed above. Indeed, the corresponding polynomial $f(x) = 1 + x^2$ does not contain any first powers of x . However, the constructions either start with a polynomial containing x , or add x , or take a product of two such polynomials. None of these operations could cancel first powers of x since all coefficients are types and cannot be subtracted.

By the same logic, we can conclude that $F^A \triangleq 1 + A \times A \times A$ cannot be obtained through monad constructions. It is likely that $F^A \triangleq 1 + A \times A \times A$, $F^A \triangleq 1 + A \times A \times A \times A$, and all other similarly constructed functors are not monads.

10.3.6 Constructions of M -filterable functors and contrafunctors

In Chapter 9, we have used the Opt-Kleisli category (that is, the M -Kleisli category with M set to the `option` monad) to formulate the laws of filterable (contra)functors. We found that the laws of a filterable (contra)functor F are equivalent to the requirement that the function `liftOpt` be a lawful (categorical) functor from the Opt-Kleisli category to an F -lifted category:

$$\begin{aligned} \text{filterable functor } F : \quad & \text{liftOpt}_F : (A \rightarrow \text{Opt}^B) \rightarrow F^A \rightarrow F^B \quad , \\ \text{filterable contrafunctor } F : \quad & \text{liftOpt}_F : (A \rightarrow \text{Opt}^B) \rightarrow F^B \rightarrow F^A \quad . \end{aligned}$$

It is natural to generalize this formulation from the `option` monad to an arbitrary monad M ,

$$\begin{aligned} M\text{-filterable functor } F : \quad & \text{lift}_{M,F} : (A \rightarrow M^B) \rightarrow F^A \rightarrow F^B \quad , \\ M\text{-filterable contrafunctor } F : \quad & \text{lift}_{M,F} : (A \rightarrow M^B) \rightarrow F^B \rightarrow F^A \quad . \end{aligned}$$

This gave us the definitions of M -filterable functors and contrafunctors.

We have shown some simple examples of M -filterable functors in Statement 10.2.8.7. Structural analysis can discover other examples of such functors systematically. As in the case of ordinary filterable functors, it turns out that we must at the same time analyze M -filterable contrafunctors.

In the following constructions, we always assume that M is a fixed, lawful monad.

We omit the proofs of all following statements because they are fully analogous to the proofs of filterable functor and contrafunctor constructions in Sections 9.2.4 and 9.2.6. In those proofs, we only used the properties of the functions liftOpt_F , which are fully analogous to the properties of the function $\text{lift}_{M,F}$ except for replacing the `Option` monad by the monad M and the operation \diamond_{Opt} by \diamond_M .

Type parameters A constant functor $F^A \triangleq Z$ is M -filterable with $\text{lift}_{M,F}(f) = \text{id}$ (Statement 10.2.8.7). The same statement shows that $F^A \triangleq M^A \rightarrow Z$ and $F^A \triangleq A \rightarrow M^Z$ are M -filterable contrafunctors.

The monad M^A itself is M -filterable; $\text{lift}_{M,M}(f) \triangleq \text{flm}_M(f)$.

The identity functor is not M -filterable except when M is the identity monad, $M^A = \text{Id}^A \triangleq A$. (However, with $M = \text{Id}$, the concept of M -filterable functor becomes trivial because all functors and all contrafunctors are Id -filterable. So, we will assume that $M \neq \text{Id}$.)

The (contra)functor $F^A \triangleq G^{H^A}$ is M -filterable when H is M -filterable and G is any (contra)functor (Statements 9.2.4.1 and 9.2.6.1), where G and H can be functors or contrafunctors independently.

Products If F^A and G^A are M -filterable then $L^A \triangleq F^A \times G^A$ is M -filterable (Statement 9.2.4.2).

Co-products If F^A and G^A are M -filterable then $L^A \triangleq F^A + G^A$ is M -filterable (Statement 9.2.4.3).

Function types If F^A is an M -filterable functor and G^A is an M -filterable contrafunctor then $F^A \rightarrow G^A$ and $G^A \rightarrow F^A$ are M -filterable (contra)functors (Statement 9.2.4.5 and 9.2.6.2).

The filterable contrafunctor $M^A \rightarrow Z$ from Statement 10.2.8.7 is obtained from this construction if we set $F^A = M^A$ and $G^A = Z$; both F and G are filterable as we have already seen.

Recursive types If $S^{A,R}$ is a bifunctor that is M -filterable with respect to A , the recursive functor F^A defined by the type equation $F^A \triangleq S^{A,F^A}$ is M -filterable (Statement 9.2.4.7).

If $S^{A,R}$ is a profunctor contravariant in A and covariant in R , and additionally $S^{\bullet,R}$ is M -filterable (with the type parameter R fixed), then the recursive contrafunctor $F^A \triangleq S^{A,F^A}$ is M -filterable (Statement 9.2.6.5).

In summary, any recursively exponential-polynomial type expression F^A will be an M -filterable (contra)functor if it depends on a type parameter A only through type expressions M^A or $A \rightarrow M^Z$ (where Z is a fixed type). While there may be other M -filterable (contra)functors, the structural analysis covers a broad class of type expressions of the form

$$S^{M^A, A \rightarrow M^{Z_1}, \dots, A \rightarrow M^{Z_n}},$$

where S^{A, B_1, \dots, B_n} is a type constructor covariant in A and contravariant in B_1, \dots, B_n (or vice versa), and Z_1, \dots, Z_n are fixed types.

11 Applicative functors, contrafunctors, and profunctors

11.1 Slides, Part I

Motivation for applicative functors Monads are inconvenient for expressing *independent* effects
Monads perform effects *sequentially* even if effects are independent:

```
x ← Future { c1 }
y ← Future { c2 }
z ← Future { c3 }

Future { c1 }.flatMap { x →
  Future { c2 }.flatMap { y →
    Future { c3 }.map { z → ... }
  }
}
```

We would like to parallelize independent computations

We would like to accumulate *all* errors, rather than stop at the first one

Changing the order of monad's effects will (generally) change the result:

```
for {
  x ← List(1, 2)
  y ← List(10, 20)
} yield f(x, y)
// f(1, 10), f(1, 20), f(2, 10), f(2, 20)

for {
  y ← List(10, 20)
  x ← List(1, 2)
} yield f(x, y)
// f(1, 10), f(2, 10), f(1, 20), f(2, 20)
```

We would like to express a computation where effects are unordered

This can be done using a method `map2`, *not* defined via `flatMap`: the desired type signature is `map2 : FA × FB → (A × B → C) → FC`

Applicative functor has `map2` and `pure` but is not necessarily a monad

Defining `map2`, `map3`, etc. Consider 1, 2, 3, ... commutative and independent “effects”

```
for { x1 ← c1
} yield f(x1)                                c1.map(f)



---


for { x1 ← c1
  x2 ← c2
} yield f(x1, x2)                            (c1, c2).map2(f)



---


for { x ← c1
  x2 ← c2
  x3 ← c3
} yield f(x1, x2, x3)                      (c1, c2, c3).map3(f)
```

Generalize from `map`, `map2`, `map3` to `mapN`:

$$\begin{aligned} \text{map}_1 &: F^A \rightarrow (A \rightarrow Z) \rightarrow F^Z \\ \text{map}_2 &: F^A \times F^B \rightarrow (A \times B \rightarrow Z) \rightarrow F^Z \\ \text{map}_3 &: F^A \times F^B \times F^C \rightarrow (A \times B \times C \rightarrow Z) \rightarrow F^Z \end{aligned}$$

Practical examples of using `mapN` $F^A \equiv Z + A$ where Z is a monoid: collect all errors

$F^A = Z + A$: Create a validated case class out of validated parts

$F^A \equiv \text{Future}[A]$: perform several computations concurrently

$F^A \equiv E \rightarrow A$: pass standard arguments to functions more easily

$F^A \equiv \text{List}^A$: transposing a matrix by using `map2`

Applicative contrafunctors and applicative profunctors

defining an instance of `Semigroup` type class from `Semigroup` parts
 implement `imap2` for non-disjunctive profunctors, e.g. $Z \times A \rightarrow A \times A$

“Fused `fold`”: automatically merge several `folds` into one
 compute several running averages in one traversal (*scala-folds*)

The difference between applicative and monadic functors
 define monadic folds using the “free functor” construction
 compute running averages that depend on previous running averages
 do not confuse this with *monad-valued* folds (*origami*)!

applicative parsers vs. monadic parsers
 applicative: parse independent data, collecting all errors

monadic: parse depends on previous results, stops on errors

monads, applicatives, arrows paper: (Lindley, Wadler, Yallop)

mumbo-jumbo, clap-trap, abra-cadabra, algebra-cadalgaebra, gobbley-gook

11.1.1 Exercises I

Implement `map2`, or `imap2` if appropriate, for these type constructors F^A :

$$F^A \equiv 1 + A + A \times A$$

$$F^A \equiv E \rightarrow A \times A$$

$$F^A \equiv Z \times A \rightarrow A$$

$$F^A \equiv A \rightarrow A \times Z \text{ where } Z \text{ is a } \text{Monoid}$$

Write a function that defines an instance of the `Monoid` type class for a tuple (A, B) whose parts already have a `Monoid` instance.

Define a `Monoid` instance for the type F^S where F is an applicative functor that has `map2` and `pure`, while S is itself a monoid type.

Define a “regexp extractor” as a type constructor R^A describing extraction of various data from strings; the extracted data is converted to a value of type `Option[A]`. Implement `zip` and `map2` for R^A .

Use parser combinators to implement an evaluator for arithmetic language containing integers and `+` symbols, for example $1 + 321 + 20$.

Use folding combinators to implement a `Fold` that computes the standard deviation of a sequence in one traversal.

11.2 Slides, Part II

Deriving the `ap` operation from `map2` Can we avoid having to define map_n separately for each n ?

- Use curried arguments, $\text{fmap}_2 : (A \rightarrow B \rightarrow Z) \rightarrow F^A \rightarrow F^B \rightarrow F^Z$
 - Set $A \equiv (B \rightarrow Z)$ and apply fmap_2 to the identity $\text{id}^{(B \rightarrow Z) \rightarrow (B \rightarrow Z)}$: obtain $\text{ap}^{[B, Z]} : F^{B \rightarrow Z} \rightarrow F^B \rightarrow F^Z \equiv \text{fmap}_2(\text{id})$
 - The functions `fmap2` and `ap` are computationally equivalent:

$$\begin{array}{ccc} \text{fmap}_2 f^{A \rightarrow B \rightarrow Z} & = & \text{fmap } f \circ \text{ap} \\ \text{fmap } f \nearrow \quad \quad \quad \text{ap} \searrow & & \\ F^A & \xrightarrow{\quad \quad \quad \quad \quad \quad \quad \quad} & (F^B \rightarrow F^Z) \\ \text{fmap}_2(f^{A \rightarrow B \rightarrow Z}) & & \end{array}$$

- The functions `fmap3`, `fmap4` etc. can be defined similarly:

$$\begin{array}{ccc} \text{fmap}_3 f^{A \rightarrow B \rightarrow C \rightarrow Z} & = & \text{fmap } f \circ \text{ap} \circ \text{fmap}_{F^B \rightarrow ?} \text{ap} \\ \text{fmap } f \nearrow \quad \quad \quad \text{ap}^{[B, C \rightarrow Z]} \searrow & & \\ F^A & \xrightarrow{\quad \quad \quad \quad \quad \quad \quad \quad \quad} & (F^B \rightarrow F^{C \rightarrow Z}) \\ \text{fmap}_3(f^{A \rightarrow B \rightarrow C \rightarrow Z}) & & \end{array}$$

$$\begin{array}{ccc} \text{fmap } f \nearrow \quad \quad \quad \text{ap}^{[B, C \rightarrow Z]} \searrow & & \\ F^B \rightarrow C \rightarrow Z & \xrightarrow{\quad \quad \quad \quad \quad \quad \quad \quad \quad} & (F^B \rightarrow F^{C \rightarrow Z}) \\ \text{fmap } f \nearrow \quad \quad \quad \text{fmap}_{F^B \rightarrow ?} \text{ap}^{[C, Z]} \searrow & & \\ F^A & \xrightarrow{\quad \quad \quad \quad \quad \quad \quad \quad \quad} & (F^B \rightarrow F^C \rightarrow F^Z) \\ \text{fmap}_3(f^{A \rightarrow B \rightarrow C \rightarrow Z}) & & \end{array}$$

- Using the infix syntax will get rid of $\text{fmap}_{F^B \rightarrow ?} \text{ap}$ (see example code)
 - * Note the pattern: a natural transformation is equivalent to a lifting

Deriving the `zip` operation from `map2`

- The types $A \rightarrow B \rightarrow C$ and $A \times B \rightarrow C$ are equivalent (curry/uncurry)
 - Uncurry fmap_2 to $\text{fmap2} : (A \times B \rightarrow C) \rightarrow F^A \times F^B \rightarrow F^C$
 - Compute $\text{fmap2}(f)$ with $f = \text{id}^{A \times B \rightarrow A \times B}$, expecting to obtain a simpler natural transformation:

$$\text{zip} : F^A \times F^B \rightarrow F^{A \times B}$$

- This is quite similar to `zip` for lists:

`List(1, 2).zip(List(10, 20)) = List((1, 10), (2, 20))`

- The functions `zip` and `fmap2` are computationally equivalent:

$$\begin{aligned} \text{zip} &= \text{fmap2}(\text{id}) \\ \text{fmap2}(f^{A \times B \rightarrow C}) &= \text{zip} \circ \text{fmap } f \end{aligned}$$

$$\begin{array}{ccc} & F^{A \times B} & \\ \text{zip} \nearrow & \searrow \text{fmap } f^{A \times B \rightarrow C} & \\ F^A \times F^B & \xrightarrow{\quad \text{fmap2}(f^{A \times B \rightarrow C}) \quad} & F^C \end{array}$$

- The functor F is **zippable** if such a `zip` exists (with appropriate laws)
 - * The same pattern: a natural transformation is equivalent to a lifting

* Equivalence of the operations `ap` and `zip`

- Set $A \equiv B \rightarrow C$, get $\text{zip}^{[B \rightarrow C, B]} : F^{B \rightarrow C} \times F^B \rightarrow F^{(B \rightarrow C) \times B}$
 - Use `eval` : $(B \rightarrow C) \times B \rightarrow C$ and $\text{fmap}(\text{eval}) : F^{(B \rightarrow C) \times B} \rightarrow F^C$
 - Uncurry: $\text{app}^{[B, C]} : F^{B \rightarrow C} \times F^B \rightarrow F^C \equiv \text{zip} \circ \text{fmap}(\text{eval})$
 - The functions `zip` and `app` are computationally equivalent:
 - * use pair : $(A \rightarrow B \rightarrow A \times B) = a^A \rightarrow b^B \rightarrow a \times b$
 - * use $\text{fmap}(\text{pair}) \equiv \text{pair}^\uparrow$ on an fa^{F^A} , get $(\text{pair}^\uparrow fa) : F^{B \rightarrow A \times B}$; then

$$\begin{aligned} \text{zip}(fa \times fb) &= \text{app}((\text{pair}^\uparrow fa) \times fb) \\ \text{app}^{[B, C]} &= \text{zip}^{[B \rightarrow C, B]} \circ \text{fmap}(\text{eval}) \end{aligned}$$

$$\begin{array}{ccc} & F^{(B \rightarrow C) \times B} & \\ \text{zip} \nearrow & \searrow \text{fmap}(\text{eval}) & \\ F^{B \rightarrow C} \times F^B & \xrightarrow{\quad \text{app}^{[B, C]} \quad} & F^C \end{array}$$

- Rewrite this using curried arguments: $\text{fzip}^{[A, B]} : F^A \rightarrow F^B \rightarrow F^{A \times B}$; $\text{ap}^{[B, C]} : F^{B \rightarrow C} \rightarrow F^B \rightarrow F^C$; then $\text{ap } f = \text{fzip } f \circ \text{fmap}(\text{eval})$.
 - Now $\text{fzip } p^{F^A} q^{F^B} = \text{ap}((\text{pair}^\uparrow p) q)$, hence we may omit the argument q : $\text{fzip} = \text{pair}^\uparrow \circ \text{ap}$. With explicit types: $\text{fzip}^{[A, B]} = \text{pair}^\uparrow \circ \text{ap}^{[B, A \rightarrow B]}$.

Motivation for applicative laws. Naturality laws for `map2` Treat `map2` as a replacement for a monadic block with independent effects:

<pre>for { x ← cont1 y ← cont2 } yield g(x, y)</pre>	<pre>map2 (cont1, cont2) { (x, y) → g(x, y) }</pre>
--	---

- Main idea: Formulate the monad laws in terms of `map2` and `pure`

Naturality laws: Manipulate data in one of the containers

<pre>for { x ← cont1.map(f) y ← cont2 } yield g(x, y)</pre>	<pre>for { x ← cont1 y ← cont2 } yield g(f(x), y)</pre>
---	---

and similarly for `cont2` instead of `cont1`; now rewrite in terms of `map2`:

- **Left naturality for `map2`:**

```
map2(cont1.map(f), cont2)(g)
= map2(cont1, cont2){ (x, y) → g(f(x), y) }
```

- **Right naturality for `map2`:**

```
map2(cont1, cont2.map(f))(g)
= map2(cont1, cont2){ (x, y) → g(x, f(y)) }
```

Associativity and identity laws for `map2` Inline two generators out of three, in two different ways:

<pre>for { x ← cont1 (y, z) ← for { yy ← cont2 zz ← cont3 } yield (yy, zz) } yield g(x, y, z)</pre>	<pre>for { (x, y) ← for { xx ← cont1 yy ← cont2 } yield (xx, yy) z ← cont3 } yield g(x, y, z)</pre>
---	---

Write this in terms of `map2` to obtain the **associativity law for `map2`**:

```
map2(cont1, map2(cont2, cont3)((_,_)){ case(x,(y,z))→g(x,y,z) })
= map2(map2(cont1, cont2)((_,_), cont3){ case((x,y),z)→g(x,y,z) })
```

Empty context precedes a generator, or follows a generator:

<pre>for { x ← pure(a) y ← cont } yield g(x, y)</pre>	<pre>for { y ← cont } yield g(a, y)</pre>
---	---

Write this in terms of `map2` to obtain the **identity laws for `map2`** and `pure`:

```
map2(pure(a), cont)(g) = cont.map { y → g(a, y) }
map2(cont, pure(b))(g) = cont.map { x → g(x, b) }
```

Deriving the laws for `zip`: naturality law

- The laws for `map2` in a short notation; here $f \otimes g \equiv \{a \times b \rightarrow f(a) \times g(b)\}$

$$\begin{aligned}
 \text{fmap2}\left(g^{A \times B \rightarrow C}\right)\left(f^\uparrow q_1 \times q_2\right) &= \text{fmap2}((f \otimes \text{id}) \circ g)(q_1 \times q_2) \\
 \text{fmap2}\left(g^{A \times B \rightarrow C}\right)\left(q_1 \times f^\uparrow q_2\right) &= \text{fmap2}((\text{id} \otimes f) \circ g)(q_1 \times q_2) \\
 \text{fmap2}(g_{1,23})(q_1 \times \text{fmap2}(\text{id})(q_2 \times q_3)) &= \text{fmap2}(g_{12,3})(\text{fmap2}(\text{id})(q_1 \times q_2) \times q_3) \\
 \text{fmap2}\left(g^{A \times B \rightarrow C}\right)\left(\text{pure } a^A \times q_2^{F^B}\right) &= (b \rightarrow g(a \times b))^\uparrow q_2 \\
 \text{fmap2}\left(g^{A \times B \rightarrow C}\right)\left(q_1^{F^A} \times \text{pure } b^B\right) &= (a \rightarrow g(a \times b))^\uparrow q_1
 \end{aligned}$$

- Express `map2` through `zip`:

$$\begin{aligned}\text{fmap}_2 g^{A \times B \rightarrow C} (q_1^{F^A} \times q_2^{F^B}) &\equiv (\text{zip} \circ g^\uparrow)(q_1 \times q_2) \\ \text{fmap}_2 g^{A \times B \rightarrow C} &\equiv \text{zip} \circ g^\uparrow\end{aligned}$$

- Combine the two naturality laws into one by using two functions f_1, f_2 :

$$\begin{aligned}(f_1^\uparrow \otimes f_2^\uparrow) \circ \text{fmap}_2 g &= \text{fmap}_2 ((f_1 \otimes f_2)^\uparrow \circ g) \\ (f_1^\uparrow \otimes f_2^\uparrow) \circ \text{zip} \circ g^\uparrow &= \text{zip} \circ (f_1 \otimes f_2)^\uparrow \circ g^\uparrow\end{aligned}$$

- The **naturality law** for `zip` then becomes: $(f_1^\uparrow \otimes f_2^\uparrow) \circ \text{zip} = \text{zip} \circ (f_1 \otimes f_2)^\uparrow$

Deriving the laws for `zip`: associativity law

- Express `map2` through `zip` and substitute into the associativity law:

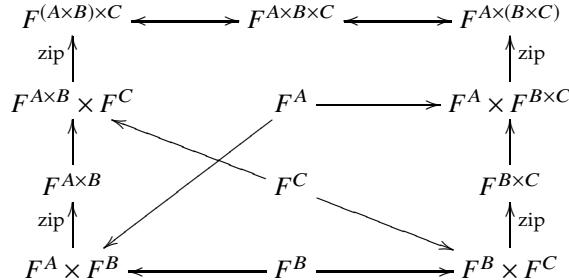
$$g_{1,2,3}^\uparrow (\text{zip} (q_1 \times \text{zip} (q_2 \times q_3))) = g_{12,3}^\uparrow (\text{zip} (\text{zip} (q_1 \times q_2) \times q_3))$$

- The arbitrary function g is preceded by transformations of the tuples,

$$a \times (b \times c) \equiv (a \times b) \times c \quad (\text{type isomorphism})$$

- Assume that the isomorphism transformations are applied as needed, then we may formulate the **associativity law** for `zip` more concisely:

$$\text{zip} (\text{zip} (q_1 \times q_2) \times q_3) \cong \text{zip} (q_1 \times \text{zip} (q_2 \times q_3))$$



Deriving the laws for `zip`: identity laws

- Identity laws seem to be complicated, e.g. the left identity:

$$g^\uparrow (\text{zip} (\text{pure} a \times q)) = (b \rightarrow g (a \times b))^\uparrow q$$

- Replace `pure` by an *equivalent* “wrapped unit” method `wu: F[Unit]`

$$\text{wu}^{F^1} \equiv \text{pure}(1); \quad \text{pure}(a^A) = (1 \rightarrow a)^\uparrow \text{wu}$$

Then the left identity law can be simplified using left naturality:

$$g^\uparrow (\text{zip} ((1 \rightarrow a)^\uparrow \text{wu} \times q)) = g^\uparrow (((1 \rightarrow a) \otimes \text{id})^\uparrow \text{zip} (\text{wu} \times q))$$

- Denote $\phi^{B \rightarrow 1 \times B} \equiv b \rightarrow 1 \times b$ and $\beta_a^{1 \times B \rightarrow A \times B} \equiv (1 \rightarrow a) \otimes \text{id}$; then the function $b \rightarrow g (a \times b)$ can be expressed more simply as $\phi \circ \beta_a \circ g$, and the identity law becomes

$$g^\uparrow (\beta_a^\uparrow \text{zip} (\text{wu} \times q)) = (\beta_a \circ g)^\uparrow (\text{zip} (\text{wu} \times q)) = (\phi \circ \beta_a \circ g)^\uparrow q = (\beta_a \circ g)^\uparrow (\phi^\uparrow q)$$

Omitting the common prefix $(\beta_a \circ g)^\uparrow$, we obtain the **left identity law**:

$$\text{zip} (\text{wu} \times q) = \phi^\uparrow q$$

- * Note that ϕ^\uparrow is an isomorphism between F^B and $F^{1 \times B}$

- Assume that this isomorphism is applied as needed, then we may write

$$\text{zip} (\text{wu} \times q) \cong q$$

- * Similarly, the **right identity law** can be written as $\text{zip} (q \times \text{wu}) \cong q$

Similarity between applicative laws and monoid laws

- Define infix syntax for `zip` and write $\text{zip}(p \times q) \equiv p \bowtie q$
 - Then the associativity and identity laws may be written as

$$\begin{aligned} q_1 \bowtie (q_2 \bowtie q_3) &\cong (q_1 \bowtie q_2) \bowtie q_3 \\ (\text{wu} \bowtie q) &\cong q \\ (q \bowtie \text{wu}) &\cong q \end{aligned}$$

These are the laws of a monoid (with some assumed transformations)

- Naturality law for `zip` written in the infix syntax:

$$f_1^\uparrow q_1 \bowtie f_2^\uparrow q_2 = (f_1 \otimes f_2)^\uparrow (q_1 \bowtie q_2)$$

- `wu` has no laws; the naturality for `pure` follows automatically
- The laws are simplest when formulated in terms of `zip` and `wu`
 - * Naturality for `zip` will usually follow from parametricity
 - A third naturality law for `map2` follows from defining `map2` through `zip!`
- “Zippable” functors have only the associativity and naturality laws
- Applicative functors are a strict superset of monadic functors
 - * There are applicative functors that *cannot* be monads
 - * Applicative functor implementation may disagree with the monad

A third naturality law for `map2`

- There must be one more naturality law for `map2`
 - Transform the result of a `map2`:

<pre>(for { x ← cont1 y ← cont2 } yield g(x, y)).map(f)</pre>	<pre>for { x ← cont1 y ← cont2 } yield f(g(x, y))</pre>
---	---

- Write this in terms of `map2`, obtain a third naturality law:

$$\begin{aligned} \text{map2}(\text{cont1}, \text{cont2})(g) \cdot \text{map}(f) &= \text{map2}(\text{cont1}, \text{cont2})(g \text{ andThen } f) \\ \text{fmap2}(g) \circ f^\uparrow &= \text{fmap2}(g \circ f) \\ f^\uparrow(\text{fmap2}(g)(p \times q)) &= \text{fmap2}(g \circ f)(p \times q) \end{aligned}$$

- This law automatically follows if we define `map2` through `zip`:

$$\text{fmap2}(g) \circ f^\uparrow = \text{zip} \circ g^\uparrow \circ f^\uparrow = \text{zip} \circ (g \circ f)^\uparrow$$

- Note: We always have one naturality law per type parameter

Applicative operation `ap` as a “lifting”

- Consider `ap` as a “lifting” since it has type $F^{A \rightarrow B} \rightarrow (F^A \rightarrow F^B)$
 - A “lifting” should obey the identity and the composition laws
 - * An “identity” value of type $F^{A \rightarrow A}$, mapped to $\text{id}^{F^{A \rightarrow F^A}}$ by `ap`
 - A good candidate for that value is $\text{id}_\circ \equiv \text{pure}(\text{id}^{A \rightarrow A})$
 - * A “composition” of an $F^{A \rightarrow B}$ and an $F^{B \rightarrow C}$, yielding an $F^{A \rightarrow C}$

- We can use `map2` to implement this composition, denoted $g \odot h$:

$$g^{F^{A \rightarrow B}} \odot h^{F^{B \rightarrow C}} \equiv \text{fmap2}(p^{A \rightarrow B} \times q^{B \rightarrow C} \rightarrow p \circ q)(g, h)$$

- What are the laws that follow for $g \odot h$ from the `map2` laws?

$$\begin{aligned} \text{id}_\odot \odot h &= h; \quad g \odot \text{id}_\odot = g \\ g^{F^{A \rightarrow B}} \odot (h^{F^{B \rightarrow C}} \odot k^{F^{C \rightarrow D}}) &= (g \odot h) \odot k \\ \left((x^{B \rightarrow C} \rightarrow f^{A \rightarrow B} \circ x)^\dagger g^{F^{B \rightarrow C}} \right) \odot h^{F^{C \rightarrow D}} &= (x^{B \rightarrow D} \rightarrow f^{A \rightarrow B} \circ x)^\dagger (g \odot h) \\ g^{F^{A \rightarrow B}} \odot \left((x^{B \rightarrow C} \rightarrow x \circ f^{C \rightarrow D})^\dagger h^{F^{B \rightarrow C}} \right) &= (x^{A \rightarrow C} \rightarrow x \circ f^{C \rightarrow D})^\dagger (g \odot h) \end{aligned}$$

- * The first 3 laws are the identity & associativity laws of a *category*
 - The morphism type is $A \rightsquigarrow B \equiv F^{A \rightarrow B}$, the composition is \odot
- * The last 2 laws are naturality laws, connecting `fmap` and \odot
- Therefore `ap` is a functor's “lifting” of morphisms from two categories

Deriving the category laws for (id_\odot, \odot) The five laws for id_\odot and \odot follow from the five `map2` laws

- Consider $\text{id}_\odot \odot h$ and substitute the definition of \odot via `map2`, cf. slide 7: $\text{id}_\odot \odot h = \text{fmap2}(p \times q \rightarrow p \circ q)(\text{pure}(b \rightarrow \text{id} \circ b)^\dagger h = h)$
 - The law $g \odot \text{id}_\odot = g$ is derived similarly
 - Associativity law: $g \odot (h \odot k) = \text{fmap2}(\circ)(g \times \text{fmap2}(\circ)(h \times k))$ The 3rd naturality law gives: $\text{fmap2}(\circ)(h \times k) = (\circ)^\dagger(\text{fmap2}(\text{id})(h \times k))$, and then:

$$\begin{aligned} g \odot (h \odot k) &= \text{fmap2}(x \times (y \times z) \rightarrow x \circ y \circ z)(g \times \text{fmap2}(\text{id})(h \times k)) \\ (g \odot h) \odot k &= \text{fmap2}((x \times y) \times z \rightarrow x \circ y \circ z)(\text{fmap2}(\text{id})(g \times h) \times k) \end{aligned}$$

Now the associativity law for `fmap2` yields $g \odot (h \odot k) = (g \odot h) \odot k$

- Derive naturality laws for \odot from the three `map2` naturality laws: $\left((x \rightarrow f \circ x)^\dagger g \right) \odot h = \text{fmap2}(\circ) \left((x \rightarrow f \circ x)^\dagger \text{fmap2}(\circ)(h \times k) \right) = (x \rightarrow f \circ x)^\dagger (h \odot k)$
- The law is $g \odot (x \rightarrow f \circ x)^\dagger h = (x \rightarrow f \circ x)^\dagger (g \odot h)$ is derived similarly

Deriving the functor laws for `ap` Now that we established the laws for \odot , we have `ap` laws:

$$\text{ap}^{[B, Z]} : F^{B \rightarrow Z} \rightarrow F^B \rightarrow F^Z = \text{fmap}_2 \left(\text{id}^{(B \rightarrow Z) \rightarrow (B \rightarrow Z)} \right)$$

Identity law: $\text{ap}(\text{id}_\odot) = \text{id}^{F^A \rightarrow F^A}$

- Derivation: $\text{ap}(\text{id}_\odot^{F^A \rightarrow A})(q^{F^A}) = \text{fmap}_2(\text{id}^{(A \rightarrow A) \rightarrow A \rightarrow A})(\text{pure}(\text{id}^{A \rightarrow A}))(q^{F^A}) = \text{fmap2}(f \times x \rightarrow f(x))(\text{pure}(\text{id}) \times q) = (x \rightarrow \text{id}(x))^\dagger q = \text{id}^\dagger q = q$
 - Easier derivation: first, express `ap` via \odot using the isomorphisms

$$A \cong 1 \rightarrow A; \quad F^A \cong F^{1 \rightarrow A}$$

Then $\text{ap}(p^{F^{B \rightarrow Z}})(q^{F^B}) \cong q^{F^{1 \rightarrow B}} \odot p^{F^{B \rightarrow Z}}$ and so $\text{ap}(\text{id}_\odot)(q) \cong q \odot \text{id}_\odot = q$

Composition law: $\text{ap}(g \odot h) = \text{ap}(g) \circ \text{ap}(h)$

- Derivation: use $\text{ap}(p \circ q) \cong q \odot p$ to get $\text{ap}(g \odot h)(q) \cong q \odot (g \odot h)$ while $(\text{ap}(g) \circ \text{ap}(h))q = \text{ap}(h)(\text{ap}(g)(q)) \cong \text{ap}(h)(q \odot g) \cong (q \odot g) \odot h$

Constructions of applicative functors

- All monadic constructions still hold for applicative functors
 - Additionally, there are some non-monadic constructions
 - $F^A \equiv 1$ (constant functor) and $F^A \equiv A$ (identity functor)
 - $F^A \equiv G^A \times H^A$ for any applicative G^A and H^A
 - but $G^A + H^A$ is in general *not* applicative
 - $F^A \equiv A + G^A$ for any applicative G^A (**free pointed** over G)
 - $F^A \equiv A + G^{F^A}$ (recursive) for *any* functor G^A (**free monad** over G)
 - $F^A \equiv H^A \rightarrow A$ for *any* contrafunctor H^A
 - Constructions that do not correspond to monadic ones:
 - $F^A \equiv Z$ (constant functor, Z a monoid)
 - $F^A \equiv Z + G^A$ for any applicative G^A and monoid Z
 - $F^A \equiv G^{H^A}$ when both G and H are applicative
 - Applicative that disagrees with its monad: $F^A \equiv 1 + (1 \rightarrow A \times F^A)$
 - Examples of non-applicative functors: $F^A \equiv (P \rightarrow A) + (Q \rightarrow A)$, $F^A \equiv (A \rightarrow P) \rightarrow Q$, $F^A \equiv (A \rightarrow P) \rightarrow 1 + A$

All non-parameterized exp-poly types are monoids

- Using known monoid constructions (Chapter 7), we can implement $X + Y$, $X \times Y$, $X \rightarrow Y$ as monoids when X and Y are monoids

By constructions 1, 2, 6, 7, *all* polynomial F^A with monoidal coefficients are applicative: write $F^A = Z_1 + A \times (Z_2 + A \times \dots)$ with some monoids Z_i

- Examples: $F^A = 1 + A \times A$ (this F^A cannot be a monad!)
- $F^A = A + A \times A \times Z$ where Z is a monoid (this F^A is a monad)

Previous examples of non-applicative functors are all *non-polynomial*

Definition and constructions of applicative contrafunctors

- The applicative functor laws, if formulated via `zip` and `wu`, do not use `map` and therefore can be formulated for contrafunctors

- Define an **applicative contrafunctor** C^A as having `zip` and `wu`:

$$\text{zip} : C^A \times C^B \rightarrow C^{A \times B}; \quad \text{wu} : C^1$$

- Identity and associativity laws must hold for `zip` and `wu`
 - * Note: applying `contramap` to the function $a \times b \rightarrow a$ will yield some $C^A \rightarrow C^{A \times B}$, but this will *not* give a valid implementation of `zip`!
- Naturality must hold for `zip`, but with `contramap` instead of `map`
 - * There are no corresponding `pure` or `contraap!` But have `cpure` : $\forall A : C^A$

Applicative contrafunctor constructions:

- $C^A \equiv Z$ (constant functor, Z a monoid)
- $C^A \equiv G^A \times H^A$ for any applicative contrafunctors G^A and H^A
- $C^A \equiv G^A + H^A$ for any applicative contrafunctors G^A and H^A
- $C^A \equiv H^A \rightarrow G^A$ for *any* functor H^A and applicative contrafunctor G^A
- $C^A \equiv G^{H^A}$ if a functor G^A and contrafunctor H^A are both applicative
 - *All* exponential-polynomial contrafunctors with monoidal coefficients are applicative! (These constructions cover all exp-poly cases.)

Definition and laws of profunctors

- **Profunctors** have the type parameter in both contravariant and covariant positions; they can have neither `map` nor `contramap`
 - Examples of profunctors: $P^A \triangleq \mathbb{1} + \text{Int} \times A \rightarrow A$; $P^A \triangleq A + (A \rightarrow \text{String})$
 - Example of non-profunctor: a GADT, $F^A \triangleq \text{String}^{F^{\text{Int}}} + \text{Int}^{F^1}$

Definition and constructions of applicative profunctors

- Definition of **applicative profunctor**: has `zip` and `wu` with the laws
 - There is no corresponding `ap!` But have `pure` : $A \rightarrow P^A$

Applicative profunctors admit all previous constructions, and in addition:

1. $P^A \equiv G^A \times H^A$ for any applicative profunctors G^A and H^A
2. $P^A \equiv Z + G^A$ for any applicative profunctor G^A and monoid Z
3. $P^A \equiv A + G^A$ for any applicative profunctor G^A
4. $P^A \equiv F^A \rightarrow Q^A$ for *any functor* F^A and applicative profunctor Q^A
 - Non-working construction: $P^A \equiv H^A \rightarrow A$ for a profunctor H^A
5. $P^A \equiv G^{H^A}$ for a functor G^A and a profunctor H^A , both applicative

Commutative applicative functors

- The monoidal operation \oplus can be **commutative** w.r.t. its arguments:

$$x \oplus y = y \oplus x$$

- Applicative operation `zip` can be **commutative** w.r.t. its arguments:

$$(a \times b \rightarrow b \times a)^\uparrow (fa \bowtie fb) = fb \bowtie fa$$

or $fa \bowtie fb \cong fb \bowtie fa$, implicitly using the isomorphism $a \times b \rightarrow b \times a$

- Applicative functor is commutative if the second effect is independent of the first effect (not only of the first value)

- Examples:

- * List is commutative; applicative parsers are not
- * If defined through the monad instance, `zip` is usually not commutative
- * All polynomial functors with *commutative* monoidal coefficients are commutative applicative functors

- Most applicative constructions preserve commutativity

- The same applies to applicative contrafunctors and profunctors

- Commutativity makes proving associativity easier:

$$(fa \bowtie fb) \bowtie fc \cong fc \bowtie (fb \bowtie fa)$$

so it's sufficient to swap fa and fc and show equivalence

Categorical overview of “regular” functor classes The “liftings” show the types of category’s morphisms

class name	lifting's name and type signature	category's morphism
functor	$fmap : (A \rightarrow B) \rightarrow F^A \rightarrow F^B$	$A \rightarrow B$
filterable	$liftOpt : (A \rightarrow 1 + B) \rightarrow F^A \rightarrow F^B$	$A \rightarrow 1 + B$
monad	$flm : (A \rightarrow F^B) \rightarrow F^A \rightarrow F^B$	$A \rightarrow F^B$
applicative	$ap : F^{A \rightarrow B} \rightarrow F^A \rightarrow F^B$	$F^{A \rightarrow B}$
contrafunctor	$cmap : (B \rightarrow A) \rightarrow F^A \rightarrow F^B$	$B \rightarrow A$
profunctor	$xmap : (A \rightarrow B) \times (B \rightarrow A) \rightarrow F^A \rightarrow F^B$	$(A \rightarrow B) \times (B \rightarrow A)$
contra-filterable	$liftOpt : (B \rightarrow 1 + A) \rightarrow F^A \rightarrow F^B$	$B \rightarrow 1 + A$
Not yet considered:		
comonad	$coflm : (F^A \rightarrow B) \rightarrow F^A \rightarrow F^B$	$F^A \rightarrow B$

Need to define each category's composition and identity morphism

Then impose the category laws, the naturality laws, and the functor laws

- Obtained a systematic picture of the “regular” type classes
 - Some classes (e.g. contra-applicative) aren’t covered by this scheme
 - Some of the possibilities (e.g. “contramonad”) don’t actually work out

11.2.1 Exercises

1. Show that pure will be automatically a natural transformation when it is defined using wu as shown in the slides.
2. Use naturality of pure to show that $\text{pure } f \odot \text{pure } g = \text{pure } (f \circ g)$
3. Show that $F^A \equiv (A \rightarrow Z) \rightarrow (1 + A)$ is a functor but not applicative.
4. Show that P^S is a monoid if S is a monoid and P is any applicative functor, contrafunctor, or profunctor.
5. Implement an applicative instance for $F^A = 1 + \text{Int} \times A + A \times A \times A$.
6. Using applicative constructions, show (without need for verifying laws) that $F^A = G^A + H^{G^A}$ is applicative if G and H are applicative functors.
7. Explicitly implement contrafunctor construction 2 and prove the laws.
8. For any contrafunctor H^A , construction 5 says that $F^A \equiv H^A \rightarrow A$ is applicative. Implement the code of $\text{zip}(fa, fb)$ for this construction.
9. Show that the recursive functor $F^A \equiv 1 + G^{A \times F^A}$ is applicative if G^A is applicative and wu_F is defined recursively as $0 + \text{pure}_G (1 \times \text{wu}_F)$.
10. Explicitly implement profunctor construction 5 and prove the laws.
11. Prove rigorously that all exponential-polynomial type constructors are profunctors.
12. Implement profunctor and applicative instances for $P^A \equiv A + Z \times G^A$ where G^A is a given applicative profunctor and Z is a monoid.
13. Show that, for any profunctor P^A , one can implement a function of type $A \rightarrow P^B \rightarrow P^{A \times B}$ but not of type $A \rightarrow P^B \rightarrow P^A$.

11.3 Practical use

Discussion

11.4 Laws and structure

12 Traversable functors and profunctors

12.1 Slides

Motivation for the `traverse` operation Consider data of type List^A and processing $f : A \rightarrow \text{Future}^B$

Typically, we want to wait until the entire data set is processed

What we need is $\text{List}^A \rightarrow (A \rightarrow \text{Future}^B) \rightarrow \text{Future}^{\text{List}^B}$

Generalize: $L^A \rightarrow (A \rightarrow F^B) \rightarrow F^{L^B}$ for some type constructors F, L

This operation is called `traverse`

How to implement it: for example, a 3-element list is $A \times A \times A$

Consider $L^A \equiv A \times A \times A$, apply map f and get $F^B \times F^B \times F^B$

We will get $F^{L^B} \equiv F^{B \times B \times B}$ if we can apply `zip` as $F^B \times F^B \rightarrow F^{B \times B}$

So we need to assume that F is applicative

In Scala, we have `Future.traverse()` that assumes L to be a sequence

This is the easy-to-remember example that fixes the requirements

Questions:

Which functors L can have this operation?

Can we express `traverse` through a simpler operation?

What are the required laws for `traverse`?

What about contrafunctors or profunctors?

Deriving the `sequence` operation The type signature of `traverse` is a complicated “lifting”

A “lifting” is often equivalent to a simpler natural transformation

To derive it, ask: what is missing from `fmap` to do the job of `traverse`?

$$\text{fmap}_L : (A \rightarrow F^B) \rightarrow L^A \rightarrow L^{F^B}$$

We need F^{L^B} , but the `traverse` operation gives us L^{F^B} instead

What's missing is a natural transformation `sequence` : $L^{F^B} \rightarrow F^{L^B}$

The functions `traverse` and `sequence` are computationally equivalent:

$$\text{trav } f \xrightarrow{A \rightarrow F^B} = \text{fmap}_L f \circ \text{seq}$$

$$\begin{array}{ccc} & L^{F^B} & \\ \text{fmap}_L f \xrightarrow{A \rightarrow F^B} & \nearrow & \searrow \text{seq} \\ L^A & \xrightarrow{\text{trav } f \xrightarrow{A \rightarrow F^B}} & F^{L^B} \end{array}$$

Here F is an *arbitrary* applicative functor

Keep in mind the example `Future.sequence` : $\text{List}^{\text{Future}^X} \rightarrow \text{Future}^{\text{List}^X}$

Examples: $L^A \equiv A \times A \times A$; $L^A = \text{List}^A$; finite trees

Non-traversable: $L^A \equiv R \rightarrow A$; lazy list (“infinite product”)

Note: We *cannot have* the opposite transformation $F^{L^B} \rightarrow L^{F^B}$

Polynomial functors are traversable Generalize from the example $L^A \equiv A \times A \times A$ to other polynomial functors

Polynomial functors have the form

$$L^A \equiv Z \times A \times \dots \times A + Y \times A \times \dots \times A + \dots + Q \times A + P$$

To implement `seq` : $L^{F^B} \rightarrow F^{L^B}$, consider monomial $L^A \equiv Z \times A \times \dots \times A$

We have $L^{F^B} = Z \times F^B \times \dots \times F^B$; apply `zip` and get $Z \times F^{B \times \dots \times B}$

Lift Z into the functor F using $Z \rightarrow F^A \rightarrow F^{Z \times A}$ (or with $F.\text{pure}$)

The result is $F^{Z \times B \times \dots \times B} \equiv F^{L^B}$

For a polynomial L^A , do this to each monomial, then lift to F^{L^B}

Note that we could apply `zip` in various different orders

The traversal order is arbitrary, may be application-specific

Non-polynomial functors are not traversable (see [Bird et al., 2013](#))

Example: $L^A \equiv E \rightarrow A$; $F^A \equiv 1 + A$; can't have $\text{seq} : L^{F^B} \rightarrow F^{L^B}$

All polynomial functors are traversable, and usually in several ways

It is still useful to have a type class for traversable functors

Motivation for the laws of the `traverse` operation The “[law of traversals](#)”{} paper (2012) argues that `traverse` should “visit each element” of the container L^A exactly once, and evaluate each corresponding “effect” F^B exactly once; then they formulate the laws

To derive the laws, use the “lifting” intuition for `traverse`,

$$\text{trav} : (A \rightarrow F^B) \rightarrow L^A \rightarrow F^{L^B}$$

Look for “identity” and “composition” laws:

“Identity” as `pure` : $A \rightarrow F^A$ must be lifted to $\text{pure} : L^A \rightarrow F^{L^A}$

“Identity” as $\text{id}^{\underline{A \rightarrow A}}$ with $F^A \equiv A$ (identity functor) lifted to $\text{id}^{\underline{L^A \rightarrow L^A}}$

“Compose” $f : A \rightarrow F^B$ and $g : B \rightarrow G^C$ to get $h : A \rightarrow F^{G^C}$, where F, G are applicative; a traversal with h maps L^A to F^{G^C} and must be equal to the composition of traversals with f and then with $g^{F \uparrow}$

Questions:

Are the laws for the `sequence` operation simpler?

Are all these laws independent?

What functors L satisfy these laws for all applicative functors F ?

Formulation of the laws for `traverse` Identity law: For any applicative functor F ,

$$\text{trav}(\text{pure}) = \text{pure}$$

$$L^A \xrightarrow[\text{trav}(\text{pure}^{\underline{A \rightarrow F^A}})]{\text{pure}^{\underline{L^A \rightarrow F^{L^A}}}} F^{L^A}$$

Second identity law: $\text{trav}^{\text{Id}}(\text{id}^A) = \text{id}^{L^A}$ is a consequence with $F = \text{Id}$

So, we need only one identity law

Composition law: For any $f^{\underline{A \rightarrow F^B}}$ and $g^{\underline{B \rightarrow G^C}}$, & applicative F and G ,

$$\text{trav } f \circ (\text{trav } g)^{F \uparrow} = \text{trav} (f \circ g^{F \uparrow})$$

$$L^A \xrightarrow[\text{trav}^F h^{\underline{A \rightarrow F^G C}}]{\text{trav}^F f^{\underline{A \rightarrow F^B}} \text{ fmap}_F (\text{trav}^G g)^{\underline{B \rightarrow G^C}}} F^{G^C}$$

where $h^{\underline{A \rightarrow F^G C}} \equiv f \circ g^{F \uparrow}$. (Note: $H^A \equiv F^{G^A}$ is applicative!)

Derivation of the laws for `sequence` Express $\text{trav } f = f^{L^A \uparrow} ; \text{seq}$ and substitute into the laws for trav :

Identity law: $\text{trav}(\text{pure}) = \text{pure}^{L^A \uparrow} ; \text{seq} = \text{pure}$

$$L^A \xrightarrow[\text{pure}^{L^A}]{\text{fmap}_L \text{pure}^A} L^{F^A} \xrightarrow{\text{seq}} F^{L^A}$$

Naturality law: $\text{seq} ; g^{F \uparrow L \uparrow} = g^{L \uparrow F \uparrow} ; \text{seq}$ with $g^{A \rightarrow B}$, mapping $L^{F^A} \rightarrow F^{L^B}$

Composition law:

$$\begin{aligned} \text{trav } f ; (\text{trav } g)^{F \uparrow} &= f^{L \uparrow} ; \text{seq} ; \left(g^{L \uparrow} ; \text{seq} \right)^{F \uparrow} \\ &= f^{L \uparrow} ; \text{seq} ; g^{L \uparrow F \uparrow} ; \text{seq}^{F \uparrow} = f^{L \uparrow} ; g^{F \uparrow L \uparrow} ; \text{seq} ; \text{seq}^{F \uparrow} \\ \text{trav } (f ; g^{F \uparrow}) &= (f ; g^{F \uparrow})^{L \uparrow} ; \text{seq} = f^{L \uparrow} ; g^{F \uparrow L \uparrow} ; \text{seq} \end{aligned}$$

Now omit the common prefix $f \cdots ; g \cdots$ and obtain: $\text{seq} ; \text{seq}^{F \uparrow} = \text{seq}$

$$\begin{array}{ccc} & \xrightarrow{\text{seq}^F} & F^{L^{G^A}} \\ L^{F^{G^A}} & \xrightarrow{\text{seq}^{F^G?}} & \xrightarrow{(\text{seq}^G)^{F \uparrow}} F^{G^{L^A}} \end{array}$$

Constructions of traversable and bitraversable functors Constructions of traversable functors:

$L^A \equiv Z$ (constant functor) and $L^A \equiv A$ (identity functor)

$L^A \equiv G^A \times H^A$ for any traversable G^A and H^A

$L^A \equiv G^A + H^A$ for any traversable G^A and H^A

$L^A \equiv S^{A,L^A}$ (recursive) for a bitraversable bifunctor $S^{A,B}$

If L^A is infinite, laws will appear to hold but `seq` will not terminate

A bifunctor $S^{A,B}$ is **bitraversable** if `biseq` exists such that

$$\text{biseq} : S^{F^A, F^B} \rightarrow F^{S^{A,B}}$$

for any applicative functor F ; the analogous laws must hold

Constructions of bitraversable bifunctors:

$S^{A,B} \equiv Z$, $S^{A,B} \equiv A$, and $S^{A,B} = B$

$S^{A,B} \equiv G^{A,B} \times H^{A,B}$ for any bitraversable G and H

$S^{A,B} \equiv G^{A,B} + H^{A,B}$ for any bitraversable G and H

All polynomial bifunctors are bitraversable

All polynomial functors, including recursive functors, are traversable

Foldable functors: traversing with respect to a monoid Take $F^A \equiv Z$ where Z is a monoid

The `zip` operation is the monoid operation \oplus

The type signature of `traverse` becomes $(A \rightarrow Z) \rightarrow L^A \rightarrow Z$

This method is called `foldMap`

The type signature of `seq` becomes $L^Z \rightarrow Z$

This is called `mconcat` – combines all values in L^Z with Z 's \oplus

It is convenient to define the `Foldable` type class

But it has no laws any more

All traversable functors are also foldable

The `foldLeft` method can be defined via `foldMap` with $Z \equiv (B \rightarrow B)$:

$$\text{foldl} : (A \rightarrow B \rightarrow B) \rightarrow L^A \rightarrow B \rightarrow B$$

Traversable contrafunctors and profunctors are not useful Traversing profunctors with respect to functors F : effects of F are ignored

All contrafunctors C^A are traversable w.r.t. applicative profunctors F^A ,

$$\text{seq} : C^{F^A} \rightarrow F^{C^A} \equiv \text{pure}^{C \downarrow} ; \text{pure}$$

$$C^{F^A} \xrightarrow{\text{cmap}_C \text{pure}_F^A} C^A \xrightarrow{\text{pure}_F^{C^A}} F^{C^A}$$

But not profunctors that are neither functors nor contrafunctors

Counterexample: $P^A \equiv A \rightarrow A$; need $\text{seq} : (F^A \rightarrow F^A) \rightarrow F^{A \rightarrow A}$; we can't get an $A \rightarrow A$, so the only implementation is to return $\text{pure}_F(\text{id})$, which ignores its argument and so will fail the identity law

12.2 Discussion

Part III

Advanced level

13 “Free” type constructions

13.1 Slides

The interpreter pattern I. Expression trees Main idea: Represent a program as a data structure, run it later

Example: a simple DSL for complex numbers

```
val a = "1+2*i".toComplex
val b = a * "3-4*i".toComplex
b.conj
Conj(
  Mul(
    Str("1+2*i"), Str("3-4*i"))
  ))
```

Unevaluated operations `Str`, `Mul`, `Conj` are defined as case classes:

```
sealed trait Prg
case class Str(s: String) extends Prg
case class Mul(p1: Prg, p2: Prg) extends Prg
case class Conj(p: Prg) extends Prg
```

An *interpreter* will “run” the program and return a complex number

```
def run(prg: Prg): (Double, Double) = ...
```

Benefits: programs are data, can compose & transform before running

Shortcomings: this DSL works only with simple expressions

Cannot represent variable binding and conditional computations

Cannot use any non-DSL code (e.g. a numerical algorithms library)

The interpreter pattern II. Variable binding A DSL with variable binding and conditional computations

Example: imperative API for reading and writing files

Need to bind a *non-DSL variable* to a value computed by DSL

Later, need to use that non-DSL variable in DSL expressions

The rest of the DSL program is a (Scala) function of that variable

```
val p = path("/file")
val str: String = read(p)
if (str.nonEmpty)
  read(path(str))
else "Error: empty path"
Bind(
  Read(Path(Literal("/file"))),
  { str → // read value ‘str’
    if (str.nonEmpty)
      Read(Path(Literal(str)))
    else Literal("Error: empty path")
  })
```

Unevaluated operations are implemented via case classes:

```
sealed trait Prg
case class Bind(p: Prg, f: String → Prg) extends Prg
case class Literal(s: String) extends Prg
case class Path(s: Prg) extends Prg
case class Read(p: Prg) extends Prg
```

Interpreter: `def run(prg: Prg): String = ...`

The interpreter pattern III. Type safety So far, the DSL has no type safety: every value is a `Prg`

We want to avoid errors, e.g. `Read(Read(...))` should not compile

Let `Prg[A]` denote a DSL program returning value of type `A` when run:

```
sealed trait Prg[A]
case class Bind(p: Prg[String], f: String → Prg[String])
  extends Prg[String]
case class Literal(s: String) extends Prg[String]
case class Path(s: Prg[String]) extends Prg[nio.file.Path]
case class Read(p: Prg[nio.file.Path]) extends Prg[String]
Interpreter: def run(prg: Prg[String]): String = ...
```

Our example DSL program is type-safe now:

```
val prg: Prg[String] = Bind(
  Read(Path(Literal("/file"))),
  { str: String →
    if (str.nonEmpty)
      Read(Path(Literal(str)))
    else Literal("Error: empty path")
  })
```

The interpreter pattern IV. Cleaning up the DSL Our DSL so far:

```
sealed trait Prg[A]
case class Bind(p: Prg[String], f: String → Prg[String])
  extends Prg[String]
case class Literal(s: String) extends Prg[String]
case class Path(s: Prg[String]) extends Prg[nio.file.Path]
case class Read(p: Prg[nio.file.Path]) extends Prg[String]
```

Problems with this DSL:

Cannot use `Read(p: nio.file.Path)`, only `Read(p: Prg[nio.file.Path])`
 Cannot bind variables or return values other than `String`

To fix these problems, make `Literal` a fully parameterized operation and replace `Prg[A]` by `A` in case class arguments

```
sealed trait Prg[A]
case class Bind[A, B](p: Prg[A], f: A → Prg[B]) extends Prg[B]
case class Literal[A](a: A) extends Prg[A]
case class Path(s: String) extends Prg[nio.file.Path]
case class Read(p: nio.file.Path) extends Prg[String]
```

The type signatures of `Bind` and `Literal` are like `flatMap` and `pure`

The interpreter pattern V. Define Monad-like methods We can actually define the methods `map`, `flatMap`, `pure`:

```
sealed trait Prg[A] {
  def flatMap[B](f: A → Prg[B]): Prg[B] = Bind(this, f)
  def map[B](f: A → B): Prg[B] = flatMap(this, f andThen Prg.pure)
}
object Prg { def pure[A](a: A): Prg[A] = Literal(a) }
```

These methods don't run anything, only create unevaluated structures

DSL programs can now be written as functor blocks and composed:

```
def readPath(p: String): Prg[String] = for {
  path ← Path(p)
  str ← Read(path)
} yield str
```

```
val prg: Prg[String] = for {
  str ← readPath("/file")
  result ← if (str.nonEmpty)
    readPath(str)
  else Prg.pure("Error: empty path")
} yield result
```

Interpreter: `def run[A](prg: Prg[A]): A = ...`

The interpreter pattern VI. Refactoring to an abstract DSL Write a DSL for complex numbers in a similar way:

```
sealed trait Prg[A] { def flatMap ... } // no code changes
case class Bind[A, B](p: Prg[A], f: A → Prg[B]) extends Prg[B]
case class Literal[A](a: A) extends Prg[A]
type Complex = (Double, Double) // custom code starts here
case class Str(s: String) extends Prg[Complex]
case class Mul(c1: Complex, c2: Complex) extends Prg[Complex]
case class Conj(c: Complex) extends Prg[Complex]
```

Refactor this DSL to separate common code from custom code:

```
sealed trait DSL[F[_], A] { def flatMap ... } // no code changes
type Prg[A] = DSL[F, A] // just for convenience
case class Bind[A, B](p: Prg[A], f: A → Prg[B]) extends Prg[B]
case class Literal[A](a: A) extends Prg[A]
```

```
case class Ops[A](f: F[A]) extends Prg[A] // custom operations here
Interpreter is parameterized by a "value extractor"  $Ex^F \equiv \forall A. (F^A \rightarrow A)$ 
def run[F[_], A](ex: Ex[F])(prg: DSL[F, A]): A = ...
The constructor DSL[F[_], A] is called a free monad over F
```

The interpreter pattern VII. Handling errors To handle errors, we want to evaluate `DSL[F[_], A]` to `Either[Err, A]`

Suppose we have a value extractor of type $Ex^F \equiv \forall A. (F^A \rightarrow Err + A)$

The code of the interpreter is almost unchanged:

```
def run[F[_], A](extract: Ex[F])(prg: DSL[F, A]): Either[Err, A] =
  prg match {
    case b: Bind[F, _, A] => b match { case Bind(p, f) =>
      run(extract)(p).flatMap(f andThen run(extract))
    } // Here, the .flatMap is from Either.
    case Literal(a) => Right(a) // pure: A → Err + A
    case Ops(f) => extract(f)
  }
```

The code of `run` only uses `flatMap` and `pure` from `Either`

We can generalize to any other monad M^A instead of `Either[Err, A]`

The resulting construction:

Start with an "operations type constructor" F^A (often not a functor)

Use $DSL^{F,A}$ and interpreter $run^{M,A} : (\forall X. F^X \rightarrow M^X) \rightarrow DSL^{F,A} \rightarrow M^A$

Create a DSL program `prg : DSLF,A` and an extractor `exX : FX → MX`

Run the program with the extractor: `run(ex)(prg)`; get a value M^A

The interpreter pattern VIII. Monadic DSLs: summary Begin with a number of operations, which are typically functions of fixed known types such as $A_1 \rightarrow B_1, A_2 \rightarrow B_2$ etc.

Define a type constructor (typically not a functor) encapsulating all the operations as case classes, with or without type parameters

```
sealed trait F[A]
case class Op1(a1: A1) extends F[B1]
case class Op2(a1: A2) extends F[B2]
```

Use `DSL[F, A]` with this `F` to write monadic DSL programs `prg: DSL[F, A]`

Choose a target monad `M[A]` and implement an extractor `ex: F[A] → M[A]`

Run the program with the extractor, `val res: M[A] = run(ex)(prg)`

Further directions (out of scope for this chapter):

May choose another monad `N[A]` and use interpreter `M[A] → N[A]`

E.g. transform into another monadic DSL to optimize, test, etc.

Since `DSL[F, A]` has a monad API, we can use monad transformers on it

Can combine two or more DSLs in a disjunction: $DSL^{F+G+H,A}$

Monad laws for DSL programs Monad laws hold for DSL programs only after evaluating them

Consider the law $flm(\text{pure}) = id$; both functions $DSL^{F,A} \rightarrow DSL^{F,A}$

Apply both sides to some `prg : DSLF,A` and get the new value

```
prg.flatMap(pure) == Bind(prg, a → Literal(a))
```

This new value is *not equal* to `prg`, so this monad law fails!

Other laws fail as well because operations never reduce anything

After interpreting this program into a target monad M^A , the law holds:

```
run(ex)(prg).flatMap((a → Literal(a)) andThen run(ex))
  == run(ex)(prg).flatMap(a → run(ex)(Literal(a)))
  == run(ex)(prg).flatMap(a → pure(a))
  == run(ex)(prg)
```

Here we have assumed that the laws hold for M^A

All other laws also hold after interpreting into a lawful monad M^A

The monad law violations are "not observable"

***1

¹"A function `launch()` that launches real-world missiles can run out of missiles." (A quote attributed to Simon Peyton Jones.)

Free constructions in mathematics: Example I Consider the Russian letter Π (tsè) and the Chinese word 水 (shui)

We want to *multiply* Π by 水. Multiply how?

Say, we want an associative (but noncommutative) product of them
So we want to define a *semigroup* that *contains* Π and 水 as elements
while we still know nothing about Π and 水

Consider the set of all *unevaluated expressions* such as $\Pi \cdot \text{水} \cdot \text{水} \cdot \Pi \cdot \text{水}$

Here $\Pi \cdot \text{水}$ is different from $\text{水} \cdot \Pi$ but $(a \cdot b) \cdot c = a \cdot (b \cdot c)$

All these expressions form a **free semigroup** generated by Π and 水

This is the most unrestricted semigroup that contains Π and 水

Example calculation: $(\text{水} \cdot \text{水}) \cdot (\Pi \cdot \text{水}) \cdot \Pi = \text{水} \cdot \text{水} \cdot \Pi \cdot \text{水} \cdot \Pi$

How to represent this as a data type:

Tree encoding: the full expression tree: $((\text{水}, \text{水}), (\Pi, \text{水})), \Pi$

Implement the operation $a \cdot b$ as pair constructor (easy)

Reduced encoding, as a “smart” structure: List(水, 水, Π , 水, Π)

Implement $a \cdot b$ by concatenating the lists (more expensive)

Free constructions in mathematics: Example II Want to define a product operation for n -dimensional vectors: $\mathbf{v}_1 \otimes \mathbf{v}_2$

The \otimes must be linear and distributive (but not commutative):

$$\begin{aligned} \mathbf{u}_1 \otimes \mathbf{v}_1 + (\mathbf{u}_2 \otimes \mathbf{v}_2 + \mathbf{u}_3 \otimes \mathbf{v}_3) &= (\mathbf{u}_1 \otimes \mathbf{v}_1 + \mathbf{u}_2 \otimes \mathbf{v}_2) + \mathbf{u}_3 \otimes \mathbf{v}_3 \\ \mathbf{u} \otimes (a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2) &= a_1 (\mathbf{u} \otimes \mathbf{v}_1) + a_2 (\mathbf{u} \otimes \mathbf{v}_2) \\ (a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2) \otimes \mathbf{u} &= a_1 (\mathbf{v}_1 \otimes \mathbf{u}) + a_2 (\mathbf{v}_2 \otimes \mathbf{u}) \end{aligned}$$

We have such a product for 3-dimensional vectors; ignore that

Consider *unevaluated expressions* of the form $\mathbf{u}_1 \otimes \mathbf{v}_1 + \mathbf{u}_2 \otimes \mathbf{v}_2 + \dots$

A free vector space generated by pairs of vectors

Impose the equivalence relationships shown above

The result is known as the **tensor product**

Tree encoding: full unevaluated expression tree

A list of any number of vector pairs $\sum_i \mathbf{u}_i \otimes \mathbf{v}_i$

Reduced encoding: an $n \times n$ matrix

Reduced encoding requires proofs and more complex operations

Worked example I: Free semigroup Implement a free semigroup **FSIS** generated by two types **Int** and **String**

A value of **FSIS** can be an **Int**; it can also be a **String**

If **x**, **y** are of type **FSIS** then so is **x** **|+|** **y**

```
sealed trait FSIS // tree encoding: full expression tree
case class Wrap1(x: Int) extends FSIS
case class Wrap2(x: String) extends FSIS
case class Comb(x: FSIS, y: FSIS) extends FSIS
```

Short type notation: $\text{FSIS} \equiv \text{Int} + \text{String} + \text{FSIS} \times \text{FSIS}$

For a semigroup S and given $\text{Int} \rightarrow S$ and $\text{String} \rightarrow S$, map $\text{FSIS} \rightarrow S$

Simplify and generalize this construction by setting $Z = \text{Int} + \text{String}$

The tree encoding is $\text{FSIS}^Z \equiv Z + \text{FSIS}^Z \times \text{FSIS}^Z$

```
def |+|(x: FS[Z], y: FS[Z]): FS[Z] = Comb(x, y)
def run[S: Semigroup, Z](extract: Z → S): FS[Z] → S = {
  case Wrap(z) → extract(z)
  case Comb(x, y) → run(extract)(x) |+| run(extract)(y)
}
```

run() // Semigroup laws will hold after applying **run()**.

The reduced encoding is $\text{FSR}^Z \equiv Z \times \text{List}^Z$ (non-empty list of Z 's)

x **|+|** **y** requires concatenating the lists, but **run()** is faster

Worked example II: Free monoid Implement a free monoid $\text{FM}[Z]$ generated by type Z

A value of $\text{FM}[Z]$ can be the empty value; it can also be a Z

If x, y are of type $\text{FM}[Z]$ then so is $x \mid+| y$

```
sealed trait FM[Z] // tree encoding
case class Empty[Z]() extends FM[Z]
case class Wrap[Z](z: Z) extends FM[Z]
case class Comb[Z](x: FM[Z], y: FM[Z]) extends FM[Z]
```

Short type notation: $\text{FM}^Z \equiv 1 + Z + \text{FM}^Z \times \text{FM}^Z$

For a monoid M and given $Z \rightarrow M$, map $\text{FM}^Z \rightarrow M$

```
def |+|(x: FM[Z], y: FM[Z]): FM[Z] = Comb(x, y)
def run[M: Monoid, Z](extract: Z → M): FM[Z] → M = {
  case Empty() → Monoid[M].empty
  case Wrap(z) → extract(z)
  case Comb(x, y) → run(extract)(x) |+| run(extract)(y)
}
```

// Monoid laws will hold after applying run().

The reduced encoding is $\text{FMR}^Z \equiv \text{List}^Z$ (list of Z 's)

Implementing $|+|$ requires concatenating the lists

Reduced encoding and tree encoding give identical results after `run()`

Mapping a free semigroup to different targets What if we interpret FS^X into *another* free semigroup?

Given $Y \rightarrow Z$, can we map $\text{FS}^Y \rightarrow \text{FS}^Z$?

Need to map $\text{FS}^Y \equiv Y + \text{FS}^Y \times \text{FS}^Y \rightarrow Z + \text{FS}^Z \times \text{FS}^Z$

This is straightforward since FS^X is a functor in X :

```
def fmap[Y, Z](f: Y → Z): FS[Y] → FS[Z] = {
  case Wrap(y) → Wrap(f(y))
  case Comb(a, b) → Comb(fmap(f)(a), fmap(f)(b))
}
```

Now we can use `run` to interpret $\text{FS}^X \rightarrow \text{FS}^Y \rightarrow \text{FS}^Z \rightarrow S$, etc.

Functor laws hold for FS^X , so `fmap` is composable as usual

The "interpreter" commutes with `fmap` as well (naturality law):

$$\begin{array}{ccc} & \text{fmap } f: X \rightarrow Y & \text{FS}^Y \\ & \swarrow & \searrow \text{run } g: Y \rightarrow S \\ \text{FS}^X & \xrightarrow{\quad \quad} & S \\ & \searrow \text{run } g \circ f: X \rightarrow S & \end{array}$$

Combine two free semigroups: FS^{X+Y} ; inject parts: $\text{FS}^X \rightarrow \text{FS}^{X+Y}$

Church encoding I: Motivation Multiple target semigroups S_i require many "extractors" $\text{ex}_i: Z \rightarrow S_i$

Refactor extractors ex_i into evidence of a typeclass constraint on S_i

// Typeclass `ExZ[S]` has a single method, `extract: Z → S`.

```
implicit val exZ: ExZ[MySemigroup] = { z → ... }
def run[S: ExZ : Semigroup](fs: FS[Z]): S = fs match {
  case Wrap(z) → implicitly[ExZ[S]].extract(z)
  case Comb(x, y) → run(x) |+| run(y)
}
```

`run()` replaces case classes by fixed functions parameterized by $S: \text{ExZ}$; instead we can represent

$\text{FS}[Z]$ directly by such functions, for example:

```
def wrap[S: ExZ](z: Z): S = implicitly[ExZ[S]].extract(z)
def x[S: ExZ : Semigroup]: S = wrap(1) |+| wrap(2)
```

The type of x is $\forall S. (Z \rightarrow S) \times (S \times S \rightarrow S) \rightarrow S$; an equivalent type is

$$\forall S. ((Z + S \times S) \rightarrow S) \rightarrow S$$

This is the "Church encoding" (of the free semigroup over Z)

The Church encoding is based on the theorem $A \cong \forall X. (A \rightarrow X) \rightarrow X$

this *resembles* the type of the continuation monad, $(A \rightarrow R) \rightarrow R$

but $\forall X$ makes the function fully generic, like a natural transformation

Church encoding II: Disjunction types Consider the Church encoding for the disjunction type $P + Q$

The encoding is $\forall X. (P + Q \rightarrow X) \rightarrow X \cong \forall X. (P \rightarrow X) \rightarrow (Q \rightarrow X) \rightarrow X$

```
trait Disj[P, Q] { def run[X](cp: P → X)(cq: Q → X): X }
```

Define some values of this type:

```
def left[P, Q](p: P) = new Disj[P, Q] {
  def run[X](cp: P → X)(cq: Q → X): X = cp(p)
}
```

Now we can implement the analog of the `case` expression simply as

```
val result = disj.run {p → ...} {q → ...}
```

This works in programming languages that have no disjunction types

General recipe for implementing the Church encoding:

```
trait Blah { def run[X](cont: ... → X): X }
```

For convenience, define a type class `Ex` describing the inner function:

```
trait Ex[X] { def cp: P → X; def cq: Q → X }
```

Different methods of this class return `X`; convenient with disjunctions

Church-encoded types have to be “run” for pattern-matching

Church encoding III: How it works Why is the type $\text{Ch}^A \equiv \forall X. (A \rightarrow X) \rightarrow X$ equivalent to the type A ?

```
trait Ch[A] { def run[X](cont: A → X): X }
```

- If we have a value of A , we can get a Ch^A

```
def a2c[A](a: A): Ch[A] = new Ch[A] {
  def run[X](cont: A → X): X = cont(a)
}
```

$$\begin{array}{ccc} \text{id} : (A \rightarrow A) & \xrightarrow{\text{ch.run}^A} & A \\ \downarrow \text{fmapReader}_A(f) & & \downarrow f \\ f : (A \rightarrow X) & \xrightarrow{\text{ch.run}^X} & X \end{array}$$

- If we have a $\text{ch} : \text{Ch}^A$, we can get an $a : A$

```
def c2a[A](ch: Ch[A]): A = ch.run[A](a → a)
```

The functions `a2c` and `c2a` are inverses of each other

To implement a value $\text{ch} : \text{Ch}^A$, we must compute an $x : X$ given $f : A \rightarrow X$, for *any* X , which *requires* having a value $a : A$ available

To show that $\text{ch} = \text{a2c}(\text{c2a}(\text{ch}))$, apply both sides to an $f : A \rightarrow X$ and get $\text{ch.run}(f) = \text{a2c}(\text{c2a}(\text{ch})).\text{run}(f) = f(\text{c2a}(\text{ch})) = f(\text{ch.run}(a → a))$

This is naturality of `ch.run` as a transformation between `Reader` and `Id`

Naturality of `ch.run` follows from parametricity of its code

It is straightforward to compute $\text{c2a}(\text{a2c}(a)) = \text{identity}(a) = a$

Church encoding satisfies laws: it is built up from parts of `run` method

Worked example III: Free functor I The `Functor` type class has one method, `fmap`: $(Z \rightarrow A) \rightarrow F^Z \rightarrow F^A$

The tree encoding of a free functor over F^* needs two case classes:

```
sealed trait FF[F[_], A]
case class Wrap[F[_], A](fa: F[A]) extends FF[F, A]
case class Fmap[F[_], A, Z](f: Z => A)(ffz: FF[F, Z]) extends FF[F, A]
```

The constructor `Fmap` has an extra type parameter Z , which is “hidden”

Consider a simple example of this:

```
sealed trait Q[A]; case class QZ[A, Z](a: A, z: Z) extends Q[A]
```

Need to use specific type Z when constructing a value of `Q[A]`, e.g.,

```
val q: Q[Int] = QZ[Int, String](123, "abc")
```

The type Z is hidden inside $q : Q^{\text{Int}}$; all we know is that Z “exists”

Type notation for this: $Q^A \equiv \exists Z. A \times Z$

The existential quantifier applies to the “hidden” type parameter

The constructor `QZ` has type $\exists Z. (A \times Z \rightarrow Q^A)$

It is not $\forall Z$ because a specific Z is used when building up a value

The code does not show $\exists Z$ explicitly! We need to keep track of that

Encoding with an existential type: How it works Show that $P^A \equiv \exists Z. Z \times (Z \rightarrow A) \cong A$

```
sealed trait P[A]; case class PZ[A, Z](z: Z, f: Z → A) extends P[A]
```

How to construct a value of type P^A for a given A ?

Have a function $Z \rightarrow A$ and a Z , construct $Z \times (Z \rightarrow A)$

Particular case: $Z \equiv A$, have $a : A$ and build $a \times \text{id}^{A \rightarrow A}$

```
def a2p[A](a: A): P[A] = PZ[A, A](a, identity)
```

Cannot extract Z out of P^A – the type Z is hidden

Can extract A out of P^A – do not need to know Z

```
def p2a[A]: P[A] → A = { case PZ(z, f) → f(z) }
```

Cannot transform P^A into anything else other than A

A value of type P^A is observable only via `p2a`

Therefore the functions `a2p` and `p2a` are "observational" inverses (i.e. we need to use `p2a` in order to compare values of type P^A)

If F^\bullet is a functor then $Q^A \equiv \exists Z. F^Z \times (Z \rightarrow A) \cong F^A$

A value of Q^A can be observed only by extracting an F^A from it

Can define `f2q` and `q2f` and show that they are observational inverses

Worked example III: Free functor II Tree encoding of `FF` has type $\text{FF}^{F^\bullet, A} \equiv F^A + \exists Z. \text{FF}^{F^\bullet, Z} \times (Z \rightarrow A)$

Derivation of the reduced encoding:

A value of type $\text{FF}^{F^\bullet, A}$ must be of the form

$$\exists Z_1. \exists Z_2. \dots \exists Z_n. F^{Z_n} \times (Z_n \rightarrow Z_{n-1}) \times \dots \times (Z_2 \rightarrow Z_1) \times (Z_1 \rightarrow A)$$

The functions $Z_1 \rightarrow A$, $Z_2 \rightarrow Z_1$, etc., must be composed associatively

The equivalent type is $\exists Z_n. F^{Z_n} \times (Z_n \rightarrow A)$

Reduced encoding: $\text{FreeF}^{F^\bullet, A} \equiv \exists Z. F^Z \times (Z \rightarrow A)$

Substituted F^Z instead of $\text{FreeF}^{F^\bullet, Z}$ and eliminated the case F^A

The reduced encoding is non-recursive

Requires a proof that this encoding is equivalent to the tree encoding

If F^\bullet is already a functor, can show $F^A \cong \exists Z. F^Z \times (Z \rightarrow A)$

Church encoding (starting from the tree encoding): $\text{FreeF}^{F^\bullet, A} \equiv \forall P^\bullet. (\forall C. (F^C + \exists Z. P^Z \times (Z \rightarrow C)) \rightsquigarrow P^C) \rightarrow P^A$

The structure of the type expression: $\forall P^\bullet. (\forall C. (\dots)^C \rightsquigarrow P^C) \rightarrow P^A$

Cannot move $\forall C$ or $\exists Z$ to the outside of the type expression!

Church encoding IV: Recursive types and type constructors Consider the recursive type $P \equiv Z + P \times P$ (tree with Z -valued leaves)

The Church encoding is $\forall X. ((Z + X \times X) \rightarrow X) \rightarrow X$

This is *non-recursive*: the inductive use of P is replaced by X

Generalize to recursive type $P \equiv S^P$ where S^\bullet is a "induction functor":

The Church encoding of P is $\forall X. (S^X \rightarrow X) \rightarrow X$

Church encoding of recursive types is non-recursive

Example: Church encoding of `List[Int]`

Church encoding of a type constructor P^\bullet :

Notation: P^\bullet is a type function; Scala syntax is `P[_]`

The Church encoding is $\text{Ch}^{P^\bullet, A} = \forall F^\bullet. (\forall X. P^X \rightarrow F^X) \rightarrow F^A$

Note: $\forall X. P^X \rightarrow F^X$ or $P^\bullet \rightsquigarrow F^\bullet$ resembles a natural transformation

Except that P^\bullet and F^\bullet are not necessarily functors, so no naturality law

Example: Church encoding of `Option[_]`

Church encoding of a *recursively* defined type constructor P^\bullet :

Definition: $P^A \equiv S^{P^\bullet, A}$ where $S^{P^\bullet, A}$ describes the "induction principle"

Notation: S^\bullet, A is a higher-order type function; Scala syntax: `S[_[_], A]`

Example: $\text{List}^A \equiv 1 + A \times \text{List}^A \equiv S^{\text{List}^\bullet, A}$ where $S^{\text{List}^\bullet, A} \equiv 1 + A \times P^A$

The Church encoding of P^A is $\text{Ch}^{P^\bullet, A} = \forall F^\bullet. (S^{F^\bullet} \rightsquigarrow F^\bullet) \rightarrow F^A$

The Church encoding of `List[_]` is non-recursive

Details: Why Church encoding of a free semigroup is a semigroup - it's not obvious

$FS = \text{forall } S. (Z \Rightarrow S) \times (S \times S \Rightarrow S) \Rightarrow S$ is a semigroup. We need to define the binary operation $\text{I} + \text{I}$ on values of type FS . A value f of type FS is a function with a type parameter, that we can use as $f[S](e, c)$ to compute a value of any given type S from arguments $e : Z \Rightarrow S$ and $c : S \times S \Rightarrow S$. Scala code for f will be

```
def f[S](empty: Z => S, combine: (S, S) => S): S = ???
```

So, given f and g of this type, we need to somehow define a new function $h = f \text{ I} + \text{I} g$ also of the same type. Begin to write code for that function:

```
def h[S](empty: Z => S, combine: (S, S) => S): S = ???
```

The free semigroup in the tree encoding defines the binary operation as a formal operation that does not compute anything. In the Church encoding, however, we have the binary operation as the argument "combine" of h , and so we should call that function. So we use it:

```
def h[S](empty: Z => S, combine: (S, S) => S): S = combine(???, ???)
```

We need to fill the typed holes $???$ of type S . It is clear that we should use f and g somehow. We can use f and g simply by calling those functions on the arguments 'empty' and 'combine'. Since f and g both have a universally quantified type parameter, we can just use the given type S for them.

```
def h[S](empty: Z => S, combine: (S, S) => S): S = combine(f[S](empty, combine), g[S](empty, combine))
```

The types match, and we have used both functions f and g in a way that is intuitively correct. We have preserved information. So, this is likely the correct implementation. It remains to verify the associativity law. To do that, we need to assume that 'combine' is associative for the actual type S on which we use the Church encoding (i.e. a non-free, lawful semigroup)

To show equivalence between FSZ and ChZ , write code for the type ChZ and the two directions of the isomorphism,

```
trait FSCh[Z] { def run[S](empty: Z => S, combine: (S, S) => S): S }
def fsz2ch[Z](fsz: NEList[Z]): FSCh[Z] = ???
def ch2fsz[Z](ch: FSCh[Z]): NEList[Z] = ???
```

Church encoding V: Type classes

Look at the Church encoding of the free semigroup:

$$ChFS^Z \equiv \forall X. (Z \rightarrow X) \times (X \times X \rightarrow X) \rightarrow X$$

If X is constrained to the [Semigroup](#) typeclass, we will already have a value $X \times X \rightarrow X$, so we can omit it: $ChFS^Z = \forall X: \text{Semigroup}. (Z \rightarrow X) \rightarrow X$

The "induction functor" for "semigroup over Z " is $\text{SemiG}^X \equiv Z + X \times X$

So the Church encoding is $\forall X. (\text{SemiG}^X \rightarrow X) \rightarrow X$

Generalize to arbitrary type classes:

Type class C is defined by its operations $C^X \rightarrow X$ (with a suitable C^\bullet)

call C^\bullet the **method functor** of the inductive typeclass C

Tree encoding of "free C over Z " is recursive, $\text{FreeC}^Z \equiv Z + C^{\text{FreeC}^Z}$

Church encoding is $\text{FreeC}^Z \equiv \forall X. (Z + C^X \rightarrow X) \rightarrow X$

Equivalently, $\text{FreeC}^Z \equiv \forall X: C. (Z \rightarrow X) \rightarrow X$

Laws of the typeclass are satisfied automatically after "running"

Works similarly for type constructors: operations $C^{P^\bullet, A} \rightarrow P^A$

Free typeclass C over F^\bullet is $\text{FreeC}^{F^\bullet, A} \equiv \forall P^\bullet: C. (F^\bullet \rightsquigarrow P^\bullet) \rightarrow P^A$

Properties of free type constructions

Generalizing from our examples so far:

We "enriched" Z to a monoid FM^Z , and F^A to a monad $DSL^{F, A}$

The "enrichment" adds case classes representing the needed operations

Works for a generating type Z and for a generating type constructor F^A

Obtain a **free type construction**, which performs no computations

FM^Z wraps Z in "just enough" stuff to make it look like a monoid

$FreeF^{F^\bullet, A}$ wraps F^A in "just enough" stuff to make it look like a functor

A value of a free construction can be "run" to yield non-free values

Questions:

Can we construct a free typeclass C over any type constructor F^A ?

Yes, with typeclasses: (contra)functor, filterable, monad, applicative

Which of the possible encodings to use?

Tree encoding, reduced encodings, Church encoding

What are the laws for the $\text{FreeC}^{F,A}$ – "free instance of C over F "?

For all F^\bullet , must have $\text{wrap}[A] : F^A \rightarrow \text{FreeC}^{F,A}$ or $F^\bullet \rightsquigarrow \text{FreeC}^{F^\bullet}$

For all $M^\bullet : C$, must have $\text{run} : (F^\bullet \rightsquigarrow M^\bullet) \rightarrow \text{FreeC}^{F^\bullet} \rightsquigarrow M^\bullet$

The laws of typeclass C must hold after interpreting into an $M^\bullet : C$

Given any $t : F^\bullet \rightsquigarrow G^\bullet$, must have $\text{fmap}(t) : \text{FreeC}^{F^\bullet} \rightsquigarrow \text{FreeC}^{G^\bullet}$

Recipes for encoding free typeclass instances Build a free instance of typeclass C over F^\bullet , as a type constructor P^\bullet

The typeclass C can be functor, contrafunctor, monad, etc.

Assume that C has methods m_1, m_2, \dots , with type signatures $m_1 : Q_1^{P^\bullet, A} \rightarrow P^A$, $m_2 : Q_2^{P^\bullet, A} \rightarrow P^A$, etc., where $Q_i^{P^\bullet, A}$ are covariant in P^\bullet

Inductive typeclass is defined via a methods functor, $S^{P^\bullet} \rightsquigarrow P^\bullet$

The tree encoded FC^A is a disjunction defined recursively by

$$\text{FC}^A \equiv F^A + Q_1^{\text{FC}^\bullet, A} + Q_2^{\text{FC}^\bullet, A} + \dots$$

```
sealed trait FC[A]; case class Wrap[A](fa: F[A]) extends FC[A]
case class Q1[A](...) extends FC[A]
case class Q2[A](...) extends FC[A]; ...
```

Any type parameters within Q_i are then existentially quantified

$\text{run}()$ maps $F^\bullet \rightsquigarrow M^\bullet$ in the disjunction and recursively for other parts

Derive a reduced encoding via reasoning about possible values of FC^A and by taking into account the laws of the typeclass C

A Church encoding can use the tree encoding or the reduced encoding

Church encoding is "automatically reduced", but performance may differ

Properties of inductive typeclasses If a typeclass C is inductive with methods $C^X \rightarrow X$ then:

A free instance of C over Z can be tree-encoded as $\text{FreeC}^Z \equiv Z + C^{\text{FreeC}^Z}$

All inductive typeclasses have free instances, FreeC^Z

If $P:C$ and $Q:C$ then $P \times Q$ and $Z \rightarrow P$ also belong to typeclass C

but not necessarily $P + Q$ or $Z \times P$

Proof: can implement $(C^P \rightarrow P) \times (C^Q \rightarrow Q) \rightarrow C^{P \times Q} \rightarrow P \times Q$ and $(C^P \rightarrow P) \rightarrow C^{Z \rightarrow P} \rightarrow Z \rightarrow P$, but cannot implement $(...) \rightarrow P + Q$

Analogous properties hold for type constructor typeclasses

Methods described as $C^{F^\bullet, A} \rightarrow F^A$ with type constructor parameter F^\bullet

What typeclasses *cannot* be tree-encoded (or have no "free" instances)?

Any typeclass with a method *not ultimately returning* a value of P^A

Example: a typeclass with methods $\text{pt} : A \rightarrow P^A$ and $\text{ex} : P^A \rightarrow A$

Such typeclasses are not inductive

Typeclasses with methods of the form $P^A \rightarrow \dots$ are **co-inductive**

Worked example IV: Free contrafunctor Method contramap : $C^A \times (B \rightarrow A) \rightarrow C^B$

Tree encoding: $\text{FreeCF}^{F^\bullet, B} \equiv F^B + \exists A. \text{FreeCF}^{F^\bullet, A} \times (B \rightarrow A)$

Reduced encoding: $\text{FreeCF}^{F^\bullet, B} \equiv \exists A. F^A \times (B \rightarrow A)$

A value of type $\text{FreeCF}^{F^\bullet, B}$ must be of the form

$$\exists Z_1. \exists Z_2. \dots \exists Z_n. F^{Z_1} \times (B \rightarrow Z_n) \times (Z_n \rightarrow Z_{n-1}) \times \dots \times (Z_2 \rightarrow Z_1)$$

The functions $B \rightarrow Z_n$, $Z_n \rightarrow Z_{n-1}$, etc., are composed associatively

The equivalent type is $\exists Z_1. F^{Z_1} \times (B \rightarrow Z_1)$

The reduced encoding is non-recursive

Example: $F^A \equiv A$, "interpret" into the contrafunctor $C^A \equiv A \rightarrow \text{String}$

```
def prefixLog[A](p: A): A → String = a → p.toString + a.toString
```

If F^\bullet is already a contrafunctor then $\text{FreeCF}^{F^\bullet, A} \cong F^A$

Worked example V: Free pointed functor Over an arbitrary type constructor F^\bullet :Pointed functor methods $\text{pt} : A \rightarrow P^A$ and $\text{map} : P^A \times (A \rightarrow B) \rightarrow P^B$ Tree encoding: $\text{FreeP}^{F^\bullet, A} \equiv A + F^A + \exists Z. \text{FreeP}^{F^\bullet, Z} \times (Z \rightarrow A)$

Derivation of the reduced encoding:

The tree encoding of a value $\text{FreeP}^{F^\bullet, A}$ is either

$$\exists Z_1. \exists Z_2. \dots \exists Z_n. F^{Z_n} \times (Z_n \rightarrow Z_{n-1}) \times \dots \times (Z_2 \rightarrow Z_1) \times (Z_1 \rightarrow A)$$

or

$$\exists Z_1. \exists Z_2. \dots \exists Z_n. Z_n \times (Z_n \rightarrow Z_{n-1}) \times \dots \times (Z_2 \rightarrow Z_1) \times (Z_1 \rightarrow A)$$

Compose all functions by associativity; one function $Z_n \rightarrow A$ remainsThe case $\exists Z_n. Z_n \times (Z_n \rightarrow A)$ is equivalent to just A Reduced encoding: $\text{FreeP}^{F^\bullet, A} \equiv A + \exists Z. F^Z \times (Z \rightarrow A)$, non-recursiveThis reuses the free functor as $\text{FreeP}^{F^\bullet, A} = A + \text{FreeF}^{F^\bullet, A}$ If the type constructor F^\bullet is already a functor, $\text{FreeF}^{F^\bullet, A} \cong F^A$ and so:Free pointed functor over a functor F^\bullet is simplified: $A + F^A$ If F^\bullet is already a pointed functor, need not use the free constructionIf we do, we will have $\text{FreeP}^{F^\bullet, A} \not\cong F^A$

only functors and contrafunctors do not change under “free”

Worked example VI: Free filterable functor (See Chapter 6.) Methods:

$$\text{map} : F^A \rightarrow (A \rightarrow B) \rightarrow F^B$$

$$\text{mapOpt} : F^A \rightarrow (A \rightarrow 1 + B) \rightarrow F^B$$

We can recover `map` from `mapOpt`, so we keep only `mapOpt`Tree encoding: $\text{FreeFi}^{F^\bullet, A} \equiv F^A + \exists Z. \text{FreeFi}^{F^\bullet, Z} \times (Z \rightarrow 1 + A)$ If F^\bullet is already a functor, can simplify the tree encoding using the identity $\exists Z. P^Z \times (Z \rightarrow 1 + A) \cong P^A$ and obtain $\text{FreeFi}^{F^\bullet, A} \equiv F^A + \text{FreeFi}^{F^\bullet, 1+A}$, which is equivalent to $\text{FreeFi}^{F^\bullet, A} = F^A + F^{1+A} + F^{1+1+A} + \dots$ Reduced encoding: $\text{FreeFi}^{F^\bullet, A} \equiv \exists Z. F^Z \times (Z \rightarrow 1 + A)$, non-recursiveDerivation: $\exists Z_1. \dots \exists Z_n. F^{Z_n} \times (Z_n \rightarrow 1 + Z_{n-1}) \times \dots \times (Z_1 \rightarrow 1 + A)$ is simplified using the laws of `mapOpt` and Kleisli composition, and yields $\exists Z_n. F^{Z_n} \times (Z_n \rightarrow 1 + A)$. Encode F^A as $\exists Z. F^Z \times (Z \rightarrow 0 + Z)$.If F^\bullet is already a functor, the reduced encoding is $\text{FreeFi}^{F^\bullet, A} = F^{1+A}$ Free filterable over a filterable functor F^\bullet is not equivalent to F^\bullet

Free filterable contrafunctor is constructed in a similar way

Worked example VII: Free monad Methods:

$$\text{pure} : A \rightarrow F^A$$

$$\text{flatMap} : F^A \rightarrow (A \rightarrow F^B) \rightarrow F^B$$

Can recover `map` from `flatMap` and `pure`, so we keep only `flatMap`Tree encoding: $\text{FreeM}^{F^\bullet, A} \equiv F^A + A + \exists Z. \text{FreeM}^{F^\bullet, Z} \times (Z \rightarrow \text{FreeM}^{F^\bullet, A})$

Derivation of reduced encoding:

can simplify $A \times (A \rightarrow \text{FreeM}^{F^\bullet, B}) \cong \text{FreeM}^{F^\bullet, B}$ use associativity to replace $\text{FreeM}^A \times (A \rightarrow \text{FreeM}^B) \times (B \rightarrow \text{FreeM}^C)$ by $\text{FreeM}^A \times (A \rightarrow \text{FreeM}^B \times (B \rightarrow \text{FreeM}^C))$ therefore we can replace $\exists Z. \text{FreeM}^{F^\bullet, Z} \times \dots$ by $\exists Z. F^Z \times \dots$ Reduced encoding: $\text{FreeM}^{F^\bullet, A} \equiv A + \exists Z. F^Z \times (Z \rightarrow \text{FreeM}^{F^\bullet, A})$ “Final Tagless style” means “Church encoding of free monad over F^\bullet ”Free monad over a functor F^\bullet is $\text{FreeM}^{F^\bullet, A} \equiv A + F^{\text{FreeM}^{F^\bullet, A}}$ Free monad $\text{FreeM}^{M^\bullet, A}$ over a monad M^\bullet is not equivalent to M^\bullet Free monad over a pointed functor F^\bullet is $\text{FreeM}^{F^\bullet, A} \equiv F^A + F^{\text{FreeM}^{F^\bullet, A}}$ start from half-reduced encoding $F^A + \exists Z. F^Z \times (Z \rightarrow \text{FreeM}^{F^\bullet, A})$ replace the existential type by an equivalent type $F^{\text{FreeM}^{F^\bullet, A}}$

Worked example VIII: Free applicative functor Methods:

$$\begin{aligned} \text{pure} &: A \rightarrow F^A \\ \text{ap} &: F^A \rightarrow F^{A \rightarrow B} \rightarrow F^B \end{aligned}$$

We can recover `map` from `ap` and `pure`, so we omit `map`

Tree encoding: $\text{FreeAp}^{F^\bullet, A} \equiv F^A + A + \exists Z. \text{FreeAp}^{F^\bullet, Z} \times \text{FreeAp}^{F^\bullet, Z \rightarrow A}$

Reduced encoding: $\text{FreeAp}^{F^\bullet, A} \equiv A + \exists Z. F^Z \times \text{FreeAp}^{F^\bullet, Z \rightarrow A}$

Derivation: a FreeAp^A is either $\exists Z_1 \dots \exists Z_n. Z_1 \times \text{FreeAp}^{Z_1 \rightarrow Z_2} \times \dots$ or $\exists Z_1 \dots \exists Z_n. F^{Z_1} \times \text{FreeAp}^{Z_1 \rightarrow Z_2} \times \dots$; encode $Z_1 \times \text{FreeAp}^{Z_1 \rightarrow Z_2}$ equivalently as $\text{FreeAp}^{Z_1 \rightarrow Z_2} \times ((Z_1 \rightarrow Z_2) \rightarrow Z_2)$ using the identity law; so the first FreeAp^Z is always F^A , or we have a pure value

Free applicative over a functor F^\bullet :

$$\begin{aligned} \text{FreeAp}^{F^\bullet, A} &\equiv A + \text{FreeZ}^{F^\bullet, A} \\ \text{FreeZ}^{F^\bullet, A} &\equiv F^A + \exists Z. F^Z \times \text{FreeZ}^{F^\bullet, Z \rightarrow A} \end{aligned}$$

$\text{FreeZ}^{F^\bullet, \bullet}$ is the reduced encoding of "free zippable" (no `pure`)

$\text{FreeAp}^{F^\bullet, \bullet}$ over an applicative functor F^\bullet is not equivalent to F^\bullet

Laws for free typeclass constructions Consider an inductive typeclass C with methods $C^A \rightarrow A$

Define a free instance of C over Z recursively, $\text{FreeC}^Z \equiv Z + C^{\text{FreeC}^Z}$

FreeC^Z has an instance of C , i.e. we can implement $C^{\text{FreeC}^Z} \rightarrow \text{FreeC}^Z$

FreeC^Z is a P^C over Z , implement the function $\text{run}^C : \text{FreeC}^Y \rightarrow \text{FreeC}^Z$

$$\begin{aligned} \text{run}^P &: (Z \rightarrow P) \rightarrow \text{FreeC}^Z \rightarrow P \\ \text{wrap} &: Z \rightarrow \text{FreeC}^Z \end{aligned}$$

$$\begin{array}{ccc} & \text{FreeC}^Y & \\ & \downarrow \text{fmap } f: Y \rightarrow Z & \\ \text{FreeC}^Z & \xrightarrow{\text{run}(f \circ g)} & P \\ & \text{run}(g: Z \rightarrow P) \searrow & \end{array}$$

Law 1: $\text{run}(\text{wrap}) = \text{id}$; law 2: $\text{fmap } f \circ \text{run } g = \text{run}(f \circ g)$ (naturality of `run`)

For any $P^C, Q^C, g: Z \rightarrow P$, and a typeclass-preserving $f: P \rightarrow Q$, we have

$$\text{run}^P(g \circ f) = \text{run}^Q(g \circ f) \quad - \text{"universal property" of run}$$

$$\begin{array}{ccc} \text{FreeC}^Z & & \\ \downarrow \text{run}^P(g: Z \rightarrow P) & \searrow \text{run}^Q(g \circ f) & \\ P & \xrightarrow{f: P \rightarrow Q} & Q \end{array}$$

$$\begin{array}{ccc} C^P & \xrightarrow{\text{ops}_P} & P \\ \downarrow \text{fmap}_S f & & \downarrow f \\ C^Q & \xrightarrow{\text{ops}_Q} & Q \end{array}$$

$f: P \rightarrow Q$ preserves typeclass C if the diagram on the right commutes

Combining the generating constructors in a free typeclass Consider FreeC^Z for an inductive typeclass C with methods $C^X \rightarrow X$

We would like to combine generating constructors Z_1, Z_2 , etc.

In a monadic DSL – combine different operations defined separately

Note: monads do not compose in general

To combine generators, use $\text{FreeC}^{Z_1 + Z_2}$; an "instance over Z_1 and Z_2 "

but need to inject parts into disjunction, which is cumbersome

Church encoding makes this easier to manage:

$\text{FreeC}^Z \equiv \forall X. (Z \rightarrow X) \times (C^X \rightarrow X) \rightarrow X$ and then

$$\text{FreeC}^{Z_1 + Z_2} \equiv \forall X. (Z_1 \rightarrow X) \times (Z_2 \rightarrow X) \times (C^X \rightarrow X) \rightarrow X$$

Encode the functions $Z_i \rightarrow X$ via typeclasses `ExZ1`, `ExZ2`, etc., where typeclass `ExZ1` has method $Z_1 \rightarrow X$, etc.

Then

$$\text{FreeC}^{Z_1 + Z_2} = \forall X^{\text{E}Z_1, \text{E}Z_2}. (C^X \rightarrow X) \rightarrow X$$

or equivalently

$$\text{FreeC}^{Z_1 + Z_2} = \forall X^C. \forall X^{\text{E}Z_1, \text{E}Z_2}. X$$

The code is easier to maintain

This works for all typeclasses C and any number of generators Z_i

Combining different free typeclasses To combine free instances of different typeclasses C_1 and C_2 :

Option 1: use functor composition, $\text{FreeC}_{12}^Z \equiv \text{FreeC}_1^{\text{FreeC}_2^Z}$

Order of composition matters!

Operations of C_2 need to be lifted into C_1

Works only for inductive typeclasses

Encodes $C_1^{C_2}$ but not $C_2^{C_1}$

Option 2: use disjunction of method functors, $C^X \equiv C_1^X + C_2^X$, and build the free typeclass instance using C^X

Church encoding: $\text{FreeC}_{12}^Z \equiv \forall X. (Z \rightarrow X) \times (C_1^X + C_2^X \rightarrow X) \rightarrow X$

Example 1: C_1 is functor, C_2 is contrafunctor

Interpret a free functor/contrafunctor into a profunctor

Example 2: C_1 is monad, C_2 is applicative functor

Interpret into a monad that has a non-standard `zip` implementation

Example: interpret into `Future` and convert `zip` into parallel execution

Each `zip` creates parallel branch, each `flatMap` creates sequential chain

13.1.1 Exercises

Implement a free semigroup generated by a type Z in the tree encoding and in the reduced encoding. Show that the semigroup laws hold for the reduced encoding but not for the tree encoding before interpreting into a lawful semigroup S .

Type P is of typeclass Mod_L (called “ L -module”) if a fixed monoid L “acts” on P via act: $L \rightarrow P \rightarrow P$, with laws $\text{act } x \circ \text{act } y = \text{act } (x \circ y)$ and $\text{act } (1^L) = \text{id}$. Show that Mod_L is an inductive typeclass. Implement a free L -module over a type Z .

Implement a monadic DSL with operations put: $A \rightarrow 1$ and get: A ; run examples.

Implement the Church encoding of the type constructor $P^A \equiv \text{Int} + A \times A$. For the resulting type constructor, implement a Functor instance.

Describe the monoid type class via a method functor C^\bullet (such that the monoid’s operations are combined into the type $S^M \rightarrow M$). Using S^\bullet , implement the free monoid over a type Z in the Church encoding.

Assuming that F^\bullet is a functor, define $Q^A \equiv \exists Z. F^Z \times (Z \rightarrow A)$ and implement f2q: $F^A \rightarrow Q^A$ and q2f: $Q^A \rightarrow F^A$. Show that these functions are natural transformations, and that they are inverses of each other “observationally”, i.e. after applying q2f in order to compare values of Q^A .

Show: $\forall X. X = 0; \exists Z. Z \cong 1; \exists Z. Z \times A \cong A; \forall A. (A \times A \times A \rightarrow A) \cong 1 + 1 + 1$.

Derive a reduced encoding for a free applicative functor over a pointed functor.

Implement a “free pointed filterable” typeclass (combining pointed and filterable) over a type constructor F^\bullet in the tree encoding. Derive a reduced encoding. Simplify these encodings when F^\bullet is already a functor.

Corrections The slides say that the “universal property” of the runner is $\text{run}^P g \circ f = \text{run}^Q (g \circ f)$, however, this is not true; it is the right naturality property of $\text{run}^P : (Z \rightarrow P) \rightarrow \text{FreeC}^Z \rightarrow P$ with respect to the type parameter P . The universal property is $f = \text{wrap} \circ \text{run}^P f$ for any $f : Z \rightarrow P$ and any type P that belongs to the typeclass C .

The “logarithm” $\text{Lg}(F^\bullet) \triangleq \forall A. F^A \rightarrow A$ is an operation with bizarre properties. Examples: $\forall A. (Z \rightarrow A) \rightarrow A \cong Z$. $\text{Lg}(F^\bullet + G^\bullet) = \text{Lg}(F) \times \text{Lg}(G)$. $\forall A. (Z \times A \times A) \rightarrow A \cong Z \times 2$. $\text{Lg}(\text{Opt}) = \emptyset$.

13.2 Discussion

14 Computations in functor blocks. III. Monad transformers

14.1 Practical use

14.1.1 Combining monadic effects via functor composition

Monads describe effects that depend on previously computed values. It is often necessary to combine effects of several monads in one value. For example, a value of type `Future` can be computed concurrently, while `Option` represents a possibly missing value. To describe a possibly missing value that is also computed concurrently, we need somehow to combine `Future` with `Option`.

The type constructors `Future` and `Option` are not subtypes of each other, so we cannot simply use them together in a functor block:

```
for {
  x <- Future(1)
  y <- Option(x + 1) // Type error.
  ...
}
```

One way of combining two effects is by using functor composition, `Future[Option[A]]`. However, we run into a problem if we want to chain a computation to the value within the nested type. For instance, this code will not compile:

```
for {
  xOpt <- Future(Option(1))      // xOpt: Option[Int]
  x    <- xOpt    // We would like to get x: Int here.
  ...
}
```

The source type must be consistent within the entire functor block; we cannot mix source lines of `Future` and `Option` types. If we try working directly with the nested data type, we will have to write code like this:

```
val result: Future[Option[Int]] = for {
  xOpt <- Future(Option(1)) // xOpt: Option[Int]
  tOpt <- xOpt match {      // yOpt: Option[Int]
    case None    => Future.successful(None) // This expression must be a Future[Option[Int]].
    case Some(x) => for {
      yOpt <- Future(Option(x + 1))
      zOpt <- yOpt match { // zOpt: Option[Int]
        case None    => Future.successful(None) // Must return a Future[Option[Int]] everywhere.
        case Some(y) => Future(Option(y + 3))
      }
    } yield zOpt
  } yield tOpt
}
```

The nested type constructor forces us to use pattern matching with nested functor blocks, since that is the only way of getting access to values of type `A` within `Future[Option[A]]`. The code is repetitive and deeply nested, which makes it hard to read and to change.

The first step towards solving the problem is by rewriting this monadic program as a direct chain of computations depending on the results of previous ones:

```
val result: Future[Option[Int]] = for { // This will not compile!
  x <- Future(Option(1))           // The type is x: Option[Int], although we need x: Int.
```

```

y <- Future(Option(x + 1))           // Type error: x + 1 is undefined.
z <- Future(Option(y + 3))
} yield z

```

To be able to write code in this way, we need to implement the `flatMap` operation directly on the type `Future[Option[A]]` with respect to the type parameter `A`. Let us wrap the type `Future[Option[A]]` into a class and implement `map` and `flatMap` as methods:

```

final case class FutOpt[A](nested: Future[Option[A]]) {
  def map[B](f: A => B)(implicit ec: ExecutionContext): FutOpt[B] = FutOpt(nested.map(_.map(f)))
  def flatMap[B](f: A => Future[B])(implicit ec: ExecutionContext): Future[B] =
    FutOpt(nested.flatMap {           // Provide a function Option[A] => Future[Option[B]].
      case None    => Future.successful(None) // Must return a Future[Option[B]] here.
      case Some(x) => f(x).nested
    })
}

import scala.concurrent.ExecutionContext.Implicits.global
val result: FutOpt[Int] = for {           // This code compiles now.
  x <- FutOpt(Future(Option(1)))        // x: Int
  y <- FutOpt(Future(Option(x + 1)))    // y: Int
  z <- FutOpt(Future(Option(y + 3)))    // z: Int
} yield z

```

To improve usability, we may define conversions from `Future` to `FutOpt` and from `Option` to `FutOpt`:

```

implicit class FutOptLift1[A](fut: Future[A])(implicit ec: ExecutionContext) {
  def lift: FutOpt[A] = FutOpt(fut.map(x => Some(x)))
}
implicit class FutOptLift2[A](opt: Option[A]) {
  def lift: FutOpt[A] = FutOpt(Future.successful(opt))
}

```

After defining the “lifts”, we may rewrite the previous code in a more readable form:

```

val result: FutOpt[Int] = for {
  x <- Option(1).lift           // x: Int
  y <- Future(x + 1).lift       // y: Int
  z <- Option(y + 3).lift       // z: Int
} yield z // After some time, this evaluates to 'FutOpt(Future(Success(Some(5))))'.

```

We may hope to apply the same technique when mixing effects from arbitrary monads, such as:

```

val result: SomeBigMonad[Int] = for { // After some definitions, this code will work:
  x <- (1 to n).lift             // x: Int
  y <- Future { computation(x) }.lift // y: Int
  z <- Try { maybeError(y) }.lift    // z: Int
} yield z

```

This code will work if we somehow create a new monad that “unifies” `Seq`, `Future`, and `Try`.

It turns out that there is no simple recipe for implementing a “unified” monad for a given set of monads. Developing the necessary techniques is the main focus of this chapter.

We have seen that `Future[A]` and `Option[A]` are combined in a new monad as `Future[Option[A]]`. Does functor composition always work? If M_1 and M_2 are monads then at least one of the compositions, $M_1 \circ M_2$ or $M_2 \circ M_1$, is often a monad. However, `Option[Future[A]]` is not a monad. By trial and error, we may find further examples where the functor composition works and where it does not (see Table 14.1, where the incorrect examples are marked with a strike-through line).

It turns out that some monads do not compose in either order, while others have a “preferred” order (we will see proofs of these properties later in this chapter). For instance, the `Reader` monad $Z \rightarrow A$ always composes on the outside (Statement 10.2.8.5) but not necessarily on the inside. On the other hand, pass/fail monads $M^A \triangleq Z + A$ composes when it is on the inside.

When the monad composition works both ways, the two resulting monads are usually not equivalent: the two effects are combined in quite different ways.

For example, the `Writer` monad $W \times A$ (appending an output message) and the pass/fail monads $Z + A$ (allowing failure) can be composed in two ways. The monad $M^A \triangleq W \times (Z + A)$ describes a computation that may fail but always outputs a message; but the monad $M^A \triangleq Z + W \times A$ will not output any message on failure.

Monad M_1^A	Monad M_2^A	$M_1 \circ M_2$	$M_2 \circ M_1$
<code>Future</code> A	$Z + A$	Future^{Z+A}	$Z + \text{Future}^A$
<code>Future[A]</code>	<code>State[S, A]</code>	<code>Future[State[S, A]]</code>	<code>State[S, Future[A]]</code>
<code>State[S, A]</code>	$Z \rightarrow A$	<code>State[S, Z \rightarrow A]</code>	$Z \Rightarrow \text{State}[S, A]$
<code>State[S, A]</code>	$Z + A$	<code>State[S, Either[Z, A]]</code>	<code>Either[Z, State[S, A]]</code>
<code>List</code> A	$Z \rightarrow A$	<code>List^{Z \rightarrow A}</code>	$Z \rightarrow \text{List}^A$
<code>List</code> A	$Z + A$	List^{Z+A}	$Z + \text{List}^A$
$\mathbb{1} + A$	$Z \rightarrow A$	$\mathbb{1} + (Z \rightarrow A)$	$Z \rightarrow \mathbb{1} + A$
$W \times A$	$Z + A$	$W \times (Z + A)$	$Z + W \times A$

Table 14.1: Correct and incorrect compositions of monads.

We will see throughout this chapter that, even when monads do not compose, effects can still be combined in different orders. The programmer needs to choose carefully the desired order of composition, according to the desired meaning of the effects.

When functor composition works, implementing the `lift` methods is straightforward:

Example 14.1.1.1 Implement a composition of `List` and `Try` monads for use in functor blocks.

Solution Since `Try` is a pass/fail monad, we must compose it *inside* other monads; so we define the “unified” monad type as `List[Try[A]]`. We wrap the type `List[Try[A]]` into a class, implement `map` and `flatMap` as methods, and then define the two `lift` functions as extension methods:

```
final case class ListTry[A](value: List[Try[A]]) {
  def map[B](f: A => B): ListTry[B] = ListTry(value.map(_.map(f)))
  def flatMap[B](f: A => ListTry[B]): ListTry[B] = ListTry(
    value.flatMap {           // Provide a function Try[A] => List[Try[B]] here.
      case Failure(t)    => List(Failure(t))
      case Success(a)   => f(a).value
    }
  )
}                                // Rename 'lift' to 'up' because 'lift' is already defined for 'List' class.
implicit class ListTryLift1[A](l: List[A]) { def up: ListTry[A] = ListTry(l.map(Success(_))) }
implicit class ListTryLift2[A](t: Try[A]) { def up: ListTry[A] = ListTry(List(t)) }
```

After these definitions, we can write functor blocks that use `List` and `Try` as needed:

```
scala> for {
  x <- List(1, 2, 3).up
  y <- Try(x + 100).up
  } yield y
res0: ListTry[Int] = ListTry(List(Success(101), Success(102), Success(103)))
```

14.1.2 Combining monads via monad transformers

Not all monads combine with others via functor composition. But even for those that do, the method used in Example 14.1.1.1 requires writing custom wrapper classes for each *pair* of monads. The length of the required code would be quadratic in the number of supported monads. To avoid the code explosion, we will use a different approach to organizing the code that combines monads.

The main idea of that approach is to fix the first monad L (the “base” monad) and to write code that combines L with an arbitrary second monad M (the “foreign” monad). The result is a new, “transformed” monad, denoted by T_L^M . It turns out that the code of T_L^M is fully parametric in M . In other words, the code of the `pure` and `flatMap` methods for T_L^M will work in the same way for all foreign monads M and will only depend on M ’s monadic methods (pu_M and flm_M). The monad T_L^M is called the “ L -transformer applied to M ” or the “transformed monad M ”. The **monad transformer** of the monad L is the type constructor T_L^M viewed as a function of its type parameters M and A .

To illustrate this approach, let us implement the transformer for the `Try` monad. We begin with Example 14.1.1.1 that combines `Try` with `List`, but we now view `Try` as the base monad and `List` as the foreign monad. The goal is to replace `List` by an arbitrary foreign monad M , so that the resulting code works in the same way for all monads M . To achieve that, we annotate the code of Example 14.1.1.1 to make explicit the usage of the `List` monad’s methods `pure` and `flatMap`

```
final case class ListTry[A](value: List[Try[A]]) { // We are using List.map, List.pure, List.flatMap.
  def map[B](f: A => B): ListTry[B] = ListTry(value.map(_.map(f))) // Using List.map here.
  def flatMap[B](f: A => ListTry[B]): ListTry[B] = ListTry(
    value.flatMap { // Using List.flatMap here.
      case Failure(t) => List(Failure(t)) // This is List.pure(Failure(t)).
      case Success(a) => f(a).value
    }
  )
}
```

It is now straightforward to generalize that code to an arbitrary foreign monad M :

```
final case class TryT[M[_]: Monad : Functor, A](value: M[Try[A]]) {
  def map[B](f: A => B): TryT[M, B] = TryT(value.map(_.map(f))) // Using M\text{map here}.
  def flatMap[B](f: A => TryT[M, B]): TryT[M, B] = TryT(
    value.flatMap { // Using M\text{flatMap here}.
      case Failure(t) => Monad[M].pure(Failure(t)) // Use 'Monad[M].pure' instead of List().
      case Success(a) => f(a).value
    }
  )
}
```

The “foreign lift” and the “base lift” also use the methods of the foreign monad M :

```
implicit class TryTLift[M[_]: Monad : Functor, A](m: M[A]) {
  def up: TryT[M, A] = TryT(m.map(Success(_)))
}
implicit class TryTBaseLift[M[_]: Monad : Functor, A](t: Try[A]) {
  def up: TryT[M, A] = TryT(Monad[M].pure(t))
}
```

After these definitions, we are able to combine `Try` with any other monad M , as long as an implicit values of type `Monad[M]` and `Functor[M]` are in scope. The code will look like this:

```
implicit val functorList: Functor[List] = ... // Create a Functor typeclass instance for List.
implicit val monadList: Monad[List] = ... // Create a Monad typeclass instance for List.

scala> val result = for {
  x <- List(1, 2, 3).up
  y <- Try(x + 100).up
} yield y
res0: TryT[List, Int] = TryT(List(Success(101), Success(102), Success(103)))
```

This example shows how to define the monad transformer T_{Try}^M that works parametrically in the foreign monad M . The result is a type constructor that we denote by $T_{\text{Try}}^{M,A}$. We can use that type as a monad with respect to the type parameter A . (The foreign monad M is usually the first type parameter in the transformer.) Similar code is required for T_{Read}^M , T_{Write}^M , T_{State}^M , and all other monads. The advantage of the transformer approach is that the code for T_L^M needs to be written once for every

base monad L (and not for every pair L, M).

14.1.3 Monad transformers for standard monads

While the code of `TryT` works, it leaves us with further questions. The code of `TryT` is specific to the `Try` monad and uses its internal structure such as the `Failure` and `Success` subtypes. It is not obvious how to define the type of a monad transformer T_L^M for an arbitrary given base monad L . The code for T_L^M 's methods is parametric in M but not in L (this is reflected in the notation: L is a *subscript* in T_L^M). Changing the base monad L to another monad requires a complete rewrite of T_L^M 's code.

There seems to be no general recipe for implementing a monad transformer T_L^M given a monad L , even when the code for L is known and is fully parametric. On the other hand, monad transformers are available for every known example of an explicitly defined, fully parametric monad. For some monads, functor composition works straightforwardly; for other monads, transformers need to be defined by special tricks. Table 14.2 shows the types of monad transformers that are known to work.

We can see from this table that monad transformers are created in different ways: composing inside the foreign monad (`Either`, `Writer`); composing outside the foreign monad (`Reader`, `Sel`); moving the recursive definition inside the foreign monad (`List`); and mixing the foreign monad in a special way with the original monad (`State`, `Cont`).

The specific choices of the transformer type constructors in Table 14.2 are far from obvious.¹ The only justification for those choices is that they work and obey the required laws, while no alternative transformers are known. We will derive and prove the required laws for these and other transformers later in this chapter. For now, we accept the results of Table 14.2.

Table 14.2: Known monad transformers for some monads.

Table 14.2 as correct and turn to other tasks regarding the practical use of monad transformers. The first task is to obtain the implementations of monad methods and lifts for the transformers from Table 14.2.

The type constructors for monad transformers are conventionally named `ReaderT`, `EitherT`, etc.

The `ReaderT` transformer The corresponding monad was already derived in Statement 10.2.8.5.

```
final case class ReaderT[M[_]: Monad : Functor, R, A](run: R => M[A]) {
  def map[B](f: A => B): ReaderT[M, R, B] = ReaderT(r => run(r).map(f))
  def flatMap[B](f: A => ReaderT[M, R, B]): ReaderT[M, R, B] = ReaderT {
    r => run(r)           // Type is M[A].
      .flatMap(f)         // Type is M[ReaderT[M, R, B]].
      .map(_.run(r))     // Type is M[B].
  }
}
```

The “foreign lift” and the “base lift” for `ReaderT` are implemented by

```
implicit class ReaderTLift[M[_]: Monad : Functor, R, A](m: M[A]) {
  def up: ReaderT[M, R, A] = ReaderT(_ => m) // This is the Reader monad\textsf{'s} 'pure' method.
}
implicit class ReaderTBaseLift[M[_]: Monad : Functor, R, A](t: R => A) {
```

¹The difficulty of this task is illustrated by the history of the `ListT` transformer in the Haskell standard library. The type constructor for `ListT` was chosen incorrectly in 2002 and remained so until it was corrected in 2007. See a discussion in https://wiki.haskell.org/index.php?title=ListT_done_right

```
def up: ReaderT[M, R, A] = ReaderT(r => Monad[M].pure(t(r)))
}
```

The lifts are written in the code notation as

foreign lift: $\text{flift} : M^A \rightarrow T_{\text{Reader}}^{M,A}$, $\text{flift}(m^{M^A}) \triangleq \underline{\cdot}^R \rightarrow m = \text{pu}_{\text{Reader}}(m)$,
 base lift: $\text{blift} : (R \rightarrow A) \rightarrow T_{\text{Reader}}^{M,A}$, $\text{blift}(t^{R \rightarrow A}) \triangleq r^R \rightarrow \text{pu}_M(t(r)) = t \circ \text{pu}_M$.

We have seen in Section 10.1.5 that getting data out of the `Reader` monad requires a runner θ_{Reader} , which is a function that calls `run` on some value of type R (i.e., injects `Reader`'s dependency value). In later sections 10.1.7 and 10.1.9, we have seen other monads that need runners. Generally, a runner for a monad M is a function of type $M^A \rightarrow A$. So, we may expect that the foreign monad M could have its own runner θ_M . How can we combine M 's runner (θ_M) with `Reader`'s runner? Since the type of T_{Reader}^M is a functor composition of `Reader` and M , the runners can be used independently of each other. We can first run the effect of M and then run the effect of the `Reader`:

$$(\theta_M^{\uparrow \text{Reader}} \circ \theta_{\text{Reader}}) : (R \rightarrow M^A) \rightarrow A .$$

Alternatively, we can run `Reader` first (injecting the dependency) and then run M 's effect:

$$(\theta_{\text{Reader}} \circ \theta_M) : (R \rightarrow M^A) \rightarrow A .$$

These runners commute because of the naturality law of θ_{Reader} , which holds for any $f^{A \rightarrow B}$:

$$f^{\uparrow \text{Reader}} \circ \theta_{\text{Reader}} = \theta_{\text{Reader}} \circ f , \text{ so } \theta_M^{\uparrow \text{Reader}} \circ \theta_{\text{Reader}} = \theta_{\text{Reader}} \circ \theta_M .$$

The `EitherT` transformer is similar to `TryT` since the type `Try[A]` is equivalent to `Either[Throwable, A]`.

The `WriterT` transformer It is easier to begin with the `flatten` method for the transformed monad $T^A \triangleq M^{A \times W}$, which has type signature

$$\text{ftn}_T : M^{M^{A \times W} \times W} \rightarrow M^{A \times W} , \text{ ftn}_T(t^{M^{M^{A \times W} \times W}}) = ???^{M^{A \times W}} .$$

Since M is an unknown, arbitrary monad, the only way of computing a value of type $M^{A \times W}$ is by using the given value t . The only way to get a value of type A wrapped in $M^{A \times W}$ is by extracting the type A from inside $M^{A \times W}$. So, we need to flatten the two layers of M that are present in the type of t . However, we cannot immediately apply M 's `flatten` method to t because t 's type is not of the form M^{M^X} with some X . To bring it to that form, we use M 's `map` method:

$$t \triangleright (m^{M^{A \times W}} \times w^W \rightarrow m \triangleright (p^{A \times W} \rightarrow p \times w)^{\uparrow M})^{\uparrow M} : M^{M^{A \times W} \times W} .$$

Now the type is well adapted to using both M 's and `Writer`'s flatten methods:

$$\begin{aligned} \text{ftn}_T(t^{M^{M^{A \times W} \times W}}) &= t \triangleright (m^{M^{A \times W}} \times w^W \rightarrow m \triangleright (p^{A \times W} \rightarrow p \times w)^{\uparrow M})^{\uparrow M} \triangleright \text{ftn}_M \triangleright (\text{ftn}_{\text{Writer}})^{\uparrow M} , \\ \text{ftn}_T &= (m^{M^{A \times W}} \times w^W \rightarrow m \triangleright (p^{A \times W} \rightarrow p \times w)^{\uparrow M})^{\uparrow M} ; \text{ftn}_M ; (\text{ftn}_{\text{Writer}})^{\uparrow M} \\ &= \text{flm}_M(m \times w \rightarrow m \triangleright (p \rightarrow p \times w)^{\uparrow M} ; \text{ftn}_{\text{Writer}}^{\uparrow M}) \\ &= \text{flm}_M(m \times w \rightarrow m \triangleright (a \times w_2 \rightarrow a \times (w \oplus w_2))) . \end{aligned}$$

Translating this formula to Scala, we obtain the code of `flatMap`:

```
final case class WriterT[M[_]: Monad, W: Monoid, A](t: M[(A, W)]) {
  def map[B](f: A => B): WriterT[M, W, B] = WriterT(t.map { case (a, w) => (f(a), w) })
  def flatMap[B](f: A => WriterT[M, W, B]): WriterT[M, W, B] = WriterT(
    t.flatMap { case (a, w) => f(a).t.map { case (b, w2) => (b, w |+| w2) } }
  )
}
```

The “foreign lift” and the “base lift” for `WriterT` are (omitting some Scala boilerplate):

```
def flift[A]: M[A] => M[(A, W)] = _.map { a => (a, Monoid[W].empty) }
def blift[A]: ((A, W)) => M[(A, W)] = { t => Monad[M].pure(t) }
```

$$\begin{aligned} \text{foreign lift} : \text{flift} : M^A \rightarrow T_{\text{Writer}}^{M,A} & , \quad m : M^A \triangleright \text{flift} \triangleq m \triangleright (a \rightarrow a \times e_w)^{\uparrow M} & , \\ \text{base lift} : \text{blift} : (R \rightarrow A) \rightarrow T_{\text{Writer}}^{M,A} & , \quad \text{blift} \triangleq \text{pu}_M & . \end{aligned}$$

Comparing the lifts for `ReaderT` and `WriterT`, we notice some common features of lifts for compositional transformers (`ReaderT`, `EitherT`, `WriterT`): one of the lifts is equal to a `pure` method, and the other is equal to a lifted `pure` method. We will see later that this pattern applies to all monad transformers defined via functor composition.

The `ListT` transformer is more complicated because `List` is a recursive type. To shorten the code, we use the `Option` type to represent $\mathbb{1} + A \times T_{\text{List}}^{M,A}$ in the recursive type $T_{\text{List}}^{M,A} \triangleq M^{\mathbb{1} + A \times T_{\text{List}}^{M,A}}$:

```
final case class ListT[M[_]: Monad : Functor, A](value: M[Option[(A, ListT[M, A])]])
```

The `List` monad was defined using the list concatenation function `concat`; “flattening” a `List[List[A]]` means concatenating all nested lists. If the ordinary `List` were defined via `Option`, the code for `concat` would be

```
final case class List0[A](value: Option[(A, List0[A])])
def concat[A](t1: List0[A], t2: List0[A]): List0[A] = t1.value match {
  case None => t2
  case Some((a, tail)) => Some((a, concat(tail, t2))) // Recursive call.
}
```

We need to implement the analogous concatenation function for `ListT`. To visualize the type of `ListT`, consider an example where the foreign monad is $M^A \triangleq R \rightarrow A$. Expand a few steps of recursion:

$$T_{\text{List}}^{M,A} = R \rightarrow \mathbb{1} + A \times (R \rightarrow \mathbb{1} + A \times (R \rightarrow \dots (R \rightarrow \mathbb{1} + A \times T_{\text{List}}^{M,A}) \dots)) \quad .$$

This type describes an M -effect wrapped around each next value of type A in the list. To extract further values in the list, we need to run the nested effects.

This understanding helps us implement the concatenation function for `ListT`:

```
def concatT[M[_]: Monad : Functor, A](t1: ListT[M, A], t2: ListT[M, A]): ListT[M, A] = ListT(for {
  first <- t1.value // Run the first effect and get 'first: Option[(A, ListT[M, A])]'.
  rest = first match {
    case None => t2.value
    case Some((a, tail)) => Monad[M].pure(Some((a, concatT(tail, t2)))) // Recursive call.
  }
  y <- rest
} yield y)
```

We are now ready to write the code of `ListT`’s `flatMap` method:

```
final case class ListT[M[_]: Monad : Functor, A](value: M[Option[(A, ListT[M, A])]]) {
  def map[B](f: A => B): ListT[M, B] = ListT(t.map(_.map { // Use M\text{tsf}\{'s and Option\text{tsf}\{'s
    map methods.
    case (head, tail) => (f(head), tail.map(f))
  }))
  def flatMap[B](f: A => ListT[M, B]): ListT[M, B] = ListT(
    t.flatMap { // Here we need a function of type Option[(A, ListT)] => M[Option[(A, ListT)]].
      case None => Monad[M].pure(None)
      case Some((head, tail)) => concatT(f(head), tail).value // Type is M[Option[(A, ListT)]].
    })
}
```

The lifts for `ListT` are:

```
def flift[A]: M[A] => ListT[M, A] = { m => ListT(m) }
def blift[A]: ((A, W)) => M[(A, W)] = { t => Monad[M].pure(t) }
```

$$\begin{aligned} \text{foreign lift : } \text{flift} : M^A \rightarrow T_{\text{Writer}}^{M,A} , \quad m : M^A \triangleright \text{flift} &\triangleq m \triangleright (a \rightarrow a \times e_w)^{\uparrow M} , \\ \text{base lift : } \text{blift} : (R \rightarrow A) \rightarrow T_{\text{Writer}}^{M,A} , \quad \text{blift} &\triangleq \text{pu}_M . \end{aligned}$$

Finally, we consider the runners for `ListT`. List-like monads do not usually need runners ($\theta_{\text{List}} : \text{List}^A \rightarrow A$) because the meaning of the monad's value is usually the entire set or collection of useful results rather than a single value, and also because the collection may be empty, which is incompatible with having a fully parametric and lawful runner. However, the foreign monad M may come with its runner θ_M . In that case, we will need to run the M 's effects if we are to extract values from a `ListT`.

Intuitively, we may expect to be able to apply M 's runner (θ_M) to a value of type `ListT[M, A]` and obtain a `List[A]`. The code may be written as:

```
def runListT[A](runner: M[A] => A)(listT: ListT[M, A]): List[A] = runner(listT.value) match {
  case None => List()
  case Some((head, tailT)) => head :: run(runner)(tailT)
}
```

However, this code will not always work as expected because the type of `ListT` contains a layer of M in every recursive step. So, extracting a value of type `List[A]` from a value of type `ListT[M, A]` involves applying θ_M once per list value. When M has a function type, a value of type `ListT[M, A]` is an “on-call” sequence whose elements are not available in advance but are evaluated on demand, by calling θ_M . Because of this, the total number of elements in `ListT` may not be known in advance, and may be even unlimited. An example of an “infinite” collection uses the `Reader` monad as M :

```
def ascend(a: Int): ListT[Reader[Int, ?], Int] = ListT(Lazy(m => Some((a, ascend(a + m)))))
```

This is a well-defined value of type `ListT` that represents integer sequences $[a, a + m, a + 2 * m, \dots]$. The parameter m is passed via the `Reader` monad's dependency injection. It is clear that `runListT` will never terminate when applied to such a sequence.

This example shows how to use `ListT` for creating effectful “infinite” streams.

The `StateT` transformer The formula for that transformer's type, $\text{StateT}[M, S, A] = S \Rightarrow M[S, A]$, requires us to place the foreign monad M at a specific place in the type of the `State` monad. Let us implement the monad methods for `StateT`:

```
final case class StateT[M[_]: Monad : Functor, S, A](run: S => M[(A, S)]) {
  def map[B](f: A => B): StateT[M, S, B] = StateT(s1 => run(s1).map f // Use M\text{mapf}{'}s map method.
    case (a, s2) => (f(a), s2)
  )
  def flatMap[B](f: A => StateT[M, S, B]): StateT[M, S, B] = StateT(s1 => run(s1).flatMap {
    case (a, s2) => f(a).run(s2) // Use the updated intermediate state s2.
  })
}
```

$$T_{\text{State}}^{M,A} \triangleq T^A \triangleq S \rightarrow M^{A \times S} , \quad \text{pu}_T \triangleq a : A \rightarrow s : S \rightarrow \text{pu}_M(a \times s) , \quad (14.1)$$

$$\text{flm}_T(f : A \rightarrow T^B) \triangleq t : T^A \rightarrow s : S \rightarrow t(s_1) \triangleright \text{flm}_M(a \times s_2 \rightarrow f(a)(s_2)) . \quad (14.2)$$

The implementations of the lift functions needs to be adapted to the type of `StateT`:

```
def flift[M[_]: Functor, S, A](m: M[A]): StateT[M, S, A] = StateT(s => m.map { a => (a, s) })
def blift[M[_]: Monad, S, A](p: State[S, A]): StateT[M, S, A] = StateT(s => Monad[M].pure(p.run(s)))
```

$$\text{foreign lift : } \text{flift} : M^A \rightarrow T_{\text{State}}^{M,A} , \quad m : M^A \triangleright \text{flift} \triangleq s \rightarrow m \triangleright (a \rightarrow a \times s)^{\uparrow M} ,$$

$$\text{base lift : } \text{blift} : (S \rightarrow A \times S) \rightarrow T_{\text{State}}^{M,A} , \quad \text{blift} \triangleq p : S \rightarrow A \times S \rightarrow p \circ \text{pu}_M .$$

The `State` monad's runner (θ_{State}) substitutes a given initial state and extracts the final value:

```
def runState[S, A](init: S)(p: State[S, A]): A = p.run(init)._1
```

$$\theta_{\text{State}}(i^S) \triangleq p^{S \rightarrow A \times S} \rightarrow i \triangleright p \triangleright \pi_1 \quad .$$

We cannot use θ_{State} directly with `StateT` because that transformer is not a functor composition. There is no general “base runner” `brun` (θ_{State}) that would take an *arbitrary* runner $\theta_{\text{State}} : \text{State}^{S,A} \rightarrow A$ and convert it into a function of type $T_{\text{State}}^{M,A} \rightarrow M^A$ by running the `State` monad’s effects but keeping the effects of the foreign monad M . The type of T_{State}^M does not allow base runners parametric in θ_{State} . Instead, we need to implement a specialized base runner that converts $T_{\text{State}}^{M,A}$ into M^A directly:

```
def brunStateT[M[_]: Functor, S, A](init: S)(t: StateT[M, S, A]): M[A] = t.run(init).map(_._1)

brunState(i^S) \triangleq t^{S \rightarrow M^{A \times S}} \rightarrow i \triangleright t \triangleright \pi_1^M \quad .
```

If we need to run only the effects of M but keep the `State` monad’s effect, we need different code:

```
def frunStateT[M[_], S, A](runner: M[A] => A)(t: StateT[M, S, A]): State[S, A] = State { s =>
  val m: M[(A, S)] = t.run(s)
  runner(m)
}

frunState(\theta^{M^A \rightarrow A}) \triangleq t^{S \rightarrow M^{A \times S}} \rightarrow (s \rightarrow s \triangleright t \triangleright \theta) = t^{S \rightarrow M^{A \times S}} \rightarrow t \triangleright \theta \quad .
```

The `ContT` transformer For the continuation monad (`Cont[R, A]`), the transformer’s type constructor `ContT[M, R, A]` has a peculiar formula where the foreign monad’s constructor M is applied to the result type R but not to any values of type A . The code for the monad methods is:

```
final case class ContT[M[_], R, A](run: (A => M[R]) => M[R]) {
  def map[B](f: A => B): ContT[M, R, B] = ContT { (k: B => M[R]) => run(f andThen k) }
  def flatMap[B](f: A => ContT[M, R, B]): ContT[M, R, B] = ContT { (k: B => M[R]) =>
    val g: A => M[R] = a => f(a).run(k)
    run(g)
  }

  def blift[A](p: ContT[R, A]): ContT[M, R, A] = ???
}
```

Turning now to the code for lifts, we can easily write the “foreign lift”:

```
def flift[M[_]: Monad, R, A](m: M[A]): ContT[M, R, A] = ContT { (k: A => M[R]) => m.flatMap(k) }

flift : M^A \rightarrow (A \rightarrow M^R) \rightarrow M^R = flm_M \quad .
```

However, implementing the “base lift” is impossible: the required type signature

$$\text{blift} : ((A \rightarrow R) \rightarrow R) \rightarrow (A \rightarrow M^R) \rightarrow M^R$$

has no implementation that could work with all monads M . The reason is that, in general, we cannot produce a value of type R out of M^R (not all monads have runners), and so we cannot convert $A \rightarrow M^R$ into $A \rightarrow R$. This prevents us from using the first argument of `blift` (a function having type $(A \rightarrow R) \rightarrow R$), while creating a value of type $(A \rightarrow M^R) \rightarrow M^R$ from scratch is impossible.

To produce an explicit example where `blift` cannot be implemented, we use the `curryhoward` library and choose a monad M^R that has no runner, e.g., $M^R \triangleq Z + R$:

```
import io.chymyst.ch._

scala> def blift[R, A]: Cont[R, A] => ContT[Either[Int, ?], R, A] = implement
Error:(43, 65) type ((A => R) => R) => (A => Either[<c>Int, R]) => Either[<c>Int, R] cannot be
  implemented
  def blift[R, A]: Cont[R, A] => ContT[Either[Int, ?], R, A] = implement
```

The absence of “base lift” for the continuation monad transformer means that we cannot use the continuation monad in a functor block as freely as other monads. We cannot lift an arbitrary given value of type `Cont` (for instance, a value produced by a third-party library that cannot be modified) into the transformed monad `ContT`. This is a serious deficiency of the continuation monad that makes

it unsuitable for certain applications. Concrete values of type `Cont[R, A]` are often interfaces to system libraries implemented in non-fully parametric code. It may be impossible to rewrite that code to return `ContT[M, R, A]` with an arbitrary monad `M`.

The `Sel` transformer Although the selector monad $\text{Sel}^{Z,A} \triangleq (A \rightarrow Z) \rightarrow A$ resembles the continuation monad, its transformer is simpler because it consists of composing outside the foreign monad M (i.e., we replace A by M^A in the type of `sel`). Only one implementation fits the type:

```
final case class SelT[M[_]: Monad : Functor, Z, A](run: (M[A] => Z) => M[A]) {
  def map[B](f: A => B): SelT[M, Z, B] = SelT { (mbz: M[B] => Z) => run(ma => mbz(ma.map(f))).map(f) }
  def flatMap[B](f: A => SelT[M, Z, B]): SelT[M, Z, B] = SelT { (mbz: M[B] => Z) =>
    val amb: A => M[B] = a => f(a).run(mbz)
    val maz: M[A] => Z = ma => mbz(ma.flatMap(amb))
    run(maz).flatMap(amb)
  }
}
```

Since $T_{\text{Sel}}^M = \text{Sel} \circ M$ is a functor composition, the lifts and the runners work as in the `Reader` monad.

```
def flift[A](m: M[A]): SelT[M, Z, A] = SelT(_ => m) // This is the Sel monad\text{tsf{'}}s 'pure' method.
def blift[A](sel: SelT[Z, A]): SelT[M, Z, A] = SelT { (k: M[A] => Z) =>
  val az: A => Z = Monad[M].pure andThen k
  Monad[M].pure(sel(az))
}
```

$$\begin{aligned} \text{foreign lift : flift : } M^A &\rightarrow T_{\text{Sel}}^{M,A} \quad , \quad m^{:M^A} \triangleright \text{flift} \triangleq \underline{:_{M^A} \rightarrow Z} \rightarrow m = m \triangleright \text{pu}_{\text{Sel}} \quad , \\ \text{base lift : blift : } ((A \rightarrow Z) \rightarrow A) &\rightarrow T_{\text{Sel}}^{M,A} \quad , \quad \text{blift} \triangleq \text{pu}_M^{\uparrow \text{Sel}} \quad . \end{aligned}$$

The runners can be applied in any order because naturality makes them commute:

$$\theta_M^{\uparrow \text{Sel}} \circ \theta_{\text{Sel}} = \theta_{\text{Sel}} \circ \theta_M \quad .$$

The `EvalT` transformer The evaluation monad $\text{Eval}^A \triangleq A + (\mathbb{1} \rightarrow A)$ (Section 10.1.8) describes a choice between eager and lazy evaluation. The corresponding transformer is defined by

$$T_{\text{Eval}}^{M,A} \triangleq M^{A+(\mathbb{1} \rightarrow M^A)} \quad .$$

Let us implement the `EvalT` monad using `Either` and write the monad methods. The code for the `Eval` monad in Section 10.1.8 used a helper function that we called `get`:

$$\text{get : Eval}^A \rightarrow A \quad .$$

We see now that `get` is `Eval`'s runner. A similar function for `EvalT` plays the role of the “base runner” and helps us implement `flatMap`:

```
final case class EvalT[M[_]: Monad : Functor, A](value: M[Either[A, Unit => M[A]]]) {
  def map[B](f: A => B): EvalT[M, B] = EvalT(value.map {
    case Left(a) => Left(f(a))
    case Right(g) => Right(_ => g().map(f))
  })
  def flatMap[B](f: A => EvalT[M, B]): EvalT[M, B] = EvalT(value.flatMap {
    case Left(a) => f(a).value
    case Right(g) => Monad[M].pure(Right(_ => g().flatMap(a => EvalT.brunit(f(a).value)) ))
  })
}

object EvalT {
  def pure[M[_]: Monad, A](a: A): EvalT[M, A] = EvalT(Monad[M].pure(Left(a)))
}
```

```
def brun[M[_]: Monad, A]: EvalT[M, A] = _.value.flatMap { // Either[A, Unit=>M[A]] => M[A].
  case Left(a) => Monad[M].pure(a)
  case Right(g) => g(())
}
```

The lifts are implemented straightforwardly:

```
def flift[A](m: M[A]): EvalT[M, A] = EvalT(m.map(a => Left(a))) // 'Left(a)' means 'Eval.pure(a)'.
def blift[A](e: Either[A, Unit => A]): EvalT[M, A] = EvalT(Monad[M].pure(e match {
  case Left(a) => Left(a)
  case Right(g) => Right { _ => Monad[M].pure(g(())) }
}))
```

Given a runner for M , we can run M 's effects while keeping Eval 's effects intact:

```
def runEvalT[A](runner: M[A] => A)(t: EvalT[M, A]): Eval[A] = runner(t) match {
  case Left(a) => Eager(a) // An eager value is ready.
  case Right(g) => Lazy(_ => runner(g(()))) // A lazy value remains lazy.
}
```

To run Eval 's effects but keep M 's effects intact, we need a base runner:

```
def brunEvalT[A](t: EvalT[M, A]): M[A] = t.value.flatMap {
  case Left(a) => Monad[M].pure(a)
  case Right(g) => g(())
}
```

14.1.4 Combining more than two monads

How can we combine *three or more* monads in a single functor block? For example, we may want to define some transformers and lifts so that the following code works:

```
for { // Could this code compile?
  i <- List(1, 2, 3).up
  j <- Writer(...).up
  x <- Try(i + j + 100).up
  y <- Future(x / 2).up
  z <- Reader(...).up
} yield z
```

A monad transformer combines a base monad L with any given foreign monad M . The result is a new, “larger” monad T_L^M that can describe the effects of both L and M . If we need to combine monad T_L^M with a third monad (say, K), we could use K 's transformer and obtain the monad $T_K^{T_L^M}$ that can describe the effects of K, L , and M . This technique is called making a **stack** of monads.

As an example, let us build a stack of `Reader`, `Writer`, and `State` monads. We set $K^\bullet = \text{Reader}^{R, \bullet}$, $L^\bullet = \text{Writer}^{W, \bullet}$, and $M^\bullet = \text{State}^{S, \bullet}$, and begin with T_L^M . Since the `Writer`'s transformer composes inside the foreign monad, we write

$$T_{\text{Writer}}^{\text{State}, A} = S \rightarrow A \times W \times S \quad .$$

Now we apply the `Reader`'s transformer, which composes outside the foreign monad:

$$T_{\text{Reader}}^{T_{\text{Writer}}^{\text{State}, A}} = R \rightarrow S \rightarrow A \times W \times S \quad .$$

This monad has an injected dependency of type R , an internally maintained state of type S , and an output value of type W . A value of type $R \rightarrow S \rightarrow A \times W \times S$ describes all three effects at once.

If we already implemented the transformers `ReaderT` and `WriterT`, we may write the Scala type

```
type RWS[R, W, S, A] = ReaderT[WriterT[State[S, ?], W, ?], R, A]
```

Nested transformers are substituted into the first argument of the outer transformer, which has the type constructor kind. So, we need to use the “kind projector” plugin to represent those type constructors such as `State[S, ?]`.

Note that the transformer composition does not necessarily represent functor composition. For instance, the plain functor composition `Reader[R, Writer[W, State[S, A]]]` is not a monad.

Base monad L^A	Foreign monad M^A	Monad T_L^M	Monad T_M^L	Same?
Reader [R, A]	Either [E, A]	$T_L^{M,A} = R \rightarrow E + A$	$T_M^{L,A} = R \rightarrow E + A$	Yes
Reader [R, A]	Writer [W, A]	$T_L^{M,A} = R \rightarrow A \times W$	$T_M^{L,A} = R \rightarrow A \times W$	Yes
Reader [R, A]	List [A]	$T_L^{M,A} = R \rightarrow \text{List}^A$	$T_M^{L,A} \triangleq R \rightarrow \mathbb{1} + A \times T_M^{L,A}$	No
Reader [R, A]	State [S, A]	$T_L^{M,A} = R \rightarrow S \rightarrow A \times S$	$T_M^{L,A} = S \rightarrow R \rightarrow A \times S$	Yes
Either [E, A]	State [S, A]	$T_L^{M,A} = S \rightarrow (E + A) \times S$	$T_M^{L,A} = S \rightarrow E + A \times S$	No
Either [E, A]	Cont [R, A]	$T_L^{M,A} = (E + A \rightarrow R) \rightarrow R$	$T_M^{L,A} = (A \rightarrow E + R) \rightarrow E + R$	No

Table 14.3: Examples of monads stacked in different orders.

We could stack the monads K , L , and M in a different order, e.g., $T_L^{T_M^K}$ or $T_M^{T_K^L}$. In general, different orders will produce inequivalent monads that combine effects in different ways, although in certain cases the results are independent of the stacking order. Table 14.3 shows some examples of stacking two monads in different orders.

The type of $T_{\text{State}}^{\text{Reader}}$ is equivalent to that of $T_{\text{Reader}}^{\text{State}}$: the only difference is the flipped order of curried arguments. To verify that the implementations of `pure` and `flatMap` in these transformers are equivalent up to that order, we begin by writing `flatMap` for the `Reader` and `State` monads:

$$\begin{aligned} \text{flm}_{\text{Reader}}(g^{A \rightarrow R \rightarrow B}) &= p^{R \rightarrow A} \rightarrow r^R \rightarrow (p(r) \triangleright g)(r) , \\ \text{flm}_{\text{State}}(g^{A \rightarrow \text{State}^{S,B}}) &= p^{S \rightarrow A \times S} \rightarrow s_1^S \rightarrow p(s_1) \triangleright (a \times s_2 \rightarrow g(a)(s_2)) . \end{aligned}$$

Now we write the code for $T_{\text{Reader}}^{\text{State}}$'s methods:

$$\begin{aligned} T_{\text{Reader}}^{\text{State},A} &\triangleq T_1^A = R \rightarrow S \rightarrow A \times S , \\ \text{pu}_{T_1} = a^{A \rightarrow} &_^R \rightarrow \text{pu}_{\text{State}}(a) = a^{A \rightarrow} _^R \rightarrow s_1^S \rightarrow a \times s , \\ \text{flm}_{T_1}(f^{A \rightarrow T_1^B}) &= t^{T_1^A} \rightarrow r^R \rightarrow t(r) \triangleright \text{flm}_{\text{State}}(a^{A \rightarrow} f(a)(r)) \\ &= t \rightarrow r \rightarrow t(r) \triangleright (p^{S \rightarrow A \times S} \rightarrow s_1^S \rightarrow p(s_1) \triangleright (a \times s_2 \rightarrow f(a)(r)(s_2))) \\ &= t \rightarrow r \rightarrow s_1 \rightarrow t(r)(s_1) \triangleright (a \times s_2 \rightarrow f(a)(r)(s_2)) . \end{aligned}$$

The code for $T_{\text{State}}^{\text{Reader}}$ is found by using Eq. (14.2):

$$\begin{aligned} T_{\text{State}}^{\text{Reader},A} &\triangleq T_2^A = S \rightarrow R \rightarrow A \times S , \\ \text{pu}_{T_2} = a^{A \rightarrow} &s_1^S \rightarrow \text{pu}_{\text{Reader}}(a \times s) = a^{A \rightarrow} s_1^S \rightarrow _^R \rightarrow a \times s , \\ \text{flm}_{T_2}(f^{A \rightarrow T_2^B}) &= t^{T_2^A} \rightarrow s_1^S \rightarrow t(s_1) \triangleright \text{flm}_{\text{Reader}}(a \times s_2 \rightarrow f(a)(s_2)) \\ &= t \rightarrow s_1 \rightarrow t(s_1) \triangleright (p \rightarrow r \rightarrow (p(r) \triangleright (a \times s_2 \rightarrow f(a)(s_2)))(r)) \\ &= t \rightarrow s_1 \rightarrow r \rightarrow (t(s_1)(r) \triangleright (a \times s_2 \rightarrow f(a)(s_2)))(r) . \end{aligned}$$

We see that the codes differ only by flipping the curried arguments r^R and s_1^S .

An example of inequivalent transformers is the pair $T_{\text{Reader}}^{\text{List}}$ and $T_{\text{List}}^{\text{Reader}}$. These two monads describe different ways of combining the effects of `Reader` and `List`. In a value of type $T_{\text{Reader}}^{\text{List}}$, the injected dependency (of type R) is used to produce an “eager” list of values. The length of the resulting list, as well as every value in it, are determined at once depending on the injected value of type R . In contrast, a data structure of type $T_{\text{List}}^{\text{Reader}}$ represents a “lazy” (on-call) sequence because getting the value of each element is obtained by calling a *new* function of type $R \rightarrow \mathbb{1} + A \times T_{\text{List}}^{\text{Reader}}$. The resulting on-call sequence may be unlimited in length.

The types of $T_{\text{State}}^{\text{Either}}$ and $T_{\text{Either}}^{\text{State}}$ are also not equivalent. The effect of $\text{State}^{S,A} = S \rightarrow A \times S$ is to produce an updated internal state alongside with a result value (of type A). Let us look at the types of the transformed monads:

$$T_{\text{State}}^{\text{Either},A} = S \rightarrow E + A \times S \quad , \quad T_{\text{Either}}^{\text{State},A} = S \rightarrow (E + A) \times S \quad .$$

The first type describes a computation that may fail with an error description (of type E) and produce no new state and no result value. The second type describes a computation that never fails to update the state, although it could fail to give a result value.

This example shows that effects of the two monads are not always independent within the larger transformed monad. An effect of one monad (`Either`) could “cancel” an effect of the other (`State`).

The monad stack $T_K^{T_L^M}$ is a monad whose code is parametric in M and thus can be viewed as the monad transformer for the monad T_K^L :

$$T_{T_K^L}^M \triangleq T_K^{T_L^M} \quad . \quad (14.3)$$

To combine $T_K^{T_L^M}$ with a new monad N , we write the stack $T_K^{T_L^{T_M^N}}$ (where we used M ’s transformer, T_M). With this technique, we may combine any number of monads in a monad stack. The only requirement is the need to know the transformers for each monad (except possibly for the last one).

We may also view monad stacking as a binary operation (denoted by \triangleleft) that combines monads K and L into a new monad $K \triangleleft L \triangleq T_K^L$. This operation is associative by construction: the transformer for the monad T_K^L is defined via Eq. (14.3), which we can rewrite as

$$(K \triangleleft L) \triangleleft M = K \triangleleft (L \triangleleft M) \quad .$$

A monad stack is then denoted unambiguously by, e.g., $K \triangleleft L \triangleleft M \triangleleft N$ without need for parentheses.

Does the stacking operation (\triangleleft) have a unit element? We notice from Table 14.2 that each transformer T_L^M somehow mixes the foreign monad M with the type of the base monad L . In every example, if we choose M as the identity monad `Id` (defined by $\text{Id}^A \triangleq A$), the transformed type will become equivalent to the base monad L . Intuitively, the identity monad is a “trivial effect”; combining it with another monad should not change that monad’s effects. So, we may expect this property to hold for every transformer:

$$T_L^{\text{Id}} \cong L \quad , \quad L \triangleleft \text{Id} \cong L \quad .$$

We also define the monad transformer for the identity monad as equal to the foreign monad:

$$T_{\text{Id}}^M \triangleq M \quad , \quad \text{Id} \triangleleft M \triangleq M \quad .$$

Because of these properties of the identity monad, the base monad L can be recovered as as T_L^{Id} if we know the formula for the transformer $T_L^{\bullet,\bullet}$. So, if we already have the type and an implementation for a monad’s transformer, we do not need to define the monad’s type and methods separately. This saves time when implementing monad code in libraries.

It is unknown² whether a transformer exists for every monad relevant to functional programming; say, for every monad defined by purely functional code.³ However, this chapter will show that lawful transformers exist for every monad derived anywhere in this book (except for the continuation monad’s transformer, which violates some of the required laws).

14.1.5 Lift operations for monad stacks

It remains to see how we could define the lifts and the runners for a given monad stack.***

²For a discussion, see <https://stackoverflow.com/questions/24515876/>

³Monad transformers may certainly fail to exist for monads defined using low-level, non-purely functional code with side effects. A lawful transformer must allow the foreign monad to cancel the effect of the base monad, as we have seen in an example with `State` and `Either`. However, non-referentially transparent code may perform side effects (picture them as “launching missiles”) that cannot be undone or canceled. See <https://stackoverflow.com/questions/11792275/>

14.2 Laws of monad transformers

14.2.1 Motivation for the laws of lift functions

A monad transformer creates a new monad that combines the effects of other monads. Our next goal is to derive the properties required for the new monad to work well in practice. So, let us look at the programmer's intuitions about monadic programs written using transformers.

To be specific, assume that we have combined the monads L and M into a transformed monad $T \triangleq L \otimes M \triangleq T_L^M$ and defined the necessary lifts, $up_L : L^A \rightarrow T^A$ and $up_M : M^A \rightarrow T^A$. After these definitions, a programmer should be able to write functor blocks with lifted values of L and M .

Programmers will certainly expect all previously accepted properties of functor blocks to remain valid after lifting all monadic values of types L and M to the transformed monad T . A monadic program in T could look like this:

```
for {
  i <- someValueL.up
  x <- someValueT(i)
  y <- anotherValueT(y)
  z <- someValueM(x).up
} yield z
```

This program combines values of L and M lifted to T with some directly available values of the monad T (say, obtained from other monadic programs). Refactoring such programs involves, as a special case, refactoring of functor blocks in the monad T alone. It follows that T must itself satisfy the monad laws, — the laws we derived in Chapter 10 by analyzing various functor block refactorings.

There are some additional code refactorings specific to lifted monadic values. Lifted values may occur before or after a source line with an arbitrary value of T . Similarly to the standard monadic refactorings, we consider three cases: a lifted `pure` method before a source line, a lifted `pure` method after a source line, and refactoring two adjacent lifted source lines into a separate functor block.

If L 's `pure` method is used before a source line, we expect this refactoring to work:

<pre>// Anywhere inside a for/yield: y <- Monad[L].pure(x).up // Assume x: A z <- f(y) // f: A => T[B]</pre>	<pre>// Must be equivalent to... y = x // x: A z <- f(y) // f: A => T[B]</pre>
---	--

The corresponding law is $pure(x).up.flatMap(f) == f(x)$ or

$$pu_L \circ blift \circ flm_T(f) = f .$$

If L 's `pure` method is used after a source line, we expect the following code equivalence:

<pre>// Anywhere inside a for/yield: x <- t // Assume t: T[A] y <- Monad[L].pure(x).up</pre>	<pre>// Must be equivalent to... x <- t // Assume t: T[A] y = x</pre>
--	--

The corresponding law is $t.flatMap(x => L.pure(x).up) == t$ or

$$flm_T(pu_L \circ blift) = id .$$

The third law comes from refactoring a functor block when two adjacent source lines are lifted from L . We may equivalently put these lines in a functor block in L and then lift its result to T :

<pre>// Anywhere inside a for/yield: x <- p.up // Assume p: L[A] y <- q(x).up // q: A => L[B]</pre>	<pre>// Must be equivalent to... pq = for { x <- p; y <- q(x) } yield y y <- pq.up // Lift a refactored block.</pre>
--	---

The corresponding law is $p.up.flatMap(x => q(x).up) == p.flatMap(q).up$ or

$$blift \circ flm_T(q \circ blift) = flm_L(q) \circ blift .$$

The form of these three laws can be simplified if we rewrite them in terms of the Kleisli composition operations \diamond_L and \diamond_T . Recall that `flatMap` is related to the Kleisli composition by

$$f:A \rightarrow T^B \diamond_T g:B \rightarrow T^C = f \circ flm_T(g) .$$

The first law is then written as

$$\text{pu}_L \circ \text{blift} \circ \text{flm}_T(f) = (\text{pu}_L \circ \text{blift}) \diamond_T f \stackrel{!}{=} f \quad .$$

The second law is not of the form $f \circ \text{flm}(\dots)$, so let us pre-compose it with an arbitrary $f : A \rightarrow T^B$:

$$f \circ \text{flm}_T(\text{pu}_L \circ \text{blift}) = f \diamond_T (\text{pu}_L \circ \text{blift}) \stackrel{!}{=} f \circ \text{id} = f \quad .$$

Pre-compose also the third law with an arbitrary function $p : A \rightarrow L^B$:

$$p \circ \text{blift} \circ \text{flm}_T(q \circ \text{blift}) = (p \circ \text{blift}) \diamond_T (q \circ \text{blift}) \stackrel{!}{=} p \circ \text{flm}_L(q) \circ \text{blift} = (p \diamond_L q) \circ \text{blift} \quad .$$

The first two laws are now easier to interpret: they require the function $\text{pu}_L \circ \text{blift}$ to be both a left and a right unit for the Kleisli composition \diamond_T . However, the monad laws of T already provide such a unit element, namely pu_T . An associative operation can have only one two-sided unit. To see that, it is sufficient to write the equation

$$(\text{pu}_L \circ \text{blift}) = (\text{pu}_L \circ \text{blift}) \diamond_T \text{pu}_T = \text{pu}_T \quad .$$

So, the first two laws are equivalent to a single **identity law**,

$$\text{pu}_L \circ \text{blift} = \text{pu}_T \quad .$$

The third law says that the post-composition with blift will map a Kleisli composition in L to a Kleisli composition in T . This is the **composition law** of lifts. This law is sometimes easier to verify when expressed through ftn , although this formulation of the law is less visually clear:

$$\text{blift} \circ \text{blift}^{\uparrow T} \circ \text{ftn}_T = \text{ftn}_L \circ \text{blift} \quad .$$

The function blift maps values L^A to $T_L^{M,A}$, and the identity and composition law enforce its compatibility with the monad methods of L and T_L^M . Laws of this form were introduced in Section 10.3.4 where we considered general monad morphisms $\phi : M \rightsquigarrow N$ between two monads M and N . So, we have now shown that the expected properties of functor blocks are equivalent to the requirement that blift should be a monad morphism $L \rightsquigarrow T_L^M$.

Since we expect to be able to use the monads L and M equally well within a functor block, the same refactorings must apply when we use values of type M instead of L . Thus, the laws of monad morphisms must hold also for the foreign lift function ($\text{flift} : M \rightsquigarrow T_L^M$).

Since monad morphisms are automatically natural transformations (Statement 10.3.4.4), both lifts also obey the corresponding naturality laws. The naturality laws can be also motivated by considering functor blocks where a source line with a lift occurs next to a non-source line. We will omit the details, keeping in mind that naturality laws always hold for fully parametric code.

14.2.2 Motivation for the laws of runner functions

A monad's runner $\theta_M : M^A \rightarrow A$ extracts results of type A from monadic values of type M^A . The use of a runner will agree with programmer's intuition if the runner obeys the laws of monad morphisms (see Definition 10.3.4.1); a runner θ_M must be a monad morphism $M \rightsquigarrow \text{Id}$ between M and the identity monad (Id). Heuristically, a runner "executes" or "runs" the effects of the monad M and delivers a pure value as a result.

How can we extract the result values from a monad stack that combines the effects of several monads? Let us try to create a runner $\theta_T : T_L^{M,A} \rightarrow A$ for the transformed monad T_L^M assuming that both L and M have known runners θ_L and θ_M .

As a first example, assume a compositional transformer such as $T_L^M = L \circ M$; this is the case for $L = \text{Reader}$ and some other monads. Given a runner θ_M , we may first run M 's effects by lifting θ_M to L ,

$$\theta_M^{\uparrow L} : L^{M^A} \rightarrow L^A \quad .$$

The result of type A can be now extracted by using L 's runner. So, we may write

$$\theta_T \triangleq \theta_M^{\uparrow L} ; \theta_L \quad .$$

If we first run the effects of L and then those of M , the result will be the same: $\theta_L^{\uparrow L} ; \theta_M = \theta_M ; \theta_L$ by the (assumed) naturality of θ_L . So, the runners are commutative, which is intuitively to be expected since the two runners work on different effects that are separately present in a value of type T_L^M .

A similar solution is available for composed-inside transformers $T_L^M = M \circ L$, which is the case for pass/fail monads L . We may run just the effects of one of the monads, or both if needed, and the runners are commutative.

Now consider a non-compositional transformer such as `StateT`:

$$T_{\text{State}}^{M,A} = S \rightarrow M^{A \times S} \quad .$$

In this case, we cannot run the effects of `State` using θ_{State} or the effects of M using $\theta_M^{\uparrow \text{State}}$: the types do not match.

$$\text{frun} : T_{\text{State}}^{M,A} \rightarrow M^A = (S \rightarrow M^{A \times S}) \rightarrow M^A \quad \text{vs.} \quad \theta_{\text{State}} : (S \rightarrow A \times S) \rightarrow A \quad ,$$

$$\text{brun} : T_{\text{State}}^{M,A} \rightarrow \text{State}^{S,A} = (S \rightarrow M^{A \times S}) \rightarrow S \rightarrow A \times S \quad \text{vs.} \quad \theta_M^{\uparrow \text{State}} : (S \rightarrow M^A \times S) \rightarrow S \rightarrow A \times S \quad .$$

We need to use new runners specially adapted to `StateT`, as shown in Section 14.1.3. The foreign runner (`frun`) is a function of θ_M and works for all foreign monads M . The base runner `brun`, however, is not a function of θ_{State} but has custom code. The commutativity of runners,

$$\text{frun}_{\text{State}}(\theta_M) ; \theta_{\text{State}}(i) \stackrel{?}{=} \text{brun}_{\text{State}}(i) ; \theta_M \quad ,$$

is no longer automatic and needs to be verified. Apply both sides to a value $t : T_{\text{State}}^{M,A}$:

$$\text{left-hand side : } t \triangleright \text{frun}_{\text{State}}(\theta_M) ; \theta_{\text{State}}(i) = t \triangleright (t \triangleright t ; \theta_M) ; (t \rightarrow i \triangleright t \triangleright \pi_1)$$

$$\text{compute composition : } = i \triangleright (t ; \theta_M) \triangleright \pi_1 = i \triangleright t ; \theta_M ; \pi_1 \quad ,$$

$$\text{left-hand side : } t \triangleright \text{brun}_{\text{State}}(i) ; \theta_M = t \triangleright (t \rightarrow i \triangleright t \triangleright \pi_1^{\uparrow M}) ; \theta_M = i \triangleright t ; \pi_1^{\uparrow M} ; \theta_M$$

$$\text{naturality of } \theta_M : = i \triangleright t ; \theta_M ; \pi_1 \quad .$$

Both sides are now equal, which proves the commutativity of runners for `StateT`.

These examples motivate the requirement that a monad transformer should come equipped with a base runner $\text{brun}_L : T_L^M \rightsquigarrow M$ and a foreign runner $\text{frun}_L(\theta_M) : T_L^M \rightsquigarrow L$ as long as the monads L and M have runners θ_L and θ_M . The foreign runner must be a function of M 's runner (θ_M). The effects of the base monad and of the foreign monad may be run in any order, and the results must be the same:

$$\text{brun}_L ; \theta_M = \text{frun}_L(\theta_M) ; \theta_L \quad .$$

Some monads M do not admit monad morphisms $M \rightsquigarrow \text{Id}$ but instead have useful monad morphisms $M \rightsquigarrow N$ with suitable target monads N . A monad morphism $M \rightsquigarrow N$ may partially run the effects of M and represent the results using N 's effects. For instance, the continuation monad can be mapped to a task monad such as `Future`; the `Option` monad can be mapped to the `Either` monad.

When a foreign monad M requires a “extended” runner of the form $\phi : M \rightsquigarrow N$ with a nontrivial target $N \not\cong \text{Id}$, we would like to be able to map the monad stacks accordingly as $T_L^M \rightsquigarrow T_L^N$, partially

running the effects of M but keeping the effects of L unchanged. It is natural to expect that the same foreign runner (frun_L) should be able to use ϕ instead of θ_M . We can then compute a map $\text{frun}_L(\phi) : T_L^M \rightsquigarrow T_L^N$ that runs the stacks as required.

When the base monad L requires a runner with a nontrivial target K , we would similarly need a monad morphism $T_L^M \rightsquigarrow T_K^M$, in order to partially run the effects of L . Since T_K^M depends on K nonparametrically, we cannot expect to have a universal base runner that uses $\phi : L \rightsquigarrow K$ as a parameter. The extended base runner $\text{brun}_{L,K}(\phi) : T_L^M \rightsquigarrow T_K^M$ will work for all foreign monads M but needs to have custom code specially adapted to the monads L and K .

The pass/fail monads (such as `Option` and `Try`) show one example where the extended base runner exists, because the monad transformers for pass/fail monads work by functor composition. If $L^A = \mathbb{1} + A$ and $K^A \triangleq E + A$ (where E is a fixed type representing error information), we can choose an arbitrary runner $\phi : L \rightsquigarrow K$ and compute the corresponding base runner as

$$\text{brun}_{L,K}(\phi) : (T_L^M \rightsquigarrow T_K^M) \triangleq \phi^{\uparrow M} \quad .$$

Apart from the laws of monad morphisms, do we need additional properties for all these runners? For intuition, recall the example with the continuation monad's runner (Section 10.1.9) that works by first converting `Cont` \rightsquigarrow `Future` and then `Future` \rightsquigarrow `Id`. Denote the two extended runners by

$$\phi_1 : \text{Cont}^{R,A} \rightarrow \text{Future}^A \quad , \quad \phi_2 : \text{Future}^A \rightarrow A \quad .$$

The runner for `Cont` is then defined by $\theta_{\text{Cont}} \triangleq \phi_1 \circ \phi_2$. Now consider how we would run a monad stack T_L^{Cont} . We may use the foreign runner with θ_{Cont} as

$$\text{frun}_L(\theta_{\text{Cont}}) : T_L^{\text{Cont}} \rightsquigarrow L \quad .$$

We could also use the extended runners to map $T_L^{\text{Cont}} \rightsquigarrow T_L^{\text{Future}}$ and then $T_L^{\text{Future}} \rightsquigarrow T_L^{\text{Id}} \cong L$:

$$\text{frun}_L(\phi_1) : T_L^{\text{Cont}} \rightsquigarrow T_L^{\text{Future}} \quad , \quad \text{frun}_L(\phi_2) : T_L^{\text{Future}} \rightsquigarrow L \quad .$$

Intuitively, we would expect that the composition of these two runners should be the same as when running T_L^{Cont} directly into L :

$$\text{frun}_L(\phi_1) \circ \text{frun}_L(\phi_2) \stackrel{!}{=} \text{frun}_L(\theta_{\text{Cont}}) = \text{frun}_L(\phi_1 \circ \phi_2) \quad .$$

The composition $\phi_1 \circ \phi_2$ of any two monad morphisms ϕ_1 and ϕ_2 is again a monad morphism (Statement 10.3.4.5). This suggests a **composition law**,

$$\text{frun}_L(\phi_1) \circ \text{frun}_L(\phi_2) \stackrel{!}{=} \text{frun}_L(\phi_1 \circ \phi_2) \quad . \tag{14.4}$$

This law is similar to the composition law of `map`. By analogy, we also impose the **identity law**:

$$\text{frun}_L(\text{id}^{M^A \rightarrow M^A}) \stackrel{!}{=} \text{id}^{T_L^{M,A} \rightarrow T_L^{M,A}} \quad .$$

Finally, let us consider the usage of a monad stack T_L^M within a functor block. Typically, we would lift some values of types L^A or M^A to the stack; at the end of the computation, we may run some (or all) of the effects. There are two possible cases: we lift a monadic value and then run the effects of the same monad; or we run the effects of the other monad.

In the first case, we lift a value $m : M^A$ to the stack T_L^M and then run the effects of M using $\text{frun}_L(\theta_M)$, getting a value of type L^A . The result is a transformation

$$\text{flift} \circ \text{frun}_L(\theta_M) : M^A \rightarrow L^A \quad .$$

Note that both `flift` and `frun` are monad morphisms, so their composition is again a monad morphism $M \rightsquigarrow L$. Since M is an arbitrary foreign monad, we cannot expect the monad L to describe

any effects of the foreign monad M . So, the result value of type L^A must have an empty L -effect, that is, it must be created from a value a^A via pu_L . The only way to obtain a value a^A is by applying a runner θ_M to the value $m : M^A$. So, we derived the “purity” law:

$$m \triangleright \text{flift} \triangleright \text{frun}_L(\theta_M) = m \triangleright \theta_M \triangleright \text{pu}_L \quad , \quad \text{flift} ; \text{frun}_L(\theta_M) = \theta_M ; \text{pu}_L \quad .$$

The analogous purity law for blift and brun is

$$\text{blift} ; \text{brun} = \theta_L ; \text{pu}_M \quad .$$

In the second case, we lift a foreign value $m : M^A$ to the stack, obtaining a value $t \triangleq \text{flift}(m)$, and then apply the *base* runner to t (running just the effects of L). This yields a monad morphism $M^A \rightarrow M^A$. We expect this to be an identity function because the monad M is arbitrary, and the code of flift and brun must be generic in M . This requirement is written as the law

$$m \triangleright \text{flift} \triangleright \text{brun} \stackrel{!}{=} m \quad , \quad \text{flift} ; \text{brun} \stackrel{!}{=} \text{id} \quad ,$$

called the **non-degeneracy law** of flift and brun . The law requires flift and brun to preserve all information about M -effects. The analogous non-degeneracy law for blift and brun is

$$\text{blift} ; \text{frun}_L(\theta_M) = \text{id} \quad .$$

14.2.3 Category-theoretic properties of lifts and runners

We have derived many laws of lifts and runners from heuristic considerations. Let us turn to category theory for guidance about creating a simplified and coherent picture of those laws.

The composition law (14.4) suggests that frun is a lifting from monad morphisms $M \rightsquigarrow N$ to monad morphisms $T_L^M \rightsquigarrow T_L^N$. In the language of category theory, a monad morphism is a morphism in the category of monads. The category of monads has all possible monads (M, N, \dots) as its objects and all possible monad morphisms (functions of type $M \rightsquigarrow N$) as its morphisms.

Note that frun obeys the laws of identity and composition; these are the laws of a (categorical) functor. Looking at the law (14.4), we find that frun maps morphisms between monads $(M, N, \text{etc.})$ to morphisms between monad stacks with base L (such as $T_L^M, T_L^N, \text{etc.}$).

So, let us consider a *category of monad stacks* with a fixed base monad L . The objects of that category are monad stacks $T_L^M, T_L^N, \text{etc.}$, all with the same base monad L . The morphisms of that category are monad morphisms between those monad stacks. (Since monad stacks are themselves monads, the category of monad stacks is a sub-category of the category of monads.)

We can now observe that frun plays the role of the map function for a (categorical) functor from the category of monads to the category of monad stacks with base L . That functor (which we may denote by T_L^\bullet) maps each monad M to the stack T_L^M , and each monad morphism $\phi : M \rightsquigarrow N$ to a monad morphism $\text{frun}_L(\phi) : T_L^M \rightsquigarrow T_L^N$. We may view the functor T_L^\bullet as the “bare” monad transformer, not yet applied to a foreign monad M .

Note that the functor T_L maps the monad M to the stack T_L^M ; it maps objects of one category to objects of another. It is not a *function* from values of type M^A to values of type $T_L^{M,A}$. That function is flift . What is the role of flift in the category of monads? The function flift works in the same way for any foreign monad M because the code of flift is *parametric* in the type constructor M . So, let us look at the relationship between the functions flift applied to different foreign monads.

Assume M and N are some monads for which a monad morphism $\phi : M^A \rightsquigarrow N^A$ exists. We may use flift to transform M^A into $T_L^{M,A}$ or to transform N^A into $T_L^{N,A}$. But we can also obtain a monad morphism $T_L^M \rightsquigarrow T_L^N$ as $\text{frun}_L(\phi)$. All these monad morphisms can be drawn as a square diagram.

$$\begin{array}{ccc} M^A & \xrightarrow{\text{flift}} & T_L^{M,A} \\ \phi \downarrow & & \downarrow \text{frun}_L(\phi) \\ N^A & \xrightarrow{\text{flift}} & T_L^{N,A} \end{array}$$

The diagram will be commutative if the following **monadic naturality law** holds:

$$\text{flift}_L^{M,A} ; \text{frun}_L(\phi : M^A \rightsquigarrow N^A) = \phi ; \text{flift}_L^{N,A} \quad . \quad (14.5)$$

Is it reasonable to require this law? To aid intuition, consider three special cases where the law is simplified, namely $M = \text{Id}$, $N = \text{Id}$, or $M = N$.

When $M = \text{Id}$, we must choose ϕ as the monad morphism $\phi \triangleq \text{pu}_N : \text{Id} \rightsquigarrow N$. (Exercise 14.9.2.2 shows that there are no other monad morphisms between Id and N .) The monad stack T_L^M is simplified to $T_L^{\text{Id}} \cong L$. The function $\text{flift}_L^{M,A}$ has the type signature $\text{Id}^A \rightarrow T_L^{M,A}$, which is the same as $A \rightarrow L^A$. For a general monad L , there is no other function of this type except pu_L ; so, $\text{flift}_L^{\text{Id}}$ must be equal to pu_L . The monad morphism $\text{frun}_L(\phi)$ has type signature $T_L^M \rightsquigarrow T_L^N$, which is the same as $L \rightsquigarrow T_L^N$, the type of blift . Since the monad N is arbitrary, it is reasonable to assume that

$$\text{frun}_L(\phi) = \text{frun}_L(\text{pu}_N) = \text{blift}_L^N \quad .$$

In this way, the monadic naturality law (14.5) gives the equation

$$\text{pu}_L \circ \text{blift}_L^N = \text{pu}_N \circ \text{flift}_L^N \quad .$$

This equation is satisfied due to the identity laws of lifts:

$$\text{pu}_L \circ \text{blift}_L^N = \text{pu}_{T_L^N} \quad , \quad \text{pu}_N \circ \text{flift}_L^N = \text{pu}_{T_L^N} \quad .$$

When $N = \text{Id}$, we choose $\phi = \theta_M$, assuming that M has a runner. The monad stack T_L^N is simplified to $T_L^{\text{Id}} \cong L$. The function flift_L^N has type signature $A \rightarrow L^A$ and is equal to pu_L . So, we get:

$$\text{flift}_L^M \circ \text{frun}_L(\theta_M) = \theta_M \circ \text{pu}_L \quad .$$

This is the “purity” law of flift that we derived earlier.

When $M = N$, we may choose $\phi = \text{id}$. The monadic naturality law (14.5) becomes

$$\text{flift} \circ \text{frun}_L(\text{id}) = \text{flift} \quad .$$

This equation holds as long as frun satisfies the identity law.

So, we found that the law (14.5) agrees with previously derived laws and, moreover, covers one of the laws of runners. The interpretation of the monadic naturality law becomes clear from viewing frun as a lifting (“map”) corresponding to the functor T_L in the category of monads. Denoting that lifting temporarily by $\uparrow T_L$, we rewrite the law (14.5) as

$$\text{flift} \circ \phi^{\uparrow T_L} = \phi \circ \text{flift} \quad .$$

This is a standard form of a naturality law if we view $\text{flift}^M : M \rightsquigarrow T_L^M$ as a *natural transformation* between the identity functor $\text{Id}^M \triangleq M$ (in the category of monads) to the functor T_L^M , the monad transformer functor with the base monad L . The law expresses the property that the code of flift works in the same way for all foreign monads M . We call it a “monadic” naturality law because $\text{flift}_L^{M,A} : M^A \rightarrow T_L^{M,A}$ also satisfies the ordinary naturality law with respect to the type parameter A , for a fixed monad M .

What other runners have monadic naturality properties? Consider the base runner, $\text{brun}_L : T_L^M \rightsquigarrow M$, assuming it exists for the chosen base monad L . Viewed in the category of monads, this function looks like a natural transformation between the functor T_L^M and the identity functor (going in the opposite way compared with flift). The corresponding **monadic naturality law** is

$$\text{brun}_L^M \circ \phi^{M \rightsquigarrow N} = \phi^{\uparrow T_L} \circ \text{brun}_L^N = \text{frun}_L(\phi) \circ \text{brun}_L^N \quad .$$

To see whether it is reasonable to assume this law, let us derive some consequences using specific choices of M and N .

With $M = \text{Id}$, the function ϕ must be $\phi = \text{pu}_N$ (Exercise 14.9.2.2), while $\text{frun}_L(\phi) : T_L^{\text{Id}} \rightsquigarrow T_L^N$ is equal to the base lift function (blift_L^N) . We also have $\text{brun}_L^{\text{Id}} : T_L^{\text{Id}} \rightsquigarrow \text{Id}$, so this function is equal to L 's runner ($\text{brun}_L^{\text{Id}} = \theta_L$). The resulting equation is the purity law of blift and brun :

$$\theta_L ; \text{pu}_N = \text{blift}_L^N ; \text{brun}_L^N .$$

With $N = \text{Id}$, the function ϕ must be a runner (θ_M), while $\text{brun}_L^{\text{Id}} = \theta_L$ as before. We get

$$\text{brun}_L^M ; \theta_M = \text{frun}_L(\theta_M) ; \theta_L .$$

This is the commutativity law of runners we obtained earlier.

So, the monadic naturality of brun covers two of the previously formulated laws.

The monadic naturality of frun is equivalent to frun 's composition law:

$$\text{frun}_L(\phi : M \rightsquigarrow N) ; \underline{\chi \uparrow T_L} = \text{frun}_L(\phi) ; \text{frun}_L(\chi) = \text{frun}_L(\phi ; \chi) .$$

To summarize, we are considering a functor T_L in the category of monads, such that a natural transformation $\text{flift}_L : \text{Id}^M \rightsquigarrow T_L^M$ exists (in the sense of the category of monads, i.e., parametric in the monad M).

Ordinary functors F^A having a natural transformation $\text{pu}_F : \text{Id}^A \rightarrow F^A$ are *pointed* functors; those with a natural transformation $\text{ex}_F : F^A \rightarrow \text{Id}^A$ are co-pointed. Naturality forces the non-degeneracy law $\text{pu}_F ; \text{ex}_F = \text{id}$. So, by analogy, we see that the functor T_L is pointed in the category of monads. If the transformer T_L has a base runner, the functor T_L will be also *co-pointed*, with $\text{brun}_L : T_L^{\bullet} \rightsquigarrow \text{Id}^{\bullet}$ being a natural transformation in the category of monads.

It is remarkable that all the laws of monad transformers can be derived from a single (but more abstract) definition: a monad transformer T is an *arbitrary* pointed and (optionally) co-pointed functor in the category of monads. Such a functor T maps monads $M, N, \text{etc.}$, to $T^M, T^N, \text{etc.}$ Given such a T , we first define the base monad as the image of the identity monad:

$$L \triangleq T^{\text{Id}} .$$

The definition of T already defines functions flift , frun , and brun as the natural transformations

$$\text{flift}_T^M : \text{Id}^M \rightsquigarrow T^M , \quad \text{brun}_T^M : T^M \rightsquigarrow \text{Id}^M , \quad \text{frun}_T^{M,N} : (M \rightsquigarrow N) \rightarrow T^M \rightsquigarrow T^N .$$

The functor laws and the monadic naturality laws will enforce all the laws of monad transformers.

Is the non-degeneracy law ($\text{flift}_T ; \text{brun}_T = \text{id}$) a consequence of monadic naturality of brun ? This may be the case, but this book does not have a proof. The composition of two natural transformations is again a natural transformation; this is true in any category, including the category of monads. So, the composition $\text{flift} ; \text{brun}$ has to be a natural transformation of type $\text{Id}^{\bullet} \rightsquigarrow \text{Id}^{\bullet}$. To show that this natural transformation is the identity function, we need to prove that the category of monads admits no other natural transformations between identity functors except identity transformations. A natural transformation of type $\text{Id}^{\bullet} \rightsquigarrow \text{Id}^{\bullet}$ means a family of monad morphisms $\varepsilon^M : M \rightsquigarrow M$ that work the same way for every monad M . One such morphism is the identity function, $\varepsilon = \text{id}$. However, it is not clear (Problem 14.9.2.13) whether there exist non-identity monad morphisms $M \rightsquigarrow M$ that work the same way for every monad M . For instance, Exercise 14.9.2.4 shows a failed attempt to define such a monad morphism.

Since we have not shown that the non-degeneracy law is an automatic consequence of monadic naturality, we will impose this law separately whenever the base runner exists (which is not the case for all base monads L).

14.2.4 Summary of the laws of monad transformers

We can now propose a formal definition: A **monad transformer** $T_L^{M,A}$ is a functor with a type parameter A and a monad parameter M , such that the following laws hold:

1. **Monad construction law:** $T_L^{M,\bullet}$ is a lawful monad for any monad M . In other words, the transformed monad $T_L^{M,\bullet}$ has methods pu_T and ftn_T that satisfy the monad laws.
2. **Identity law:** $T_L^{\text{Id},\bullet} \cong L^\bullet$ via a monad isomorphism, where Id is the identity monad, $\text{Id}^A \triangleq A$.
3. **Lifting law:** There is a function $\text{flift}_L : M^A \rightarrow T_L^{M,A}$ (in a shorter notation, $\text{flift}_L : M \rightsquigarrow T_L^M$) that works for any monad M and obeys the laws of monad morphisms.
4. **Runner laws:** There is a “foreign runner”⁴ $\text{frun}_L(\phi)$ such that for any monads M, N and any monad morphism $\phi : M \rightsquigarrow N$, the function $\text{frun}(\phi) : T_L^M \rightsquigarrow T_L^N$ is a monad morphism. The function frun lifts monad morphisms from $M \rightsquigarrow N$ to $T_L^M \rightsquigarrow T_L^N$ and must satisfy the corresponding **functor laws**:

$$\text{frun}(\text{id}) = \text{id} \quad , \quad \text{frun}(\phi) \circ \text{frun}(\chi) = \text{frun}(\phi \circ \chi) \quad .$$

It follows from the identity law $T_L^{\text{Id}} \cong L$ that the base monad L can be lifted into T_L^M : Setting $\phi = \text{pu}_M : \text{Id} \rightsquigarrow M$, we obtain

$$\text{frun}(\text{pu}_M) : T_L^{\text{Id}} \rightsquigarrow T_L^M = L \rightsquigarrow T_L^M.$$

This function is called the **base lift**, $\text{blift} \triangleq \text{frun}(\text{pu}_M) : L^A \rightarrow T_L^{M,A}$. The base lift automatically satisfies the non-degeneracy law,

$$\text{blift} \circ \text{frun}(\phi^{M \rightsquigarrow \text{Id}}) = \text{id} \quad ,$$

for any monad morphism $\phi : M \rightsquigarrow \text{Id}$, because the left-hand side equals $\text{frun}(\text{pu}_M \circ \phi)$, and the composition law for monad morphisms gives $\text{pu}_M \circ \phi = \text{pu}_{\text{Id}} = \text{id}$.

5. **Base runner laws:** There is a **base runner**, $\text{brun} : T_L^M \rightsquigarrow M$, must be defined for any monad M . The base runner must also satisfy the **non-degeneracy law**,

$$\text{flift} \circ \text{brun} = \text{id} \quad .$$

Since the monad transformer is specific to the base monad L and does not support an arbitrary other base monad, there are in general no functor laws for brun , unlike frun . Some monads have a general base runner, $\text{brun}(\theta_L)$, parameterized by an arbitrary runner $\theta_L : L \rightsquigarrow \text{Id}$. If so, the function $\text{brun}(\theta_L)$ must obey the base runner laws for any fixed θ_L .

6. **Monadic naturality laws:** The functions flift and brun must satisfy the monadic naturality law with respect to the monad parameter M . For an arbitrary monad morphism $\phi : M \rightsquigarrow N$,

$$\text{monadic naturality of flift} : \text{flift}^M \circ \text{frun}(\phi^{M \rightsquigarrow N}) = \phi^{M \rightsquigarrow N} \circ \text{flift}^N \quad ,$$

$$\text{monadic naturality of brun} : \text{brun}^M \circ \phi^{M \rightsquigarrow N} = \text{frun}(\phi^{M \rightsquigarrow N}) \circ \text{brun}^N \quad .$$

In total, we found 18 laws for monad transformers. Are all these laws necessary?

The main use of the laws is to verify correctness of the code. The next section shows some examples of incorrect implementations of monad transformers and indicates the laws that are violated.

⁴This function is called `hoist` in Haskell standard libraries and in `scalaz`.

14.2.5 Examples of trivially incorrect monad transformers

The laws of monad transformers guarantee that the transformed monad is able to represent, without loss of information, the operations of the base monad as well as the operations of the foreign monad. If some of these laws are omitted, we risk accepting a transformer that does not work correctly although it has the methods with the required type signatures.

The simplest example of an incorrect monad transformer is obtained by defining the new monad to be the unit monad, $T_L^{M,A} \triangleq \mathbb{1}$, for any monads L and M . This “transformer” is clearly useless: it cannot possibly describe the effects of L and M , because the methods of the unit monad discard *all* information and return 1. However, the type constructor $T_L^{M,A}$ has all the required methods pu_T , ftn_T , flift_L , frun_L , and brun_L (they are constant functions returning 1). All these functions are automatically monad morphisms, since a function from any monad to the unit monad is always a monad morphism. So, the fake “transformer” actually satisfies many of the monad transformer laws! However, the identity law $T_L^{\text{Id}} \cong L$ and the non-degeneracy law ($\text{flift} ; \text{brun} = \text{id}$) are violated since $T_L^{\text{Id}} = \mathbb{1} \not\cong L$ and $\text{flift} ; \text{brun} = (_ \rightarrow 1) \neq \text{id}$. For this reason, the unit monad is not a lawful monad transformer.

This simple example demonstrates the importance of the monad transformer laws. We could be given a fake implementation of a “transformer” that appears to have all the methods with the correct type signatures but, instead of a bigger monad, constructs a unit monad dressed up as a type constructor $T_L^{M,A}$. The only way for us to detect the fraud is to find that the identity law and the non-degeneracy law are violated.

Other examples of incorrect “transformers” violating some of the laws are $T_L^M \triangleq L$ (no lifting law) and $T_L^M \triangleq M$ (no identity law).

In these cases, it is intuitively clear that the transformer definitions are incorrect because the information about either L or M is missing in T_L^M . A potentially working definition of T_L^M must be a type constructor that somehow combines both L and M . Many such definitions are possible, but few will satisfy the monad transformer laws.

14.2.6 Examples of failure to define a generic monad transformer

It appears to be impossible to define T_L^M as a generic construction that works in the same way for all monads L and M . We will now consider a few ways of combining the type constructors L and M in a way that is independent of their structure. In all these cases, we will find that some of the monad transformer laws are violated.

General ways of combining two type constructors L^\bullet and M^\bullet are functor composition L^{M^\bullet} or M^{L^\bullet} , disjunction $L^\bullet + M^\bullet$, and product $L^\bullet \times M^\bullet$.

Functor composition A general way of combining two type constructors L^\bullet and M^\bullet is the functor composition L^{M^\bullet} or M^{L^\bullet} . However, the functor composition works only for certain monads and only in a certain order; so it cannot work as a generic monad transformer. A simple counterexample is $L^A \triangleq \mathbb{1} + A$ and $M^A \triangleq A \times A$ where $M^{L^A} = (\mathbb{1} + A) \times (\mathbb{1} + A)$ is a monad but $L^{M^A} = \mathbb{1} + A \times A$ is not (see Exercise 10.2.9.8). Another counterexample is the `State` monad, $\text{State}_S^A \triangleq S \rightarrow S \times A$, for which we have already shown that $\mathbb{1} + \text{State}_S^A$ is not a monad and $\text{State}_S^{Z \rightarrow A}$ is not a monad (see Section ???). In other words, the `State` monad does not compose with arbitrary monads M in either order.

Functor product We know from Chapter 10 that the functor product $M_1^A \times M_2^A$ is a monad when M_1 and M_2 are themselves monads. However, the product $M_1^A \times M_2^A$ describes two separate computations with two separate effects. Instead, we need a single computation with a combined effect. Formally, we find that there is no naturally defined $\text{flift} : M^\bullet \rightsquigarrow L^\bullet \times M^\bullet$ because we cannot create values of type L^A out of values of type M^A for arbitrary monads L and M .

Functor disjunction The functor disjunction $L^\bullet + M^\bullet$ is in general not a monad when L and M are arbitrary monads. An immediate counterexample is found by using two `Reader` monads, $L^A \triangleq P \rightarrow A$ and $M^A \triangleq Q \rightarrow A$. The disjunction $(P \rightarrow A) + (Q \rightarrow A)$ is not a monad (Exercise 10.2.9.7). But even if

$L^\bullet + M^\bullet$ were a monad, the result would be a choice between two different effectful computations, not a single computation with a combined effect.

Using the free monad The functor composition L^{M^\bullet} and the disjunction $L^\bullet + M^\bullet$ may not always be monads, but they are always functors. So we can make monads out of them, by using the free monad construction. We get $\text{Free}^{L^{M^\bullet}}$, the free monad over L^{M^\bullet} , and $\text{Free}^{L^\bullet + M^\bullet}$, the free monad over $L^\bullet + M^\bullet$. Many laws of the monad transformers are satisfied by these constructions. However, the identity laws fail because

$$\text{Free}^{L^{\text{Id}^\bullet}} \cong \text{Free}^{L^\bullet} \not\cong L \quad , \quad \text{Free}^{L^\bullet + \text{Id}^\bullet} \not\cong L \quad ,$$

and the lifting laws are also violated because $\text{flift} : M^A \rightarrow \text{Free}^{L^\bullet + M^\bullet, A}$ is not a monad morphism: it maps pu_M into a non-pure value of the free monad. Nevertheless, these constructions are not useless. Once we run the free monad into a concrete (non-free) monad, we could arrange to hide the violations of these laws, so that the monad laws hold for the resulting (non-free) monad.

“Monoidal convolution” The construction called “monoidal convolution” defines a new functor $L \star M$ via

$$(L \star M)^A \triangleq \exists P \exists Q. (P \times Q \rightarrow A) \times L^P \times M^Q \quad . \quad (14.6)$$

This formula can be seen as a combination of the co-Yoneda identities

$$L^A \cong \exists P. L^P \times (P \rightarrow A) \quad , \quad M^A \cong \exists Q. M^Q \times (Q \rightarrow A) \quad .$$

The functor product $L \times M$ is equivalent to

$$\begin{aligned} & L^A \times M^A \\ \text{co-Yoneda identities for } L^A \text{ and } M^A : & \cong \exists P. L^P \times \underline{(P \rightarrow A)} \times \exists Q. M^Q \times \underline{(Q \rightarrow A)} \\ \text{equivalence in Eq. (14.8) :} & \cong \exists P. \exists Q. L^P \times M^Q \times (P + Q \rightarrow A) \end{aligned} \quad (14.7)$$

where we used the type equivalence

$$(P \rightarrow A) \times (Q \rightarrow A) \cong P + Q \rightarrow A \quad . \quad (14.8)$$

If we (arbitrarily) replace $P + Q \rightarrow A$ by $P \times Q \rightarrow A$ in Eq. (14.7), we will obtain Eq. (14.6).

The monoidal convolution $L \star M$ always produces a functor since Eq. (14.6) is covariant in A . An example where the monoidal convolution fails to produce a monad transformer is $L^A \triangleq 1 + A$ and $M^A \triangleq R \rightarrow A$. We compute the functor $L \star M$ and establish that it is not a monad:

$$\begin{aligned} & (L \star M)^A \\ \text{definitions of } L, M, \star : & \cong \exists P \exists Q. \underline{(P \times Q \rightarrow A)} \times (1 + P) \times (R \rightarrow Q) \\ \text{curry the arguments, move quantifier :} & \cong \exists P. (1 + P) \times \underline{(Q \rightarrow P \rightarrow A)} \times (R \rightarrow Q) \\ \text{co-Yoneda identity with } \exists Q : & \cong \exists P. (1 + P) \times \underline{(R \rightarrow P \rightarrow A)} \\ \text{swap curried arguments :} & \cong \exists P. (1 + P) \times (P \rightarrow R \rightarrow A) \\ \text{co-Yoneda identity with } \exists P : & \cong 1 + (R \rightarrow A) \quad . \end{aligned}$$

This functor is not a monad (see, e.g., Exercise 10.2.9.8).

Codensity tricks***

- codensity monad over L^{M^\bullet} : $F^A \triangleq \forall B. (A \rightarrow L^{M^B}) \rightarrow L^{M^B}$ – no lift
- Codensity- L transformer: $\text{Cod}_L^{M,A} \triangleq \forall B. (A \rightarrow L^B) \rightarrow L^{M^B}$ – no lift
 - applies the continuation transformer to $M^A \cong \forall B. (A \rightarrow B) \rightarrow M^B$
- Codensity composition: $F^A \triangleq \forall B. (M^A \rightarrow L^B) \rightarrow L^B$ – not a monad
 - Counterexample: $M^A \triangleq R \rightarrow A$ and $L^A \triangleq S \rightarrow A$

14.2.7 Functor composition with transformed monads

Suppose we are working with a base monad L and a foreign monad M , and we have constructed the transformed monad T_L^M . In this section, let us denote the transformed monad simply by T .

A useful property of monad transformers is that the monad T adequately describes the effects of both monads L and M at the same time. Suppose we are working with a deeply nested type constructor involving many functor layers of monads L , M , and T such as

$$T^{M^T L^M A} = (T \circ M \circ T \circ L \circ M \circ L)^A .$$

The properties of the transformer allow us to convert this type to a single layer of the transformed monad T . In this example, we will have a natural transformation

$$(T \circ M \circ T \circ L \circ M \circ L)^A \rightarrow T^A .$$

To achieve this, we first use the methods `blift` and `flift` to convert each layer of L or M to a layer of T , lifting into functors as necessary. The result will be a number of nested layers of T . Second, we use `ftnT` as many times as necessary to flatten all nested layers of T into a single layer. The result will be a value of type T^A , as required.

14.2.8 Stacking two monads

Suppose we know the transformers T_P and T_Q for some given monads P and Q . We can transform Q with P and obtain a monad $R^A \triangleq T_P^{Q,A}$. What would be the monad transformer T_R for the monad R ?

A simple solution is to first transform the foreign monad M with T_Q , obtaining a new monad $T_Q^{M,\bullet}$, and then to transform that new monad with T_P . So the formula for the transformer T_R is

$$T_R^{M,A} = T_P^{T_Q^{M,\bullet},A} .$$

Here the monad $T_Q^{M,\bullet}$ was substituted into $T_P^{M,A}$ as the foreign monad M (not as the type parameter A). This way of composition is called **stacking** the monad transformers.

In Scala code, this “stacking” construction is written as

```
type RT[M, A] = PT[QT[M, ?], A]
```

The resulting monad is a **stack** of three monads P , Q , and M . The order of monads in the stack is significant since, in general, there will be no monad isomorphism between differently ordered stacks.

We will now show that the transformer T_R is lawful (satisfies the laws stated in Section 14.2.4), as long as both T_P and T_Q satisfy the same laws. To shorten the notation, we talk about a “monad T_P^M ” meaning the monad defined as $T_P^{M,\bullet}$ or, more verbosely, the monad $G^A \triangleq T_P^{M,A}$.

Monad construction law We need to show that T_P^M is a monad for any monad M . The monad construction law for T_Q says that T_Q^M is a monad. The monad construction law for T_P says that T_P^S is a monad for any monad S ; in particular, for $S = T_Q^M$. Therefore, $T_P^S = T_P^{T_Q^M}$ is a monad, as required.

Identity law We need to show that $T_P^{T_Q^{\text{Id}}}$ $\cong T_P^Q$ via a monad isomorphism. The identity law for T_Q says that $T_Q^{\text{Id}} \cong Q$ via a monad isomorphism. So, we already have a monad morphism $\phi : Q \rightsquigarrow T_Q^{\text{Id}}$ and its inverse, $\chi : T_Q^{\text{Id}} \rightsquigarrow Q$. The runner `frunP` for T_P can be applied to both ϕ and χ since they are monad morphisms. So we obtain two new monad morphisms,

$$\text{frun}_P(\phi) : T_P^Q \rightsquigarrow T_P^{T_Q^{\text{Id}}} ; \quad \text{frun}_P(\chi) : T_P^{T_Q^{\text{Id}}} \rightsquigarrow T_P^Q .$$

Are these two monad morphisms inverses of each other? To show this, we need to verify that

$$\text{frun}_P(\phi) ; \text{frun}_P(\chi) = \text{id} \quad , \quad \text{frun}_P(\chi) ; \text{frun}_P(\phi) = \text{id} \quad .$$

By the runner law for T_P , we have $\text{frun}_P(f) ; \text{frun}_P(g) = \text{frun}_P(f ; g)$ for any two monad morphisms f and g . We also have $\text{frun}_P(\text{id}) = \text{id}$ by the same law. So,

$$\begin{aligned} \text{frun}_P(\phi) ; \text{frun}_P(\chi) &= \text{frun}_P(\phi ; \chi) = \text{frun}_P(\text{id}) = \text{id} \quad , \\ \text{frun}_P(\chi) ; \text{frun}_P(\phi) &= \text{frun}_P(\chi ; \phi) = \text{frun}_P(\text{id}) = \text{id} \quad . \end{aligned}$$

We have indeed obtained a monad isomorphism between T_P^Q and $T_P^{T_Q^{\text{Id}}}$.

Lifting law We need to show that there exists a monad morphism $\text{lift}_R : M \rightsquigarrow T_P^{T_Q^M}$ for any monad M . The lifting law for T_Q gives a monad morphism $\text{lift}_Q : M \rightsquigarrow T_Q^M$. The lifting law for T_P can be applied to the monad T_Q^M , which gives a monad morphism

$$\text{lift}_P : T_Q^M \rightsquigarrow T_P^{T_Q^M} \quad .$$

The composition $\text{lift}_Q ; \text{lift}_P$ has the required type $M \rightsquigarrow T_P^Q$ and is a monad morphism by Statement 10.3.4.5.

Runner law We need to show that there exists a lifting

$$\text{frun}_R : (M \rightsquigarrow N) \rightarrow T_P^{T_Q^M} \rightsquigarrow T_P^{T_Q^N} \quad .$$

First, we have to define $\text{frun}_R(\phi)$ for any given $\phi : M \rightsquigarrow N$. We use the lifting law for T_Q to get a monad morphism

$$\text{frun}_Q(\phi) : T_Q^M \rightsquigarrow T_Q^N \quad .$$

Now we can apply the lifting law for T_P to this monad morphism and obtain

$$\text{frun}_P(\text{frun}_Q(\phi)) : T_P^{T_Q^M} \rightsquigarrow T_P^{T_Q^N} \quad .$$

This function has the correct type signature. So we can define

$$\text{frun}_R \triangleq \text{frun}_Q ; \text{frun}_P \quad .$$

It remains to prove that frun_R is a lawful lifting. We use the fact that both lift_P and lift_Q are lawful liftings; we need to show that their composition is also a lawful lifting. To verify the identity law of lifting, apply lift_R to an identity function $\text{id} : M \rightsquigarrow M$,

$$\begin{aligned} \text{frun}_R \left(\text{id}^{M \rightsquigarrow M} \right) &= \text{frun}_P \left(\underline{\text{frun}_Q(\text{id}^{M \rightsquigarrow M})} \right) \\ \text{identity law for } \text{frun}_Q : \quad &= \text{frun}_P \left(\text{id}^{T_Q^M \rightsquigarrow T_Q^M} \right) \\ \text{identity law for } \text{frun}_P : \quad &= \text{id} \quad . \end{aligned}$$

To verify the composition law of lifting, apply frun_R to a composition of two monad morphisms $\phi : L \rightsquigarrow M$ and $\chi : M \rightsquigarrow N$,

$$\begin{aligned} \text{frun}_R(\phi ; \chi) &= \text{frun}_P \left(\underline{\text{frun}_Q(\phi ; \chi)} \right) \\ \text{composition law for } \text{frun}_Q : \quad &= \underline{\text{frun}_P(\text{frun}_Q(\phi) ; \text{frun}_Q(\chi))} \\ \text{composition law for } \text{frun}_P : \quad &= \underline{\text{frun}_P(\text{frun}_Q(\phi))} ; \underline{\text{frun}_P(\text{frun}_Q(\chi))} \\ \text{definition of } \text{frun}_R : \quad &= \text{frun}_R(\phi) ; \text{frun}_R(\chi) \quad . \end{aligned}$$

Base runner law We need to show that for any monad morphism $\theta : T_P^Q \rightsquigarrow \text{Id}$ and for any monad M , there exists a monad morphism $\text{brun}_R(\theta) : T_P^{T_Q^M} \rightsquigarrow M$. To define this morphism for a given θ , we clearly need to use the base runners for T_P and T_Q . The base runner for T_Q has the type signature

$$\text{brun}_Q : (Q \rightsquigarrow \text{Id}) \rightarrow T_Q^M \rightsquigarrow M .$$

We can apply the base runner for T_P to T_Q^M as the foreign monad,

$$\text{brun}_P : (P \rightsquigarrow \text{Id}) \rightarrow T_P^{T_Q^M} \rightsquigarrow T_Q^M .$$

It is now clear that we could obtain a monad morphism $T_P^{T_Q^M} \rightsquigarrow M$ if we had some monad morphisms $\phi : P \rightsquigarrow \text{Id}$ and $\chi : Q \rightsquigarrow \text{Id}$,

$$\text{brun}_P(\phi) ; \text{brun}_Q(\chi) : T_P^{T_Q^M} \rightsquigarrow M .$$

However, we are only given a single monad morphism $\theta : T_P^Q \rightsquigarrow \text{Id}$. How can we compute ϕ and χ out of θ ? We can use the liftings $\text{blift}_P : P \rightsquigarrow T_P^Q$ and $\text{flift}_P : Q \rightsquigarrow T_P^Q$, which are both monad morphisms, and compose them with θ :

$$(\text{blift}_P ; \theta) : P \rightsquigarrow \text{Id} ; \quad (\text{flift}_P ; \theta) : Q \rightsquigarrow \text{Id} .$$

So we can define the monad morphism $\text{brun}_R(\theta)$ as

$$\begin{aligned} \text{brun}_R(\theta) &: T_P^{T_Q^M} \rightsquigarrow M , \\ \text{brun}_R(\theta) &\triangleq \text{brun}_P(\text{blift}_P ; \theta) ; \text{brun}_Q(\text{flift}_P ; \theta) . \end{aligned}$$

Since we have defined $\text{brun}_R(\theta)$ as a composition of monad morphisms, $\text{brun}_R(\theta)$ is itself a monad morphism by Statement 10.3.4.5.

To verify the non-degeneracy law of the base runner, $\text{flift}_R ; \text{brun}_R(\theta) = \text{id}$, we need to use the non-degeneracy laws for the base runners of T_P and T_Q , which are

$$\text{flift}_P ; \text{brun}_P(\chi : P \rightsquigarrow \text{Id}) = \text{id} , \quad \text{flift}_Q ; \text{brun}_Q(\psi : Q \rightsquigarrow \text{Id}) = \text{id} .$$

Then we can write

$$\begin{aligned} &\underline{\text{flift}_R ; \text{brun}_R(\theta)} \\ \text{expand definitions : } &= \text{flift}_Q ; \underline{\text{flift}_P ; \text{brun}_P(\text{blift}_P ; \theta)} ; \text{brun}_Q(\text{flift}_P ; \theta) \\ \text{non-degeneracy for } \text{brun}_P : &= \underline{\text{flift}_Q ; \text{brun}_Q(\text{flift}_P ; \theta)} \\ \text{non-degeneracy for } \text{brun}_Q : &= \text{id} . \end{aligned}$$

Monadic naturality laws The monadic naturality law of flift_R is

$$\text{flift}_R ; \text{frun}_R(\phi) = \phi ; \text{flift}_R .$$

We have defined $\text{flift}_R \triangleq \text{flift}_Q ; \text{flift}_P$, and we may assume that the monad naturality laws hold for flift_P and flift_Q . A composition of natural transformations is again a natural transformation; this holds for any category, including the category of monads. We can also verify this law directly:

$$\begin{aligned} \text{expect to equal } \phi ; \text{flift}_R : & \text{flift}_R ; \text{frun}_R(\phi) = \text{flift}_Q ; \underline{\text{flift}_P ; \text{frun}_P(\text{frun}_Q(\phi))} \\ \text{monadic naturality of } \text{flift}_P : &= \underline{\text{flift}_Q ; \text{frun}_Q(\phi)} ; \text{flift}_P \\ \text{monadic naturality of } \text{flift}_Q : &= \phi ; \text{flift}_Q ; \text{flift}_P = \phi ; \text{flift}_R . \end{aligned}$$

The monadic naturality of brun_R is verified similarly, assuming the same law for brun_P and brun_Q :

$$\begin{aligned} \text{expect to equal } \text{brun}_R(\theta) ; \phi : & \text{frun}_R(\phi) ; \text{brun}_R(\theta) = \underline{\text{frun}_P(\text{frun}_Q(\phi))} ; \text{brun}_P(\text{blift}_P ; \theta) ; \text{brun}_Q(\text{flift}_P ; \theta) \\ \text{same law for } \text{brun}_P : &= \underline{\text{brun}_P(\text{blift}_P ; \theta)} ; \underline{\text{frun}_Q(\phi)} ; \text{brun}_Q(\text{flift}_P ; \theta) \\ \text{same law for } \text{brun}_Q : &= \underline{\text{brun}_P(\text{blift}_P ; \theta)} ; \text{brun}_Q(\text{flift}_P ; \theta) ; \phi = \text{brun}_R(\theta) ; \phi . \end{aligned}$$

14.2.9 Stacking any number of monads

The monad transformer for T_P^Q can be applied to another monad K ; the result is the transformed monad

$$S^A \triangleq T_P^{T_Q^K A} = (P \otimes Q \otimes K)^A.$$

What is the monad transformer for the monad S ? Assuming that we know the monad transformer T_K , we could stack the transformers one level higher:

$$T_S^{M,A} \triangleq T_P^{T_Q^{T_K^M A}} = (P \otimes Q \otimes K \otimes M)^A .$$

This looks like a stack of four monads P , Q , K , and M . Note that the type parameter A is used as $T_P^{(\dots),A}$, that is, it belongs to the *outer* transformer $T_P^{(\dots),A}$.

We can now define a transformer stack for any number of monads P, Q, \dots, Z in a similar way,

$$T_S^{M,A} \triangleq T_P^{T_Q^{T_Z^{(\dots),A}}} = (P \otimes Q \otimes \dots \otimes Z \otimes M)^A . \quad (14.9)$$

The type parameter A will always remain at the outer transformer level, while the foreign monad M will be in the innermost nested position.

It turns out that T_S is a lawful monad transformer for *any* number of stacked monads. We can prove this by induction on the number of monads. In the previous section, we have derived the transformer laws for any *three* stacked monads (two monads P, Q within the transformer and one foreign monad M). Now we need to derive the same laws for a general transformer stack, such as that in Eq. (14.9). Let us temporarily denote by J the monad

$$J \triangleq T_Q^{(\dots), \text{Id}} = Q \otimes \dots \otimes Z \otimes \text{Id} \cong Q \otimes \dots \otimes Z ,$$

where we used the identity monad Id in the place normally taken by a foreign monad M . The monad J is a shorter transformer stack than S , so the inductive assumption tells us that the transformer laws already hold for the transformer T_J defined as

$$T_J^M \triangleq T_Q^{(\dots), T_Z^M} = (Q \otimes \dots \otimes Z) \otimes M .$$

Since both T_P^{\bullet} and T_J^{\bullet} are lawful transformers, their stacking composition $T_P^{T_J^{\bullet}}$ is also a lawful transformer (this was shown in the Section 14.2.8). In our notation, $T_S^{M,A} = T_P^{T_J^M A}$, and so we have shown that $T_S^{M,A}$ is a lawful transformer.

14.3 Monad transformers via functor composition: General properties

We have seen examples of monad transformers that work via functor composition, either as composed-inside or as composed-outside. The simplest examples are the `OptionT` transformer,

$$L^A \triangleq \mathbb{1} + A, \quad T_L^{M,A} \triangleq M^{L^A} = M^{\mathbb{1} + A} ,$$

which puts the base monad L *inside* the monad M , and the `ReaderT` transformer,

$$L^A \triangleq R \rightarrow A, \quad T_L^{M,A} \triangleq L^{M^A} = R \rightarrow M^A ,$$

which puts the base monad L *outside* the foreign monad M .

We can prove many properties of both kinds of monad transformers via a single derivation if we temporarily drop the distinction between the base monad and the foreign monad. We simply assume that two different monads, L and M , have a functor composition $T^\bullet \triangleq L^M$ that also happens to be a monad. Since the assumptions on the monads L and M are the same, the resulting properties of the composed monad T will apply equally to both kinds of monad transformers.

To interpret the results, we will assume that L is the base monad for the composed-outside transformers, and that M is the base monad for the composed-inside transformers. For instance, we will be able to prove the laws of liftings $L \rightsquigarrow T$ and $M \rightsquigarrow T$ regardless of the choice of the base monad.

What properties of monad transformers will *not* be derivable in this way? Monad transformers depend on the structure on the base monad, but not on the structure of the foreign monad; the transformer's methods `pure` and `flatten` are generic in the foreign monad. This is expressed via the monad transformer laws for the runners `run` and `brun`, which we will need to derive separately for each of the two kinds of transformers.

14.3.1 Motivation for the `swap` function

The first task is to show that the composed monad $T^\bullet \triangleq L^M$ obeys the monad laws. For this, we need to define the methods for the monad T , namely `pure` (short notation “`puT`”) and `flatten` (short notation “`ftnT`”), with the type signatures

$$\text{pu}_T : A \rightarrow L^{M^A} , \quad \text{ftn}_T : L^{M^{L^M}} \rightarrow L^{M^A} .$$

How can we implement these methods? *All we know* about L and M is that they are monads with their own methods `puL`, `ftnL`, `puM`, and `ftnM`. We can easily implement

$$\text{pu}_T \triangleq \text{pu}_M ; \text{pu}_L . \quad (14.10)$$

$$\begin{array}{ccc} A & \xrightarrow{\text{pu}_M} & M^A \\ & \searrow \text{pu}_T \triangleq & \downarrow \text{pu}_L \\ & & L^{M^A} \end{array}$$

It remains to implement `ftnT`. In the type $L^{M^{L^M}}$, we have two layers of the functor L and two layers of the functor M . We could use the available method `ftnL` to flatten the two layers of L if we could *somehow* bring these nested layers together. However, these layers are separated by a layer of the functor M . To show this layered structure in a more visual way, let us employ another notation for the functor composition,

$$(L \circ M)^A \triangleq L^{M^A} .$$

In this notation, the type signature for `flatten` is written as

$$\text{ftn}_T : L \circ M \circ L \circ M \rightsquigarrow L \circ M .$$

If we had $L \circ L \circ M \circ M$ here, we would have applied `ftnL` and flattened the two layers of the functor L . Then we would have flattened the remaining two layers of the functor M . How can we achieve this? The trick is to *assume* that we have a function called `swap` (short notation “`sw`”), which can interchange the order of the layers. The type signature of `swap` is

$$\text{sw} : M \circ L \rightsquigarrow L \circ M ,$$

which is equivalently written in a more verbose notation as

$$\text{sw}_{L,M}^A : M^{L^A} \rightarrow L^{M^A} .$$

If the `swap` operation were *somehow* defined for the two monads L and M , we could implement ftn_T by first swapping the order of the inner layers M and L as

$$L \circ M \circ L \circ M \rightsquigarrow L \circ L \circ M \circ M \quad ,$$

and then applying the `flatten` methods of the monads L and M . The resulting code for the function ftn_T and the corresponding type diagram are

$$\text{ftn}_T \triangleq \text{sw}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L} \quad . \quad (14.11)$$

$$\begin{array}{ccccc} L^{M^{L^M A}} & \xrightarrow{\text{sw}^{\uparrow L}} & L^{L^{M^M A}} & \xrightarrow{\text{ftn}_L} & L^{M^{M^A}} \\ & \searrow \text{ftn}_T \triangleq & & & \downarrow \text{ftn}_M^{\uparrow L} \\ & & & & L^{M^A} \end{array}$$

It turns out that in *both* cases (the composed-inside and the composed-outside transformers), the new monad's `flatten` method can be defined through the `swap` operation. For the two kinds of transformers, the type signatures of these functions are

$$\begin{aligned} \text{composed-inside : } & \text{ftn}_T : M^{L^{M^L A}} \rightarrow M^{L^A} \quad , \quad \text{sw}_{M,L}^A : L^{M^A} \rightarrow M^{L^A} \quad , \\ \text{composed-outside : } & \text{ftn}_T : L^{M^{L^M A}} \rightarrow L^{M^A} \quad , \quad \text{sw}_{L,M}^A : M^{L^A} \rightarrow L^{M^A} \quad . \end{aligned}$$

The difference between the operations `swap` and `sequence` There is a certain similarity between the `swap` operation introduced here and the `sequence` operation introduced in Chapter 12 for traversable functors. Indeed, the type signature of the `sequence` operation is

$$\text{seq} : L^{F^A} \rightarrow F^{L^A} \quad ,$$

where F is an arbitrary applicative functor (which could be M , since monads are applicative functors) and L is a traversable functor. However, the similarity stops here. The laws required for the `swap` operation to yield a monad T are different from the laws of traversable functors. In particular, if we wish M^L [•] to be a monad, it is insufficient to require the monad L to be a traversable functor. A simple counterexample is found with $L^A \triangleq A \times A$ and $M^A \triangleq 1 + A$. Both L and M are traversable (since they are polynomial functors); but their composition $Q^A \triangleq 1 + A \times A$ is not a monad (see Exercise 10.2.9.8).

Another difference between `swap` and `sequence` is that the `swap` operation needs to be generic in the foreign monad, which may be either L or M according to the type of the monad transformer; whereas `sequence` is always generic in the applicative functor F .

To avoid confusion, we call the function $\text{sw}_{L,M} : M^L \rightsquigarrow L^M$ “swap” rather than “sequence” in the context of monad transformers. Let us now find out what laws are required for the `swap` operation.⁵

14.3.2 Deriving the necessary laws for `swap`

The first law is that `swap` must be a natural transformation. Since `swap` has only one type parameter, there is one naturality law: for any function $f : A \rightarrow B$,

$$f^{\uparrow L \uparrow M} ; \text{sw}_{L,M} = \text{sw}_{L,M} ; f^{\uparrow M \uparrow L} \quad . \quad (14.12)$$

⁵The `swap` operation was used in the paper “Composing monads” (1993) by M. P. Jones and L. Duponcheel; see <http://web.cecs.pdx.edu/~mpj/pubs/composing.html>

$$\begin{array}{ccc}
 M^{L^A} & \xrightarrow{f^{\uparrow L \uparrow M}} & M^{L^B} \\
 \text{sw} \downarrow & & \downarrow \text{sw} \\
 L^{M^A} & \xrightarrow{f^{\uparrow M \uparrow L}} & L^{M^B}
 \end{array}$$

To derive further laws for `swap`, consider the requirement that the transformed monad T should satisfy the monad laws:

$$\begin{aligned}
 \text{pu}_T \circ \text{ftn}_T &= \text{id} , & \text{pu}_T^{\uparrow T} \circ \text{ftn}_T &= \text{id} , \\
 \text{ftn}_T^{\uparrow T} \circ \text{ftn}_T &= \text{ftn}_T \circ \text{ftn}_T .
 \end{aligned}$$

Additionally, T must satisfy the laws of a monad transformer. We will now discover the laws for `swap` that make the laws for ftn_T hold automatically, as long as ftn_T is derived from `swap` using Eq. (14.11).

We substitute Eq. (14.11) into the left identity law for ftn_T and simplify:

$$\begin{aligned}
 \text{id} &= \text{pu}_T \circ \text{ftn}_T \\
 \text{replace } \text{ftn}_T \text{ using Eq. (14.11)} : &= \text{pu}_T \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \\
 \text{replace } \text{pu}_T \text{ using Eq. (14.10)} : &= \text{pu}_M \circ \text{pu}_L \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \\
 \text{naturality of } \text{pu}_L : &= \text{pu}_M \circ \text{sw} \circ \text{pu}_L \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \\
 \text{left identity law for } L : &= \text{pu}_M \circ \text{sw} \circ \text{ftn}_M^{\uparrow L} . \tag{14.13}
 \end{aligned}$$

How could the last expression in Eq. (14.13) be equal to id ? We know nothing about the `pure` and `flatten` methods of the monads L and M , except that these methods satisfy their monad laws. We could satisfy the law in Eq. (14.13) if we somehow reduce that expression to

$$\text{pu}_M^{\uparrow L} \circ \text{ftn}_M^{\uparrow L} = (\text{pu}_M \circ \text{ftn}_M)^{\uparrow L} = \text{id} .$$

This will be possible only if we are able to interchange the order of function compositions with `sw` and eliminate `swap` from the expression. So, we must require the “outer-identity law” for `swap`,

$$\text{pu}_M \circ \text{sw} = \text{pu}_M^{\uparrow L} . \tag{14.14}$$

$$\begin{array}{ccc}
 L^A & \xrightarrow{\text{pu}_M} & M^{L^A} \\
 & \searrow \text{pu}_M^{\uparrow L} & \downarrow \text{sw} \\
 & & L^{M^A}
 \end{array}$$

Intuitively, this law says that a pure layer of the monad M remains pure after interchanging the order of layers with `swap`.

With this law, we can finish the derivation in Eq. (14.13) as

$$\begin{aligned}
 \text{outer-identity law for } \text{sw} : & \text{pu}_M \circ \text{sw} \circ \text{ftn}_M^{\uparrow L} \\
 &= \text{pu}_M^{\uparrow L} \circ \text{ftn}_M^{\uparrow L} \\
 \text{functor composition law for } L : & (\text{pu}_M \circ \text{ftn}_M)^{\uparrow L} \\
 \text{left identity law for } M : & = \text{id}^{\uparrow L} \\
 \text{functor identity law for } L : & = \text{id} .
 \end{aligned}$$

So, the M -identity law for `swap` entails the left identity law for T .

In the same way, we motivate the “inner-identity” law for `swap`,

$$\text{pu}_L^{\uparrow M} ; \text{sw} = \text{pu}_L \quad . \quad (14.15)$$

$$\begin{array}{ccc} M^A & \xrightarrow{\text{pu}_L^{\uparrow M}} & M^{L^A} \\ & \searrow \text{pu}_L & \downarrow \text{sw} \\ & & L^{M^A} \end{array}$$

This law expresses the idea that a pure layer of the functor L remains pure after swapping the order of layers.

Assuming this law, we can derive the right identity law for T :

$$\begin{aligned} \text{pu}_T^{\uparrow T} ; \text{ftn}_T & \\ \left(\text{by definition, } f^{\uparrow T} \triangleq f^{\uparrow M \uparrow L} \right) : & = (\text{pu}_T)^{\uparrow M \uparrow L} ; \text{ftn}_T \\ \text{definitions of } \text{pu}_T \text{ and } \text{ftn}_T : & = \text{pu}_M^{\uparrow M \uparrow L} ; \underline{\text{pu}_L^{\uparrow M \uparrow L} ; \text{sw}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L}} \\ \text{inner-identity law for } \text{sw}, \text{ under }^{\uparrow L} : & = \text{pu}_M^{\uparrow M \uparrow L} ; \underline{\text{pu}_L^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L}} \\ \text{right identity law for } L : & = \text{pu}_M^{\uparrow M \uparrow L} ; \text{ftn}_M^{\uparrow L} = \underline{(\text{pu}_M^{\uparrow M} ; \text{ftn}_M)^{\uparrow L}} \\ \text{right identity law for } M : & = \text{id}^{\uparrow L} = \text{id} \quad . \end{aligned}$$

Deriving the monad associativity law for T ,

$$\text{ftn}_T^{\uparrow T} ; \text{ftn}_T = \text{ftn}_T ; \text{ftn}_T \quad ,$$

turns out to require *two* further laws for `swap`. Let us see why.

Substituting the definition of ftn_T into the associativity law, we get

$$\begin{aligned} & (\text{sw}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L})^{\uparrow M \uparrow L} ; \text{sw}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L} \\ & = \text{sw}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L} ; \text{sw}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L} \quad . \end{aligned} \quad (14.16)$$

The only hope of proving this law is being able to interchange ftn_L as well as ftn_M with sw . In other words, the `swap` function should be in some way adapted to the `flatten` methods of both L and M .

Let us look for such interchange laws. One possibility is to have a law involving $\text{ftn}_M ; \text{sw}$, which is a function of type $M^{M^{L^A}} \rightarrow L^{M^A}$ or, in another notation, $M \circ M \circ L \rightsquigarrow L \circ M$. This function first flattens the two adjacent layers of M , obtaining $M \circ L$, and then swaps the two remaining layers, moving the L layer outside. Let us think about what law could exist for this kind of transformation. It is plausible that we may obtain the same result if we first swap the layers twice, so that the L layer moves to the outside, obtaining $L \circ M \circ M$, and then flatten the two inner M layers. Writing this assumption in code, we obtain the “outer-interchange” law

$$\text{ftn}_M ; \text{sw} = \text{sw}^{\uparrow M} ; \text{sw} ; \text{ftn}_M^{\uparrow L} \quad . \quad (14.17)$$

$$\begin{array}{ccccc} & M^{M^{L^A}} & \xrightarrow{\text{ftn}_M} & M^{L^A} & \\ \text{sw}^{\uparrow M} \swarrow & \nearrow & & & \downarrow \text{sw} \\ M^{L^{M^A}} & \xrightarrow{\text{sw}} & L^{M^{M^A}} & \xrightarrow{\text{ftn}_M^{\uparrow L}} & L^{M^A} \end{array}$$

The analogous “inner-interchange” law involving two layers of L and a transformation $M \circ L \circ L \rightsquigarrow L \circ M$ is written as

$$\text{ftn}_L^{\uparrow M} ; \text{sw} = \text{sw} ; \text{sw}^{\uparrow L} ; \text{ftn}_L \quad . \quad (14.18)$$

$$\begin{array}{ccccc}
 & & M^{L^A} & \xrightarrow{\text{ftn}_L^M} & M^{L^A} \\
 & \swarrow & & & \downarrow \swarrow \\
 L^{M^A} & \xrightarrow{\text{ftn}_L^M} & L^{L^M} & \xrightarrow{\text{ftn}_L} & L^{M^A}
 \end{array}$$

At this point, we have simply written down these two interchange laws, hoping that they will help us derive the associativity law for T . We will now verify that this is indeed so.

Both sides of the law in Eq. (14.16) involve compositions of several `flattens` and `swaps`. The heuristic idea of the proof is to use various laws to move all `flattens` to right of the composition, while moving all `swaps` to the left. In this way we will bring both sides of Eq. (14.16) to a similar form, hoping to prove that they are equal.

We begin with the right-hand side of Eq. (14.16) since it is simpler than the left-hand side, and look for ways of using the interchange laws. At every step of the calculation, there happens to be only one place where some law can be applied:

$$\begin{aligned}
 & \text{composition for } L : \quad \text{sw}^{\uparrow L} ; \text{ftn}_L ; \underline{\text{ftn}_M^L} ; \text{sw}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^L \\
 & \text{outer-interchange for sw} : \quad = \text{sw}^{\uparrow L} ; \text{ftn}_L ; \underline{(\text{ftn}_M ; \text{sw})^{\uparrow L}} ; \text{ftn}_L ; \text{ftn}_M^L \\
 & \text{composition for } L : \quad = \text{sw}^{\uparrow L} ; \text{ftn}_L ; \underline{\text{sw}^{\uparrow M} \text{ftn}_M^L} ; \text{sw}^{\uparrow L} ; \underline{\text{ftn}_M^L} ; \text{ftn}_L ; \text{ftn}_M^L \\
 & \text{naturality of ftn}_L : \quad = \text{sw}^{\uparrow L} ; \text{ftn}_L ; \underline{(\text{sw}^{\uparrow M} ; \text{sw})^{\uparrow L}} ; \text{ftn}_L ; \text{ftn}_M^L ; \text{ftn}_M^L \\
 & \text{naturality of ftn}_L : \quad = \text{sw}^{\uparrow L} ; \underline{(\text{sw}^{\uparrow M} ; \text{sw})^{\uparrow L}} ; \text{ftn}_L ; \text{ftn}_L ; \underline{\text{ftn}_M^L} ; \text{ftn}_M^L .
 \end{aligned}$$

Now all `swaps` are on the left and all `flattens` on the right of the expression.

Transform the right-hand side of Eq. (14.16) in the same way as

$$\begin{aligned}
 & \text{functor composition} : \quad \left(\text{sw}^{\uparrow L} ; \text{ftn}_L ; \underline{\text{ftn}_M^L} \right)^{\uparrow M \uparrow L} ; \text{sw}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^L \\
 & \text{naturality of sw} : \quad = \left(\text{sw}^{\uparrow L} ; \text{ftn}_L \right)^{\uparrow M \uparrow L} ; \underline{(\text{ftn}_M^L ; \text{sw})^{\uparrow L}} ; \text{ftn}_L ; \text{ftn}_M^L \\
 & \text{naturality of ftn}_L : \quad = \left(\text{sw}^{\uparrow L} ; \text{ftn}_L \right)^{\uparrow M \uparrow L} ; \underline{(\text{sw} ; \text{ftn}_M^L)^{\uparrow L}} ; \text{ftn}_L ; \text{ftn}_M^L \\
 & \text{associativity of ftn}_M : \quad = \text{sw}^{\uparrow L \uparrow M \uparrow L} ; \underline{\text{ftn}_L^{\uparrow M \uparrow L}} ; \text{sw}^{\uparrow L} ; \text{ftn}_L ; \underline{\text{ftn}_M^{\uparrow M \uparrow L}} ; \text{ftn}_M^L \\
 & \text{inner-interchange for sw} : \quad = \text{sw}^{\uparrow L \uparrow M \uparrow L} ; \underline{(\text{sw} ; \text{sw}^{\uparrow L} ; \text{ftn}_L)^{\uparrow L}} ; \underline{\text{ftn}_L} ; \text{ftn}_M^L ; \text{ftn}_M^L \\
 & \text{associativity of ftn}_L : \quad = (\text{sw}^{\uparrow L \uparrow M} ; \text{sw} ; \text{sw}^{\uparrow L})^{\uparrow L} ; \underline{\text{ftn}_L} ; \text{ftn}_L ; \underline{\text{ftn}_M^{\uparrow L}} ; \text{ftn}_M^L .
 \end{aligned}$$

We have again managed to move all `swaps` to the left and all `flattens` to the right of the expression.

Comparing now the two sides of the associativity law, we see that all the `flattens` occur in the same combination: $\text{ftn}_L ; \text{ftn}_L ; \text{ftn}_M^L ; \text{ftn}_M^L$. It remains to show that

$$\text{sw}^{\uparrow L} ; (\text{sw}^{\uparrow M} ; \text{sw})^{\uparrow L \uparrow L} = (\text{sw}^{\uparrow L \uparrow M} ; \text{sw} ; \text{sw}^{\uparrow L})^{\uparrow L} .$$

or equivalently

$$(\text{sw} ; \text{sw}^{\uparrow M \uparrow L} ; \text{sw}^{\uparrow L})^{\uparrow L} = (\text{sw}^{\uparrow L \uparrow M} ; \text{sw} ; \text{sw}^{\uparrow L})^{\uparrow L} .$$

The two sides are equal due to the naturality law of `swap`,

$$\text{sw} ; \text{sw}^{\uparrow M \uparrow L} = \text{sw}^{\uparrow L \uparrow M} ; \text{sw}.$$

This completes the proof of the following statement:

Statement 14.3.2.1 For two monads L and M , assume that there exists a function

$$\text{sw}_{L,M} : M^{L^A} \rightarrow L^{M^A}$$

(called “`swap`”), which is a natural transformation satisfying four additional laws:

$$\begin{aligned} \text{outer-identity} : \text{pu}_L^{\uparrow M} ; \text{sw}_{L,M} &= \text{pu}_L & , \\ \text{inner-identity} : \text{pu}_M ; \text{sw}_{L,M} &= \text{pu}_M^{\uparrow L} & , \\ \text{outer-interchange} : \text{ftn}_L^{\uparrow M} ; \text{sw}_{L,M} &= \text{sw}_{L,M} ; \text{sw}_{L,M}^{\uparrow L} ; \text{ftn}_L & , \\ \text{inner-interchange} : \text{ftn}_M ; \text{sw}_{L,M} &= \text{sw}_{L,M}^{\uparrow M} ; \text{sw}_{L,M} ; \text{ftn}_M^{\uparrow L} & . \end{aligned}$$

Then the functor composition

$$T^A \triangleq L^{M^A} = (L \circ M)^A$$

is a monad with the methods `pure` and `flatten` defined by

$$\text{pu}_T \triangleq \text{pu}_M ; \text{pu}_L & , \tag{14.19}$$

$$\text{ftn}_T \triangleq \text{sw}_{L,M}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L} & . \tag{14.20}$$

14.3.3 Intuition behind the laws of `swap`

The interchange laws for `swap` guarantee that any functor composition built up from L and M , e.g., like this,

$$M \circ M \circ L \circ M \circ L \circ L \circ M \circ M \circ L & ,$$

can be simplified to a value of type $T^A = L^{M^A}$ by flattening the layers using ftn_L , ftn_M , or ftn_T , or by interchanging the layers with `swap`. We may apply flattening or interchange in any order, and we will always get the same final value of type T^A .

In other words, the monadic effects of the monads L and M can be arbitrarily interleaved, swapped, and flattened in any order, with no change to the final results. The programmer is free to refactor a monadic program, say, by first computing some L -effects in a separate functor block of L -`flatMapS` and only then combining the result with the rest of the computation in the monad T . Regardless of the refactoring, the monad T computes all the effects correctly. This is what programmers would expect of the monad T , if it is to be regarded as a useful monad transformer.

We will now derive the properties of the monad T that follow from the interchange laws. We will find that it is easier to formulate these laws in terms of `swap` than in terms of ftn_T . In practice, all known examples of compositional monad transformers (the `pass/fail` monads, the “rigid” monads, and `Writer`) are defined via `swap`.

14.3.4 Deriving `swap` from `flatten`

We have shown that the `flatten` method of the monad $T^\bullet = L^{M^\bullet}$ can be defined via the `swap` method. However, we have seen examples of some composable monads (such as `Reader` and `Option`) where we already know the definitions of the `flatten` method for the composed monad T . Does a suitable `swap` function exist for these examples? In other words, if a `flatten` function for the monad $T = L \circ M$ is already known, can we establish whether a `swap` function exists such that the *given* `flatten` function is expressed via Eq. (14.11)?

To answer this question, let us look at the type signature of `flatten` for T :

$$\text{ftn}_T : L \circ M \circ L \circ M \rightsquigarrow L \circ M \quad .$$

This type signature is different from $\text{sw} : M \circ L \rightsquigarrow L \circ M$ only because the argument of ftn_T has extra layers of the functors L and M that are placed outside the $M \circ L$ composition. We can use the `pure` methods of M and L to add these extra layers to a value of type $M \circ L$, without modifying any monadic effects present in $M \circ L$. This will allow us to apply ftn_T and to obtain a value of type $L \circ M$. The resulting code for the function ftn_T and the corresponding type diagram are

$$\text{sw} = \text{pu}_M^{\uparrow L \uparrow M} ; \text{pu}_L ; \text{ftn}_T \quad . \quad (14.21)$$

$$\begin{array}{ccccc} M^{L^A} & \xrightarrow{\text{pu}_M^{\uparrow L \uparrow M}} & M^{L^{M^A}} & \xrightarrow{\text{pu}_L} & L^{M^{L^M}} \\ & \searrow & & & \downarrow \text{ftn}_T \\ & & \text{sw} \triangleq & & L^{M^A} \end{array}$$

We have expressed ftn_T and $\text{sw}_{L,M}$ through each other. Are these functions always equivalent?

Statement 14.3.4.1 The equivalence of `flatten` and `swap` holds under the following conditions:

- (a) Given an implementation of `swap` satisfying the naturality law, we may define ftn_T using Eq. (14.11) and then substitute that ftn_T into Eq. (14.21) to define a new `swap` function. The new function will be equal to the initial `swap`.
- (b) Given an implementation of ftn_T satisfying the additional “compatibility laws” (14.23)–(14.24), we may define `swap` using Eq. (14.21) and then substitute that `swap` into Eq. (14.11) to define a new ftn_T function. The new function will be equal to the initial ftn_T .

Proof (a) Substitute ftn_T from Eq. (14.11) into Eq. (14.21):

$$\begin{aligned} & \text{pu}_M^{\uparrow L \uparrow M} ; \text{pu}_L ; \text{ftn}_T \\ \text{substitute } \text{ftn}_T : &= \text{pu}_M^{\uparrow L \uparrow M} ; \text{pu}_L ; \text{sw}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L} \\ \text{naturality of } \text{pu}_L : &= \text{pu}_M^{\uparrow L \uparrow M} ; \text{sw} ; \text{pu}_L ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L} \\ \text{left identity law for } L : &= \text{pu}_M^{\uparrow L \uparrow M} ; \text{sw} ; \text{ftn}_M^{\uparrow L} \\ \text{naturality of } \text{sw} : &= \text{sw} ; \text{pu}_M^{\uparrow M \uparrow L} ; \text{ftn}_M^{\uparrow L} \\ \text{functor composition for } L : &= \text{sw} ; \left(\text{pu}_M^{\uparrow M} ; \text{ftn}_M \right)^{\uparrow L} \\ \text{right identity law for } M : &= \text{sw} \quad . \end{aligned}$$

We recovered the initial `swap` function.

(b) Substitute `sw` from Eq. (14.21) into Eq. (14.11):

$$\begin{aligned} & \text{sw}^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L} \\ \text{substitute } \text{sw} : &= \left(\text{pu}_M^{\uparrow L \uparrow M} ; \text{pu}_L ; \text{ftn}_T \right)^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L} \\ \text{functor composition} : &= \text{pu}_M^{\uparrow L \uparrow M \uparrow L} ; \text{pu}_L^{\uparrow L} ; \text{ftn}_T^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_M^{\uparrow L} \quad . \end{aligned} \quad (14.22)$$

At this point, we are stuck: we can find no laws to transform the last expression. Without assuming additional laws, it *does not follow* that the right-hand side of Eq. (14.22) is equal to ftn_T . Let us now derive those additional laws.

The only sub-expression in Eq. (14.22) that we could possibly transform is the composition $\text{ftn}_T^{\uparrow L} ; \text{ftn}_L$. So, we need to assume a law involving the expression

$$(\text{ftn}_T^{\uparrow L} ; \text{ftn}_L) : L \circ L \circ M \circ L \circ M \rightsquigarrow L \circ M \quad .$$

This function flattens the two layers of $(L \circ M)$ and then flattens the remaining two layers of L . Another function with the same type signature could first flatten the two *outside* layers of L and then flatten the two remaining layers of $(L \circ M)$:

$$(\text{ftn}_L ; \text{ftn}_T) : L \circ L \circ M \circ L \circ M \rightsquigarrow L \circ M \quad .$$

So we conjecture that a possibly useful additional law for ftn_T is

$$\text{ftn}_L ; \text{ftn}_T = \text{ftn}_T^{\uparrow L} ; \text{ftn}_L \quad . \quad (14.23)$$

$$\begin{array}{ccc} L^{L^M L^{M^A}} & \xrightarrow{\text{ftn}_L} & L^{M^L M^A} \\ \downarrow \text{ftn}_T^{\uparrow L} & & \downarrow \text{ftn}_T \\ L^{L^{M^A}} & \xrightarrow{\text{ftn}_L} & L^{M^A} \end{array}$$

This law expresses a kind of “compatibility” between the monads L and T .

With this law, the right-hand side of Eq. (14.22) becomes

$$\begin{aligned} & \text{pu}_M^{\uparrow L \uparrow M \uparrow L} ; \text{pu}_L^{\uparrow L} ; \text{ftn}_L ; \text{ftn}_T ; \text{ftn}_M^{\uparrow L} \\ \text{right identity law of } L : & = \text{pu}_M^{\uparrow L \uparrow M \uparrow L} ; \text{ftn}_T ; \text{ftn}_M^{\uparrow L} \quad . \end{aligned}$$

Again, we cannot proceed unless we assume a law involving the expression

$$(\text{ftn}_T ; \text{ftn}_M^{\uparrow L}) : L \circ M \circ L \circ M \rightsquigarrow L \circ M \quad .$$

This function first flattens the two layers of $(L \circ M)$ and then flattens the remaining two layers of M . An alternative order of flattenings is to first flatten the *innermost* two layers of M :

$$(\text{ftn}_M^{\uparrow L \uparrow M \uparrow L} ; \text{ftn}_T) : L \circ M \circ L \circ M \rightsquigarrow L \circ M \quad .$$

The second conjectured law is therefore

$$\text{ftn}_T ; \text{ftn}_M^{\uparrow L} = \text{ftn}_M^{\uparrow L \uparrow M \uparrow L} ; \text{ftn}_T \quad . \quad (14.24)$$

$$\begin{array}{ccc} L^{M^L M^{M^A}} & \xrightarrow{\text{ftn}_T} & L^{M^{M^A}} \\ \downarrow \text{ftn}_M^{\uparrow L \uparrow M \uparrow L} & & \downarrow \text{ftn}_M^{\uparrow L} \\ L^{M^{L^M A}} & \xrightarrow{\text{ftn}_T} & L^{M^A} \end{array}$$

This law expresses a kind of “compatibility” between the monads M and T .

Assuming this law, we can finally complete the derivation:

$$\begin{aligned} & \text{pu}_M^{\uparrow L \uparrow M \uparrow L} ; \text{ftn}_T ; \text{ftn}_M^{\uparrow L} \\ \text{substitute the second conjecture :} & = \text{pu}_M^{\uparrow L \uparrow M \uparrow L} ; \text{ftn}_M^{\uparrow L \uparrow M \uparrow L} ; \text{ftn}_T \\ \text{functor composition :} & = (\text{pu}_M ; \text{ftn}_M)^{\uparrow L \uparrow M \uparrow L} ; \text{ftn}_T \\ \text{left identity law of } M : & = \text{ftn}_T \quad . \end{aligned}$$

We recovered the initial ftn_T by assuming two additional laws.

It turns out that these additional laws will always hold when ftn_T is defined via `swap` (see Exercise 14.9.2.7).

It may be hard to verify directly the monad laws for $L \circ M$ because of deeply nested type constructors, e.g., $L \circ M \circ L \circ M \circ L \circ M$. If the monad $L \circ M$ has a `swap` method (in practice, this is always the case), it is simpler to verify the laws of `swap` and then obtain the monad laws of $L \circ M$ via Statement 14.3.2.1.

14.3.5 Monad transformer identity law: Proofs

Statement 14.3.5.1 Assume that pu_T and ftn_T are defined by Eqs. (14.19)–(14.20), and that the two identity laws of `swap` hold (see Statement 14.3.2.1),

$$\begin{aligned} \text{outer-identity: } \text{pu}_L^{\uparrow M} ; \text{sw}_{L,M} &= \text{pu}_L & , \\ \text{inner-identity: } \text{pu}_M ; \text{sw}_{L,M} &= \text{pu}_M^{\uparrow L} & . \end{aligned}$$

Then the identity law holds, $T_L^M \cong L$ for $M = \text{Id}$, via monad isomorphisms.

Proof We will now prove this law assuming that M is the foreign monad for composed-outside transformers, $T_L^M = L \circ M$. Setting $M = \text{Id}$ in the inner-identity law, we obtain

$$\text{pu}_{\text{Id}} ; \text{sw}_{L,\text{Id}} = \text{pu}_{\text{Id}}^{\uparrow L} .$$

Since $\text{pu}_{\text{Id}} = \text{id}$, it follows that $\text{sw}_{L,\text{Id}} = \text{id}$. In a similar way, for composed-inside transformers $T_L^M = M \circ L$ we need to switch the roles of M and L in the same computation and substitute $L = \text{id}$ into the outer-identity law,

$$\text{pu}_{\text{Id}}^{\uparrow M} ; \text{sw}_{\text{Id},M} = \text{pu}_{\text{Id}} .$$

We obtain $\text{sw}_{\text{Id},M} = \text{id}$.

Note that $\text{sw}_{L,\text{Id}} : L^A \rightarrow L^A$ is a natural transformation for a monad L , so one may heuristically expect $\text{sw}_{L,\text{Id}}$ to be equal to the identity map (the only natural transformation $L^A \rightarrow L^A$ that exists for all monads L). Similarly, one may expect that $\text{sw}_{\text{Id},M} : M^A \rightarrow M^A = \text{id}$ since it is a natural transformation. But these are only heuristic expectations, while we have just shown that the properties $\text{sw}_{L,\text{Id}} = \text{id}$ and $\text{sw}_{\text{Id},M} = \text{id}$ follow from the previously established laws of `swap` without any new assumptions. These properties will be needed in the proofs below.

To demonstrate a monadic isomorphism between the monads T_L^{Id} and L , we will consider separately the cases of composed-inside and composed-outside transformers.

For composed-inside transformers $T_L^M = M \circ L$, we set $M = \text{Id}$ and find that the monad $T_L^{\text{Id}} = \text{Id} \circ L = L$ is the same type constructor as L . So, the isomorphism maps between T_L^{Id} and L are simply the identity maps in both directions, $\text{id} : T_L^{\text{Id},A} \rightarrow L^A$ and $\text{id} : L^A \rightarrow T_L^{\text{Id},A}$.

For composed-outside transformers $T_L^M = L \circ M$, the monad $T_L^{\text{Id}} = L \circ \text{Id} = L$ is again the same type constructor as L . The isomorphisms between T_L^{Id} and L are again the identity maps in both directions, $\text{id} : T_L^{\text{Id},A} \rightarrow L^A$ and $\text{id} : L^A \rightarrow T_L^{\text{Id},A}$.

We have found the isomorphism maps between T_L^{Id} and L . However, we still need to verify that the monad structure of T_L^{Id} is the same as that of L ; otherwise the isomorphism would not be a *monad* isomorphism (i.e., an isomorphism that preserves the structure of the monads). To verify this, it is sufficient to show that the methods pu_T and ftn_T defined by Eqs. (14.19)–(14.20) for the monad T_L^{Id} are *the same functions* as the given methods pu_L and ftn_L of the monad L . If the monad's methods are the same functions, i.e. $\text{pu}_L = \text{pu}_T$ and $\text{ftn}_L = \text{ftn}_T$, then the identity map $\text{id} : T^A \rightarrow L^A$ will satisfy the laws of the monad morphism,

$$\text{pu}_T ; \text{id} = \text{pu}_L , \quad \text{ftn}_T ; \text{id} = \text{id}^{\uparrow T} ; \text{id} ; \text{ftn}_L .$$

In the same way, the laws of the monad morphism will hold for the identity map in the direction $L \rightsquigarrow T$.

For composed-inside transformers: We need to show that $\text{pu}_M = \text{pu}_T$ and $\text{ftn}_M = \text{ftn}_T$. Designate L as the foreign monad and M as the base monad in Eq. (14.19), as appropriate for the composed-inside transformer $T_L^M = M \circ L$. Setting the foreign monad to identity, $L = \text{Id}$, in Eq. (14.19) gives

$$\text{pu}_T = \text{pu}_M \circ \text{pu}_{\text{Id}} = \text{pu}_M.$$

To show that $\text{ftn}_M = \text{ftn}_T$, we use Eq. (14.20) with $L = \text{Id}$:

$$\begin{aligned} \text{ftn}_T &= \text{sw}_{\text{Id}, M}^{\uparrow \text{Id}} \circ \text{ftn}_{\text{Id}} \circ \text{ftn}_M^{\uparrow \text{Id}} \\ \text{use } \text{ftn}_{\text{Id}} = \text{id} \text{ and } \text{sw}_{\text{Id}, M} = \text{id} : \quad &= \text{ftn}_M \quad . \end{aligned}$$

For composed-outside transformers: We need to show that $\text{pu}_L = \text{pu}_T$ and $\text{ftn}_L = \text{ftn}_T$. Designate M as the foreign monad and L as the base monad in Eq. (14.19), as appropriate for the composed-outside transformer $T_L^M = L \circ M$. Setting the foreign monad to identity, $M = \text{Id}$, in Eq. (14.19) gives

$$\text{pu}_T = \text{pu}_{\text{Id}} \circ \text{pu}_L = \text{pu}_L.$$

To show that $\text{ftn}_L = \text{ftn}_T$, use Eq. (14.20) with $M = \text{Id}$:

$$\begin{aligned} \text{ftn}_T &= \text{sw}_{L, \text{Id}}^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_{\text{Id}}^{\uparrow L} \\ \text{use } \text{ftn}_{\text{Id}} = \text{id} \text{ and } \text{sw}_{L, \text{Id}} = \text{id} : \quad &= \text{ftn}_L \quad . \end{aligned}$$

14.3.6 Monad transformer lifting laws: Proofs

Statement 14.3.6.1 The monad transformer lifting laws hold if the monad methods of T are defined using Eqs. (14.19)–(14.20), and if the `swap` function obeys the outer-identity and inner-identity laws (Statement 14.3.2.1).

Proof To be specific, let us assume that L is the base monad of the transformer. For the other choice of the base monad, we only need to interchange the definitions of `flift` and `blift`.

The lift functions of a compositional monad transformer are:

$$\begin{aligned} \text{flift} &= \text{pu}_L : M^A \rightarrow L^{M^A} \quad , \\ \text{blift} &= \text{pu}_M^{\uparrow L} : L^A \rightarrow L^{M^A} \quad . \end{aligned}$$

Their laws of monad morphisms (the identity and the composition laws) are

$$\begin{aligned} \text{pu}_M \circ \text{flift} &= \text{pu}_T \quad , \quad \text{pu}_L \circ \text{blift} = \text{pu}_T \quad , \\ \text{ftn}_M \circ \text{flift} &= \text{flift}^{\uparrow M} \circ \text{flift} \circ \text{ftn}_T \quad , \quad \text{ftn}_L \circ \text{blift} = \text{blift}^{\uparrow L} \circ \text{blift} \circ \text{ftn}_T \quad . \end{aligned}$$

The identity laws are verified quickly,

$$\begin{aligned} \text{expect to equal } \text{pu}_T : \quad &\text{pu}_M \circ \text{flift} = \text{pu}_M \circ \text{pu}_L \\ \text{definition of } \text{pu}_T : \quad &= \text{pu}_T \quad , \\ \text{expect to equal } \text{pu}_T : \quad &\text{pu}_L \circ \text{blift} = \text{pu}_L \circ \text{pu}_M^{\uparrow L} \\ \text{naturality of } \text{pu}_L : \quad &= \text{pu}_M \circ \text{pu}_L = \text{pu}_T \quad . \end{aligned}$$

To verify the composition laws, we need to start from their right-hand sides because the left-hand sides cannot be simplified. We then substitute the definition of ftn_T in terms of `swap`. The composition

law for flift :

$$\begin{aligned}
 \text{expect to equal } \text{ftn}_M \circ \text{pu}_L : & \text{ flift}^{\uparrow M} \circ \text{flift} \circ \text{ftn}_T \\
 \text{definitions of flift and ftn}_T : & = \text{pu}_L^{\uparrow M} \circ \text{pu}_L \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \\
 \text{naturality of pu}_L : & = \text{pu}_L^{\uparrow M} \circ \text{sw} \circ \text{pu}_L \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \\
 \text{left identity law of } L : & = \text{pu}_L^{\uparrow M} \circ \text{sw} \circ \text{ftn}_M^{\uparrow L} \\
 \text{inner-identity law of sw} : & = \text{pu}_L \circ \text{ftn}_M^{\uparrow L} \\
 \text{naturality of pu}_L : & = \text{ftn}_M \circ \text{pu}_L .
 \end{aligned}$$

The composition law for blift :

$$\begin{aligned}
 \text{expect to equal } \text{ftn}_L \circ \text{pu}_M^{\uparrow L} : & \text{ blift}^{\uparrow L} \circ \text{blift} \circ \text{ftn}_T \\
 \text{definitions of blift and ftn}_T : & = \text{pu}_M^{\uparrow L \uparrow L} \circ \text{pu}_M^{\uparrow L} \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \\
 \text{functor composition in } L : & = \text{pu}_M^{\uparrow L \uparrow L} \circ (\text{pu}_M \circ \text{sw})^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \\
 \text{outer-identity law of sw} : & = (\text{pu}_M^{\uparrow L \uparrow L} \circ \text{pu}_M^{\uparrow L \uparrow L}) \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \\
 \text{naturality of ftn}_L : & = \text{ftn}_L \circ (\text{pu}_M^{\uparrow L} \circ \text{pu}_M^{\uparrow L}) \circ \text{ftn}_M^{\uparrow L} \\
 \text{right identity law of } M : & = \text{ftn}_L \circ \text{pu}_M^{\uparrow L} .
 \end{aligned}$$

So, the lifting laws for T follow from the laws of swap .

14.3.7 Monad transformer runner laws: Proofs

Statement 14.3.7.1 The runners (frun and brun) satisfy the runner laws and the base runner laws, as long as the swap function is monadically natural as given by Eqs. (14.25) or (14.26) below. Compositional monad transformers have general base runners, $\text{brun}(\theta)$, that satisfy the laws for any lawful runner $\theta : L \rightsquigarrow \text{Id}$. The functions frun and brun also satisfy the monadic naturality laws.

Proof The laws of runners are not symmetric with respect to the base monad and the foreign monad: the runners are parametric in the foreign monad but not in the base monad. In each case, the swap function must be monadically natural with respect to the *foreign* monad. So, the laws need to be written differently, depending on the choice of the base monad. Let us consider separately the cases where either L or M is the base monad.

If the base monad is L , the runners are defined by

$$\begin{aligned}
 \text{frun}(\phi : M \rightsquigarrow N) : L \circ M \rightsquigarrow L \circ N , \quad \text{frun}(\phi) = \phi^{\uparrow L} ; \\
 \text{brun}(\theta : L \rightsquigarrow \text{Id}) : L \circ M \rightsquigarrow M , \quad \text{brun}(\theta) = \theta .
 \end{aligned}$$

The laws of runners require that $\text{frun}(\phi)$ and $\text{brun}(\theta)$ must be monad morphisms, i.e., the identity and composition laws must hold:

$$\begin{aligned}
 \text{pu}_{L \circ M} \circ \text{frun}(\phi) &= \text{pu}_{L \circ N} , \\
 \text{ftn}_{L \circ M} \circ \text{frun}(\phi) &= (\text{frun}(\phi))^{\uparrow M \uparrow L} \circ \text{frun}(\phi) \circ \text{ftn}_{L \circ N} , \\
 \text{pu}_{L \circ M} \circ \text{brun}(\theta) &= \text{pu}_M , \\
 \text{ftn}_{L \circ M} \circ \text{brun}(\theta) &= (\text{brun}(\theta))^{\uparrow M \uparrow L} \circ \text{brun}(\theta) \circ \text{ftn}_M .
 \end{aligned}$$

To derive these laws, we may use the identity and composition laws of monad morphisms for ϕ and θ . We also use Eqs. (14.19)–(14.20) as definitions of the monad T . Additionally, we assume the **monadic naturality** of swap with respect to ϕ and θ ,

$$\text{sw}_{L,M} \circ \phi^{\uparrow L} = \phi \circ \text{sw}_{L,N} , \quad \text{sw}_{L,M} \circ \theta = \theta^{\uparrow M} . \quad (14.25)$$

$$\begin{array}{ccc}
 M^{L^A} & \xrightarrow{\text{sw}_{L,M}} & L^{M^A} \\
 \phi \downarrow & & \downarrow \phi^{\uparrow L} \\
 N^{L^A} & \xrightarrow{\text{sw}_{L,N}} & L^{N^A}
 \end{array}
 \quad
 \begin{array}{ccc}
 M^{L^A} & \xrightarrow{\text{sw}_{L,M}} & L^{M^A} \\
 & \searrow \theta^{\uparrow M} & \downarrow \theta \\
 & & M^A
 \end{array}$$

The first law to be shown is the identity law for $\text{frun } \phi$:

$$\begin{aligned}
 \text{expect this to equal } \text{pu}_{L \circ M} \circ \text{frun } (\phi) : & \quad \text{pu}_{L \circ M} \circ \text{frun } (\phi) \\
 \text{definitions of frun and } \text{pu}_{L \circ M} : & = \text{pu}_M \circ \underline{\text{pu}_L \circ \phi^{\uparrow L}} \\
 \text{naturality of } \text{pu}_L : & = \underline{\text{pu}_M \circ \phi \circ \text{pu}_L} \\
 \text{identity law for } \phi : & = \text{pu}_N \circ \text{pu}_L \\
 \text{definition of } \text{pu}_{L \circ N} : & = \text{pu}_{L \circ N} .
 \end{aligned}$$

The next law to be verified is the composition law for $\text{frun } (\phi)$:

$$\begin{aligned}
 \text{expect this to equal } \text{ftn}_T \circ \phi^{\uparrow L} : & \quad (\text{frun } (\phi))^{\uparrow M \uparrow L} \circ \text{frun } (\phi) \circ \text{ftn}_{L \circ N} \\
 \text{definitions of frun and } \text{ftn}_{L \circ N} : & = \phi^{\uparrow L \uparrow M \uparrow L} \circ \phi^{\uparrow L} \circ \underline{\text{sw}_{L,N}^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_N^{\uparrow L}} \\
 \text{monadic naturality of } \text{sw}_{L,M} : & = \underline{\phi^{\uparrow L \uparrow M \uparrow L} \circ \text{sw}_{L,M}^{\uparrow L} \circ \phi^{\uparrow L \uparrow L} \circ \text{ftn}_L \circ \text{ftn}_N^{\uparrow L}} \\
 \text{naturality of } \text{sw}_{L,M} : & = \text{sw}_{L,M}^{\uparrow L} \circ \underline{\phi^{\uparrow M \uparrow L \uparrow L} \circ \phi^{\uparrow L \uparrow L} \circ \text{ftn}_L \circ \text{ftn}_N^{\uparrow L}} \\
 \text{naturality of } \text{ftn}_L : & = \text{sw}_{L,M}^{\uparrow L} \circ \text{ftn}_L \circ \underline{(\phi^{\uparrow M} \circ \phi \circ \text{ftn}_N)^{\uparrow L}} \\
 \text{composition law for } \phi : & = \underline{\text{sw}_{L,M}^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \circ \phi^{\uparrow L}} \\
 \text{definition of } \text{ftn}_T : & = \underline{\text{ftn}_T \circ \phi^{\uparrow L}} .
 \end{aligned}$$

The next law is the identity law for brun :

$$\begin{aligned}
 \text{expect this to equal } \text{pu}_M : & \quad \text{pu}_{L \circ M} \circ \text{brun } \theta \\
 \text{definitions of brun and } \text{pu}_{L \circ M} : & = \text{pu}_M \circ \underline{\text{pu}_L \circ \theta} \\
 \text{identity law for } \theta : & = \text{pu}_M .
 \end{aligned}$$

The last law to be shown is the composition law for brun . Begin with its right-hand side since it is longer,

$$\begin{aligned}
 & (\text{brun } (\theta))^{\uparrow M \uparrow L} \circ \text{brun } (\theta) \circ \text{ftn}_M \\
 \text{definition of brun} : & = \theta^{\uparrow M \uparrow L} \circ \theta \circ \text{ftn}_M .
 \end{aligned}$$

We cannot simplify this expression any more, and yet it is still different from the left-hand side. So let us transform the left-hand side, hoping to obtain the same expression. In particular, we need to move ftn_M to the right and θ to the left:

$$\begin{aligned}
 & \text{ftn}_{L \circ M} \circ \text{brun } (\theta) \\
 \text{definitions of } \text{ftn}_{L \circ M} \text{ and } \text{brun} : & = \text{sw}_{L,M}^{\uparrow L} \circ \text{ftn}_L \circ \underline{\text{ftn}_M^{\uparrow L} \circ \theta} \\
 \text{naturality of } \theta : & = \text{sw}_{L,M}^{\uparrow L} \circ \underline{\text{ftn}_L \circ \theta \circ \text{ftn}_M} \\
 \text{composition law for } \theta : & = \underline{\text{sw}_{L,M}^{\uparrow L} \circ \theta^{\uparrow L} \circ \theta \circ \text{ftn}_M} \\
 \text{functor composition} : & = \underline{(\text{sw}_{L,M} \circ \theta)^{\uparrow L} \circ \theta \circ \text{ftn}_M} \\
 \text{monadic naturality of } \text{sw}_{L,M} : & = \underline{\theta^{\uparrow M \uparrow L} \circ \theta \circ \text{ftn}_M} .
 \end{aligned}$$

We have transformed both sides of the law into the same expression.

The functor laws for `frun` are

$$\text{frun}(\text{id}) = \text{id} , \quad \text{frun}(\phi) ; \text{frun}(\chi) = \text{frun}(\phi ; \chi) .$$

Since $\text{frun}(\phi) = \phi^{\uparrow L}$ in our case, these laws hold because they are the same as the functor laws of L .

To verify the non-degeneracy law for `brun`:

$$\begin{aligned} \text{expect to equal id : } & \text{flift} ; \text{brun}(\theta) \\ \text{definitions of flift and brun : } & = \text{pu}_L ; \theta \\ \text{identity law for } \theta : & = \text{id} . \end{aligned}$$

Finally, we need to check the monadic naturality laws for `flift` and `brun`:

$$\begin{aligned} \phi^{M \rightsquigarrow N} ; \text{flift}_L^N &= \text{flift}_L^M ; (\phi^{M \rightsquigarrow N})^{\uparrow L} , \\ (\phi^{M \rightsquigarrow N})^{\uparrow L} ; \text{brun}_L^N(\theta_L) &= \text{brun}_L^M(\theta_L) ; \phi^{M \rightsquigarrow N} . \end{aligned}$$

To verify these laws, use the naturality laws of pu_L and θ_L :

$$\begin{aligned} \phi ; \text{flift} &= \phi ; \text{pu}_L = \text{pu}_L ; \phi^{\uparrow L} = \text{flift} ; \phi^{\uparrow L} , \\ \phi^{\uparrow L} ; \text{brun}(\theta_L) &= \phi^{\uparrow L} ; \theta_L = \theta_L ; \phi = \text{brun}(\theta_L) ; \phi . \end{aligned}$$

If the base monad is M , the runners are defined by

$$\begin{aligned} \text{frun}(\phi^{L \rightsquigarrow N}) : L \circ M \rightarrow N \circ M , \quad \text{frun}(\phi) = \phi ; \\ \text{brun} \theta^{M \rightsquigarrow \text{Id}} : L \circ M \rightsquigarrow L , \quad \text{brun}(\theta) = \theta^{\uparrow L} . \end{aligned}$$

The laws of runners require that $\text{frun}(\phi)$ and $\text{brun}(\theta)$ must be monad morphisms, i.e., the identity and composition laws must hold:

$$\begin{aligned} \text{pu}_{L \circ M} ; \text{frun}(\phi) &= \text{pu}_{N \circ M} , \\ \text{ftn}_{L \circ M} ; \text{frun}(\phi) &= (\text{frun}(\phi))^{\uparrow M \uparrow L} ; \text{frun}(\phi) ; \text{ftn}_{N \circ M} , \\ \text{pu}_{L \circ M} ; \text{brun}(\theta) &= \text{pu}_L , \\ \text{ftn}_{L \circ M} ; \text{brun}(\theta) &= (\text{brun}(\theta))^{\uparrow M \uparrow L} ; \text{brun}(\theta) ; \text{ftn}_L . \end{aligned}$$

The monadic naturality laws for `swap` with respect to ϕ and χ are

$$\text{sw}_{L,M} ; \phi = \phi^{\uparrow M} ; \text{sw}_{N,M} , \quad \text{sw}_{L,M} ; \theta^{\uparrow L} = \theta . \quad (14.26)$$

$$\begin{array}{ccc} M^{L^A} & \xrightarrow{\text{sw}_{L,M}} & L^{M^A} \\ \phi^{\uparrow M} \downarrow & & \downarrow \phi \\ M^{N^A} & \xrightarrow{\text{sw}_{N,M}} & N^{M^A} \end{array} \quad \begin{array}{ccc} M^{L^A} & \xrightarrow{\text{sw}_{L,M}} & L^{M^A} \\ & \searrow \theta & \downarrow \theta^{\uparrow L} \\ & & L^A \end{array}$$

The first law to be proved is

$$\begin{aligned} \text{expect to equal } \text{pu}_{N \circ M} : & \text{pu}_{L \circ M} ; \text{frun}(\phi) \\ \text{definitions of frun and } \text{pu}_{L \circ M} : & = \text{pu}_M ; \text{pu}_L ; \phi \\ \text{identity law for } \phi : & = \text{pu}_M ; \text{pu}_N \\ \text{definition of } \text{pu}_{N \circ M} : & = \text{pu}_{N \circ M} . \end{aligned}$$

The next law is the composition law for $\text{frun}(\phi)$:

$$\begin{aligned}
 & \text{expect this to equal } \text{ftn}_T \circ \phi : \quad (\text{frun}(\phi))^{\uparrow M \uparrow L} \circ \text{frun}(\phi) \circ \text{ftn}_{N \circ M} \\
 & \text{definitions of } \text{frun} \text{ and } \text{ftn}_{N \circ M} : \quad = \phi^{\uparrow M \uparrow L} \circ \phi \circ \text{sw}_{N,M}^{\uparrow N} \circ \text{ftn}_N \circ \text{ftn}_M^{\uparrow N} \\
 & \text{naturality of } \phi : \quad = \phi^{\uparrow M \uparrow L} \circ \text{sw}_{N,M}^{\uparrow L} \circ \phi \circ \text{ftn}_N \circ \text{ftn}_M^{\uparrow N} \\
 & \text{monadic naturality of } \text{sw}_{N,M} \text{ raised to } L : \quad = \text{sw}_{L,M}^{\uparrow L} \circ \phi^{\uparrow L} \circ \phi \circ \text{ftn}_N \circ \text{ftn}_M^{\uparrow N} \\
 & \text{composition law for } \phi : \quad = \text{sw}_{L,M}^{\uparrow L} \circ \text{ftn}_L \circ \phi \circ \text{ftn}_M^{\uparrow N} \\
 & \text{naturality of } \phi : \quad = \text{sw}_{L,M}^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \circ \phi \\
 & \text{definition of } \text{ftn}_T : \quad = \text{ftn}_T \circ \phi \quad .
 \end{aligned}$$

The next law is the identity law for $\text{brun}(\theta)$:

$$\begin{aligned}
 & \text{expect this to equal } \text{pu}_L : \quad \text{pu}_{L \circ M} \circ \text{brun}(\theta) \\
 & \text{definitions of } \text{brun} \text{ and } \text{pu}_{L \circ M} : \quad = \text{pu}_M \circ \text{pu}_L \circ \theta^{\uparrow L} \\
 & \text{naturality of } \text{pu}_L : \quad = \text{pu}_M \circ \theta \circ \text{pu}_L \\
 & \text{identity law for } \theta : \quad = \text{pu}_L \quad .
 \end{aligned}$$

The last law is the composition law for $\text{brun}(\theta)$. Begin with its right-hand side,

$$\begin{aligned}
 & (\text{brun}(\theta))^{\uparrow M \uparrow L} \circ \text{brun}(\theta) \circ \text{ftn}_L \\
 & \text{definition of } \text{brun} : \quad = \theta^{\uparrow L \uparrow M \uparrow L} \circ \theta^{\uparrow L} \circ \text{ftn}_L \\
 & \text{functor composition} : \quad = (\theta^{\uparrow L \uparrow M} \circ \theta)^{\uparrow L} \circ \text{ftn}_L \\
 & \text{naturality of } \theta : \quad = (\theta \circ \theta^{\uparrow L})^{\uparrow L} \circ \text{ftn}_L \quad .
 \end{aligned}$$

We now transform the left-hand side, hoping to obtain the same expression. We need to move ftn_L to the right and θ to the left:

$$\begin{aligned}
 & \text{expect to equal } \theta^{\uparrow L} \circ \theta^{\uparrow L \uparrow L} \circ \text{ftn}_L : \quad \text{ftn}_{L \circ M} \circ \text{brun}(\theta) \\
 & \text{definitions of } \text{ftn}_{L \circ M} \text{ and } \text{brun} : \quad = \text{sw}_{L,M}^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \circ \theta^{\uparrow L} \\
 & \text{composition law for } \theta : \quad = \text{sw}_{L,M}^{\uparrow L} \circ \text{ftn}_L \circ (\theta \circ \theta)^{\uparrow L} \\
 & \text{naturality of } \text{ftn}_L : \quad = \text{sw}_{L,M}^{\uparrow L} \circ (\theta \circ \theta)^{\uparrow L \uparrow L} \circ \text{ftn}_L \\
 & \text{functor composition} : \quad = (\text{sw}_{L,M} \circ \theta^{\uparrow L})^{\uparrow L} \circ \theta^{\uparrow L \uparrow L} \circ \text{ftn}_L \\
 & \text{composition law for } \theta : \quad = \theta^{\uparrow L} \circ \theta^{\uparrow L \uparrow L} \circ \text{ftn}_L \quad .
 \end{aligned}$$

The functor laws for frun are

$$\text{frun}(\text{id}) = \text{id} \quad , \quad \text{frun}(\phi) \circ \text{frun}(\chi) = \text{frun}(\phi \circ \chi) \quad .$$

Since $\text{frun}(\phi) = \phi$ in our case, these laws are trivially satisfied.

To verify the non-degeneracy law for brun :

$$\begin{aligned}
 & \text{expect to equal } \text{id} : \quad \text{flift} \circ \text{brun}(\theta) \\
 & \text{definitions of } \text{flift} \text{ and } \text{brun} : \quad = \text{pu}_M^{\uparrow L} \circ \theta^{\uparrow L} \\
 & \text{identity law for } \theta : \quad = \text{id}^{\uparrow L} = \text{id} \quad .
 \end{aligned}$$

Finally, we need to check the monadic naturality laws for `flift` and `brun`:

$$\begin{aligned}\phi^{L \rightsquigarrow K} ; \text{flift}_M^K &= \text{flift}_M^L ; \phi^{L \rightsquigarrow K} , \\ \phi^{L \rightsquigarrow K} ; \text{brun}_M^K(\theta_M) &= \text{brun}_M^L(\theta_M) ; \phi^{L \rightsquigarrow K} .\end{aligned}$$

To verify these laws, use the naturality law of $\phi^{L \rightsquigarrow K}$:

$$\begin{aligned}\phi ; \text{flift}_M^K &= \phi ; \text{pu}_M^{\uparrow K} = \text{pu}_M^{\uparrow L} ; \phi = \text{flift}_N^L ; \phi , \\ \phi ; \text{brun}_M^K(\theta_M) &= \phi ; \theta_M^{\uparrow K} = \theta_M^{\uparrow L} ; \phi = \text{brun}_M^L(\theta_M) ; \phi .\end{aligned}$$

14.3.8 Summary of results

The following two statements summarize the derivations in Section 14.3:

Statement 14.3.8.1 (composed-outside) For a base monad L and a foreign monad M , the functor composition $L \circ M$ is a lawful monad transformer if a `swap` function $\text{sw}_{L,M} : M \circ L \rightsquigarrow L \circ M$ exists, satisfying the conditions of Statement 14.3.2.1 and the monadic naturality laws

$$\text{sw}_{L,M} ; \phi^{\uparrow L} = \phi ; \text{sw}_{L,N} , \quad \text{sw}_{L,M} ; \theta = \theta^{\uparrow M} ,$$

with respect to arbitrary monad morphisms $\phi : M \rightsquigarrow N$ and $\theta : L \rightsquigarrow \text{Id}$. An additional law, $\text{sw}_{L,\text{Id}} = \text{id}$, follows from the conditions of Statement 14.3.2.1.

Statement 14.3.8.2 (composed-inside) For a base monad M and a foreign monad L , the functor composition $L \circ M$ is a lawful monad transformer if a `swap` function $\text{sw}_{L,M} : M \circ L \rightsquigarrow L \circ M$ exists, satisfying the conditions of Statement 14.3.2.1 and the monadic naturality laws

$$\text{sw}_{L,M} ; \phi = \phi^{\uparrow M} ; \text{sw}_{N,M} , \quad \text{sw}_{L,M} ; \theta^{\uparrow L} = \theta ,$$

with respect to arbitrary monad morphisms $\phi : L \rightsquigarrow N$ and $\theta : M \rightsquigarrow \text{Id}$. An additional law, $\text{sw}_{\text{Id},M} = \text{id}$, follows from the conditions of Statement 14.3.2.1.

Statements 14.3.8.1–14.3.8.2 enable us to check more easily whether a given base monad admits a monad transformer via functor composition. It is easier to verify the 6 laws for the `swap` function than to verify the 18 monad transformer laws directly. Also, the laws of `swap` use simpler types than the monad transformer laws.

14.4 Composed-inside transformers: Linear monads

A monad M is **linear** if it is of the form $M^A \triangleq P + Q \times A$, where P and Q are fixed types, and Q is a monoid. (The polynomial $P + Q \times A$ is linear in its type parameter A .) Well-known examples of linear monads are `Option`, `Either`, and `Writer`. The general case $M^A \triangleq P + Q \times A$ represents a computation that can fail and at the same time append a log message. So, M can be seen as a composition of `Either` and `Writer`.

A different (but also linear) monad is obtained from the composition of `Writer` and `Either`. The type constructor of this monad is $Q \times (P + A)$.

In general, composition of two linear monads $M_1^A \triangleq P_1 + Q_1 \times A$ and $M_2^A \triangleq P_2 + Q_2 \times A$ is again linear because

$$\begin{aligned}P_1 + Q_1 \times (P_2 + Q_2 \times A) \\ \text{expand brackets : } &\cong P_1 + Q_1 \times P_2 + Q_1 \times Q_2 \times A \\ \text{define new } P, Q : &\cong P + Q \times A .\end{aligned}$$

Note that we need to define $Q \triangleq Q_1 \times Q_2$, and Q is a monoid since, by assumption, Q_1 and Q_2 are monoids.

For a linear monad M and any foreign monad L , the functor composition $L \circ M$ is a monad. For example, the type constructor for the `OptionT` monad transformer can be defined as

```
type OptionT[L[_], A] = L[Option[A]]
```

The `Option` type constructor must be composed *inside* the foreign monad L . This is the case for all linear monads. Linear monads are the only known examples of monads whose transformers are composed inside the foreign monad.

14.4.1 Definitions of `swap` and `flatten`

To show that the monad transformer for the base monad $M^A \triangleq P + Q \times A$ is $T_M^{L,A} = L^{M^A}$, we will implement a suitable `swap` function having the type signature

$$\text{sw}_{N,M} : M^{L^A} \rightarrow L^{M^A} \quad ,$$

for the base monad $M^A \triangleq P + Q \times A$ and an arbitrary foreign monad L . We will then prove that `swap` satisfies all the required laws stated in Statement 14.3.8.2. This will guarantee that $T_M^{L,A} = L^{M^A}$ is a lawful monad transformer.

Expanding the definition of the type constructor M^\bullet , we can write the type signature of the `swap` function as

$$\text{sw}_{L,M} : P + Q \times L^A \rightarrow L^{P+Q \times A} \quad .$$

We can map P to L^P by applying pu_L . We can also map $Q \times L^A \rightarrow L^{Q \times A}$ since L is a functor,

$$q \times l \rightarrow (a \rightarrow q \times a)^{\uparrow L} l \quad .$$

It remains to unite these two functions. In the matrix notation, we write

$$\text{sw}_{L,M} = \left| \begin{array}{c} P \\ Q \times L^A \end{array} \right\| \left| \begin{array}{c} (x:P \rightarrow x + \mathbb{0}:Q \times A) \circ \text{pu}_L \\ q \times l \rightarrow (a:A \rightarrow \mathbb{0}:P + q \times a)^{\uparrow L} l \end{array} \right| \quad . \quad (14.27)$$

In Scala, the code is

```
type M[A, P, Q] = Either[P, (Q, A)]
def swap[L[_]: Monad, A, P, Q]: M[L[A]] => L[M[A]] = {
  case Left(p) => Monad[L].pure(Left(p))
  case Right((q, la)) => la.map(a => Right((q, a)))
}
```

Given this `swap` function, we define the `flatten` method for the transformed monad T (short notation ftn_T) by the standard formula

$$\text{ftn}_T = \text{sw}^{\uparrow L} \circ \text{ftn}_L \circ \text{ftn}_M^{\uparrow L} \quad .$$

The `pure` method for T (short notation pu_T) is $\text{pu}_T = \text{pu}_M \circ \text{pu}_L$. In Scala:

```
def pure[L[_]: Monad, A, P, Q: Monoid](x: A): L[M[A]] =
  Monad[L].pure(Right((Monoid[Q].empty, x)))
def flatten[L[_]: Monad, A, P, Q: Monoid](tt: L[M[L[M[A]]]]): L[M[A]] =
  tt.map(swap).flatten.map(_.flatten) // Assuming suitable implicits in scope.
```

Now we will verify that the laws of `swap` hold. We will need to use the code for the methods $fmap_M$, ftn_M , and pu_M of the monad M :

$$\begin{aligned}
 f\text{map}_M f^{A \rightarrow B} &= f^{\uparrow M} = \left| \begin{array}{c|cc} & P & Q \times B \\ \hline P & \text{id} & \mathbb{0} \\ Q \times A & \mathbb{0} & q \times a \rightarrow q \times f(a) \end{array} \right| , \\
 pu_M a^A &= 0^P + q_0 \times a , \quad pu_M = \left| \begin{array}{c|cc} & P & Q \times A \\ \hline A & \mathbb{0} & a \rightarrow q_0 \times a \end{array} \right| , \\
 f\text{tn}_M^{M^{MA} \rightarrow M^A} &= \left| \begin{array}{c|cc} & P & Q \times A \\ \hline P & \text{id} & \mathbb{0} \\ Q \times P & q \times p \rightarrow p & \mathbb{0} \\ Q \times Q \times A & \mathbb{0} & q_1 \times q_2 \times a \rightarrow (q_1 \oplus q_2) \times a \end{array} \right| .
 \end{aligned}$$

14.4.2 Laws of `swap`

We do not need to verify naturality since `swap` is defined as a fully parametric function.

The inner-identity law We need to show that $pu_L^{\uparrow M} \circ sw = pu_L$:

$$\begin{aligned}
 pu_L^{\uparrow M} \circ sw &= \left\| \begin{array}{c|cc} \text{id} & \mathbb{0} \\ \hline \mathbb{0} & q \times a \rightarrow q \times pu_L a \end{array} \right\| \circ \left\| \begin{array}{c} (x^P \rightarrow x + \mathbb{0}^{Q \times A}) \circ pu_L \\ q \times l \rightarrow (x^A \rightarrow \mathbb{0}^P + q \times x)^{\uparrow L} l \end{array} \right\| \\
 \text{composition :} &= \left\| \begin{array}{c} (x^P \rightarrow x + \mathbb{0}^{Q \times A}) \circ pu_L \\ q \times a \rightarrow (a^A \rightarrow \mathbb{0}^P + q \times a)^{\uparrow L} (pu_L a) \end{array} \right\| \\
 pu_L \text{'s naturality :} &= \left| \begin{array}{c|c} P & x^P \rightarrow pu_L(x + \mathbb{0}^{Q \times A}) \\ Q \times A & q \times a \rightarrow pu_L(\mathbb{0}^P + q \times a) \end{array} \right| \\
 \text{matrix notation :} &= pu_L .
 \end{aligned}$$

The outer-identity law We need to show that $pu_M \circ sw = pu_M^{\uparrow L}$:

$$\begin{aligned}
 pu_M \circ sw &= \left\| \begin{array}{c|c} \mathbb{0} & l^{L^A} \rightarrow q_0 \times l \end{array} \right\| \circ \left\| \begin{array}{c} (x^P \rightarrow x + \mathbb{0}^{Q \times A}) \circ pu_L \\ q \times l \rightarrow (x^A \rightarrow \mathbb{0}^P + q \times x)^{\uparrow L} l \end{array} \right\| \\
 \text{composition :} &= l^{L^A} \rightarrow (x^A \rightarrow \mathbb{0}^P + q_0 \times x)^{\uparrow L} l \\
 \text{definition of } pu_M : &= l \rightarrow pu_M^{\uparrow L} l = pu_M .
 \end{aligned}$$

The inner-interchange law Show that $\text{ftn}_L^M \circ \text{sw} = \text{sw} \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L$:

$$\begin{aligned}
 \text{ftn}_L^M \circ \text{sw} &= \left\| \begin{array}{c} \text{id} \quad \emptyset \\ \emptyset \quad q \times l \rightarrow q \times \text{ftn}_L \end{array} \right\| \circ \left\| \begin{array}{c} (x:P \rightarrow x + \emptyset:Q \times A) \circ \text{pu}_L \\ q \times l \rightarrow (a \rightarrow \emptyset:P + q \times a)^{\uparrow L}l \end{array} \right\| \\
 &= \left\| \begin{array}{c} (x:P \rightarrow x + \emptyset:Q \times A) \circ \text{pu}_L \\ q \times l \rightarrow (a \rightarrow \emptyset:P + q \times a)^{\uparrow L}(\text{ftn}_L l) \end{array} \right\|, \\
 \text{sw} \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L &= \left\| \begin{array}{c} (x:P \rightarrow x + \emptyset:Q \times A) \circ \text{pu}_L \\ q \times l \rightarrow (a \rightarrow \emptyset:P + q \times a)^{\uparrow L}l \end{array} \right\| \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L \\
 &= \left\| \begin{array}{c} (x:P \rightarrow x + \emptyset:Q \times A) \circ \text{pu}_L \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L \\ (q \times l \rightarrow (a \rightarrow \emptyset:P + q \times a)^{\uparrow L}l) \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L \end{array} \right\|.
 \end{aligned} \tag{14.28}$$

It is quicker to simplify each expression in the last column separately and then to compare with the column in Eq. (14.28). Simplify the upper expression:

$$\begin{aligned}
 &(x:P \rightarrow x + \emptyset:Q \times A) \circ \text{pu}_L \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L \\
 \text{naturality of } \text{pu}_L : &= (x:P \rightarrow x + \emptyset:Q \times A) \circ \text{sw} \circ \text{pu}_L \circ \text{ftn}_L \\
 \text{identity law of } L : &= (x:P \rightarrow x + \emptyset:Q \times A) \circ \text{sw} \\
 \text{definition of } \text{sw} : &= (x:P \rightarrow x + \emptyset:Q \times A) \circ \text{pu}_L.
 \end{aligned}$$

This equals the upper expression in Eq. (14.28). Simplify the lower expression;

$$\begin{aligned}
 &(q \times l \rightarrow (a \rightarrow \emptyset:P + q \times a)^{\uparrow L}l) \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L \\
 \text{definition of } \triangleright : &= q \times l \rightarrow l \triangleright (a \rightarrow \emptyset:P + q \times a)^{\uparrow L} \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L.
 \end{aligned} \tag{14.29}$$

Simplify the expression $(a \rightarrow \emptyset:P + q \times a)^{\uparrow L} \circ \text{sw}^{\uparrow L}$ separately:

$$\begin{aligned}
 &(a \rightarrow \emptyset:P + q \times a) \circ \text{sw} \\
 \text{composition} : &= a \rightarrow \text{sw}(\emptyset:P + q \times a) \\
 \text{definition of } \text{sw} : &= a \rightarrow a \triangleright (x \rightarrow \emptyset:P + q \times x)^{\uparrow L} \\
 \text{omit argument} : &= (x \rightarrow \emptyset:P + q \times x)^{\uparrow L}.
 \end{aligned} \tag{14.30}$$

Then we continue simplifying Eq. (14.29):

$$\begin{aligned}
 &q \times l \rightarrow l \triangleright (a \rightarrow \emptyset:P + q \times a)^{\uparrow L} \circ \text{sw}^{\uparrow L} \circ \text{ftn}_L \\
 \text{use Eq. (14.30)} : &= q \times l \rightarrow l \triangleright (x \rightarrow \emptyset:P + q \times x)^{\uparrow L} \circ \text{ftn}_L \\
 \text{naturality of } \text{ftn}_L : &= q \times l \rightarrow l \triangleright \text{ftn}_L \circ (x \rightarrow \emptyset:P + q \times x)^{\uparrow L} \\
 \text{definition of } \triangleright : &= q \times l \rightarrow (x \rightarrow \emptyset:P + q \times x)^{\uparrow L}(\text{ftn}_L l).
 \end{aligned}$$

This equals the lower expression in Eq. (14.28) after renaming x to a .

The outer-interchange law Show that $\text{ftn}_M ; \text{sw} = \text{sw}^{\uparrow M} ; \text{sw} ; \text{ftn}_M^{\uparrow L}$. The left-hand side is written using the matrices for ftn_M and sw :

$$\begin{aligned}
 & \text{ftn}_M ; \text{sw} \\
 &= \left\| \begin{array}{cc} \text{id} & \emptyset \\ q \times p \rightarrow p & \emptyset \\ \emptyset & q_1 \times q_2 \times a \rightarrow (q_1 \oplus q_2) \times a \end{array} \right\| \circ \left\| \begin{array}{c} (x:P \rightarrow x + \emptyset) ; \text{pu}_L \\ q \times l \rightarrow (x \rightarrow \emptyset + q \times x)^{\uparrow L} l \end{array} \right\| \\
 &= \left\| \begin{array}{c} (x:P \rightarrow x + \emptyset) ; \text{pu}_L \\ (q \times p \rightarrow p + \emptyset) ; \text{pu}_L \\ q_1 \times q_2 \times a \rightarrow (x \rightarrow \emptyset + (q_1 \oplus q_2) \times x)^{\uparrow L} a \end{array} \right\|. \tag{14.31}
 \end{aligned}$$

We cannot simplify this any more, so we hope to transform the right-hand side, $\text{sw}^{\uparrow M} ; \text{sw} ; \text{ftn}_M^{\uparrow L}$, to the same column expression. Begin by writing the matrix for $\text{sw}^{\uparrow M}$, expanding the rows for the input type $M^{M^{L^A}}$:

$$\begin{aligned}
 \text{sw}^{\uparrow M} &= \left\| \begin{array}{c} P \\ Q \times P \\ Q \times Q \times L^A \end{array} \right\| \begin{array}{cc} \text{id} & \emptyset \\ \emptyset & q \times p \rightarrow q \times \text{sw}(p + \emptyset) \\ \emptyset & q_1 \times q_2 \times l \rightarrow q_1 \times \text{sw}(\emptyset + q_2 \times l) \end{array} \right\| \\
 &= \left\| \begin{array}{c} P \\ Q \times P \\ Q \times Q \times L^A \end{array} \right\| \begin{array}{cc} \text{id} & \emptyset \\ \emptyset & q \times p \rightarrow q \times \text{pu}_L(p + \emptyset) \\ \emptyset & q_1 \times q_2 \times l \rightarrow q_1 \times (x \rightarrow \emptyset + q_2 \times x)^{\uparrow L} l \end{array} \right\|.
 \end{aligned}$$

Then compute the composition $\text{sw}^{\uparrow M} ; \text{sw}$ as

$$\begin{aligned}
 & \text{sw}^{\uparrow M} ; \text{sw} \\
 &= \left\| \begin{array}{cc} \text{id} & \emptyset \\ \emptyset & q \times p \rightarrow q \times \text{pu}_L(p + \emptyset) \\ \emptyset & q_1 \times q_2 \times l \rightarrow q_1 \times (x \rightarrow \emptyset + q_2 \times x)^{\uparrow L} l \end{array} \right\| \circ \left\| \begin{array}{c} (x:P \rightarrow x + \emptyset) ; \text{pu}_L \\ q \times l \rightarrow (x \rightarrow \emptyset + q \times x)^{\uparrow L} l \end{array} \right\| \\
 &= \left\| \begin{array}{c} (x:P \rightarrow x + \emptyset) ; \text{pu}_L \\ q \times p \rightarrow (x^{M^A} \rightarrow \emptyset^P + q \times x)^{\uparrow L} (\text{pu}_L(p + \emptyset)) \\ q_1 \times q_2 \times l \rightarrow (x^{M^A} \rightarrow \emptyset^P + q_1 \times x)^{\uparrow L} (x \rightarrow \emptyset + q_2 \times x)^{\uparrow L} l \end{array} \right\| \\
 &= \left\| \begin{array}{c} (x:P \rightarrow x + \emptyset) ; \text{pu}_L \\ q \times p \rightarrow \text{pu}_L(\emptyset^P + q \times (p + \emptyset)) \\ q_1 \times q_2 \times l \rightarrow (x^{M^A} \rightarrow \emptyset + q_1 \times (\emptyset + q_2 \times x))^{\uparrow L} l \end{array} \right\|.
 \end{aligned}$$

Now we need to post-compose $\text{ftn}_M^{\uparrow L}$ with this column:

$$\begin{aligned}
 \text{sw}^{\uparrow M} \circ \text{sw} \circ \text{ftn}_M^{\uparrow L} &= \left\| \begin{array}{c} (x:P \rightarrow x + 0) \circ \text{pu}_L \circ \text{ftn}_M^{\uparrow L} \\ (q \times p \rightarrow 0:P + q \times (p + 0)) \circ \text{pu}_L \circ \text{ftn}_M^{\uparrow L} \\ q_1 \times q_2 \times l \rightarrow l \triangleright (x:M^A \rightarrow 0 + q_1 \times (0 + q_2 \times x))^{\uparrow L} \circ \text{ftn}_M^{\uparrow L} \end{array} \right\| \\
 \text{pu}_L \text{'s naturality : } &= \left\| \begin{array}{c} (x:P \rightarrow x + 0) \circ \text{ftn}_M \circ \text{pu}_L \\ (q \times p \rightarrow 0:P + q \times (p + 0)) \circ \text{ftn}_M \circ \text{pu}_L \\ q_1 \times q_2 \times l \rightarrow l \triangleright (x:M^A \rightarrow \text{ftn}_M(0 + q_1 \times (0 + q_2 \times x)))^{\uparrow L} \end{array} \right\| \\
 \text{compute ftn}_M(\dots) : &= \left\| \begin{array}{c} (x:P \rightarrow \text{ftn}_M(x + 0)) \circ \text{pu}_L \\ (q \times p \rightarrow \text{ftn}_M(0:P + q \times (p + 0))) \circ \text{pu}_L \\ q_1 \times q_2 \times l \rightarrow l \triangleright (x:M^A \rightarrow 0 + (q_1 \oplus q_2) \times x)^{\uparrow L} \end{array} \right\| \\
 \text{compute ftn}_M(\dots) : &= \left\| \begin{array}{c} (x:P \rightarrow x + 0) \circ \text{pu}_L \\ (q \times p \rightarrow p + 0) \circ \text{pu}_L \\ q_1 \times q_2 \times l \rightarrow (x:M^A \rightarrow 0 + (q_1 \oplus q_2) \times x)^{\uparrow L} l \end{array} \right\| .
 \end{aligned}$$

After renaming l to a , this is the same as the column in Eq. (14.31).

Monadic naturality laws Verify the laws of Statement 14.3.8.2,

$$\text{sw}_{\text{Id},M} = \text{id} , \quad \text{sw}_{L,M} \circ \phi = \phi^{\uparrow M} \circ \text{sw}_{N,M} , \quad \text{sw}_{L,M} \circ \theta^{\uparrow L} = \theta .$$

for arbitrary monad morphisms $\phi : L \rightsquigarrow N$ and $\theta : M \rightsquigarrow \text{Id}$.

The first law is the swap identity law, $\text{sw}_{\text{Id},M} = \text{id}$:

$$\begin{aligned}
 \text{sw}_{\text{Id},M} &= \left\| \begin{array}{c} P \\ Q \times \text{Id}^A \end{array} \right\| \left\| \begin{array}{c} (x:P \rightarrow x + 0:Q \times A) \circ \text{pu}_{\text{Id}} \\ q \times l \rightarrow (a:A \rightarrow 0:P + q \times a)^{\uparrow \text{Id}} l \end{array} \right\| \\
 \text{Eq. (14.27) with } L = \text{Id} : &= \left\| \begin{array}{c} P \\ Q \times \text{Id}^A \end{array} \right\| \left\| \begin{array}{c} (x:P \rightarrow x + 0:Q \times A) \circ \text{pu}_{\text{Id}} \\ q \times l \rightarrow (a:A \rightarrow 0:P + q \times a)^{\uparrow \text{Id}} l \end{array} \right\| \\
 \text{matrix notation :} &= \left\| \begin{array}{c} P \\ Q \times A \end{array} \right\| \left\| \begin{array}{c} x:P \rightarrow x + 0:Q \times A \\ q \times a \rightarrow 0:P + q \times a \end{array} \right\| \\
 &= \text{id} .
 \end{aligned}$$

Begin with the left-hand side of the second law,

$$\begin{aligned}
 \text{sw}_{L,M} \circ \phi &= \left\| \begin{array}{c} (x:P \rightarrow x + 0:Q \times A) \circ \text{pu}_L \\ q \times l \rightarrow (a \rightarrow 0:P + q \times a)^{\uparrow L} l \end{array} \right\| \circ \phi \\
 \text{definition of } \text{sw}_{L,M} : &= \left\| \begin{array}{c} (x:P \rightarrow x + 0:Q \times A) \circ \text{pu}_L \\ q \times l \rightarrow (a \rightarrow 0:P + q \times a)^{\uparrow L} l \end{array} \right\| \circ \phi \\
 \text{compose with } \phi : &= \left\| \begin{array}{c} (x:P \rightarrow x + 0:Q \times A) \circ \text{pu}_L \circ \phi \\ q \times l \rightarrow l \triangleright (a \rightarrow 0:P + q \times a)^{\uparrow L} \circ \phi \end{array} \right\| \\
 \text{naturality of } \phi : &= \left\| \begin{array}{c} (x:P \rightarrow x + 0:Q \times A) \circ \text{pu}_N \\ q \times l \rightarrow l \triangleright \phi \circ (a \rightarrow 0:P + q \times a)^{\uparrow N} \end{array} \right\| .
 \end{aligned}$$

The right-hand side is

$$\begin{aligned}
 & \phi^{\uparrow M} ;_{\text{SW}_{N,M}} \\
 &= \left\| \begin{array}{cc} \text{id} & 0 \\ 0 & q \times l \rightarrow q \times \phi(l) \end{array} \right\| ; \left\| \begin{array}{c} (x:P \rightarrow x + 0:Q \times A) ;_{\text{pu}_N} \\ q \times n \rightarrow n \triangleright (a \rightarrow 0:P + q \times a)^{\uparrow N} \end{array} \right\| \\
 \text{composition :} &= \left\| \begin{array}{c} (x:P \rightarrow x + 0:Q \times A) ;_{\text{pu}_N} \\ q \times l \rightarrow n \triangleright \phi ;_{(a \rightarrow 0:P + q \times a)^{\uparrow N}} \end{array} \right\| .
 \end{aligned}$$

Both sides of the second law are now shown to be equal.

The left-hand side of the third law is

$$\begin{aligned}
 & \text{sw}_{L,M} ;_{\theta^{\uparrow L}} \\
 \text{compose with } \theta^{\uparrow L} : &= \left\| \begin{array}{c} (x:P \rightarrow x + 0:Q \times A) ;_{\text{pu}_L ; \theta^{\uparrow L}} \\ q \times l \rightarrow l \triangleright (a \rightarrow 0:P + q \times a)^{\uparrow L} ; \theta^{\uparrow L} \end{array} \right\| \\
 \text{naturality of } \text{pu}_L : &= \left\| \begin{array}{c} (x:P \rightarrow x + 0:Q \times A) ;_{\theta ; \text{pu}_L} \\ q \times l \rightarrow l \triangleright (a \rightarrow \theta(0:P + q \times a))^{\uparrow L} \end{array} \right\| .
 \end{aligned} \tag{14.32}$$

We expect this to equal the right-hand side, which we write as

$$\begin{aligned}
 & m^{M^{LA}} \rightarrow \theta(m) \\
 \text{matrix notation :} &= \left\| \begin{array}{c} x:P \rightarrow \theta(x + 0:Q \times L^A) \\ q \times l \rightarrow \theta(0:P + q \times l) \end{array} \right\| .
 \end{aligned} \tag{14.33}$$

Now consider each line in Eq. (14.32) separately. The upper line can be transformed as

$$\begin{aligned}
 & (x:P \rightarrow x + 0:Q \times A) ;_{\theta ; \text{pu}_L} \\
 \text{naturality of } \theta : &= (x:P \rightarrow x + 0:Q \times A) ;_{\text{pu}_L^{\uparrow M} ; \theta} \\
 \text{definition of } \uparrow M : &= x:P \rightarrow \left\| \begin{array}{cc} x & 0 \\ 0 & q \times l \rightarrow q \times \text{pu}_L l \end{array} \right\| ; \theta \left\| \begin{array}{cc} \text{id} & 0 \\ 0 & q \times l \rightarrow q \times \text{pu}_L l \end{array} \right\| ;_{\theta} \\
 \text{matrix notation :} &= x:P \rightarrow (x + 0:Q \times L^A) \triangleright \theta .
 \end{aligned}$$

This is now equal to the upper line of Eq. (14.33).

To proceed with the proof for the lower line of Eq. (14.32), we need to evaluate the monad morphism $\theta : M^A \rightarrow A$ on a specific value of type M^A of the form $0 + q \times a$. We note that the value $\theta(0 + q \times a)$ must be of type A and must be computed in the same way for all types A , because θ is a natural transformation. It seems clear that the result cannot depend on the value $q:Q$ since Q is a type not related to A . In other words, we expect that $\theta(0 + q \times a) = a$ as a consequence of naturality of θ . To derive this formally, we use the trick of starting with a unit type, $\mathbb{1}$, and mapping it to a within the naturality law. For any values $q:Q$, $a:A$, we define

$$\begin{aligned}
 m^{P+Q \times A} &\triangleq 0:P + q \times a , \\
 m_1^{P+Q \times \mathbb{1}} &\triangleq 0:P + q \times 1 .
 \end{aligned}$$

We can compute m from m_1 if we replace 1 by a under the functor M . To write this as a formula, define the function $f:\mathbb{1} \rightarrow A$ as $f \triangleq (_ \rightarrow a)$ using the fixed value a . Then we have $m = f^{\uparrow M} m_1$. Now

we apply both sides of the naturality law $f \uparrow^M \circ \theta = \theta \circ f$ to the value m_1 :

$$m_1 \triangleright f \uparrow^M \circ \theta = m_1 \triangleright \theta \circ f \quad .$$

Simplify the left-hand side to

$$m_1 \triangleright f \uparrow^M \circ \theta = \theta(f \uparrow^M m_1) = \theta(m) = \theta(0^P + q \times a) \quad .$$

Simplify the right-hand side to

$$m_1 \triangleright \theta \circ f = f(\theta(m_1)) = a \quad ,$$

since f always returns a . Therefore

$$\theta(0^P + q \times a) = a \quad . \quad (14.34)$$

We can now compute the second line in Eq. (14.32) as

$$\begin{aligned} q \times l \rightarrow l \triangleright (a \rightarrow \theta(0^P + q \times a))^{\uparrow L} \\ \text{use Eq. (14.34)} : = q \times l \rightarrow l \triangleright (a \rightarrow a)^{\uparrow L} \\ \text{identity law} : = q \times l \rightarrow l \quad . \end{aligned}$$

The second line in Eq. (14.33) is the same function, $q \times l \rightarrow l$.

This concludes the proof of the swap laws for linear monads. It follows that linear monads have monad transformers that compose inside the foreign monad.

14.4.3 Composition of transformers for linear monads

We have just shown that the linear monad $M^A \triangleq P + Q \times A$ has the `swap` function that satisfies the laws necessary for a composed-inside transformer. The other type of linear monad is $W^A \triangleq Q \times (P + A)$. Do we need to show separately that the monad W has a lawful `swap` function? Actually, that follows from the stacking property of monad transformers (see Sections 14.2.8–14.2.9). The monad W is a functor composition of the `Writer` monad $Q \times A$ with the `Either` monad $P + A$, which is the same as applying the `Either` monad's transformer to the `Writer` monad. Because of the transformer stacking property, the monad transformer of W works as composed-inside.

We can show in general that the functor composition of any two linear monads has a composed-inside transformer. Suppose M_1 and M_2 are linear monads, so their transformers are of the composed-inside kind:

$$T_{M_1}^N = N \circ M_1 \quad , \quad T_{M_2}^N = N \circ M_2 \quad .$$

The functor composition of M_1 and M_2 can be seen as a monad stack,

$$M_1 \circ M_2 = T_{M_2}^{M_1} \quad .$$

What is the transformer for the monad $M_1 \circ M_2$? For any foreign monad N , we have the transformer stack

$$T_{M_2}^{T_{M_1}^N} = T_{M_2}^{N \circ M_1} = N \circ (M_1 \circ M_2) \quad .$$

Since this is a transformer stack, it is a lawful monad transformer, as we have seen in Section 14.2.8. So, this is the monad transformer for $M_1 \circ M_2$, and it is of the composed-inside kind.

14.5 Composed-outside transformers: Rigid monads

Section 14.4 shows that the composed-inside monad transformers are available for a limited subset of monads, namely the monads of the form $M^A = P + Q \times A$ and $M^A = Q \times (P + A)$, called “linear”. No other examples of composed-inside transformers are known. It turns out that the composed-outside transformers are available for a significantly wider range of monads. Those monads are called “rigid” because one of their general properties is having a single “shape” (Statement 14.5.5.8 in Section 14.5.5). (There does not seem to be another already accepted name for monads of this kind.)

Definition of rigid monads A monad R is **rigid** if it has a lawful composed-outside monad transformer, $T_R^M = R \circ M$, where M is a foreign monad.

This definition just states the required properties but does not explain what monads are rigid or how to recognize a non-rigid monad. We will answer these questions in the rest of this section.

Two examples of rigid monads are the `Reader` monad and the `sel` (selector) monad,⁶

$$\begin{aligned} \text{(the Reader monad)} : \quad R^A &\triangleq Z \rightarrow A \quad , \\ \text{(the Sel monad)} : \quad S^A &\triangleq (A \rightarrow Z) \rightarrow A \quad , \end{aligned}$$

where Z is a fixed type. These monads have composed-outside transformers:

$$\begin{aligned} \text{(the ReaderT transformer)} : \quad T_R^{M,A} &\triangleq Z \rightarrow M^A \quad , \\ \text{(the SelT transformer)} : \quad T_S^{M,A} &\triangleq (M^A \rightarrow Z) \rightarrow M^A \quad . \end{aligned}$$

To build intuition for rigid monads, we will perform structural analysis looking for general constructions that create new rigid monads or combine existing rigid monads into new ones. In this section, we will prove that the following four constructions produce rigid monads:

1. Product: $P^A \times R^A$ is a rigid monad if P and R are rigid monads.
2. Composition: $P \circ R$ is a rigid monad if P and R are rigid monads.
3. Choice: $C^A \triangleq H^A \rightarrow A$ is a rigid monad if H is any contrafunctor.
4. Selector: $S^A \triangleq H^A \rightarrow R^A$ is a rigid monad for any rigid monad R and any R -filterable contrafunctor H .

Construction 3 is a special case of construction 4 because we can set $R = \text{Id}$ (the identity monad) in construction 4, and any contrafunctor H is Id -filterable (see Section 10.3.6). So, we will prove only constructions 1, 2, and 4.

It is not known whether these four constructions are the only possible ways of creating new rigid monads. Other open questions about rigid monads will be given below.

14.5.1 Rigid monad construction 1: choice

The construction called the **choice** monad, $R^A \triangleq H^A \rightarrow A$, defines a rigid monad R for *any* given contrafunctor H .

This monad chooses a value of type A given a contrafunctor H that may *consume* values of type A (and presumably could check some conditions on those values). The contrafunctor H could be a constant contrafunctor $H^A \triangleq Q$, a function such as $H^A \triangleq A \rightarrow Q$, or a more complicated contrafunctor.

Different choices of the contrafunctor H give specific examples of rigid monads, such as $R^A \triangleq \mathbb{1}$ (the unit monad), $R^A \triangleq A$ (the identity monad), $R^A \triangleq Z \rightarrow A$ (the reader monad), as well as the selector monad, $\text{Sel}^{Q,A} \triangleq (A \rightarrow Q) \rightarrow A$.

The selector monad represents the effect of selecting a value of type A that satisfies a condition expressed through a function of type $A \rightarrow Q$. The simplest example of a selector monad is found by setting $Q \triangleq \mathbb{2}$ (the `Boolean` type). One may implement a function of type $(A \rightarrow \mathbb{2}) \rightarrow A$ that *somehow* finds a value of type A that might satisfy the given predicate of type $A \rightarrow \mathbb{2}$. The intention is to return a value that, if possible, satisfies the predicate. If no such value can be found, *some* value of type A is still returned.

A closely related monad is the **search** monad, $R^A \triangleq (A \rightarrow \mathbb{2}) \rightarrow \mathbb{1} + A$. This (non-rigid) monad returns an empty value $\mathbb{1} + \mathbb{0}^A$ if no value satisfying the predicate is found. There is a monad morphism from the selector monad to the search monad, implemented by checking whether the value returned by the selector monad does actually satisfies the predicate.

⁶See <http://math.andrej.com/2008/11/21/>

Statement 14.5.1.1 Define the search monad as $\text{Search}^A \triangleq (A \rightarrow 2) \rightarrow \mathbb{1} + A$. The function defined by

```
def finder[A](s: (A => Boolean) => A): (A => Boolean) => Option[A]
```

		1	A
finder : $\text{Sel}^{2,A} \rightarrow \text{Search}^A$	$\text{finder}(s: \text{Sel}^{2,A} \rightarrow \text{Search}^A) \triangleq p: A \rightarrow 2 \rightarrow p(s(p)) \triangleright$	1 (false)	id 0
		1 (true)	0 $s(p)$

is a monad morphism $\text{Sel}^{2,\bullet} \leadsto \text{Search}^\bullet$.

Proof ***

Assume that H is a contrafunctor and M is a monad, and denote for brevity

$$T^A \triangleq R^{M^A} \triangleq H^{M^A} \rightarrow M^A .$$

We will first give a self-contained proof that T is a monad. To verify the laws of monad transformers for T , we will derive the `swap` function and verify its laws.

Statement 14.5.1.2 $T^\bullet \triangleq R^{M^\bullet} \triangleq H^{M^\bullet} \rightarrow M^\bullet$ is a monad if M is any monad and H is any contrafunctor. (If we set $M^A \triangleq A$, this also proves that R itself is a monad.)

Proof We need to define the monad instance for T and prove the identity and the associativity laws for T , assuming that the monad M satisfies these laws.

To define the monad instance for T , it is convenient to use the Kleisli formulation of the monad. In this formulation, we consider Kleisli morphisms of type $A \rightarrow T^B$ and then define the Kleisli identity morphism, $\text{pu}_T : A \rightarrow T^A$, and the Kleisli product operation \diamond_T ,

$$f: A \rightarrow T^B \diamond_T g: B \rightarrow T^C : A \rightarrow T^C .$$

We are then required to define the operation \diamond_T and to prove identity and associativity laws for it.

We notice that since the type constructor R is itself a function type $H^A \rightarrow A$, the type of the Kleisli morphism $A \rightarrow T^B$ is actually $A \rightarrow T^B \triangleq A \rightarrow H^{M^B} \rightarrow M^B$. While proving the monad laws for T , we will need to use the monad laws for M (since M is an arbitrary, unknown monad). In order to use the monad laws for M , it would be helpful if we had the Kleisli morphisms for M of type $A \rightarrow M^B$ more easily available. If we flip the curried arguments of the Kleisli morphism type $A \rightarrow H^{M^B} \rightarrow M^B$ and instead consider the **flipped Kleisli** morphisms of type $H^{M^B} \rightarrow A \rightarrow M^B$, the type $A \rightarrow M^B$ will be easier to reason about. Since the type $A \rightarrow H^{M^B} \rightarrow M^B$ is equivalent to $A \rightarrow H^{M^B} \rightarrow M^B$, any laws we prove for the flipped Kleisli morphisms will yield the corresponding laws for the standard Kleisli morphisms. The use of flipped Kleisli morphisms makes the proof significantly shorter.

We temporarily denote by $\tilde{\text{pu}}_T$ and $\tilde{\diamond}_T$ the flipped Kleisli operations:

$$\begin{aligned} \tilde{\text{pu}}_T : H^{M^A} &\rightarrow A \rightarrow M^A \\ f: H^{M^B} \rightarrow A \rightarrow M^B \tilde{\diamond}_T g: H^{M^C} \rightarrow B \rightarrow M^C &: H^{M^C} \rightarrow A \rightarrow M^C . \end{aligned}$$

To define the operations $\tilde{\text{pu}}_T$ and $\tilde{\diamond}_T$, we may use the methods pu_M and flm_M as well as the Kleisli product \diamond_M for the given monad M . The definitions are

$$\begin{aligned} \tilde{\text{pu}}_T &= - \rightarrow \text{pu}_M \quad (\text{the argument is unused}) , \\ f \tilde{\diamond}_T g &= q \rightarrow (f p) \diamond_M (g q) \quad \text{where} \\ p: H^{M^B} &= (\text{flm}_M (g q)) \downarrow^H q . \end{aligned}$$

This definition works by using the Kleisli product \diamond_M on values $f p : A \rightarrow M^B$ and $g q : B \rightarrow M^C$. To obtain a value $p : H^{M^B}$, we use the function $\text{flm}_M (g q) : M^B \rightarrow M^C$ to H -contramap $q : H^{M^C}$ into $p : H^{M^B}$.

Written as a single expression, the definition of $\tilde{\diamond}_T$ is

$$f \tilde{\diamond}_T g = q \rightarrow f \left((\text{flm}_M (g q))^{\downarrow H} q \right) \diamond_M (g q) \quad . \quad (14.35)$$

Checking the left identity law:

$$\begin{aligned} & \tilde{\text{pu}}_T \tilde{\diamond}_T g \\ \text{definition of } \tilde{\diamond}_T : &= q \rightarrow \tilde{\text{pu}}_T \left((\text{flm}_M (g q))^{\downarrow H} q \right) \diamond_M (g q) \\ \text{definition of } \tilde{\text{pu}}_T : &= q \rightarrow \text{pu}_M \diamond_M g q \\ \text{left identity law for } M : &= q \rightarrow g q \\ \text{function expansion :} &= g \end{aligned}$$

Checking the right identity law:

$$\begin{aligned} & f \tilde{\diamond}_T \tilde{\text{pu}}_T \\ \text{definition of } \tilde{\diamond}_T : &= q \rightarrow f \left((\text{flm}_M (\tilde{\text{pu}}_T q))^{\downarrow H} q \right) \diamond_M (\tilde{\text{pu}}_T q) \\ \text{definition of } \tilde{\text{pu}}_T : &= q \rightarrow f \left((\text{flm}_M (\text{pu}_M))^{\downarrow H} q \right) \diamond_M \text{pu}_M \\ \text{right identity law for } M : &= q \rightarrow f \left((\text{id})^{\downarrow H} q \right) \\ \text{identity law for } H : &= q \rightarrow f q \\ \text{function expansion :} &= f \end{aligned}$$

Checking the associativity law: $(f \tilde{\diamond}_T g) \tilde{\diamond}_T h$ must equal $f \tilde{\diamond}_T (g \tilde{\diamond}_T h)$. We have

$$\begin{aligned} & (f \tilde{\diamond}_T g) \tilde{\diamond}_T h \\ &= (s \rightarrow f \left((\text{flm}_M (g s))^{\downarrow H} s \right) \diamond_M (g s)) \tilde{\diamond}_T h \\ &= q \rightarrow f \left((\text{flm}_M (g r))^{\downarrow H} r \right) \diamond_M (g r) \diamond_M (h q) \quad \text{where} \\ & \quad r \triangleq (\text{flm}_M (h q))^{\downarrow H} q \quad ; \end{aligned}$$

while

$$\begin{aligned} & f \tilde{\diamond}_T (g \tilde{\diamond}_T h) \\ &= f \tilde{\diamond}_T \left(q \rightarrow g \left((\text{flm}_M (h q))^{\downarrow H} q \right) \diamond_M (h q) \right) \\ &= q \rightarrow f \left((\text{flm}_M k)^{\downarrow H} q \right) \diamond_M u \quad \text{where} \\ & \quad r \triangleq (\text{flm}_M (h q))^{\downarrow H} q \quad \text{and} \\ & \quad u \triangleq (g r) \diamond_M (h q) \quad . \end{aligned}$$

It remains to show that the following two expressions are equal,

$$\begin{aligned} & f \left((\text{flm}_M (g r))^{\downarrow H} r \right) \diamond_M (g r) \diamond_M (h q) \quad \text{and} \\ & f \left((\text{flm}_M ((g r) \diamond_M (h q)))^{\downarrow H} q \right) \diamond_M (g r) \diamond_M (h q), \quad \text{where} \\ & \quad r \triangleq (\text{flm}_M (h q))^{\downarrow H} q \quad . \end{aligned}$$

These two expressions differ only by the following sub-expressions,

$$(\text{flm}_M (g r))^{\downarrow H} r$$

and

$$(\text{flm}_M ((g r) \diamond_M (h q)))^{\downarrow H} q \quad ,$$

where $r \triangleq (\text{flm}_M (h q))^{\downarrow H} q$. Writing out the value r in the last argument of $(\text{flm}_M (g r))^{\downarrow H} r$ but leaving r unexpanded everywhere else, we now rewrite the differing sub-expressions as

$$\begin{aligned} & (\text{flm}_M (g r))^{\downarrow H} (\text{flm}_M (h q))^{\downarrow H} q \quad \text{and} \\ & (\text{flm}_M ((g r) \diamond_M (h q)))^{\downarrow H} q \quad . \end{aligned}$$

Now it becomes apparent that we need to put the two “ flm_M ”s closer together and to combine them by using the associativity law of the monad M . Then we can rewrite the first sub-expression and transform it into the second one:

$$\begin{aligned} & \frac{(\text{flm}_M (g r))^{\downarrow H} (\text{flm}_M (h q))^{\downarrow H} q}{\text{composition law for } H : \quad = (\text{flm}_M (g r) ; \text{flm}_M (h q))^{\downarrow H} q} \\ & \text{associativity law for } M : \quad = \underline{(\text{flm}_M ((g r) ; \text{flm}_M (h q)))^{\downarrow H} q} \\ & \text{definition of } \diamond_M \text{ via } \text{flm}_M : \quad = (\text{flm}_M ((g r) \diamond_M (h q)))^{\downarrow H} q \quad . \end{aligned}$$

This proves the associativity law for \diamond_T .

Statement 14.5.1.3 The monad methods for T defined in Statement 14.5.1.2 can be written equivalently as

$$\begin{aligned} \text{pu}_T(a^{:A}) &: H^{M^A} \rightarrow M^A \quad , \\ \text{pu}_T(a) &\triangleq (_ \rightarrow \text{pu}_M a) \quad ; \\ \text{flm}_T(f^{:A \rightarrow H^{M^B} \rightarrow M^B}) &: (H^{M^A} \rightarrow M^A) \rightarrow H^{M^B} \rightarrow M^B \quad , \\ \text{flm}_T f &\triangleq t^{:R^{M^A} \rightarrow q^{:H^{M^B} \rightarrow (\text{flm}_M(x^{:A} \rightarrow f x q))^{\uparrow R} t q}} \quad . \end{aligned} \tag{14.36}$$

Expressed through R ’s `flatMap` method, which is implemented as

$$\text{flm}_R g^{:A \rightarrow R^B} = t^{:R^A \rightarrow q^{:H^B \rightarrow (x^{:A} \rightarrow g x q))^{\uparrow R} t q}} \quad , \tag{14.37}$$

the method flm_T can be written as

$$\text{flm}_T f = \text{flm}_R(y \rightarrow q \rightarrow \text{flm}_M(x \rightarrow f x q))y \quad . \tag{14.38}$$

Proof The definition of \diamond_T in Statement 14.5.1.2 used the flipped types of Kleisli morphisms, which is not the standard way of defining the methods of a monad. To restore the standard type signatures, we need to unflip the arguments:

$$\begin{aligned} f^{:A \rightarrow H^{M^B} \rightarrow M^B} \diamond_T g^{:B \rightarrow H^{M^C} \rightarrow M^C} &: A \rightarrow H^{M^C} \rightarrow M^C \quad ; \\ f \diamond_T g &= t \rightarrow q \rightarrow \left(\tilde{f} \left((\text{flm}_M(b \rightarrow g b q))^{\downarrow H} q \right) \diamond_M (b \rightarrow g b q) \right) t \quad , \end{aligned}$$

where $\tilde{f} \triangleq h \rightarrow k \rightarrow f k h$ is the flipped version of f . To replace \diamond_M by flm_M , express $x \diamond_M y = x ; \text{flm}_M y$ to find

$$f \diamond_T g = t \rightarrow q \rightarrow \left(\tilde{f} \left(p^{\downarrow H} q \right) ; p \right) t \quad \text{where } p = \text{flm}_M(x \rightarrow g x q) \quad .$$

To obtain an implementation of flm_T , express flm_T through \diamond_T as

$$\text{flm}_T g^{:A \rightarrow T^B} = \text{id}^{:T^A \rightarrow T^A} \diamond_T g \quad .$$

Now we need to substitute $f^{:T^A \rightarrow T^A} = \text{id}$ into $f \diamond_T g$. Noting that \tilde{f} will then become

$$\tilde{f} = (h \rightarrow k \rightarrow \text{id } k \ h) = (h \rightarrow k \rightarrow k \ h) \quad ,$$

we get

$$\begin{aligned} \text{flm}_T g^{:A \rightarrow T^B} &= \text{id} \circ \text{flm}_T g \\ \text{definition of } \diamond_T : &= t^{:T^A} \rightarrow q^{H^{M^B}} \rightarrow \left(\tilde{f} \left(p^{\downarrow H} q \right) \circ p \right) t \\ &\quad \text{where } p \triangleq \text{flm}_M (x \rightarrow g \ x \ q) \\ \text{substitute } f = \text{id} : &= t \rightarrow q \rightarrow \left((h \rightarrow k \rightarrow k \ h) \left(p^{\downarrow H} q \right) \circ p \right) t \\ \text{apply } k \text{ to } p^{\downarrow H} q : &= t \rightarrow q \rightarrow \left(\left(k \rightarrow k \left(p^{\downarrow H} q \right) \right) \circ p \right) t \\ \text{definition of } \circ : &= t \rightarrow q \rightarrow p \left(t \left(p^{\downarrow H} q \right) \right) \quad . \end{aligned}$$

By definition of the functor $R^A \triangleq H^A \rightarrow A$, we raise any function $p^{:A \rightarrow B}$ into R as

$$\begin{aligned} p^{\uparrow R} : (H^A \rightarrow A) &\rightarrow H^B \rightarrow B \quad , \\ p^{\uparrow R} r^{H^A \rightarrow A} &\triangleq p^{\downarrow H} \circ r \circ p \\ &= q^{H^B} \rightarrow p \left(r \left(p^{\downarrow H} q \right) \right) \quad . \end{aligned}$$

Finally, renaming g to f , we obtain the desired code,

$$\text{flm}_T f = t \rightarrow q \rightarrow p^{\uparrow R} t \ q \quad \text{where } p \triangleq \text{flm}_M (x \rightarrow f \ x \ q) \quad .$$

To express flm_T via flm_R , we just need to choose the value of g such that Eq. (14.37) becomes equal to Eq. (14.36). Comparing these two expressions, we find that we need

$$\text{flm}_M (x \rightarrow f \ x \ q) = (y \rightarrow g \ y \ q) \quad .$$

This is achieved if we define $g \ y \ q = \text{flm}_M (x^{:A} \rightarrow f \ x \ q) \ y$, or equivalently

$$g = y \rightarrow q \rightarrow \text{flm}_M (x \rightarrow f \ x \ q) \ y \quad .$$

This gives the desired Eq. (14.38).

Statement 14.5.1.4 The monad T has the methods `flatten` and `swap` defined by

$$\text{ftn}_T = t^{:T^{T^A}} \rightarrow q^{:H^{M^A}} \rightarrow q \triangleright \left(t \triangleright \left(\text{flm}_M (x^{:R^{M^A}} \rightarrow x \ q) \right)^{\uparrow R} \right) \quad , \quad (14.39)$$

$$\text{sw}_{R,M} = m^{:M^{R^A}} \rightarrow q^{:H^{M^A}} \rightarrow \left(r^{:R^A} \rightarrow r(\text{pu}_M^{\downarrow H} q) \right)^{\uparrow M} m \quad . \quad (14.40)$$

These functions are computationally equivalent (can be derived from each other). In the \triangleright -notation, the formula for $\text{sw}_{R,M}$ is

$$q \triangleright (m \triangleright \text{sw}_{R,M}) = m \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M} \quad . \quad (14.41)$$

Proof Using Eq. (14.36) and the relationship $\text{ftn}_T = \text{flm}_T \text{id}^{:T^A \rightarrow T^A}$, we find

$$\begin{aligned} \text{ftn}_T t^{:T^A} &= \text{flm}_T (\text{id}) t \\ \text{use Eq. (14.36)} : &= q \rightarrow (\text{flm}_M (x^{:A} \rightarrow f \ x \ q))^{\uparrow R} t \ q \\ \text{definition of } \triangleright : &= q \rightarrow q \triangleright (t \triangleright (\text{flm}_M (x^{:A} \rightarrow f \ x \ q))^{\uparrow R}) \quad . \end{aligned}$$

Using Eq. (14.38) instead of Eq. (14.36), we get

$$\begin{aligned} \text{ftn}_T &= \text{flm}_T(\text{id}) \\ &= \text{flm}_R(y \rightarrow q \rightarrow \text{flm}_M(x \rightarrow x q) y) \quad . \end{aligned}$$

Deriving the formulas for $\text{sw}_{R,M}$ We start with ftn_T as just obtained and substitute into Eq. (14.21):

$$\begin{aligned} \text{sw}_{R,M}(m) &= m \triangleright \text{pu}_M^{\uparrow R \uparrow M} ; \text{pu}_R ; \text{ftn}_T \\ \text{use Eq. (14.38)} : &= m \triangleright \text{pu}_M^{\uparrow R \uparrow M} ; \text{pu}_R ; \text{flm}_R(y \rightarrow q \rightarrow \text{flm}_M(x \rightarrow x q) y) \\ \text{left identity law of } R : &= m \triangleright \text{pu}_M^{\uparrow R \uparrow M} ; (y \rightarrow q \rightarrow \text{flm}_M(x \rightarrow x q) y) \\ \triangleright \text{ notation} : &= m \triangleright \text{pu}_M^{\uparrow R \uparrow M} ; (y \rightarrow q \rightarrow y \triangleright \text{flm}_M(x \rightarrow x q)) \quad (14.42) \end{aligned}$$

$$\text{apply to argument } y : = q \rightarrow m \triangleright \text{pu}_M^{\uparrow R \uparrow M} ; \text{flm}_M(x \rightarrow x q)$$

$$\text{express } \text{flm}_M \text{ via } \text{ftn}_M : = q \rightarrow m \triangleright \text{pu}_M^{\uparrow R \uparrow M} ; (x \rightarrow x q)^{\uparrow M} ; \text{ftn}_M$$

$$\text{composition law of } M : = q \rightarrow m \triangleright (\text{pu}_M^{\uparrow R} ; (x \rightarrow x q))^{\uparrow M} ; \text{ftn}_M \quad (14.43)$$

It appears that simplifying this expression requires to rewrite the function $\text{pu}_M^{\uparrow R} ; (x \rightarrow x q)$. To proceed further, we need to use the definition of raising a function $f : A \rightarrow B$ to the functor R ,

$$f^{\uparrow R} \triangleq r : R^A \rightarrow f \downarrow^H ; r ; f \quad ,$$

so we can write

$$\begin{aligned} &\text{pu}_M^{\uparrow R} ; (x \rightarrow x q) \\ \text{function composition} : &= r \rightarrow \text{pu}_M^{\uparrow R} r q \\ \text{definition of } \uparrow^R : &= r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; r ; \text{pu}_M \\ \text{forward composition} : &= (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; r) ; \text{pu}_M \quad (14.44) \end{aligned}$$

In deriving Eq. (14.44), we used the general property of the forward composition,

$$x \rightarrow y \triangleright f(x, y) ; g = (x \rightarrow y \triangleright f(x, y)) ; g \quad ,$$

where g must not depend on x or y . We can now rewrite Eq. (14.43) as

$$\begin{aligned} &q \rightarrow m \triangleright (\text{pu}_M^{\uparrow R} ; (x \rightarrow x q))^{\uparrow M} ; \text{ftn}_M \\ \text{use Eq. (14.44)} : &= q \rightarrow m \triangleright ((r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; r) ; \text{pu}_M)^{\uparrow M} ; \text{ftn}_M \\ \text{functor composition for } M : &= q \rightarrow m \triangleright ((r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; r))^{\uparrow M} ; \text{pu}_M^{\uparrow M} ; \text{ftn}_M \\ \text{identity law of } M : &= q \rightarrow m \triangleright ((r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; r))^{\uparrow M} \\ \triangleright \text{ notation} : &= q \rightarrow (r \rightarrow r(\text{pu}_M^{\downarrow H} q))^{\uparrow M} m \quad . \end{aligned}$$

The last expression coincides with Eq. (14.40).

The formula (14.41) follows by applying Eq. (14.40) to the arguments m and q . To make the computation clearer, we rename the bound variables m and q inside Eq. (14.40) to m_1 and q_1 :

$$\begin{aligned} &q \triangleright (m \triangleright (m_1 \rightarrow q_1 \rightarrow (r \rightarrow r(\text{pu}_M^{\downarrow H} q_1))^{\uparrow M} m_1)) \\ \text{apply to argument } m : &= q \triangleright (q_1 \rightarrow m \triangleright (r \rightarrow r(\text{pu}_M^{\downarrow H} q_1))^{\uparrow M}) \\ \text{apply to argument } q : &= m \triangleright (r \rightarrow r(\text{pu}_M^{\downarrow H} q))^{\uparrow M} \\ \triangleright \text{ notation} : &= m \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; r)^{\uparrow M} \quad . \end{aligned}$$

Deriving ftn_T from $\text{sw}_{R,M}$ Given the swap function defined by Eq. (14.40), we can recover the original ftn_T function from Eq. (14.39) via the standard formula (14.11), $\text{ftn}_T = \text{sw}^{\uparrow R} \circ \text{ftn}_R \circ \text{ftn}_M^{\uparrow R}$:

$$\begin{aligned}
 & \text{naturality of } \text{ftn}_R : \quad \underline{\text{sw}^{\uparrow R} \circ \text{ftn}_R \circ \text{ftn}_M^{\uparrow R}} \\
 & \text{composition under } R : \quad = \underline{\text{sw}^{\uparrow R} \circ \text{ftn}_M^{\uparrow R \uparrow R} \circ \text{ftn}_R} \\
 & \text{relating } \text{flm}_R \text{ and } \text{ftn}_R : \quad = \underline{\text{flm}_R(\text{sw} \circ \text{ftn}_M^{\uparrow R})} \\
 & \text{use Eq. (14.37)} : \quad = t \rightarrow q \rightarrow (x^A \rightarrow (\text{sw} \circ \text{ftn}_M^{\uparrow R}) x q)^{\uparrow R} t q \quad . \tag{14.45}
 \end{aligned}$$

To proceed, we need to transform $\text{sw} \circ \text{ftn}_M^{\uparrow R}$ in some way:

$$\begin{aligned}
 & \text{sw} \circ \text{ftn}_M^{\uparrow R} \\
 & \text{definitions} : \quad = (m \rightarrow q \rightarrow m \triangleright ((r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r))^{\uparrow M}) \circ \underline{(r \rightarrow \text{ftn}_M^{\downarrow H} \circ r \circ \text{ftn}_M)} \\
 & \text{composition} : \quad = m \rightarrow \underline{\text{ftn}_M^{\downarrow H} \circ (q \rightarrow m \triangleright ((r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r))^{\uparrow M})} \circ \text{ftn}_M \\
 & \text{expansion} : \quad = m \rightarrow (q \rightarrow \underline{q \triangleright \text{ftn}_M^{\downarrow H}}) \circ (q \rightarrow m \triangleright ((r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r))^{\uparrow M}) \circ \text{ftn}_M \\
 & \text{composition} : \quad = m \rightarrow (q \rightarrow m \triangleright ((r \rightarrow \underline{q \triangleright \text{ftn}_M^{\downarrow H} \triangleright \text{pu}_M^{\downarrow H} \circ r})^{\uparrow M}) \circ \text{ftn}_M \quad . \tag{14.46}
 \end{aligned}$$

We can transform the sub-expression $(r \rightarrow q \triangleright \text{ftn}_M^{\downarrow H} \triangleright \text{pu}_M^{\downarrow H} \circ r)$ to

$$\begin{aligned}
 & \triangleright \text{ notation} : \quad r \rightarrow q \triangleright \underline{\text{ftn}_M^{\downarrow H} \circ \text{pu}_M^{\downarrow H} \circ r} \\
 & \text{composition law of } H : \quad = r \rightarrow q \triangleright (\text{pu}_M \circ \text{ftn}_M)^{\downarrow H} \circ r \\
 & \text{left identity law of } M : \quad = r \rightarrow \underline{q \triangleright r} \\
 & \triangleright \text{ notation} : \quad = r \rightarrow r(q) \quad . \tag{14.47}
 \end{aligned}$$

Using this simplification, we continue transforming Eq. (14.46) as

$$\begin{aligned}
 & m \rightarrow (q \rightarrow m \triangleright ((r \rightarrow \underline{q \triangleright \text{ftn}_M^{\downarrow H} \triangleright \text{pu}_M^{\downarrow H} \circ r})^{\uparrow M}) \circ \text{ftn}_M \\
 & \text{use Eq. (14.47)} : \quad = m \rightarrow (\underline{q \rightarrow m \triangleright (r \rightarrow r(q))^{\uparrow M}}) \circ \text{ftn}_M \\
 & \text{composition} : \quad = m \rightarrow q \rightarrow m \triangleright (r \rightarrow r(q))^{\uparrow M} \circ \text{ftn}_M \\
 & \text{relating } \text{flm}_M \text{ and } \text{ftn}_M : \quad = m \rightarrow q \rightarrow \text{flm}_M(r \rightarrow r(q))m \quad .
 \end{aligned}$$

Substituting this instead of $\text{sw} \circ \text{ftn}_M^{\uparrow R}$ into Eq. (14.45), we get

$$\begin{aligned}
 & t \rightarrow q \rightarrow (x \rightarrow (\underline{\text{sw} \circ \text{ftn}_M^{\uparrow R}}) x q)^{\uparrow R} t q \\
 & = t \rightarrow q \rightarrow (x \rightarrow \text{flm}_M(r \rightarrow r(q)) x)^{\uparrow R} t q \quad .
 \end{aligned}$$

The last expression is the same as Eq. (14.39).

Statement 14.5.1.5 Without assuming the monad laws for the function ftn_T , the laws in Statement 14.3.2.1 hold for the swap function defined by Eq. (14.40).

Proof After replacing the base monad L by R , the required laws are

$$\begin{aligned}
 & \text{pu}_R^{\uparrow M} \circ \text{sw} = \text{pu}_R \quad , \quad \text{pu}_M \circ \text{sw} = \text{pu}_M^{\uparrow R} \quad , \\
 & \text{ftn}_R^{\uparrow M} \circ \text{sw} = \text{sw} \circ \text{sw}^{\uparrow R} \circ \text{ftn}_R \quad , \quad \text{ftn}_M \circ \text{sw} = \text{sw}^{\uparrow M} \circ \text{sw} \circ \text{ftn}_M^{\uparrow R} \quad .
 \end{aligned}$$

Proof of the inner-identity law Compute

$$\begin{aligned}
 & \text{pu}_R^{\uparrow M} ; \text{sw} \\
 \text{use Eq. (14.40)} : &= (m \rightarrow \underline{m} ; \text{pu}_R^{\uparrow M}) ; (m \rightarrow q \rightarrow \underline{m} ; (r \rightarrow q ; \text{pu}_M^{\downarrow H} ; r))^{\uparrow M} \\
 \text{function composition} : &= m \rightarrow q \rightarrow m ; \underline{\text{pu}_R^{\uparrow M} ; (r \rightarrow q ; \text{pu}_M^{\downarrow H} ; r))^{\uparrow M}} \\
 \text{functor law of } M : &= m \rightarrow q \rightarrow m ; (\text{pu}_R ; (r \rightarrow q ; \text{pu}_M^{\downarrow H} ; r))^{\uparrow M} \quad . \tag{14.48}
 \end{aligned}$$

To proceed, we simplify the expression $\text{pu}_R ; (r \rightarrow \dots)$:

$$\begin{aligned}
 & \text{pu}_R ; (r \rightarrow q ; \text{pu}_M^{\downarrow H} ; r) \\
 \text{argument expansion} : &= (m \rightarrow m ; \text{pu}_R) ; (r \rightarrow q ; \text{pu}_M^{\downarrow H} ; r) \\
 \text{function composition} : &= m \rightarrow q ; \text{pu}_M^{\downarrow H} ; (m ; \text{pu}_R) \quad . \tag{14.49}
 \end{aligned}$$

We now have to use the definition of pu_R , which is $\text{pu}_R = x \rightarrow y \rightarrow x$, or in the pipe notation,

$$y ; (x ; \text{pu}_R) = x \quad . \tag{14.50}$$

With this simplification at hand, we continue from Eq. (14.49) to

$$\begin{aligned}
 & m \rightarrow \underline{q ; \text{pu}_M^{\downarrow H}} ; (m ; \text{pu}_R) \\
 \text{use Eq. (14.50)} : &= m \rightarrow m = \text{id} \quad .
 \end{aligned}$$

Therefore, Eq. (14.48) becomes

$$\begin{aligned}
 & m \rightarrow q \rightarrow m ; \underline{(\text{pu}_R ; (r \rightarrow q ; \text{pu}_M^{\downarrow H} ; r))^{\uparrow M}} \\
 &= (m \rightarrow q \rightarrow \underline{m ; \text{id}}) \\
 &= (m \rightarrow q \rightarrow m) = \text{pu}_R \quad .
 \end{aligned}$$

This proves the inner-identity law.

Proof of the outer-identity law The left-hand side of this law is

$$\begin{aligned}
 & \text{pu}_M ; \text{sw} \\
 \text{Eq. (14.40)} : &= (m \rightarrow \underline{m ; \text{pu}_M}) ; (m \rightarrow q \rightarrow \underline{m ; (r \rightarrow q ; \text{pu}_M^{\downarrow H} ; r))^{\uparrow M}}) \\
 \text{function composition} : &= m \rightarrow q \rightarrow m ; \underline{\text{pu}_M ; (r \rightarrow q ; \text{pu}_M^{\downarrow H} ; r))^{\uparrow M}} \\
 \text{naturality of } \text{pu}_M : &= m \rightarrow q \rightarrow m ; (r \rightarrow q ; \text{pu}_M^{\downarrow H} ; r) ; \text{pu}_M \\
 \triangleright \text{ notation} : &= m \rightarrow q \rightarrow m ; (r \rightarrow q ; \text{pu}_M^{\downarrow H} ; r ; \text{pu}_M) \\
 \text{apply function to } m : &= m \rightarrow q \rightarrow q ; \text{pu}_M^{\downarrow H} ; m ; \text{pu}_M \\
 \text{argument expansion} : &= m \rightarrow \text{pu}_M^{\downarrow H} ; m ; \text{pu}_M \\
 \text{definition of } \uparrow^R : &= \text{pu}_M^{\uparrow R} \quad .
 \end{aligned}$$

This is equal to the right-hand side of the law.

Proof of the inner-interchange law The law is written as

$$\text{ftn}_R^{\uparrow M} ; \text{sw} = \text{sw} ; \text{sw}^{\uparrow R} ; \text{ftn}_R \quad . \tag{14.51}$$

We will apply both sides of the law to arbitrary $m^{\cdot M^R}{}^{RA}$ and $q^{\cdot H^M}{}^A$, and transform both sides to the same expression.

Below, we will need a simplified formula for ftn_R derived from Eq. (14.37):

$$\begin{aligned}
 \text{ftn}_R &= \text{flm}_R(\text{id}) \\
 \text{use Eq. (14.37)} : &= t \rightarrow q \rightarrow (x \rightarrow x q)^{\uparrow R} t q \\
 \text{definition of } \uparrow^R : &= t \rightarrow q \rightarrow (\underline{r \rightarrow (x \rightarrow x q)^{\downarrow H} ; r ; (x \rightarrow x)}) t q \\
 \text{apply to argument} : &= t \rightarrow q \rightarrow ((x \rightarrow q \triangleright x)^{\downarrow H} ; t ; (x \rightarrow x q)) q \\
 \text{use } \triangleright \text{ notation} : &= t \rightarrow q \rightarrow \underline{q \triangleright (q \triangleright (x \rightarrow q \triangleright x)^{\downarrow H} ; t)} \quad . \tag{14.52}
 \end{aligned}$$

We first apply the left-hand side of the law (14.51) to m and q :

$$\begin{aligned}
 q \triangleright (m \triangleright \text{ftn}_R^{\uparrow M} ; \text{sw}) \\
 \triangleright \text{ notation} : &= q \triangleright (m \triangleright \text{ftn}_R^{\uparrow M} \triangleright \text{sw}) \\
 \text{use Eq. (14.41)} : &= m \triangleright \underline{\text{ftn}_R^{\uparrow M} ; (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; r)^{\uparrow M}} \\
 \text{composition law for } M : &= m \triangleright (\text{ftn}_R ; (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; r))^{\uparrow M} \quad .
 \end{aligned}$$

We now need to simplify the sub-expression under $(\dots)^{\uparrow M}$:

$$\begin{aligned}
 \text{ftn}_R ; (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \triangleright r) \\
 \text{function composition} : &= r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \triangleright \underline{\text{ftn}_R(r)} \\
 \text{use Eq. (14.52)} : &= r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \triangleright (q \triangleright \underline{\text{pu}_M^{\downarrow H} \triangleright (x \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \triangleright x)^{\downarrow H} ; r}) \\
 \text{composition law for } H : &= r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; (q \triangleright (x \rightarrow q \triangleright \underline{\text{pu}_M^{\downarrow H} ; x ; \text{pu}_M})^{\downarrow H} ; r) \\
 \text{definition of } \uparrow^R : &= r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; (q \triangleright (x \rightarrow q \triangleright \text{pu}_M^{\uparrow R}(x))^{\downarrow H} ; r) \quad .
 \end{aligned}$$

The left-hand side of the law (14.51) then becomes

$$m \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} ; (q \triangleright (x \rightarrow q \triangleright \text{pu}_M^{\uparrow R}(x))^{\downarrow H} ; r))^{\uparrow M} \quad .$$

Now apply the right-hand side of the law (14.51) to m and q :

$$\begin{aligned}
 q \triangleright (m \triangleright \text{sw} ; \text{sw}^{\uparrow R} ; \text{ftn}_R) \\
 \text{definition of } \uparrow^R : &= q \triangleright (\underline{m \triangleright \text{sw} \triangleright (x \rightarrow \text{sw}^{\downarrow H} ; x ; \text{sw})} \triangleright \text{ftn}_R) \\
 \text{apply to arguments} : &= q \triangleright (\underline{\text{ftn}_R(\text{sw}^{\downarrow H} ; \text{sw}(m) ; \text{sw})}) \\
 \text{use Eq. (14.52)} : &= q \triangleright (q \triangleright (\underline{x \rightarrow q \triangleright x}^{\downarrow H} ; \text{sw}^{\downarrow H} ; \text{sw}(m) ; \text{sw})) \\
 \text{composition law of } H : &= q \triangleright (q \triangleright (\text{sw} ; (x \rightarrow q \triangleright x))^{\downarrow H} ; \text{sw}(m) ; \text{sw}) \quad . \tag{14.53}
 \end{aligned}$$

To proceed, we simplify the sub-expression $\text{sw}(m) ; \text{sw}$ separately by computing the function compositions:

$$\begin{aligned}
 \text{sw}(m) ; \text{sw} \\
 &= (q_1 \rightarrow m \triangleright (r \rightarrow q_1 \triangleright \text{pu}_M^{\downarrow H} ; r)^{\uparrow M}) ; (y \rightarrow q_2 \rightarrow y \triangleright (r \rightarrow q_2 \triangleright \text{pu}_M^{\downarrow H} ; r)^{\uparrow M}) \\
 &= q_1 \rightarrow q_2 \rightarrow (m \triangleright (r \rightarrow q_1 \triangleright \text{pu}_M^{\downarrow H} ; r)^{\uparrow M}) \triangleright (r \rightarrow q_2 \triangleright \text{pu}_M^{\downarrow H} ; r)^{\uparrow M} \\
 &= q_1 \rightarrow q_2 \rightarrow m \triangleright ((r \rightarrow q_1 \triangleright \text{pu}_M^{\downarrow H} ; r) ; (r \rightarrow q_2 \triangleright \text{pu}_M^{\downarrow H} ; r))^{\uparrow M} \quad .
 \end{aligned}$$

Using this formula, we can write, for any z of a suitable type,

$$q \triangleright (z \triangleright \text{sw}(m) \circ \text{sw}) = m \triangleright ((r \rightarrow z \triangleright \text{pu}_M^{\downarrow H} \circ r) \circ (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r))^{\uparrow M}$$

function composition : $= m \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ (z \triangleright \text{pu}_M^{\downarrow H} \circ r))^{\uparrow M}$.

(14.54)

Now we can substitute this into Eq. (14.53):

$$\begin{aligned} & q \triangleright (q \triangleright (\text{sw} \circ (x \rightarrow q \triangleright x))^{\downarrow H} \circ \text{sw}(m) \circ \text{sw}) \\ \text{use Eq. (14.54)} : & = m \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ (q \triangleright (\text{sw} \circ (x \rightarrow q \triangleright x))^{\downarrow H} \circ \text{pu}_M^{\downarrow H} \circ r))^{\uparrow M} \\ H\text{'s composition} : & = m \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ (q \triangleright (\text{pu}_M \circ \text{sw} \circ (x \rightarrow q \triangleright x))^{\downarrow H} \circ r))^{\uparrow M} \\ \text{outer-identity} : & = m \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ (q \triangleright (\text{pu}_M^{\uparrow R} \circ (x \rightarrow q \triangleright x))^{\downarrow H} \circ r))^{\uparrow M} \\ \text{composition} : & = m \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ (q \triangleright (x \rightarrow q \triangleright \text{pu}_M^{\uparrow R} \circ (x))^{\downarrow H} \circ r))^{\uparrow M} . \end{aligned}$$

We arrived at the same expression as the left-hand side of the law.

Proof of the outer-interchange law The law is written as

$$\text{ftn}_M \circ \text{sw} = \text{sw}^{\uparrow M} \circ \text{sw} \circ \text{ftn}_M^{\uparrow R} . \quad (14.55)$$

We will apply both sides of the law to arbitrary $m^{\circ M^{\circ R^A}}$ and $q^{\circ H^{\circ A}}$, and transform both sides to the same expression. We begin with the more complicated right-hand side:

$$\begin{aligned} & q \triangleright (m \triangleright \text{sw}^{\uparrow M} \circ \text{sw} \circ \text{ftn}_M^{\uparrow R}) \\ \triangleright \text{notation} : & = q \triangleright ((m \triangleright \text{sw}^{\uparrow M} \triangleright \text{sw}) \triangleright \text{ftn}_M^{\uparrow R}) \\ \text{definition of } \uparrow^R : & = q \triangleright (\text{ftn}_M^{\downarrow H} \circ (m \triangleright \text{sw}^{\uparrow M} \triangleright \text{sw}) \circ \text{ftn}_M) \\ \triangleright \text{notation} : & = (q \triangleright \text{ftn}_M^{\downarrow H} \triangleright (m \triangleright \text{sw}^{\uparrow M} \triangleright \text{sw})) \triangleright \text{ftn}_M \\ \text{use Eq. (14.41)} : & = (m \triangleright \text{sw}^{\uparrow M} \triangleright (r \rightarrow q \triangleright \text{ftn}_M^{\downarrow H} \circ \text{pu}_M^{\downarrow H} \circ r))^{\uparrow M} \triangleright \text{ftn}_M \\ \text{composition for } H \text{ and } M : & = m \triangleright (\text{sw} \circ (r \rightarrow q \triangleright (\text{pu}_M \circ \text{ftn}_M)^{\downarrow H} \circ r))^{\uparrow M} \circ \text{ftn}_M \\ \text{left identity law of } M : & = m \triangleright (\text{sw} \circ (r \rightarrow q \triangleright r))^{\uparrow M} \circ \text{ftn}_M . \end{aligned} \quad (14.56)$$

Let us simplify the sub-expression $\text{sw} \circ (r \rightarrow q \triangleright r)$ separately:

$$\begin{aligned} & \text{sw} \circ (r \rightarrow q \triangleright r) = (x \rightarrow x \triangleright \text{sw}) \circ (r \rightarrow q \triangleright r) \\ \text{function composition} : & = (x \rightarrow q \triangleright (x \triangleright \text{sw})) \\ \text{use Eq. (14.41)} : & = x \rightarrow x \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M} \\ \text{expand argument} : & = (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M} . \end{aligned} \quad (14.57)$$

Substituting this expression into Eq. (14.56), we get

$$\begin{aligned} & m \triangleright (\text{sw} \circ (r \rightarrow q \triangleright r))^{\uparrow M} \circ \text{ftn}_M \\ \text{use Eq. (14.57)} : & = m \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M \uparrow M} \circ \text{ftn}_M \\ \text{naturality of } \text{ftn}_M : & = m \triangleright \text{ftn}_M \circ (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M} . \end{aligned}$$

Now write the left-hand side of the law:

$$\begin{aligned} & q \triangleright (m \triangleright \text{ftn}_M \circ \text{sw}) = q \triangleright (m \triangleright \text{ftn}_M \triangleright \text{sw}) \\ \text{use Eq. (14.41)} : & = m \triangleright \text{ftn}_M \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M} . \end{aligned}$$

This is equal to the right-hand side we just obtained.

Statement 14.5.1.6 The monadic naturality laws in Statement 14.3.8.1 hold for the `swap` function defined by Eq. (14.40) and the base monad $L \triangleq R$.

Proof The monadic naturality laws are

$$\text{sw}_{R,\text{Id}} = \text{id} \quad , \quad \text{sw}_{R,M} \circ \phi^{\uparrow R} = \phi \circ \text{sw}_{R,N} \quad , \quad \text{sw}_{R,M} \circ \theta = \theta^{\uparrow M} \quad ,$$

where $\phi : M \rightsquigarrow N$ and $\theta : R \rightsquigarrow \text{Id}$ are arbitrary monad morphisms.

To verify the first law, set $M = \text{Id}$ in Eq. (14.41) and get ***

$$\begin{aligned} q \triangleright (m \triangleright \text{sw}_{R,\text{Id}}) \\ \text{use Eq. (14.41)} : &= m \triangleright (r \rightarrow q \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M} \\ \text{use } M = \text{Id} \text{ and } \text{pu}_M = \text{id} : &= m \triangleright (r \rightarrow q \triangleright r) \\ \text{apply to argument } m : &= q \triangleright m = q \triangleright (m \triangleright \text{id}) \quad . \end{aligned}$$

So, $\text{sw}_{R,\text{Id}} = \text{id}$ when applied to arbitrary argument values m and q .

To verify the second law, apply both sides to arbitrary m and q . The left-hand side:

$$\begin{aligned} q \triangleright (m \triangleright \text{sw}_{R,M} \circ \phi^{\uparrow R}) &= q \triangleright (m \triangleright \text{sw}_{R,M} \triangleright \phi^{\uparrow R}) \\ \text{definition of } \uparrow R : &= q \triangleright (\phi^{\downarrow H} \circ (m \triangleright \text{sw}_{R,M}) \circ \phi) \\ \triangleright \text{ notation} : &= (q \triangleright \phi^{\downarrow H}) \triangleright (m \triangleright \text{sw}_{R,M}) \triangleright \phi \\ \text{use Eq. (14.41)} : &= m \triangleright (r \rightarrow q \triangleright \phi^{\downarrow H} \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M} \triangleright \phi \\ \text{composition law for } H : &= m \triangleright (r \rightarrow q \triangleright (\text{pu}_M \circ \phi)^{\downarrow H} \circ r)^{\uparrow M} \circ \phi \\ \text{identity law for } \phi : &= m \triangleright (r \rightarrow q \triangleright \text{pu}_N^{\downarrow H} \circ r)^{\uparrow M} \circ \phi \\ \text{naturality of } \phi : &= m \triangleright \phi \circ (r \rightarrow q \triangleright \text{pu}_N^{\downarrow H} \circ r)^{\uparrow N} \quad . \end{aligned}$$

The right-hand side, when applied to m and q , gives the same expression:

$$\begin{aligned} q \triangleright (m \triangleright \phi \circ \text{sw}_{R,N}) &= q \triangleright (m \triangleright \phi \triangleright \text{sw}_{R,N}) \\ \text{use Eq. (14.41)} : &= m \triangleright \phi \triangleright (r \rightarrow q \triangleright \text{pu}_N^{\downarrow H} \circ r)^{\uparrow N} \quad . \end{aligned}$$

To argue that the third law holds,⁷ apply the left-hand side to m and q :

$$\begin{aligned} q \triangleright (m \triangleright \text{sw}_{R,M} \circ \theta) &= q \triangleright (m \triangleright \text{sw}_{R,M} \triangleright \theta) \\ &= q \triangleright ((q_1 \rightarrow m \triangleright (r \rightarrow q_1 \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M}) \triangleright \theta) \quad . \end{aligned} \tag{14.58}$$

This expression cannot be simplified any further; and neither can the right-hand side $q \triangleright (m \triangleright \theta^{\uparrow M})$. We need more detailed information about the function θ .

The type of θ is

$$\theta : \forall A. (H^A \rightarrow A) \rightarrow A \quad .$$

To implement a function of this type, we need to write code that takes an argument of type $H^A \rightarrow A$ and returns a value of type A . Since the type A is arbitrary, the code of θ cannot store a fixed value of type A to use as the return value. The only possibility to implement a function θ with the required type signature seems to be by substituting a value of type H^A into the given argument of type $H^A \rightarrow A$, which will return the result of type A . So,⁸ we need to produce a value of type H^A for an arbitrary

⁷This is not a fully rigorous proof of the third monadic naturality law. The place where the proof lacks rigor will be shown.

⁸This is where rigor is lacking: we did not prove that the type $\forall A. (H^A \rightarrow A) \rightarrow A$ is really *equivalent* to $\forall A. H^A$. Up to that assumption, the proof is rigorous.

type A , that is, a value of type $\forall A. H^A$. Using the contravariant Yoneda identity, we can simplify this type expression to the type H^1 :

$$\begin{aligned} \forall A. H^A &\cong \forall A. \mathbb{1} \rightarrow H^A \\ \text{use identity } (A \rightarrow \mathbb{1}) \cong \mathbb{1} : &\quad \cong \forall A. (A \rightarrow \mathbb{1}) \rightarrow H^A \\ \text{contravariant Yoneda identity :} &\quad \cong H^1 \quad . \end{aligned}$$

So, we can construct a θ if we store a value h_1 of type H^1 and compute $h : H^A$ as

$$h^{H^A} = h_1^{H^1} \triangleright (a^A \rightarrow 1)^{\downarrow H} \quad .$$

Given a fixed value $h_1 : H^1$, the code of θ is therefore

$$(r^{H^A \rightarrow A}) \triangleright \theta \triangleq h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \triangleright r \quad . \quad (14.59)$$

Let us check whether this θ is a monad morphism $R \rightsquigarrow \text{Id}$. We need to verify the two laws of monad morphisms,

$$\text{pu}_R \circ \theta = \text{id} \quad , \quad \text{ftn}_R \circ \theta = \theta^{\uparrow R} \circ \theta = \theta \circ \theta \quad .$$

The identity law, applied to an arbitrary $x : A$, is

$$\begin{aligned} x \triangleright \text{pu}_R \circ \theta &= (x \triangleright \text{pu}_R) \triangleright \theta \\ \text{definition of } r \triangleright \theta : &= h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \triangleright (x \triangleright \text{pu}_R) \\ \text{definition of } x \triangleright \text{pu}_R : &= (h_1 \triangleright (_ \rightarrow 1)^{\downarrow H}) \triangleright (_ \rightarrow x) \\ \text{function composition :} &= x \quad . \end{aligned}$$

This verifies the identity law.

The composition law, applied to an arbitrary $r : R^{R^A}$, expands to

$$\begin{aligned} r \triangleright \text{ftn}_R \circ \theta &= r \triangleright \text{ftn}_R \triangleright \theta \\ \text{definition of } \text{ftn}_R : &= r \triangleright (t \rightarrow q \rightarrow q \triangleright (q \triangleright (x \rightarrow q \triangleright x)^{\downarrow H} \circ r)) \triangleright \theta \\ \text{apply to } r : &= (q \rightarrow q \triangleright (q \triangleright (x \rightarrow q \triangleright x)^{\downarrow H} \circ r)) \triangleright \theta \\ \text{definition (14.59) of } \theta : &= h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \triangleright (q \rightarrow q \triangleright (q \triangleright (x \rightarrow q \triangleright x)^{\downarrow H} \circ r)) \\ \text{apply to } h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} : &= h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \triangleright (h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \circ (x \rightarrow \dots)^{\downarrow H} \circ r) \\ \text{composition under } H : &= h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \triangleright (h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \triangleright r) \\ \text{definition (14.59) of } \theta : &= r \triangleright \theta \triangleright \theta = r \triangleright \theta \circ \theta \quad . \end{aligned}$$

This verifies the composition law; so θ is indeed a monad morphism.

Using the code of θ defined in Eq. (14.59), we can now verify the monadic naturality law of $\text{sw}_{R,M}$ (with respect to the runners θ of that form). The left-hand side of the law is given by Eq. (14.58) and is rewritten as

$$\begin{aligned} q \triangleright ((q_1 \rightarrow m \triangleright (r \rightarrow q_1 \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M}) \triangleright \theta) \\ \text{definition of } \theta : &= q \triangleright (h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \triangleright (q_1 \rightarrow m \triangleright (r \rightarrow q_1 \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M})) \\ \text{apply to argument :} &= q \triangleright (m \triangleright (r \rightarrow h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \triangleright \text{pu}_M^{\downarrow H} \circ r)^{\uparrow M}) \\ H\text{'s composition law :} &= q \triangleright (m \triangleright (r \rightarrow h_1 \triangleright (\text{pu}_M \circ (_ \rightarrow 1))^{\downarrow H} \circ r)^{\uparrow M}) \\ \text{compose functions :} &= q \triangleright (m \triangleright (r \rightarrow h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \triangleright r)^{\uparrow M}) \quad . \end{aligned}$$

The right-hand side is

$$\begin{aligned}
 & q \triangleright (m \triangleright \theta^{\uparrow M}) \\
 \text{function expansion : } &= q \triangleright (m \triangleright (r \rightarrow \underline{r \triangleright \theta})^{\uparrow M}) \\
 \text{definition (14.59) of } \theta : &= q \triangleright (m \triangleright (r \rightarrow h_1 \triangleright (_ \rightarrow 1)^{\downarrow H} \triangleright r))^{\uparrow M} .
 \end{aligned}$$

This expression is now the same as the left-hand side.

14.5.2 Rigid monad construction 2: composition

Functor composition is the second construction that produces rigid monads. This is a consequence of the properties of monad transformer stacks.

Statement 14.5.2.1 The composition $R_1^{R_2^*}$ of two rigid monads R_1 and R_2 is also a rigid monad.

Proof Since R_1 is rigid, its outside-composition $R_1 \circ M$ with any other monad M is a monad. So $R_1 \circ R_2$ is a monad. To show that $R_1 \circ R_2$ is a rigid monad, we need to show that its monad transformer is of the composed-outside kind. By Statement 14.2.8, the stacking of monad transformers T_{R_1} and T_{R_2} is a lawful monad transformer. Since the transformers for R_1 and R_2 are of the composed-outside kind, $T_{R_1}^M = R_1 \circ M$ and $T_{R_2}^M = R_2 \circ M$, the stack of transformers is

$$T_{R_1}^{T_{R_2}^M} = R_1 \circ T_{R_2}^M = R_1 \circ (R_2 \circ M) = R_1 \circ R_2 \circ M .$$

Therefore $T^M \triangleq R_1 \circ R_2 \circ M$ is a monad transformer applied to the foreign monad M . This shows, by definition of a rigid monad, that $R_1 \circ R_2$ is a rigid monad.

Example 14.5.2.2 Consider the functor composition of the `se1` monad $R_1^A \triangleq (A \rightarrow Q) \rightarrow A$ and the `Reader` monad $R_2^A \triangleq Z \rightarrow A$:

$$P^A \triangleq ((Z \rightarrow A) \rightarrow Q) \rightarrow Z \rightarrow A .$$

It follows from Statement 14.5.2.1 that the functor P^* is a rigid monad; so P 's transformer is of the composed-outside kind. The the transformed monad for any foreign monad M is

$$T^A \triangleq ((Z \rightarrow M^A) \rightarrow Q) \rightarrow Z \rightarrow M^A .$$

To define the monad methods for T , we need to have the definitions of the transformers $T_{R_1}^M$ and $T_{R_2}^M$. Since both the `se1` and the `Reader` monads are special cases of the `Choice` monad construction (Section 14.5.1) where the contrafunctor H is chosen to be $H^A \triangleq A \rightarrow Q$ and $H^A \triangleq Z$ respectively, we can use Eq. (14.36) to define the `flatMap` methods for the transformers $T_{R_1}^M$ and $T_{R_2}^M$:

```

type R1[A] = (A => Q) => A
def map_R1[A, B](r1: R1[A])(f: A => B): R1[B] = { (b2q: B => Q) => f(r1(f andThen b2q)) }
def flatMap_R1[A, B, M[_]: Monad](r1: R1[M[A]])(f: A => R1[M[B]]): R1[M[B]] = {
  (q: M[B]) => map_R1(r1){ (m: M[A]) => m.flatMap(x => f(x)(q)) }(q)
}

type R2[A] = Z => A
def map_R2[A, B](r2: R2[A])(f: A => B): R2[B] = { r2 andThen f }
def flatMap_R2[A, B, M[_]: Monad](r2: R2[M[A]])(f: A => R2[M[B]]): R2[M[B]] = {
  z => map_R2(r2){ (m: M[A]) => m.flatMap(x => f(x)(z)) }(z)
}

```

Now we can define the `flatMap` method for T by using the monad $T_{R_2}^M$ instead of M in the `flatMap` method for $T_{R_1}^M$:

```

type T[A] = R1[R2[A]]
def flatMap_T[A, B, M[_]: Monad](t: T[M[A]])(f: A => T[M[B]]): T[M[B]] = {
  (q: R2[M[B]]) => map_R1(t){ (m: R2[M[A]]) => flatMap_R2(m)(x => f(x)(q)) }(q)
}

```

Does the composed monad have `swap`? The definitions of the monad methods for the composed monads are somewhat complicated. In Section 14.5.1, we have proved the monad transformer laws for $T_R^M \triangleq R \circ M$ by defining a `swap` function with type signature

$$\text{sw}_{R,M} : M \circ R \rightsquigarrow R \circ M ,$$

and proving its laws. Suppose S is also a rigid monad; then the composed monad $R \circ S$ is a rigid monad. Does $T \triangleq R \circ S$ have a suitable `swap` function,

$$\text{sw}_{T,M} : M \circ R \circ S \rightsquigarrow R \circ S \circ M ,$$

satisfying all the required laws? If so, we may be able to find a simpler definition of `flatten` for the monad stack $R \circ S \circ M$. Let us briefly investigate this question. However, keep in mind that the absence of a suitable `swap` function will not invalidate the composition properties of rigid monad stacks: those properties were established without assuming the existence of `swap` for the composed monad.

It turns out that we need yet another law for `swap` (the “3-swap” law) if we wish to prove that the composed monad also has a lawful `swap`.

Statement 14.5.2.3 Assume that two monads R and S both have `swap` methods, $\text{sw}_{R,M}$ and $\text{sw}_{S,M}$, satisfying the 8 laws listed in Statements 14.3.2.1 and 14.3.8.1. Additionally, assume that the **3-swap law** holds with respect to an arbitrary monad M ,

$$\text{sw}_{R,S}^{\uparrow M} ; \text{sw}_{R,M} ; \text{sw}_{S,M}^{\uparrow R} = \text{sw}_{S,M} ; \text{sw}_{R,M}^{\uparrow S} ; \text{sw}_{R,S} . \quad (14.60)$$

$$\begin{array}{ccccc} & \text{sw}_{R,S}^{\uparrow M} & & & \\ M \circ S \circ R & \xrightarrow{\quad \text{sw}_{R,S}^{\uparrow M} \quad} & M \circ R \circ S & \xrightarrow{\quad \text{sw}_{R,M} \quad} & R \circ M \circ S \\ \text{sw}_{S,M} \downarrow & & & & \downarrow \text{sw}_{S,M}^{\uparrow R} \\ S \circ M \circ R & \xrightarrow{\quad \text{sw}_{R,M}^{\uparrow S} \quad} & S \circ R \circ M & \xrightarrow{\quad \text{sw}_{R,S} \quad} & R \circ S \circ M \end{array}$$

Then the composed monad $T \triangleq R \circ S$ also has a `swap` method defined by

$$\text{sw}_{T,M} = \text{sw}_{R,M} ; \text{sw}_{S,M}^{\uparrow R} , \quad (14.61)$$

$$\begin{array}{ccc} M \circ R \circ S & \xrightarrow{\quad \text{sw}_{R,M} \quad} & R \circ M \circ S \\ & \searrow \text{sw}_{T,M} \triangleq & \downarrow \text{sw}_{S,M}^{\uparrow R} \\ & & R \circ S \circ M \end{array}$$

which satisfies the same 8 laws.

Proof We need to verify the 8 laws for $\text{sw}_{T,M}$ (one naturality law, two identity laws, two interchange laws, and three monadic naturality laws), assuming that these 8 laws hold for $\text{sw}_{R,M}$ and $\text{sw}_{S,M}$. In addition, we assume that Eq. (14.60) holds, where M is an arbitrary monad. The monad methods of T are defined by Eqs. (14.19)–(14.20) after renaming $L \triangleq R$ and $M \triangleq S$:

$$\text{pu}_T = \text{pu}_S ; \text{pu}_R , \quad (14.62)$$

$$\text{ftn}_T = \text{sw}_{R,S}^{\uparrow R} ; \text{ftn}_R ; \text{ftn}_S^{\uparrow R} . \quad (14.63)$$

As usual, we do not need to verify the naturality law for $\text{sw}_{T,M}$.

Identity laws To verify the outer-identity law:

$$\begin{aligned}
 \text{expect to equal } \text{pu}_T &: \text{pu}_T^{\uparrow M} ; \text{sw}_{T,M} \\
 \text{use Eqs. (14.61)-(14.62)} &: = \text{pu}_S^{\uparrow M} ; \underline{\text{pu}_R^{\uparrow M} ; \text{sw}_{R,M}} ; \text{sw}_{S,M}^{\uparrow R} \\
 \text{outer-identity law for } \text{sw}_{R,M} &: = \text{pu}_S^{\uparrow M} ; \underline{\text{pu}_R^{\uparrow M} ; \text{sw}_{S,M}^{\uparrow R}} \\
 \text{naturality of } \text{pu}_R &: = \underline{\text{pu}_S^{\uparrow M} ; \text{sw}_{S,M} ; \text{pu}_R} \\
 \text{outer-identity law for } \text{sw}_{S,M} &: = \text{pu}_S ; \text{pu}_R \\
 \text{use Eq. (14.62)} &: = \text{pu}_T \quad .
 \end{aligned}$$

To verify the inner-identity law:

$$\begin{aligned}
 \text{expect to equal } \text{pu}_M^{\uparrow T} &: \text{pu}_M ; \underline{\text{sw}_{T,M}} \\
 \text{use Eq. (14.61)} &: = \underline{\text{pu}_M ; \text{sw}_{R,M}} ; \text{sw}_{S,M}^{\uparrow R} \\
 \text{inner-identity law for } \text{sw}_{R,M} &: = \text{pu}_M^{\uparrow R} ; \text{sw}_{S,M}^{\uparrow R} = (\underline{\text{pu}_M ; \text{sw}_{S,M}})^{\uparrow R} \\
 \text{inner-identity law for } \text{sw}_{S,M} &: = \underline{\text{pu}_M^{\uparrow S \uparrow R}} = \text{pu}_M^{\uparrow T} \quad .
 \end{aligned}$$

Interchange laws The outer-interchange law is

$$\text{ftn}_T^{\uparrow M} ; \text{sw}_{T,M} = \text{sw}_{T,M} ; \text{sw}_{T,M}^{\uparrow T} ; \text{ftn}_T \quad . \quad (14.64)$$

We will use the outer-interchange laws for $\text{sw}_{R,M}$ and $\text{sw}_{S,M}$:

$$\text{ftn}_R^{\uparrow M} ; \text{sw}_{R,M} = \text{sw}_{R,M} ; \text{sw}_{R,M}^{\uparrow R} ; \text{ftn}_R \quad , \quad (14.65)$$

$$\text{ftn}_S^{\uparrow M} ; \text{sw}_{S,M} = \text{sw}_{S,M} ; \text{sw}_{S,M}^{\uparrow S} ; \text{ftn}_S \quad . \quad (14.66)$$

Begin with the right-hand side of Eq. (14.64) since it is more complicated:

$$\begin{aligned}
 & \text{sw}_{T,M} ; \text{sw}_{T,M}^{\uparrow T} ; \text{ftn}_T \\
 \text{definitions} &: = \text{sw}_{R,M} ; \underline{\text{sw}_{S,M}^{\uparrow R} ; (\text{sw}_{R,M} ; \text{sw}_{S,M}^{\uparrow R})^{\uparrow S \uparrow R}} ; \text{sw}_{R,S}^{\uparrow R} ; \text{ftn}_R ; \text{ftn}_S^{\uparrow R} \\
 \uparrow R\text{-composition} &: = \text{sw}_{R,M} ; (\text{sw}_{S,M} ; \text{sw}_{R,M}^{\uparrow S} ; \underline{\text{sw}_{S,M}^{\uparrow R \uparrow S} ; \text{sw}_{R,S}})^{\uparrow R} ; \text{ftn}_R ; \text{ftn}_S^{\uparrow R} \\
 \text{sw's naturality} &: = \text{sw}_{R,M} ; (\underline{\text{sw}_{S,M} ; \text{sw}_{R,M}^{\uparrow S} ; \text{sw}_{R,S} ; \text{sw}_{S,M}^{\uparrow S \uparrow R}})^{\uparrow R} ; \text{ftn}_R ; \text{ftn}_S^{\uparrow R} \\
 \text{Eq. (14.60)} &: = \text{sw}_{R,M} ; (\underline{\text{sw}_{R,S}^{\uparrow M} ; \text{sw}_{R,M} ; \text{sw}_{S,M}^{\uparrow R} ; \text{sw}_{S,M}^{\uparrow S \uparrow R}})^{\uparrow R} ; \text{ftn}_R ; \text{ftn}_S^{\uparrow R} \quad .
 \end{aligned}$$

In the last expression, all `flattens` are to the right of all `swaps`. So we look to rewrite the left-hand side of Eq. (14.64) into the same form:

$$\begin{aligned}
 & \text{ftn}_T^{\uparrow M} ; \text{sw}_{T,M} \\
 (14.61), (14.63) &: = (\text{sw}_{R,S}^{\uparrow R} ; \text{ftn}_R ; \underline{\text{ftn}_S^{\uparrow R}})^{\uparrow M} ; \text{sw}_{R,M} ; \text{sw}_{S,M}^{\uparrow R} \\
 \text{naturality} &: = \text{sw}_{R,S}^{\uparrow R \uparrow M} ; \underline{\text{ftn}_R^{\uparrow M} ; \text{sw}_{R,M}} ; \text{ftn}_S^{\uparrow M \uparrow R} ; \text{sw}_{S,M}^{\uparrow R} \\
 \text{Eq. (14.65)} &: = \text{sw}_{R,S}^{\uparrow R \uparrow M} ; \text{sw}_{R,M} ; \text{sw}_{R,M}^{\uparrow R} ; \text{ftn}_R ; (\underline{\text{ftn}_S^{\uparrow M} ; \text{sw}_{S,M}})^{\uparrow R} \\
 \text{Eq. (14.66)} &: = \text{sw}_{R,M} ; \text{sw}_{R,S}^{\uparrow M \uparrow R} ; \text{sw}_{R,M}^{\uparrow R} ; \text{ftn}_R ; (\text{sw}_{S,M} ; \text{sw}_{S,M}^{\uparrow S} ; \text{ftn}_S)^{\uparrow R} \\
 \text{naturality} &: = \text{sw}_{R,M} ; \text{sw}_{R,S}^{\uparrow M \uparrow R} ; \text{sw}_{R,M}^{\uparrow R} ; (\text{sw}_{S,M} ; \text{sw}_{S,M}^{\uparrow S})^{\uparrow R \uparrow R} ; \text{ftn}_R ; \text{ftn}_S^{\uparrow R} \\
 \text{composition} &: = \text{sw}_{R,M} ; (\underline{\text{sw}_{R,S}^{\uparrow M} ; \text{sw}_{R,M} ; \text{sw}_{S,M}^{\uparrow R} ; \text{sw}_{S,M}^{\uparrow S \uparrow R}})^{\uparrow R} ; \text{ftn}_R ; \text{ftn}_S^{\uparrow R} \quad .
 \end{aligned}$$

Both sides of the outer-interchange law (14.64) are now equal.

The proof of the inner-interchange law is simpler: the law says

$$\text{ftn}_M \circ \text{sw}_{T,M} = \text{sw}_{T,M}^{\uparrow M} \circ \text{sw}_{T,M} \circ \text{ftn}_M^{\uparrow T} . \quad (14.67)$$

We will use the inner-interchange laws for $\text{sw}_{R,M}$ and $\text{sw}_{S,M}$:

$$\text{ftn}_M \circ \text{sw}_{R,M} = \text{sw}_{R,M}^{\uparrow M} \circ \text{sw}_{R,M} \circ \text{ftn}_M^{\uparrow R} , \quad (14.68)$$

$$\text{ftn}_M \circ \text{sw}_{S,M} = \text{sw}_{S,M}^{\uparrow M} \circ \text{sw}_{S,M} \circ \text{ftn}_M^{\uparrow S} . \quad (14.69)$$

Begin with the left-hand side of Eq. (14.67):

$$\begin{aligned} \text{ftn}_M \circ \text{sw}_{T,M} &= \text{ftn}_M \circ \text{sw}_{R,M} \circ \text{sw}_{S,M}^{\uparrow R} \\ \text{use Eq. (14.68)} : &= \text{sw}_{R,M}^{\uparrow M} \circ \text{sw}_{R,M} \circ \text{ftn}_M^{\uparrow R} \circ \text{sw}_{S,M}^{\uparrow R} \\ \text{use Eq. (14.69) under } \uparrow^R : &= \text{sw}_{R,M}^{\uparrow M} \circ \text{sw}_{R,M} \circ (\text{sw}_{S,M}^{\uparrow M} \circ \text{sw}_{S,M} \circ \text{ftn}_M^{\uparrow S})^{\uparrow R} \\ \text{composition under } \uparrow^R : &= \text{sw}_{R,M}^{\uparrow M} \circ \text{sw}_{R,M} \circ \text{sw}_{S,M}^{\uparrow M \uparrow R} \circ \text{sw}_{S,M}^{\uparrow R} \circ \text{ftn}_M^{\uparrow S \uparrow R} \end{aligned}$$

The right-hand side of Eq. (14.67) is

$$\begin{aligned} \text{sw}_{T,M}^{\uparrow M} \circ \text{sw}_{T,M} \circ \text{ftn}_M^{\uparrow T} \\ \text{use Eq. (14.61)} : &= (\text{sw}_{R,M} \circ \text{sw}_{S,M}^{\uparrow R})^{\uparrow M} \circ \text{sw}_{R,M} \circ \text{sw}_{S,M}^{\uparrow R} \circ \text{ftn}_M^{\uparrow S \uparrow R} \\ \text{naturality of } \text{sw}_{R,M} : &= \text{sw}_{R,M}^{\uparrow M} \circ \text{sw}_{R,M} \circ \text{sw}_{S,M}^{\uparrow M \uparrow R} \circ \text{sw}_{S,M}^{\uparrow R} \circ \text{ftn}_M^{\uparrow S \uparrow R} . \end{aligned}$$

The right-hand side is now equal to the left-hand side.

Monadic naturality laws We need to verify that the three laws,

$$\text{sw}_{T,\text{Id}} = \text{id} , \quad \text{sw}_{T,M} \circ \phi^{\uparrow T} = \phi \circ \text{sw}_{T,N} , \quad \text{sw}_{T,M} \circ \theta = \theta^{\uparrow M} ,$$

hold for any monad morphisms $\phi : M \rightsquigarrow N$ and $\theta : T \rightsquigarrow \text{Id}$. We may assume that these laws already hold for $\text{sw}_{R,M}$ and $\text{sw}_{S,M}$.

To verify the first law, write

$$\begin{aligned} \text{expect to equal } \text{id} : & \text{sw}_{T,\text{Id}} \\ \text{use Eq. (14.61)} : &= \text{sw}_{R,\text{Id}} \circ \text{sw}_{S,\text{Id}}^{\uparrow R} \\ \text{first law for } \text{sw}_{R,\text{Id}} \text{ and } \text{sw}_{S,\text{Id}} : &= \text{id} \circ \text{id}^{\uparrow R} = \text{id} . \end{aligned}$$

To verify the second law, write

$$\begin{aligned} \text{expect to equal } \phi \circ \text{sw}_{T,N} : & \text{sw}_{T,M} \circ \phi^{\uparrow T} \\ \text{use Eq. (14.61)} : &= \text{sw}_{R,M} \circ \text{sw}_{S,M}^{\uparrow R} \circ \phi^{\uparrow S \uparrow R} \\ \text{second law for } \text{sw}_{S,M} \text{ under } \uparrow^R : &= \text{sw}_{R,M} \circ (\phi \circ \text{sw}_{S,N})^{\uparrow R} \\ \text{second law for } \text{sw}_{R,M} : &= \phi \circ \text{sw}_{R,N} \circ \text{sw}_{S,N}^{\uparrow R} = \phi \circ \text{sw}_{T,N} . \end{aligned}$$

To verify the third law, we begin with the left-hand side,

$$\text{sw}_{T,M} \circ \theta = \text{sw}_{R,M} \circ \text{sw}_{S,M}^{\uparrow R} \circ \theta .$$

At this point, no relationship or law applies to the intermediate expression $\text{sw}_{S,M}^{\uparrow R} ; \theta$, so we need additional information to proceed. This information is given by Lemma 14.5.2.4, which expresses $\theta = \theta_R ; \theta_S$ where θ_R and θ_S are monad morphisms. So, we may use θ_R and θ_S with the monad morphism laws for R and S , in particular with the third law:

$$\text{sw}_{R,M} ; \theta_R = \theta_R^{\uparrow M} , \quad \text{sw}_{S,M} ; \theta_S = \theta_S^{\uparrow M} .$$

Now we can finish the proof of the third monadic naturality law:

$$\begin{aligned} \text{expect to equal } \theta^{\uparrow M} : & \text{ sw}_{T,M} ; \theta \\ \text{Lemma 14.5.2.4} : & = \text{sw}_{R,M} ; \text{sw}_{S,M}^{\uparrow R} ; \theta_R ; \theta_S \\ \text{naturality of } \theta_R : & = \text{sw}_{R,M} ; \theta_R ; \text{sw}_{S,M} ; \theta_S \\ \text{third law for } R \text{ and } S : & = \theta_R^{\uparrow M} ; \theta_S^{\uparrow M} = \theta^{\uparrow M} . \end{aligned}$$

Lemma If R is a rigid monad with a lawful `swap` method $\text{sw}_{R,M}$ and S is any (not necessarily rigid) monad then any monad morphism $\theta : R \circ S \rightsquigarrow \text{Id}$ can be expressed as a composition $\theta = \theta_R ; \theta_S$ where

$$\theta_R \triangleq \text{pu}_S^{\uparrow R} ; \theta , \quad \theta_S \triangleq \text{pu}_R ; \theta$$

are monad morphisms. In other words, all “runners” θ for the composed monad $R \circ S$ can be written as a function composition of some “runners” $\theta_R : R \rightsquigarrow \text{Id}$ and $\theta_S : S \rightsquigarrow \text{Id}$ for the monads R and S .

Proof How can we find θ_R and θ_S from the given “runner” θ ? Consider that θ will evaluate the operations of both monads R and S , while θ_R can evaluate only the operations of R . To obtain $\theta_R : R \rightsquigarrow \text{Id}$ from $\theta : R \circ S \rightsquigarrow \text{Id}$, we need to prepend a function $R \rightsquigarrow R \circ S$. A suitable function of this type is

$$\text{pu}_S^{\uparrow R} : R^A \rightarrow R^{S^A} .$$

So, $\theta_R = \text{pu}_S^{\uparrow R} ; \theta$ has the correct type signature, $R^A \rightarrow A$. Similarly, $\theta_S = \text{pu}_R ; \theta$ has the correct type signature, $S^A \rightarrow A$. So, we can define θ_R and θ_S from the given “runner” θ as

$$\theta_R \triangleq \text{pu}_S^{\uparrow R} ; \theta , \quad \theta_S \triangleq \text{pu}_R ; \theta .$$

Since pu_R and $\text{pu}_S^{\uparrow R}$ are the lifting and the base lifting of the monad transformer $T_R^S = R \circ S$, Statement 14.3.8.1 shows that pu_R and $\text{pu}_S^{\uparrow R}$ are monad morphisms. Since the composition of monad morphisms is again a monad morphism (Statement 10.3.4.5), we see that θ_R and θ_S as defined above are monad morphisms.

It remains to verify that $\theta = \theta_R ; \theta_S$:

$$\begin{aligned} \text{expect to equal } \theta : & \theta_R ; \theta_S = \text{pu}_S^{\uparrow R} ; \theta ; \text{pu}_R ; \theta \\ \text{naturality of } \theta : & = \text{pu}_S^{\uparrow R} ; \text{pu}_R^{\uparrow T} ; \theta ; \theta = \text{pu}_S^{\uparrow R} ; \text{pu}_R^{\uparrow S \uparrow R} ; \theta ; \theta \\ \text{composition law of } \theta : & = \text{pu}_S^{\uparrow R} ; \text{pu}_R^{\uparrow S \uparrow R} ; \text{ftn}_T ; \theta . \end{aligned}$$

The last line differs from the required result, θ , by the function $\text{pu}_S^{\uparrow R} ; \text{pu}_R^{\uparrow S \uparrow R} ; \text{ftn}_T$. We will finish the proof if we show that this function is identity:

$$\begin{aligned} \text{expect to equal id} : & = \text{pu}_S^{\uparrow R} ; \text{pu}_R^{\uparrow S \uparrow R} ; \text{ftn}_T . \\ \text{use Eq. (14.20)} : & = \text{pu}_S^{\uparrow R} ; \text{pu}_R^{\uparrow S \uparrow R} ; \text{sw}_{R,S}^{\uparrow R} ; \text{ftn}_R ; \text{ftn}_S^{\uparrow R} \\ \text{outer-identity law for } \text{sw}_{R,S} : & = \text{pu}_S^{\uparrow R} ; \text{pu}_R^{\uparrow R} ; \text{ftn}_R ; \text{ftn}_S^{\uparrow R} \\ \text{identity law for } R : & = \text{pu}_S^{\uparrow R} ; \text{ftn}_S^{\uparrow R} \\ \text{identity law for } S : & = \text{id} . \end{aligned}$$

14.5.3 Rigid monad construction 3: product

Statement 14.5.3.1 The product of rigid monads, $R_1^A \times R_2^A$, is a rigid monad.

Proof If we show that $R_1^A \times R_2^A$ has a composed-outside transformer, it will follow (by definition) that $R_1^A \times R_2^A$ is a rigid monad. It follows from Statement 14.7.1.1 (whose proof does not depend on any of the results of this section) that the lawful monad transformer $T^{M,A}$ for the product monad $R_1^A \times R_2^A$ is the product of transformers

$$T_{R_1}^{M,A} \times T_{R_2}^{M,A} = R_1^{M^A} \times R_2^{M^A} \quad ,$$

where M is the foreign monad. This is the required composed-outside transformer for the monad $R_1^A \times R_2^A$.

14.5.4 Rigid monad construction 4: selector

The **selector monad** $S^A \triangleq F^{A \rightarrow R^Q} \rightarrow R^A$ is rigid if R^\bullet is a rigid monad, F^\bullet is any functor, and Q is any fixed type.

14.5.5 Rigid functors

The properties of rigid monads can be extended to a (possibly) larger class of rigid functors. We begin with a definition of a rigid functor that, unlike the definition of rigid monads (Section 14.5), does not refer to any monad transformers.

Definition R is a **rigid functor** if there exists a natural transformation `fuseIn` (denoted by fi_R) with the type signature

$$\text{fi}_R : \forall(A, B). (A \rightarrow R^B) \rightarrow R^{A \rightarrow B}$$

satisfying the non-degeneracy law (14.71) shown below.

Not all functors admit a natural transformation with the type signature of `fuseIn`. For example, the functor $F^A \triangleq Z + A$ is not rigid because the required type signature

$$\text{fi}^{A,B} : (A \rightarrow Z + B) \rightarrow Z + (A \rightarrow B)$$

cannot be implemented. However, any functor R admits the opposite natural transformation `fuseOut` (denoted by fo_R):

$$\text{fo}^{A,B} : R^{A \rightarrow B} \rightarrow A \rightarrow R^B \quad , \quad \text{fo}(r) \triangleq a \rightarrow r \triangleright (f^{A \rightarrow B} \rightarrow f(a))^{\uparrow R} \quad . \quad (14.70)$$

The method `fuseIn` must satisfy the nondegeneracy law

$$\text{fi}_R \circ \text{fo}_R = \text{id}^{(A \rightarrow R^B) \rightarrow (A \rightarrow R^B)} \quad . \quad (14.71)$$

The opposite relation does not hold in general, $\text{fo}_R \circ \text{fi}_R \neq \text{id}$ (see Example 14.5.5.2).

Note that the type signature of fi_R is the same as the type signature of `swap` with respect to the Reader monad,

$$\text{sw}_{R,M} : M^{R^A} \rightarrow R^{M^A} \cong (Z \rightarrow R^A) \rightarrow R^{Z \rightarrow A} \text{ if we set } M^A \triangleq Z \rightarrow A \quad .$$

So we are prompted to ask whether any rigid monad having a `swap` method might also admit `fuseIn`. It turns out that all rigid monads (regardless of the existence of `swap`) are also rigid functors. While proving that statement, we will not assume that `swap` exists for the monad R , but will use directly the definition of rigid monads via the composed-outside monad transformer.

Statement 14.5.5.1 A rigid monad R is also a rigid functor.

Proof By assumption, the monad R has the composed-outside monad transformer $T_R^M = R \circ M$ and the corresponding monad method ftn_T . We can define the transformation fi_R in the same way as Eq. (14.21) defined the `swap` method via ftn_T . To use that formula, we need to set the foreign monad M to be the `Reader` monad, $M^B \triangleq A \rightarrow B$, with the fixed environment type A :

$$\text{fi}_R(f:A \rightarrow R^B) = \text{pu}_M^{\uparrow R \uparrow M} \circ \text{pu}_R \circ \text{ftn}_T \quad (14.72)$$

$$\begin{array}{ccccc} A \rightarrow R^B & \xrightarrow{\text{pu}_M^{\uparrow R \uparrow M}} & A \rightarrow R^{A \rightarrow B} & \xrightarrow{\text{pu}_R} & R^{A \rightarrow R^{A \rightarrow B}} \\ & \searrow & & & \downarrow \text{ftn}_T \\ & & \text{fi}_R \triangleq & & R^{A \rightarrow B} \end{array}$$

Since M is the `Reader` monad with the environment type A , we have

$$\text{pu}_M(x) = (\underline{}^A \rightarrow x) \quad , \quad (f^{X \rightarrow Y})^{\uparrow M}(r^{A \rightarrow X}) = r \circ f \quad .$$

The type signature of the function `fuseOut`,

$$\text{fo} : R^{A \rightarrow B} \rightarrow A \rightarrow R^B \quad ,$$

resembles “running” the composed monad R^{M^B} into R^B , given a value of the environment (of type A). Indeed, given a fixed value a^A , we can “run” the `Reader` monad into the identity monad. The corresponding “runner” $\phi_a^{M \rightsquigarrow \text{Id}}$ is

$$\phi_a^{M^X \rightarrow X} = (m^{A \rightarrow X} \rightarrow m(a)) \quad . \quad (14.73)$$

So we are inspired to use the runner law for the monad transformer T_R^M . That law (which holds since we assumed that T_R^M satisfies all laws) says that the lifted “runner” ϕ_a^R is a monad morphism $T_R^M \rightsquigarrow T_R^{\text{Id}} \cong T_R^M \rightsquigarrow R$. The monad morphism law for ϕ_a^R is then written as

$$\text{ftn}_T \circ \phi_a^R = \phi_a^{\uparrow R \uparrow M \uparrow R} \circ \phi_a^R \circ \text{ftn}_R \quad . \quad (14.74)$$

How could we use this law to obtain Eq. (14.71)? Compare Eqs. (14.70) and (14.73) and derive the connection between the “runner” ϕ_a and the `fuseOut` function (which always exists for any functor R),

$$\text{fo}_R = (r \rightarrow a \rightarrow r \triangleright \phi_a^R) \quad ,$$

and then rewrite the law (14.71) as

$$\begin{aligned} \text{expect to equal } m : & m^{M^R^B} \triangleright \text{fi}_R \circ \text{fo}_R = (m \triangleright \text{fi}_R) \triangleright \text{fo}_R \\ \text{use Eq. (14.70)} : & = a \rightarrow m \triangleright \underline{\text{fi}_R} \triangleright \phi_a^R \\ \text{use Eq. (14.72)} : & = a \rightarrow m \triangleright \text{pu}_M^{\uparrow R \uparrow M} \circ \text{pu}_R \circ \underline{\text{ftn}_T \circ \phi_a^R} \\ \text{use Eq. (14.74)} : & = a \rightarrow m \triangleright \text{pu}_M^{\uparrow R \uparrow M} \circ \text{pu}_R \circ \underline{\phi_a^{\uparrow R \uparrow M \uparrow R} \circ \phi_a^R \circ \text{ftn}_R} \\ \text{naturality of } \text{pu}_R : & = a \rightarrow m \triangleright \text{pu}_M^{\uparrow R \uparrow M} \circ \phi_a^{\uparrow R \uparrow M} \circ \phi_a \circ \underline{\text{pu}_R \circ \text{ftn}_R} \\ \text{left identity law of } R : & = a \rightarrow m \triangleright (\text{pu}_M \circ \phi_a)^{\uparrow R \uparrow M} \circ \phi_a \\ \text{identity law for } \phi_a : & = a \rightarrow m \triangleright \underline{\phi_a} \\ \text{definition (14.73) for } \phi_a : & = a \rightarrow m(a) = m \quad . \end{aligned}$$

Here we used the monad morphism identity law for ϕ_a ,

$$\text{pu}_M \circ \phi_a = (x \rightarrow (\underline{} \rightarrow x)) \circ (m \rightarrow a \triangleright m) = (x \rightarrow a \triangleright (\underline{} \rightarrow x)) = (x \rightarrow x) = \text{id} \quad .$$

Example 14.5.5.2 Show that the functions fuseIn and fuseOut are not always inverses:

$$\text{fo}_R \circ \text{fi}_R \neq \text{id} \quad .$$

Solution Consider the rigid monads $P^A \triangleq Z \rightarrow A$ and $R^A \triangleq (A \rightarrow Q) \rightarrow A$, where Q and Z are fixed types. Since all rigid monads are rigid functors, it follows that the monads P and R have methods fi_P , fo_P , fi_R , fo_R satisfying the non-degeneracy law (14.71). It turns out that additionally $\text{fo}_P \circ \text{fi}_P = \text{id}$, i.e. the methods fo_P and fi_P are inverses of each other, but $\text{fo}_R \circ \text{fi}_R \neq \text{id}$.

To show that $\text{fo}_P \circ \text{fi}_P = \text{id}$, consider the type signatures of fo_P and fi_P :

$$\begin{aligned} \text{fo}_P : P^{A \rightarrow B} &\rightarrow A \rightarrow P^B \cong (Z \rightarrow A \rightarrow B) \rightarrow (A \rightarrow Z \rightarrow B) \quad , \\ \text{fi}_P : (A \rightarrow P^B) &\rightarrow P^{A \rightarrow B} \cong (A \rightarrow Z \rightarrow B) \rightarrow (Z \rightarrow A \rightarrow B) \quad . \end{aligned}$$

The implementations of these functions are derived uniquely from type signatures, as long as we require that these implementations are natural transformations. The functions fo_P and fi_P switch the curried arguments of types A and Z of a function that returns values of type B . It is clear that these two functions are inverses of each other. To show this directly, consider the type signature of $\text{fo}_P \circ \text{fi}_P$,

$$(\text{fo}_P \circ \text{fi}_P) : P^{A \rightarrow B} \rightarrow P^{A \rightarrow B} \cong (Z \rightarrow A \rightarrow B) \rightarrow (Z \rightarrow A \rightarrow B) \quad .$$

There is only one implementation for this type signature as a natural transformation, namely the identity function $\text{id}^{Z \rightarrow A \rightarrow B}$.

For the monad R , the type signatures of fi_R and fo_R are

$$\begin{aligned} \text{fi}_R : (A \rightarrow (B \rightarrow Q) \rightarrow B) &\rightarrow ((A \rightarrow B) \rightarrow Q) \rightarrow A \rightarrow B \quad , \\ \text{fo}_R : (((A \rightarrow B) \rightarrow Q) \rightarrow A \rightarrow B) &\rightarrow A \rightarrow (B \rightarrow Q) \rightarrow B \quad , \end{aligned}$$

and the implementations are again derived uniquely from type signatures,

$$\begin{aligned} f : A \rightarrow (B \rightarrow Q) \rightarrow B \triangleright \text{fi}_R &= x : (A \rightarrow B) \rightarrow Q \rightarrow a : A \rightarrow f(a)(b : B \rightarrow x(_ \rightarrow b)) \quad , \\ g : ((A \rightarrow B) \rightarrow Q) \rightarrow A \rightarrow B \triangleright \text{fo}_R &= a : A \rightarrow y : B \rightarrow Q \rightarrow g(h : A \rightarrow B \rightarrow y(h(a)))(a) \quad . \end{aligned}$$

We notice that the implementation of fi_R uses a constant function, $(_ \rightarrow b)$, which is likely to lose information. Indeed, while $\text{fi}_R \circ \text{fo}_R = \text{id}$ as it must be due to the rigid non-degeneracy law, we find that

$$\begin{aligned} \text{expect not to equal } g : \quad g \triangleright \text{fo}_R \circ \text{fi}_R &= (g \triangleright \text{fo}_R) \triangleright \text{fi}_R \\ \text{definition of } \text{fi}_R : \quad &= x \rightarrow a \rightarrow \text{fo}_R(g)(a)(b \rightarrow x(_ \rightarrow b)) \\ \text{definition of } \text{fo}_R : \quad &= x \rightarrow a \rightarrow g(h : A \rightarrow B \rightarrow h(a) \triangleright (b \rightarrow x(_ \rightarrow b)))(a) \\ \text{apply to argument } b : \quad &= x \rightarrow a \rightarrow g(h \rightarrow x(_ \rightarrow h(a)))(a) \quad . \end{aligned}$$

We cannot simplify the last line any further: the functions g , h , and x are unknown, and we cannot calculate symbolically, say, the value of $x(_ \rightarrow h(a))$. If the last line were equal to g , we would expect it to be $x \rightarrow a \rightarrow g(x)(a)$. The difference is in the first argument of g , namely we have $h \rightarrow x(_ \rightarrow h(a))$ instead of x . The two last expressions are not always equal; they would be equal if we had

$$(h \rightarrow x(h)) = (h \rightarrow x(k \rightarrow h(k)))$$

instead of $h \rightarrow x(_ \rightarrow h(a))$. Consider again the argument of x in the two last expressions: $k \rightarrow h(a)$ instead of $k \rightarrow h(k)$. Since h is not always a constant function (h is an arbitrary function of type $A \rightarrow B$), the two expressions $k \rightarrow h(a)$ and $k \rightarrow h(k)$ are generally not equal. So, we must conclude that $\text{fo}_R \circ \text{fi}_R \neq \text{id}$.

Exercise 14.5.5.3 Show that the functor $F^A \triangleq A \times A \times A$ is rigid.

Exercise 14.5.5.4 Show that the functor $F^A \triangleq \mathbb{1} + A \times A$ is not rigid.

Exercise 14.5.5.5 Show that the functor $F^A \triangleq (A \rightarrow Z) \rightarrow Z$ is not rigid. (Here Z is a fixed type.)

Since all rigid monads are rigid functors, we can reuse all the rigid monad constructions to obtain new rigid functors. The following statement demonstrates a construction of rigid functors that does not assume any monadic properties. It shows that the set of all rigid functors is larger than the set of all rigid monads.

Statement 14.5.5.6 The functor $S^\bullet \triangleq H^\bullet \rightarrow P^\bullet$ is rigid when H is any contrafunctor and P is any rigid functor. (Note that P does not need to be a monad.)

Proof We assume that fi_P and fo_P are known and satisfy the non-degeneracy law (14.71). The function fi_S is then defined by

$$\begin{aligned}\text{fi}_S : (A \rightarrow H^B \rightarrow P^B) &\rightarrow H^{A \rightarrow B} \rightarrow P^{A \rightarrow B} , \\ \text{fi}_S = f^{A \rightarrow H^B \rightarrow P^B} &\rightarrow h^{H^{A \rightarrow B}} \rightarrow \text{fi}_P(a \rightarrow f(a)((b \rightarrow _ \rightarrow b)^{\downarrow H} h)) ,\end{aligned}$$

or, using the forwarding notation,

$$h \triangleright \text{fi}_S(f) = (a \rightarrow h \triangleright (b \rightarrow _ \rightarrow b)^{\downarrow H} \triangleright f(a)) \triangleright \text{fi}_P .$$

Let us write the definition of fo_S as well,

$$\begin{aligned}\text{fo}_S : (H^{A \rightarrow B} \rightarrow P^{A \rightarrow B}) &\rightarrow A \rightarrow H^B \rightarrow P^B , \\ \text{fo}_S = g^{H^{A \rightarrow B} \rightarrow P^{A \rightarrow B}} &\rightarrow a^A \rightarrow h^{H^B} \rightarrow \text{fo}_P\left(g\left(\left(p^{A \rightarrow B} \rightarrow p(a)\right)^{\downarrow H} h\right)\right)(a) ,\end{aligned}$$

or, using the forwarding notation,

$$\text{fo}_S(g) = a \triangleright (h \triangleright (p \rightarrow p(a))^{\downarrow H} ; g ; \text{fo}_P) .$$

To verify the non-degeneracy law for S , apply both sides to some arguments; we expect $f \triangleright (\text{fi}_S ; \text{fo}_S)$ to equal f for an arbitrary $f : A \rightarrow H^B \rightarrow P^B$. To compare values, we need to apply both sides further to some arguments $a : A$ and $h : H^B$. So we expect the following expression to equal $f(a)(h)$:

$$\begin{aligned}(f \triangleright \text{fi}_S ; \text{fo}_S)(a)(h) &= (f \triangleright \text{fi}_S \triangleright \text{fo}_S)(a)(h) \\ \text{expand fo}_S : &= a \triangleright (h \triangleright (p \rightarrow a \triangleright p)^{\downarrow H} \triangleright \text{fi}_S(f) \triangleright \text{fo}_P) \\ \text{expand fi}_S : &= a \triangleright ((a \rightarrow h \triangleright (p \rightarrow p(a))^{\downarrow H} ; (b \rightarrow _ \rightarrow b)^{\downarrow H} ; f(a)) \triangleright \text{fi}_P \triangleright \text{fo}_P) \\ \text{compose }^{\downarrow H} : &= a \triangleright ((a \rightarrow h \triangleright ((b \rightarrow _ \rightarrow b) ; (p \rightarrow p(a)))^{\downarrow H} ; f(a)) \triangleright \text{fi}_P \triangleright \text{fo}_P) .\end{aligned}$$

Computing the function composition

$$(b \rightarrow _ \rightarrow b) ; (p \rightarrow p(a)) = (b \rightarrow (_ \rightarrow b)(a)) = (b \rightarrow b) = \text{id} ,$$

and using the non-degeneracy law $\text{fi}_P ; \text{fo}_P = \text{id}$, we can simplify further:

$$\begin{aligned}a \triangleright ((a \rightarrow h \triangleright ((b \rightarrow _ \rightarrow b) ; (p \rightarrow p(a)))^{\downarrow H} ; f(a)) \triangleright \text{fi}_P \triangleright \text{fo}_P) \\ \text{identity law for } H : &= a \triangleright ((a \rightarrow h \triangleright f(a)) \triangleright \text{fi}_P \triangleright \text{fo}_P) \\ \text{non-degeneracy :} &= a \triangleright (a \rightarrow h \triangleright f(a)) = h \triangleright f(a) .\end{aligned}$$

This equals $f(a)(h)$, as required.

Statement 14.5.5.7 A rigid functor R is pointed; the method `pure` can be defined as

$$\text{pu}_R(x^A) \triangleq \text{id}^{R^A \rightarrow R^A} \triangleright \text{fi}_R \triangleright (_ \rightarrow x)^{\uparrow R} .$$

In particular, there is a selected value r_1 of type $R^\mathbb{1}$ (“wrapped unit”), computed as

$$r_1 \triangleq \text{pu}_R(1) = \text{id} \triangleright \text{fi}_R \triangleright (_ \rightarrow 1)^{\uparrow R} . \quad (14.75)$$

Proof The method $\text{fi}_R : (X \rightarrow R^Y) \rightarrow R^{X \rightarrow Y}$ with type parameters $X = R^A$ and $Y = A$ is applied to the identity function $\text{id} : R^A = R^A$, considered as a value of type $X \rightarrow R^Y$. The result is a value

$$\text{fi}_R(\text{id}) : R^{R^A \rightarrow A} .$$

The result is transformed via the raised constant function $(_ \rightarrow x)^{\uparrow R}$, which takes $R^{R^A \rightarrow A}$ and returns R^A . The resulting code can be written as

$$\text{pu}_R(x) \triangleq (_ \rightarrow x)^{\uparrow R}(\text{fi}_R(\text{id})) = \text{id} \triangleright \text{fi}_R \triangleright (_ \rightarrow x)^{\uparrow R} .$$

The function pu_R defined in this way is a natural transformation since fi_R is one. Applying pu_R to a unit value, we obtain the selected value r_1 of type R^1 .

The next statement shows that r_1 is the *only* distinct value of the type R^1 . This means, in particular, that a rigid functor cannot be a disjunctive type defined with several constructors, such as $\mathbb{1} + A + A$ or List^A . A rigid functor's type definition must have a single constructor. This is one motivation for the name “rigid”: the container R^A has a fixed shape and bears no extra information other than holding some values of type A .

Statement 14.5.5.8 If a functor R has a method fi_R satisfying the non-degeneracy law (14.71) then the type R^1 is equivalent to the unit type: $R^1 \cong \mathbb{1}$. The value r_1 defined by Eq. (14.75) is the only available distinct value of type R^1 . The isomorphism between $\mathbb{1}$ and R^1 is the function $(1 \rightarrow r_1)$.

Proof The idea of the proof is to apply both sides of the non-degeneracy law $\text{fi}_R \circ \text{fo}_R = \text{id}$ to the identity function of type $R^1 \rightarrow R^1$. To adapt the type parameters, consider the type signature of $\text{fi}_R \circ \text{fo}_R$,

$$(\text{fi}_R \circ \text{fo}_R) : (A \rightarrow R^B) \rightarrow (A \rightarrow R^B) ,$$

and set $A = R^1$ and $B = \mathbb{1}$. The left-hand side of the law can be now applied to the identity function $\text{id} : R^1 \rightarrow R^1$, which yields a value of type $R^1 \rightarrow R^1$, i.e. a function

$$f_1 : R^1 \rightarrow R^1 , \quad f_1 \triangleq \text{fo}_R(\text{fi}_R(\text{id})) .$$

We will show that f_1 is a constant function, $f_1 = (_ \rightarrow r_1)$, always returning the same value r_1 defined in Statement 14.5.7. However, the right-hand side of the non-degeneracy law applied to id is the identity function of type $R^1 \rightarrow R^1$. So, the non-degeneracy law means that $f_1 = \text{id}$. If the identity function of type $R^1 \rightarrow R^1$ always returns the same value (r_1), it means that r_1 is the only distinct value of the type R^1 .

To begin the proof, note that for any fixed type A , the function type $A \rightarrow \mathbb{1}$ is equivalent to $\mathbb{1}$. This is so because there exists only one pure function of type $A \rightarrow \mathbb{1}$, namely $(_ \rightarrow 1)$. In other words, there is only one distinct value of the type $A \rightarrow \mathbb{1}$, and the value is the function $(_ \rightarrow 1)$. The code of this function is uniquely determined by its type signature.

The isomorphism between the types $A \rightarrow \mathbb{1}$ and $\mathbb{1}$ is realized by the functions $u : \mathbb{1} \rightarrow A \rightarrow \mathbb{1}$ and $v : (A \rightarrow \mathbb{1}) \rightarrow \mathbb{1}$. The code of these functions is also uniquely determined by their type signatures:

$$u = (1 \rightarrow _ : A \rightarrow 1) , \quad v = (_ : A \rightarrow \mathbb{1} \rightarrow 1) .$$

Applying fi_R to the identity function of type $R^1 \rightarrow R^1$, we obtain a value g ,

$$g : R^{R^1 \rightarrow \mathbb{1}} , \quad g \triangleq \text{id} \triangleright \text{fi}_R .$$

Since the type $R^1 \rightarrow \mathbb{1}$ is equivalent to $\mathbb{1}$, the type $R^{R^1 \rightarrow \mathbb{1}}$ is equivalent to R^1 . To use this equivalence explicitly, we need to raise the isomorphisms u and v into the functor R . The isomorphism will then map $g : R^{R^1 \rightarrow \mathbb{1}}$ to some $g_1 : R^1$ by

$$g_1 \triangleq v^{\uparrow R}(g) .$$

Substituting the definitions of g , v , and r_1 , we find that actually $g_1 = r_1$:

$$g_1 = \text{id} \triangleright \text{fi}_R \triangleright v^{\uparrow R} = \text{id} \triangleright \text{fi}_R \triangleright (_ \rightarrow 1)^{\uparrow R} = r_1 \quad .$$

We can now map g_1 back to g via the raised isomorphism u :

$$\begin{aligned} g &= g_1 \triangleright u^{\uparrow R} = r_1 \triangleright u^{\uparrow R} \\ \text{definition of } u : &= r_1 \triangleright (1 \rightarrow _ \rightarrow 1)^{\uparrow R} \quad . \end{aligned} \quad (14.76)$$

Compute $\text{fo}_R(g)$ as

$$\begin{aligned} \text{fo}_R(g) &= g \triangleright \text{fo}_R \\ \text{use Eq. (14.70)} : &= a^{\cdot R^{\dagger}} \rightarrow g \triangleright (f^{\cdot A \rightarrow \mathbb{1}} \rightarrow f(a))^{\uparrow R} \\ \text{use Eq. (14.76)} : &= a \rightarrow r_1 \triangleright (1 \rightarrow _^{\cdot A} \rightarrow 1)^{\uparrow R} \triangleright (f^{\cdot A \rightarrow \mathbb{1}} \rightarrow f(a))^{\uparrow R} \\ \text{composition under } \uparrow^R : &= a \rightarrow r_1 \triangleright (\underline{1 \rightarrow 1})^{\uparrow R} \\ (1 \rightarrow 1) \text{ is identity} : &= (a \rightarrow r_1 \triangleright \text{id}) = (a^{\cdot R^{\dagger}} \rightarrow r_1) = (_^{\cdot R^{\dagger}} \rightarrow r_1) \quad . \end{aligned}$$

So, $\text{fo}_R(g) = \text{fo}_R(\text{fi}_R(\text{id}))$ is a function of type $R^{\dagger} \rightarrow R^{\dagger}$ that ignores its argument and always returns the same value r_1 .

By virtue of the non-degeneracy law, $\text{fo}_R(g) = \text{id}$. We see that the identity function $\text{id} : R^{\dagger} \rightarrow R^{\dagger}$ always returns the same value r_1 . Applying this function to an arbitrary value $x : R^{\dagger}$, we get

$$x = x \triangleright \text{id} = x \triangleright \text{fo}_R(g) = x \triangleright (_ \rightarrow r_1) = r_1 \quad .$$

It means that all values of type R^{\dagger} are equal to r_1 . So the function $1 \rightarrow r_1$ is indeed an isomorphism between the types $\mathbb{1}$ and R^{\dagger} .

It follows from Statement 14.5.5.8 that a rigid functor cannot be a disjunctive type with more than one constructor. So, functors such as `Option` or `List` are not rigid. A rigid functor must have a fixed “shape” of data wrapped in it. This partially explains the choice of the name “rigid”.

Statement 14.5.5.9 The product of rigid functors is a rigid functor.

Proof We assume that two rigid functors P and Q are given, with their methods fi_P and fi_Q satisfying the non-degeneracy law. The functor product $R \triangleq P \times Q$ needs its own fi_R method, with the type signature

$$\text{fi}_R : (A \rightarrow P^B \times Q^B) \rightarrow P^{A \rightarrow B} \times Q^{A \rightarrow B} \quad .$$

To implement that method, first note that the function type $A \rightarrow P \times Q$ is isomorphic to $(A \rightarrow P) \times (A \rightarrow Q)$. The isomorphism is given by a “split” function,

$$\text{split} : (A \rightarrow B \times C) \rightarrow (A \rightarrow B) \times (A \rightarrow C) \quad , \quad \text{split} \triangleq f^{\cdot A \rightarrow B \times C} \rightarrow (f \circ \pi_1) \times (f \circ \pi_2) \quad .$$

Then the method fi_R is implemented concisely:

$$\text{fi}_R \triangleq \text{split} \circ (\text{fi}_P \boxtimes \text{fi}_Q) \quad .$$

To check the non-degeneracy law ($\text{fi}_R \circ \text{fo}_R = \text{id}$), we need the code of fo_R . It is convenient to define fo_R via a “merge” function:

$$\text{merge} : (A \rightarrow B) \times (A \rightarrow C) \rightarrow A \rightarrow B \times C \quad , \quad \text{merge} \triangleq f^{\cdot A \rightarrow B} \times g^{\cdot A \rightarrow C} \rightarrow a^{\cdot A} \rightarrow f(a) \times g(a) \quad .$$

The code of fo_R is then written as:

$$\text{fo}_R : P^{A \rightarrow B} \times Q^{A \rightarrow C} \rightarrow A \rightarrow P^B \times Q^C \quad , \quad \text{fo}_R \triangleq (\text{fo}_P \boxtimes \text{fo}_Q) \circ \text{merge} \quad .$$

The functions `split` and `merge` are mutual inverses. Now we can verify the non-degeneracy law:

$$\begin{aligned} \text{fi}_R \circ \text{fo}_R &= \text{split} \circ (\text{fi}_P \boxtimes \text{fi}_Q) \circ (\text{fo}_P \boxtimes \text{fo}_Q) \circ \text{merge} \\ &= \text{split} \circ ((\text{fi}_P \circ \text{fo}_P) \boxtimes (\text{fi}_Q \circ \text{fo}_Q)) \circ \text{merge} = \text{split} \circ (\text{id} \boxtimes \text{id}) \circ \text{merge} = \text{split} \circ \text{merge} = \text{id} \quad . \end{aligned}$$

Statement 14.5.5.10 The composition of rigid functors is a rigid functor.

Proof We will show that the non-degeneracy law ($\text{fi}_R \circ \text{fo}_R = \text{id}$) holds for the functor $R^\bullet \triangleq P^{Q^\bullet}$ as long as P, Q are rigid. Begin by defining the methods fi_R and fo_R , assuming that the corresponding methods of P, Q are known and satisfy the non-degeneracy law. We write:

$$\begin{aligned}\text{fi}_R : (A \rightarrow P^{Q^B}) &\rightarrow P^{Q^{A \rightarrow B}} , \quad \text{fi}_R \triangleq \text{fi}_P \circ \text{fi}_Q^P , \\ \text{fo}_R : P^{Q^{A \rightarrow B}} &\rightarrow A \rightarrow P^{Q^B} , \quad \text{fo}_R \triangleq \text{fo}_Q^P \circ \text{fo}_P .\end{aligned}$$

To verify the non-degeneracy law for R :

$$\text{fi}_R \circ \text{fo}_R = \text{fi}_P \circ \text{fi}_Q^P \circ \text{fo}_Q^P \circ \text{fo}_P = \text{fi}_P \circ (\text{fi}_Q \circ \text{fo}_Q)^P \circ \text{fo}_P = \text{fi}_P \circ \text{fo}_P = \text{id} .$$

Some more use cases for rigid functors are shown in the next statements. The “ R -valued `flatMap`” is a generalization of `flatMap` that can handle multiple M -effects at once; the collection of M -effects is described by a functor R .

Statement 14.5.5.11 ***For a rigid functor R and a monad M , an “ R -valued `flatMap`” can be defined,

$$\text{rflm}_{M,R} : (A \rightarrow R^{M^B}) \rightarrow M^A \rightarrow R^{M^B} .$$

A “refactoring” is a program transformation that does not significantly change the functionality.

Statement 14.5.5.12 ***Given a rigid functor R , a refactoring function can be implemented,

$$\text{refactor} : ((A \rightarrow B) \rightarrow C) \rightarrow (A \rightarrow R^B) \rightarrow R^C .$$

This function transforms a program $p(f:A \rightarrow B) : C$ into a program $\tilde{p}(\tilde{f}:A \rightarrow R^B) : R^C$.

14.6 Recursive monad transformers

14.6.1 Transformer for the free monad `FreeT`

14.6.2 Transformer for the list monad `ListT`

14.7 Monad transformers for monad constructions

14.7.1 Product of monad transformers

Statement 14.7.1.1 The transformer for a product of two monads is the product of transformers.

Proof Given two monads G and H whose transformers T_G and T_H are known, we need to show that the transformer for the monad $L^\bullet \triangleq G^\bullet \times H^\bullet$ is given by

$$T_L^{M,A} \triangleq T_G^{M,A} \times T_H^{M,A} .$$

We may assume that all monad transformer laws hold for T_G and T_H .

Monad construction law Since T_G^M and T_H^M are (by assumption) lawful monads, T_L^M is a lawful monad by the product construction (Statement 10.2.8.2).

Identity law With $M = \text{Id}$, we have $T_L^{\text{Id}} = T_G^{\text{Id}} \times T_H^{\text{Id}} \cong G \times H$ by the pair product of the monad morphisms $T_G^{\text{Id}} \cong G$ and $T_H^{\text{Id}} \cong H$. The pair product of two monad morphisms is again a monad morphism (Statement 14.9.1.2).

Lifting law We need to define the foreign lift function (`flift`) and show that it is a monad morphism:

$$\text{flift}_L^{\cdot M \rightsquigarrow T_G^M \times T_H^M} \triangleq \Delta \circ (\text{flift}_G \boxtimes \text{flift}_H) \quad .$$

To show that flift_L is a monad morphism, we use the fact that $\Delta : M \rightsquigarrow M \times M$ is one (Statement 14.9.1.1), that the pair product preserves monad morphisms (Statement 14.9.1.2), and that the composition also preserves monad morphisms (Statement 10.3.4.5).

Runner laws We need to define the foreign runner (`frun`) and verify its functor laws.

$$\text{frun}_L(\phi^{\cdot M \rightsquigarrow N}) : T_G^{M,A} \times T_H^{M,A} \rightsquigarrow T_G^{N,A} \times T_H^{N,A} \quad , \quad \text{frun}_L(\phi) \triangleq \text{frun}_G(\phi) \boxtimes \text{frun}_H(\phi) \quad .$$

The result of applying $\text{frun}_L(\phi)$ is a monad morphism by Statement 14.9.1.2.

The functor laws hold due to the properties of the pair product:

$$\text{expect to equal id} : \text{frun}_L(\text{id}) = \text{frun}_G(\text{id}) \boxtimes \text{frun}_H(\text{id}) = \text{id} \boxtimes \text{id} = \text{id} \quad ,$$

$$\text{expect to equal } \text{frun}_L(\phi \circ \chi) : \text{frun}_L(\phi) \circ \text{frun}_L(\chi) = (\text{frun}_G(\phi) \circ \text{frun}_G(\chi)) \boxtimes (\text{frun}_H(\phi) \circ \text{frun}_H(\chi))$$

$$\text{functor laws for } \text{frun}_G, \text{frun}_H : \quad = \text{frun}_G(\phi \circ \chi) \boxtimes \text{frun}_H(\phi \circ \chi) = \text{frun}_L(\phi \circ \chi) \quad .$$

Base runner laws We may define a base runner in two different ways, dropping either the first or the second part of the product. Both base runners are lawful. Consider the first one:

$$\text{brun}_L : T_G^M \times T_H^M \rightsquigarrow M \quad , \quad \text{brun}_L \triangleq \pi_1 \circ \text{brun}_G \quad .$$

It is a monad morphism because it is a composition of a projection (Statement 14.9.1.3) and a monad morphism brun_G . Function composition preserves monad morphisms (Statement 10.3.4.5).

The non-degeneracy law holds because it holds for brun_G :

$$\text{flift}_L \circ \text{brun}_L = \Delta \circ (\text{flift}_G \boxtimes \text{flift}_H) \circ \pi_1 \circ \text{brun}_G = \Delta \circ \pi_1 \circ \text{flift}_G \circ \text{brun}_G = \text{id} \quad .$$

Analogous proofs work for the other definition, $\text{brun}_L \triangleq \pi_2 \circ \text{brun}_H$.

A deficiency of these base runners is that the effects of one of the monads (G or H) are being ignored, which loses information. If we would like to run the effects of both monads G and H , we can instead implement a base runner into the monad $M \times M$. This corresponds to running the monad $G \times H$ into the double-valued monad $D^A \triangleq A \times A$, since $T_D^M = M \times M$. Define the base runner as:

$$\text{brun}_L : T_G^M \times T_H^M \rightsquigarrow M \times M \quad , \quad \text{brun}_L \triangleq \text{brun}_G \boxtimes \text{brun}_H \quad .$$

This is a monad morphism due to Statement 14.9.1.2. The non-degeneracy law has the form

$$\text{flift}_L \circ \text{brun}_L = \Delta^{\cdot M \rightsquigarrow M \times M} \quad ,$$

since the target monad is now $M \times M$. To verify the non-degeneracy law:

$$\begin{aligned} \text{flift}_L \circ \text{brun}_L &= \Delta \circ (\text{flift}_G \boxtimes \text{flift}_H) \circ (\text{brun}_G \boxtimes \text{brun}_H) \\ &= \Delta \circ (\text{flift}_G \circ \text{brun}_G) \boxtimes (\text{flift}_H \circ \text{brun}_H) = \Delta \quad . \end{aligned}$$

Monadic naturality laws To verify the monadic naturality of flift_L :

$$\begin{aligned} \text{flift}_L \circ \text{frun}_L(\phi) &= \Delta \circ (\text{flift}_G \boxtimes \text{flift}_H) \circ (\text{frun}_G(\phi) \boxtimes \text{frun}_H(\phi)) \\ &= \Delta \circ (\text{flift}_G \circ \text{frun}_G(\phi)) \boxtimes (\text{flift}_H \circ \text{frun}_H(\phi)) = \Delta \circ (\phi \circ \text{flift}_G) \boxtimes (\phi \circ \text{flift}_H) \\ &= \Delta \circ (\phi \boxtimes \phi) \circ (\text{flift}_G \boxtimes \text{flift}_H) = \phi \circ \Delta \circ (\text{flift}_G \boxtimes \text{flift}_H) = \phi \circ \text{flift}_L \quad . \end{aligned}$$

To verify the monadic naturality of an information-losing definition of brun_L :

$$\begin{aligned} \text{frun}_L(\phi) \circ \text{brun}_L &= (\text{frun}_G(\phi) \boxtimes \text{frun}_H(\phi)) \circ \pi_1 \circ \text{brun}_G = \pi_1 \circ \text{frun}_G(\phi) \circ \text{brun}_G \quad , \\ \text{brun}_L \circ \phi &= \pi_1 \circ \text{brun}_G \circ \phi = \pi_1 \circ \text{frun}_G(\phi) \circ \text{brun}_G \quad . \end{aligned}$$

The runner $\text{brun}_L : T_L^M \rightsquigarrow M \times M$ also obeys a monadic naturality law involving $\phi \boxtimes \phi$ instead of ϕ :

$$\begin{aligned} \text{frun}_L(\phi) \circ \text{brun}_L &= (\text{frun}_G(\phi) \boxtimes \text{frun}_H(\phi)) \circ (\text{brun}_G \boxtimes \text{brun}_H) \\ &= (\text{frun}_G(\phi) \circ \text{brun}_G) \boxtimes (\text{frun}_H(\phi) \circ \text{brun}_H) = (\text{brun}_G \circ \phi) \boxtimes (\text{brun}_H \circ \phi) \\ &= (\text{brun}_G \boxtimes \text{brun}_H) \circ (\phi \boxtimes \phi) \quad . \end{aligned}$$

14.7.2 Free pointed monad transformer

For an arbitrary monad K , the **free pointed monad on K** is defined as the monad $L^A \triangleq A + K^A$ (see Statement 10.2.8.4).

Statement 14.7.2.1 The monad L 's transformer is defined by the type formula

$$T_L^{M,A} \triangleq M^{A+T_K^{M,A}} \quad ,$$

where T_K is the monad K 's transformer, which is assumed to be known.

Proof To verify the monad transformer laws for T_L , we begin by observing that $A + T_K^{M,A}$ is a free pointed monad on T_K^M . Let us denote for brevity

$$N^A \triangleq T_K^{M,A} \quad , \quad P^A \triangleq A + T_K^{M,A} = A + N^A \quad , \quad T_L^{M,A} \triangleq M^{P^A} = (M \circ P)^A \quad .$$

We know from Statement 10.2.8.4 that

$$\text{pu}_P = a^A \rightarrow a + \mathbb{0}^{N^A} \quad , \quad \text{ftn}_P = \left| \begin{array}{c|cc} & A & N^A \\ \hline A & \text{id} & \mathbb{0} \\ N^A & \mathbb{0} & \text{id} \\ N^{P^A} & \mathbb{0} & \gamma^{\uparrow N} \circ \text{ftn}_N \end{array} \right| \quad , \quad \text{where} \quad \gamma^{P^A \rightarrow N^A} \triangleq \left| \begin{array}{c|c} & N^A \\ \hline A & \text{pu}_N \\ N^A & \text{id} \end{array} \right| \quad .$$

The transformer T_L is a functor composition of the foreign monad M outside P . To show that T_L is a monad, it is convenient to use Statement 14.3.2.1 with a suitable `swap` function (denoted by “sw”):

$$\text{sw} : P^{M^A} \rightarrow M^{P^A} \quad , \quad \text{sw} \triangleq \left| \begin{array}{c|c} & M^{A+N^A} \\ \hline M^A & \text{pu}_P^{\uparrow M} \\ N^{M^A} & (t \rightarrow \mathbb{0}^A + \text{merge}(t)) \circ \text{pu}_M \end{array} \right| \quad ,$$

where the helper function temporarily denoted by `merge` will combine all effects contained within N^{M^A} into an overall N -effect in a value of type N^A :

$$\text{merge} : N^{M^A} \rightarrow N^A \quad , \quad \text{merge} \triangleq \text{flift}_K^{\uparrow N} \circ \text{ftn}_N \quad .$$

We will now verify the monad transformer laws for T_L .

Monad construction law The monad methods of T_L are defined via `sw` as per Statement 14.3.2.1:

$$\text{pu}_T \triangleq \text{pu}_P \circ \text{pu}_M = (a^A \rightarrow a + \mathbb{0}^{N^A}) \circ \text{pu}_M \quad , \quad \text{ftn}_T \triangleq \text{sw}^{\uparrow M} \circ \text{ftn}_M \circ \text{ftn}_P^{\uparrow M} \quad .$$

The functor composition $T_L^M \triangleq M \circ P$ is a monad as long as the `swap` function is a natural transformation and satisfies the four laws given in Statement 14.3.2.1, where that function was denoted by $\text{sw}_{M,P}$. Let us verify those laws.

The `swap` function is a natural transformation because it is a combination of various natural transformations such as pu_M and `merge`.

The outer-identity law is $\text{pu}_M^{\uparrow P} \circ \text{sw} = \text{pu}_M$. We rewrite the left-hand side, expecting to obtain pu_M :

$$\begin{aligned}
 \text{pu}_M^{\uparrow P} \circ \text{sw} &= \left| \begin{array}{c|cc|c|c} & M^A & N^{M^A} & & M^{A+N^A} \\ \hline A & \text{pu}_M & \emptyset & \circ & (a \rightarrow a + \emptyset^{N^A})^{\uparrow M} \\ N^A & \emptyset & \text{pu}_M^{\uparrow N} & & (t \rightarrow \emptyset^A + \text{merge}(t)) \circ \text{pu}_M \end{array} \right| \\
 &= \left| \begin{array}{c|c|c} & M^{A+N^A} & & M^{A+N^A} \\ \hline A & \text{pu}_M \circ (a \rightarrow a + \emptyset^{N^A})^{\uparrow M} & = & (a \rightarrow a + \emptyset^{N^A}) \circ \text{pu}_M \\ N^A & \text{pu}_M^{\uparrow N} \circ (t \rightarrow \emptyset^A + \text{merge}(t)) \circ \text{pu}_M & & (n \rightarrow \emptyset^A + \text{merge}(n \triangleright \text{pu}_M^{\uparrow N})) \circ \text{pu}_M \end{array} \right| \\
 &= \left| \begin{array}{c|cc|c|c} & A + N^A & & A & N^A \\ \hline A & a \rightarrow a + \emptyset^{N^A} & \circ \text{pu}_M & \text{id} & \emptyset \\ N^A & n \rightarrow \emptyset^A + n \triangleright \text{pu}_M^{\uparrow N} \circ \text{merge} & & \emptyset & \text{pu}_M^{\uparrow N} \circ \text{merge} \end{array} \right| \circ \text{pu}_M
 \end{aligned}$$

It remains to show that the matrix element $\text{pu}_M^{\uparrow N} \circ \text{merge}$ is equal to the identity function:

$$\begin{aligned}
 \text{expect to equal id : } \text{pu}_M^{\uparrow N} \circ \text{merge} &= \text{pu}_M^{\uparrow N} \circ \text{flift}_K^{\uparrow N} \circ \text{ftn}_N \\
 \text{monad morphism law of flift}_K : \quad &= \text{pu}_N^{\uparrow N} \circ \text{ftn}_N = \text{id} \quad .
 \end{aligned}$$

The inner-identity law is $\text{pu}_P \circ \text{sw} = \text{pu}_P^{\uparrow M}$. Begin with the left-hand side:

$$\begin{aligned}
 \text{pu}_P \circ \text{sw} &= (m^{M^A} \rightarrow m + \emptyset^{N^{M^A}}) \circ \text{sw} \\
 &= \left| \begin{array}{c|cc|c|c} & M^A & N^{M^A} & & M^{A+N^A} \\ \hline M^A & \text{id} & \emptyset & \circ & \text{pu}_P^{\uparrow M} \\ & & & & (n \rightarrow \emptyset^A + \text{merge}(n)) \circ \text{pu}_M \end{array} \right| = \text{pu}_P^{\uparrow M} \quad .
 \end{aligned}$$

The outer-interchange law is

$$\text{ftn}_M^{\uparrow P} \circ \text{sw} = \text{sw} \circ \text{sw}^{\uparrow M} \circ \text{ftn}_M \quad .$$

Begin with the right-hand side since it is more complicated, and use the identity law we just proved:

$$\begin{aligned}
 \text{sw} \circ \text{sw}^{\uparrow M} \circ \text{ftn}_M &= \left| \begin{array}{c|c|c} & & M^{A+N^A} \\ \hline M^{M^A} & \text{pu}_P^{\uparrow M} \circ \text{sw}^{\uparrow M} \circ \text{ftn}_M & \\ N^{M^{M^A}} & (n \rightarrow \emptyset + \text{merge}(n)) \circ \text{pu}_M \circ \text{sw}^{\uparrow M} \circ \text{ftn}_M & \end{array} \right| \\
 &= \left| \begin{array}{c|c} \text{pu}_P^{\uparrow M} \circ \text{ftn}_M & \text{ftn}_M \circ \text{pu}_P^{\uparrow M} \\ \hline (t \rightarrow \emptyset + \text{merge}(t)) \circ \text{sw} \circ \text{pu}_M \circ \text{ftn}_M & (t \rightarrow \emptyset + \text{merge}(t)) \circ \text{sw} \end{array} \right| \quad .
 \end{aligned}$$

The left-hand side of the law is written as:

$$\begin{aligned}
 \text{ftn}_M^{\uparrow P} \circ \text{sw} &= \left| \begin{array}{c|cc|c|c} & M^A & N^{M^A} & & M^{A+N^A} \\ \hline M^{M^A} & \text{ftn}_M & \emptyset & \circ & \text{pu}_P^{\uparrow M} \\ N^{M^{M^A}} & \emptyset & \text{ftn}_M^{\uparrow N} & & (t \rightarrow \emptyset + \text{merge}(t)) \circ \text{pu}_M \end{array} \right| \\
 &= \left| \begin{array}{c|c} & M^{A+N^A} \\ \hline M^{M^A} & \text{ftn}_M \circ \text{pu}_P^{\uparrow M} \\ N^{M^{M^A}} & \text{ftn}_M^{\uparrow N} \circ (t \rightarrow \emptyset + \text{merge}(t)) \circ \text{pu}_M \end{array} \right| \quad .
 \end{aligned}$$

The first rows of the matrices for the two sides of the law are now equal (to $\text{ftn}_M \circ \text{pu}_P^{\uparrow M}$). It remains to show that the second rows of the matrices are also equal:

$$(t \rightarrow \mathbb{0} + \text{merge}(t)) \circ \text{sw} \stackrel{?}{=} \text{ftn}_M^{\uparrow N} \circ (t \rightarrow \mathbb{0} + \text{merge}(t)) \circ \text{pu}_M ,$$

apply to t : $(\mathbb{0} + t \triangleright \text{merge}) \triangleright \text{sw} \stackrel{?}{=} (\mathbb{0} + t \triangleright \text{ftn}_M^{\uparrow N} \triangleright \text{merge}) \triangleright \text{pu}_M ,$

apply sw : $(\mathbb{0} + t \triangleright \text{merge} \triangleright \text{merge}) \triangleright \text{pu}_M \stackrel{?}{=} (\mathbb{0} + t \triangleright \text{ftn}_M^{\uparrow N} \triangleright \text{merge}) \triangleright \text{pu}_M .$

We will prove the last equality if we show that the function `merge` has the following property:

$$\text{merge} \circ \text{merge} = \text{ftn}_M^{\uparrow N} \circ \text{merge} , \text{ or : } \text{flift}_K^{\uparrow N} \circ \text{ftn}_N \circ \text{flift}_K^{\uparrow N} \circ \text{ftn}_N = \text{ftn}_M^{\uparrow N} \circ \text{flift}_K^{\uparrow N} \circ \text{ftn}_N .$$

$$\begin{array}{ccc} N^{M^M A} & \xrightarrow{\text{merge}^{M^A}} & N^{M^A} \\ \text{ftn}_M^{\uparrow N} \downarrow & & \downarrow \text{merge}^A \\ N^{M^A} & \xrightarrow{\text{merge}^A} & N^A \end{array} \quad \begin{array}{l} \text{We know nothing about the transformer } T_K \text{ 's } \text{flift} \text{ function except for} \\ \text{its laws. The law connecting } \text{flift} \text{ and } \text{flatten} \text{ is the lifting law:} \\ \text{flift}_K \circ \text{flift}_K^{\uparrow N} \circ \text{ftn}_N = \text{ftn}_M \circ \text{flift}_K . \end{array} \quad (14.77)$$

To make use of that law, we use other laws to transform the expression $\text{merge} \circ \text{merge}$, so that the two `flift` functions are composed together and are followed by ftn_N :

$$\begin{array}{ll} \text{expect to equal } \text{ftn}_M^{\uparrow N} \circ \text{merge} : & \text{merge} \circ \text{merge} = \text{flift}_K^{\uparrow N} \circ \text{ftn}_N \circ \text{flift}_K^{\uparrow N} \circ \text{ftn}_N \\ \text{naturality of } \text{ftn}_N : & = \text{flift}_K^{\uparrow N} \circ \text{flift}_K^{\uparrow N \uparrow N} \circ \underline{\text{ftn}_N \circ \text{ftn}_N} \\ \text{associativity law of } \text{ftn}_N : & = \underline{\text{flift}_K^{\uparrow N} \circ \text{flift}_K^{\uparrow N \uparrow N} \circ \text{ftn}_N^{\uparrow N}} \circ \text{ftn}_N = (\text{flift}_K \circ \text{flift}_K^{\uparrow N} \circ \text{ftn}_N)^{\uparrow N} \circ \text{ftn}_N \\ \text{lifting law (14.77)} : & = (\text{ftn}_M \circ \text{flift}_K)^{\uparrow N} \circ \text{ftn}_N = \text{ftn}_M^{\uparrow N} \circ \underline{\text{flift}_K^{\uparrow N} \circ \text{ftn}_N} = \text{ftn}_M^{\uparrow N} \circ \text{merge} . \end{array}$$

This completes the proof of the outer-interchange law of `swap`.

The inner-interchange law is

$$\text{ftn}_P \circ \text{sw} = \text{sw}^{\uparrow P} \circ \text{sw} \circ \text{ftn}_P^{\uparrow M} .$$

The left-hand side of that law is rewritten as:

$$\begin{aligned} \text{ftn}_P \circ \text{sw} &= \left| \begin{array}{c|cc|c|c} & M^A & N^{M^A} & & M^{P^A} \\ \hline M^A & \text{id} & \mathbb{0} & & \text{pu}_P^{\uparrow M} \\ N^{M^A} & \mathbb{0} & \text{id} & & \\ \hline N^{P^{M^A}} & \mathbb{0} & \gamma^{\uparrow N} \circ \text{ftn}_N & & (t \rightarrow \mathbb{0} + \text{merge}(t)) \circ \text{pu}_M \end{array} \right| \\ &= \left| \begin{array}{c|c} M^{P^A} & \\ \hline M^A & \text{pu}_P^{\uparrow M} \\ N^{M^A} & (t \rightarrow \mathbb{0} + \text{merge}(t)) \circ \text{pu}_M \\ N^{P^{M^A}} & \gamma^{\uparrow N} \circ \text{ftn}_N \circ (t \rightarrow \mathbb{0} + \text{merge}(t)) \circ \text{pu}_M \end{array} \right| . \end{aligned}$$

The right-hand side is transformed using the identity law of ftn_P and the naturality law of pu_P :

$$\begin{aligned}
 & \text{sw}^{\uparrow P} ; \text{sw} ; \text{ftn}_P^{\uparrow M} \\
 = & \left| \begin{array}{c|cc|c} & M^{P^A} & N^{M^{P^A}} & \\ \hline M^A & \text{pu}_P^{\uparrow M} & \emptyset & \\ N^{M^A} & (t \rightarrow \emptyset + \text{merge}(t)) ; \text{pu}_M & \emptyset & \\ N^{P^{M^A}} & \emptyset & \text{sw}^{\uparrow N} & \end{array} \right| ; \left| \begin{array}{c|cc} M^{P^A} & \\ \hline M^{P^A} & \text{pu}_P^{\uparrow M} ; \text{ftn}_P^{\uparrow M} \\ N^{M^{P^A}} & (t \rightarrow \emptyset + \text{merge}(t)) ; \text{pu}_M ; \text{ftn}_P^{\uparrow M} \end{array} \right| \\
 = & \left| \begin{array}{c|cc} & M^{P^A} & \\ \hline M^A & \text{pu}_P^{\uparrow M} & \\ N^{M^A} & (t \rightarrow \emptyset + \text{merge}(t)) ; \text{pu}_M & \\ N^{P^{M^A}} & \text{sw}^{\uparrow N} ; (t \rightarrow \emptyset + \text{merge}(t)) ; \text{ftn}_P ; \text{pu}_M & \end{array} \right| .
 \end{aligned}$$

The difference between the two sides is only in the third rows of the matrices:

$$\gamma^{\uparrow N} ; \text{ftn}_N ; (t \rightarrow \emptyset + \text{merge}(t)) ; \text{pu}_M \stackrel{?}{=} \text{sw}^{\uparrow N} ; (t \rightarrow \emptyset + \text{merge}(t)) ; \text{ftn}_P ; \text{pu}_M .$$

We can omit the composition with pu_M if we can prove a stronger condition:

$$\gamma^{\uparrow N} ; \text{ftn}_N ; (t \rightarrow \emptyset + \text{merge}(t)) \stackrel{?}{=} \text{sw}^{\uparrow N} ; (t \rightarrow \emptyset + \text{merge}(t)) ; \text{ftn}_P . \quad (14.78)$$

Apply both sides of Eq. (14.78) to an arbitrary $n : N^{P^{M^A}}$. The left-hand side becomes:

$$n \triangleright \gamma^{\uparrow N} ; \text{ftn}_N ; (t \rightarrow \emptyset + \text{merge}(t)) = \emptyset + n \triangleright \gamma^{\uparrow N} \triangleright \text{ftn}_N \triangleright \text{merge} .$$

The right-hand side of Eq. (14.78) applied to n is

$$\begin{aligned}
 & n \triangleright \text{sw}^{\uparrow N} ; (t \rightarrow \emptyset + \text{merge}(t)) ; \text{ftn}_P = (\emptyset + n \triangleright \text{sw}^{\uparrow N} \triangleright \text{merge}) \triangleright \text{ftn}_P \\
 = & (\emptyset + n \triangleright \text{sw}^{\uparrow N} ; \text{merge}) \triangleright \left| \begin{array}{c|cc} & A & N^A \\ \hline A & \text{id} & \emptyset \\ N^A & \emptyset & \text{id} \\ N^{P^A} & \emptyset & \gamma^{\uparrow N} ; \text{ftn}_N \end{array} \right| = \emptyset + n \triangleright \text{sw}^{\uparrow N} ; \text{merge} ; \gamma^{\uparrow N} ; \text{ftn}_N .
 \end{aligned}$$

The remaining difference is between two functions of type $N^{P^{M^A}} \rightarrow N^A$:

$$\gamma^{\uparrow N} ; \text{ftn}_N ; \text{merge} \stackrel{?}{=} \text{sw}^{\uparrow N} ; \text{merge} ; \gamma^{\uparrow N} ; \text{ftn}_N . \quad (14.79)$$

The left-hand side of Eq. (14.79) can be transformed as

$$\begin{aligned}
 \gamma^{\uparrow N} ; \text{ftn}_N ; \text{merge} &= \gamma^{\uparrow N} ; \text{ftn}_N ; \text{flift}_K^{\uparrow N} ; \text{ftn}_N \\
 \text{naturality of } \text{ftn}_N : &= \gamma^{\uparrow N} ; \text{flift}_K^{\uparrow N \uparrow N} ; \text{ftn}_N ; \text{ftn}_N \\
 \text{associativity law of } \text{ftn}_N : &= \gamma^{\uparrow N} ; \text{flift}_K^{\uparrow N \uparrow N} ; \text{ftn}_N^{\uparrow N} ; \text{ftn}_N = (\gamma ; \text{flift}_K^{\uparrow N} ; \text{ftn}_N)^{\uparrow N} ; \text{ftn}_N .
 \end{aligned}$$

The right-hand side of Eq. (14.79) is

$$\begin{aligned}
 & \text{sw}^{\uparrow N} ; \text{merge} ; \gamma^{\uparrow N} ; \text{ftn}_N \\
 \text{naturality of } \text{merge} : &= \text{sw}^{\uparrow N} ; \gamma^{\uparrow M \uparrow N} ; \text{merge} ; \text{ftn}_N = \text{sw}^{\uparrow N} ; \gamma^{\uparrow M \uparrow N} ; \text{flift}_K^{\uparrow N} ; \text{ftn}_N ; \text{ftn}_N \\
 \text{associativity law of } \text{ftn}_N : &= (\text{sw} ; \gamma^{\uparrow M} ; \text{flift}_K ; \text{ftn}_N)^{\uparrow N} ; \text{ftn}_N .
 \end{aligned}$$

The remaining difference is now between two functions of type $P^{M^A} \rightarrow N^A$:

$$\gamma \circ \text{flift}_K^N \circ \text{ftn}_N \stackrel{?}{=} \text{sw} \circ \gamma^M \circ \text{flift}_K \circ \text{ftn}_N . \quad (14.80)$$

The left-hand side of Eq. (14.80) is

$$\begin{aligned} \gamma \circ \text{flift}_K^N \circ \text{ftn}_N &= \left| \begin{array}{c|c} & N^{M^A} \\ \hline M^A & \text{pu}_N \\ N^{M^A} & \text{id} \end{array} \right| \circ \text{flift}_K^N \circ \text{ftn}_N = \left| \begin{array}{c|c} & N^A \\ \hline M^A & \text{pu}_N \circ \text{flift}_K^N \circ \text{ftn}_N \\ N^{M^A} & \text{id} \circ \text{flift}_K^N \circ \text{ftn}_N \end{array} \right| \\ &= \left| \begin{array}{c|c} & N^A \\ \hline M^A & \text{flift}_K \circ \text{pu}_N \circ \text{ftn}_N \\ N^{M^A} & \text{id} \circ \text{flift}_K^N \circ \text{ftn}_N \end{array} \right| = \left| \begin{array}{c|c} & N^A \\ \hline M^A & \text{flift}_K \\ N^{M^A} & \text{merge} \end{array} \right| . \end{aligned}$$

To proceed with the right-hand side of Eq. (14.80), we compute some intermediate expressions:

$$\begin{aligned} \text{sw} \circ \gamma^M &= \left| \begin{array}{c|c} & M^{N^A} \\ \hline M^A & \text{pu}_P^M \circ \gamma^M \\ N^{M^A} & (t \rightarrow \mathbb{0} + \text{merge}(t)) \circ \text{pu}_M \circ \gamma^M \end{array} \right| , \\ \text{pu}_P \circ \gamma &= (a^A \rightarrow a + \mathbb{0}) \circ \left| \begin{array}{c|c} & N^A \\ \hline A & \text{pu}_N \\ N^A & \text{id} \end{array} \right| = a^A \rightarrow \text{pu}_N(a) = \text{pu}_N , \\ (t \rightarrow \mathbb{0} + \text{merge}(t)) \circ \text{pu}_M \circ \gamma^M &= (t \rightarrow \mathbb{0} + \text{merge}(t)) \circ \gamma \circ \text{pu}_M \\ &= (t \rightarrow \mathbb{0} + \text{merge}(t)) \circ \left| \begin{array}{c|c} & N^A \\ \hline A & \text{pu}_N \\ N^A & \text{id} \end{array} \right| \circ \text{pu}_M = t \rightarrow \text{merge}(t) \triangleright \text{pu}_M = \text{merge} \circ \text{pu}_M . \end{aligned}$$

So we can reduce the right-hand side of Eq. (14.80) to:

$$\begin{aligned} \text{sw} \circ \gamma^M \circ \text{flift}_K \circ \text{ftn}_N &= \left| \begin{array}{c|c} & N^A \\ \hline M^A & \text{pu}_N^M \circ \text{flift}_K \circ \text{ftn}_N \\ N^{M^A} & \text{merge} \circ \text{pu}_M \circ \text{flift}_K \circ \text{ftn}_N \end{array} \right| \\ &= \left| \begin{array}{c|c} & N^A \\ \hline M^A & \text{flift}_K \circ \text{pu}_N^M \circ \text{ftn}_N \\ N^{M^A} & \text{merge} \circ \text{pu}_N \circ \text{ftn}_N \end{array} \right| = \left| \begin{array}{c|c} & N^A \\ \hline M^A & \text{flift}_K \\ N^{M^A} & \text{merge} \end{array} \right| . \end{aligned}$$

This now equals the left-hand side of Eq. (14.80). The proof is concluded.

Identity law The identity law holds: setting $M = \text{Id}$ in the type of T_L^M gives the isomorphisms

$$T_L^{\text{Id}, A} \cong \text{Id}^{A+T_K^{\text{Id}, A}} \cong A + K^A \cong L^A .$$

These isomorphisms are monad morphisms due to Statement 14.3.5.1 under the assumption that the identity law already holds for the transformer T_K . We also need an additional property that the free pointed monad construction preserves monad morphisms (Exercise 14.9.2.10). Then the assumed monad isomorphism $T_K^{\text{Id}} \cong K$ extends to a monad isomorphism $A + T_K^{\text{Id}, A} \cong A + K^A$.

Lifting law The `flift` function for T_L is defined as

$$\text{flift}_L : M^A \rightarrow M^{P^A} , \quad \text{flift}_L \triangleq \text{pu}_P^{\uparrow M} .$$

To show that `flift` is a monad morphism, we refer to Statement 14.3.6.1, where we need to use `blift` instead of `flift` and consider P as the base monad and M as the foreign monad.

Other laws To verify the runner laws and the monadic naturality laws for T_L , we can refer to Statement 14.3.7.1, where we need to replace $\text{sw}_{L,M}$ by $\text{sw}_{M,P}$ and consider P as the base monad and M as the foreign monad. However, the transformer $T_L^M = M \circ P$ is not a purely compositional transformer because the monad P also depends on M ,

$$P^A \triangleq A + T_K^{M,A} .$$

For this reason, the monadic naturality of $\text{sw}_{M,P}$ cannot be formulated as simply the monadic naturality with respect to the monad type M . Mapping the monad M to M' forces us also to map P to P' .

14.8 Irregular and incomplete monad transformers

14.8.1 The state monad transformer (StateT)

The transformer for the `State` monad is not compositional; as shown in Section ???, the `State` monad does not generally compose with other monads. Instead, `StateT` needs a custom definition (Section 14.1.3).

Statement 14.8.1.1 The monad transformer for the `State` monad is

$$T_{\text{State}}^{M,A} \triangleq S \rightarrow M^{A \times S} .$$

It satisfies all monad transformer laws except the monad morphism laws for base runners.

Proof We need to verify all the monad transformer laws (Section 14.2.4) separately.

As we will see below, the `StateT` transformer has one problem: its base runner violates the monad morphism composition law. The standard runner for the `State` monad,

$$\text{run}_{\text{State}}(s_0) : \text{State}^{S,A} \rightarrow A , \quad \text{run}_{\text{State}}(s_0) \triangleq p : S \rightarrow A \times S \rightarrow p(s_0) \triangleright \pi_1 ,$$

also violates that law because it discards the updated state, while the composition law requires to pass the updated state to the next monad value. For the `State` monad, this problem can be solved by starting the runner from an updated state instead of the fixed state s_0 . However, that solution cannot be used with the `StateT` transformer whose updated state is wrapped as $M^{A \times S}$ in an arbitrary foreign monad M . The monad M 's effect could wrap several values or no values. The updated state is not well-defined in that case.

For this reason, a monadic program involving `StateT` cannot be refactored to run in several steps. To get correct results, the runner must be applied once to the entire monadic program. Alternatively, the runner needs to be modified to return a value of type $M^{A \times S}$ rather than a value of type M^A .

Monad construction law We need to show that T_{State}^M is a lawful monad, assuming that M is one. The monad operations for T_{State}^M are defined by Eqs. (14.1)–(14.2). To verify the monad laws, it is faster to use the uncurried Kleisli formulation, as we did for the `State` monad (Statement 10.2.7.2).

The type signature for a T_{State} -Kleisli function is

$$A \rightarrow T_{\text{State}}^{M,B} = A \rightarrow S \rightarrow M^{B \times S} .$$

When we uncurry this type signature, we get the type $A \times S \rightarrow M^{B \times S}$. It is easy to define the Kleisli composition (\diamond_T) for uncurried Kleisli functions because \diamond_T is the same as the M -Kleisli composition (\diamond_M) except applied to types $A \times S$, $B \times S$, and $C \times S$:

$$f : A \times S \rightarrow M^{B \times S} \quad \diamond_T \quad g : B \times S \rightarrow M^{C \times S} \triangleq f \diamond_M g \quad .$$

The proper T_{State} -Kleisli composition (\diamond_T) is then defined by

$$f : A \rightarrow S \rightarrow M^{B \times S} \quad \diamond_T \quad g : B \rightarrow S \rightarrow M^{C \times S} \triangleq \text{uncu}(f) \diamond_M \text{uncu}(g) \quad .$$

The uncurrying transformation

$$\text{uncu} : (A \rightarrow B \rightarrow C) \rightarrow A \times B \rightarrow C \quad , \quad \text{uncu}(f) \triangleq a : A \times b : B \rightarrow f(a)(b) \quad ,$$

is an isomorphism with the inverse transformation

$$\text{cu} : (A \times B \rightarrow C) \rightarrow A \rightarrow B \rightarrow C \quad , \quad \text{cu}(f) \triangleq a : A \rightarrow b : B \rightarrow f(a \times b) \quad .$$

So, it is sufficient to verify the monad laws in the uncurried Kleisli formulation.

To verify the identity laws, we need to define the uncurried `pure` method for T_{State} :

$$\tilde{\text{pu}}_T \triangleq \text{uncu}(\text{pu}_T) = a : A \times s : S \rightarrow \text{pu}_M(a \times s) = \text{pu}_M^{A \times S} \quad .$$

It is now quick to show that all three monad laws for T_{State} hold due to M 's monad laws:

$$\begin{aligned} \tilde{\text{pu}}_T \diamond_T f &= \text{pu}_M \diamond_M f = f \quad , \quad f \diamond_T \tilde{\text{pu}}_T = f \diamond_M \text{pu}_M = f \quad , \\ (f \diamond_T g) \diamond_T h &= (f \diamond_M g) \diamond_M h = f \diamond_M (g \diamond_M h) = f \diamond_T (g \diamond_T h) \quad . \end{aligned}$$

Identity law When $M = \text{Id}$, the type of $T_{\text{State}}^{M,A}$ becomes equal to State^A , and the monad operations become equal to those of the `state` monad.

Lifting law The `flift` function is defined by

$$m : M^A \triangleright \text{flift} \triangleq s : S \rightarrow m \triangleright (a : A \rightarrow a \times s)^{\uparrow M} \quad . \quad (14.81)$$

We need to show that `flift` is a monad morphism $M \rightsquigarrow T_{\text{State}}^M$. The identity law:

$$a : A \triangleright \text{pu}_M \triangleright \text{flift} = s \rightarrow a \triangleright \text{pu}_M \circ (a : A \rightarrow a \times s)^{\uparrow M}$$

$$\text{naturality of } \text{pu}_M : \quad = s \rightarrow a \triangleright (a : A \rightarrow a \times s) \circ \text{pu}_M = s \rightarrow \text{pu}_M(a \times s) = \text{pu}_T \quad .$$

The composition law in terms of the Kleisli composition (\diamond_T) is:

$$(f : A \rightarrow M^B \circ \text{flift}) \diamond_T (g : B \rightarrow M^C \circ \text{flift}) \stackrel{?}{=} (f \diamond_M g) \circ \text{flift} \quad .$$

In terms of \diamond_T , this equation is reformulated as

$$\text{uncu}(f : A \rightarrow M^B \circ \text{flift}) \diamond_T \text{uncu}(g : B \rightarrow M^C \circ \text{flift}) \stackrel{?}{=} \text{uncu}((f \diamond_M g) \circ \text{flift}) \quad .$$

To show that this equation holds, begin by uncurrying the arguments of \diamond_T :

$$\begin{aligned} \text{uncu}(f : A \rightarrow M^B \circ \text{flift}) &= \text{uncu}(a : A \rightarrow f(a) \triangleright \text{flift}) = \text{uncu}(a : A \rightarrow s : S \rightarrow a \triangleright f \triangleright (b : B \rightarrow b \times s)^{\uparrow M}) \\ &= a : A \times s \rightarrow a \triangleright f \circ (b : B \rightarrow b \times s)^{\uparrow M} \quad . \end{aligned}$$

Using this formula, we get:

$$\begin{aligned} \text{uncu}(f : A \rightarrow M^B \circ \text{flift}) \diamond_T \text{uncu}(g : B \rightarrow M^C \circ \text{flift}) &= \text{uncu}(f : A \rightarrow M^B \circ \text{flift}) \diamond_M \text{uncu}(g : B \rightarrow M^C \circ \text{flift}) \\ &= (a : A \times s \rightarrow a \triangleright f \circ (b : B \rightarrow b \times s)^{\uparrow M}) \diamond_M (b : B \times s \rightarrow b \triangleright g \circ (c : C \rightarrow c \times s)^{\uparrow M}) \quad . \end{aligned}$$

We need somehow to move $(b \rightarrow b \times s)^{\uparrow M}$ to the right of \diamond_M . However, the naturality law of \diamond_M cannot be applied because $(b \rightarrow b \times s)^{\uparrow M}$ contains the variable s , which is bound in the scope of the left argument of \diamond_M . To proceed, let us apply the expression to an arbitrary value $a \times s$, which will enable us to move the rest of the expression $(\diamond_M \dots)$ into the scope of the left argument of \diamond_M :

$$\begin{aligned} & (a \times s) \triangleright (a \times s \rightarrow a \triangleright f \circ (b \rightarrow b \times s)^{\uparrow M}) \diamond_M \dots \\ &= (a \times s) \triangleright (a \times s \rightarrow a \triangleright f \circ (b \rightarrow b \times s)^{\uparrow M}) \circ \text{flm}_M(\dots) \\ &= a \triangleright f \circ (b \rightarrow b \times s)^{\uparrow M} \circ \text{flm}_M(\dots) = a \triangleright f \circ (b \rightarrow b \times s)^{\uparrow M} \diamond_M \dots \end{aligned}$$

This trick allows us to complete the derivation:

$$\begin{aligned} & a \triangleright f \circ (b \rightarrow b \times s)^{\uparrow M} \diamond_M (b \times s \rightarrow b \triangleright g \circ (c \rightarrow c \times s)^{\uparrow M}) \\ \text{naturality of } \diamond_M : &= a \triangleright f \diamond_M (b \rightarrow b \times s) \circ (b \times s \rightarrow b \triangleright g \circ (c \rightarrow c \times s)^{\uparrow M}) \\ \text{compute composition :} &= a \triangleright f \diamond_M (b \rightarrow b \triangleright g \circ (c \rightarrow c \times s)^{\uparrow M}) \\ \triangleright\text{-notation :} &= a \triangleright f \diamond_M (g \circ (c \rightarrow c \times s)^{\uparrow M}) \\ \text{naturality of } \diamond_M : &= a \triangleright (f \diamond_M g) \circ (c \rightarrow c \times s)^{\uparrow M} = (a \times s) \triangleright \text{uncu}((f \diamond_M g) \circ \text{flift}) \end{aligned} .$$

Foreign runner laws The foreign runner (frun) is defined by

$$\text{frun}(\phi: M \rightsquigarrow N) \triangleq p: S \rightarrow M^{A \times S} \rightarrow s: S \rightarrow s \triangleright p \triangleright \phi = p: S \rightarrow M^{A \times S} \rightarrow p \circ \phi .$$

First, we need to prove that $\text{frun}(\phi)$ is a monad morphism of type $T_{\text{State}}^M \rightsquigarrow T_{\text{State}}^N$. For convenience, denote $P \triangleq T_{\text{State}}^M$ and $Q \triangleq T_{\text{State}}^N$. To verify the identity law:

$$\begin{aligned} & \text{pu}_P \circ \text{frun}(\phi: M \rightsquigarrow N) = (a: A \rightarrow S \rightarrow (a \times s) \triangleright \text{pu}_M) \circ (p \rightarrow p \circ \phi) \\ \text{compute composition :} &= a \rightarrow s \rightarrow (a \times s) \triangleright \text{pu}_M \circ \phi \\ \text{identity law of } \phi : &= a \rightarrow s \rightarrow (a \times s) \triangleright \text{pu}_N = \text{pu}_Q . \end{aligned}$$

The composition law is more convenient to verify in the uncurried Kleisli formulation. For brevity, denote $r \triangleq \text{frun}(\phi)$; this is a monad morphism of type $P \rightsquigarrow Q$. We need to show that

$$(f: A \rightarrow P^B \circ r) \diamond_Q (g: B \rightarrow P^C \circ r) \stackrel{?}{=} (f \diamond_P g) \circ r .$$

Express this through uncurried functions and the operations $\diamond_P, \tilde{\diamond}_Q$:

$$\text{uncu}(f: A \rightarrow P^B \circ r) \tilde{\diamond}_Q \text{uncu}(g: B \rightarrow P^C \circ r) \stackrel{?}{=} \text{uncu}((f \diamond_P g) \circ r) .$$

Uncurrying a function composition such as $f \circ r$ works like this:

$$\begin{aligned} & \text{uncu}(f: A \rightarrow S \rightarrow M^{B \times S} \circ r: P^B \rightsquigarrow Q^B) = a: A \times S \rightarrow s \triangleright (a \triangleright f \circ r) = a: A \times S \rightarrow s \triangleright (f(a) \triangleright r) \\ \text{definition of } r : &= a \times s \rightarrow s \triangleright (f(a) \triangleright (p \rightarrow p \circ \phi)) = a \times s \rightarrow s \triangleright f(a) \circ \phi \\ &= (a \times s \rightarrow s \triangleright f(a)) \circ \phi = \text{uncu}(f) \circ \phi . \end{aligned} \tag{14.82}$$

The right-hand side of the composition law then becomes:

$$\text{uncu}((f \diamond_P g) \circ r) = \text{uncu}(f \diamond_P g) \circ \phi = (\text{uncu}(f) \diamond_M \text{uncu}(g)) \circ \phi .$$

We can now proceed to simplify the left-hand side of the composition law:

$$\begin{aligned} \text{expect to equal } \text{uncu}((f \diamond_P g) \circ r) : & \frac{\text{uncu}(f: A \rightarrow P^B \circ r) \tilde{\diamond}_Q \text{uncu}(g: B \rightarrow P^C \circ r)}{\text{use Eq. (14.82)} : \quad = (\text{uncu}(f) \circ \phi) \tilde{\diamond}_Q (\text{uncu}(g) \circ \phi)} \\ & \text{definition of } \tilde{\diamond}_Q : \quad = (\text{uncu}(f) \circ \phi) \diamond_N (\text{uncu}(g) \circ \phi) \\ \text{composition law of } \phi : & \quad = (\text{uncu}(f) \diamond_M \text{uncu}(g)) \circ \phi = \text{uncu}((f \diamond_P g) \circ r) . \end{aligned}$$

It remains to verify that $\text{frun}(\phi^{M \rightsquigarrow N})$ satisfies the functor laws with respect to the monad parameters M, N . The functor identity law:

$$\text{frun}(\text{id}^{M \rightsquigarrow M}) = p \rightarrow p \circ \text{id} = p \rightarrow p = \text{id} .$$

The functor composition law:

$$\text{frun}(\phi^{L \rightsquigarrow M}) \circ \text{frun}(\chi^{M \rightsquigarrow N}) = (p \rightarrow p \circ \phi) \circ (q \rightarrow q \circ \chi) = p \rightarrow p \circ \phi \circ \chi = \text{frun}(\phi \circ \chi) .$$

Base runner laws The base runner is parameterized by an initial state value s_0 :

$$\text{brun}(s_0) \triangleq p^{S \rightarrow M^{A \times S}} \rightarrow p(s_0) \triangleright \pi_1^M .$$

We will check the laws for a fixed s_0 . The non-degeneracy law holds:

$$\begin{aligned} \text{expect to equal } m : & m \triangleright \text{flift} \triangleright \text{brun}(s_0) = m \triangleright \text{flift} \triangleright (p \rightarrow p(s_0) \triangleright \pi_1^M) = s_0 \triangleright (m \triangleright \text{flift}) \triangleright \pi_1^M \\ \text{use Eq. (14.81)} : & = m \triangleright (a \rightarrow a \times s_0) \triangleright \pi_1^M = m \triangleright (a \rightarrow a) = m . \end{aligned}$$

We now turn to the monadic morphism laws. The identity law holds:

$$\begin{aligned} \text{expect to equal } \text{pu}_M : & \text{pu}_P \circ \text{brun}(s_0) = (a \rightarrow s \rightarrow (a \times s) \triangleright \text{pu}_M) \circ (p \rightarrow p(s_0) \triangleright \pi_1^M) \\ \text{compute composition} : & = a \rightarrow (a \times s_0) \triangleright \text{pu}_M \triangleright \pi_1^M = a \rightarrow (a \times s_0) \triangleright \underline{\text{pu}_M \circ \pi_1^M} \\ \text{naturality of } \text{pu}_M : & = a \rightarrow (a \times s_0) \triangleright \pi_1 \circ \text{pu}_M = a \rightarrow a \triangleright \text{pu}_M = \text{pu}_M . \end{aligned}$$

The composition law is easier to check in the Kleisli formulation:

$$(f^{A \rightarrow S \rightarrow M^{B \times S}} \circ \text{brun}(s_0)) \diamond_M (g^{B \rightarrow S \rightarrow M^{C \times S}} \circ \text{brun}(s_0)) \stackrel{?}{=} (f \diamond_P g) \circ \text{brun}(s_0) .$$

Since it is easier to define \diamond_P in terms of the uncurried Kleisli functions, let us express the composition $f \circ \text{brun}(s_0)$ in terms of the uncurried version of f :

$$\begin{aligned} f \circ \text{brun}(s_0) &= a \rightarrow a \triangleright f \triangleright (p \rightarrow p(s_0) \triangleright \pi_1^M) = a \rightarrow f(a)(s_0) \triangleright \pi_1^M \\ &= a \rightarrow (a \times s_0) \triangleright \text{uncu}(f) \circ \pi_1^M = (a \rightarrow a \times s_0) \circ \text{uncu}(f) \circ \pi_1^M . \end{aligned} \tag{14.83}$$

We can now write the left-hand side of the composition law as:

$$\begin{aligned} (f \circ \text{brun}(s_0)) \diamond_M (g \circ \text{brun}(s_0)) &= ((a \rightarrow a \times s_0) \circ \text{uncu}(f) \circ \pi_1^M) \diamond_M (g \circ \text{brun}(s_0)) \\ \text{naturality of } \diamond_M : & = (a \rightarrow a \times s_0) \circ \text{uncu}(f) \diamond_M (\pi_1 \circ g \circ \text{brun}(s_0)) \\ \text{use Eq. (14.83)} : & = (a \rightarrow a \times s_0) \circ \text{uncu}(f) \diamond_M (\pi_1 \circ (b \rightarrow b \times s_0) \circ \text{uncu}(g) \circ \pi_1^M) . \end{aligned}$$

The right-hand side of the law is:

$$\begin{aligned} (f \diamond_P g) \circ \text{brun}(s_0) &= (a \rightarrow a \times s_0) \circ \text{uncu}(f \diamond_P g) \circ \pi_1^M \\ \text{definition of } \diamond_P : & = (a \rightarrow a \times s_0) \circ (\text{uncu}(f) \diamond_M \text{uncu}(g)) \circ \pi_1^M \\ \text{right naturality of } \diamond_M : & = (a \rightarrow a \times s_0) \circ \text{uncu}(f) \diamond_M (\text{uncu}(g) \circ \pi_1^M) . \end{aligned}$$

We find that the two sides of the law are *not* equal. The difference is the presence of a function $\pi_1 \circ (b \rightarrow b \times s_0)$ of type $B \times S \rightarrow B \times S$. That function replaces the updated state by the fixed initial state s_0 before applying the uncurried function g . This replacement erases the updated state, which leads to an incorrect composition of effects in the `StateT` monad.

If we set $M = \text{Id}$, we will obtain the same violation for the `State` monad's runner.

Monadic naturality laws To verify the monadic naturality of `flift`, apply to an arbitrary $m^{:M^A}$:

$$\begin{aligned} \text{expect to equal } m \triangleright \phi \triangleright \text{flift}^N : \quad & m \triangleright \text{flift}^M \triangleright \text{frun}(\phi) = m \triangleright \text{flift}^M \triangleright (p \rightarrow p \circ \phi) = (m \triangleright \text{flift}^M) \circ \phi \\ \text{use Eq. (14.81)} : \quad & = (s \rightarrow m \triangleright (a \rightarrow a \times s)^{\uparrow M}) \circ \phi = s \rightarrow m \triangleright (a \rightarrow a \times s)^{\uparrow M} \circ \phi \\ \text{naturality of } \phi : \quad & = s \rightarrow m \triangleright \phi \circ (a \rightarrow a \times s)^{\uparrow N} = m \triangleright \phi \triangleright \text{flift}^N . \end{aligned}$$

To verify the monadic naturality of `brun`:

$$\begin{aligned} \text{expect to equal } \text{brun}(s_0) \circ \phi : \quad & \text{frun}(\phi) \circ \text{brun}(s_0) = (p \rightarrow p \circ \phi) \circ (p \rightarrow s_0 \triangleright p \circ \pi_1^{\uparrow N}) \\ \text{compute composition} : \quad & = p \rightarrow s_0 \triangleright p \circ \phi \circ \pi_1^{\uparrow N} \\ \text{naturality of } \phi : \quad & = p \rightarrow s_0 \triangleright p \circ \pi_1^{\uparrow M} \circ \phi = \text{brun}(s_0) \circ \phi . \end{aligned}$$

14.8.2 The continuation monad transformer (ContT)

The transformer's type is defined by

$$T_{\text{Cont}}^{M,A} \triangleq (A \rightarrow M^R) \rightarrow M^R .$$

Unlike most other monad transformers, T_{Cont} is not covariant in the foreign monad M because its type contains M both in covariant and contravariant positions. As a result, we cannot define the foreign runner or the base lift functions. The corresponding monad transformer laws (the foreign runner laws, the base runner laws, and the monadic naturality laws) do not apply to T_{Cont} . This significantly limits the usability of the continuation monad transformer.

Another limitation is that the `Cont` monad does not have a fully parametric runner (there is no implementation of the type signature $\text{Cont}^{R,A} \rightarrow A$). The runners constructed in Section 10.1.9 will terminate only for values of type $\text{Cont}^{R,A}$ that eventually call their callbacks (of type $A \rightarrow R$). Those runners may be used only in applications where the available values of type $\text{Cont}^{R,A}$ are suitably restricted. Accordingly, the transformer `ContT` does not have fully parametric base runners, and it is not possible to formulate their laws. We will now verify the remaining laws of `ContT`.

Monad construction law The transformed monad T_{Cont}^M differs from the continuation monad `Cont` only in the type of the result (R), replacing $(A \rightarrow R) \rightarrow R$ by $(A \rightarrow M^R) \rightarrow M^R$. The monad's methods work the same way for all result types, so T_{Cont}^M is a monad for all type constructors M . This remains true even if M is itself not a functor.

Identity law If we set $M = \text{Id}$, the type of $T_{\text{Cont}}^{M,A}$ is equivalent to the base monad, $(A \rightarrow R) \rightarrow R$.

Lifting law The foreign lift function (`flift`) is defined by

$$\text{flift} : M^A \rightarrow T_{\text{Cont}}^{M,A} = M^A \rightarrow (A \rightarrow M^R) \rightarrow M^R , \quad \text{flift} \triangleq \text{flatMap}_M = m^{:M^A} \rightarrow f^{:A \rightarrow M^R} \rightarrow m \triangleright \text{flm}_M(f) .$$

To verify that `flift` is a monad morphism, we use the flipped Kleisli formulation for T_{Cont} . In that formulation, Kleisli functions of type $A \rightarrow T_{\text{Cont}}^{M,B} = A \rightarrow (B \rightarrow M^R) \rightarrow M^R$ are flipped into the type signature $(B \rightarrow M^R) \rightarrow A \rightarrow M^R$. For the `Cont` monad, the flipped Kleisli composition is equal to the ordinary *backward* function composition (see Statement 10.2.7.1):

$$f^{:(B \rightarrow M^R) \rightarrow A \rightarrow M^R} \circ_T g^{:(C \rightarrow M^R) \rightarrow B \rightarrow M^R} = (g \circ f)^{:(C \rightarrow M^R) \rightarrow A \rightarrow M^R} . \quad (14.84)$$

Let us also compute the result of flipping a lifted M -Kleisli function:

$$\text{flip}(f^{:A \rightarrow M^B} \circ \text{flift}) = p^{:B \rightarrow M^R} \rightarrow a^{:A} \rightarrow f(a) \triangleright \text{flm}_M(p) = p \rightarrow f \circ \text{flm}_M(p) = p \rightarrow f \diamond_M p . \quad (14.85)$$

The identity law of `flift` is now easy to verify. The flipped `pure` method for T_{Cont} is

$$\tilde{\text{pu}}_T : (A \rightarrow M^R) \rightarrow A \rightarrow M^R , \quad \tilde{\text{pu}}_T = \text{id}^{A \rightarrow M^R} .$$

To show that the law holds, we write:

$$\text{expect to equal } \tilde{pu}_T : \text{ flip}(\tilde{pu}_M ; \text{flift}) = p \rightarrow \tilde{pu}_M \diamond_M p = p \rightarrow p = \text{id} = \tilde{pu}_T \quad .$$

The composition law in the ordinary Kleisli formulation is

$$(f^{A \rightarrow M^B} ; \text{flift}) \diamond_T (g^{B \rightarrow M^C} ; \text{flift}) \stackrel{?}{=} (f \diamond_M g) ; \text{flift} \quad .$$

To transform the composition law to the flipped Kleisli formulation, we write:

$$\text{flip}(f ; \text{flift}) \tilde{\diamond}_T \text{flip}(g ; \text{flift}) \stackrel{?}{=} \text{flip}((f \diamond_M g) ; \text{flift}) \quad .$$

We complete the proof by using Eqs. (14.82)–(14.83) and the associativity of \diamond_M :

$$\begin{aligned} \text{flip}(f ; \text{flift}) \tilde{\diamond}_T \text{flip}(g ; \text{flift}) &= (q \rightarrow g \diamond_M q) ; (p \rightarrow f \diamond_M p) = q \rightarrow \underline{f \diamond_M (g \diamond_M q)} \\ &= q \rightarrow (f \diamond_M g) \diamond_M q = \text{flip}((f \diamond_M g) ; \text{flift}) \quad . \end{aligned}$$

14.8.3 The choice monad transformer

14.8.4 The co-density monad transformer (CodT)

The **co-density monad** over a functor F is defined as

$$\text{Cod}^{F,A} \triangleq \forall B. (A \rightarrow F^B) \rightarrow F^B$$

Properties:***

- $\text{Cod}^{F,\bullet}$ is a monad for any type constructor F .
- If F^\bullet is itself a monad then we have monad morphisms $\text{inC} : F^\bullet \rightsquigarrow \text{Cod}^{F,\bullet}$ and $\text{outC} : \text{Cod}^{F,\bullet} \rightsquigarrow F^\bullet$ such that $\text{inC} ; \text{outC} = \text{id}$. However, these functions are not isomorphisms. For example, when $F^A \triangleq R \rightarrow A$ (the `Reader` monad), we get (using the Yoneda identity):

$$\text{Cod}^{F,A} = \forall B. (\underline{A \rightarrow R} \rightarrow B) \rightarrow \underline{R \rightarrow B} = R \rightarrow \forall B. (A \times R \rightarrow B) \rightarrow B = R \rightarrow A \times R \quad .$$

The result is the `State` monad which describes the effects of `Reader` but additionally the effect of changing a value of type R and passing it to the next computation (which `Reader` cannot do). We find that, in general, Cod^F can describe all F -effects and some additional effects.

- If F is a monad with a known transformer T_F , we can write the monad transformer for Cod^F as

$$T_{\text{Cod}}^{M,A} = \forall B. (A \rightarrow T_F^{M,B}) \rightarrow T_F^{M,B} \quad .$$

However, this transformer does not have the base lifting morphism (is that so???)

$$\text{blift} : \left(\forall B. (A \rightarrow F^B) \rightarrow F^B \right) \rightarrow \forall C. (A \rightarrow T_F^{M,C}) \rightarrow T_F^{M,C} \quad ,$$

since this type signature cannot be implemented. The co-density transformer also does not have the required “runner” transformations `frun` and `brun`, (is that so???)

$$\begin{aligned} \text{frun} : (M^\bullet \rightsquigarrow N^\bullet) &\rightarrow \left(\forall B. (A \rightarrow T_F^{M,B}) \rightarrow T_F^{M,B} \right) \rightarrow \forall C. (A \rightarrow T_F^{N,C}) \rightarrow T_F^{N,C} \quad , \\ \text{brun} : \left(\left(\forall B. (A \rightarrow F^B) \rightarrow F^B \right) \rightarrow A \right) &\rightarrow \left(\forall C. (A \rightarrow T_F^{M,C}) \rightarrow T_F^{M,C} \right) \rightarrow M^A \quad . \end{aligned}$$

14.9 Summary and discussion

14.9.1 Some properties of monad morphisms

In this section we prove certain properties of monad morphisms used elsewhere in this chapter.

Statement 14.9.1.1 For any monad M , the function $\Delta : M^A \rightarrow M^A \times M^A$ is a monad morphism between monads M and $M \times M$.

Proof We use the definition of the product monad. The method $\text{pu}_{M \times M}$ is defined by

$$x \triangleright \text{pu}_{M \times M} = \text{pu}_M(x) \times \text{pu}_M(x) = x \triangleright \text{pu}_M \circ \Delta \quad ,$$

which is the identity law for Δ . To verify the composition law for Δ ,

$$\text{ftn}_M \circ \Delta = \Delta^{\uparrow M} \circ \Delta \circ \text{ftn}_{M \times M} \quad ,$$

we use the definition (10.17) of $\text{ftn}_{M \times M}$:

$$\text{ftn}_{M \times M} : M^{M^\bullet \times M^\bullet} \times M^{M^\bullet \times M^\bullet} \rightarrow M^\bullet \times M^\bullet = (\pi_1^{\uparrow M} \circ \text{ftn}_M) \boxtimes (\pi_2^{\uparrow M} \circ \text{ftn}_M) \quad ,$$

and compute:

$$\begin{aligned} \text{expect to equal } \text{ftn}_M \circ \Delta : & \quad \underline{\Delta^{\uparrow M} \circ \Delta} \circ \text{ftn}_{M \times M} \\ \text{naturality of } \Delta : & \quad = \Delta \circ (\Delta^{\uparrow M} \boxtimes \Delta^{\uparrow M}) \circ \underline{\text{ftn}_{M \times M}} \\ \text{definition of } \text{ftn}_{M \times M} : & \quad = \Delta \circ (\Delta^{\uparrow M} \boxtimes \Delta^{\uparrow M}) \circ (\pi_1^{\uparrow M} \circ \text{ftn}_M) \boxtimes (\pi_2^{\uparrow M} \circ \text{ftn}_M) \\ \text{composition law of } \boxtimes : & \quad = \Delta \circ (\underline{\Delta^{\uparrow M} \circ \pi_1^{\uparrow M} \circ \text{ftn}_M}) \boxtimes (\underline{\Delta^{\uparrow M} \circ \pi_2^{\uparrow M} \circ \text{ftn}_M}) \\ \text{simplify } \Delta \circ \pi_i = \text{id} : & \quad = \Delta \circ \text{ftn}_M \boxtimes \text{ftn}_M \\ \text{duplication law of } \Delta : & \quad = \text{ftn}_M \circ \Delta \quad . \end{aligned}$$

Statement 14.9.1.2 For any monads K, L, M, N and monad morphisms $\phi : K \rightsquigarrow M$ and $\chi : L \rightsquigarrow N$, the pair product $\phi \boxtimes \chi : K^\bullet \times L^\bullet \rightsquigarrow M^\bullet \times N^\bullet$ is a monad morphism between the product monads $K \times L$ and $M \times N$.

Proof The definitions of $\text{pu}_{K \times L}$ and $\text{pu}_{M \times N}$ are

$$\text{pu}_{K \times L} = \text{pu}_K \boxtimes \text{pu}_L \quad , \quad \text{pu}_{M \times N} = \text{pu}_M \boxtimes \text{pu}_N \quad .$$

The identity law for $\phi \boxtimes \chi$ is verified by

$$\begin{aligned} \text{expect to equal } \text{pu}_{M \times N} : & \quad \underline{\text{pu}_{K \times L} \circ \phi \boxtimes \chi} \\ \text{definition of } \text{pu}_{K \times L} : & \quad = \text{pu}_K \boxtimes \text{pu}_L \circ \phi \boxtimes \chi \\ \text{composition law of } \boxtimes : & \quad = (\underline{\text{pu}_K \circ \phi}) \boxtimes (\underline{\text{pu}_L \circ \chi}) \\ \text{identity laws for } \phi, \chi : & \quad = \text{pu}_M \boxtimes \text{pu}_N = \text{pu}_{M \times N} \quad . \end{aligned}$$

To verify the composition law for $\phi \boxtimes \chi$, we use the definitions

$$\begin{aligned} \text{ftn}_{K \times L} &= (\pi_1^{\uparrow K} \circ \text{ftn}_K)^{K^{K^\bullet \times L^\bullet} \rightarrow K^\bullet} \boxtimes (\pi_2^{\uparrow L} \circ \text{ftn}_L)^{L^{K^\bullet \times L^\bullet} \rightarrow L^\bullet} \quad , \\ \text{ftn}_{M \times N} &= (\pi_1^{\uparrow M} \circ \text{ftn}_M) \boxtimes (\pi_2^{\uparrow N} \circ \text{ftn}_N) \quad . \end{aligned}$$

Denote $\psi \triangleq \phi \boxtimes \chi$ for brevity. The required law is

$$\text{ftn}_{K \times L} \circ \psi = \psi^{\uparrow(K \times L)} \circ \psi \circ \text{ftn}_{M \times N} \quad .$$

The left-hand side of this law is

$$\begin{aligned}
 & \text{ftn}_{K \times L} ; \psi \\
 \text{definition of ftn}_{M \times N} : &= (\pi_1^{\uparrow K} ; \text{ftn}_K) \boxtimes (\pi_2^{\uparrow L} ; \text{ftn}_L) ; \phi \boxtimes \chi \\
 \text{composition law of } \boxtimes : &= (\pi_1^{\uparrow K} ; \text{ftn}_K ; \phi) \boxtimes (\pi_2^{\uparrow L} ; \text{ftn}_L ; \chi)
 \end{aligned}$$

The right-hand side is

$$\begin{aligned}
 & \psi^{\uparrow(K \times L)} ; \psi ; \text{ftn}_{M \times N} \\
 \text{lifting to } K \times L : &= (\psi^{\uparrow K} \boxtimes \psi^{\uparrow L}) ; (\phi \boxtimes \chi) ; \text{ftn}_{M \times N} \\
 \text{definition of ftn}_{M \times N} : &= \psi^{\uparrow K} \boxtimes \psi^{\uparrow L} ; \phi \boxtimes \chi ; (\pi_1^{\uparrow M} ; \text{ftn}_M) \boxtimes (\pi_2^{\uparrow N} ; \text{ftn}_N) \\
 \text{composition law of } \boxtimes : &= (\psi^{\uparrow K} ; \phi ; \pi_1^{\uparrow M} ; \text{ftn}_M) \boxtimes (\psi^{\uparrow L} ; \chi ; \pi_2^{\uparrow N} ; \text{ftn}_N) .
 \end{aligned}$$

Consider now the first part of the pair product in the last line:

$$\begin{aligned}
 & \psi^{\uparrow K} ; \phi ; \pi_1^{\uparrow M} ; \text{ftn}_M \\
 \text{naturality of } \phi : &= \phi ; \psi^{\uparrow M} ; \pi_1^{\uparrow M} ; \text{ftn}_M \\
 \text{projection law of } \pi_1 : &= \phi ; \pi_1^{\uparrow M} ; \phi^{\uparrow M} ; \text{ftn}_M \\
 \text{naturality of } \phi : &= \pi_1^{\uparrow K} ; \phi ; \phi^{\uparrow M} ; \text{ftn}_M \\
 \text{composition law of } \phi : &= \pi_1^{\uparrow K} ; \text{ftn}_K ; \phi .
 \end{aligned}$$

In the same way, we find that the second part of the pair product is $\pi_2^{\uparrow L} ; \text{ftn}_L ; \chi$, and so the composition law holds.

Statement 14.9.1.3 For any monads M and N , the projection function $\pi_1 : M^\bullet \times N^\bullet \rightsquigarrow M^\bullet$ is a monad morphism. Same for $\pi_2 : M^\bullet \times N^\bullet \rightsquigarrow N^\bullet$.

Proof It is sufficient to verify the laws for π_1 ; the proof for π_2 will be similar. The identity law:

$$\begin{aligned}
 \text{expect to equal } \text{pu}_M : & \text{pu}_{M \times N} ; \pi_1 \\
 \text{definition of } \text{pu}_{M \times N} : &= (\text{pu}_M \boxtimes \text{pu}_N) ; \pi_1 \\
 \text{projection law of } \pi_1 : &= \text{pu}_M .
 \end{aligned}$$

The composition law:

$$\begin{aligned}
 \text{expect to equal } \pi_1 ; \pi_1^{\uparrow M} ; \text{ftn}_M : & \text{ftn}_{M \times N} ; \pi_1 \\
 \text{definition of ftn}_{M \times N} : &= (\pi_1^{\uparrow M} ; \text{ftn}_M) \boxtimes (\pi_2^{\uparrow N} ; \text{ftn}_N) ; \pi_1 \\
 \text{projection law of } \pi_1 : &= \pi_1 ; \pi_1^{\uparrow M} ; \text{ftn}_M .
 \end{aligned}$$

Statement 14.9.1.4 For any monads M and N , the component-swapping function $\sigma : M^\bullet \times N^\bullet \rightsquigarrow N^\bullet \times M^\bullet$ is a monad morphism.

Proof The code for σ can be written as a combination of other functions as $\sigma = \Delta ; (\pi_2 \boxtimes \pi_1)$. The functions Δ , π_1 , and π_2 are monad morphisms by Statements 14.9.1.1 and 14.9.1.3. The function product $\pi_1 \boxtimes \pi_2$ is a monad morphism by Statement 14.9.1.2. So σ is a composition of monad morphisms; by Statement 10.3.4.5, σ is a monad morphism.

14.9.2 Exercises

Exercise 14.9.2.1 Statement 14.9.1.1 shows that the duplicating function $\Delta : M \rightsquigarrow M \times M$ is a monad morphism for an arbitrary monad M . Prove that the same holds for all similarly defined functions $M \rightsquigarrow M \times M \times M$, $M \rightsquigarrow M \times M \times M \times M$, and so on.

Exercise 14.9.2.2 We know from Statement 10.3.4.6 that pu_M is a monad morphism the identity monad (Id) to a monad M . Show that there are no other such monad morphisms ϕ ,

$$\phi : \text{Id} \rightsquigarrow M \quad , \quad \phi^A : A \rightarrow M^A \quad .$$

In other words, show that any such monad morphism ϕ must be equal to M 's `pure` method: $\phi = \text{pu}_M$.

Exercise 14.9.2.3 Assume that there exists a monad morphism $\phi : \text{Id} \rightsquigarrow M$ between the unit monad $L^A \triangleq \text{Id}$ and some monad M , and show that M must be itself a unit monad ($M^A \triangleq \text{Id}$).

Exercise 14.9.2.4 For any given monad M , consider a function $\text{dbl} : M^A \rightarrow M^A$ defined as

```
def double[M[_]: Monad, A](m: M[A]): M[A] = for {
  _ <- m
  x <- m
} yield x
```

$$m : M^A \triangleright \text{dbl} \triangleq m \triangleright \text{flm}_M(_^A \rightarrow m) \\ = m \triangleright (_ \rightarrow m)^{\uparrow M} ; \text{ftn}_M \quad .$$

This function “doubles” the M -effect in a given monadic value m , for an arbitrary monad M .

(a) Show that the function `double` is a *monadically natural* transformation $M^A \rightarrow M^A$ with respect to the monad type parameter M .

(b) Show that (at least for some monads M) the function `double` is *not* a monad morphism $M \rightsquigarrow M$.

Exercise 14.9.2.5 Suppose T_L^M is a lawful monad transformer for the base monad L and a foreign monad M . Can we modify T_L^M and construct another monad transformer for L that still satisfies all transformer laws?

One possibility is to compose T with an extra layer of the monads L or M . Define U and V by

$$U^A \triangleq L \circ T_L^M \quad , \quad V^A \triangleq M \circ T_L^M \quad .$$

In a shorter notation, $U \triangleq L \circ T_L^M$ and $V \triangleq M \circ T_L^M$. We have the `swap` functions

$$\begin{aligned} \text{sw}_{L,T} &: T_L^M \circ L \rightsquigarrow L \circ T_L^M \quad , \\ \text{sw}_{M,T} &: T_L^M \circ M \rightsquigarrow M \circ T_L^M \quad , \end{aligned}$$

defined using the already given methods `flift` and `blift` of T_L^M as

$$\begin{aligned} \text{sw}_{L,T} &= \text{blift}^{\uparrow T} ; \text{ftn}_T ; \text{pu}_L \quad , \\ \text{sw}_{M,T} &= \text{flift}^{\uparrow T} ; \text{ftn}_T ; \text{pu}_M \quad . \end{aligned}$$

We can define the monad methods ftn_U and ftn_V using these `swap` functions. Will U and/or V be lawful monad transformers for L ?

Exercise 14.9.2.6 Show that there exist monad morphisms between the selection $(A \rightarrow R) \rightarrow A$ and the continuation $(A \rightarrow R) \rightarrow R$ monads. Are those morphisms (left, right, or full) inverses of each other?

Exercise 14.9.2.7 Assuming that

- L and M are monads,
- the method `swap` is a natural transformation $M \circ L \rightsquigarrow L \circ M$,
- the method ftn_T of the monad $T = L \circ M$ is *defined* via `swap` by Eq. (14.11),

show that the two interchange laws must hold for ftn_T :

$$\begin{aligned} \text{inner-interchange} &: \text{ftn}_L ; \text{ftn}_T = \text{ftn}_T^{\uparrow L} ; \text{ftn}_L \quad , \\ \text{outer-interchange} &: \text{ftn}_T ; \text{ftn}_M^{\uparrow L} = \text{ftn}_M^{\uparrow L \uparrow M \uparrow L} ; \text{ftn}_T \quad . \end{aligned}$$

Exercise 14.9.2.8 With the same assumptions as Exercise 14.9.2.7 and additionally assuming the inner and outer identity laws for `swap` (see Statement 14.3.2.1), show that the monad $T^\bullet \triangleq L^{M^\bullet}$ satisfies two “pure compatibility” laws,

$$\begin{aligned} \text{inner-pure-compatibility : } \text{ftn}_L &= \text{pu}_M^{\uparrow L} ; \text{ftn}_T : L^{L^{M^A}} \rightarrow L^{M^A} , \\ \text{outer-pure-compatibility : } \text{ftn}_M^{\uparrow L} &= \text{pu}_L^{\uparrow T} ; \text{ftn}_T : L^{M^{M^A}} \rightarrow L^{M^A} , \end{aligned}$$

or, expressed equivalently through the `flatMap` methods instead of `flatten`:

$$\begin{aligned} \text{flm}_L(f^{:A \rightarrow L^{M^B}}) &= \text{pure}_M^{\uparrow L} ; \text{flm}_T(f) , \\ (\text{flm}_M(f^{:A \rightarrow M^B}))^{\uparrow L} &= \text{pure}_L^{\uparrow T} ; \text{flm}_T(f^{\uparrow L}) . \end{aligned}$$

Exercise 14.9.2.9 The construction of the free pointed monad on M , namely $N^A \triangleq A + M^A$ (where M is a given monad), uses a helper function γ_M defined by

$$\gamma_M : A + M^A \rightarrow M^A , \quad \gamma_M \triangleq \begin{array}{c|c} & M^A \\ \hline A & \text{pu}_M \\ M^A & \text{id} \end{array} .$$

Show that γ_M is a monad morphism between the monads N and M .

Exercise 14.9.2.10 For any monads K and M and a given monad morphism $\phi : K \rightsquigarrow M$, show that one can implement a corresponding monad morphism $\psi : L \rightsquigarrow N$ between free pointed monads $L^A \triangleq A + K^A$ and $N^A \triangleq A + M^A$.

Exercise 14.9.2.11 For a given arbitrary monad K , consider the free pointed monad $L^A \triangleq A + K^A$. Show that the monad transformer T_L^M for L cannot be defined as the monad $P^A \triangleq A + T_K^{M,A}$ because (at least for some monads M) the required monad morphism $\text{lift}_L : M \rightsquigarrow P$ cannot be implemented.

Exercise 14.9.2.12 Assume that M is a given monad, and define N as the free pointed monad $N^A \triangleq A + M^A$.

(a) Given a monad morphism $\delta : M \rightsquigarrow \text{Opt}$ (from M to the `Option` monad), show that one can implement a monad morphism $\phi : M \rightsquigarrow N$.

(b) Given a monad morphism $\phi : M \rightsquigarrow N$, show that one can implement a monad morphism $\delta : M \rightsquigarrow \text{Opt}$.

Problem 14.9.2.13 Consider a monad morphism $\varepsilon^{M,A} : M^A \rightarrow M^A$ that is defined for all monads M and is monadically natural in the monad parameter M . So, ε must satisfy the laws listed here:

$$\begin{aligned} \text{naturality law : } \varepsilon^{M,A} ; (f^{:A \rightarrow B})^{\uparrow M} &= (f^{:A \rightarrow B})^{\uparrow M} ; \varepsilon^{M,B} , \\ \text{monad morphism laws : } \text{pu}_M ; \varepsilon &= \text{pu}_M , \quad \varepsilon^{\uparrow M} ; \varepsilon ; \text{ftn}_M = \text{ftn}_M ; \varepsilon , \\ \text{monadic naturality law : } \varepsilon^{M,A} ; \phi^{:M^A \rightarrow N^A} &= \phi^{:M^A \rightarrow N^A} ; \varepsilon^{N,A} , \end{aligned}$$

where $f^{:A \rightarrow B}$ is an arbitrary function, M, N are arbitrary monads, and $\phi : M \rightsquigarrow N$ is an arbitrary monad morphism. Prove that any such ε must be an identity function, $\varepsilon = \text{id}^{:M^A \rightarrow M^A}$, or show an example of such ε not equal to identity.⁹

⁹The author of this book does not know the solution.

Part IV

Discussions

15 Sample problems

1. Compute the smallest integer expressible as a sum of two cubed integers in more than one way.
2. Read a text file, split it by spaces into words, and print the word counts, sorted by decreasing count.
3. FPIS exercise 2.2: Check whether a sequence `Seq[A]` is sorted according to a given ordering function of type `(A, A) => Boolean`.
4. FPIS exercise 3.24: Implement a function `hasSubsequence` that checks whether a `List` contains another `List` as a subsequence. For instance, `List(1,2,3,4)` would have `List(1,2)`, `List(2,3)`, and `List(4)` as subsequences, among others. (Dynamic programming?)
5. (Bird, de Moor page 20) Derive the following identity between functions $F^A \rightarrow F^A$, for any filterable functor F and any predicate $p^{A \rightarrow 2}$:

$$\text{filt}_F(p) = (\Delta ; \text{id} \boxtimes p)^{\uparrow F} ; \text{filt}_F(\pi_2) ; \pi_1^{\uparrow F} .$$

6. Define a monoid of partial functions with fixed types $P \rightarrow Q$.

```
final case class PFM[P, Q](pf: PartialFunction[P, Q])
// After defining a monoid instance, the following code must work:
val p1 = PFM[Option[Int], String] { case Some(3) => "three" }
val p2 = PFM[Option[Int], String] {
  case Some(20)  => "twenty"
  case None      => "empty"
}
p1 |+| p2 // Must be the same as the concatenation of all 'case' clauses.
```

7. Consider a typeclass called “Splittable” for functors F^\bullet that have an additional method

$$\text{split}^{A,B} : F^{A+B} \rightarrow F^A + B$$

with the non-degeneracy law for functions $F^A \rightarrow F^A$,

$$(x^{A \rightarrow x} + \mathbb{0}^B)^{\uparrow F} ; \text{split} = y^{F^A \rightarrow y} + \mathbb{0}^B$$

and the special associativity law for functions $F^{A+B+C} \rightarrow F^A + B + C$,

$$\text{split}^{A+B,C} ; \left| \begin{array}{c|cc} & F^A + B & C \\ \hline F^{A+B} & \text{split}^{A,B} & \mathbb{0} \\ C & \mathbb{0} & \text{id} \end{array} \right| = \text{split}^{A,B+C} .$$

Show that all polynomial functors F^\bullet belong to this typeclass. Show that exponential functors such as $F^A \triangleq Z \rightarrow A$ do not.

8. Given two fixed types P, Q that are not known to be the same, consider the contrafunctors $F^A \triangleq ((A \rightarrow P) \rightarrow P) \rightarrow Q$ and $G^A \triangleq A \rightarrow Q$. Show that there exist natural transformations $F^\bullet \rightsquigarrow G^\bullet$ and $G^\bullet \rightsquigarrow F^\bullet$. Show that these transformations are not isomorphisms.

9. Given two fixed types P, Q that are not known to be the same, show that the functor $L^A \triangleq (((A \rightarrow P) \rightarrow Q) \rightarrow Q) \rightarrow P$ is a semimonad but not a full monad. (When $P \cong Q$, the functor L is full monad because L is equivalent to a composition of the continuation monad with itself. See Exercise 10.2.9.18.)
10. When M is a given monad, show that $M \circ M \circ \dots \circ M$ (with finitely many M) is also a monad.

16 “Applied functional type theory”: A proposal

What exactly is the extent of “theory” that a software engineer must know in order to be a proficient functional programmer? This book proposes an answer to that question by presenting a coherent body of theoretical knowledge that, in the author’s view, is the theory that is practically useful for writing code. This body of knowledge may be viewed as a new branch of computer science, called **applied functional type theory** (AFTT). This is the area of theoretical computer science serving the practical needs of functional programmers working as software engineers.

It is for those practitioners, rather than for academic researchers, that this book sets out to examine the functional programming inventions over the last 30 years, — such as the “functional pearls” papers¹ — and to determine the scope of theoretical material that has demonstrated its pragmatic usefulness and thus belongs to AFTT, as opposed to material that is purely academic and may be tentatively omitted. This book is a first step towards formulating AFTT.

In this book, code is written in Scala because the author is fluent in that language. However, most of this material will work equally well in Haskell, OCaml, and other FP languages. This is because the science of functional programming, called AFTT, is not a set of tricks specific to Scala or Haskell. An advanced user of any other functional programming language will have to face the same questions and struggle with the same practical issues.

16.1 AFTT is not covered by computer science curricula

Traditional courses of theoretical computer science (algorithms and data structures, complexity theory, distributed systems, databases, network systems, compilers, operating systems) are largely not relevant to AFTT. Courses in programming language theory are more relevant but are not presented at an appropriate level. To an academic computer scientist, the “theory behind Haskell” is the polymorphic lambda-calculus, the type-theoretic “System $F\omega$ ”, and formal semantics. These theories guided the design of the Haskell language and define rigorously what a Haskell program “means” in a mathematical sense. The “theory behind Scala” is the dependent object type (DOT) calculus.²

However, a practicing Haskell or Scala programmer is not concerned with designing Haskell or Scala, or with proving any theoretical properties of those languages. A practicing programmer is mainly concerned with *using* a chosen programming language to write code.

Neither the theory of lambda-calculus, nor proofs of type-theoretical properties of “System $F\omega$ ”, nor theories of formal semantics, nor domain theory will actually help a programmer to write code. So, all these theories are not within the scope of AFTT. Functional programming does not require any graduate-level theoretical studies.

As an example of theoretical material that *is* within the scope of AFTT, consider applicative functors (Chapter 11). It is essential for a practicing functional programmer to be able to recognize and use applicative functors. An applicative functor is a data structure specifying declaratively a set of operations that can run independently of each other. Programs may combine these operations, for example, to execute them in parallel, or to refactor the program for better maintainability.

To use this functionality, the programmer must begin by checking whether a given data structure satisfies the laws of applicative functors. In a given application, a data structure may be dictated

¹https://wiki.haskell.org/Research_papers/Functional_pearls

²<https://www.scala-lang.org/blog/2016/02/03/essence-of-scala.html>

in part by the business logic rather than by a programmer's choice. The programmer first writes down the type of that data structure and the code implementing the required methods; then checks that the laws hold. The data structure may need to be adjusted in order to fit the definition of an applicative functor or its laws.

So, before starting to write the actual code, the programmer needs to perform a certain amount of symbolic derivations. That work is best done using pen and paper, writing equations in a mathematical notation. Once the applicative laws are verified, the programmer proceeds to write code using that data structure.

The mathematical proofs and derivations assure that the data structure will satisfy the laws of applicative functors, no matter how the rest of the program is written. So, for example, it is assured that the relevant effects can be automatically parallelized and will still work correctly. In this way, AFTT directly guides the programmer and helps write correct code.

Applicative functors were discovered by practitioners who were using Haskell in programs such as parser combinators, compilers, and domain-specific languages for parallel computations. However, applicative functors are not a feature of Haskell: they can be used in Scala or any other functional programming language. And yet, no standard computer science textbook defines applicative functors, motivates their laws, explores their structure on examples, or shows data structures that are *not* applicative functors and explains why. Books on category theory rarely mention applicative functors (known in mathematics as "lax monoidal" functors³).

16.2 AFTT is not category theory, type theory, or formal logic

It appears that AFTT includes a small selection of results from category theory, formal logic, and type theory. However, software engineers would not benefit from traditional academic courses in these subjects, because their presentation is too abstract and at the same time lacks specific results necessary for practical programming. In other words, the traditional academic courses answer questions that academic computer scientists have, not questions that practicing software engineers have.

There exist several books intended as presentations of category theory "for computer scientists"⁴ or "for programmers".⁵ However, those books do not explain many concepts relevant to programming, such as applicative or traversable functors. Instead, those books dwell on concepts (e.g., limits, enriched categories, topoi) that have no applications in practical functional programming today.

Typical questions in academic books are "Is X an introduction rule or an elimination rule" and "Does property Y hold in non-small categories or only in the category of sets". Questions a Scala programmer might ask are "Can we compute a value of type `Either[Z, R => A]` from a value of type `R => Either[Z, A]`" and "Is the type constructor `F[A] = Option[(A, A, A)]` a monad or only an applicative functor". The scope of AFTT includes answering the last two questions but *not* the first two.

A software engineer hoping to understand the science of functional programming will not find the concepts of filterable, applicative, or traversable functors in any books on category theory, including books intended for programmers. And yet, these concepts are necessary to obtain mathematically correct implementations of `filter`, `zip`, and `fold`, which are important and widely used operations.

To compensate for the lack of AFTT textbooks, programmers have written many online tutorials for each other, trying to explain the theoretical concepts necessary for practical work. The term "monad tutorial" became infamous because so many were written and posted online.⁶ Tutorials were also written about applicative functors, traversable functors, free monads, etc., showing a real unfulfilled need for presenting relevant theory in an applied setting.

For example, "free monads" became popular in the Scala community around 2015. Many talks about free monads were presented at Scala engineering conferences, giving different implementation but never formulating rigorously the properties required for a piece of code to be a valid

³https://en.wikipedia.org/wiki/Monoidal_functor

⁴<https://www.amazon.com/dp/0262660717>

⁵<https://github.com/hmemcpy/milewski-ctfp-pdf>

⁶<https://www.johndcook.com/blog/2014/03/03/monads-are-hard-because/>

implementation of the free monad. Without knowing the required mathematical properties of free monads, a programmer cannot make sure that a given implementation is correct. However, books on category theory define free monads in a way that is unsuitable for programming applications: a free monad is an adjoint functor to a forgetful functor from a Kleisli category to the category of sets.⁷ Such “academic” definitions can be used neither as guidance for writing code and ensuring that the code is correct, nor as a conceptual explanation that a learner would understand.

Perhaps the best selection of AFTT tutorial material today can be found in the Haskell Wikibooks.⁸ However, those tutorials are incomplete and limited to explaining the use of Haskell. Many of them are suitable neither as a first introduction nor as a reference on AFTT. Also, the Haskell Wikibooks tutorials rarely show any derivations of laws or explain the required techniques.

Apart from referring to some notions from category theory, AFTT also uses concepts from type theory and formal logic. However, existing textbooks on type theory and formal logic focus on domain theory and proof theory — which presents a lot of information that will be difficult to learn for practicing programmers and yet will never be applicable in their daily work. At the same time, those academic books never mention practical techniques used in many functional programming libraries today, such as reasoning about and implementing quantified types, types parameterized by type constructors, or partial type-level functions (known as “typeclasses”).

The proper scope of AFTT is to help the programmer with practical tasks such as:

- Deciding whether two data types are equivalent and implementing the isomorphism transformations. For example, the Scala type `(A, Either[B, C])` is equivalent to `Either[(A, B), (A, C)]`.
- Checking whether a definition of a recursive type is “reasonable”, i.e., does not lead to infinite loops. An example of an “unreasonable” recursive type definition is `case class Bad(x: Bad)`.
- Deciding whether a function with a given type signature can be implemented. For example,

```
def f[Z,A,R]: (R => Either[Z, A]) => Either[Z, R => A] = ??? // Cannot be implemented.
def g[Z,A,R]: Either[Z, R => A] => (R => Either[Z, A]) = ??? // Can be implemented.
```

- Deriving an implementation of a function from its type signature and checking required laws. For example, deriving the `flatMap` method and checking the laws for the `Reader` monad,

```
def flatMap[Z, A, B](r: Z => A)(f: A => Z => B): Z => B = ???
```

- Deriving a simpler but equivalent code by calculating with functions, equations, and laws.

These are real-world applications of type theory and the Curry-Howard correspondence, but existing books on type theory and logic do not give practical recipes for performing these tasks.

Books such as *Scala with Cats*⁹, *Functional programming simplified*¹⁰, and *Functional programming for mortals*¹¹ are primarily focused on explaining practical aspects of functional programming and do not derive the mathematical laws for e.g., applicative, monadic, or traversable functors.

The only Scala-based AFTT textbook is *Functional Programming in Scala*¹². It balances practical coding with theoretical developments and laws. *Program design by calculation*¹³ is another (Haskell-oriented) AFTT book in progress. The present book, *Science of Functional Programming*¹⁴, is written at about the same level but aims at better motivation for mathematical concepts and a wider range of pedagogical examples that help build the necessary intuition and formal technique.

⁷“A monad is just a monoid in the category of endofunctors. What’s the problem?” as the joke goes. For background information about that joke, see <https://stackoverflow.com/questions/3870088/>

⁸<https://en.wikibooks.org/wiki/Haskell>

⁹<https://underscore.io/books/scala-with-cats/>

¹⁰<https://alvinalexander.com/scala/functional-programming-simplified-book>

¹¹<http://www.lulu.com/shop/search.ep?contributorId=1600066>

¹²<https://www.manning.com/books/functional-programming-in-scala>

¹³<http://www4.di.uminho.pt/~jno/ps/pdbc.pdf>

¹⁴<https://github.com/winitzki/sof>

17 Essay: Software engineers and software artisans

Let us look at the differences between the kind of activities we ordinarily call engineering, as opposed to artisanship or craftsmanship. It will then become apparent that today's computer programmers are better understood as "software artisans" rather than software engineers.

17.1 Engineering disciplines

Consider what kinds of process a mechanical engineer, a chemical engineer, or an electrical engineer follows in their work, and what kind of studies they require for proficiency in their work.

A mechanical engineer studies¹ calculus, linear algebra, differential geometry, and several areas of physics such as theoretical mechanics, thermodynamics, and elasticity theory, and then uses calculations to guide the design of a bridge, say. A chemical engineer studies² chemistry, thermodynamics, calculus, linear algebra, differential equations, some areas of physics such as thermodynamics and kinetic theory, and uses calculations to guide the design of a chemical process, say. An electrical engineer studies³ advanced calculus, linear algebra, and several areas of physics such as electrodynamics and quantum theory, and uses calculations to design an antenna or a microchip.

The pattern here is that an engineer uses mathematics and natural sciences in order to design new devices. Mathematical calculations and scientific reasoning are required *before* drawing a design, let alone building a real device or machine.

Some of the studies required for engineers include arcane abstract concepts such as a "rank-4 elasticity tensor"⁴ (used in calculations of elasticity of materials), "Lagrangian with non-holonomic constraints"⁵ (used in robotics), the "Gibbs free energy" (for chemical reactor design⁶), or the "Fourier transform of the delta function"⁷ and the "inverse Z-transform"⁸ (for digital signal processing).

To be sure, a significant part of what engineers do is not covered by any theory: the *know-how*, the informal reasoning, the traditional knowledge passed on from expert to novice, — all those skills that are hard to formalize are important. Nevertheless, engineering is crucially based on natural science and mathematics for some of its decision-making about new designs.

17.2 Artisanship: Trades and crafts

Now consider what kinds of things shoemakers, plumbers, or home painters do, and what they have to learn in order to become proficient in their profession.

A novice shoemaker, for example, begins by copying some drawings⁹ and goes on to cutting leather in a home workshop. Apprenticeships proceed via learning by doing, with comments and

¹<https://www.colorado.edu/mechanical/undergraduate-students/curriculum>

²<https://www.colorado.edu/engineering/sample-undergraduate-curriculum-chemical>

³<http://archive.is/XYLyE>

⁴https://serc.carleton.edu/NAGTWorkshops/mineralogy/mineral_physics/tensors.html

⁵<https://arxiv.org/abs/math/0008147>

⁶<https://www.amazon.com/Introduction-Chemical-Engineering-Kinetics-Reactor/dp/1118368258>

⁷<https://www.youtube.com/watch?v=KAbqISZ6SHQ>

⁸<http://archive.is/SsJqP>

⁹<https://youtu.be/cY5MY0czMAk?t=141>

instructions from an expert. After a few years of study (for example, a painter apprenticeship in California¹⁰ can be as short as 2 years), a new artisan is ready to start productive work.

All these trades operate entirely from tradition and practical experience. The trades do not require academic study because there is no formal theory from which to proceed. Of course, there is *a lot* to learn in the crafts, and it takes prolonged effort to become a good artisan in any profession. But there are no rank-4 tensors to calculate, nor any differential equations to solve; no Fourier transforms applied to delta functions, and no Lagrangians with non-holonomic constraints.

Artisans do not study science or mathematics because their professions do not make use of any formal theory for guiding their designs or processes.

17.3 Programmers today are artisans, not engineers

Programmers are *not engineers* in the sense we normally see the engineering professions.

17.3.1 No requirement of formal study

According to a recent Stack Overflow survey¹¹, about 56% of working programmers do not have a CS degree. The author of this book is a self-taught programmer who has degrees in physics but never formally studied computer science or taken any academic courses in algorithms, data structures, computer networks, compilers, programming languages, or other computer science topics.

A large fraction of successful programmers have no college degrees and perhaps *never* studied formally. They acquired all their knowledge and skills through self-study and practical work. A prominent example is Robert C. Martin¹², an outspoken guru in the arts of programming. He routinely refers to programmers as artisans¹³ and uses the appropriate imagery: novices and masters, trade and craft, the honor of the guild, etc. He compares programmers to plumbers, electricians, lawyers, and surgeons, but never to mathematicians, physicists, or engineers of any kind. According to one of his blog posts¹⁴, he started working at age 17 as a self-taught programmer, and then went on to more jobs in the software industry; he never mentioned going to college. It is clear that R. C. Martin *is* an expert craftsman and that he did *not* need academic study to master his craft.

Here is another opinion¹⁵ (emphasis is theirs):

Software Engineering is unique among the STEM careers in that it absolutely does *not* require a college degree to be successful. It most certainly does not require licensing or certification. *It requires experience.*

This description fits a career in crafts — but certainly not a career, say, in electrical engineering.

The high demand for software developers gave rise to “developer boot camps”¹⁶ — vocational schools that educate new programmers in a few months through purely practical training, with no formal theory or mathematics involved. These vocational schools are successful¹⁷ in job placement. But it is unimaginable that a 6-month crash course or even a 2-year vocational school could prepare engineers to work successfully on designing, e.g., analog quantum computers¹⁸ *without* ever teaching them quantum physics or calculus.

¹⁰http://www.calapprenticeship.org/programs/painter_apprenticeship.php

¹¹<https://thenextweb.com/insider/2016/04/23/dont-need-go-college-anymore-programmer/>

¹²https://en.wikipedia.org/wiki/Robert_C._Martin

¹³<https://blog.cleancoder.com/uncle-bob/2013/02/01/The-Humble-Craftsman.html>

¹⁴<https://blog.cleancoder.com/uncle-bob/2013/11/25/Novices-Coda.html>

¹⁵<http://archive.is/tAKQ3>

¹⁶<http://archive.is/Gk0L9>

¹⁷<http://archive.is/E9FXP>

¹⁸<https://www.dwavesys.com/quantum-computing>

17.3.2 No mathematical formalism guides software development

Most books on software engineering contain no formulas or equations, no mathematical derivations, and no precise definitions of the various technical terms they are using (such as “object-oriented” or “module’s responsibilities”). Some of those books¹⁹ also have almost no program code in them; instead they are filled with words and illustrative diagrams. These books talk about how programmers should approach their job, how to organize the work flow and the code architecture, etc., in vague and general terms: “code is about detail”, “you must never abandon the big picture”, “avoid tight coupling in your modules”, “a class must serve a single responsibility”, and so on. Practitioners such as R. C. Martin never studied any formalisms and do not think in terms of formalisms; instead, they summarize their programming experience in vaguely formulated heuristic “principles”.²⁰

In contrast, textbooks on mechanical or electrical engineering include a significant amount of mathematics. The design of a microwave antenna is guided not by an “open and closed module principle” but by solving the relevant differential equations²¹ of electrodynamics.

Donald Knuth’s classic textbook is called “*The Art of Programming*”. It is full of tips and tricks about how to program; but it does not provide any formal theory that could guide programmers in actually *writing* programs. There is nothing in that book that would be similar to the way mathematical formalism guides designs in electrical or mechanical engineering. If Knuth’s books were based on such formalism, they would have looked quite differently: some theory would be first explained and then applied to help us write code.

Knuth’s books provide many rigorously derived algorithms. But algorithms are similar to patented inventions: they can be used immediately without further study. Understanding an algorithm is not similar to understanding a mathematical theory. Knowing one algorithm does not make it easier to develop another algorithm in an unrelated domain. In comparison, knowing how to solve differential equations will be applicable to thousands of different areas of science and engineering.

A book exists²² with the title “Science of Programming”, but the title is misleading. The author does not propose a science, similar to physics, at the foundation of the process of designing programs, similarly to how calculations in quantum physics predict the properties of a quantum device. The book claims to give precise methods that guide programmers in writing code, but the scope of proposed methods is narrow: the design of simple algorithms for iterative manipulation of data. The procedure suggested in that book is far from a formal mathematical *derivation* of programs from specifications. (A book with that title²³ similarly disappoints.) In any case, programmers today are oblivious to these books and do not use the methods explained there.

Standard computer science courses today do not teach a true *engineering* approach to software construction. They do teach analysis of programs using formal mathematical methods; the main such methods are complexity analysis²⁴ (the “big-*O* notation”) and formal verification²⁵. But programs are analyzed or verified only *after* they are somehow written. Theory does not guide the actual *process* of writing code, does not suggest good ways of organizing the code (e.g., how to decompose the code into modules, classes, or functions), does not tell programmers which data structures and type signatures of functions will be useful to implement. Programmers make such design decisions purely on the basis of experience and intuition, trial-and-error, copy-paste, and guesswork.

In this context, program analysis and verification is analogous to writing mathematical equations describing the surface of a shoe made by a fashion designer. True, the “shoe surface equations” are mathematically rigorous and can be “analyzed” or “verified”; but the equations are written after the fact and do not guide the fashion designers in actually making shoes. It is understandable that fashion designers do not study the mathematical theory of surfaces.

¹⁹E.g., <https://www.amazon.com/Object-Oriented-Software-Engineering-Unified-Methodology/dp/0073376256>

²⁰<https://blog.cleancoder.com/uncle-bob/2016/03/19/GivingUpOnTDD.html>

²¹<https://youtu.be/8KpfVsJ5Jw4?t=447>

²²<https://www.amazon.com/Science-Programming-Monographs-Computer/dp/0387964800>

²³<https://www.amazon.com/Program-Derivation-Development-Specifications-International/dp/0201416247>

²⁴<https://www.cs.cmu.edu/~adamchik/15-121/lectures/Algorithmic%20Complexity/complexity.html>

²⁵https://en.wikipedia.org/wiki/Formal_verification

17.3.3 Programmers avoid academic terminology

Programmers jokingly grumble about terms such as “functor”, “monad”, or “lambda-functions”:

Those fancy words used by functional programmers purists really annoy me. Monads, functors...
Nonsense!!!²⁶

Perhaps only a small minority of software developers actually complain about this; the vast majority seems to remain unaware of “traversable functors” or “free monads”.

However, chemical engineers accept the need for studying differential equations and do not mind using the terms “phase diagram” or “Gibbs free energy”. Electrical engineers do not complain that the word “Fourier” is difficult to spell, or that “delta-function” is a weird thing to say. Mechanical engineers take it for granted that they need to calculate with “tensors” and “Lagrangians” and “non-holonomic constraints”. The arcane terminology seems to be the least of their difficulties, as their textbooks are full of complicated equations and long derivations.

Similarly, software engineers would not complain about the word “functor”, or about having to study the derivation of the algebraic laws for “monads,” — if they were true engineers. Textbooks on true software engineering would have been full of equations and derivations, in order to teach engineers how to perform certain calculations that are required *before* starting to write code.

17.4 Towards true engineering in software

It is now clear that we do not presently have true software engineering. The people employed under that job title are actually artisans. They work using artisanal methods, and their culture and processes are that of a crafts guild.

True software engineering means having a mathematical theory that guides the process of writing programs, — not theory that describes or analyzes programs after they are *somehow* written.

It is true that the numerical methods required for physics or the matrix calculations required for data science are “mathematical”. These programming tasks are indeed formulated using mathematical theory. However, mathematical *subject matter* (aerospace control, physics or astronomy simulations, or statistics) does not mean that engineering is used for the process of writing code. Data scientists, aerospace engineers, and physicists often write programs — but they almost always work as artisans when implementing their computations in program code.

We expect that software engineers’ textbooks should be full of equations and derivations. What theory would those equations represent?

This theory is what this book calls **applied functional type theory**. It is the algebraic foundation of the modern practice of functional programming, as implemented in languages such as OCaml, Haskell, and Scala. This theory is a blend of type theory, category theory, and logical proof theory, adapted for the needs of programmers. It has been in development since late 1990s and is still being actively worked on by a community of software practitioners and academic computer scientists.

To appreciate that functional programming, unlike any other programming paradigm, has a theory that guides coding, we can look at some recent software engineering conferences such as “Scala By the Bay”²⁷ or BayHac²⁸, or at the numerous FP-related online tutorials and blogs. We cannot fail to notice that speakers devote significant time to a peculiar kind of applied mathematical reasoning. Rather than focusing on one or another API or algorithm, as it is often the case with other software engineering blogs or presentations, an FP speaker describes a *mathematical structure* — such as the “applicative functor”²⁹ or the “free monad”³⁰ — and illustrates its use for practical coding.

²⁶<http://archive.is/65K3D>

²⁷<http://2015.scala.bythebay.io/>

²⁸<http://bayhac.org/>

²⁹<http://www.youtube.com/watch?v=bmIxIslimVY>

³⁰<http://www.youtube.com/watch?v=U01KOhnbc4U>

These people are not graduate students showing off their theoretical research; they are practitioners, software engineers who use FP on their jobs. It is just the nature of FP that certain mathematical tools — coming from formal logic and category theory — are now directly applicable to practical programming tasks.

These mathematical tools are not mere tricks for a specific programming language; they apply equally to all FP languages. Before starting to write code, the programmer can jot down certain calculations in a mathematical notation (see Fig. 17.1). The results of those calculations will help design the code fragment the programmer is about to write. This activity is similar to that of an engineer who performs some mathematical calculations before embarking on a design project.

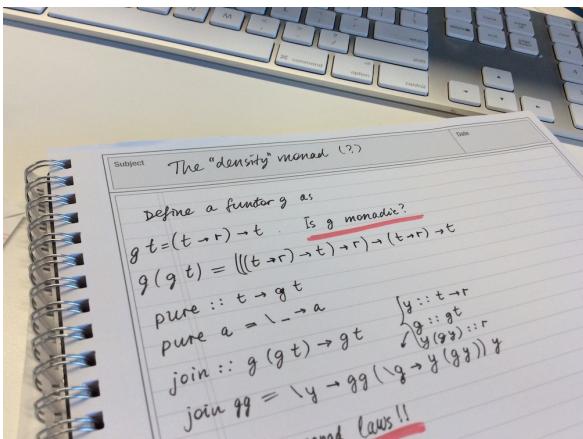


Figure 17.1: A programmer performs a derivation before writing Haskell code.

that has the necessary algebraic properties. Guided by the resulting type formula, they wrote the code that was guaranteed to work.

Another example of applied functional type theory is the “tagless final” encoding of effects, first described³⁴ in 2009. That technique (called “Church-encoded free monad” in the present book) has several advantages over the ordinary free monad and can improve upon it in a number of cases — just as the free monad itself was designed to cure certain problems with monad transformers³⁵. The new encoding is not tied to a specific programming language. Rather, it is a language-agnostic construction that was originally described in OCaml and later used in Haskell and Scala, but can be made to work even in Java³⁶, which is not an FP language.

This example shows that we may need several more years of work before the practical aspects of using applied functional type theory are sufficiently well understood by the FP community. The theory is in active development, and its design patterns — as well as the exact scope of the requisite theoretical material — are still being figured out. If the 40-year gap hypothesis³⁷ holds, we should expect applied functional type theory (perhaps under a different name) to become mainstream by 2030. This book is a step towards a clear designation of the scope of that theory.

³¹<https://arxiv.org/pdf/1403.0749.pdf>

³²<https://github.com/typelevel/cats/issues/983>

³³<https://elvishjerrico.github.io/2016/04/08/applicative-effects-in-free-monads.html>

³⁴<http://okmij.org/ftp/tagless-final/index.html>

³⁵<http://blog.ezyang.com/2013/09/if-youre-using-lift-youre-doing-it-wrong-probably/>

³⁶<http://archive.is/rLAh9>

³⁷<http://archive.is/rJc4A>

A recent example of a development in applied functional type theory is the “free applicative functor” construction. It was first described in a 2014 paper³¹; a couple of years later, a combined free applicative / free monad data type was designed and its implementation proposed³² as well as in Haskell³³. This technique allows programmers to implement declarative side-effect computations where some parts are sequential but other parts are computed in parallel, and to achieve the parallelism *automatically* while maintaining the composability of the resulting programs. The new technique has distinct advantages over using monad transformers, which was a previously used method of composing declarative side-effects. The combined “free applicative / free monad” was designed and implemented by true software engineers. They first derived the type constructor

17.5 Does software need engineers, or are artisans good enough?

The demand for programmers is growing. “Software developer” was #1 best job³⁸ in the US in 2018. But is there a demand for engineers or just for artisans?

We do not seem to be able to train enough software artisans.³⁹ So, it is probably impossible to train as many software engineers in the true sense of the word. Modern computer science courses do not actually train engineers in that sense; at best, they train academic researchers who write code as software artisans. Recalling the situation in construction business, with a few architects and hundreds of construction workers, we might also conclude that, perhaps, only a few software engineers are required per hundred software artisans.

What is the price of *not* having engineers, of replacing them with artisans?

Software practitioners have long bemoaned the mysterious difficulty of software development. Code “rots with time”, its complexity grows “out of control”, and operating systems have been notorious for ever-appearing security flaws⁴⁰ despite many thousands of programmers and testers employed. Clearly, we overestimated the capacity of the human brain for artisanal programming.

It is precisely in designing large and robust software systems that we would benefit from true engineering. Artisans has been building bridges and using chemical reactions by trial and error and via tradition, long before mechanical or chemical engineering disciplines were developed and founded upon rigorous theory. But once the theory became available, engineers were able to design unimaginably more powerful and complicated structures, devices, and processes. It is clear that trial, error, and adherence to tradition is inadequate for some of the software development tasks in front of us.

To build large and reliable software, such as new mobile or embedded operating systems or distributed peer-to-peer trust architectures, we will most likely need the qualitative increase in productivity and reliability that can only come from replacing artisanal programming by a true engineering discipline. Applied functional type theory and functional programming are steps in that direction.

³⁸<http://archive.is/cGJ2T>

³⁹<http://archive.is/137b8>

⁴⁰<http://archive.fo/HtQzw>

18 Essay: Towards functional data engineering with Scala

Data engineering is among the most in-demand¹ novel occupations in the IT world today. Data engineers create software pipelines that process large volumes of data efficiently. Why did the Scala programming language emerge as a premier tool² for crafting the foundational data engineering technologies such as Spark or Akka? Why is Scala in high demand³ within the world of big data?

There are reasons to believe that the choice of Scala was not accidental.

18.1 Data is math

Humanity has been working with data at least since Babylonian tax tables⁴ and the ancient Chinese number books⁵. Mathematics summarizes several millennia's worth of data processing experience in a few fundamental tenets:

- Data is *immutable*, because facts are immutable.
- Each *type* of values (population count, land area, distances, prices, times, etc.) needs to be handled separately; it is meaningless to add a distance to a population count.
- Data processing should be performed according to *mathematical formulas*.

Violating these tenets produces nonsense (see Fig. 18.1 for a real-life illustration).



Figure 18.1: Mixing incompatible data types produces nonsensical results.

The power of the principles of mathematics extends over all epochs and all cultures; math is the same in San Francisco, in Rio de Janeiro, in Kuala-Lumpur, and in Pyongyang (Fig. 18.2).

18.2 Functional programming is math

The functional programming paradigm is based on mathematical principles: values are immutable, data processing is coded through formula-like expressions, and each type of data is required to match correctly during the computations. The type-checking process automatically prevents programmers from making many kinds of coding errors. In addition, programming languages such as Scala and Haskell have a set of features adapted to building powerful abstractions and domain-specific languages. This power of abstraction is not accidental. Since mathematics is the ultimate art of building abstractions, math-based functional programming languages capitalize on the advantage of millennia of mathematical experience.

¹<http://archive.is/mK59h>

²<https://www.slideshare.net/noootsab/scala-the-unpredicted-lingua-franca-for-data-science>

³<https://techcrunch.com/2016/06/14/scala-is-the-new-golden-child/>

⁴<https://www.nytimes.com/2017/08/29/science/trigonometry-babylonian-tablet.html>

⁵<http://quatr.us/china/science/chinamath.htm>

A prominent example of how mathematics informs the design of programming languages is the connection between constructive logic⁶ and the programming language's type system, called the Curry-Howard (CH) correspondence⁷. The main idea of the CH correspondence is to think of programs as mathematical formulas that compute a value of a certain type A . The CH correspondence is between programs and logical propositions: To any program that computes a value of type A , there corresponds a proposition stating that "a value of type A can be computed".

This may sound rather theoretical so far. To see the real value of the CH correspondence, recall that formal logic has operations "*and*", "*or*", and "*implies*". For any two propositions A, B , we can construct the propositions " A *and* B ", " A *or* B ", " A *implies* B ". These three logical operations are foundational; without one of them, the logic is *incomplete* (you cannot derive some theorems).

A programming language **obeys the CH correspondence** with the logic if for any types A, B , the language also contains composite types corresponding to the logical formulas " A *or* B ", " A *and* B ", " A *implies* B ". In Scala, these composite types are `Either[A, B]`, the tuple `(A, B)`, and the function type, `A → B`. All modern functional languages such as OCaml, Haskell, Scala, F#, Swift, Elm, and PureScript support these three type constructions and thus are faithful to the CH correspondence. Having a *complete* logic in a language's type system enables declarative domain-driven code design⁸.

It is interesting to note that most older programming languages (C/C++, Java, JavaScript, Python) do not support some of these composite types. In other words, these programming languages have type systems based on an incomplete logic. As a result, users of these languages have to implement burdensome workarounds that make for error-prone code. Failure to follow mathematical principles has real costs (Figure 18.2).

18.3 The power of abstraction

Early adopters of Scala, such as Netflix, LinkedIn, and Twitter, were implementing what is now called "big data engineering". The required software needs to be highly concurrent, distributed, and resilient to failure. Those software companies used Scala as their main implementation language and reaped the benefits of functional programming.

What makes Scala suitable for big data tasks? The only reliable way of managing massively concurrent code is to use sufficiently high-level abstractions that make application code declarative. The two most important such abstractions are the "resilient distributed dataset" (RDD) of Apache Spark and the "reactive stream" used in systems such as Kafka, Akka Streams, and Apache Flink. While these abstractions are certainly implementable in Java or Python, a fully declarative and type-safe usage is possible only in a programming language with a sophisticated type system. Among the currently available mature functional languages, only Scala and Haskell are technically adequate for that task, due to their support for typeclasses and higher-order types. The early adopters of Scala were able to benefit from the powerful abstractions Scala supports. In this way, Scala enabled those businesses to engineer and to scale up their massively concurrent computations.

It remains to see why Scala and not, say, Haskell became the *lingua franca* of big data.



Figure 18.2: The Pyongyang method of error-free software engineering.

⁶https://en.wikipedia.org/wiki/Intuitionistic_logic

⁷https://en.wikipedia.org/wiki/Curry%2FHoward_correspondence

⁸<https://fsharpforfunandprofit.com/ddd/>

18.4 Scala is Java on math

The recently invented general-purpose functional programming languages can be grouped into “industrial” (F#, Scala, Swift) and “academic” (OCaml, Haskell).

The “academic” languages are clean-room implementations of well-researched mathematical principles of programming language design (the CH correspondence being one such principle). These languages are unencumbered by requirements of compatibility with any existing platform or libraries. Because of this, the “academic” languages are perfect playgrounds for taking various mathematical ideas to their logical conclusion. At the same time, software practitioners struggle to adopt these languages due to a steep learning curve, a lack of enterprise-grade libraries and tool support, and immature package management.

The languages from the “industrial” group are based on existing and mature software ecosystems: F# on .NET, Scala on JVM, and Swift on the MacOS/iOS platform. One of the important design requirements for these languages is 100% binary compatibility with their “parent” platforms and languages (F# with C#, Scala with Java, and Swift with Objective-C). Because of this, developers can immediately take advantage of the existing tooling, package management, and industry-strength libraries, while slowly ramping up the idiomatic usage of new language features. However, the same compatibility requirements dictated certain limitations in the languages, making their design less than fully satisfactory from the functional programming viewpoint.

It is now easy to see why the adoption rate of the “industrial” group of languages is much higher⁹ than that of the “academic” languages. The transition to the functional paradigm is also smoother for software developers because F#, Scala, and Swift seamlessly support the familiar object-oriented programming paradigm. At the same time, these new languages still have logically complete type systems, which gives developers an important benefit of type-safe domain modeling.

Nevertheless, the type systems of these languages are not equally powerful. For instance, F# and Swift are similar to OCaml in many ways but omit OCaml’s parameterized modules and some other features. Of all mentioned languages, only Scala and Haskell directly support typeclasses and higher-order types, which are helpful for expressing abstractions such as automatically parallelized data sets or asynchronous data streams.

To see the impact of these advanced features, consider LINQ, a domain-specific language for database queries on .NET, implemented in C# and F# through a special built-in syntax supported by Microsoft’s compilers. Analogous functionality is provided in Scala as a *library*, without need to modify the Scala compiler, by several open-source projects such as Slick, Squirly, or Quill. Similar libraries exist for Haskell — but not in languages with less powerful type systems.

18.5 Summary

The decisive advantages of Scala over other contenders (such as OCaml, Haskell, F#, or Swift) are:

1. Functional collections in the standard library.
2. A sophisticated type system with support for typeclasses and higher-order types.
3. Seamless compatibility with a mature software ecosystem (JVM).

Based on this assessment, we may be confident in Scala’s great future as a main implementation language for big data engineering.

⁹<https://www.tiobe.com/tiobe-index/>, archived in 2019 at <http://archive.is/RsNH8>

Part V

Appendices

A Notations

Certain notations and terms were chosen in this book differently from what the functional programming community currently uses. The proposed notation is well adapted to reasoning about types and code, and especially for designing data types and proving the laws of various typeclasses.

A.1 Summary of notations for types and code

F^A type constructor F with type argument A . In Scala, `F[A]`

$x:A$ value x has type A ; in Scala, `x:A`

$\mathbb{1}, 1$ the unit type and its value; in Scala, `Unit` and `()`

$\mathbb{0}$ the void type. In Scala, `Nothing`

$A + B$ a disjunctive type. In Scala, this type is `Either[A, B]`

$x:A + \mathbb{0}:B$ a value of a disjunctive type $A + B$. In Scala, `Left(x)`

$A \times B$ a product (tuple) type. In Scala, this type is `(A, B)`

$a:A \times b:B$ value of a tuple type $A \times B$. In Scala, `(a, b)`

$A \rightarrow B$ the function type, mapping from A to B

$x:A \rightarrow f$ a nameless function (as a value). In Scala, `{ x:A => f }`

id the identity function; in Scala, `identity[A]`

\triangleq “is defined to be” or “equal by definition”

$\stackrel{!}{=}$ “must be equal according to what we know”

$\stackrel{?}{=}$ “we ask — is it equal? — because we still need to prove that”

\cong for types, a natural isomorphism between types; for values, “equivalent” values according to an established isomorphism

A^{F^B} type annotation, used for defining unfunctors (GADTs)

\wedge logical conjunction; $\alpha \wedge \beta$ means “both α and β are true”

\vee logical disjunction; $\alpha \vee \beta$ means “either α or β or both are true”

\Rightarrow logical implication; $\alpha \Rightarrow \beta$ means “if α is true then β is true”

fmap $_F$ the standard method `fmap` of a functor F . In Scala, `Functor[F].fmap`

flatMap $_F$, **flatten** $_F$, **pure** $_F$ the standard methods `flatMap`, `flatten`, and `pure` of a monad F

F^\bullet the type constructor F understood as a type-level function. In Scala, `F[_]`

A Notations

$F^\bullet \rightsquigarrow G^\bullet$ or $F \rightsquigarrow G$ a natural transformation between functors F and G . In Scala, `F ~> G`

$\forall A.P^A$ a universally quantified type expression. In Scala 3, `[A] => P[A]`

$\exists A.P^A$ an existentially quantified type expression. In Scala, `{ type A; val x: P[A] }`

◦ the forward composition of functions: $f \circ g$ is $x \rightarrow g(f(x))$. In Scala, `f andThen g`

◦ the backward composition of functions: $f \circ g$ is $x \rightarrow f(g(x))$. In Scala, `f compose g`

◦ the backward composition of type constructors: $F \circ G$ is F^G

▷ use a value as the argument of a function: $x \triangleright f$ is $f(x)$. In Scala, `x.pipe(f)`

$f^{\uparrow G}$ a function f raised to a functor G ; same as `fmap_G f`

$f^{\uparrow G \uparrow H}$ a function raised first to G and then to H . In Scala, `h.map(_.map(f))`

$f^{\downarrow H}$ a function f raised to a contrafunctor

\diamond_M the Kleisli product operation for the monad M

\oplus the binary operation of a monoid. In Scala, `x |+| y`

Δ the “diagonal” function of type $\forall A. A \rightarrow A \times A$

π_1, π_2, \dots the projections from a tuple to its first, second, ..., parts

\boxtimes pair product of functions, $(f \boxtimes g)(a \times b) \triangleq f(a) \times g(b)$

$[a, b, c]$ an ordered sequence of values. In Scala, `Seq(a, b, c)`

$$\left| \begin{array}{cc} x \rightarrow x & \emptyset \\ \emptyset & a \rightarrow a \times a \end{array} \right|$$
 a function that works with disjunctive types (a “disjunctive function”)

A.2 Detailed explanations

F^A means a type constructor F with a type parameter A . In Scala, this is `f[A]`. Type constructors with multiple type parameters are denoted by $F^{A,B,C}$.

x^A means a value x that has type A ; this is a **type annotation**. In Scala, a type annotation is `x:A`. The colon symbol, `:`, in the superscript shows that A is not a type argument (as it would be in a type constructor, F^A). The notation $x : A$ can be used as well, but x^A is easier to read when x is inside a larger code expression.

$\mathbb{1}$ means the unit type, and 1 means the value of the unit type. In Scala, the unit type is `Unit`, and its value is `()`. Example of using the unit type is `1 + A`, which corresponds to `Option[A]` in Scala.

\emptyset means the void type (the type with no values). In Scala, this is the type `Nothing`. Example of using the void type is to denote the empty part of a disjunction. For example, in the disjunction $\mathbb{1} + A$ the non-empty part is $\emptyset + A$, which in Scala corresponds to `Some[A]`. The empty part $\mathbb{1} + \emptyset$ corresponds to `None`. Similarly, $A + \emptyset$ denotes the left part of the type $A + B$ (in Scala, `Left[A]`), while $\emptyset + B$ denotes its right part (in Scala, `Right[B]`). Values of disjunctive types are denoted similarly. For instance, $x^A + \emptyset^B$ denotes a value of the left part of the type $A + B$; in Scala, this value is written as `Left[A,B](x)`.

$A + B$ means the disjunctive type made from types A and B (or, a disjunction of A and B). In Scala, this is the type `Either[A, B]`.

$x^A + \emptyset^B$ denotes a value of a disjunctive type $A + B$, where x is the value of type A , which is the chosen case, and \emptyset stands for other possible cases. For example, $x^A + \emptyset^B$ is `Left[A,B](x)` in Scala. Type annotations A and B may be omitted if the types are unambiguous from the context.

$A \times B$ means the product type made from types A and B . In Scala, this is the tuple type `(A, B)`.

$a:A \times b:B$ means a value of a tuple type $A \times B$; in Scala, this is the tuple value `(a, b)`. Type annotations $:A$ and $:B$ may be omitted if the types are unambiguous from the context.

$A \rightarrow B$ means a function type from A to B . In Scala, this is the function type `A => B`.

$x:A \rightarrow y$ means a nameless function with argument x of type A and function body y . (Usually, the body y will be an expression that uses x . In Scala, this is `{ x: A => y }`. Type annotation $:A$ may be omitted if the type is unambiguous from the context.

id means the identity function. The type of its argument should be either specified as id^A or $\text{id}^{A \rightarrow A}$, or else should be unambiguous from the context. In Scala, `identity[A]` corresponds to id^A .

\triangleq means “equal by definition”. Examples:

- $f \triangleq (x:\text{Int} \rightarrow x + 10)$ is a definition of a function f . In Scala, this is `val f = { x: Int => x + 10 }`.
- $F^A \triangleq \mathbb{1} + A$ is a definition of a type constructor F . In Scala, this is `type F[A] = Option[A]`.

\cong for types means an equivalence (an isomorphism) of types. For example, $A + A \times B \cong A \times (\mathbb{1} + B)$. The same symbol \cong for *values* means “equivalent” according to an equivalence relation that needs to be established in the text. For example, if we have established an equivalence that allows nested tuples to be reordered whenever needed, we can write $(a \times b) \times c \cong a \times (b \times c)$, meaning that these values are mapped to each other by the established isomorphism functions.

A^{F^B} in type expressions means that the type constructor F^\bullet assigns the type F^B to the type expression A . This notation is used for defining unfunctors (GADTs). For example, the Scala code

```
sealed trait F[A]
case class F1() extends F[Int]
case class F2[A](a: A) extends F[(A, String)]
```

defines an unfunctor, which is denoted by $F^A \triangleq \mathbb{1}^{F^{\text{Int}}} + A^{F^{\text{A} \times \text{String}}}$.

\wedge (conjunction), \vee (disjunction), and \Rightarrow (implication) are used in formulas of Boolean as well as constructive logic in Chapter 5, e.g., $\alpha \wedge \beta$, where Greek letters stand for logical propositions.

fmap_F means the standard method `fmap` of the `Functor` typeclass, implemented for the functor F . In Scala, this may be written as `Functor[F].fmap`. Since each functor F has its own specific implementation of fmap_F , the subscript “ F ” is not a type parameter of fmap_F . The method fmap_F actually has *two* type parameters, which can be written out as $\text{fmap}_F^{A,B}$. Then the type signature of `fmap` is written in full as $\text{fmap}_F^{A,B} : (A \rightarrow B) \rightarrow F^A \rightarrow F^B$. For clarity, we may sometimes write out the type parameters A, B in the expression $\text{fmap}_F^{A,B}$, but in most cases these type parameters A, B can be omitted without loss of clarity.

pu_F denotes a monad F ’s method `pure`. This function has type signature $A \rightarrow F^A$ that contains a type parameter A . In the code notation, the type parameter may be either omitted or denoted as pu_F^A . If we are using the `pure` method with a complicated type, e.g., $\mathbb{1} + P^A$, instead of the type parameter A , we might want to write this type parameter for clarity and write $\text{pu}_F^{\mathbb{1} + P^A}$. The type signature of that function is then

$$\text{pu}_F^{\mathbb{1} + P^A} : \mathbb{1} + P^A \rightarrow F^{\mathbb{1} + P^A} .$$

But in most cases we will not need to write out the type parameters.

flm_F denotes a monad F ’s method `flatMap` with the type signature

$$\text{flm}_F : (A \rightarrow F^B) \rightarrow F^A \rightarrow F^B .$$

Note that Scala’s standard `flatMap` type signature is not curried. The curried method flm_F is easier to use in calculations involving the monad’s laws.

ftn_F denotes a monad F ’s method `flatten` with the type signature

$$\text{ftn}_F : F^{F^A} \rightarrow F^A .$$

F^\bullet means the type constructor F understood as a type-level function, — that is, with a type parameter unspecified. In Scala, this is `F[_]`. The bullet symbol, \bullet , is used as a placeholder for the missing type parameter. When no type parameter is needed, F means the same as F^\bullet . (For example, “a functor F ” and “a functor F^\bullet ” mean the same thing.) However, it is useful for clarity to be able to indicate the place where the type parameter would appear. For instance, functor composition is denoted as F^{G^\bullet} ; in Scala 2, this is `Lambda[x => F[G[x]]]` when using the “kind projector” plugin.¹ When the first type parameter of a bifunctor $P^{A,B}$ is fixed to Z , the resulting functor may be denoted by $P^{Z,\bullet}$. As another example, $T_L^{M,\bullet}$ denotes a monad transformer for the base monad L and the foreign monad M . The foreign monad M is a type parameter in $T_L^{M,\bullet}$, and so is the missing type parameter denoted by the placeholder symbol \bullet . (However, the base monad L is not a type parameter in $T_L^{M,\bullet}$ because the construction of the monad transformer depends sensitively on the internal details of L .)

$F^\bullet \sim G^\bullet$ or $F \sim G$ means a natural transformation between two functors F and G . In some Scala libraries, this is denoted by `F ~> G`.

$\forall A.P^A$ is a universally quantified type expression, in which A is a bound type parameter.

$\exists A.P^A$ is an existentially quantified type expression, in which A is a bound type parameter.

\circ means the forward composition of functions: $f \circ g$ (reads “ f before g ”) is the function defined as $x \rightarrow g(f(x))$.

\circ means the backward composition of functions: $f \circ g$ (reads “ f after g ”) is the function defined as $x \rightarrow f(g(x))$.

\circ with type constructors means their (backward) composition, for example $F \circ G$ denotes the type constructor F^{G^\bullet} . In Scala, this is `F[G[A]]`.

$x \triangleright f$ (the **pipe notation**) means that x is inserted as the argument into the function f . The expression $x \triangleright f$ means the same as $f(x)$. In Scala, the expression $x \triangleright f$ is written as `x.pipe(f)` or, if f is a method, `x.f`. This syntax is used with many standard methods such as `size` or `toSeq`. Because the function f is to the *right* of x in this notation, forward compositions of functions such as $x \triangleright f \triangleright g$ are naturally grouped to the left as it is done in Scala code, for example `x.toSeq.sorted`. The operation \triangleright (pronounced “pipe”) groups weaker than the forward composition (\circ), and so we have $x \triangleright f \circ g = x \triangleright f \triangleright g$ in this notation. Reasoning about code in the pipe notation uses the identities

$$\begin{aligned} x \triangleright f &= f(x), & (x \triangleright f) \triangleright g &= x \triangleright f \triangleright g & , \\ x \triangleright f \circ g &= x \triangleright (f \circ g), & x \triangleright f \triangleright g &= x \triangleright f \circ g & . \end{aligned}$$

The pipe symbol groups stronger than the function arrow, so $x \rightarrow y \triangleright f = x \rightarrow (y \triangleright f)$. Here are some examples of reasoning with functions in the pipe notation:

$$\begin{aligned} (a \rightarrow a \triangleright f) &= (a \rightarrow f(a)) = f & , \\ f \triangleright (y \rightarrow a \triangleright y) &= a \triangleright f = f(a) & , \\ f(y(x)) &= x \triangleright y \triangleright f \neq x \triangleright (y \triangleright f) = f(y)(x) & . \end{aligned}$$

The correspondence between the forward composition and the backward composition:

$$\begin{aligned} f \circ g &= g \circ f & , \\ x \triangleright (f \circ g) &= x \triangleright f \circ g = x \triangleright f \triangleright g = g(f(x)) = (g \circ f)(x) & . \end{aligned}$$

$f^{\uparrow G}$ means a function f lifted to a functor G . For a function $f^{A \rightarrow B}$, the application of $f^{\uparrow G}$ to a value g^{G^A} is written as $f^{\uparrow G}(g)$ or as $g \triangleright f^{\uparrow G}$. In Scala, this is `g.map(f)`. Nested lifting (i.e., lifting to the functor composition $H \circ G$) can be written as $f^{\uparrow G \uparrow H}$, which means $(f^{\uparrow G})^{\uparrow H}$ and produces a function of type $H^{G^A} \rightarrow H^{G^B}$. Applying a nested lifting to a value h of type H^{G^A} is written as $f^{\uparrow G \uparrow H} h$ or $h \triangleright f^{\uparrow G \uparrow H}$. In Scala, this is `h.map(_ map(f))`. The functor composition law is written as

$$p^{\uparrow G} \circ q^{\uparrow G} = (p \circ q)^{\uparrow G} \quad .$$

¹<https://github.com/typelevel/kind-projector>

Note the similarity between Scala code `x.map(p).map(q)` and the notation $x \triangleright p^{\uparrow G} \triangleright q^{\uparrow G}$.

$f^{\downarrow H}$ means a function f lifted to a contrafunctor H . For a function $f: A \rightarrow B$, the application of $f^{\downarrow H}$ to a value $h : H^B$ is written as $f^{\downarrow H}h$ or $h \triangleright f^{\downarrow H}$, and yields a value of type H^A . In Scala, this is `h.contramap(f)`. Nested lifting is denoted as $f^{\downarrow H \uparrow G} \triangleq (f^{\downarrow H})^{\uparrow G}$.

\diamond_M means the Kleisli product operation for a given monad M . This is a binary operation working on two Kleisli functions of types $A \rightarrow M^B$ and $B \rightarrow M^C$ and yields a new function of type $A \rightarrow M^C$.

\oplus means the binary operation of a monoid, for example $x \oplus y$. The specific monoid type should be defined for this expression to make sense. For example, in Scala the monoidal operation is usually denoted by `x |+| y`.

Δ means the “diagonal” function of type $\forall A. A \rightarrow A \times A$. There is only one implementation of this type signature,

```
def delta[A](a: A): (A, A) = (a, a)
```

π_1, π_2, \dots denote the functions extracting the first, second, ..., parts in a tuple. In Scala, π_1 is `_1`.

\boxtimes means the pair product of functions, where the result is a pair of the values of the two functions: $(f \boxtimes g)(a \times b) = f(a) \times g(b)$. In Scala, this operation can be defined by

```
def pair_product[A,B,P,Q](f: A => P, g: B => Q): ((A, B)) => (P, Q) = {
  case (a, b) => (f(a), g(b))
}
```

The operations Δ, π_i (where $i = 1, 2, \dots$), and \boxtimes allow us to express any function operating on tuples. Useful properties for reasoning about code of such functions:

identity law : $\Delta ; \pi_i = \text{id}$,

naturality law : $f ; \Delta = \Delta ; (f \boxtimes f)$,

left projection law : $(f \boxtimes g) ; \pi_1 = \pi_1 ; f$,

right projection law : $(f \boxtimes g) ; \pi_2 = \pi_2 ; g$,

composition law : $(f \boxtimes g) ; (p \boxtimes q) = (f ; p) \boxtimes (g ; q)$,

as well as the functor lifting laws for Δ and π_i :

$$\begin{aligned} f^{\uparrow F} ; \Delta &= \Delta ; f^{\uparrow (F \times F)} = \Delta ; (f^{\uparrow F} \boxtimes f^{\uparrow F}) , \\ (f^{\uparrow F} \boxtimes f^{\uparrow G}) ; \pi_1 &= f^{\uparrow (F \times G)} ; \pi_1 = \pi_1 ; f^{\uparrow F} . \end{aligned}$$

`[a, b, c]` means an ordered sequence of values, such as a list or an array. In Scala, this can be `List(a, b, c)`, `Vector(a, b, c)`, `Array(a, b, c)`, or another collection type.

$f: Z+A \rightarrow Z+A \times A \triangleq \begin{array}{|c|c|} \hline z \rightarrow z & 0 \\ \hline 0 & a \rightarrow a \times a \\ \hline \end{array}$ is the **matrix notation** for a function whose input and/or output type is a disjunctive type (a **disjunctive function**). In Scala, the function f is implemented as

```
def f[Z, A]: Either[Z, A] => Either[Z, (A, A)] = {
  case Left(z)  => Left(z) // Identity function on Z.
  case Right(a) => Right((a, a)) // Delta on A.
}
```

The rows of the matrix indicate the different `cases` in the function’s code, corresponding to the different parts of the input disjunctive type. If the input type is not disjunctive, there will be only one row. The columns of the matrix indicate the parts of the output disjunctive type. If the the output type is not disjunctive, there will be only one column.

A matrix may show all parts of the disjunctive types in separate “type row” and “type column”:

$$f: Z+A \Rightarrow Z+A \times A \triangleq \begin{array}{|c|c|c|} \hline & Z & A \times A \\ \hline Z & \text{id} & 0 \\ \hline A & 0 & a \rightarrow a \times a \\ \hline \end{array} . \quad (\text{A.1})$$

This notation clearly indicates the input and the output types of the function and is useful at some stages of reasoning about the code. The vertical double line separates input types from the function code. In the code above, the “type column” shows the parts of the input disjunctive type $Z + A$. The “type row” shows the parts of the output disjunctive type $Z + A \times A$.

The matrix notation is adapted to *forward* function composition $f \circ g$. Assume that A is a monoid type, and consider the composition of the function f shown above and the function g defined as

```
def g[Z, A: Monoid]: Either[Z, (A, A)] => A = {
  case Left(_) => Monoid[A].empty
  case Right((a1, a2)) => a1 |+| a2
}
```

In the matrix notation, the function g is written (with and without types) as

$$g \triangleq \left| \begin{array}{c|c} & A \\ Z & _ \rightarrow e^A \\ A \times A & a_1 \times a_2 \rightarrow a_1 \oplus a_2 \end{array} \right|, \quad g \triangleq \left| \begin{array}{c} _ \rightarrow e^A \\ a_1 \times a_2 \rightarrow a_1 \oplus a_2 \end{array} \right|.$$

The forward composition $f \circ g$ is computed by forward-composing the matrix elements using the rules of the ordinary matrix multiplication, omitting any terms containing $\mathbb{0}$:

$$\begin{aligned} f \circ g &= \left| \begin{array}{cc} \text{id} & \mathbb{0} \\ \mathbb{0} & a \rightarrow a \times a \end{array} \right| \left| \begin{array}{c} _ \rightarrow e^A \\ a_1 \times a_2 \rightarrow a_1 \oplus a_2 \end{array} \right| \\ &= \left| \begin{array}{c} \text{id} \circ (_ \rightarrow e^A) \\ (a \rightarrow a \times a) \circ (a_1 \times a_2 \rightarrow a_1 \oplus a_2) \end{array} \right| = \left| \begin{array}{c} _ \rightarrow e^A \\ a \rightarrow a \oplus a \end{array} \right|. \end{aligned}$$

Applying a function to a disjunctive value such as x^{Z+A} is computed by writing x as a row vector:

$$x = z^Z + \mathbb{0}^A = \left| \begin{array}{c} z^Z \\ \mathbb{0} \end{array} \right|,$$

and the computation $x \triangleright f \circ g$ again follows the rules of matrix multiplication:

$$x \triangleright f \circ g = \left| \begin{array}{c} z^Z \\ \mathbb{0} \end{array} \right| \triangleright \left| \begin{array}{c} _ \rightarrow e^A \\ a \rightarrow a \oplus a \end{array} \right| = z \triangleright (_ \rightarrow e) = e.$$

Since the standard rules of matrix multiplication are associative, the properties of the \triangleright -notation such as $x \triangleright (f \circ g) = (x \triangleright f) \triangleright g$ are guaranteed to hold.

To use the matrix notation with *backward* compositions $(f \circ g)$, all function matrices need to be transposed. (A standard identity of matrix calculus is that the transposition reverses the order of composition, $(AB)^T = B^T A^T$.) The argument types will then appear in the top row and the result types in the left column; the double line is above the matrix since that is where the function inputs come from. The above calculations are then rewritten as

$$\begin{aligned} g \circ f &= \left| \begin{array}{c|c} & Z \\ & A \times A \\ \hline Z & _ \rightarrow e^A \\ A & a_1 \times a_2 \rightarrow a_1 \oplus a_2 \end{array} \right| \circ \left| \begin{array}{c|c} Z & A \\ \hline Z & \text{id} \\ A \times A & \mathbb{0} \\ \hline \mathbb{0} & a \rightarrow a \times a \end{array} \right| \\ &= \left| \begin{array}{c} \text{id} \circ (_ \rightarrow e^A) \\ (a \rightarrow a \times a) \circ (a_1 \times a_2 \rightarrow a_1 \oplus a_2) \end{array} \right| = \left| \begin{array}{c} _ \rightarrow e^A \\ a \rightarrow a \oplus a \end{array} \right|. \\ (g \circ f)(x) &= \left| \begin{array}{c} _ \rightarrow e^A \\ a \rightarrow a \oplus a \end{array} \right| \left| \begin{array}{c} z^Z \\ \mathbb{0} \end{array} \right| = (_ \rightarrow e^A)(z) = e. \end{aligned}$$

The *forward* composition seems to be easier to read and to reason about in the matrix notation.

B Glossary of terms

Code notation A mathematical notation developed in this book for deriving properties of code in functional programs. Variables have optional type annotations, such as $x:A$ or $f:A \rightarrow B$. Nameless functions are denoted by $x:A \rightarrow f$, products by $a \times b$, and values of a disjunctive type $A + B$ are written as $x:A + 0:B$ or $0:A + y:B$. Functions working with disjunctive types are denoted by matrices. Lifting of functions to functors, such as $\text{fmap}_L(f)$, is denoted by $f^{\uparrow L}$; function compositions are denoted by $f \circ g$ (forward composition) and $f \circ g$ (backward composition); and function applications by $f(x)$ or equivalently $x \triangleright f$. See Appendix A for details.

Contrafunctor A type constructor having the properties of a contravariant functor with respect to a type parameter. Instead of “contravariant functor”, this book uses the shorter name “contrafunctor”.

Disjunctive type A type representing one of several distinct possibilities. In Scala, this is usually implemented as a sealed trait extended by several case classes. The standard Scala disjunction types are `Option[A]` and `Either[A, B]`. Also known as **sum type**, **tagged union type**, **co-product type**, and **variant type** (in Object Pascal and in OCaml). The shortest name is “sum type,” but the English word “disjunctive” is less ambiguous to the ear than “sum”.

Exponential-polynomial type A type constructor built using products, disjunctions (sums or co-products), and function types (“exponentials”), as well as type parameters and fixed types. For example, `type F[A] = Either[(A, A), Int => A]` is an exponential-polynomial type constructor. Such type constructors are always profunctors and can also be functors or contrafunctors.

Functor block A short syntax for composing several `map`, `flatMap`, and `filter` operations applied to a functor-typed value. The type constructor corresponding to that value must be a functor and is fixed throughout the entire functor block. For example, the Scala code

```
for { x <- List(1,2,3); y <- List(10, x); if y > 2 }
  yield 2 * y
```

is equivalent to the code

```
List(1, 2, 3).flatMap(x => List(10, x))
  .filter(y => y > 1).map(y => 2 * y)
```

and computes the value `List(20, 20, 20, 6)`. This is a functor block that “raises” computations to the `List` functor. Similar syntax exists in a number of languages and is called a “**for-comprehension**” or a “list comprehension” in Python, “**do-notation**” in Haskell, and “**computation expressions**” in F#. I use the name “functor block” in this book because it is shorter and more descriptive. (The type constructor used in a functor block needs to be at least a functor but does not have to be a monad.)

Kleisli function A function with type signature $A \rightarrow M^B$ for some fixed monad M . Also called a Kleisli morphism (a morphism in the Kleisli category corresponding to the monad M). The standard monadic method $\text{pure}_M : A \rightarrow M^A$ has the type signature of a Kleisli function. The Kleisli composition operation, \diamond_M , is a binary operation that combines two Kleisli functions (of types $A \rightarrow M^B$ and $B \rightarrow M^C$) into a new Kleisli function (of type $A \rightarrow M^C$).

Method This word is used in two ways: 1) A `method_1` is a Scala function defined as a member of a typeclass. For example, `flatMap` is a method defined in the `Monad` typeclass. 2) A `method_2` is a

Scala function defined as a member of a data type declared as a Java-compatible `class` or `trait`. Trait methods₂ are necessary in Scala when implementing functions whose arguments have type parameters (because Scala function values defined via `val` cannot have type parameters). So, many typeclasses such as `Functor` or `Monad`, whose methods₁ require type parameters, will use Scala `traits` with methods₂ for their implementation. The same applies to type constructions with quantified types, such as the Church encoding.

Nameless function An expression of function type, representing a function. For example, `(x: Int) => x * 2`. Also known as function expression, function literal, anonymous function, closure, lambda-function, lambda-expression, or simply a “lambda”.

Partial type-to-value function (PTVF) A function with a type parameter but defined only for a certain subset of types. In Scala, PTVFs are implemented via a typeclass constraint:

```
def double[T: Semigroup](t: T): T = implicitly[Semigroup[T]].combine(t, t)
```

This PTVF is defined only for types τ for which a `Semigroup` typeclass instance is available.

Polynomial functor A type constructor built using disjunctions (sums), products (tuples), type parameters and fixed types. For example, in Scala, `type F[A] = Either[Int, A]` is a polynomial functor with respect to the type parameter A , while `Int` is a fixed type (not a type parameter). Polynomial functors are also known as **algebraic data types**.

Product type A type representing several values given at once. In Scala, product types are the tuple types, for example `(Int, String)`, and case classes. Also known as **tuple** type, **struct** (in C and C++), and **record**.

Profunctor A type constructor whose type parameter occurs in both covariant and contravariant positions and satisfying the appropriate laws; see Section 6.4.1.

Type notation A mathematical notation for type expressions developed in this book for easier reasoning about types in functional programs. Disjunctive types are denoted by $+$, product types by \times , and function types by \rightarrow . The unit type is denoted by $\mathbb{1}$, and the void type by $\mathbb{0}$. The function arrow \rightarrow groups weaker than $+$, which in turn groups weaker than \times . This means

$$Z + A \rightarrow Z + A \times A \quad \text{is the same as} \quad (Z + A) \rightarrow (Z + (A \times A)) \quad .$$

Type parameters are denoted by superscripts. As an example, the Scala definition

```
type F[A] = Either[(A, A => Option[Int]), String => List[A]]
```

is written in the type notation as

$$F^A \triangleq A \times (A \rightarrow \mathbb{1} + \text{Int}) + (\text{String} \rightarrow \text{List}^A) \quad .$$

Unfunctor A type constructor that cannot possibly be a functor, nor a contrafunctor, nor a profunctor. An example is a type constructor with explicitly indexed type parameters, such as $F^A \triangleq (A \times A)^{F^{\text{Int}}} + (\text{Int} \times A)^{F^{\mathbb{1}}}$. The Scala code for this type constructor is

```
sealed trait F[A]
final case class F1[A](x: A, y: A) extends F[Int]
final case class F2[A](s: Int, t: A) extends F[Unit]
```

This can be seen as a **GADT** (generalized algebraic data type) that uses specific values of type parameters instead of the type parameter A in at least some of its case classes.

B.1 On the current misuse of the term “algebra”

This book avoids using the terms “algebra” or “algebraic” because these terms are too ambiguous. The functional programming community uses the word “algebra” in at least *four* incompatible ways.

Definition 0. In mathematics, an “algebra” is a vector space with multiplication and certain standard properties. For example, we need $1 * x = x$, the addition must be commutative, the multiplication must be distributive over addition, etc. The set of all 10×10 matrices with real coefficients is an “algebra” in this sense. These matrices form a 100-dimensional vector space, and they can be multiplied and added. This definition of “algebra” is not used in functional programming.

Definition 1. For a given functor F , an “ F -algebra” is a type A and a function with type signature $F^A \rightarrow A$. (This definition comes from category theory.) The name “ F -algebra” emphasizes the dependence on a chosen functor F . There is no direct connection between this “algebra” and Definition 0, except when the functor F is defined by $F^A \triangleq A \times A$: a function of type $A \times A \rightarrow A$ may then be interpreted as a “multiplication” operation. However, A is a type and not a vector space, and there are no distributivity or commutativity laws. This book does not use F -algebras because knowing their category-theoretical properties does not help in practical programming.

Definition 2. Polynomial functors are often called “algebraic data types”. However, they are not “algebraic” in the sense of Definitions 0 or 1. For example, `Either[Option[A], Int]` is considered an “algebraic data type”, denoted by $F^A \triangleq 1 + A + \text{Int}$ in the type notation. However, values of the type F^A do not support the addition and multiplication operations required by Definition 0 of “algebra”. The type F^A may admit some binary or unary operations (e.g., that of a monoid), but these operations will likely not be commutative or distributive. Also, there is not necessarily a function with type $F^A \rightarrow A$, as required for Definition 1. Rather, the word “algebra” refers to “school-level algebra” dealing with polynomials, since these data types are built from sums and products of types. If the data contains a function type, e.g., `Option[Int => A]`, the type is no longer polynomial. This book uses more precise terms “polynomial types” and “exponential-polynomial types”.

Definition 3. People talk about the “algebra” of properties of functions such as `map` or `flatMap`, meaning that these functions satisfy certain “algebraic” laws (e.g., the composition, naturality, or associativity laws). But these laws do not make functions `map` or `flatMap` into an algebra in the sense of Definition 0 or in the sense of Definition 1. There is also no relation to the algebraic data types of Definition 2. So, this is a different usage of the word “algebra”. However, there is no general “algebra of laws” that we can use; every derivation proceeds in a different way, specific to the laws being proved. In mathematics, “algebraic” equations are distinguished from differential or integral equations. In that sense, the laws used in functional programming are *always* algebraic: they are just equations with compositions and applications of functions. So, it is not useful to call laws “algebraic” in any of these two senses. This book calls them “equations” or just “laws”.

Definition 4. The Church encoding of a free monad (also known as the “tagless final interpreter”) is the type $\forall E^*. (S^{E^*} \rightsquigarrow E^*) \rightarrow E^A$. It uses a higher-order type constructor S^{E^*} parameterized by a *type constructor* parameter E^* . In this context, one calls an “algebra” a value of type $S^{E^*} \rightsquigarrow E^*$, which is an S -algebra in the category of “type constructors” (functors in a category without any non-identity morphisms). So, Definition 4 is related to Definition 1, with a specific choice of a functor. However, knowing that $S^{E^*} \rightsquigarrow E^*$ is an S -algebra in the category of type constructors does not provide any help or additional insights for practical work with the Church encoding of a free monad.

The higher-order type constructor S is used to parameterize the effects described by a Church-encoded free monad, so this book calls it the “effect constructor”.

So, it appears that the current usage of the word “algebra” in functional programming is both inconsistent and unhelpful to practitioners. In this book, the word “algebra” always means a branch of mathematics, as in “high-school algebra”. Instead of “algebra” as in Definitions 1 to 4, this book talks about “polynomial types” or “polynomial functors” or “exponential-polynomial functors” etc.; “equations” or “laws”; and an “effect constructor” S .

C Inferring code from types with the LJT algorithm

The [Gentzen-Vorobieff-Hudelmaier algorithm](#) and its generalizations

See also the [curryhoward](#) project

C.1 Slides

In formal logic, this statement is written in the syntax

$$X, Y, \dots, Z \vdash T$$

and is called a **sequent** having the premises X, Y, \dots, Z and the goal T .

A sequent in formal logic can be proved if proof task;

The elementary proof task is represented by a **sequent**

Notation: $A, B, C \vdash G$; the **premises** are A, B, C and the **goal** is G

Proofs are achieved via axioms and derivation rules

Axioms: such and such sequents are already true

Derivation rules: this sequent is true if such and such sequents are true

To make connection with logic, represent code fragments as **sequents**

$A, B \vdash C$ represents an *expression* of type C that uses $x: A$ and $y: B$

Examples in Scala:

`(x: Int).toString + "abc"` is an expression of type `String` that uses an `x: Int` and is represented by the sequent `Int ⊢ String`

`(x: Int) → x.toString + "abc"` is an expression of type `Int → String` and is represented by the sequent $\emptyset \vdash \text{Int} \rightarrow \text{String}$

Sequents only describe the *types* of expressions and their parts

Translating language constructions into the logic II

What are the derivation rules for the logic of types?

Write all the constructions in FP languages as sequents

This will give all the derivation rules for the logic of types

Each type construction has an expression for creating it and an expression for using it

Tuple type $A \times B$

Create: $A, B \vdash A \times B$

Use: $A \times B \vdash A$ and also $A \times B \vdash B$

Function type $A \rightarrow B$

Create: if we have $A \vdash B$ then we will have $\emptyset \vdash A \rightarrow B$

Use: $A \rightarrow B, A \vdash B$

Disjunction type $A + B$

Create: $A \vdash A + B$ and also $B \vdash A + B$

Use: $A + B, A \rightarrow C, B \rightarrow C \vdash C$

Unit type 1

Create: $\emptyset \vdash 1$

Translating language constructions into the logic III

Additional rules for the logic of types

In addition to constructions that use types, we have “trivial” constructions:

a single, unmodified value of type A is a valid expression of type A

For any A we have the sequent $A \vdash A$

if a value can be computed using some given data, it can also be computed if given *additional* data

If we have $A, \dots, C \vdash G$ then also $A, \dots, C, D \vdash G$ for any D

For brevity, we denote by Γ a sequence of arbitrary premises

the order in which data is given does not matter, we can still compute all the same things given the same premises in different order

If we have $\Gamma, A, B \vdash G$ then we also have $\Gamma, B, A \vdash G$

Syntax conventions:

the implication operation associates *to the right*

$A \rightarrow B \rightarrow C$ means $A \rightarrow (B \rightarrow C)$

precedence order: implication, disjunction, conjunction

$A + B \times C \rightarrow D$ means $(A + (B \times C)) \rightarrow D$

Quantifiers: implicitly, all our type variables are universally quantified

When we write $A \rightarrow B \rightarrow A$, we mean $\forall A : \forall B : A \rightarrow B \rightarrow A$

The logic of types I

Now we have all the axioms and the derivation rules of the logic of types.

What theorems can we derive in this logic?

Example: $A \rightarrow B \rightarrow A$

Start with an axiom $A \vdash A$; add an unused extra premise B : $A, B \vdash A$

Use the “create function” rule with B and A , get $A \vdash B \rightarrow A$

Use the “create function” rule with A and $B \rightarrow A$, get the final sequent $\emptyset \vdash A \rightarrow B \rightarrow A$ showing that $A \rightarrow B \rightarrow A$ is a **theorem** since it is derived from no premises

What code does this describe?

The axiom $A \vdash A$ represents the expression x^A where x is of type A

The unused premise B corresponds to unused variable y^B of type B

The “create function” rule gives the function $y^B \rightarrow x^A$

The second “create function” rule gives $x^A \rightarrow (y^B \rightarrow x)$

Scala code: `def f[A, B]: A → B → A = (x: A) → (y: B) → x`

Any code expression’s type can be translated into a sequent

A proof of a theorem directly guides us in writing code for that type

Correspondence between programs and proofs

By construction, any theorem can be implemented in code

Proposition	Code
$\forall A : A \rightarrow A$	<code>def identity[A](x: A): A = x</code>
$\forall A : A \rightarrow 1$	<code>def toUnit[A](x: A): Unit = ()</code>
$\forall A \forall B : A \rightarrow A + B$	<code>def inLeft[A,B](x:A): Either[A,B] = Left(x)</code>
$\forall A \forall B : A \times B \rightarrow A$	<code>def first[A,B](p: (A,B)): A = p._1</code>
$\forall A \forall B : A \rightarrow B \rightarrow A$	<code>def const[A,B](x: A): B → A = (y:B) → x</code>

Also, non-theorems *cannot be implemented* in code

Examples of non-theorems:

$\forall A : 1 \rightarrow A$; $\forall A \forall B : A + B \rightarrow A$;

$\forall A \forall B : A \rightarrow A \times B$; $\forall A \forall B : (A \rightarrow B) \rightarrow A$

Given a type’s formula, can we implement it in code? Not obvious.

Example: $\forall A \forall B : (((A \rightarrow B) \rightarrow A) \rightarrow A) \rightarrow B$

Can we write a function with this type? Can we prove this formula?

The logic of types II

What kind of logic is this? What do mathematicians call this logic?

This is called “intuitionistic propositional logic”, IPL (also “constructive”)

This is a “nonclassical” logic because it is different from Boolean logic

Disjunction works very differently from Boolean logic

Example: $A \rightarrow B + C \vdash (A \rightarrow B) + (A \rightarrow C)$ does not hold in IPL

This is counter-intuitive!

We cannot implement a function with this type:

```
def q[A,B,C](f: A → Either[B, C]): Either[A → B, A → C]
```

Disjunction is “constructive”: need to supply one of the parts

But `Either[A → B, A → C]` is not a function of `A`

Implication works somewhat differently

Example: $((A \rightarrow B) \rightarrow A) \rightarrow A$ holds in Boolean logic but not in IPL

Cannot compute an `x: A` because of insufficient data

Conjunction works the same as in Boolean logic

Example:

$$A \rightarrow B \times C \vdash (A \rightarrow B) \times (A \rightarrow C)$$

The logic of types III

How to determine whether a given IPL formula is a theorem?

The IPL cannot have a truth table with a fixed number of truth values

This was shown by Gödel in 1932 (see [Wikipedia page](#))

The IPL has a decision procedure (algorithm) that either finds a proof for a given IPL formula, or determines that there is no proof

There may be several inequivalent proofs of an IPL theorem

Each proof can be *automatically translated* into code

The `curryhoward` library implements an IPL prover as a Scala macro, and generates Scala code from types

The `djinn-ghc` compiler plugin and the `JustDoIt plugin` implement an IPL prover in Haskell, and generate Haskell code from types

All these IPL provers use the same basic algorithm called LJT

and all cite the same paper [\[Dyckhoff 1992\]](#)

because most other papers on this subject are incomprehensible to non-specialists, or describe algorithms that are too complicated

Proof search I: looking for an algorithm

Why our initial presentation of IPL does not give a proof search algorithm

The FP type constructions give nine axioms and three derivation rules:

- $\Gamma, A, B \vdash A \times B$

$$\frac{\Gamma, A \vdash B}{\Gamma \vdash A \rightarrow B}$$

- $\Gamma, A \times B \vdash A$

$$\frac{\Gamma \vdash G}{\Gamma, D \vdash G}$$

- $\Gamma, A \times B \vdash B$

$$\frac{\Gamma, A, B \vdash G}{\Gamma, B, A \vdash G}$$

- $\Gamma, A \rightarrow B, A \vdash B$

- $\Gamma, A \vdash A + B$

$$\frac{\Gamma \vdash 1}{\Gamma, A \vdash A}$$

- $\Gamma, B \vdash A + B$

- $\Gamma, A + B, A \rightarrow C, B \rightarrow C \vdash C$

- $\Gamma \vdash 1$

- $\Gamma, A \vdash A$

Can we use these rules to obtain a finite and complete search tree? No.

Try proving $A, B + C \vdash A \times B + C$: cannot find matching rules

Need a better formulation of the logic

Proof search II: Gentzen's calculus LJ (1935)

A “complete and sound calculus” is a set of axioms and derivation rules that will yield all (and only!) theorems of the logic

$$\begin{array}{c}
 (X \text{ is atomic}) \frac{}{\Gamma, \textcolor{blue}{X} \vdash X} Id \\
 \frac{\Gamma, A \rightarrow B \vdash A \quad \Gamma, B \vdash C}{\Gamma, \textcolor{blue}{A} \rightarrow B \vdash C} L \rightarrow \\
 \frac{\Gamma, A \vdash C \quad \Gamma, B \vdash C}{\Gamma, \textcolor{blue}{A} + B \vdash C} L+ \\
 \frac{\Gamma, A_i \vdash C}{\Gamma, \textcolor{blue}{A}_1 \times \textcolor{blue}{A}_2 \vdash C} L \times_i \\
 \frac{\Gamma, A \vdash B \quad \Gamma, B \vdash A}{\Gamma \vdash A \rightarrow B} R \rightarrow \\
 \frac{\Gamma \vdash A_i}{\Gamma \vdash \textcolor{blue}{A}_1 + \textcolor{blue}{A}_2} R+_i \\
 \frac{\Gamma \vdash A \quad \Gamma \vdash B}{\Gamma \vdash A \times B} R \times
 \end{array}$$

Two axioms and eight derivation rules

Each derivation rule says: The sequent at bottom will be proved if proofs are given for sequent(s) at top

Use these rules “bottom-up” to perform a proof search

Sequents are nodes and proofs are edges in the proof search tree

Proof search example I

Example: to prove $((R \rightarrow R) \rightarrow Q) \rightarrow Q$

Root sequent $S_0 : \emptyset \vdash ((R \rightarrow R) \rightarrow Q) \rightarrow Q$

S_0 with rule $R \rightarrow$ yields $S_1 : (R \rightarrow R) \rightarrow Q \vdash Q$

S_1 with rule $L \rightarrow$ yields $S_2 : (R \rightarrow R) \rightarrow Q \vdash R \rightarrow R$ and $S_3 : Q \vdash Q$

Sequent S_3 follows from the Id axiom; it remains to prove S_2

S_2 with rule $L \rightarrow$ yields $S_4 : (R \rightarrow R) \rightarrow Q \vdash R \rightarrow R$ and $S_5 : Q \vdash R \rightarrow R$

We are stuck here because $S_4 = S_2$ (we are in a loop)

We can prove S_5 , but that will not help

So we backtrack (erase S_4, S_5) and apply another rule to S_2

S_2 with rule $R \rightarrow$ yields $S_6 : (R \rightarrow R) \rightarrow Q; R \vdash R$

Sequent S_6 follows from the Id axiom

Therefore we have proved S_0

Since $((R \rightarrow R) \rightarrow Q) \rightarrow Q$ is derived from no premises, it is a theorem

Q.E.D.

Proof search III: The calculus LJT

Vorobieff-Hudelmaier-Dyckhoff, 1950-1990

The Gentzen calculus LJ will loop if rule $L \rightarrow$ is applied ≥ 2 times

The calculus LJT keeps all rules of LJ except rule $L \rightarrow$

Replace rule $L \rightarrow$ by pattern-matching on A in the premise $A \rightarrow B$:

$$\begin{array}{c}
 (X \text{ is atomic}) \frac{\Gamma, X, B \vdash D}{\Gamma, \textcolor{blue}{X} \rightarrow B \vdash D} L \rightarrow_1 \\
 \frac{\Gamma, A \rightarrow B \rightarrow C \vdash D}{\Gamma, \textcolor{blue}{(A \times B)} \rightarrow \textcolor{blue}{C} \vdash D} L \rightarrow_2 \\
 \frac{\Gamma, A \rightarrow C, B \rightarrow C \vdash D}{\Gamma, \textcolor{blue}{(A + B)} \rightarrow \textcolor{blue}{C} \vdash D} L \rightarrow_3 \\
 \frac{\Gamma, B \rightarrow C \vdash A \rightarrow B \quad \Gamma, C \vdash D}{\Gamma, \textcolor{blue}{(A \rightarrow B)} \rightarrow \textcolor{blue}{C} \vdash D} L \rightarrow_4
 \end{array}$$

When using LJT rules, the proof tree has no loops and terminates

See [this paper](#) for an explicit decreasing measure on the proof tree

Proof search IV: The calculus LJT

“It is obvious that it is obvious” — a mathematician after thinking for a half-hour

Rule $L \rightarrow_4$ is based on the key theorem:

$$((A \rightarrow B) \rightarrow C) \rightarrow (A \rightarrow B) \iff (B \rightarrow C) \rightarrow (A \rightarrow B)$$

The key theorem for rule $L \rightarrow_4$ is attributed to Vorobieff (1958)

A stepping stone to this theorem:

$$((A \rightarrow B) \rightarrow C) \rightarrow B \rightarrow C$$

Proof: $f^{(A \rightarrow B) \rightarrow C} \rightarrow b^B \rightarrow f(x^A \rightarrow b)$

Proof search V: From deduction rules to code

The new rules are equivalent to the old rules, therefore...

Proof of a sequent $A, B, C \vdash G \Leftrightarrow$ code/expression $t(a, b, c) : G$

Also can be seen as a function t from A, B, C to G

Sequent in a proof follows from an axiom or from a transforming rule

The two axioms are fixed expressions, $x^A \rightarrow x$ and 1

Each rule has a *proof transformer* function: $PT_{R \rightarrow}$, PT_{L+} , etc.

Examples of proof transformer functions:

$$\frac{\Gamma, A \vdash C \quad \Gamma, B \vdash C}{\Gamma, A + B \vdash C} L+$$

$$PT_{L+}(t_1^{A \rightarrow C}, t_2^{B \rightarrow C}) = x^{A+B} \rightarrow x \text{ match } \begin{cases} a^A \rightarrow t_1(a) \\ b^B \rightarrow t_2(b) \end{cases}$$

$$\frac{\Gamma, A \rightarrow B \rightarrow C \vdash D}{\Gamma, (A \times B) \rightarrow C \vdash D} L \rightarrow_2$$

$$PT_{L \rightarrow_2}(f^{(A \rightarrow B \rightarrow C) \rightarrow D}) = g^{A \times B \rightarrow C} \rightarrow f(x^A \rightarrow y^B \rightarrow g(x, y))$$

Verify that we can indeed produce PTs for every rule of LJT

Proof search example II: deriving code

Once a proof tree is found, start from leaves and apply PTs

For each sequent S_i , this will derive a **proof expression** t_i

Example: to prove S_0 , start from S_6 backwards:

$$\begin{aligned} S_6 : (R \rightarrow R) \rightarrow Q; R \vdash R & \quad (\text{axiom } Id) \quad t_6(rrq, r) = r \\ S_2 : (R \rightarrow R) \rightarrow Q \vdash (R \rightarrow R) & \quad PT_{R \rightarrow}(t_6) \quad t_2(rrq) = (r \rightarrow t_6(rrq, r)) \\ S_3 : Q \vdash Q & \quad (\text{axiom } Id) \quad t_3(q) = q \\ S_1 : (R \rightarrow R) \rightarrow Q \vdash Q & \quad PT_{L \rightarrow}(t_2, t_3) \quad t_1(rrq) = t_3(rrq(t_2(rrq))) \\ S_0 : \emptyset \vdash ((R \rightarrow R) \rightarrow Q) \rightarrow Q & \quad PT_{R \rightarrow}(t_1) \quad t_0 = (rrq \rightarrow t_1(rrq)) \end{aligned}$$

The proof expression for S_0 is then obtained as

$$\begin{aligned} t_0 &= rrq \rightarrow t_3(rrq(t_2(rrq))) = rrq \rightarrow rrq(r \rightarrow t_6(rrq, r)) \\ &= rrq \rightarrow rrq(r \rightarrow r) \end{aligned}$$

Simplified final code having the required type:

$$t_0 : ((R \rightarrow R) \rightarrow Q) \rightarrow Q = (rrq \rightarrow rrq(r \rightarrow r))$$

To prove that there is no proof, one needs to use methods of proof theory that are beyond the scope of this book. A good introduction to the required technique is the book "Proof and Disproof in Formal Logic" by R. Bornat.¹

¹ R. Bornat, "Proof and Disproof in Formal Logic", Oxford, 2005 - link to Amazon.com

D Parametricity theorem and naturality laws

Functional programming (FP) focuses on a small set of language features — the six type constructions and the nine code constructions, introduced in Sections 5.1.2 and 5.2.3; see also Tables 7.1 and D.1–D.2. These constructions create **purely functional** programs and are sufficient to implement all design patterns of FP. At the same time, restricting programs to be written only through the purely functional constructions enables several ways of mathematical reasoning about code. Examples of such reasoning are treating programs as mathematical values (referential transparency); deriving the most general type from code (type inference); and deriving code from type (code inference).

Another property of purely functional code is automatic adherence to naturality laws. By the **parametricity theorem**, any purely functional code with a type parameter will satisfy a certain naturality law. Not having to derive and verify naturality laws by hand saves a lot of time.

Type construction	Scala example	Type notation
unit or a “named unit”	Unit or None	$\mathbb{1}$
type parameter	\mathbf{A}	A
product type	(\mathbf{A}, \mathbf{B})	$A \times B$
co-product type	Either $[\mathbf{A}, \mathbf{B}]$	$A + B$
function type	$\mathbf{A} \Rightarrow \mathbf{B}$	$A \rightarrow B$
recursive type	Fix $[\mathbf{F}[_]]$	Fix^F

Table D.1: The six type constructions of purely functional programming.

placement is done by *guessing*: the parametricity theorem no longer provides guidance at that point.

Adapting the parametricity theorem to the needs of FP practitioners, this Appendix derives all naturality laws for purely functional code without using relations and without guessing.

D.1 Commutativity laws for profunctors and bifunctors

The first result concerns type constructors with two type parameters, such as bifunctors. The bifunctor commutativity law (6.9) was introduced in Section 6.2.2. We will now prove the commutativity law by induction on the type structure of a given bifunctor.⁴ It turns out that the proof also requires the commutativity laws for type constructors with *contravariant* type parameters. All those proofs are completely analogous, so we will first prove the commutativity law for profunctors, which are type constructors with one contravariant and one covariant type parameter. The profunctor commutativity law will be important for the proof of the main parametricity theorem (Section D.2).

¹See <https://homepages.inf.ed.ac.uk/wadler/papers/free/free.ps> and some further explanations in the blog posts <https://reasonablypolymorphic.com/blog/theorems-for-free/> and <https://bartoszmilewski.com/2014/09/22/>

²See the paper by J. Reynolds, “*Types, abstraction, and parametric polymorphism*”, *Information Processing* 83, p. 513 (1983).

³Tutorials on that subject are rare; one is found in the paper by Ronald Backhouse, “*On a relation on functions*” (1990). See also the book by R. Bird and O. de Moor, “*The algebra of programming*” (1997).

⁴In the blog post <https://byorgey.wordpress.com/2018/03/30/>, Brent Yorgey gave a proof of the commutativity law for bifunctors by using the Reynolds-Wadler parametricity theorem.

Code construction	Scala example	Code notation
use unit value	<code>()</code>	1
use argument (bound variable)	<code>x</code>	x
create function	<code>x => expression</code>	$x \rightarrow expression$
use function	<code>f(x)</code>	$f(x)$ or $x \triangleright f$
create tuple	<code>(a, b)</code>	$a \times b$
use tuple	<code>{ case (a, b) => ... }</code> or <code>p._1</code> or <code>p._2</code>	$a \times b \rightarrow \dots$ or $p \triangleright \pi_1$ or $p \triangleright \pi_2$
create disjunctive value	<code>Left[A, B](a)</code>	$a:A + 0:B$ or $\boxed{a \ 0}$
use disjunctive value	<code>p match { case Left(a) => f(a) case Right(b) => g(b) }</code>	$p \triangleright \begin{array}{ c c } \hline & C \\ \hline A & f \\ \hline B & g \\ \hline \end{array}$ or $p \triangleright \begin{array}{ c } \hline f \\ \hline g \\ \hline \end{array}$
use recursive call	<code>def f(x) = { ... f(y) ... }</code>	$f(x) \triangleq \dots \overline{f}(y) \dots$

Table D.2: The nine code constructions of purely functional programming.

D.1.1 Proof of the profunctor commutativity law

Given an arbitrary type constructor $P^{A,B}$ contravariant with respect to A and covariant with respect to B , we formulate the profunctor **commutativity law** by analogy with the bifunctor law (6.9):

$$\begin{aligned} \text{commutativity law of } P : \quad & \text{cmap}_{P \bullet B}(f:A \rightarrow C) ; \text{fmap}_{P A \bullet}(g:B \rightarrow D) = \text{fmap}_{P C \bullet}(g) ; \text{cmap}_{P \bullet D}(f) . \\ \text{shorter notation : } \quad & f \downarrow^{P \bullet B} ; g \uparrow^{P A \bullet} = g \uparrow^{P C \bullet} ; f \downarrow^{P \bullet D} . \end{aligned} \quad (\text{D.1})$$

$$\begin{array}{ccc} P^{C,B} & \xrightarrow{\text{cmap}_{P \bullet B}(f:A \rightarrow C)} & P^{A,B} \\ \text{fmap}_{P C \bullet}(g) \downarrow & & \downarrow \text{fmap}_{P A \bullet}(g:B \rightarrow D) \\ P^{C,D} & \xrightarrow{\text{cmap}_{P \bullet D}(f)} & P^{A,D} \end{array}$$

We will now prove that Eq. (D.1) holds for any **purely functional profunctor** $P^{A,B}$, i.e., a profunctor defined via a combination of the six type constructions from Table D.1. We will assume that all contrafunctor liftings such as $f \downarrow^{P \bullet B}$ and all functor liftings such as $g \uparrow^{P A \bullet}$ are implemented as shown in Chapter 6 for functors and contrafunctors corresponding to each of the type constructions. (Section D.2.4

will show that the code from Chapter 6 is the only possible lawful implementation of the liftings.)

The proof goes by induction on the type structure of $P^{A,B}$. The base case contains the first two constructions (the unit type and the type parameter), which gives two possibilities: $P^{A,B} \triangleq 1$ (constant profunctor) and $P^{A,B} \triangleq B$ (identity profunctor). The other possibility of using a type parameter, $P^{A,B} \triangleq A$, does not give a profunctor since $P^{A,B}$ must be contravariant in A .

The commutativity law holds for $P^{A,B} \triangleq 1$ because all liftings are identity functions: $\text{cmap}_{P \bullet B}(f) = \text{id}$ and $\text{fmap}_{P A \bullet}(g) = \text{id}$. For the same reason, the law will also hold for the constant profunctor $P^{A,B} \triangleq Z$ where Z is a fixed type (or a type parameter other than A or B).

For the profunctor $P^{A,B} \triangleq B$, the law holds because $\text{cmap}_{P \bullet B}(f) = \text{id}$. So, the base case is proved.

The inductive step looks at the outer construction in the type expression of $P^{A,B}$, which must be one of product, co-product, function, or recursion. These constructions create larger type expressions out of smaller ones; for example, $P^{A,B} \triangleq (A \rightarrow B) \times B$ uses the product construction with sub-expressions $A \rightarrow B$ and B , which can be viewed as profunctors $R^{A,B} \triangleq A \rightarrow B$ and $S^{A,B} \triangleq B$. The inductive assumption is that the commutativity law is already proved for all sub-expression types (such as $R^{A,B}$ or $S^{A,B}$). We will then need to prove the law for the entire type expression $P^{A,B}$.

Product type If the outer level of $P^{A,B}$ is a product, we must have $P^{A,B} = R^{A,B} \times S^{A,B}$ where $R^{A,B}$ and $S^{A,B}$ are some profunctors. The code for liftings follows from Statement 6.2.3.3 and Exercise 6.3.1.2:

$$f^{\downarrow P^{\bullet,B}} \triangleq f^{\downarrow R^{\bullet,B}} \boxtimes f^{\downarrow S^{\bullet,B}} \quad , \quad g^{\uparrow P^{A,\bullet}} \triangleq g^{\uparrow R^{A,\bullet}} \boxtimes g^{\uparrow S^{A,\bullet}} \quad .$$

By inductive assumption, R and S already satisfy the commutativity law (D.1). So, we compute

$$\begin{aligned} \text{expect to equal } g^{\uparrow P^{C,\bullet}} \circ f^{\downarrow P^{\bullet,D}} : & \quad \underline{f^{\downarrow P^{\bullet,B}} \circ g^{\uparrow P^{A,\bullet}}} = (f^{\downarrow R^{\bullet,B}} \boxtimes f^{\downarrow S^{\bullet,B}}) \circ (g^{\uparrow R^{A,\bullet}} \boxtimes g^{\uparrow S^{A,\bullet}}) \\ \text{composition law (7.2)} : & \quad = (\underline{f^{\downarrow R^{\bullet,B}} \circ g^{\uparrow R^{A,\bullet}}}) \boxtimes (\underline{f^{\downarrow S^{\bullet,B}} \circ g^{\uparrow S^{A,\bullet}}}) \\ \text{inductive assumption} : & \quad = (g^{\uparrow R^{C,\bullet}} \circ f^{\downarrow R^{\bullet,D}}) \boxtimes (g^{\uparrow S^{C,\bullet}} \circ f^{\downarrow S^{\bullet,D}}) \\ \text{composition law (7.2)} : & \quad = (\underline{g^{\uparrow R^{C,\bullet}} \boxtimes g^{\uparrow S^{C,\bullet}}}) \circ (\underline{f^{\downarrow R^{\bullet,D}} \boxtimes f^{\downarrow S^{\bullet,D}}}) = g^{\uparrow P^{C,\bullet}} \circ f^{\downarrow P^{\bullet,D}} \quad . \end{aligned}$$

Co-product type In this case, we must have $P^{A,B} = R^{A,B} + S^{A,B}$ where $R^{A,B}$ and $S^{A,B}$ are some profunctors. The code for liftings follows from Statement 6.2.3.4 and Exercise 6.3.1.2:

$$f^{\downarrow P^{\bullet,B}} \triangleq \begin{vmatrix} f^{\downarrow R^{\bullet,B}} & 0 \\ 0 & f^{\downarrow S^{\bullet,B}} \end{vmatrix} \quad , \quad g^{\uparrow P^{A,\bullet}} \triangleq \begin{vmatrix} g^{\uparrow R^{A,\bullet}} & 0 \\ 0 & g^{\uparrow S^{A,\bullet}} \end{vmatrix} \quad .$$

By inductive assumption, R and S already satisfy the commutativity law (D.1). So, we compute

$$\begin{aligned} \text{expect to equal } g^{\uparrow P^{C,\bullet}} \circ f^{\downarrow P^{\bullet,D}} : & \quad f^{\downarrow P^{\bullet,B}} \circ g^{\uparrow P^{A,\bullet}} = \begin{vmatrix} f^{\downarrow R^{\bullet,B}} & 0 \\ 0 & f^{\downarrow S^{\bullet,B}} \end{vmatrix} \circ \begin{vmatrix} g^{\uparrow R^{A,\bullet}} & 0 \\ 0 & g^{\uparrow S^{A,\bullet}} \end{vmatrix} \\ \text{matrix composition} : & \quad = \begin{vmatrix} f^{\downarrow R^{\bullet,B}} \circ g^{\uparrow R^{A,\bullet}} & 0 \\ 0 & f^{\downarrow S^{\bullet,B}} \circ g^{\uparrow S^{A,\bullet}} \end{vmatrix} \\ \text{inductive assumption} : & \quad = \begin{vmatrix} g^{\uparrow R^{C,\bullet}} \circ f^{\downarrow R^{\bullet,D}} & 0 \\ 0 & g^{\uparrow S^{C,\bullet}} \circ f^{\downarrow S^{\bullet,D}} \end{vmatrix} \quad . \end{aligned}$$

The right-hand side of the law, $g^{\uparrow P^{C,\bullet}} \circ f^{\downarrow P^{\bullet,D}}$, can be rewritten as

$$\begin{aligned} g^{\uparrow P^{C,\bullet}} \circ f^{\downarrow P^{\bullet,D}} &= \begin{vmatrix} g^{\uparrow R^{C,\bullet}} & 0 \\ 0 & g^{\uparrow S^{C,\bullet}} \end{vmatrix} \circ \begin{vmatrix} f^{\downarrow R^{\bullet,D}} & 0 \\ 0 & f^{\downarrow S^{\bullet,D}} \end{vmatrix} \\ \text{matrix composition} : & \quad = \begin{vmatrix} g^{\uparrow R^{C,\bullet}} \circ f^{\downarrow R^{\bullet,D}} & 0 \\ 0 & g^{\uparrow S^{C,\bullet}} \circ f^{\downarrow S^{\bullet,D}} \end{vmatrix} \quad . \end{aligned}$$

This now coincides with the left-hand side of the law.

Function type The type expression $P^{A,B} \triangleq R^{B,A} \rightarrow S^{A,B}$ (note the swapped type parameters in $R^{B,A}$) is contravariant in A and covariant in B when $R^{A,B}$ and $S^{A,B}$ are any profunctors. The lifting code follows from Statement 6.2.3.5 and Exercise 6.3.1.1:

$$(f^{:A \rightarrow C})^{\downarrow P^{\bullet,B}} \triangleq p^{:P^{C,B}} \rightarrow f^{\uparrow R^{B,\bullet}} \circ p \circ f^{\downarrow S^{\bullet,B}} \quad , \quad (g^{:B \rightarrow D})^{\uparrow P^{A,\bullet}} \triangleq p^{:P^{A,B}} \rightarrow g^{\downarrow R^{\bullet,A}} \circ p \circ g^{\uparrow S^{A,\bullet}} \quad .$$

By inductive assumption, R and S satisfy the commutativity law (D.1). Omitting types, we compute

$$\begin{aligned} \text{left-hand side} : & \quad f^{\downarrow P} \circ g^{\uparrow P} = (p \rightarrow f^{\uparrow R} \circ p \circ f^{\downarrow S}) \circ (p \rightarrow g^{\downarrow R} \circ p \circ g^{\uparrow S}) \\ \text{compute composition} : & \quad = p \rightarrow \underline{g^{\downarrow R} \circ f^{\uparrow R} \circ p \circ f^{\downarrow S} \circ g^{\uparrow S}} \\ \text{inductive assumption} : & \quad = p \rightarrow f^{\uparrow R} \circ g^{\downarrow R} \circ p \circ g^{\uparrow S} \circ f^{\downarrow S} \quad . \end{aligned}$$

The right-hand side of the law is then transformed to the same expression:

$$g^{\uparrow P} ; f^{\downarrow P} = (p \rightarrow g^{\downarrow R} ; p ; g^{\uparrow S}) ; (p \rightarrow f^{\uparrow R} ; p ; f^{\downarrow S})$$

compute composition : $= p \rightarrow f^{\uparrow R} ; g^{\downarrow R} ; p ; g^{\uparrow S} ; f^{\downarrow S}$.

Recursive type A recursive profunctor P is defined using a type equation of the form

$$P^{A,B} \triangleq S^{A,B,PA,B} , \text{ equivalently denoted as } P^{A,B} \triangleq \text{Fix}^{S^{A,B,*}} ,$$

where $S^{A,B,R}$ must be contravariant in A but covariant in B and R . The liftings are defined by

$$(f:A \rightarrow C)^{\downarrow P^{*,B}} \triangleq f^{\downarrow S^{*,B,PC,B}} ; (f^{\downarrow P^{*,B}})^{\uparrow S^{A,B,*}} , \quad (g:B \rightarrow D)^{\uparrow P^{A,*}} \triangleq g^{\uparrow S^{A,*,PA,B}} ; (g^{\uparrow P^{A,*}})^{\uparrow S^{A,D,*}} .$$

The inductive assumption is two-fold: first, that the recursively used lifting to $P^{A,B}$ already satisfies the commutativity law. Second, that the profunctor $S^{A,B,R}$ satisfies the commutativity law with respect to *any* two parameters (we will prove that law for covariant type parameters in Section D.1.2).

Both sides of the commutativity law are functions of type $PC,B \rightarrow PA,D$ or equivalently (if we expand the type recursion) as functions of type $S^{C,B,PC,B} \rightarrow S^{A,D,PA,D}$. To avoid mistakes, we write out the type parameters in this computation:

$$\begin{aligned} \text{expect to equal } g^{\uparrow P^{C,*}} ; f^{\downarrow P^{*,D}} : & f^{\downarrow P^{*,B}} ; g^{\uparrow P^{A,*}} = f^{\downarrow S^{*,B,PC,B}} ; (f^{\downarrow P^{*,B}})^{\uparrow S^{A,B,*}} ; g^{\uparrow S^{A,*,PA,B}} ; (g^{\uparrow P^{A,*}})^{\uparrow S^{A,D,*}} \\ \text{B,R-commutativity of } S^{A,B,R} : & = f^{\downarrow S^{*,B,PC,B}} ; g^{\uparrow S^{A,*,PC,B}} ; (f^{\downarrow P^{*,B}})^{\uparrow S^{A,D,*}} ; (g^{\uparrow P^{A,*}})^{\uparrow S^{A,D,*}} \\ \text{A,B-commutativity of } S^{A,B,R} : & = g^{\uparrow S^{C,*,PC,B}} ; f^{\downarrow S^{*,D,PC,C,B}} ; (f^{\downarrow P^{*,B}} ; g^{\uparrow P^{A,*}})^{\uparrow S^{A,D,*}} \\ \text{inductive assumption} : & = g^{\uparrow S^{C,*,PC,B}} ; f^{\downarrow S^{*,D,PC,C,B}} ; (g^{\uparrow P^{C,*}} ; f^{\downarrow P^{*,D}})^{\uparrow S^{A,D,*}} \\ \text{composition under } \uparrow S^{A,D,*} : & = g^{\uparrow S^{C,*,PC,B}} ; f^{\downarrow S^{*,D,PC,C,B}} ; (g^{\uparrow P^{C,*}})^{\uparrow S^{A,D,*}} ; (f^{\downarrow P^{*,D}})^{\uparrow S^{A,D,*}} \\ \text{A,R-commutativity of } S^{A,B,R} : & = g^{\uparrow S^{C,*,PC,B}} ; (g^{\uparrow P^{C,*}})^{\uparrow S^{C,D,*}} ; f^{\downarrow S^{*,D,PC,D}} ; (f^{\downarrow P^{*,D}})^{\uparrow S^{A,D,*}} \\ \text{definitions of liftings} : & = g^{\uparrow P^{C,*}} ; f^{\downarrow P^{*,D}} . \end{aligned}$$

This concludes the proof of the profunctor commutativity law.

D.1.2 Commutativity laws for bifunctors and bi-contrafunctors

A bi-contrafunctor $P^{A,B}$ is a type constructor contravariant with respect to both A and B . The commutativity law for bi-contrafunctors is formulated as

$$\begin{aligned} \text{commutativity law of } P : \quad & \text{cmap}_{P,D}(f:A \rightarrow C) ; \text{cmap}_{P,A,*}(g:B \rightarrow D) = \text{cmap}_{PC,*}(g) ; \text{cmap}_{P,B}(f) . \\ \text{shorter notation} : \quad & f^{\downarrow P^{*,D}} ; g^{\downarrow P^{A,*}} = g^{\downarrow P^{C,*}} ; f^{\downarrow P^{*,B}} . \end{aligned} \quad (\text{D.2})$$

$$\begin{array}{ccc} PC,D & \xrightarrow{\text{cmap}_{P,D}(f:A \rightarrow C)} & PA,D \\ \text{cmap}_{PC,*}(g:B \rightarrow D) \downarrow & & \downarrow \text{cmap}_{P,A,*}(g) \\ PC,B & \xrightarrow{\text{cmap}_{P,B}(f:A \rightarrow C)} & PA,B \end{array}$$

Any bifunctor or bi-contrafunctor whose type expression is built up using the six type constructions (Table D.1) will satisfy its commutativity law. To prove that, we need to repeat the same calculations as for profunctors in Section D.1.1 except for changing `cmap` into `fmap` or back when needed. We only need to check that the proof will still work after such changes. Looking

over the proof in Section D.1.1, we find that we never used the functor or contrafunctor *composition law* for P . The only usage of the composition law was with respect to the type parameter R in the structure functor $S^{A,B,R}$ of the recursive type construction. However, in all cases $S^{A,B,R}$ needs to be *covariant* in R because that is required by the recursive type equation. So, we are assured that the same proof with minor changes will work for bifunctors and bi-contrafunctors.

The same techniques and proofs apply to type constructors with more than two type parameters.

D.2 Naturality laws for purely functional transformations

The goal of this section is to derive and prove a naturality law for any transformation implemented via purely functional code. Simple examples of such transformations are the `map`, `filter`, and `fold` methods for the `Option` functor, whose type signatures can be written as

$$\begin{aligned} \text{fmap}_{\text{Opt}}^{A,B} : (A \rightarrow B) \rightarrow \text{Opt}^A \rightarrow \text{Opt}^B & , \\ \text{filt}_{\text{Opt}}^A : (A \rightarrow 2) \rightarrow \text{Opt}^A \rightarrow \text{Opt}^A & , \\ \text{fold}_{\text{Opt}}^{A,B} : B \times (A \times B \rightarrow B) \rightarrow \text{Opt}^A \rightarrow B & . \end{aligned}$$

These methods satisfy appropriate naturality laws — one law per type parameter. When a method's type signature is that of a natural transformation between functors (or between contrafunctors), the naturality laws have the form derived in Section 9.4.1. For instance, fixing the type parameter A in the `fmap` method, we obtain a type signature of the form $F^B \rightarrow G^B$ where F^\bullet and G^\bullet are functors:

$$\text{fmap}_{\text{Opt}}^B : F^B \rightarrow G^B , \quad F^B \triangleq A \rightarrow B , \quad G^B \triangleq \text{Opt}^A \rightarrow \text{Opt}^B .$$

The corresponding naturality law (which is equivalent to the functor composition law) is

$$f^{\uparrow F} ; \text{fmap}_{\text{Opt}} = \text{fmap}_{\text{Opt}} ; f^{\uparrow G} .$$

However, the type signature of `fold` is not of the form $P^B \rightarrow Q^B$ with any functors or contrafunctors P, Q . In general, it is not obvious how to write the naturality law for transformations with such type signatures. The parametricity theorem provides a recipe for writing naturality laws and also gives a proof that purely functional transformations always satisfy their naturality laws.

The key insight for deriving that theorem is that type signatures of purely functional transformations must always contain type parameters in either covariant or contravariant positions. So, all such type signatures can be written as $P^{A,A} \rightarrow Q^{A,A}$ where P and Q are some *profunctors*. For instance, the type signature of `fold` with respect to the type parameter B (with the parameter A fixed) is written as

$$\text{fold}_{\text{Opt}}^B : P^{B,B} \rightarrow Q^{B,B} , \quad \text{where } P^{X,Y} \triangleq Y \times (A \times X \rightarrow Y) , \quad Q^{X,Y} \triangleq \text{Opt}^A \rightarrow Y . \quad (\text{D.3})$$

The next task is to motivate the naturality law for profunctor transformations of type $\forall A. P^{A,A} \rightarrow Q^{A,A}$, where the type parameters of the profunctors are set to be the same. Such transformations are called “dinatural”. (Natural transformations between profunctors P and Q have a different type signature, namely $\forall (A, B). P^{A,B} \rightarrow Q^{A,B}$, where the type parameters A, B are independent).

D.2.1 Dinatural transformations between profunctors

A **dinatural transformation** is a function $t^A : P^{A,A} \rightarrow Q^{A,A}$, where $P^{X,Y}$ and $Q^{X,Y}$ are profunctors contravariant in X and covariant in Y . A dinatural transformation t must satisfy the naturality law

$$(f : A \rightarrow B) \downarrow P^{\bullet,A} ; t^A ; f^{\uparrow Q^{\bullet,A}} = f^{\uparrow P^{B,\bullet}} ; t^B ; f \downarrow Q^{\bullet,B} . \quad (\text{D.4})$$

$$\begin{array}{ccc} P^{A,A} & \xrightarrow{t} & Q^{A,A} \\ \nearrow xmap_P(f)(id) & & \searrow xmap_Q(id)(f) \\ P^{B,A} & & Q^{A,B} \\ \searrow xmap_P(id)(f) & & \nearrow xmap_Q(f)(id) \\ P^{B,B} & \xrightarrow{t} & Q^{B,B} \end{array}$$

To build up intuition for that law (see diagram at left), compare the laws of natural transformations $t^A : F^A \rightarrow G^A$ when F^\bullet and G^\bullet are functors,

$$t^A ; (f : A \rightarrow B) \uparrow G = (f : A \rightarrow B) \uparrow F ; t^B , \quad (\text{D.5})$$

and when F^\bullet and G^\bullet are contrafunctors,

$$(f : A \rightarrow B) \downarrow F ; t^A = t^B ; (f : A \rightarrow B) \downarrow G . \quad (\text{D.6})$$

We would obtain naturality laws of that form if we could fix the type parameter A in the profunctors $P^{A,B}$ and $Q^{A,B}$ and consider then as functors with respect to the type parameter B . However, this requires us to have a transformation $\tilde{t} : P^{A,B} \rightarrow Q^{A,B}$ defined for arbitrary (not necessarily equal) type parameters A, B . We do not have such a function: we are only given a transformation with the type signature $t : P^{A,A} \rightarrow Q^{A,A}$, defined only in the “diagonal” case⁵ $A = B$. As a rule, we cannot extend the code of t to some \tilde{t} that works with arbitrary type parameters A, B . A law for t must be an equation that somehow transforms both type parameters of t by using a lifted function $f^{:A \rightarrow B}$.

The naturality law (D.4) combines the laws (D.5) and (D.6) in the way required for all types to match. On the other hand, the laws (D.5) and (D.6) will follow from Eq. (D.4) when $P^{A,A}$ and $Q^{A,A}$ are both functors or both contrafunctors in A . To further motivate the law (D.4), we will now derive the known forms of naturality laws for `filter` and `fold` for arbitrary filterable or foldable functors F .

Example D.2.1.1 (naturality law of `filter`) To derive the naturality law of `filter`, express `filter`’s type signature through profunctors P and Q as

$$\text{filt}_F^A : P^{A,A} \rightarrow Q^{A,A} \quad , \quad P^{X,Y} \triangleq (X \rightarrow 2) \quad , \quad Q^{X,Y} \triangleq F^X \rightarrow F^Y \quad ,$$

and then write the law (D.4),

$$f^{\downarrow P^{\bullet,A}} ; \text{filt}_F^A ; f^{\uparrow Q^{A,\bullet}} \stackrel{?}{=} f^{\uparrow P^{B,\bullet}} ; \text{filt}_F^B ; f^{\downarrow Q^{\bullet,B}} \quad . \quad (\text{D.7})$$

It remains to substitute the code for the liftings using the specific types of P and Q :

$$\begin{aligned} (f^{:A \rightarrow B})^{\downarrow P^{\bullet,A}} &= p^{:B \rightarrow 2} \rightarrow f ; p \quad , \quad f^{\uparrow P^{B,\bullet}} = \text{id} \quad , \\ (f^{:A \rightarrow B})^{\downarrow Q^{\bullet,B}} &= q^{:F^B \rightarrow F^B} \rightarrow f^{\uparrow F} ; q \quad , \quad f^{\uparrow Q^{A,\bullet}} = q^{:F^A \rightarrow F^A} \rightarrow q ; f^{\uparrow F} \quad . \end{aligned}$$

Then we rewrite Eq. (D.7) as

$$(p \rightarrow f ; p) ; \text{filt}_F ; (q \rightarrow q ; f^{\uparrow F}) \stackrel{?}{=} \text{id} ; \text{filt}_F ; (q \rightarrow f^{\uparrow F} ; q) \quad .$$

To simplify the form of the naturality law, apply both sides to an arbitrary $p^{:P^{B,A}} = p^{:B \rightarrow 2}$:

$$\begin{aligned} \text{left-hand side : } p \triangleright (p \rightarrow f ; p) ; \text{filt}_F ; (q \rightarrow q ; f^{\uparrow F}) \\ \triangleright\text{-notation : } &= p \triangleright (p \rightarrow f ; p) \triangleright \text{filt}_F \triangleright (q \rightarrow q ; f^{\uparrow F}) \\ \text{apply functions : } &= (f ; p) \triangleright \text{filt}_F \triangleright (q \rightarrow q ; f^{\uparrow F}) = \text{filt}_F(f ; p) \triangleright (q \rightarrow q ; f^{\uparrow F}) = \text{filt}_F(f ; p) ; f^{\uparrow F} \quad , \\ \text{right-hand side : } p \triangleright \text{id} ; \text{filt}_F ; (q \rightarrow f^{\uparrow F} ; q) &= p \triangleright \text{filt}_F \triangleright (q \rightarrow f^{\uparrow F} ; q) \\ &= \text{filt}_F(p) \triangleright (q \rightarrow f^{\uparrow F} ; q) = f^{\uparrow F} ; \text{filt}_F(p) \quad . \end{aligned}$$

So, we obtained the naturality law (9.3) of `filter`,

$$\text{filt}_F(f ; p) ; f^{\uparrow F} = f^{\uparrow F} ; \text{filt}_F(p) \quad .$$

Example D.2.1.2 (naturality law of `fold`) To derive the naturality law of `fold` with respect to the type parameter B , we begin with Eq. (D.3) that shows the type signature of `fold` as a transformation of type $P^{B,B} \rightarrow Q^{B,B}$ between profunctors P, Q . Since the type parameter A is fixed, the naturality law (D.4) now involves an arbitrary function $f^{:B \rightarrow C}$,

$$(f^{:B \rightarrow C})^{\downarrow P^{\bullet,B}} ; \text{fold}_F^B ; f^{\uparrow Q^{B,\bullet}} = f^{\uparrow P^{C,\bullet}} ; \text{fold}_F^C ; f^{\downarrow Q^{\bullet,C}} \quad . \quad (\text{D.8})$$

The lifting code required for the profunctors $P^{X,Y} \triangleq Y \times (A \times X \rightarrow Y)$ and $Q^{X,Y} \triangleq F^A \rightarrow Y$ is

$$\begin{aligned} (f^{:B \rightarrow C})^{\downarrow P^{\bullet,B}} &= \text{id}^B \boxtimes (h^{:A \times C \rightarrow B} \rightarrow a^{:A} \times b^{:B} \rightarrow h(a \times f(b))) \quad , \quad f^{\uparrow P^{C,\bullet}} = f \boxtimes (h^{:A \times C \rightarrow B} \rightarrow h ; f) \quad , \\ (f^{:B \rightarrow C})^{\downarrow Q^{\bullet,C}} &= \text{id} \quad , \quad f^{\uparrow Q^{B,\bullet}} = q^{:F^A \rightarrow B} \rightarrow q ; f \quad . \end{aligned}$$

⁵As a memory aid, we may consider the word “dinaturality” to be a shorthand for “diagonal naturality”.

Substituting this code into the law (D.8) and applying to an arbitrary $p : P^{C,B} = z : B \times h : A \times C \rightarrow B$, we get

$$\begin{aligned}
 \text{left-hand side : } & (z \times h) \triangleright (f : B \rightarrow C) \downarrow P^{\bullet,B} ; \text{fold}_F ; f \uparrow Q^{\bullet,B} \\
 \text{definitions of liftings : } & = (z \times h) \triangleright (\text{id} \boxtimes (h \rightarrow a \times b \rightarrow h(a \times f(b)))) ; \text{fold}_F ; (q \rightarrow q ; f) \\
 \triangleright\text{-notation : } & = (z \times h) \triangleright (\text{id} \boxtimes (h \rightarrow a \times b \rightarrow h(a \times f(b)))) \triangleright \text{fold}_F \triangleright (q \rightarrow q ; f) \\
 \text{apply functions : } & = \text{fold}_F(z \times (a \times b \rightarrow h(a \times f(b)))) ; f \quad , \\
 \text{right-hand side : } & (z \times h) \triangleright f \uparrow P^{\bullet,B} ; \text{fold}_F ; f \downarrow Q^{\bullet,C} = (z \times h) \triangleright (f \boxtimes (h : A \times C \rightarrow B \rightarrow h ; f)) ; \text{fold}_F ; \text{id} \\
 \text{apply functions : } & = \text{fold}_F(f(z) \times (h ; f)) \quad .
 \end{aligned}$$

We obtained a naturality law of `fold`,

$$\text{fold}_F(f(z) \times (h ; f)) = \text{fold}_F(z \times (a \times b \rightarrow h(a \times f(b)))) ; f \quad .$$

These examples illustrate how we may derive the form of the naturality law for any type signature by specializing the general law (D.4) to specific profunctors P and Q .

In the next subsections, we will prove⁶ that any purely functional transformation $t^A : P^{A,A} \rightarrow Q^{A,A}$ is dinatural, i.e., satisfies its naturality law (D.4). Since the form of the law depends only on the type signature, all transformations t^A satisfy the same naturality law.

The restriction to *purely functional* programs is essential because it excludes, for instance, any use of mutable variables, `null` values, exceptions, run-time type identification, run-time code loading, or values defined in external libraries that are not known to be purely functional. Code that uses those features of Scala is not covered by parametricity theorems and may violate naturality laws, as we have seen elsewhere in this book.

Dinatural expressions We note that the type constructor $P^{A,A} \rightarrow Q^{A,A}$ is neither covariant nor contravariant in A , but it is itself a profunctor that we may denote by T . The naturality law is simpler when formulated via T :

Statement D.2.1.3 Using the profunctor $T^{X,Y} \triangleq P^{Y,X} \rightarrow Q^{X,Y}$, the naturality law (D.4) is written as

$$\begin{array}{ccc}
 \begin{array}{c} T^{B,B} \\ \downarrow f \downarrow T^{\bullet,B} \\ T^{A,A} \xrightarrow{f \uparrow T^{\bullet,A}} T^{A,B} \end{array} & t^A \triangleright f \uparrow T^{\bullet,A} = t^B \triangleright f \downarrow T^{\bullet,B} \quad , & \text{(D.9)} \\
 \text{where the transformation } t \text{ is viewed as a value of type } \forall Z. T^{Z,Z} \text{, while the} \\
 \text{function } f : A \rightarrow B \text{ is arbitrary. This law is known as the \textbf{wedge law} for } t. \end{array}$$

Proof The two liftings for T are expressed as

$$\begin{aligned}
 f \uparrow T^{\bullet,A} : T^{A,A} & \rightarrow T^{A,B} \quad , \quad t^A \triangleright f \uparrow T^{\bullet,A} = p : P^{B,A} \rightarrow t^A(p \triangleright f \downarrow P^{\bullet,A}) \triangleright f \uparrow Q^{\bullet,A} = f \downarrow P^{\bullet,A} ; t^A ; f \uparrow Q^{\bullet,A} \quad , \\
 f \downarrow T^{\bullet,B} : T^{B,B} & \rightarrow T^{A,B} \quad , \quad t^B \triangleright f \downarrow T^{\bullet,B} = p : P^{B,A} \rightarrow t^B(p \triangleright f \uparrow P^{\bullet,B}) \triangleright f \downarrow Q^{\bullet,B} = f \uparrow P^{\bullet,B} ; t^B ; f \downarrow Q^{\bullet,B} \quad .
 \end{aligned}$$

It follows that Eq. (D.9) is equivalent to Eq. (D.4), which proves the statement.

Expressions t of type $\forall Z. T^{Z,Z}$ satisfying the wedge law (D.9) are called **dinatural expressions**. The wedge law formulates the dinaturality condition for any expression t with a type parameter, whether t is a “transformation” (i.e., a function) or not. It can be derived from the parametricity theorem that any purely functional expression of type $\forall Z. T^{Z,Z}$ is dinatural.

D.2.2 Composition of natural and dinatural transformations

In the proof of the parametricity theorem, we will need to compose some dinatural transformations. Composition of *natural* transformations is easily shown to be again natural. However, proving the same property for dinatural transformations requires additional assumptions. To show why, first consider the composition of two natural transformations $u : P^A \rightarrow Q^A$ and $v : Q^A \rightarrow R^A$.

⁶The proof builds upon ideas from the paper by E. S. Bainbridge et al., “Functorial polymorphism” (Theor. Comp. Sci. 70, p. 35, 1990), see <https://www.sciencedirect.com/science/article/pii/0304397590901517>. A more rigorous but significantly more complicated proof was given by Joachim de Lataillade in the paper “Dinatural terms in System F”, see <https://www.irif.fr/~delatail/dinat.pdf>, based on deriving the syntactic form of naturality laws in full detail.

Statement D.2.2.1 The composition $t \triangleq u \circ v$ is a natural transformation, $t : P^A \rightarrow R^A$, assuming that $u : P^A \rightarrow Q^A$ and $v : Q^A \rightarrow R^A$ satisfy their naturality laws, and that P, Q, R are functors.

Proof The naturality laws for u and v are written with an arbitrary $f : A \rightarrow B$ as

$$f \uparrow P \circ u = u \circ f \uparrow Q \quad , \quad f \uparrow Q \circ v = v \circ f \uparrow R \quad .$$

The required naturality law for t is derived by a direct calculation,

$$\begin{array}{ccccc} P^A & \xrightarrow{u} & Q^A & \xrightarrow{v} & R^A \\ \downarrow f \uparrow P & & \downarrow f \uparrow Q & & \downarrow f \uparrow R \\ P^B & \xrightarrow{u} & Q^B & \xrightarrow{v} & R^B \end{array} \quad \begin{array}{l} \text{expect to equal } t \circ f \uparrow R : f \uparrow P \circ u = f \uparrow P \circ u \circ v \\ \text{naturality of } u : u \circ f \uparrow Q \circ v \\ \text{naturality of } v : u \circ v \circ f \uparrow R = t \circ f \uparrow R \end{array} .$$

This calculation shows that the above type diagram commutes.

In the proof of Statement D.2.2.1, we derived the naturality law for the composed transformation $t \triangleq u \circ v$ directly from the naturality laws of u and v with no further assumptions. In particular, we did not need to assume that u and v are implemented by purely functional code. We will now see that the composition property for dinatural transformations cannot be derived in the same way.

Consider two dinatural transformations $u : P^{A,A} \rightarrow Q^{A,A}$ and $v : Q^{A,A} \rightarrow R^{A,A}$, where P, Q, R are any profunctors. The naturality laws of u and v are written as

$$f \downarrow P^{\bullet,A} \circ u^A \circ f \uparrow Q^{A,\bullet} = f \uparrow P^{B,\bullet} \circ u^B \circ f \downarrow Q^{\bullet,B} \quad , \quad f \downarrow Q^{\bullet,A} \circ v^A \circ f \uparrow R^{A,\bullet} = f \uparrow Q^{B,\bullet} \circ v^B \circ f \downarrow R^{\bullet,B} \quad .$$

The composition $t \triangleq u \circ v$ has type signature $t : P^{A,A} \rightarrow R^{A,A}$, and so its naturality law is

$$f \downarrow P^{\bullet,A} \circ t^A \circ f \uparrow R^{A,\bullet} = f \uparrow P^{B,\bullet} \circ t^B \circ f \downarrow R^{\bullet,B} \quad . \quad (\text{D.10})$$

Can we derive that law by combining the naturality laws of u and v ? Note that *both* sides of the laws of u and v contain a lifting of f to Q , while Eq. (D.10) does not contain that lifting. The function $f : A \rightarrow B$ does not satisfy any known equations that we could use in the proof. So, we do not have any law that would allow us to eliminate $f \uparrow Q$ or $f \downarrow Q$ from the laws of u and v . (If, e.g., we knew that f is invertible, we would use a function $g : B \rightarrow A$ satisfying $f \circ g = \text{id}$ and $g \circ f = \text{id}$ and so eliminate f . But Eq. (D.10) must hold for arbitrary functions $f : A \rightarrow B$.) No matter how we combine those laws, starting from one side or from the other, the resulting expressions will always contain $f \uparrow Q$ and/or $f \downarrow Q$. Similarly, starting from any side of Eq. (D.10), we cannot *introduce* the function $f \uparrow Q$ or $f \downarrow Q$ into the expression. So, we are unable to obtain an expression in which we could use the laws of u and v .

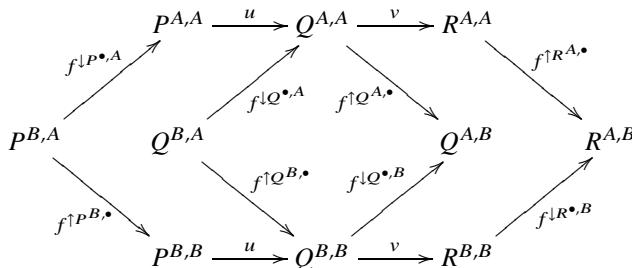


Figure D.1: Composition of dinatural transformations.

given only a “diagonal” transformation $P^{A,A} \rightarrow Q^{A,A}$ and not a function of type $P^{X,Y} \rightarrow Q^{X,Y}$ with arbitrary type parameters X, Y . So, we cannot prove in this way that the diagram commutes.

Nevertheless, it is true that the composition of dinatural transformations is dinatural — as long as the transformations are purely functional, which is always assumed by the parametricity theorem.

Statement D.2.2.2 Given any profunctors P, Q, R and any two purely functional dinatural transformations $u : P^{A,A} \rightarrow Q^{A,A}$ and $v : Q^{A,A} \rightarrow R^{A,A}$, the transformation $t \triangleq u \circ v$ is also dinatural.

The type diagram in Figure D.1 combines the diagrams for the laws of u and v in an attempt to derive the law of t . We can see a value of type $Q^{B,A}$ acting as a “source” of some arrows. It means that we are required to have a value of type $Q^{B,A}$ in order for the complete diagram to commute. But we are given only an arbitrary value $p : P^{B,A}$, and we are required to fill out the remainder of the diagram starting from that value. We cannot compute a value of type $Q^{B,A}$ from $p : P^{B,A}$ because we are

Proof We need to show that the naturality law holds for t : for any $p: P^{B,A}$ and $f: A \rightarrow B$,

$$p \triangleright f \downarrow P^{\bullet,A} ; t^A ; f \uparrow R^{\bullet,B} \stackrel{?}{=} p \triangleright f \uparrow P^{B,\bullet} ; t^B ; f \downarrow R^{\bullet,B} .$$

substitute $t \triangleq u ; v$: $p \triangleright f \downarrow P^{\bullet,A} ; u^A ; v^A ; f \uparrow R^{\bullet,B} \stackrel{?}{=} p \triangleright f \uparrow P^{B,\bullet} ; u^B ; v^B ; f \downarrow R^{\bullet,B} . \quad (\text{D.11})$

It is given that the naturality laws of u and v both hold: for any $p: P^{B,A}$ and $q: Q^{B,A}$,

$$p \triangleright f \downarrow P^{\bullet,A} ; u^A ; f \uparrow Q^{\bullet,A} = p \triangleright f \uparrow P^{B,\bullet} ; u^B ; f \downarrow Q^{\bullet,B} , \quad q \triangleright f \downarrow Q^{\bullet,A} ; v^A ; f \uparrow R^{\bullet,A} = q \triangleright f \uparrow Q^{B,\bullet} ; v^B ; f \downarrow R^{\bullet,B} .$$

We will be able to establish Eq. (D.11) only if we find a suitable value $q: Q^{B,A}$ that correctly fits the inner square of the type diagram in Figure D.1, which can be written as the equation

$$\begin{array}{ccc} & Q^{A,A} & \\ f \downarrow Q^{\bullet,A} \nearrow & & \searrow f \uparrow Q^{\bullet,A} \\ Q^{B,A} & & Q^{A,B} \\ \searrow f \uparrow Q^{B,\bullet} & & \nearrow f \downarrow Q^{\bullet,B} \\ & Q^{B,B} & \end{array}$$

$$q \triangleright f \downarrow Q^{\bullet,A} ; f \uparrow Q^{\bullet,A} = q \triangleright f \uparrow Q^{B,\bullet} ; f \downarrow Q^{\bullet,B} .$$

By the commutativity law of the profunctor Q (see Section D.1.1), this equation holds for *any* value $q: Q^{B,A}$. It remains to show that a suitable value q can be found such that we may replace $p \triangleright f \downarrow P^{\bullet,A} ; u^A$ by $q \triangleright f \downarrow Q^{\bullet,A}$ and $p \triangleright f \uparrow P^{B,\bullet} ; u^B$ by $q \triangleright f \uparrow Q^{B,\bullet}$ in Eq. (D.11). In other words, q must satisfy the equations

$$q \triangleright f \downarrow Q^{\bullet,A} = p \triangleright f \downarrow P^{\bullet,A} ; u^A , \quad q \triangleright f \uparrow Q^{B,\bullet} = p \triangleright f \uparrow P^{B,\bullet} ; u^B . \quad (\text{D.12})$$

We will complete the proof of law if we are able to compute a suitable value q for *any* given $f: A \rightarrow B$, $p: P^{B,A}$, and $u^A: P^{A,A} \rightarrow Q^{A,A}$. However, such q does not always exist without further assumptions. A simple counterexample is found by choosing the profunctor $Q^{X,Y} \triangleq X \rightarrow Y$ and setting

$$u^A \triangleq \underline{:P^{A,A}} \rightarrow \text{id}^{A \rightarrow A} .$$

With this choice of u , the value p is ignored, and so we can simplify Eq. (D.12) to

$$f^{A \rightarrow B} ; q^{B \rightarrow A} = \text{id}^{A \rightarrow A} , \quad q^{B \rightarrow A} ; f^{A \rightarrow B} = \text{id}^{B \rightarrow B} .$$

This is possible only if the function $q^{B \rightarrow A}$ is the inverse of f ; but, of course, not all functions f are invertible. Also, the existence of an inverse for a function $f: A \rightarrow B$ means that the types A and B are *equivalent* ($A \cong B$) due to the isomorphism f , which is clearly not true for arbitrary types A and B .

To proceed with the proof, we use a trick: we first prove the naturality law of t when restricted to isomorphic types $A \cong B$ and to invertible functions $f: A \rightarrow B$. Then we use the fact that a purely functional program t cannot use any type information about A and B or any information about the function f , and so the naturality law must actually hold for all types A, B and for all $f: A \rightarrow B$.

When the function $f: A \rightarrow B$ has an inverse $g: B \rightarrow A$, we may define q according to

$$q^{Q^{B,A}} \triangleq p^{P^{B,A}} \triangleright f \downarrow P^{\bullet,A} ; u^A ; g \downarrow Q^{\bullet,A} .$$

This will satisfy the first requirement in Eq. (D.12) because

$$\begin{aligned} \text{expect to equal } p \triangleright f \downarrow P^{\bullet,A} ; u^A : \quad & q \triangleright f \downarrow Q^{\bullet,A} = p^{P^{B,A}} \triangleright f \downarrow P^{\bullet,A} ; u^A ; g \downarrow Q^{\bullet,A} ; f \downarrow Q^{\bullet,A} \\ \text{composition under lifting :} \quad & = p^{P^{B,A}} \triangleright f \downarrow P^{\bullet,A} ; u^A ; (g ; f) \downarrow Q^{\bullet,A} \\ \text{use } g ; f = \text{id} : \quad & = p^{P^{B,A}} \triangleright f \downarrow P^{\bullet,A} ; u^A . \end{aligned}$$

The second of the requirements in Eq. (D.12) is then also satisfied:

$$\begin{aligned} \text{expect to equal } p \triangleright f \uparrow P^{B,\bullet} ; u^B : \quad & q \triangleright f \uparrow Q^{B,\bullet} = p^{P^{B,A}} \triangleright f \downarrow P^{\bullet,A} ; u^A ; g \downarrow Q^{\bullet,A} ; f \uparrow Q^{B,\bullet} \\ \text{commutativity law of } Q : \quad & = p^{P^{B,A}} \triangleright f \downarrow P^{\bullet,A} ; u^A ; f \uparrow Q^{A,\bullet} ; g \downarrow Q^{\bullet,B} \\ \text{naturality law of } u : \quad & = p \triangleright f \uparrow P^{B,\bullet} ; u^B ; f \downarrow Q^{\bullet,B} ; g \downarrow Q^{\bullet,B} \\ \text{composition under lifting :} \quad & = p \triangleright f \uparrow P^{B,\bullet} ; u^B ; (f ; g) \downarrow Q^{\bullet,B} = p \triangleright f \uparrow P^{B,\bullet} ; u^B . \end{aligned}$$

We can now substitute Eq. (D.12) into Eq. (D.11) and obtain

$$q \triangleright f^{\downarrow Q^{\bullet,A}} ; v^A ; f^{\uparrow R^{A,\bullet}} \stackrel{?}{=} q \triangleright f^{\uparrow Q^{B,\bullet}} ; v^B ; f^{\downarrow R^{\bullet,B}} .$$

This holds due to the given naturality law of v . So, we have shown that the naturality law (D.10) of t is satisfied for *invertible* functions $f^{A \rightarrow B}$.

It remains to remove the assumption of invertibility of $f^{A \rightarrow B}$. At this point, we need to use the fact that u and v are purely functional. Then the code of $t \triangleq u ; v$ is also purely functional. The naturality law of t has the form of Eq. (D.10). Both sides of that equation can be written out as some code that combines the code of t and the code of the required liftings of f . As we have seen in Chapter 6 and in the proof of the profunctor commutativity law (Section D.1.1), all lifting code is purely functional and uses the arbitrary function f only by applying f to some arguments. So, the code corresponding to both sides of Eq. (D.10) is purely functional and involves the function f as an unevaluated value (a “free variable”). We can symbolically write that code as the equation

$$a(f) = b(f) ,$$

where $a(f)$ and $b(f)$ are some expressions of type $P^{B,A} \rightarrow R^{A,B}$ that are built up from the nine purely functional constructions from Table D.2. It follows that the code of $a(f)$ and $b(f)$ may apply the given function f but cannot use the inverse function g assumed in the proof above. So, the only possibility for the equation $a(f) = b(f)$ to hold is when the programs $a(f)$ and $b(f)$ are *equal as symbolic expressions*⁷ up to equivalence transformations $g = x \rightarrow g(x)$ and $y \triangleright (x \rightarrow g(x)) = g(y)$. So, the equation $a(f) = b(f)$ must hold for arbitrary functions f and for arbitrary types A, B .

This concludes the proof of Statement D.2.2.2.

The proof of the main parametricity theorem is made shorter if we use the following property: If the type parameters of a natural transformation are substituted by some profunctors, we will obtain a dinatural transformation. Two versions of this property are proved next.

Statement D.2.2.3 Suppose $u^A : G^A \rightarrow H^A$ is a natural transformation, and suppose $P^{X,Y}$ is some purely functional profunctor. Define the transformation

$$t^A : G^{P^{A,A}} \rightarrow H^{P^{A,A}} , \quad t^A \triangleq u^{P^{A,A}} ,$$

by using the unmodified code of the function u^A with a substituted type parameter, $P^{A,A}$ instead of A . Then t is a dinatural transformation between profunctors $G^{P^{A,A}}$ and $H^{P^{A,A}}$.

Proof By assumption, the naturality law of u holds for any $k^{A \rightarrow B}$,

$$k^{\uparrow G} ; u^B = u^A ; k^{\uparrow H} .$$

The naturality law of t is an equality of functions of type $G^{P^{B,A}} \rightarrow H^{P^{A,B}}$ for an arbitrary $f^{A \rightarrow B}$:

$$(f^{\downarrow P^{\bullet,A}})^{\uparrow G} ; t^A ; (f^{\uparrow P^{\bullet,A}})^{\uparrow H} \stackrel{?}{=} (f^{\uparrow P^{B,\bullet}})^{\uparrow G} ; t^B ; (f^{\downarrow P^{\bullet,B}})^{\uparrow H} .$$

Begin the calculation with the left-hand side of that law:

$$\begin{aligned} \text{use } t^A = u^{P^{A,A}} : \quad & (f^{\downarrow P^{\bullet,A}})^{\uparrow G} ; \underline{t^A} ; (f^{\uparrow P^{A,\bullet}})^{\uparrow H} = (f^{\downarrow P^{\bullet,A}})^{\uparrow G} ; \underline{u^{P^{A,A}}} ; (f^{\uparrow P^{A,\bullet}})^{\uparrow H} \\ \text{naturality of } u : \quad & = u^{P^{B,A}} ; (f^{\downarrow P^{\bullet,A}})^{\uparrow H} ; (f^{\uparrow P^{A,\bullet}})^{\uparrow H} = u^{P^{B,A}} ; (f^{\downarrow P^{\bullet,A}} ; f^{\uparrow P^{A,\bullet}})^{\uparrow H} . \end{aligned}$$

Now write the right-hand side:

$$\begin{aligned} \text{use } t^B = u^{P^{B,B}} : \quad & (f^{\uparrow P^{B,\bullet}})^{\uparrow G} ; \underline{t^B} ; (f^{\downarrow P^{\bullet,B}})^{\uparrow H} = (f^{\uparrow P^{B,\bullet}})^{\uparrow G} ; \underline{u^{P^{B,B}}} ; (f^{\downarrow P^{\bullet,B}})^{\uparrow H} \\ \text{naturality of } u : \quad & = u^{P^{B,A}} ; (f^{\uparrow P^{B,\bullet}})^{\uparrow H} ; (f^{\downarrow P^{\bullet,B}})^{\uparrow H} = u^{P^{B,A}} ; (f^{\uparrow P^{B,\bullet}} ; f^{\downarrow P^{\bullet,B}})^{\uparrow H} . \end{aligned}$$

⁷This step in the proof seems plausible but is not rigorously derived in the present version of this book.

The difference between the left-hand side and the right-hand side is now only in the order of applying lifted functions f . By the profunctor commutativity law of P , we have

$$f^{\downarrow P^{\bullet,A}} ; f^{\uparrow P^{A,\bullet}} = f^{\uparrow P^{B,\bullet}} ; f^{\downarrow P^{\bullet,B}} ,$$

which completes the proof.

The following more general version of the same property will be also useful.

Statement D.2.2.4 Suppose $G^{X,Y}$ and $H^{X,Y}$ are some bifunctors, $P^{X,Y}$ and $Q^{X,Y}$ are some profunctors, all purely functional, and $u^{A,B} : G^{A,B} \rightarrow H^{A,B}$ is a natural transformation separately in the type parameters A and B . Define the transformation

$$t^A : G^{P^{A,A},Q^{A,A}} \rightarrow H^{P^{A,A},Q^{A,A}} , \quad t \triangleq u^{P^{A,A},Q^{A,A}} ,$$

by using the unmodified code of $u^{A,B}$ with substituted type parameters, $P^{A,A}$ and $Q^{A,A}$ instead of A and B . Then t is a dinatural transformation between profunctors $G^{P^{A,A},Q^{A,A}}$ and $H^{P^{A,A},Q^{A,A}}$.

Proof The assumed naturality law of u holds separately with respect to both type parameters,

$$f^{\uparrow G^{\bullet,B}} ; u^{B,B} = u^{A,B} ; f^{\uparrow H^{\bullet,B}} , \quad f^{\uparrow G^{A,\bullet}} ; u^{A,B} = u^{A,A} ; f^{\uparrow H^{A,\bullet}} .$$

The naturality law of t is written, after substituting $t = u$, as

$$\text{left-hand side : } (f^{\downarrow P})^{\uparrow G^{\bullet,Q^{B,A}}} ; (f^{\downarrow Q})^{\uparrow G^{P^{A,A},\bullet}} ; u^{P^{A,A},Q^{A,A}} ; (f^{\uparrow P})^{\uparrow H^{\bullet,Q^{A,A}}} ; (f^{\uparrow Q})^{\uparrow H^{P^{A,B},\bullet}}$$

$$\text{right-hand side : } \stackrel{?}{=} (f^{\uparrow P})^{\uparrow G^{\bullet,Q^{B,A}}} ; (f^{\uparrow Q})^{\uparrow G^{P^{B,B},\bullet}} ; u^{P^{B,B},Q^{B,B}} ; (f^{\downarrow P})^{\uparrow H^{\bullet,Q^{B,B}}} ; (f^{\downarrow Q})^{\uparrow H^{P^{A,B},\bullet}} .$$

The naturality law of u allows us to move all G -lifted functions to the right of u , where they become H -lifted. The law becomes

$$\text{left-hand side : } u^{P^{B,A},Q^{B,A}} ; (f^{\downarrow P})^{\uparrow H^{\bullet,Q^{B,A}}} ; (f^{\downarrow Q})^{\uparrow H^{P^{A,A},\bullet}} ; (f^{\uparrow P})^{\uparrow H^{\bullet,Q^{A,A}}} ; (f^{\uparrow Q})^{\uparrow H^{P^{A,B},\bullet}}$$

$$\text{right-hand side : } \stackrel{?}{=} u^{P^{B,A},Q^{B,A}} ; (f^{\uparrow P})^{\uparrow H^{\bullet,Q^{B,A}}} ; (f^{\uparrow Q})^{\uparrow H^{P^{B,B},\bullet}} ; (f^{\downarrow P})^{\uparrow H^{\bullet,Q^{B,B}}} ; (f^{\downarrow Q})^{\uparrow H^{P^{A,B},\bullet}} .$$

The remaining difference between the two sides is only in the application order of lifted functions f . All those liftings commute due to the profunctor and bifunctor commutativity laws.

D.2.3 Proof of the parametricity theorem

We will now prove that any purely functional transformation t^A expressed as

$$t^A : P^{A,A} \rightarrow Q^{A,A} , \quad t \triangleq p^{\cdot P^{A,A}} \rightarrow \text{expression} ,$$

satisfies the naturality law (D.4). It is assumed that the function body, denoted by “expression”, is some purely functional code built up from the nine code constructions of Table D.2.

The proof goes by induction on the structure of “expression”. The base case contains the first two constructions (“use unit value” and “use argument”) since those constructions do not assume any previous code. We will need to prove that the law (D.4) holds in those cases. The inductive step covers the remaining seven constructions, which create larger code expressions from smaller ones. We will need to prove that the code for t satisfies the law (D.4) under the inductive assumption that all previous code expressions already satisfy their versions of that law.

For example, the product construction (“create tuple”) combines two previously available code expressions (a and b) into a tuple $a \times b$. The corresponding transformations are $p \rightarrow a$, $p \rightarrow b$, and $p \rightarrow a \times b$. Note that the naturality laws have different forms for each of the transformations $p \rightarrow a$, $p \rightarrow b$, and $p \rightarrow a \times b$ because those functions have different types. Assuming that the profunctor transformations $p \rightarrow a$ and $p \rightarrow b$ already satisfy their respective naturality laws, we will need to prove that $p \rightarrow a \times b$ also satisfies the corresponding law. (The proof is in item 5 below.)

We will now prove the naturality law (D.4) for each of the nine code constructions.

1) Use unit value This is a base case where “expression” is just the unit value, 1. The code of t is a constant function that always returns that value: $t \triangleq p^{:P^{A,A}} \rightarrow 1$. In that case, $Q^{X,Y} \triangleq 1$ and so the liftings for Q consist of identity functions, $f \downarrow Q^{*,B} = \text{id}$ and $f \uparrow Q^{A,*} = \text{id}$. We can also write $t = _ \rightarrow 1$. Both sides of the law (D.4) are then functions that ignore their arguments and always return 1.

A quicker way of proving the dinaturality of t is to use Statement D.2.2.3 with $G^A \triangleq A$ and $H^A \triangleq 1$.

2) Use argument This is a base case where “expression” is just the given argument of the transformation, i.e., the code of t is $t \triangleq p \rightarrow p$. Since t is an identity function, it has type $t^A : P^{A,A} \rightarrow P^{A,A}$, and so we must have $P = Q$. The naturality law follows from Statement D.2.2.3 with $G^A \triangleq H^A \triangleq A$.

The rest of the proof goes over the remaining code constructions, which are inductive steps.

3) Create function The “expression” is a nameless function $z \rightarrow r$, so $t \triangleq p^{:P^{A,A}} \rightarrow z \rightarrow r$, where the sub-expression r may use both p and z as bound variables. Since the code of r is purely functional, the types of r and z must be expressible as $R^{A,A}$ and $Z^{A,A}$, where $R^{X,Y}$ and $Z^{X,Y}$ are some profunctors. The form of t implies that $Q^{X,Y} \triangleq Z^{Y,X} \rightarrow R^{X,Y}$ (note the reverse order of parameters in $Z^{Y,X}$). The inductive assumption says that the naturality law is already satisfied by *any* transformation whose “expression” is just r . Such a transformation must have code of the form $u \triangleq s \rightarrow r$, where the type of the argument s must contain the bound variables p and z that may be used in r . So, we set $s = p \times z$ and write the code of u as $u = p^{:P^{A,A}} \times z^{:Z^{A,A}} \rightarrow r$. We then have

$$t = p^{:P^{A,A}} \rightarrow z^{:Z^{A,A}} \rightarrow u(p \times z) , \quad u = p^{:P^{A,A}} \times z^{:Z^{A,A}} \rightarrow t(p)(z) . \quad (\text{D.13})$$

Denoting the profunctor $U^{X,Y} \triangleq P^{X,Y} \times Z^{X,Y}$, we write the naturality law for $u : U^{A,A} \rightarrow R^{A,A}$ as

$$f \downarrow U^{*,A} ; u^A ; f \uparrow R^{A,*} = f \downarrow U^{B,*} ; u^B ; f \uparrow R^{*,B} . \quad (\text{D.14})$$

This equation holds by the inductive assumption. We will derive the naturality law (D.4) for t from Eq. (D.14) by using the type definitions $U^{X,Y} \triangleq P^{X,Y} \times Z^{X,Y}$ and $Q^{X,Y} \triangleq Z^{Y,X} \rightarrow R^{X,Y}$ to express the liftings of U and Q through the liftings of P , R , and Z via the standard functor/contrafunctor codes:

$$f \downarrow U^{*,A} = f \downarrow P^{*,A} \boxtimes f \downarrow Z^{*,A} , \quad f \uparrow U^{B,*} = f \uparrow P^{B,*} \boxtimes f \uparrow Z^{B,*} , \quad (\text{D.15})$$

$$f \uparrow Q^{A,*} = q^{:Q^{A,A}} \rightarrow f \downarrow Z^{*,A} ; q ; f \uparrow R^{A,*} , \quad f \downarrow Q^{B,B} = q^{:Q^{B,B}} \rightarrow f \downarrow Z^{B,*} ; q ; f \downarrow R^{*,B} . \quad (\text{D.16})$$

Substituting the relevant formulas into the left-hand side of Eq. (D.4) and applying to $p^{:P^{B,A}}$, we find

$$\begin{aligned} \text{Eq. (D.13)} : \quad & p \triangleright f \downarrow P^{*,A} ; t ; f \uparrow Q^{A,*} = p \triangleright f \downarrow P^{*,A} \triangleright (p \rightarrow z \rightarrow u(p \times z)) \triangleright (q^{:Q^{A,A}} \rightarrow f \downarrow Z^{*,A} ; q ; f \uparrow R^{A,*}) \\ & = f \downarrow Z^{*,A} ; (z \rightarrow u((p \triangleright f \downarrow P^{*,A}) \times z)) ; f \uparrow R^{A,*} = z \rightarrow u((p \triangleright f \downarrow P^{*,A}) \times (z \triangleright f \downarrow Z^{*,A})) \triangleright f \uparrow R^{A,*} \\ & = z \rightarrow u((p \times z) \triangleright (f \downarrow P^{*,A} \boxtimes f \downarrow Z^{*,A})) \triangleright f \uparrow R^{A,*} = z \rightarrow (p \times z) \triangleright (f \downarrow P^{*,A} \boxtimes f \downarrow Z^{*,A}) \triangleright u \triangleright f \uparrow R^{A,*} \end{aligned}$$

$$\text{Eq. (D.15)} : \quad = z \rightarrow (p \times z) \triangleright f \downarrow U^{*,A} ; u ; f \uparrow R^{A,*}$$

$$\text{Eq. (D.14)} : \quad = z \rightarrow (p \times z) \triangleright f \downarrow U^{B,*} ; u ; f \downarrow R^{*,B}$$

$$\begin{aligned} \text{Eq. (D.15)} : \quad & = z \rightarrow (p \times z) \triangleright (f \uparrow P^{B,*} \boxtimes f \uparrow Z^{B,*}) \triangleright (p \times z \rightarrow t(p)(z)) \triangleright f \downarrow R^{*,B} \\ & = z \rightarrow t(p \triangleright f \uparrow P^{B,*})(z \triangleright f \uparrow Z^{B,*}) \triangleright f \downarrow R^{*,B} . \end{aligned}$$

Now apply the right-hand side of Eq. (D.4) to the same value p :

$$\begin{aligned} \text{use Eq. (D.16)} : \quad & p \triangleright f \uparrow P^{B,*} ; t ; f \downarrow Q^{*,B} = p \triangleright f \uparrow P^{B,*} \triangleright t \triangleright (q^{:Q^{B,B}} \rightarrow f \uparrow Z^{B,*} ; q ; f \downarrow R^{*,B}) \\ & = f \uparrow Z^{B,*} ; t(p \triangleright f \uparrow P^{B,*}) ; f \downarrow R^{*,B} = z \rightarrow z \triangleright f \uparrow Z^{B,*} \triangleright t(p \triangleright f \uparrow P^{B,*}) \triangleright f \downarrow R^{*,B} \\ & = z \rightarrow t(p \triangleright f \uparrow P^{B,*})(z \triangleright f \uparrow Z^{B,*}) \triangleright f \downarrow R^{*,B} . \end{aligned}$$

We obtain the same expression as for the left-hand side, which proves that the law holds.

The proof does not assume that the expression $t(p)(r)$ actually uses both arguments p and r . So, the law holds also for functions that ignore some of their arguments.

4) Use function The “expression” is a function application such as $f(z)$. Then we may write the code of t as $t \triangleq p : P^{A,A} \rightarrow f(z)$, where f and z are some other expressions. Since these expressions are parts of the code $p \rightarrow f(z)$, it must be possible to compute f and z separately, given p . So, we can write purely functional code for the two transformations $u \triangleq p \rightarrow f$ and $v \triangleq p \rightarrow z$. The inductive assumption is that the naturality law holds for any transformations whose code contains f or z as the function body. So, we may use the naturality laws for u and v .

What are the possible types of f and z ? Since z is computed by purely functional code, the type of z can be expressed as $z : Z^{A,A}$ via some profunctor $Z^{X,Y}$. To match the types, the function f must have type $Z^{A,A} \rightarrow Q^{A,A}$. It will be convenient to express t as a composition of two transformations: the first, denoted by s , simply gathers f and z in a tuple,

$$s : P^{A,A} \rightarrow (Z^{A,A} \rightarrow Q^{A,A}) \times Z^{A,A} \quad , \quad s \triangleq p \rightarrow u(p) \times v(p) \quad .$$

The second, denoted w , applies the function:

$$w : (Z^{A,A} \rightarrow Q^{A,A}) \times Z^{A,A} \rightarrow Q^{A,A} \quad , \quad w \triangleq f \times z \rightarrow f(z) \quad .$$

We will show separately that s and w are dinatural. Since both s and w are implemented via purely functional code, it will follow by Statement D.2.2.2 that the composition $t = s ; w$ is also dinatural.

Defining the profunctor $G^{X,Y} \triangleq Z^{Y,X} \rightarrow Q^{X,Y}$, we write the type signatures of u and v as

$$u : P^{A,A} \rightarrow G^{A,A} \quad , \quad v : P^{A,A} \rightarrow Z^{A,A} \quad .$$

Since u and v are dinatural, we may apply the derivation in item 5 below (which does not use any other assumptions) with $R \triangleq G$ and $S \triangleq Z$ to show that the transformation $s \triangleq p \rightarrow u(p) \times v(p)$ of type $P^{A,A} \rightarrow G^{A,A} \times Z^{A,A}$ is also dinatural.

It remains to show that the transformation $w \triangleq g \times z \rightarrow z \triangleright g$ is dinatural. For brevity, we will omit type annotations and write $f^{\uparrow G}$ instead of $f^{\uparrow G^{X,Y}}$ and $f^{\downarrow G}$ instead of $f^{\downarrow G^{X,Y}}$, with G and other profunctors. Since profunctors have one contravariant and one covariant type parameter, the choice of the type parameter in a lifting will remain unambiguous. The naturality law is then written as

$$(f^{\downarrow G} \boxtimes f^{\uparrow Z}) ; w ; f^{\uparrow Q} \stackrel{?}{=} (f^{\uparrow G} \boxtimes f^{\uparrow Z}) ; w ; f^{\downarrow Q} \quad . \quad (\text{D.17})$$

It helps to write out the liftings to G . For arbitrary $g : G^{B,A}$ and $f : A \rightarrow B$, we have $g \triangleright f^{\downarrow G^{B,A}}$ of type $G^{A,A} = Z^{A,A} \rightarrow Q^{A,A}$, so

$$g : G^{B,A} \triangleright f^{\downarrow G} = z : Z^{A,A} \rightarrow (z \triangleright f^{\uparrow Z}) \triangleright g ; f^{\downarrow Q} = f^{\uparrow Z} ; g ; f^{\downarrow Q} \quad .$$

Similarly

$$g : G^{B,A} \triangleright f^{\uparrow G} = z : Z^{B,B} \rightarrow (z \triangleright f^{\downarrow Z}) \triangleright g ; f^{\uparrow Q} = f^{\downarrow Z} ; g ; f^{\uparrow Q} \quad .$$

Now apply the left-hand side of the naturality law (D.17) to an arbitrary $g : G^{B,A} \times z : Z^{B,A}$:

$$\begin{aligned} \text{definitions of } \boxtimes \text{ and } w : \quad & (g : G^{B,A} \times z : Z^{B,A}) \triangleright (f^{\downarrow G} \boxtimes f^{\uparrow Z}) ; w ; f^{\uparrow Q} = (z \triangleright f^{\uparrow Z}) \triangleright (g \triangleright f^{\downarrow G}) ; f^{\uparrow Q} \\ \text{definition of } g \triangleright f^{\downarrow G} : \quad & = (z \triangleright f^{\uparrow Z}) \triangleright f^{\uparrow Z} ; g ; f^{\downarrow Q} ; f^{\uparrow Q} = z \triangleright f^{\downarrow Z} ; f^{\uparrow Z} ; g ; f^{\downarrow Q} ; f^{\uparrow Q} \quad . \end{aligned}$$

Applying the right-hand side of Eq. (D.17) to $g \times z$ gives similarly

$$\begin{aligned} \text{definitions of } \boxtimes \text{ and } w : \quad & (g : G^{B,A} \times z : Z^{B,A}) \triangleright (f^{\uparrow G} \boxtimes f^{\uparrow Z}) ; w ; f^{\downarrow Q} = (z \triangleright f^{\uparrow Z}) \triangleright (g \triangleright f^{\uparrow G}) ; f^{\downarrow Q} \\ \text{definition of } g \triangleright f^{\uparrow G} : \quad & = (z \triangleright f^{\uparrow Z}) \triangleright f^{\downarrow Z} ; g ; f^{\uparrow Q} ; f^{\downarrow Q} = z \triangleright f^{\uparrow Z} ; f^{\downarrow Z} ; g ; f^{\uparrow Q} ; f^{\downarrow Q} \quad . \end{aligned}$$

The two sides of Eq. (D.17) now differ only by the order of application of lifted functions $f^{\uparrow Z}$, $f^{\downarrow Z}$, $f^{\uparrow Q}$, and $f^{\downarrow Q}$. These applications commute by the profunctor commutativity laws of Z and Q :

$$f^{\downarrow Z} ; f^{\uparrow Z} = f^{\uparrow Z} ; f^{\downarrow Z} \quad , \quad f^{\downarrow Q} ; f^{\uparrow Q} = f^{\uparrow Q} ; f^{\downarrow Q} \quad .$$

This concludes the proof of Eq. (D.17) and so proves the entire item 4.

5) Create tuple The “expression” is a tuple, so the code of t is of the form $t \triangleq p : P^{A,A} \rightarrow r \times s$ where r and s are some expressions. In general, the types of r and s will be given by some profunctors R and S , so that $Q^{X,Y} = R^{X,Y} \times S^{X,Y}$. Since the code of t computes $r \times s$ from p , we should be able to compute r and s separately from p . It means that we have well-defined transformations $u \triangleq p \rightarrow r$ and $v \triangleq p \rightarrow s$ having types $u : P^{A,A} \rightarrow R^{A,A}$ and $v : P^{A,A} \rightarrow S^{A,A}$. The inductive assumption is that the naturality law already holds for any transformation whose function body is r or s . So, we may assume that the law holds separately for u and v :

$$f^{\downarrow P} ; u ; f^{\uparrow R} = f^{\uparrow P} ; u ; f^{\downarrow R} \quad , \quad f^{\downarrow P} ; v ; f^{\uparrow S} = f^{\uparrow P} ; v ; f^{\downarrow S} \quad .$$

The lifting for Q is expressed through the liftings for R and S as

$$f^{\uparrow Q} = f^{\uparrow R} \boxtimes f^{\uparrow S} \quad , \quad f^{\downarrow Q} = f^{\downarrow R} \boxtimes f^{\downarrow S} \quad .$$

We can now verify the naturality law of t by expressing $t = p \rightarrow u(p) \times v(p) = \Delta ; (u \boxtimes v)$:

expect $f^{\uparrow P} ; t ; f^{\downarrow Q} : f^{\downarrow P} ; \underline{L} ; f^{\uparrow Q} = f^{\downarrow P} ; \Delta ; (u \boxtimes v) ; (f^{\uparrow R} \boxtimes f^{\uparrow S})$

naturality of Δ : $= \Delta ; (f^{\downarrow P} \boxtimes f^{\downarrow P}) ; (u \boxtimes v) ; (f^{\uparrow R} \boxtimes f^{\uparrow S}) = \Delta ; (\underline{f^{\downarrow P} ; u ; f^{\uparrow R}}) \boxtimes (\underline{f^{\downarrow P} ; v ; f^{\uparrow S}})$

inductive assumption : $= \Delta ; (f^{\uparrow P} ; u ; f^{\downarrow R}) \boxtimes (f^{\uparrow P} ; v ; f^{\downarrow S}) = \underline{\Delta ; (f^{\uparrow P} \boxtimes f^{\uparrow P})} ; (u \boxtimes v) ; (\underline{f^{\downarrow R} \boxtimes f^{\downarrow S}})$

naturality of Δ : $= f^{\uparrow P} ; \underline{\Delta ; (u \boxtimes v)} ; f^{\downarrow Q} = f^{\uparrow P} ; t ; f^{\downarrow Q} \quad .$

6) Use tuple The “expression” contains a tuple accessor, such as π_1 . It is sufficient to prove the law for π_1 , since all tuple accessors work similarly. So, we assume that the type P is a product, $P^{X,Y} \triangleq Q^{X,Y} \times R^{X,Y}$, and that the code of t is of the form $t \triangleq \pi_1 = p : Q^{A,A} \times R^{A,A} \rightarrow p \triangleright \pi_1$. This is just the natural transformation π_1 with substituted type parameters, so the dinaturality of t follows from Statement D.2.2.4 with $G^{X,Y} \triangleq X \times Y$, $H^{X,Y} \triangleq X$, $P \triangleq Q$, and $Q \triangleq R$.

7) Create disjunctive value The “expression” contains a case class constructor such as `Left`, `Right`, or `Some`. We may assume that $Q^{X,Y} = R^{X,Y} + S^{X,Y}$ for some profunctors R , S , and that the code of t is of the form $t \triangleq p : P^{A,A} \rightarrow r : R^{A,A} + \mathbb{0} : S^{A,A}$ where r is some expression of type $R^{A,A}$. It is sufficient to prove the law for $t = p \rightarrow r + \mathbb{0}$, because all other disjunctive cases such as $t = p \rightarrow \mathbb{0} + s$ work analogously.

Since the code of t is purely functional, the value r can be also computed from p using purely functional code. So the inductive assumption is that the transformation $u \triangleq p \rightarrow r$ is already known to be dinatural, with the law

$$f^{\downarrow P} ; u ; f^{\uparrow R} = f^{\uparrow P} ; u ; f^{\downarrow R} \quad .$$

The naturality law for t is

$$f^{\downarrow P} ; t ; f^{\uparrow Q} \stackrel{?}{=} f^{\uparrow P} ; t ; f^{\downarrow Q} \quad .$$

The liftings to Q are disjunctive functions written in the matrix notation as

$$f^{\uparrow Q^{B,\bullet}} = \left| \begin{array}{c|cc} & R^{B,B} & S^{B,B} \\ \hline R^{B,A} & f^{\uparrow R^{B,\bullet}} & \mathbb{0} \\ S^{B,A} & \mathbb{0} & f^{\uparrow S^{B,\bullet}} \end{array} \right| \quad , \quad f^{\downarrow Q^{\bullet,A}} = \left| \begin{array}{c|cc} & R^{A,A} & S^{A,A} \\ \hline R^{B,A} & f^{\downarrow R^{\bullet,A}} & \mathbb{0} \\ S^{B,A} & \mathbb{0} & f^{\downarrow S^{\bullet,A}} \end{array} \right| \quad .$$

Omitting type annotations, we apply the naturality law to a value $p^{P^{B,A}}$ and get

$$\begin{aligned}
 \text{left-hand side : } & p \triangleright f^{\downarrow P} ; t ; f^{\uparrow Q} = ((p \triangleright f^{\downarrow P} \triangleright u) + \mathbb{0}) \triangleright \begin{vmatrix} f^{\uparrow R} & \mathbb{0} \\ \mathbb{0} & f^{\downarrow S} \end{vmatrix} \\
 \text{apply disjunctive function : } & = p \triangleright f^{\downarrow P} ; u ; f^{\uparrow R} \quad . \\
 \text{right-hand side : } & p \triangleright f^{\uparrow P} ; t ; f^{\downarrow Q} = ((p \triangleright f^{\uparrow P} \triangleright u) + \mathbb{0}) \triangleright \begin{vmatrix} f^{\downarrow R} & \mathbb{0} \\ \mathbb{0} & f^{\downarrow S} \end{vmatrix} \\
 \text{apply disjunctive function : } & = p \triangleright \underline{f^{\uparrow P} ; u ; f^{\downarrow R}} \\
 \text{dinaturality of } u : & = p \triangleright f^{\downarrow P} ; u ; f^{\uparrow R} \quad .
 \end{aligned}$$

Both sides are now equal.

8) Use disjunctive value The “expression” is a pattern match, so the code of t is of the form

$$t \triangleq p^{P^{A,A}} \rightarrow e \triangleright \begin{vmatrix} g \\ h \end{vmatrix} \quad ,$$

where g and h are known functions. The expressions e , g , and h are used as part of the code, and so they must be all purely functional and computable from p . We then choose the inductive assumption that the transformations $p \rightarrow e$, $p \rightarrow g$, $p \rightarrow h$ are dinatural. As in item 4, we can represent t as a composition of two transformations,

$$t = u ; v \quad , \quad u \triangleq p \rightarrow e \times g \times h \quad , \quad v \triangleq e \times g \times h \rightarrow e \triangleright \begin{vmatrix} g \\ h \end{vmatrix} \quad .$$

To show that the transformation u is dinatural, we apply the argument in item 5 twice. It remains to show that v is dinatural. Since the code of v is purely functional, it will follow by Statement D.2.2.2 that the composition $t = u ; v$ is dinatural.

Assume that the types of e , g , h are $e : R^{A,A} + S^{A,A}$, $g : R^{A,A} \rightarrow Q^{A,A}$, and $h : S^{A,A} \rightarrow Q^{A,A}$, where $R^{X,Y}$ and $S^{X,Y}$ are some profunctors. For convenience, let us define the profunctors K and L by

$$K^{X,Y} \triangleq R^{Y,X} \rightarrow Q^{X,Y} \quad , \quad L^{X,Y} \triangleq S^{Y,X} \rightarrow Q^{X,Y} \quad .$$

The type of v is then $v^A : (R^{A,A} + S^{A,A}) \times K^{A,A} \times L^{A,A} \rightarrow Q^{A,A}$, so its naturality law is

$$(f^{\downarrow(R+S)} \boxtimes f^{\downarrow K} \boxtimes f^{\downarrow L}) ; v ; f^{\uparrow Q} \stackrel{?}{=} (f^{\uparrow(R+S)} \boxtimes f^{\uparrow K} \boxtimes f^{\uparrow L}) ; v ; f^{\downarrow Q} \quad .$$

The code of the required liftings is defined by

$$\begin{aligned}
 f^{\downarrow(R+S)} &= \begin{vmatrix} f^{\downarrow R} & \mathbb{0} \\ \mathbb{0} & f^{\downarrow S} \end{vmatrix} \quad , \quad f^{\uparrow(R+S)} = \begin{vmatrix} f^{\uparrow R} & \mathbb{0} \\ \mathbb{0} & f^{\uparrow S} \end{vmatrix} \quad , \\
 g \triangleright f^{\downarrow K} &= f^{\uparrow R} ; g ; f^{\downarrow Q} \quad , \quad g \triangleright f^{\uparrow K} = f^{\downarrow R} ; g ; f^{\uparrow Q} \quad , \\
 h \triangleright f^{\downarrow L} &= f^{\uparrow S} ; h ; f^{\downarrow Q} \quad , \quad h \triangleright f^{\uparrow L} = f^{\downarrow S} ; h ; f^{\uparrow Q} \quad .
 \end{aligned}$$

Apply the left-hand side of the naturality law to a value $e^{R^{B,A} + S^{B,A}} \times g^{R^{A,B} \rightarrow Q^{B,A}} \times h^{S^{A,B} \rightarrow Q^{B,A}}$:

$$\begin{aligned}
 \text{expand } \boxtimes : & e \times g \times h \triangleright (f^{\downarrow(R+S)} \boxtimes f^{\downarrow K} \boxtimes f^{\downarrow L}) ; v ; f^{\uparrow Q} = (e \triangleright f^{\downarrow(R+S)}) \times (g \triangleright f^{\downarrow K}) \times (h \triangleright f^{\downarrow L}) \triangleright v ; f^{\uparrow Q} \\
 \text{definition of } v : & = e \triangleright \begin{vmatrix} f^{\downarrow R} & \mathbb{0} \\ \mathbb{0} & f^{\downarrow S} \end{vmatrix} ; \begin{vmatrix} g \triangleright f^{\downarrow K} \\ h \triangleright f^{\downarrow L} \end{vmatrix} ; f^{\uparrow Q} \\
 \text{composition : } & = e \triangleright \begin{vmatrix} f^{\downarrow R} ; f^{\uparrow R} ; g ; f^{\downarrow Q} \\ f^{\downarrow S} ; f^{\uparrow S} ; h ; f^{\downarrow Q} \end{vmatrix} ; f^{\uparrow Q} = e \triangleright \begin{vmatrix} f^{\downarrow R} ; f^{\uparrow R} ; g ; f^{\downarrow Q} ; f^{\uparrow Q} \\ f^{\downarrow S} ; f^{\uparrow S} ; h ; f^{\downarrow Q} ; f^{\uparrow Q} \end{vmatrix} \quad .
 \end{aligned}$$

Apply the right-hand side to the same value:

$$\begin{aligned}
 \text{expand } \boxtimes : \quad & e \times g \times h \triangleright (f^{\uparrow(R+S)} \boxtimes f^{\uparrow K} \boxtimes f^{\uparrow L}) ; v ; f^{\downarrow Q} = (e \triangleright f^{\uparrow(R+S)}) \times (g \triangleright f^{\uparrow K}) \times (h \triangleright f^{\uparrow L}) \triangleright v ; f^{\downarrow Q} \\
 \text{definition of } v : \quad & = e \triangleright \left\| \begin{array}{cc} f^{\uparrow R} & 0 \\ 0 & f^{\uparrow S} \end{array} \right\| ; \left\| \begin{array}{c} g \triangleright f^{\uparrow K} \\ h \triangleright f^{\uparrow L} \end{array} \right\| ; f^{\downarrow Q} \\
 \text{composition :} \quad & = e \triangleright \left\| \begin{array}{c} f^{\uparrow R} ; f^{\downarrow R} ; g ; f^{\uparrow Q} \\ f^{\uparrow S} ; f^{\downarrow S} ; h ; f^{\uparrow Q} \end{array} \right\| ; \left\| \begin{array}{c} f^{\uparrow R} ; f^{\downarrow R} ; g ; f^{\uparrow Q} ; f^{\downarrow Q} \\ f^{\uparrow S} ; f^{\downarrow S} ; h ; f^{\uparrow Q} ; f^{\downarrow Q} \end{array} \right\| .
 \end{aligned}$$

The two sides of the naturality law now differ only by the application order of lifted functions f . Applying the profunctor commutativity law of R , S , and Q , we find that the two sides of the naturality law are equal.

9) Use recursion Here the “expression” is the value of the function t itself, so the code of t is $t \triangleq p : P^{A,A} \rightarrow \bar{t}$ (the overline in \bar{t} denotes the recursive use of t). We treat \bar{t} as just code for some function that is already known (by the inductive assumption) to obey its naturality law. Assume that the type of t is $P^{A,A} \rightarrow Q^{A,A}$; then the recursive invocation \bar{t} has the type $Q^{A,A}$ (which is a recursive type assumed to be equivalent to $P^{A,A} \rightarrow Q^{A,A}$). The naturality law of t applied to an arbitrary $p : P^{B,A}$ is

$$p \triangleright f^{\downarrow P} ; t ; f^{\uparrow Q} \stackrel{?}{=} p \triangleright f^{\uparrow P} ; t ; f^{\downarrow Q} .$$

Substituting $t = p \rightarrow \bar{t}$, we find

$$p \triangleright f^{\downarrow P} \triangleright (p \rightarrow \bar{t}) \triangleright f^{\uparrow Q} = \bar{t} \triangleright f^{\uparrow Q} \stackrel{?}{=} p \triangleright f^{\uparrow P} \triangleright (p \rightarrow \bar{t}) \triangleright f^{\downarrow Q} = \bar{t} \triangleright f^{\downarrow Q} .$$

It remains to show that

$$\bar{t} \triangleright f^{\uparrow Q} \stackrel{?}{=} \bar{t} \triangleright f^{\downarrow Q} .$$

This is the “wedge law” of \bar{t} , which follows from the assumed dinaturality of \bar{t} by Statement D.2.1.3.

This concludes the proof of what is commonly known as “the” parametricity theorem: Any purely functional code with a type parameter will automatically satisfy a naturality law. If a function has several type parameters, we can fix all the type parameters except one and apply the parametricity theorem separately, obtaining one naturality law for each type parameter.

D.2.4 Uniqueness of functor and contrafunctor typeclass instances

An important consequence of the parametricity theorem is the fact that functors and contrafunctors can be implemented in only one way.

Statement D.2.4.1 A **purely functional functor**, i.e., a functor F whose type is a combination of the constructions of Table D.1, has a unique lawful and purely functional implementation of `fmap`.

Proof Section 6.2.3 derived lawful and purely functional implementations of the `fmap` method for all functors F build up from the six type constructions. The naturality laws obtained from the parametricity theorem must use precisely those “standard” implementations of `fmap`, because the proof of the parametricity theorem significantly depends on the code of those implementations. Throughout this book, the standard lifting code is denoted by $\text{fmap}_F(f)$ or by $f^{\uparrow F}$. Now suppose that there exists *another* lawful and purely functional implementation of `fmap` for F , denoted by $\text{fmap}'_F(f)$:

$$\text{fmap}'_F : (A \rightarrow B) \rightarrow F^A \rightarrow F^B , \quad \text{fmap}'_F(f : A \rightarrow B) = ???^{F^A \rightarrow F^B} .$$

We will now show that $\text{fmap}'_F = \text{fmap}_F$. Let us fix the type parameter A and apply the parametricity theorem to fmap'_F with respect to B . The resulting naturality law involves an arbitrary $g : B \rightarrow C$:

$$\text{fmap}'_F(f : A \rightarrow B) ; g : B \rightarrow C \stackrel{!}{=} \text{fmap}'_F(f) ; g^{\uparrow F} .$$

Within the naturality law, the lifting $g^{\uparrow F}$ must use the “standard” lifting code $g^{\uparrow F} \triangleq \text{fmap}_F(g)$. By assumption, fmap'_F is lawful, so we may use its composition law and write

$$\text{fmap}'_F(f \circ g) = \text{fmap}'_F(f) \circ \text{fmap}'_F(g) \stackrel{!}{=} \text{fmap}'_F(f) \circ g^{\uparrow F} \quad .$$

Since $f: A \rightarrow B$ is arbitrary, we can choose $A = B$ and $f = \text{id}: B \rightarrow B$ to obtain

$$\text{fmap}'_F(\text{id}) \circ \text{fmap}'_F(g) \stackrel{!}{=} \text{fmap}'_F(\text{id}) \circ g^{\uparrow F} \quad .$$

The identity law for fmap'_F gives $\text{fmap}'_F(\text{id}) = \text{id}$, so we can simplify the last equation to

$$\text{fmap}'_F(g) \stackrel{!}{=} g^{\uparrow F} = \text{fmap}_F(g) \quad .$$

This must hold for arbitrary $g: B \rightarrow C$, which proves that $\text{fmap}'_F = \text{fmap}_F$.

Statement D.2.4.2 A contrafunctor H whose type expression is a combination of the six type constructions (Table D.1) has a unique purely functional implementation of a lawful `cmap` method.

Proof We use similar arguments as in the proof of Statement D.2.4.1. For any lawful, purely functional alternative implementation cmap'_H , the parametricity theorem gives the naturality law

$$\text{cmap}'_H(f: A \rightarrow B \circ g: B \rightarrow C) \stackrel{!}{=} (g: B \rightarrow C) \downarrow H \circ \text{cmap}'_H(f) \quad .$$

By assumption, the identity and composition law hold for cmap'_H . Setting $f = \text{id}: B \rightarrow B$, we get

$$\text{cmap}'_H(\text{id} \circ g) = \text{cmap}'_H(g) \stackrel{!}{=} g \downarrow H \circ \text{cmap}'_H(\text{id}) = g \downarrow H \quad .$$

This must hold for arbitrary $g: B \rightarrow C$, which shows that $\text{cmap}'_H(g) = g \downarrow H = \text{cmap}_H(g)$ as required.

D.3 Summary

We have proved three parametricity results that apply to all purely functional programs:

- The lifting methods of any bifunctor, profunctor, or bi-contrafunctor obey the commutativity law such as Eq. (6.9). Because of this, any purely functional type constructor $F^{A,B}$ which is a functor separately with respect to A and B is always a bifunctor whose `bimap` method satisfies the composition law (6.10). Similar properties hold for profunctors and for bi-contrafunctors. The proof goes by induction on the exponential-polynomial type expression of $F^{A,B}$, which must be built up via the six type constructions (Table D.1).
- Any function of type $P^{A,A} \rightarrow Q^{A,A}$ (where P, Q are profunctors) obeys the general naturality law (D.4), which allows us to derive a specific naturality law for any fully parametric function. The form of the law depends only on the function’s type signature and applies to all purely functional implementations of that type signature. The proof goes by induction on the structure of the expression, which must be built up via the nine code constructions (Table D.2).
- The lifting methods of functors and contrafunctors can be implemented in only one way once the identity and composition laws are imposed. The unique correct implementations are defined by the standard procedures shown in Chapter 6, and there are no other inequivalent implementations. Here we do not distinguish *equivalent* implementations such as $f(x)$ and $(y \rightarrow y)(f)(x)$, which are syntactically different programs that will always give the same results. So, there is only one lawful implementation of the `Functor` or `Contrafunctor` typeclass instances for a given type constructor. (For many other typeclasses, such as `Filterable` or `Monad`, a given type constructor may have several inequivalent and lawful typeclass instances.)

E A humorous disclaimer

The following text is quoted in part from an anonymous online source ("Project Guten Tag") dating back at least to 1997. The original text is no longer available on the Internet.

WARRANTO LIMITENSIS; DISCLAMATANTUS DAMAGENSIS
Sonus exceptus "Rectum Replacator Refundiens" describitus ecci,

1. Projectus (etque nunquam partum quis hic etext remitibus cum PROJECT GUTEN TAG™ identifier) disclamabat omni liabilitus tuus damagensis, pecuniensisque, includibantus pecunia legalitus, et
2. REMEDIA NEGLIGENTITIA NON HABET TUUS, WARRANTUS DESTRUCTIBUS CONTRACTUS NULLIBUS NI LIABILITUS SUMUS, INCLUTATIBUS NON LIMITATUS DESTRUCTIO DIRECTIBUS, CONSEQUENTIUS, PUNITIO, O INCIDENTUS, NON SUNT SI NOS NOTIFICAT VOBIS.

Sit discubriatus defectus en etextum sic entram diaram noventam recibidio, pecuniam tuum refundatorium receptorus posset, sic scribatis vendor. Sit veniabat medium physicalis, vobis idem returnat et replacator possit copius. Sit venitabat electronicabilis, sic viri datus chansus segundibus.

HIC ETEXT VENID "COMO-ASI". NIHIL WARRANTI NUNQUAM CLASSUM, EXPRESSITO NI IMPLICATO, LE MACCHEN COMO SI ETEXTO BENE SIT O IL MEDIO BENE SIT, INCLUTAT ET NON LIMITAT WARRANTI MERCATENSIS, APPROPRIATENSIS PURPOSEM.

Statuen varias non permitatent disclambaris ni warranti implicatoren ni exclusioni limitatio damagaren consequentialis, ecco lo qua disclamatori exclusatorique non vobis applicant, et potat optia alia legali.

F GNU Free Documentation License

Version 1.2, November 2002

Copyright (c) 2000,2001,2002 Free Software Foundation, Inc. 59 Temple Place, Suite 330, Boston, MA 02111-1307 USA

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section [F.0.2](#).

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document free in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondly, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

F.0.0 Applicability and definitions

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, L^AT_EX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ to another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

F.0.1 Verbatim copying

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License

F.0.2 Copying in quantity

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

F.0.3 Modifications

You may copy and distribute a Modified Version of the Document under the conditions of sections [F.0.1](#) and [F.0.2](#) above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

F GNU Free Documentation License

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties — for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

Combining documents

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

Collections of documents

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

Aggregation with independent works

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section F0.2 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

Translation

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section F0.3. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section F0.3) to Preserve its Title (section F0.0) will typically require changing the actual title.

Termination

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

Future revisions of this license

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See <http://www.gnu.org/copyleft/>.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have

the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

Copyright (c) <year> <your name>. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

with the Invariant Sections being <list their titles>, with the Front-Cover Texts being <list>, and with the Back-Cover Texts being <list>.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Copyright

Copyright (c) 2000, 2001, 2002 Free Software Foundation, Inc. 59 Temple Place, Suite 330, Boston, MA 02111-1307, USA

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

List of Tables

1.1	Translating mathematics into code	13
1.2	Nameless functions in programming languages.	20
2.1	Implementing mathematical induction.	48
4.1	Some notation for symbolic reasoning about code.	111
5.1	The correspondence between type constructions and \mathcal{CH} -propositions.	125
5.2	Examples of logical formulas that are true theorems in Boolean logic.	130
5.3	Examples of logical formulas that are <i>not</i> true in Boolean logic.	130
5.4	Proof rules for the constructive logic.	135
5.5	Logic identities with disjunction and conjunction, and the corresponding equivalences of types.	142
5.6	Logical identities with implication, and the corresponding type equivalences and arithmetic identities.	148
5.7	Proof rules of constructive logic are true also in the Boolean logic.	167
6.1	Example translations of functor blocks into <code>map</code> methods.	180
6.2	Type constructions defining a functor L^A	197
6.3	Recursive disjunctive types defined using type equations.	203
6.4	Type constructions defining a contrafunctor C^A	205
7.1	Mathematical notation for the nine basic code constructions.	214
8.1	Type constructions producing the <code>Extractor</code> typeclass.	244
8.2	Type constructions producing the <code>Eq</code> typeclass.	247
8.3	Type constructions producing the <code>Semigroup</code> typeclass.	251
8.4	Type constructions producing the <code>Monoid</code> typeclass.	253
8.5	Type constructions producing the <code>Pointed</code> functor typeclass.	260
8.6	Type constructions producing the <code>Copointed</code> functor typeclass.	264
8.7	Type constructions producing the <code>Pointed</code> contrafunctor typeclass.	266
8.8	Structure of typeclass instance values for various inductive typeclasses.	282
14.1	Correct and incorrect compositions of monads.	436
14.2	Known monad transformers for some monads.	438
14.3	Examples of monads stacked in different orders.	445
D.1	The six type constructions of purely functional programming.	553
D.2	The nine code constructions of purely functional programming.	554

List of Figures

3.1	The disjoint domain represented by the type <code>RootsOfQ</code>	93
5.1	Proof tree for sequent (5.8).	136
8.1	Implementing a recursive instance of the <code>Monoid</code> typeclass via a trait.	254
8.2	Implementing type-safe computations with units of length and mass.	286
17.1	A programmer performs a derivation before writing Haskell code.	533
18.1	Mixing incompatible data types produces nonsensical results.	535
18.2	The Pyongyang method of error-free software engineering.	536
D.1	Composition of dinatural transformations.	560

Index

- F*-algebra, 547
- M*-filterable contrafunctors, 388
- M*-filterable functor, 338, 341
- λ -calculus, 20
- \triangleright -notation
 - see “pipe notation”, 216, 542
- “kind projector” plugin, 444
- abstract syntax tree, 86, 357
- accumulator argument, 36
- aggregation, 2, 11, 39, 40
- algebra, 547
- algebraic, 547
- algebraic data type, 165, 546
- anonymous function
 - see “nameless functions”, 20, 546
- applicative functor, 273, 526
- applied functional type theory, 526, 532
- argument list, 51, 229
- assembly language, 19
- associativity law
 - of addition, 37
 - of flatMap, 369
 - of flatMap for Option, 315
 - of function composition, 103
 - of Kleisli composition, 377
 - of semigroup, 234
- backward composition, 102, 104, 542
- bifunctor, 195
- binary search, 50
- binary tree, 83
- Boolean logic, 129
- bound variable, 8, 59
- Brent Yorgey, 553
- callback, 364
- cardinality, 145
- Cartesian product, 145
- case class, 63
- case expression, 22
- category of monads, 400
- category theory, 339
 - category, 339
 - functor, 339
- morphism, 339
- objects, 339
- closure, 98
- co-density monad, 518
- co-inductive typeclass, 283
- co-pointed bifunctor, 263
- co-pointed functor, 260
- co-product type
 - see “disjunctive type”, 545
- code inference, 120, 137, 372
- code notation, 545
- Collatz sequence, 55
- commutative diagram, 174
- commutativity law
 - of bifunctors, 196
 - of profunctors, 554
- companion object, 230
- compatibility law
 - for pointed and co-pointed functors, 261
- composition law
 - of *M*-filterable contrafunctors, 388
 - of contrafunctors, 185
 - of filter, 291
 - of filterable contrafunctor, 318
 - of flatMap, 377
 - of foreign monad runners, 450
 - of functors, 172
 - of liftOpt, 309
 - of liftOpt for contrafunctors, 323
 - of monad runners, 398
 - of monad transformer’s lift, 448
 - of pair product, 218
- computation expressions (F#), 545
- computational equivalence, 100, 256
 - examples, 301, 306, 369
- conjunction (in logic), 95
- conjunctive normal form, 348
- constant contrafunctor, 204
- constant function, 102, 213
- constant functor, 198, 285
 - example of use, 285
- constructive logic, 134, 169
- continuation-passing, 48, 364
- contrafunctor, 185, 188, 545

- contramonads, 395
- contravariant functor
 - see “contrafunctor”, 188
- contravariant position, 189
- covariant position, 189
- curried arguments, 98
- curried function, 98
- Curry-Howard correspondence, 138, 536
- data transformation, 11
- decidable logic, 154
- default value, 37
- dependency injection, 360
- dependent type, 153, 226
- derivation rule, 133
- destructuring, 22
- dictionary, 19
- dinatural expression, 559
- dinatural transformation, 557
- disjoint domain, 93
- disjoint union, 93
- disjunction (in logic), 95
- disjunctive functions, 214, 540, 543
- disjunctive normal form, 348
- disjunctive type, 69, 545
 - in matrix notation, 144, 214
- do-notation (Haskell), 545
- domain of a function, 5
- dynamic programming, 54, 57
- eager collection, 60
- eager value, 59
- eight code constructions, 131, 132, 213
- embedded ‘if’, 70
- empty case class, 66
- endofunctor, 339
- enumeration type, 88, 94
- Euler product, 15
- evidence value, 267, 284
- exception, 58, 77, 166
- exercises, 14, 32, 55, 82, 84, 85, 91, 116, 128, 145, 153, 163, 193, 209, 224, 274, 298, 311, 324, 333, 351, 367, 393, 403, 502, 520
- expanded form of a function, 103
 - use in recursive values, 243
- exponent, 147
- exponential-polynomial type, 165, 206, 253, 545
- expression, 4
- expression block, 6
- extension methods, 231
- extractor typeclass, 239
- factorial function, 4
- filterable
 - contrafunctor, 317
 - defined via `takeWhile`, 297, 317
 - functor, 291
 - type constructor, 289
- first-order logic, 154
- fixpoint type, 276
- “flipped Kleisli” technique, 379, 387, 484
- for-comprehensions (Python), 545
- formal logic, 132
- forward composition, 102, 104, 542
- free monad, 390, 533
- free pointed functor, 384
- free pointed monad, 384, 508
- free variable, 8, 213, 562
- fully parametric function, 101, 102, 181
- function composition, 102
- function value, 5, 100
- functional programming, 15
- functor, 174
 - in category theory, 339
 - laws of, 184
- functor block, 177, 288, 545
 - source, 178, 288
- functor co-product, 200
- functor composition, 201, 202
- functor exponential, 200, 201
- functor product, 198, 312
- GADT, 181
 - see “unfunctor”, 546
- generalized algebraic data types, 181
- generic functions, 102
- higher-order function, 117
- higher-order type function, 281
- identity function, 102
- identity functor, 198
- identity laws
 - of M -filterable contrafunctors, 388
 - of contrafunctors, 185
 - of deflate, 329
 - of equality, 183, 245
 - of filter, 291
 - of filterable contrafunctors, 318
 - of foreign monad runners, 450
 - of function composition, 103
 - of functors, 171
 - of inflate, 333
 - of `liftOpt`, 307, 308
 - of `liftOpt` for contrafunctors, 323

- of map, 160
- of monad runner, 398
- of monad transformer's lift, 448
- of monoid, 236
- of pair product, 218
- of pure and flatten, 376
- identity morphism, 339
- immutable value, 61
- implication (in logic), 128
- implicit value, 229
- inductive typeclass, 282
- infinite loop in type recursion, 78, 79, 276
- infix syntax, 7
- information loss, 90, 91, 140, 141, 160, 171
- interpreter, 87
- intuitionistic propositional logic, 134
- inverse function, 139, 212
- isomorphic types, 139
- iterator, 60
- Joachim de Lataillade, 559
- John de Goes, 287
- John Reynolds, 553
- “kind projector” plugin, 257, 271, 542
- kind signature, 281
- Kleisli
 - category, 341, 399
 - functions, 336, 376, 545
 - morphisms, 545
 - pronunciation of the name, 309
- Kleisli composition, 376
 - for Option, 309
 - with function composition, 378
- Kurt Gödel, 157
- lambda-function
 - see “nameless function”, 20, 546
- law of de Morgan, 157
- law of excluded middle, 168, 169
- lawful functor, 184
- lazy collection, 60
- lazy value, 60
- left inverse, 212, 329
- Lehmer algorithm, 362
- lifting, 72, 74, 172
- LJT algorithm, 137
- local scope, 6, 19, 59
- logical axiom, 132
- logical implication, 124, 128
- loop detection, 53
- Machin's formula, 14
- map/reduce programming style, 11
- mathematical induction, 14, 17, 34
 - base case, 34
 - inductive assumption, 34
 - inductive step, 34
- matrix notation, 144, 214, 543
- method, 545
- method syntax, 7
- monad morphism, 400
- monad transformers, 437
 - base lift, 454
 - base runner laws, 454
 - formal definition, 454
 - functor laws of runner, 454
 - identity law, 454
 - lifting law, 454
 - monad construction law, 454
 - monadic naturality laws, 454
 - non-degeneracy law, 451, 454
 - runner laws, 454
 - stacking, 457
- monadic naturality law
 - of base runner (brun), 452
 - of flift, 451
 - of swap, 471
- monadic program, 360, 397
- monads, 26, 338, 344
 - 3-swap law, 496
 - category of, 400
 - choice monad, 483
 - commutative, 394
 - continuation monad (Cont), 364
 - free pointed, 384, 508
 - lazy/eager evaluation monad (Eval), 363
 - linear, 475, 482
 - Reader monad, 358
 - rigid, 482
 - runner, 360, 363, 365, 398
 - Search monad, 390
 - search monad, 483
 - Sel (selector) monad, 390, 483
 - stack of, 444, 457
 - State monad, 362
 - Writer monad, 361
- monoid, 236
- monoidal convolution, 456
- morphism, 339
- nameless function, 5, 546
- nameless type-to-type function, 257
- natural transformation, 335
- naturality law

- combined with identity law, 308
- of co-pointed bifunctors, 263
- of deflate, 303
- of extract, 260
- of filter, 291
- of filter for contrafunctors, 324
- of flatMap, 368, 369
- of flatMap for Option, 316
- of inflate for contrafunctors, 324
- of liftOpt, 307, 311, 321
- of monad runners, 398
- of pure, 255
- of pure for contrafunctors, 263
- of the function Δ , 218
- negation (in logic), 167
- Newton's method, 112
- nine code constructions, 213, 553
- non-empty list, 81
- object-oriented inheritance, 287
- on-call value, 60, 149, 441
- opaque type, 63
- operator syntax, 119
- order of a function, 117
- pair product of functions, 198, 218, 543
- palindrome integer, 56
- paradigm of programming, 15
- parametric code, 65
- parametricity theorem, 337, 371, 553
- partial application, 100, 120
- partial function, 58, 70, 112, 212, 226
- partial function law
 - of filter, 291
 - reverse order in contrafunctors, 323
- partial type-to-value function, 226, 546
- pattern matching, 22
 - in matrix notation, 144
 - infallible, 58
- pattern variables, 22
- Pawel Szulc, 183
- perfect number, 57
- phantom type parameter, 285
- Philip Wadler, 553
- pipe notation, 194, 216, 316, 542
 - operator precedence, 209, 217
- planned exception, 77
- point-free calculations, 216
- pointed contrafunctor, 262
- pointed functor, 255
- pointed type, 234
- polynomial functor, 200, 207, 546
 - recursive, 177
- polynomial type, 165
- predicate, 7
- problems, 522
- procedure, 92
- product type, 145, 546
- profunctor, 211, 327, 404, 546
- proof (in logic), 122
- proposition (in logic), 95
- pure compatibility laws, 522
- pure function, 61, 62
- purely functional
 - code constructions, 554
 - functor, 568
 - profunctor, 554
 - program, 213, 553, 559
 - type constructions, 553
- purity laws
 - of monad transformer runners, 451
- recursive function, 35
 - accumulator argument, 36
 - proving laws for, 203
 - termination, 243
- recursive type, 78
 - unrolling trick, 332
- referential transparency, 61, 358
- reflexivity law, 183, 244, 245
- regular-shaped tree, 84, 299
 - is not a monad, 358
- Riemann's zeta function, 15
- rigid functor, 500
- Robert C. Martin, 530
- Ronald Backhouse, 553
- rose tree, 84
- run-time error, 58
- runner, 87
- Russell O'Connor, 273, 275
- saturated application, 100
- Scala method, 100
- Scala's Iterator, 61, 62
- search functor, 326
- selector monad, 500
- semigroup, 234
- semimonads, 344
 - example of usage, 361
- sequent (in logic), 548
 - goal, 122
 - premises, 122
- serializer, 244
- shadowed name, 59, 118
- side effect, 62, 358
- Simon Peyton Jones, 424

- Simpson's rule, 19
- six type constructions, 123, 206
- solved examples, 11, 27, 41, 48, 69, 73, 81, 85, 88, 108, 111, 125, 141, 146, 147, 154, 175, 189, 208, 232, 267, 293, 300, 322, 328, 344, 353, 356, 365, 371, 400, 436, 502, 558
- stack memory, 36
- stream, 46
- structural analysis, 204, 239
- structure functor, 282
- sum type, 146
 - see "disjunctive type", 545
- symbolic calculations, 106
- symmetry law of equality, 183
- tagged union type
 - see "disjunctive type", 545
- tagless final, 533, 547
- tail recursion, 36
- total function, 58, 226
- trace of a matrix, 351
- trampolines, 48
- transitivity law of equality, 183
- "trivial" semigroup, 247
- truth table, 129
- tuples, 21, 546
 - accessors, 21
 - as function arguments, 101
 - fields, 21
 - nested, 22
 - in pattern matching, 47
 - parts, 21
- turnstile symbol, 133
- type alias, 41, 63, 124
- type annotation, 73, 90, 540
- type casts, 169
- type checking, 111
- type constructor, 66
 - contravariant, 188
 - covariant, 188
- type conversion function, 187
- type diagram, 174
- type domain, 227
- type equivalence, 139
 - accidental, 146, 163
- type error, 21, 22, 31, 58, 64
- type inference, 110, 111, 120
- type notation, 123, 125, 546
 - operator precedence, 125
- type parameter, 24, 65
- type reflection, 183
- type relation, 284
 - many-to-many, 284
 - many-to-one, 285
- type safety, 166
- type-level function, 201
- type-to-type function, 226
- type-to-value function, 226
- typeclass, 225
 - Bifunctor, 270
 - co-inductive, 283
 - constraint, 225
 - Contrafunctor, 270
 - Copointed, 260
 - Eq, 244
 - evidence argument, 267
 - evidence value, 284
 - Extractor, 239
 - Functor, 238
 - inductive, 282
 - inheritance, 285
 - instance value, 227
 - Monoid, 236
 - Pointed, 255
 - Semigroup, 234
 - Semimonad, 369
 - Show, 244
- typed hole, 158
- types, 17
 - equivalent, 139
 - exponential-polynomial, 165
 - higher-kinded, 281
 - isomorphic, 139
 - pointed, 234
 - polynomial, 165, 174, 175
 - structural analysis, 239
 - subtype of, 187
- uncurried function, 98, 100
- undecidable logic, 154
- unevaluated expression, 86
- unfold function, 52, 57
- unfunctor, 181, 186, 211, 227, 541, 546
- unit type, 66, 540
 - named, 67, 94, 123, 126, 213
- universal quantifier, 124
- unplanned exception, 77
- unrolling trick for recursive types, 332
- value semantics, 61, 358
- variable, 16, 17
- variance annotation, 188
- verifying laws with scalacheck, 238
- void type, 72, 78, 141, 149, 166, 227, 540

Index

in matrix notation, 215, 291

Wallis product, 12

wedge law, 559

well-typed expression, 111