Ernst-Erich Doberkat

Special Topics in Mathematics for Computer Scientists

Sets, Categories, Topologies and Measures

Springer

Special Topics in Mathematics for Computer Scientists

Ernst-Erich Doberkat

Special Topics in Mathematics for Computer Scientists

Sets, Categories, Topologies and Measures

Springer

Ernst-Erich Doberkat
Math ++ Software
Lewackerstr. 6 b
Bochum, Germany
math@doberkat.de

ISBN 978-3-319-22749-8 ISBN 978-3-319-22750-4 (eBook)
DOI 10.1007/978-3-319-22750-4

Library of Congress Control Number: 2015954182

Mathematical Subject Classification (2010): 03-01, 03B45, 03B70, 18-01, 18C20, 18C50, 54-01, 54B99, 54D99, 60-01, 60B05, 60J05, 68Q87, 68Q55, 68N30

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Für Nina Luise

Preface

The idea to write this book came to me when, after having taught an undergraduate course on *concrete mathematics* using the wonderful eponymous book *Concrete Mathematics: A Foundation for Computer Science* [GKP89], I wanted to do a graduate course on Markov transition systems for computer scientists. It turned out that I had to devote most of the time to laying the foundations from sets, measures, and topology and that I could not find an adequate textbook to recommend to my students. This contrasts significantly the situation in other areas, such as the analysis of algorithms, where many fine textbooks are available. Consequently, I had to dig through the mathematical literature, taking pieces from here and there, in effect trying to nail a firm albeit makeshift mathematical scaffold the students could stand on securely.

Looking at the research literature in this area and in related fields, one also finds a lack of quotable resources. Each author has to construct her or his own foundations in order to get going, wasting considerable effort to prove the same lemmata over and over again.

So the plan for this book began to develop. I decided to focus on sets, topologies, categories, and measures. Let me tell you why.

Sets and the Axiom of Choice Sets are a universal tool for computer scientists, the tool which has been imported as a *lingua franca* from mathematics. When surveying the computer science literature, we see that sets and the corresponding constructs like maps, power sets, orders, etc., are being used freely, but there is usually no concern regarding the axiomatic basis of these objects—sets are being used, albeit in a fairly naive way. This should not be surprising, because they are just tools and often not the objects of consideration themselves. However, fairly early in education, a computer scientist encounters the phenomenon of recursion, either as a recursive function or as a recursive definition. And here immediately arises the question as to why the corresponding constructs work and, specifically, how one can be sure that a particular recursive program is actually terminating. The same question, probably a little more precisely, appears in techniques which are related to term rewriting. Here, one inquires whether a particular chain of replacements will actually lead to a result in a finite amount of time. People in term rewriting have found a way of writing this down, namely, a terminating condition which is closely related to some well ordering. This means that there are no infinitely long chains, which is of course a very similar condition to the one that is encountered when talking about the termination of a recursive procedure: Here, one does not want to have infinitely long chains of procedure or method calls. This suggests structural similarities between the termination of a recursive method and rewriting a term.

When investigating the background of all these events, we[1] find that we need to look at well orderings. These are orderings which forbid the existence of infinitely long decreasing chains. Do well orderings always exist? This question is of course fairly easy to answer when we talk about finite scenarios, but sometimes it is mandatory to consider infinite objects as well. The world may be finite, but our models of the world are not always so. Hence, the question arises whether we can take an arbitrary set and construct a well ordering for it. As it turns out, this question is very closely connected with another question, which at first glance does not look related at all: Given a collection of nonempty sets, are we able to select from each set exactly one element? One of the interesting points which indicates that things are probably a little bit more complicated than they look is the observation that the possibility of well ordering an arbitrary set is equivalent to that of the question of selection, which came to be known as the axiom of choice. It turned out during the

[1]For the usage of the first person plural in this treatise, let me quote William Goldbloom Bloch. He writes in his enjoyable book on Borges' mathematical ideas in a similar situation: "This should not be construed as a 'royal we.' It has been a construct of the community of mathematicians for centuries and it traditionally signifies two ideas: that 'we' are all in consultation with each other through space and time, making use of each other's insights and ideas to advance the ongoing human project of mathematics, and that 'we'—the author and reader—are together following the sequences of logical ideas that lead to inexorable, and sometimes poetic, conclusions." [Blo08, p. 19].

discussion in mathematics that there is a whole bag of other properties that are equivalent to this axiom. We will see that the axiom of choice is equivalent to some well-known proof principles like Zorn's Lemma or Tuckey's Maximality Principle. Because this discussion relating to the axiom of choice and similar constructions has been raging in mathematics for more than a century now, we cannot hope to be able to even completely list out all those things which we have to eliminate. Nevertheless, we try to touch upon some topics that appear to be important for developing mathematical structures within computer science. We can even show that some results do not hold if another path is pursued and the axiom of choice is replaced by another one; this refers to a game-theoretic scenario, which is of course of interest to computer scientists as well.

Because the discussion on the axiom of choice touches upon so many areas in mathematics, it gives us the opportunity to look at some of them. In this sense, the axiom of choice is a peg on which we hang our walking stick.

Categories Many areas of mathematics show surprising structural similarities, which suggests that it might be interesting and helpful to focus on an abstract view, hereby unifying concepts. This abstract view looks at the mathematical objects from the outside and studies the relationship between them, for example, groups (as objects) and homomorphisms (as an indicator of their relationship), or topological spaces together with continuous maps, or ordered sets with monotone maps, etc. It leads to the general notion of a category. A category is based on a class of objects together with morphisms for each pair of objects. Morphisms can be composed; the composition follows some laws which are considered evident and natural.

This approach has considerable appeal to a software engineer. In software engineering, the implementation details of a software system are usually not particularly important from an architectural point of view; they are encapsulated in a component. In contrast, the relationship of components with each other is of interest because this knowledge is necessary for composing a system from its components. Roughly speaking, the architecture of a software system is characterized both by its components and their interaction, the static part of which can be described by what we may perceive as morphisms.

This has been recognized fairly early in the software architecture community, witnessed by the April 1995 issue of the *IEEE Transactions on Software Engineering*. It was devoted to software architecture and demonstrated that formalisms from category theory in discussing architectures are very helpful for clarifying structures. So the language of categories offers some attractions to software engineers. We will also see that the tool set of modal logics, another area which is important to software construction, profits substantially from constructions which are firmly grounded in categories.

We are going to discuss categories here and introduce the reader to the basic constructions. The world of categories is too rich to be captured in these few pages, so we have made an attempt to provide the reader with some basic proficiency in categories, helping her or him to get a grasp of the basic techniques. This modest goal is attained by blending the abstract mathematical development with a plethora of examples.

Topological Spaces A topology formalizes the notion of an open set; call a set open iff each of its members leaves a little room, something like a breathing space, around it. This gives immediately an idea of the structure of the collection of open sets—they should be closed under finite intersections, but under arbitrary unions, yielding the base for a calculus of observable properties. This approach puts its emphasis subtly away from the classic approach, e.g., in mathematical analysis or probability theory, by stressing on different properties of a space. The traditional approach, for example, stresses on separation properties, such as being able to separate two distinct points through an open set. Such a strong emphasis is not necessarily observed in the computationally oriented use of topologies, where, to give an example, pseudometrics for measuring the conceptual distance between objects are important for finding an approximation between Markov transition systems.

We give a brief introduction to some of the main properties of topological spaces. The objective is to provide the tools and methods offered by the set-theoretic topology to an application- oriented reader; thus, we introduce the very basic notions of this topology and discuss briefly the applications of its tools. Some connections to logic and set theory are indicated. Compactness has been made available very early; thus, compact spaces serve occasionally as an exercise ground, before compactness with its ramifications is discussed in depth. Continuity is an important topic and so are the basic constructions like product or quotients which are enabled by it. Since some interesting and important topological constructions are linked to filters, we study filters and convergence, comparing through examples the sometimes more easy-to-handle nets with the occasionally more cumbersome filters. Talking about convergence, separation properties suggest themselves. Often it happens that one works with a powerful concept but that this concept requires assumptions which are too strong; hence, one has to weaken it in a sensible way. This is demonstrated in the transition from compactness to local compactness; we discuss local compact spaces, and we give an example of a compactification. Quantitative aspects come into the picture when one measures openness through a pseudometric; here, many concepts are seen in a new, brighter light; in particular, the problem of completeness emerges. Complete spaces have some very special properties, for example, the intersection of countably many open dense sets is dense again. This is Baire's Theorem; we show through a Banach–Mazur game played on a topological space that being of the first category can be determined through one of the players having a winning strategy.

This completes the overview of the basic properties of topological spaces. We present next a small gallery in which topology is in action. The reason for singling out some topics is that we want to demonstrate the techniques developed with topological spaces for some interesting applications. For example, Gödel's Completeness Theorem for (countable) first-order logic has been proved by Rasiowa and Sikorski through a combination of Baire's Theorem and Stone's topological representation of Boolean algebras. This topic is discussed in detail. The calculus of observations, which is mentioned above, leads to the notion of topological systems. This hints at an interplay of topology and order, since a topology is after all a complete Heyting algebra in the partial order provided by inclusion. Another important topic is the approximation of continuous functions by a given class of functions, like polynomials on a closed interval, leading directly to the Stone–Weierstraß Theorem on a compact topological space, a topic with a rich history. Finally, the relationship of pseudometric spaces with general topological spaces is reflected again; we introduce uniform spaces as an ample class of spaces which are more general than pseudometric spaces but less general than their topological cousins. Here, we find concepts like completeness or uniform continuity, which are formulated for metric spaces but which cannot be realized in general topological ones.

Measures for Probabilistic Systems Markov transition systems are based on transition probabilities on a measurable space. This is a generalization of discrete spaces, where certain sets are declared to be measurable. So, in contrast to assuming that we know the probability for the transition between two states, we have to model the probability of a transition going from one state to a set of states: Point-to-point probabilities are no longer available due to working in a comparatively larger space. Measurable spaces are the domains of these probabilities. This approach has the advantage of being more general than finite or countable spaces, but now one deals with a fairly involved mathematical structure.

We start off with a discussion of σ-algebras, which are also discussed in the chapter of sets, and we look at the structure of σ-algebras, in particular at its generators; it turns out that the underlying space has something to say about it. In particular, we deal with Polish and related spaces. Some constructions in this area are studied; they have immediate applications in logic and in Markov transition systems, in which measures are vital. We show also that we can construct measurable selections, which we use for an investigation into the structure of quotients in the Kleisli monad, providing an interesting and fruitful example of the interplay of arguments from measure theory and categories. This interplay is stressed upon also for the investigation of stochastic effectivity functions, which leads to an interpretation of game logics.

After having laid the groundwork, we construct the integral of a measurable function through an approximation process, very much in the tradition of the Riemann integral but with a larger scope. We also go the other way: Given an integral, we construct a measure from it. This is the elegant way proposed by P.J. Daniell for constructing measures, and it can be brought to fruit in this context for a fairly simple proof of the Riesz Representation Theorem on compact metric spaces.

Having all these tools at our disposal, we look at product measures, which can be introduced now through a kind of line sweeping—if you want to measure an area in the plane, you measure the line length as you sweep over the area; this produces a function of the abscissa, which then yields the area through integration. One of the main tools here is Fubini's Theorem. Applications include a discussion of projective systems. A case study shows that projective systems arise naturally in the study of continuous time stochastic logics.

Finally, we take up a classic: L_p-spaces. We start from Hilbert spaces, apply the representation of linear functionals on L_2 to obtain the Radon–Nikodym Theorem through von Neumann's ingenious proof, and derive from it the representation of the dual spaces.

Because we are driven by applications to Markov transition systems and similar objects, we do not strive for the most general approach to measure and integral. We usually formulate the results for finite or σ-finite measures, leaving the more general cases outside our focus. This also means that we do not deal with complex measures; we show, however, how to deal with complex measures when the occasion arises.

Things Left Out Several things had to be left out. This is an incomplete list, from which many things had to be left out. For example, I do not discuss ill-founded sets in the chapter on set theory, and I cannot take even a tiny step into forcing or infinite combinatorics. I do not cover final coalgebras or coinduction in the chapter on categories, which also excludes an extensive discussion on limits and colimits. It would have been helpful to look into hyperspaces in the chapter on topologies or to discuss topological groups with their Haar measure, let alone provide a glimpse at topological vector spaces. Talking about measures, martingales are missing, and the connections to topological measure theory are looked at through the lens of Polish or analytic spaces.

But, alas, many a choice had to be made, and since I am a confessing Westphalian, I quote this proverb:

Wat dem een sin Uhl, is dem anneren sin Nachtigal.

So I have tried to incorporate topics which to me seem useful.

Organization Each chapter derives its content in the usual strict mathematical way, with proofs and all that. It belongs certainly to the education of a computer scientist of the theoretical variety to carry out proofs, and not to rely on good faith or on the well-known art of handwaving. The development is supported by many examples, some motivating, some mathematically interesting, but most of them oriented toward applications in computer science. Larger examples are presented as *case studies*. They appear interesting from a modeling point of view as well as due to their application of the mathematical techniques at hand. The same applies to exercises, which are given at the end of each chapter. Bibliographic notes provide usually the source for some particular approach, a proof, or an idea which is pursued; they also give hints where further information may be found. It has not always been easy to attribute a development to a particular paper, book, or author, since folklore quickly spreads, and results and ideas are amended or otherwise modified, sometimes obscuring the true originators. Thus, the author apologizes if results could not always be attributed properly, or not at all.

How to Read This Book This is a book with an intended audience which is somewhat advanced, hence, it seems to be a bit out of place to give suggestions on how to read it. The advice to the reader is just to take what she or he needs, do the exercises, and, if this is not enough, look for further information in the literature. The author put in a great deal of effort to provide an ample list of references. Good luck!

Thanks Finally, I want to thank J. Bessai, B. Fuchssteiner, H.G. Kellerer, E.O. Omodeo, P. Panangaden, D. Pumplün, H. Sabadhikary and S.M. Srivastava, P. Sánchez Terraf, and F. Stetter. They helped in opening some doors—mathematically or otherwise—for me, made some insightful comments, and gave me a helping hand which, sometimes in the transitive closure, had some impact on this book. The *Deutsche Forschungsgemeinschaft* supported my research into algebraic and coalgebraic properties of stochastic relations for more than ten years; some results of this work could be used for this book. Stefan Dissmann and Alla Stankjawitschene, my former secretary, helped clear my path by taking many administrative obstacles off it.

The cooperation with Dr. Mario Aigner and Sonja Gasser of Springer-Verlag was constructive and helpful. I want to thank them all.

Above all, I owe my thanks to Gudrun for all her love and understanding. I devote this book to our youngest granddaughter Nina Luise; she will probably like the idea of playing with symbols.

Bochum, Germany Ernst-Erich Doberkat

Contents

Chapter 1

The Axiom of Choice and Some of Its Equivalents

Sets are a universal tool for computer scientists, the tool which has been imported as a *lingua franca* from mathematics. Program development, for example, starts sometimes from a mathematical description of the things to be done, the specification, and the data structures, and—you guess it—sets are the language in which these first designs are usually written down. There is even a programming language called `SETL` based on sets [SDDS86]; this language served as a prototyping tool, its development having been essentially motivated by the ambition to shorten as much as possible the road from a formal description of an object to its representation through an executable program; see [CFO89, COP01] and for practical issues [DF89].

In fact, it turned out that programming in what might be called executable set theory has the advantage of having the capability to experiment with the objects at hand, leading, for example, to the first implementation of the programming language `Ada`, the implementation of which was deemed for quite some time as nearly impossible. On the other hand it turned out that sets may be a feature *nice to have* in a programming language, but that they are probably not always the appropriate universal data structure for engineering program systems. This

E.-E. Doberkat, *Special Topics in Mathematics for Computer Scientists*, DOI 10.1007/978-3-319-22750-4_1

is witnessed by the fact that some languages, like `Haskell` [OGS09], have set-like constructs such as list comprehension, but they do not implement sets fully. As the case may be, sets are important objects when arguing about programs. They constitute an important component of the tool kit which a serious computer scientist should have at her or his disposal.

When surveying the computer science literature, we see that sets and the corresponding constructs like maps, power sets, orders, etc., are being used freely, but there is usually no concern regarding the axiomatic basis of these objects—sets are being used, albeit in a fairly naive way. This should not be surprising, because they are just tools and often not the objects of consideration themselves. A tool should be available to a computer scientist whenever needed, but it really should not bring with it complications of its own. However, fairly early in education, a computer scientist encounters the phenomenon of recursion, be it as a recursive function, be it as a recursive definition. And here immediately arises the question as to why the corresponding constructs work and, specifically, how one can be sure that a particular recursive program is actually terminating. The same question, probably a little bit more focused, appears in techniques which are related to term rewriting. Here one inquires whether a particular chain of replacements will actually lead to a result in a finite amount of time. People in term rewriting have found a way of writing this down, namely, a terminating condition which is closely related to some well ordering. This means that we do not have infinitely long chains, which is of course a very similar condition to the one that is encountered when talking about the termination of a recursive procedure: Here we do not want to have infinitely long chains of procedure or method calls. This suggests structural similarities between the termination of a recursive method and rewriting a term. If you think about it, mathematical induction enters this family of observations, the members of which show a considerable similarity.

When we investigate the background in which all this happens, we find that we need to look at well orderings. These are orderings which forbid the existence of infinitely long decreasing chains. It turns out that the mathematical ideas expressed here are fairly closely connected to ordinal numbers. It is not difficult to construct a bridge from orderings and well orders to the question whether it is actually possible to find a well order for each and every set. The bridge a computer scientist might traverse is loosely described as follows: Because we want to be able to deal with arbitrary objects and because we want to run programs with these arbitrary objects, it should be possible to construct terminating recursive methods for those objects. But in order to do that, we should make sure that no infinite chains of method invocations occur, which in turn poses the question whether or not we can impose an order on these objects that renders infinite chains impossible (admittedly somewhat indirectly, because the order is imposed actually by procedure calls). But here we are—we want to know whether such a construction is possible; mathematically this leads to the possibility of well ordering each and every set.

This question is of course fairly easy to answer when we talk about finite scenarios, but sometimes it is mandatory to consider infinite objects as well. The world may be finite, but our models of the world are not always so. Hence the question arises whether we can take an arbitrarily large set and construct a well ordering for this set. As it turns out, this question is very closely connected with another question, which at first glance does not look similar at all: We are given a collection of nonempty sets; are we able to select from each set exactly one element? This question has vexed mathematicians for more than a century now. One of the interesting points, which indicates that things are probably a little more complicated than they look, is the observation that the possibility of well ordering an arbitrary set is equivalent to that of the question of selection, which came to be known as the axiom of choice. It turned out during the discussion in mathematics that there is a whole bag of other properties that are equivalent to this axiom. We will see that the axiom of choice is equivalent to some well-known proof principles like Zorn's Lemma or Tuckey's Maximality Principle. Because this discussion relating to the axiom of choice and similar constructions has been raging in mathematics for more than a century now, we cannot hope to be able to even completely list out all those things which we have to leave out. Nevertheless, we try to touch upon some topics that appear to be important for developing mathematical structures within computer science.

Since the axiom of choice and its variants touch upon those topics in mathematics that are much in use in computer science, this presents us with the opportunity to select some of these and discuss them independently and in light of the use of the axiom of choice. We discuss, for example, lattices; introduce ideals and filters; and pose the maximality question: Is it always possible to extend a filter to a maximal filter? It turns out that the answer is in the affirmative, and this has some interesting applications in the structure theory of, for example, Boolean algebras. Because of this we are able to discuss one of the true classics in this area, namely, Stone's Representation Theorem, which says that every Boolean algebra is isomorphic to an algebra of sets. Another interesting application of Zorn's Lemma is Alexander's Theorem, which shows that for establishing compactness of the space, one may restrict one's attention to covering a topological space with subbase elements. Because we have then compactness at our disposal, we establish also compactness of the space of all prime ideals of a Boolean algebra. Quite apart from these questions, which are oriented toward order structures, we establish the Hahn–Banach Theorem, which shows that a dominated linear functional can be extended from a linear subspace to the entire space in a dominated way.

A particular class of Boolean algebras are closed even under countable infima and suprema; these algebras are called σ-algebras. Since these algebras are interesting, specifically when it comes to probabilistic models in computer science, we treat these σ-algebras in some detail, in particular with respect to measures and their extensions. The general situation in application is sometimes that one has the generator of a Boolean σ-algebra and a set function which behaves decently on this generator, and one wants to extend this set function to the whole σ-algebra. This gives rise to a fairly rich and interesting construction, which in turn has some connection with the question of the axiom of choice. The extension process extends the measure far beyond the Boolean σ-algebra generated by the family under consideration, and the question arises as to how far this extension really goes. This may be of interest, e.g., if one wants to measure some set, since one needs to know whether this set can be measured at all, hence whether it is actually in the domain of the extended measure. The axiom of choice helps in demonstrating that this is not always possible. It can be shown that there are sets which cannot be measured. This depends on a selection argument for classes of an easily constructed equivalence relation.

This will be discussed further in the chapter.

Then we turn to games, games as a model for human interaction, in which two players, *Angel* and *Demon*, play against each other. We describe how a game is played and what are the strategies. In particular, we say what constitutes winning strategies. This is done first in the context of infinite sequences of natural numbers. The model has the advantage of being fairly easy to grasp; it has the additional structural advantage in that we can map many applications to this scenario.

Actually, games become really interesting when we know that one of the participants has actually a chance to win. Hence, we postulate that our games are of this kind, so that always either Angel or Demon has a strategy to win the game. Unfortunately it turns out that this postulate, called the axiom of determinacy, is in conflict with the axiom of choice. This is of course a fairly unpleasant situation, because both axioms appear as reasonable statements. So we have to see what can be done about this. We show that if we assume the axiom of determinacy, we can actually demonstrate that each and every subset of the real line is measurable. This is in contradiction to the observation we just described.

This discussion serves two purposes. The first one is that one sometimes wants to challenge the axiom of choice in favor of other postulates, which may turn out to have more advantages (in the context of games, the postulate that one of the players has a winning strategy, no matter how the game is constructed, has certainly some advantageous aspects). But the axiom of choice is, as we will see, quite a fundamental postulate, so one has to find a balance between both. This does look terribly complicated, but on the other hand does not seem to be difficult to manage from a practical point of view—and computer scientists are by definition practical people! The second reason for introducing games and for elaborating on these results is to demonstrate that games can actually be used as tools for proofs. These tools are used in some branches of mathematics quite extensively, and it appears that this may be an attractive choice for computer scientists as well.

We work usually in what is called *naive set theory*, in which sets are used as a formal manner of speaking without much thinking about it. Sets are just tools to formally express ideas.

When mathematicians and logicians like G. Frege, G. Cantor, or B. Russell thought about the basic foundations of mathematics, they found a huge pile of unposed and unanswered questions about the basic building blocks of mathematics, e.g., the definition of a cardinal number was usually taken for granted, without a formal foundation; a foundation was even resisted or ridiculed.[1]

The Axioms of ZFC. Nevertheless, at around the turn of the century, there seems to have been some consensus among mathematicians that the following axioms are helpful for their work; they are called the *Zermelo–Fraenkel System With Choice (ZFC)* after E. Zermelo and A.A. Fraenkel.

We will discuss them briefly and informally now. Here they are.

Extensionality *Two sets are equal iff they contain the same elements.* This requires that sets exist and that we know which elements are contained in them; usually these notions (set, element) are taken for granted.

Empty set axiom *There is a set with no elements.* This is of course the empty set, denoted by $\emptyset$.

Axiom of pairs *For any two sets, there exists a set whose elements are precisely these sets.* From the extensionality axiom, we conclude that this set is uniquely determined. Without the axiom of pairs, it would be difficult to construct maps. Hence we can construct sets like $\{a, b\}$ and singleton sets $\{a\}$ (because the axiom does not talk about different elements, so we can form the set $\{a, a\}$, which, by the axiom of extensionality, equals the set $\{a\}$). We can also define an ordered pair through $\langle a, b\rangle := \{\{a\}, \{a, b\}\}$.

Axiom of separation Let φ be a statement of the formal language with a free variable z. *For any set x, there is a set containing all z in x for which $\varphi(z)$ holds.* This permits forming sets by describing the properties of their elements. Note the restriction "for any set x"; suppose we drop this and postulate "There is a set containing all z for which $\varphi(z)$ holds." Let $\varphi(z)$ be the statement $z \notin z$, then we would have postulated the existence of the set $a := \{z \mid z \notin z\}$ (is $a \in a$?). Hence we have to be a bit more modest.

[1]Frege's position, for example, was considered in the polemic by J.K. Thomae, "Gedankenlose Denker, eine Ferienplauderei" (Thinkers without a thought, a causerie for the vacations). Jahresber. Deut. Math.Ver. 15, 1906, 434–438 as somewhat hare-brained; see Thiel's treatise [Thi65].

Power set axiom *For any set x, there exists a set consisting of all subsets of x.* This set is called the power set of x and denoted by $\mathcal{P}(x)$. Of course, we have to define the notion of a subset, before we can determine all subsets of a set x. A set u is called a subset of set x (in symbols $u \subseteq x$) iff every element of u is an element of x.

Union axiom *For any set there is a set which is the union of all elements of x.* This set is denoted by $\bigcup x$. If x contains only a handful of elements like $x = \{a, b, c\}$, we write $\bigcup x$ as $a \cup b \cup c$. The notion of a union is not yet defined, although intuitively clear. We could rewrite this axiom by stating it as given a set x, there exists a set y such that $w \in y$ iff there exists a set a with $a \in x$ and $w \in a$. The intersection of two sets a and b can then be defined through the axiom of separation with the predicate $\varphi(z) := z \in a$ *and* $z \in b$, so that we obtain $a \cap b := \{z \in a \cup b \mid z \in a \text{ and } z \in b\}$.

This is the first group of axioms which are somewhat intuitive. It is possible to build from it many mathematical notions (like maps with their domains and ranges, finite Cartesian products). But it turns out that they are not yet sufficient, so an extension to them is needed.

Axiom of infinity *There is an inductive set.* This means that there exists a set x with the property that $\emptyset \in x$ and that $y \cup \{y\} \in x$ whenever $y \in x$. Apparently, this permits building infinite sets.

Axiom of replacement Let φ be a formula with two arguments. *If for every a there exists exactly one b such that $\varphi(a, b)$ holds, then for each set x there exists a set y which contains exactly those elements b for which $\varphi(a, b)$ holds for some $a \in x$.* Intuitively, if we can find for formula φ for each $a \in x$ exactly an element b such that $\varphi(a, b)$ is true, then we can collect all these elements b in a set. Let φ be the formula $\varphi(x, y)$ iff x is a set and $y = \mathcal{P}(x)$; then there exists for a given family x of sets the set of all power sets $\mathcal{P}(a)$ with $a \in x$.

Axiom of foundation *Every set contains a $\in$-minimal element.* Sets contain other sets as elements, as we have seen, so there might be the danger that a situation like $a \in b \in c \in a$ occurs, hence there is a $\in$-cycle. In some situations this might be desirable, but not in this very basic scenario, where we try to find a fixed ground to work on. A formal description of this axiom reads that for each set x there exists a set y such that $y \in x$ and $x \cap y = \emptyset$. We will have to deal with a very similar property when discussing ordinal numbers in Sect. 1.4.

Now we have recorded some axioms which provide the basis of our daily work, to be used without any qualms. They permit building up mathematical structures like relations, maps, injectivity, surjectivity, and so on. We will not do this here (it gets somewhat boring after a time if one is not seeing some special effect—then it may become awfully hard), and trust that the reader is familiar with these structures.

But there still is a catch: Look at the argumentation in the following proposition which constructs some sort of an inverse for a surjective map.

Proposition 1.0.1 *There exists for each surjective map $f : A \to B$ a function $g : B \to A$ such that $(f \circ g)(b) = b$ for all $b \in B$.*

$\dashv$ is *end of proof.*

Proof For each $b \in B$, the set $f^{-1}\left[\{b\}\right]$ is not empty, because f is surjective. Thus we can pick for each $b \in B$ an element $g(b) \in f^{-1}\left[\{b\}\right]$. Then $g : B \to A$ is a map, and $f(g(b)) = b$ by construction. $\dashv$

WHERE IS THE CATCH? The proof seems to be perfectly innocent and straightforward. We simply have a look at all the inverse images of elements of the image set B, and all these inverse images are not empty, so we pick from each of these inverse images exactly one element and construct a map from this.

Well, the catch lies in picking an element from each member of this collection. The collection of axioms above says nowhere that this selection is permitted (now you might think that mathematicians find a sneaky way of permitting such a pick, through the back door, so to speak; trust me—they cannot!).

Hence we need some additional device, and this is the axiom of choice. It will be discussed at length now; we take the opportunity to use this discussion as a kind of a peg onto which we hang some other objects as well. The general approach will be that we will discuss mathematical objects of interest, and at a crucial point the discussion of $(\mathbb{AC})$ and its equivalents will be continued (if you ever listened to a Wagner opera, you will have encountered his leitmotifs).

1.1 The Axiom of Choice

The axiom of choice states that

> ($\mathbb{AC}$) Given a family $\mathcal{F}$ of non-empty subsets of some set X, there exists a function $f : \mathcal{F} \to X$ such that $f(F) \in F$ for all $F \in \mathcal{F}$.

The function, the existence of which is postulated by this axiom, is called a *choice function* on $\mathcal{F}$.

It is at this point not quite clear why mathematicians make such a fuss about ($\mathbb{AC}$):

W. Sierpinski *It is the great and ancient problem of existence that underlies the whole controversy about the axiom of choice.*

P. Maddy *The axiom of choice has easily the most tortured history of all set-theoretic axioms.*

T. Jech *There has been some controversy about the axiom of choice, indeed.*

H. Herrlich *AC, the axiom of choice, because of its nonconstructive character, is the most controversial mathematical axiom, shunned by some, and used indiscriminately by others*

(see [Her06]). In fact, let $X = \mathbb{N}$, the set of natural numbers. If $\mathcal{F}$ is a set of nonempty subsets of $\mathbb{N}$, a choice function is immediate—just let $f(F) := \min F$. So why bother? We will see below that $\mathbb{N}$ is a special case. B. Russell gave an interesting illustration: Suppose that you have an infinite set of pairs of shoes, and you are to select systematically one shoe from each pair. You can always take either the left or the right one. But now try the same with an infinite set of pairs of socks, where the left sock cannot be told from the right one. Then you have to have a choice function.

But we do not have to turn to socks in order to see that a choice function is helpful; we rather prove Proposition 1.0.1 again.

Proof (of Proposition 1.0.1)
Define
$$\mathcal{F} := \{f^{-1}\left[\{b\}\right] \mid b \in B\};$$

then $\mathcal{F}$ is a collection of nonempty subsets of A, since f is onto. By assumption there exists a choice function $G : \mathcal{F} \to A$ on $\mathcal{F}$. Put $g(b) := G(f^{-1}[\{b\}])$, then $f(g(b)) = b$. $\dashv$

So this is a pure, simple, and direct application of ($\mathbb{AC}$), making one wonder what application the existence of a choice function will find. We will see.

1.2 Cantor's Enumeration of $\mathbb{N} \times \mathbb{N}$

We will deal in this section with the comparison of sets with respect to their size. We say that two sets A and B have the same *cardinality* iff there exists a bijection between them. This condition can sometimes be relaxed by saying that there exists an injective map $f : A \to B$ and an injective map $g : B \to A$. Intuitively, A and B have the same size, since the image of each set is contained in the other one. So we would expect that there exists a bijection between A and B. This is what the famous Schröder–Bernstein Theorem says.

Schröder-Bernstein Theorem

Theorem 1.2.1 *Let $f : A \to B$ and $g : B \to A$ be injective maps. Then there exists a bijection $h : A \to B$.*

Proof Define recursively

$$A_0 := A \setminus g[B],$$
$$A_{n+1} := g[f[A_n]]$$

and

$$B_n := f[A_n].$$

If $a \in A$ with $a \notin A_0$, there exists a unique $b =: g^*(a)$ such that $a = g(b)$, because g is an injection. Now define the map $h : A \to B$ through

$$h(a) := \begin{cases} f(a), & \text{if } a \in \bigcup_{n\geq 0} A_n \\ g^*(a), & \text{otherwise.} \end{cases}$$

Assume that $h(a) = h(a')$. If $a, a' \in \bigcup_{n\geq 0} A_n$, we may conclude that $a = a'$, since f is one to one. If $a \in A_n$ for some n and $a' \notin \bigcup_{n\geq 0} A_n$, then $h(a) = f(a), h(a') = g^*(a')$; hence $a' = g(h(a')) = g(h(a)) = g(f(a))$. This implies $a \in A_{n+1}$, contrary to our assumption. Hence h is an injection. If $b \in \bigcup_{n\geq 0} B_n$, then $b = f(a) = h(a)$. Now let

$b \notin \bigcup_{n \geq 0} B_n$. We claim that $g(b) \notin A_n$ for any $n \geq 0$. In fact, if $g(b) \in A_n$ with $n > 0$, we know that $g(b) = g(f(a))$ for some $a \in A_{n-1}$, so $b = f(a) \in f\left[A_{n-1}\right]$, contrary to our assumption. Hence $h(g(b)) = g^*(g(b)) = b$. Thus h is also onto. $\dashv$

Another proof will be suggested in Exercise 1.7 through a fixed point argument.

This is a first application of the Schröder–Bernstein Theorem.

Example 1.2.2 Let $X = \mathbb{N}$. If there exists an injection $\mathcal{P}(\mathbb{N}) \to \mathbb{N}$, then the Schröder–Bernstein Theorem implies that there exists a bijection $f : \mathbb{N} \to \mathcal{P}(\mathbb{N})$, because we have the injective map $\mathbb{N} \ni x \mapsto \{x\} \in \mathcal{P}(\mathbb{N})$. Now look at $A := \{x \in \mathbb{N} \mid x \notin f(x)\}$. Then there exists $a \in \mathbb{N}$ with $A = f(a)$. But $a \in A$ iff $a \notin A$, and thus there cannot exist an injection $\mathcal{P}(\mathbb{N}) \to \mathbb{N}$. ✌

✌ is *end of example*.

Call a set A *countably infinite* iff there exists a bijection $A \to \mathbb{N}$. By the Schröder–Bernstein Theorem 1.2.1, it then suffices to find an injective map $A \to \mathbb{N}$ and an injective map $\mathbb{N} \to A$. A set is called *countable* iff it is either finite or countably infinite. Example 1.2.2 tells us that $\mathcal{P}(\mathbb{N})$ is not countable.

We will have a closer look at countably infinite sets now and show that the set of all finite sequences of natural numbers is countable; for simplicity, we work with $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. $\mathbb{N}_0$

We start with showing that there exists a bijection from the Cartesian product $\mathbb{N}_0 \times \mathbb{N}_0 \to \mathbb{N}_0$. Cantor's celebrated procedure for producing an enumeration for $\mathbb{N}_0 \times \mathbb{N}_0$ works for an initial section as follows:

$$
\begin{tikzcd}
\langle 0,0\rangle \arrow[r] & \langle 0,1\rangle \arrow[dl] & \langle 0,2\rangle \arrow[dl] & \langle 0,3\rangle \arrow[dl] & \dots \\
\langle 1,0\rangle \arrow[urr] & \langle 1,1\rangle \arrow[dl] & \langle 1,2\rangle \arrow[dl] & \langle 1,3\rangle \arrow[dl] & \dots \\
\langle 2,0\rangle \arrow[uurrr] & \langle 2,1\rangle \arrow[dl] & \langle 2,2\rangle \arrow[dl] & \langle 2,3\rangle \arrow[dl] & \dots \\
\langle 3,0\rangle & \langle 3,1\rangle & \langle 3,2\rangle & \langle 3,3\rangle & \dots
\end{tikzcd}
\qquad
\begin{tikzcd}
0 \arrow[r] & 1 \arrow[dl] & 3 \arrow[dl] & 6 \arrow[dl] & \dots \\
2 \arrow[urr] & 4 \arrow[dl] & 7 \arrow[dl] & 11 \arrow[dl] & \dots \\
5 \arrow[uurrr] & 8 \arrow[dl] & 12 \arrow[dl] & 17 \arrow[dl] & \dots \\
9 & 13 & 18 & 24 & \dots
\end{tikzcd}
$$

Define the function

$$J(x, y) := \binom{x + y + 1}{2} + x;$$

then an easy computation shows that this yields just the enumeration scheme of Cantor's procedure. We will have a closer look at J now; note that the function $x \mapsto \binom{x}{2}$ increases monotonically.

Proposition 1.2.3 *$J : \mathbb{N}_0 \times \mathbb{N}_0 \to \mathbb{N}_0$ is a bijection.*

Proof 1. J is injective. We show first that $J(a,b) = J(x,y)$ implies $a = x$. Assume that $x > a$; then x can be written as $x = a + r$ for some positive r, so

$$\binom{a+r+y+1}{2} + r = \binom{a+b+1}{2};$$

hence $b > r + y$, so that b can be written as $b = r + y + s$ with some positive s. Abbreviating $c := a + r + y + 1$, we obtain

$$\binom{c}{2} + r = \binom{c+s}{2}.$$

But because we have $r < c$, we get

$$\binom{c}{2} + r < \binom{c}{2} + c = \binom{c+1}{2} \leq \binom{c+s}{2}.$$

This is a contradiction. Hence $x \leq a$. Interchanging the rôles of x and a, one obtains $a \leq x$, so that $x = a$ may be inferred.

Thus we obtain

$$\binom{a+y+1}{2} = \binom{a+b+1}{2}.$$

This yields the quadratic equation

$$y^2 + 2ay - (b^2 + 2ab) = 0$$

which has the solutions b and $-(2a + b)$. If $a = b = 0$, then $y = 0 = b$, if $b > 0$; the only nonnegative solution is b, so that $y = b$ also in this case. Hence we have shown that $J(a,b) = J(x,y)$ implies $\langle a,b\rangle = \langle x,y\rangle$.

2. J is onto. Define $Z := J\big[\mathbb{N}_0 \times \mathbb{N}_0\big]$, then $0 = J(0,0) \in Z$ and $1 = J(0,1) \in Z$. Assume that $n \in Z$, so that $n = J(x,y)$ for some $\langle x,y\rangle \in \mathbb{N}_0$. We consider these cases

$y > 0:\ n+1 = J(x,y)+1 = \binom{x+y+1}{2}+x+1 = J(x+1,y-1) \in Z.$

$y = 0:\ n = \binom{x}{2} + x = \binom{x+1}{2}$, and thus $n + 1 = \binom{x+1}{2} + 1$.

$x > 0:\ n+1 = \binom{x+1}{2}+1 = \binom{1+(x-1)+1}{2}+1 = J(1, x-1) \in Z.$

$x = 0:$ Then $n = 0$, so that $n + 1 = J(0, 1) \in Z$.

Thus we have shown that $0 \in Z$ and that $n \in Z$ implies $n + 1 \in Z$, from which we infer $Z = \mathbb{N}_0$. ⊣

This construction permits the construction of an enumeration for the set of all nonempty sequences of elements of $\mathbb{N}_0$. First we have a look at sequences of fixed length. For this, define inductively

$$\mathrm{t}_1(x) := x,$$
$$\mathrm{t}_{k+1}(x_1, \ldots, x_k, x_{k+1}) := J(\mathrm{t}_k(x_1, \ldots, x_k), x_{k+1})$$

($x \in \mathbb{N}_0$ and $k \in \mathbb{N}$, $\langle x_1, \ldots, x_{k+1}\rangle \in \mathbb{N}_0^{k+1}$), the idea being that an enumeration of $\mathbb{N}^k \times \mathbb{N}$ is reduced to an enumeration of $\mathbb{N} \times \mathbb{N}$, an enumeration of which in turn is known. Idea

Proposition 1.2.4 *The maps* t_k *are bijections* $\mathbb{N}_0^k \to \mathbb{N}_0$.

Proof 1. The proof proceeds by induction on k. It is trivial for $k = 0$. Now assume that we have established bijectivity for $\mathrm{t}_k : \mathbb{N}_0^k \to \mathbb{N}_0$.

2.t_{k+1} is injective: Assume $\mathrm{t}_{k+1}(x_1, \ldots, x_k, x_{k+1}) = \mathrm{t}_{k+1}(x_1', \ldots, x_k', x_{k+1}')$; this means

$$J(\mathrm{t}_k(x_1, \ldots, x_k), x_{k+1}) = J(\mathrm{t}_k(x_1', \ldots, x_k'), x_{k+1}'),$$

and hence $\mathrm{t}_k(x_1, \ldots, x_k) = \mathrm{t}_k(x_1', \ldots, x_k')$ and $x_{k+1} = x_{k+1}'$ by Proposition 1.2.3. By induction hypothesis, $\langle x_1, \ldots, x_k\rangle = \langle x_1', \ldots, x_k'\rangle$.

3.t_{k+1} is onto: Given $n \in \mathbb{N}_0$, there exists $\langle a, b\rangle \in \mathbb{N}_0 \times \mathbb{N}_0$ with $J(a, b) = n$. Given $a \in \mathbb{N}_0$, there exists $\langle x_1, \ldots, x_k\rangle \in \mathbb{N}_0^k$ with $\mathrm{t}_k(x_1, \ldots, x_k) = a$ by induction hypothesis, so

$$n = J(a, b) = J(\mathrm{t}_k(x_1, \ldots, x_k), b) = \mathrm{t}_{k+1}(x_1, \ldots, x_k, b).$$

⊣

From this, we can manufacture a bijection

$$\bigcup_{k \in \mathbb{N}} \mathbb{N}_0^k \to \mathbb{N}_0$$

in the following way. Given a finite sequence v of natural numbers, we use its length, say, k, as one parameter of an enumeration of $\mathbb{N} \times \mathbb{N}$, and the other parameter for this enumeration is $\mathrm{t}_k(v)$. This yields a bijection.

Proposition 1.2.5 *There exists a bijection* $\mathrm{s} : \bigcup_{k \in \mathbb{N}} \mathbb{N}_0^k \to \mathbb{N}_0$.

Proof Define

$$\mathrm{s}(x_1, \ldots, x_k) := J(k, \mathrm{t}_k(x_1, \ldots, x_k))$$

for $k \in \mathbb{N}$ and $\langle x_1, \ldots, x_k \rangle \in \mathbb{N}_0^k$. Because both J and t_k are injective, s is injective. Given $n \in \mathbb{N}_0$, we can find $\langle a, b \rangle \in \mathbb{N}_0 \times \mathbb{N}_0$ with $J(a, b) = n$. Given $b \in \mathbb{N}_0$, we can find $\langle x_1, \ldots, x_a \rangle \in \mathbb{N}_0^a$ with $\mathrm{t}_a(x_1, \ldots, x_a) = b$, so that

$$n = J(a, b) = J(a, \mathrm{t}_a(x_1, \ldots, x_a)).$$

Hence s is also surjective. ⊣

One wonders why we did go through this somewhat elaborate construction. First, the construction is elegant in its simplicity, but there is another, more subtle reason. When tracing the arguments leading to Proposition 1.2.5, one sees that the argumentation is elementary; it does not require any set-theoretic assumptions like $(\mathbb{AC})$. But now look at this:

Proposition 1.2.6 *Let* $\{A_n \mid n \in \mathbb{N}_0\}$ *be a sequence of countably infinite sets. Then* $(\mathbb{AC})$ *implies that* $\bigcup_{n \in \mathbb{N}_0} A_n$ *is countable.*

Proof We assume for simplicity that the A_n are mutually disjoint. Given $n \in \mathbb{N}_0$, there exists an enumeration $\psi_n : A_n \to \mathbb{N}_0$. $(\mathbb{AC})$ permits us to fix for each n such an enumeration ψ_n; then define

$$\psi : \begin{cases} \bigcup_{n \in \mathbb{N}_0} A_n & \to \mathbb{N}_0 \\ x & \mapsto J(k, \psi_k(x)), \text{ if x} \in A_k \end{cases}$$

with J as the bijection defined in Proposition 1.2.3. ⊣

This is somewhat puzzling at first; but note that the proof of Proposition 1.2.5 does not require a selection argument, because we are in a position to construct t_k for all $k \in \mathbb{N}$.

Having $(\mathbb{AC})$, hence Proposition 1.2.6 at our disposal, one shows by induction that

$$\mathbb{N}_0^{k+1} = \bigcup_{n \in \mathbb{N}_0} \mathbb{N}_0^k \times \{n\}$$

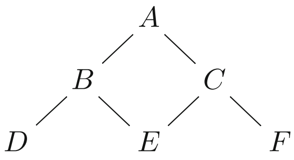
is countable for every $k \in \mathbb{N}$. This establishes the countability of $\bigcup_{k \in \mathbb{N}} \mathbb{N}_0^k$ immediately. On the other hand it can be shown that Proposition 1.2.6 is not valid if $(\mathbb{AC})$ is not assumed [KM76, p. 172] or [Her06, Sect. 3.1]. This is also true if $(\mathbb{AC})$ is weakened somewhat to postulate the existence of a choice function for *countable* families of nonempty sets (which in our case would suffice). The proof of nonvalidity, however, is in either case far beyond our scope.

1.3 Well-Ordered Sets

A relation R on a set M is called an *order relation* iff it is *reflexive* (thus xRx holds for all $x \in M$), *antisymmetric* (this means that xRy and yRx imply $x = y$ for all $x, y \in M$), and *transitive* (hence xRy and yRz imply xRz for all $x, y, z \in M$). The relation R is called *linear* iff one of the cases $x = y$, xRy, or yRx applies for all $x, y \in M$, and it is called *strict* iff xRx is false for each $x \in M$. If R is strict and transitive, then it is called a *strict order*.

Let R be an order relation; then $x \in M$ is called a *lower bound* for $\emptyset \neq A \subseteq M$ iff xRz holds for all $z \in M$ and a *smallest element* for A iff it is both a lower bound for A and a member of A. *Upper bounds* and *largest elements* are defined similarly. An element y is called *maximal* iff there exists no element x with yRx; *minimal* elements are defined similarly. A *minimal upper bound* for a set $A \neq \emptyset$ is called the *supremum* of A and is denoted by $\sup A$; similarly, a *maximal lower bound* for A is called the *infimum* of A and is denoted by $\inf A$. Neither infimum nor supremum of a nonempty set needs to exist. sup, inf

Example 1.3.1 Look at this ordered set:



Here A is the maximum, because every element is smaller than A; the minimal elements are D, E, and F, but there is no minimum. The minimal elements cannot be compared to each other.

✌

Example 1.3.2 Define $a \leq_d b$ iff a divides b for $a, b \in \mathbb{N}$; thus $a \leq_d b$ iff there exists $k \in \mathbb{N}$ such that $b = k \cdot a$. Let g be the greatest common divisor of a and b, then $g = \inf\{a, b\}$, and if s is the smallest common multiple of a and b, then $s = \sup\{a, b\}$. Here is why: One notes first that both $g \leq_d a$ and $g \leq_d b$ hold, because g is a common divisor of a and b. Let g' be another common divisor of a and b, and then one shows easily that g' divides g, so that $g' \leq_d g$ holds. Thus g is in fact the greatest common divisor. One argues similarly for the lowest common multiple of a and b. ✌

Example 1.3.3 Order $S := \mathcal{P}(\mathbb{N}) \setminus \{\mathbb{N}\}$ by inclusion. Then $\mathbb{N} \setminus \{k\}$ is maximal in S for every $k \in \mathbb{N}$. We obtain from the definition of S and its order that each element which contains $\mathbb{N} \setminus \{k\}$ properly would be outside the basic set S. The set $A := \{\{n, n+2\} \mid n \in \mathbb{N}\}$ is unbounded in S. Assume that T is an upper bound for A; then $n \in \{n, n+2\} \subseteq T$ and for each $n \in \mathbb{N}$, so that $T = \mathbb{N} \notin S$. ✌

Usually strict orders are written as $<$ (or $<_M$, if the basis set is to be emphasized) and order relations as $\leq$ or $\leq_M$, resp.

Let $<_M$ be a strict order on M and $<_N$ be a strict order on N; then a map $f : M \to N$ is called *increasing* iff $x <_M y$ implies $f(x) <_N f(y)$; M and N are called *similar* iff f is a bijection such that $x <_M y$ is equivalent to $f(x) <_N f(y)$. An *order isomorphism* is a bijection which together with its inverse is increasing.

Definition 1.3.4 *The strict linear order $<$ on a set M is called a* well ordering *on M iff each nonempty subset of M has a smallest element. M is then called* well ordered *(under $<$).*

These are simple examples of well-ordered sets.

Example 1.3.5 $\mathbb{N}$ (this shows the special rôle of $\mathbb{N}$ alluded to above), finite linearly ordered sets, and $\{1 - \frac{1}{n} \mid n \in \mathbb{N}\}$ are well ordered. ✌

Not every ordered set, however, is well ordered, witnessed by these simple examples.

Example 1.3.6 $\mathbb{Z}$ is not well ordered, because it does not have a minimal element. $\mathbb{R}$ is neither, because, e.g., the open interval $]0, 1[$ does not have a smallest element. The power set of $\mathbb{N}$, denoted by $\mathcal{P}(\mathbb{N})$, is not well ordered by inclusion because a well order is linear, and $\{1, 2\}$ and $\{3, 4\}$ are not comparable. Finally, $\{1 + \frac{1}{n} \mid n \in \mathbb{N}\}$ is not well ordered, because the set does not contain a smallest element. ✌

Example 1.3.7 A *reduction system* $\mathcal{R} = (A, \rightarrow)$ is a set A with a relation $\rightarrow \subseteq A \times A$; the intent is to have a set of rewrite rules, say, $\langle a, b\rangle \in \rightarrow$ such that a may be replaced by b in words over an alphabet which includes the carrier A of $\mathcal{R}$. Usually, one writes $a \rightarrow b$ iff $\langle a, b\rangle \in \rightarrow$. Denote by $\stackrel{+}{\rightarrow}$ the reflexive-transitive closure of relation $\rightarrow$, i.e., $x \stackrel{+}{\rightarrow} y$ iff $x = y$ or there exists a chain $x = a_0 \rightarrow \ldots \rightarrow a_k = y$.

We call $\mathcal{R}$ *terminating* iff there are no infinite chains $a_0 \rightarrow a_1 \rightarrow \ldots \rightarrow a_k \rightarrow \ldots$. The following proof rule is associated with a reduction system:

$$\text{(WFI)}\ \frac{\forall x \in A : \big(\forall y \in A : x \stackrel{+}{\rightarrow} y \Rightarrow P(y)\big) \Rightarrow P(x)}{\forall x \in A : P(x).}$$

Here P is a predicate on A so that $P(x)$ is true iff x has property P. The rule (WFI) says that if we can conclude for every x that $P(x)$ holds, provided the property holds for all predecessors of x, then we may conclude that P holds for each element of A.

This rule is equivalent to termination. In fact

- If $\rightarrow$ terminates, then (WFI) holds. Assume that (WFI) does not hold, then we find $x_0 \in A$ such that $P(x_0)$ does not hold, hence we can find some x_1 with $x_0 \stackrel{+}{\rightarrow} x_1$ and $P(x_1)$ does not hold. For x_1 we find x_2 for which P does not hold with $x_1 \stackrel{+}{\rightarrow} x_2$, etc. Hence we construct an infinite chain $x_0 \stackrel{+}{\rightarrow} x_1 \stackrel{+}{\rightarrow} \ldots$ of elements for which P does not hold. But this means that $\rightarrow$ does not terminate.

- If (WFI) holds, then $\rightarrow$ terminates. Take the predicate $P(x)$ iff *there is no infinite chain starting from* x. Now (WFI) says that if $y \stackrel{+}{\rightarrow} x$, and if $P(y)$ holds, then $P(x)$ holds. This means that no infinite chain starts from y, and x is a successor to y, and so no infinite chain starts from x either. Hence, to conclude this rule, no x is the starting point of an infinite chain; consequently $\rightarrow$ terminates.

Now let $(A, \rightarrow)$ be a terminating reduction system; then each nonempty subset $B \subseteq A$ has a minimal element, because if this is not the case, we can construct an infinite descending chain. But $(A, \rightarrow)$ is usually not well ordered, because $\stackrel{+}{\rightarrow}$ is not necessarily strict. ✌

There are some helpful ways of producing a new well order from old ones.

Example 1.3.8 Let M and N be well-ordered and disjoint sets, define on $M \cup N$

$$a < b \text{ iff } \begin{cases} a <_M b, & \text{if } a, b \in M, \\ a <_N b, & \text{if } a, b \in N, \\ a \in M, b \in N, & \text{otherwise.} \end{cases}$$

Then $M \cup N$ is well ordered; this well-ordered set is usually denoted by $M + N$. Note that $M + N$ is not the same as $N + M$.

If the sets are not disjoint, make a copy of each upon setting $M' := M \times \{1\}, N' := N \times \{2\}$, and order these sets through, e.g., $\langle m, 1\rangle <_{M'} \langle m', 1\rangle$ iff $m <_M m'$. ✌

Example 1.3.9 Define on the Cartesian product $M \times N$

$$\langle m, n\rangle < \langle m', n'\rangle \text{ iff } \begin{cases} m < m' \\ n < n', & \text{if } m = m'. \end{cases}$$

This *lexicographic order* yields a well ordering again. ✌

Example 1.3.10 Let Z be well ordered, and assume that for each $z \in Z$ the set M_z is well ordered so that the sets $(M_z)_{z\in Z}$ are mutually disjoint. Then $\bigcup_{z\in Z} M_z$ is well ordered. ✌

Having a look at $(\mathbb{AC})$ again, we see that it holds in a well-ordered set.

Proposition 1.3.11 *Let $\mathcal{F}$ be a family of nonempty subsets of the well-ordered set M. Then there exists a choice function on $\mathcal{F}$.*

Proof For each $F \in \mathcal{F}$, there exists a smallest element $m_F \in \mathcal{F}$. Put $f(F) := m_F$; then $f : \mathcal{F} \to M$ is a choice function on $\mathcal{F}$. $\dashv$

Thus, if we can find a well order on a set, then we know that we can find choice functions. We formulate first this property.

$$\boxed{(\mathbb{WO}) \text{ Each set can be well ordered.}}$$

We will refer to this property as $(\mathbb{WO})$. Hence we can rephrase Proposition 1.3.11 as

$$(\mathbb{WO}) \Longrightarrow (\mathbb{AC}).$$

Establishing the converse will turn out to be more interesting, since it will require the introduction of a new class of objects, viz., the ordinal numbers. This is what we will undertake in the next section.

We start with some preparations which deal with properties of well orders.

Lemma 1.3.12 *Let M be well ordered and $f : M \to M$ be an increasing map. Then $x \leq f(x)$ (thus $x < f(x)$ or $x = f(x)$) holds for all $x \in M$.*

Proof Suppose that the set $Y := \{y \in M \mid f(y) < y\}$ is not empty; then it has a smallest element z. Since $f(z) < z$, we obtain $f(f(z)) < f(z) < z$, because f is increasing. This contradicts the choice of z. $\dashv$

Let M be well ordered; then define for $x \in M$ the *initial segment* $O(x)$ (or $O_M(x)$) for x as $O(x) := \{z \in M \mid z < x\}$. $O(x), O_M(x)$

We obtain as a consequence

Corollary 1.3.13 *No well-ordered set is order isomorphic to an initial segment of itself.*

Proof An isomorphism $f : M \to O_M(x)$ for some $x \in M$ would have $f(x) < x$, which contradicts Lemma 1.3.12. $\dashv$

A surprising consequence of Lemma 1.3.12 is that there exists at most one isomorphism between well-ordered sets.

Corollary 1.3.14 *Let A and B be well-ordered sets. If $f : A \to B$ and $g : A \to B$ are order isomorphisms, then $f = g$.*

Proof Clearly both $g^{-1} \circ f$ and $f^{-1} \circ g$ are increasing, yielding $x \leq (g^{-1} \circ f)(x)$ and $x \leq (f^{-1} \circ g)(x)$ for each $x \in A$, which means $g(x) \leq f(x)$ and $f(x) \leq g(x)$ for each $x \in A$. $\dashv$

This is an important property of well-ordered sets akin to induction in the set of natural numbers. Accordingly it is called the *principle of transfinite induction*, sometimes also called *Noetherian induction* (after the eminent German mathematician Emmy Noether) or *well-founded induction* (after the virtually unknown Chinese mathematician Wēl Fǔn Dèd).

Theorem 1.3.15 *Let M be well ordered and $B \subseteq M$ be a set which has for each $x \in M$ the property that $O(x) \subseteq B$ implies $x \in B$. Then $B = M$.*

Proof Assume that $M \setminus B \neq \emptyset$; then there exists a smallest element x in this set. Since x is minimal, all elements smaller than x are elements of B; hence $O(x) \subseteq B$. But this implies $x \in B$, a contradiction. $\dashv$

Let us have a second look at a proof of Lemma 1.3.12, this time using the principle of transfinite induction. Put $B := \{z \in M \mid z \leq f(z)\}$, and assume $O(x) \subseteq B$. If $y \in B$ with $y \neq f(y)$, then $y < f(y)$ and $y < x$, so that $f(y) < f(x)$, and thus $y < f(x)$. Hence $f(x)$ is larger than any element of $O(x)$, and thus $f(x) \in M \setminus O(x)$. But x is the smallest element of the latter set, which implies $x < f(x)$, so $x \in B$. From Theorem 1.3.15, we see now that $B = M$.

We will show now that each set can be well ordered. In order to do this, we construct a prototypical well order and show that each set can be mapped bijectively to this set. This then will serve as the basis for the construction of a well order for this set.

Carrying out this program requires the prototype. This will be considered next.

1.4 Ordinal Numbers

Following von Neumann [KM76, §VII.9], ordinal numbers are defined as sets with these special properties.

Definition 1.4.1 *A set α is called an* ordinal number *iff these conditions are satisfied:*

① *Every element of α is a set.*

② *If $\beta \in \alpha$, then $\beta \subseteq \alpha$.*

③ *If $\beta, \gamma \in \alpha$, then $\beta = \gamma$ or $\beta \in \gamma$ or $\gamma \in \beta$.*

④ *If $\emptyset \neq B \subseteq \alpha$, then there exists $\gamma \in B$ with $\gamma \cap B = \emptyset$.*

Hence in order to show that a given set is an ordinal, we have to show that the properties ①, ②, ③, and ④ hold. We will demonstrate this principle for some examples. Obligation

Example 1.4.2 Consider this definition of the *somewhat natural numbers* $\mathfrak{N}_0$

$$
\begin{aligned}
0 &:= \emptyset, \\
n + 1 &:= \{0, \ldots, n\}, \\
\mathfrak{N}_0 &:= \{0, 1, \ldots\}.
\end{aligned}
$$

Then $\mathfrak{N}_0$ is an ordinal number. Each element of $\mathfrak{N}_0$ is a set by definition. Let $\beta \in \mathfrak{N}_0$. If $\beta = 0$, $\beta = \emptyset \subseteq \mathfrak{N}_0$, and if $\beta \neq 0$, $\beta = n = \{0, \ldots, n-1\} \subseteq \mathfrak{N}_0$. One argues similarly for property ③. Finally, let $\emptyset \neq \beta \subseteq \mathfrak{N}_0$, and let γ be the smallest element of β. If $\delta \in \gamma \cap \beta$, then δ is both an element of β and smaller than γ, and this is a contradiction. Hence $\gamma \cap \delta = \emptyset$. ✌

Example 1.4.3 Let α be an ordinal number, and then $\beta := \alpha \cup \{\alpha\}$ is an ordinal. It is the smallest ordinal which is greater than α. Property ① is evident, so is property ②. Let $\gamma, \gamma' \in \beta$ with $\gamma \neq \gamma'$, and assume that $\gamma = \alpha$, then $\gamma' \neq \alpha$, and consequently $\gamma' \in \gamma$. If neither γ nor γ' is equal to α, property ③ trivially holds. Assume finally that $\emptyset \neq B \subseteq \beta$. If $B \cap \alpha \neq \emptyset$, property ④ for β follows from this property for α; if, however, $B = \{\alpha\}$, observe that $\alpha \in \beta$ with $B \cap \alpha = \emptyset$. Hence this property holds for β as well. ✌

$\alpha + 1$

Definition 1.4.4 *Let α be an ordinal, and then $\alpha \cup \{\alpha\}$ is called the* successor to α *denoted by* $\alpha + 1$.

It is clear from this definition that no ordinal can be squeezed in between an ordinal α and its successor $\alpha + 1$.

Lemma 1.4.5 *If M is a nonempty set of ordinals, then*

1. *$\alpha_* := \bigcap M$ is an ordinal; it is the largest ordinal contained in all elements of M.*
2. *$\alpha^* := \bigcup M$ is an ordinal; it is the smallest ordinal which contains all elements of M.*

Proof We iterate over the defining properties of an ordinal number for $\bigcap M$. Since every element γ of $\bigcap M$ is also an element of every $\alpha \in M$, we may conclude that γ is a set and that $\gamma \subseteq \bigcap M$. If $\gamma, \delta \in \bigcap M \subseteq \alpha$ for each $\alpha \in M$, we have either $\gamma = \delta$, $\gamma \in \delta$ or $\delta \in \gamma$. Finally, if $\emptyset \neq B \subseteq \bigcap M \subseteq \alpha$ for each $\alpha \in M$, we find $\eta \in B$ such that $\eta \cap B = \emptyset$. Thus $\alpha_* := \bigcap M$ has all the properties of an ordinal number from Definition 1.4.1. It is clear that α_* is the largest ordinal contained in all elements of M.

The proof for $\bigcup M$ works along the same lines. $\dashv$

Corollary 1.4.6 *Given a nonempty set M of ordinals, there is always an ordinal which is strictly larger than all the elements of M.*

Proof If $\alpha^* := \bigcup M \in M$, then $\alpha^* + 1$ is the desired ordinal; otherwise α^* is suitable. $\dashv$

This is an interesting consequence.

Corollary 1.4.7 *There is no set of all ordinals.*

Proof If Z is the set of all ordinals, then Lemma 1.4.5 shows that $\alpha^* := \bigcup Z$ is an ordinal. But the successor $\alpha^* + 1$ to α^* is an ordinal as well by Example 1.4.3, which, however, is not an element of Z. This is a contradiction. $\dashv$

Limit ordinal

Definition 1.4.8 *An ordinal λ is called a* limit ordinal *iff $\alpha < \lambda$ implies $\alpha + 1 < \lambda$ for all ordinals α.*

Thus a limit ordinal is not reachable through the successor operation. This is a convenient characterization of limit ordinals.

Proposition 1.4.9 *Let λ be an ordinal. Then*

1. *If λ is a limit ordinal, then $\bigcup \lambda = \lambda$.*
2. *If $\bigcup \lambda = \lambda$, then λ is a limit ordinal.*

Proof 1. Assume first that λ is a limit ordinal. Let $\beta \in \bigcup \lambda$, and then $\beta \in \alpha$ for some $\alpha \in \lambda$. Since α is an ordinal, we conclude $\beta \in \alpha \subseteq \lambda$, so $\bigcup \lambda \subseteq \lambda$. On the other hand, if $\alpha \in \lambda$, then $\alpha + 1 \in \lambda$, since λ is a limit ordinal. Thus $\alpha \in \bigcup \lambda$, so $\bigcup \lambda \supseteq \lambda$. This proves part 1.

2. Let $\alpha < \lambda = \bigcup \lambda$, and then $\alpha \in \beta$ for some $\beta \in \lambda$. Then either $\alpha + 1 \in \beta$ or $\alpha + 1 = \beta$, in any case $\alpha + 1 \subseteq \beta$, so that $\alpha + 1 \in \lambda$. Thus λ is a limit ordinal. This establishes part 2. $\dashv$

Ordinals can be *odd* or *even*: A limit ordinal is said to be even; if the ordinal ζ can be written as $\zeta = \xi + 1$ and ξ is even, then ζ is odd, and if ξ is odd, ζ is even. This classification is sometimes helpful, and some constructions involving ordinals depend on it; see, for example, Sect. 1.6.1 on page 69. Odd, even

Several properties of ordinal numbers are established now; this is required for carrying out the program sketched above. The first property states that the $\in$-relation is not cyclic, which seems to be trivial. But since ordinal numbers have the dual face of being elements and subsets of the same set, we will need to exclude this property explicitly by showing that the properties of ordinals prevent this undesired behavior.

Lemma 1.4.10 *If α is an ordinal number, then there does not exist a sequence $\beta_1, \ldots, \beta_k$ of sets with $\beta_k \in \beta_1 \in \ldots \beta_{k-1} \in \beta_k \in \alpha$.*

Proof If there exist such sets $\beta_1, \ldots, \beta_k$, put $\gamma := \{\beta_1, \ldots, \beta_k\}$; then γ is the smallest ordinal containing $\beta_1, \ldots, \beta_k$. Now $\beta_k \in \alpha$ implies $\beta_k \subseteq \alpha$, thus $\beta_{k-1} \in \alpha$, and hence $\beta_{k-1} \subseteq \alpha$, so that $\beta_1, \ldots, \beta_k \in \alpha$. But now $\beta_{i-1} \in \beta_i \cap \gamma$ for $2 \leq i \leq k$ and $\beta_k \in \beta_1 \cap \gamma$, so that property ④ in Definition 1.4.1 is violated. Hence γ is not an ordinal at all. $\dashv$

Lemma 1.4.11 *If α is an ordinal, then each $\beta \in \alpha$ is an ordinal as well.*

Proof 1. The properties of ordinal numbers from Definition 1.4.1 are inherited. This is immediate for properties ①, ③, and ④, so we have to take care of property ②.

2. Let $\gamma \in \beta$, and we have to show that $\gamma \subseteq \beta$. So if $\eta \in \gamma$, we have by property ③ for α in the definition of ordinals either $\eta = \gamma$ (which would imply $\gamma \in \gamma \in \beta \in \alpha$, contradicting Lemma 1.4.10) or $\gamma \in \eta$ (which would yield $\gamma \in \eta \in \gamma \in \beta \in \alpha$, contradicting Lemma 1.4.10 again). Thus $\eta \in \beta$, so that property ② also holds. $\dashv$

Lemma 1.4.12 *Let α and β be ordinals, and then these properties are equivalent:*

1. $\alpha \in \beta$.
2. $\alpha \subseteq \beta$ *and* $\alpha \neq \beta$.

Proof 1 $\Rightarrow$ 2: We obtain $\alpha \subseteq \beta$ from $\alpha \in \beta$ and from property ② and $\alpha \neq \beta$ from Lemma 1.4.10, for otherwise we could conclude $\beta \in \beta$.

2 $\Rightarrow$ 1: Because α is a proper subset of β, thus $\emptyset \neq \beta \setminus \alpha \subseteq \beta$, and we infer from property ④ for ordinals that we can find $\gamma \in \beta \setminus \alpha$ such that $\gamma \cap \beta \setminus \alpha = \emptyset$. We claim that $\gamma = \alpha$.

"$\subseteq$": Since $\gamma \in \beta$, we know that $\gamma \subseteq \beta$, and since $\gamma \cap \beta \setminus \alpha = \emptyset$, it follows $\gamma \subseteq \alpha$.

"$\supseteq$": We will show that the assumption $\alpha \setminus \gamma \neq \emptyset$ is contradictory. Because $\emptyset \neq \alpha \setminus \gamma \subseteq \alpha$, we find $\eta \in \alpha \setminus \gamma$ with $\eta \cap \alpha \setminus \gamma = \emptyset$. Because $\eta \in \alpha \setminus \gamma \subseteq \alpha \subseteq \beta$, we conclude $\eta \in \beta$. From property ③, we infer that the cases $\eta = \gamma$, $\eta \in \gamma$ and $\gamma \in \eta$ may occur. Look at these cases in turn

- $\eta = \gamma$: This is impossible, because we would have then $\eta \in \alpha$ and $\eta \in \beta \setminus \alpha$.
- $\eta \in \gamma$: This is impossible because $\eta \in \alpha \setminus \gamma$.
- $\gamma \in \eta$: We know that $\gamma \notin \alpha \setminus \gamma$, but $\gamma \in \alpha$, which implies $\gamma \in \gamma \in \alpha$, contradicting Lemma 1.4.10.

Thus we conclude that the assumption $\alpha \setminus \gamma \neq \emptyset$ leads us to a contradiction, from which the desired inclusion is inferred.

$\dashv$

Consequently, the containment relation $\in$ yields a total order on an ordinal number.

Lemma 1.4.13 *If α and β are ordinals, then either $\alpha \subseteq \beta$ or $\beta \subseteq \alpha$. Thus $\alpha \in \beta$ or $\beta \in \alpha$ for $\alpha \neq \beta$.*

Proof Suppose $\alpha \neq \alpha \cap \beta \neq \beta$, then $\alpha \cap \beta \in \alpha$ and $\alpha \cap \beta \in \beta$ by Lemma 1.4.12, and hence $\alpha \cap \beta \in \alpha \cap \beta$, contradicting Lemma 1.4.10. $\dashv$

Since we want to use the ordinals as prototypes for well orders, we have to show that they constitute a well order themselves; inclusion, or, what amounts to be the same, containment suggests itself as an order relation.

Lemma 1.4.14 *Every ordinal is well ordered by the inclusion relation.*

Proof Let α be an ordinal, we show first that α is linearly ordered by inclusion. Take $\beta, \gamma \in \alpha$, then either $\beta = \gamma$, $\beta \in \gamma$, or $\gamma \in \beta$. The last two conditions translate to $\beta \subseteq \gamma$ or $\gamma \subseteq \beta$ because of property ②. Now let B be a nonempty subset of α, and then we know from property ④ that there exists $\gamma \in B$ with $\gamma \cap B = \emptyset$. This is the smallest element of B. In fact, let $\eta \in B$ with $\eta \neq \gamma$; then either $\gamma \in \eta$ or $\eta \in \gamma$. But $\gamma \in \eta$ is impossible, since otherwise $\gamma \in B \cap \eta$. So $\eta \in \gamma$, hence $\eta \subseteq \gamma$. $\dashv$

We can describe this strict order even a bit more precise.

Lemma 1.4.15 *If α and β are distinct ordinals, then either α is an initial segment of β or β is an initial segment of α.*

Proof Because $\alpha \neq \beta$, we have either $\alpha \subseteq \beta$ or $\beta \subseteq \alpha$ by Lemma 1.4.13. Assume that $\alpha \subseteq \beta$ holds. If $\gamma \in \alpha$, then $\gamma \subseteq \alpha$, and thus all elements of α precede the element α; conversely, if $\eta \in \beta$ with $\eta \subseteq \alpha$, then $\eta \in \alpha$. Hence α is a segment of β. It cannot be similar to β because of Corollary 1.3.13. $\dashv$

Historically, ordinal numbers have been introduced as some sort of equivalence classes of well-ordered sets under order isomorphisms (note that *some sort of equivalence classes* is a cautionary expression, alluding to the fact that there is no such thing as a set of all sets). We show now that the present definition is not too far away from the traditional definition. Loosely speaking, the ordinals defined here may serve as representatives for those classes of well-ordered sets. We want to establish

Theorem 1.4.16 *If M is a well-ordered set, then there exists an ordinal α such that M and α are isomorphic.*

Outline for the proof

The proof will be done in several steps. Call two well-ordered sets A and B *similar* ($A \sim B$) iff there exists an isomorphism between them. Recall that isomorphisms preserve order relations in both directions.

Define the set H as all elements of M, the initial segment of which is similar to some ordinal number, i.e.,

$$H := \{z \in M \mid \alpha_z \sim O(z) \text{ for some ordinal } \alpha_z\}.$$

In view of Lemma 1.4.15, if $\alpha_z \sim O(z)$ and $\alpha'_z \sim O(z)$, then $\alpha_z = \alpha'_z$, so the ordinal α_z is uniquely determined, if it exists. We first show by induction that $H = M$. For this, assume that $O(z) \subseteq H$; then we have to show that $z \in H$, so we have to find an ordinal α_z with $\alpha_z \sim O(z)$. In fact, the natural choice is

$$\alpha_z := \{\alpha_x \in M \mid x < z\},$$

so we show that this is an ordinal number by going through the properties according to Definition 1.4.1:

- Since each element of α_z is an ordinal, property ① is satisfied.
- Let $\alpha_x \in \alpha_z$, and then $x < z$; if $\eta \in \alpha_x$, then η is an ordinal number, hence an initial segment of α_x by Lemma 1.4.15; thus $\eta \sim O(t)$ for some t. Hence $t < x < z$, so that $\alpha_t = \eta \in \alpha_z$. Thus property ② is satisfied.
- Property ③ follows from Lemma 1.4.13: Take $\alpha_x, \alpha_y \in \alpha_z$; then α_x and α_y are ordinals. Assume that they are different; then either $\alpha_x \subseteq \alpha_y$ nor $\alpha_y \subseteq \alpha_x$, so that by Lemma 1.4.12 $\alpha_x \in \alpha_y$ or $\alpha_y \in \alpha_x$ follows.
- Finally, let $\emptyset \neq B \subseteq \alpha_x$. Then B corresponds to a nonempty subset of M with a smallest element y. Then $\alpha_y \in \alpha_z$, because $y < z$, and we claim that $\alpha_y \cap B = \emptyset$. In fact, if $\eta \in \alpha_y \cap B$, then $\eta = \alpha_t$ for some $t \in B$, so that y would not be minimal. This shows that Property ④ is satisfied.

Hence α_z is an ordinal. In order to establish that $z \in H$ we have to show that α_z is similar to $O(z)$. But this follows from the construction. Consequently we know that the initial segment for each element of M is similar to an ordinal.

We are now in a position to complete the proof.

Proof (for Theorem 1.4.16) Let

$$\alpha := \{\alpha_z \mid \alpha_z \sim O(z) \text{ for some } z \in M\};$$

then one shows with exactly the arguments from above that α is an ordinal. Moreover, α is similar to M: Consider the map $z \mapsto \alpha_z$, provided $\alpha_z \sim O(z)$. It is clear that it is one to one, since $x < y$ implies $\alpha_x \in \alpha_y$, for $O(x)$ is a (proper) initial segment of $O(y)$. It is also onto, because given $\eta \in \alpha$, we find $z \in M$ with $\eta \sim O(z)$, so that $z \mapsto \eta$. ⊣

Let us have a brief look at all countable ordinals. They will be used later on for a particular construction in Sect. 1.7 for the construction of a game and in Sect. 1.6.1 for the construction of a σ-algebra.

Proposition 1.4.17 *Let $\omega_1 := \{\alpha \mid \alpha$ is a countable ordinal$\}$. Then ω_1 is an ordinal, the first uncountable ordinal.* ω_1

Proof Exercise 1.11; the proof will have to look at the properties ① through ④. ⊣

Denote by $W(\alpha) := \{\zeta \mid \zeta < \alpha\}$ all ordinals smaller than α, hence the initial segment of ordinals determined by α. Given an arbitrary nonempty set S, a map $f : W(\alpha) \to S$ is called an *α-sequence over S* and sometimes denoted by $\langle a_\zeta \mid \zeta < \alpha\rangle$, where $a_\zeta := f(\zeta)$. The next very general statement says that these sequences can be defined by transfinite recursion in the following manner:

Theorem 1.4.18 *Let S be a nonempty set, and let Φ be the set of all α-sequences over S for some ordinal α. Moreover, assume that $h : \Phi \to S$ is a map of α-sequences over S to S. Then there exists a uniquely determined $(\alpha + 1)$-sequence $\langle a_\zeta \mid \zeta \leq \alpha\rangle$ such that*

$$a_\zeta = h(\langle a_\xi \mid \xi < \zeta\rangle)$$

for all $\zeta \leq \alpha$.

Proof 0. The proof works by induction. We show first that the sequence is uniquely determined, and then we define this uniquely determined sequence inductively. Outline

1. We show uniqueness first. Assume that we have two $\alpha + 1$-sequences $\langle a_\zeta \mid \zeta \leq \alpha\rangle$ and $\langle b_\zeta \mid \zeta \leq \alpha\rangle$ such that

$$\begin{aligned} a_\zeta &= h(\langle a_\eta \mid \eta < \zeta\rangle), \\ b_\zeta &= h(\langle b_\eta \mid \eta < \zeta\rangle) \end{aligned}$$

for all $\zeta \leq \alpha$. Then we show by induction on ζ that $a_\zeta = b_\zeta$. The induction begins at the smallest ordinal $\zeta = \emptyset$, so that $a_\emptyset = h(\emptyset) = b_\emptyset$, and the induction step is trivial.

2. The sequence $\langle a_\zeta \mid \zeta \leq \alpha\rangle$ is defined now by induction on ζ. If $\langle \alpha_\eta \mid \eta \leq \zeta\rangle$ is defined, then define

$$\alpha_{\zeta+1} := h(\langle a_\eta \mid \eta \leq \zeta\rangle).$$

If, however, λ is a limit ordinal such that $\langle \alpha_\eta \mid \eta \leq \zeta\rangle$ is defined for each $\zeta < \lambda$, then one notes that $\langle a_\xi \mid \xi < \zeta'\rangle$ is the restriction of $\langle a_\xi \mid \xi < \zeta\rangle$ for $\zeta < \zeta' < \lambda$ by uniqueness, so that

$$\alpha_\lambda := h(\langle a_\zeta \mid \zeta < \lambda\rangle)$$

defines α_λ uniquely. $\dashv$

Idea: exhaust S

We are now in a position to show that the existence of a choice function implies that each set S can be well ordered. The idea of the proof is to find for some suitable ordinal α an α-sequence $\langle a_\zeta \mid \zeta < \alpha\rangle$ over S which exhausts S, so that $S = \{a_\zeta \mid \zeta < \alpha\}$, and then to use the well ordering of the ordinals by saying that $a_\zeta < a_\xi$ iff $\zeta < \xi$.

Constructing the sequence will use the choice function, selecting an element in such a way that one can be sure that it has not been selected previously.

Theorem 1.4.19 *If* $(\mathbb{AC})$ *holds, then each set* S *can be well ordered.*

Proof Let $f : \mathcal{P}(S) \setminus \{\emptyset\} \to S$ be a choice function on the nonempty subset of S. Extend f by putting $f(\emptyset) := p$, where $p \notin S$. This element p serves as an indicator that we are done with constructing the sequence. Let $\mathcal{C}$ be the set of all ordinals ζ such that there exists a well order $<_B$ on a subset $B \subseteq S$ with $(B, <_B)$ similar to $(O(\zeta)$, cp. Theorem 1.4.16. Since $\mathcal{C}$ is a set of ordinals, there exists a smallest ordinal α not in $\mathcal{C}$ by Corollary 1.4.6.

By Theorem 1.4.18, there exists an α-sequence $\langle a_\zeta \mid \zeta < \alpha\rangle$ over S such that

$$a_\zeta := f\big(S \setminus \langle a_\eta \mid \eta < \zeta\rangle\big) \in S \setminus \langle a_\eta \mid \eta < \zeta\rangle$$

for all $\zeta < \alpha$. Now if $S \setminus \langle a_\eta \mid \eta < \zeta\rangle \neq \emptyset$, then $a_\zeta \neq p$, and $a_\zeta \notin \{a_\eta \mid \eta < \zeta\}$, so that the a_ζ are mutually different. Suppose that this process does not exhaust S; then $a_\zeta \neq p$ for all $\zeta < \alpha$. Construct the corresponding well order $<$ on $\{a_\zeta \mid \zeta < \alpha\}$; then $(\{a_\zeta \mid \zeta < \alpha\}, <) \sim O(\alpha)$. Thus $\alpha \in \mathcal{C}$, contradicting the choice of α. Hence there exists a smallest ordinal $\xi < \alpha$ with $a_\xi = p$, which implies that $S = \{a_\zeta \mid \zeta < \xi\}$ so that elements having different labels are in fact different. This yields a well order on S. $\dashv$

Hence we have shown

Theorem 1.4.20 *The following statements are equivalent:*

($\mathbb{AC}$) *The axiom of choice.*

($\mathbb{WO}$) *Each set can be well ordered.*

$\dashv$

($\mathbb{AC}$) has other important and often used equivalent formulations, which we will discuss now.

1.5 Zorn's Lemma and Tuckey's Maximality Principle

Let A be an ordered set, and then $B \subseteq A$ is called a *chain* iff it is linearly ordered, an ordered set in which each chain has an upper bound sometimes called *inductively ordered*. Then Zorn's Lemma states

> ($\mathbb{ZL}$) If A is an ordered set in which every chain has an upper bound, then A has a maximal element.

Proposition 1.5.1 ($\mathbb{ZL}$) *implies* ($\mathbb{AC}$).

Proof 0. Given a family of nonempty sets $\mathcal{F}$, we want to find a choice function for it. In order to apply Zorn's Lemma, we have to put ourselves in a position that we have an ordered set at our disposal in which every chain has an upper bound. We take as the ordered set all functions f which are defined on subsets of $\mathcal{F}$, for which $f(F) \in F$ holds, whenever F is in the domain of f. This set can be ordered in a natural way, and it is then not difficult to see that each chain has an upper bound. Thus we obtain from ($\mathbb{ZL}$) a maximal element, which easily is shown to be a map with $\mathcal{F}$ as its domain. This is the plan: Proof outline

1. Let $\mathcal{F} \neq \emptyset$ be a family of nonempty subsets of a set S; we want to find a choice function on $\mathcal{F}$. Define

$$R := \{\langle F, s\rangle \mid s \in F \in \mathcal{F}\};$$

then $R \subseteq \mathcal{F} \times A$ is a relation. Put

$$\mathcal{C} := \{f \mid f \text{ is a function with } f \subseteq R\}$$

(note that we use functions here as sets of pairs). Then $\mathcal{C} \neq \emptyset$, because $\langle F, s\rangle \in \mathcal{C}$ for each $\langle F, s\rangle \in R$. $\mathcal{C}$ is ordered by inclusion, and each chain has an upper bound in $\mathcal{C}$. In fact, if $K \subseteq \mathcal{C}$ is a chain, then $\bigcup K$ is a map: Let $\langle F, s\rangle, \langle F, s'\rangle \in \bigcup K$; then there exists $f_1, f_2 \in K$ with $\langle F, s\rangle \in f_1, \langle F, s'\rangle \in f_2$. Because K is a chain, either $f_1 \subseteq f_2$ or vice versa; let us assume $f_1 \subseteq f_2$. Thus $\langle F, s\rangle \in f_2$, and since f_2 is a map, we may conclude that $s = s'$. Hence $\bigcup K$ is an upper bound to K in $\mathcal{C}$.

2. By $(\mathbb{ZL})$, $\mathcal{C}$ has a maximal element f^*. We prove that f^* is the desired choice function and hence that there exists for each and every $F \in \mathcal{F}$ some $s \in F$ with $f^*(F) = s$, or, equivalently, $\langle F, s\rangle \in f^*$. Consequently, the domain of f^* should be all of $\mathcal{F}$. Assume that the domain of f^* does not contain some $F^* \in \mathcal{F}$, and then the map $f^* \cup \{\langle F^*, a\rangle\}$ contains for each $a \in F^*$ the map f^* properly. This is a contradiction; thus $f^* : \mathcal{F} \to S$ with $f^*(F) \in F$ for all $F \in \mathcal{F}$. $\dashv$

$(\mathbb{ZL})$: Pattern

We will encounter this pattern over and over again when applying $(\mathbb{ZL})$. We need an ordered set, for which we can establish the chain condition. The maximal element obtained from $(\mathbb{ZL})$ will then have to be checked for its suitability, usually by bringing the assumption that it is not the one we are looking for to a contradiction.

A proof for $(\mathbb{AC}) \Rightarrow (\mathbb{ZL})$ uses a well-ordering argument for constructing a maximal chain.

Proposition 1.5.2 *Assume that A is an ordered set in which each chain has an upper bound, and assume that there exists a choice function on $\mathcal{P}(A) \setminus \{\emptyset\}$. Then A has a maximal element.*

Proof 0. We construct from the choice function a maximal element for A by transfinite induction.

1. As in the proof of Theorem 1.4.19, let $\mathcal{C}$ be the set of all ordinals ζ such that there exists a well order $<_B$ on a subset $B \subseteq A$ with $(B, <_B) \sim O(\zeta)$, and let α be the smallest ordinal not in $\mathcal{C}$; see Corollary 1.4.6. Extend the choice function f on $\mathcal{P}(A) \setminus \{\emptyset\}$ upon setting $f(\emptyset) := p$ with $p \notin A$. This element will again serve as a marker, indicating that the selection process is finished.

2. Define by induction a transfinite sequence $\langle a_\zeta \mid \zeta < \alpha \rangle$ such that $a_\emptyset \in A$ is arbitrary and

$$a_\zeta := f(\{x \in A \mid x > a_\eta \text{ for all } \eta < \zeta\}).$$

Assume that $a_\zeta \neq p$; then $a_\zeta > a_\eta$ for all $\eta < \zeta$. As in the proof of Theorem 1.4.19, there is a smallest ordinal $\beta < \alpha$ such that $a_\beta = p$. The selection process makes sure that $\langle a_\zeta \mid \zeta < \beta \rangle$ is an increasing sequence, and that there does not exists an element $x \in A$ such that $a_\zeta < x$ for all $\zeta < \beta$.

3. Let t be an upper bound for the chain $\langle a_\zeta \mid \zeta < \beta \rangle$. If t is not a maximal element for A, then there exists x with $x > t$; hence $x > a_\zeta$ for all $\zeta < \beta$, which is a contradiction. $\dashv$

Call a subset $\mathcal{F} \subseteq \mathcal{P}(A)$ of *finite character* iff the following condition holds: F is a member of $\mathcal{F}$ iff each finite subset of F is a member of $\mathcal{F}$. The following statement is known as *Tuckey's Lemma* or as *Tuckey's Maximality Principle*:

$$\boxed{\text{(}\mathbb{MP}\text{) Each family of finite character has a maximal element.}}$$

This is another equivalent to ($\mathbb{AC}$).

Proposition 1.5.3 ($\mathbb{MP}$) $\Leftrightarrow$ ($\mathbb{AC}$).

Proof 0. We show ($\mathbb{MP}$) $\Rightarrow$ ($\mathbb{AC}$); the other direction is delegated to the exercises.

1. Let $\mathcal{F} \subseteq \mathcal{P}(S) \setminus \{\emptyset\}$ be a family of nonempty sets. We construct a choice function for $\mathcal{F}$. Consider

$$\mathcal{G} := \{f \mid f \text{ is a choice function for some } \mathcal{E} \subseteq \mathcal{F}\}.$$

Then $\mathcal{G}$ is of finite character. In fact, let f be map from $\mathcal{E} \subseteq \mathcal{F}$ to S such that each finite subset $f_0 \subseteq f$ is a choice function for some $\mathcal{E}_0 \subseteq \mathcal{E}$; then f is itself a choice function for $\mathcal{E}$. Conversely, if $f : \mathcal{E} \to S$ is a choice function for $\mathcal{E} \subseteq \mathcal{F}$, then each finite subset of f is a choice function for its domain. Thus there exists by the Maximality Principle a maximal element $f^* \in \mathcal{G}$. The domain of f^* is all of $\mathcal{F}$, because otherwise f^* could be extended as in the proof of Proposition 1.5.1, and it is clear that f^* is a choice function on $\mathcal{F}$. $\dashv$

Thus we have shown

Theorem 1.5.4 *The following statements are equivalent:*

($\mathbb{AC}$) *The axiom of choice.*

($\mathbb{WO}$) *Each set can be well ordered.*

($\mathbb{ZL}$) *If A is an ordered set in which every chain has an upper bound, then A has a maximal element (Zorn's Lemma).*

($\mathbb{MP}$) *Each family of finite character has a maximal element (Tuckey's Maximality Principle).*

We will discuss some applications of Zorn's Lemma and the Maximality Principle now. From Theorem 1.5.4 we know that in each case we could use also ($\mathbb{AC}$) or ($\mathbb{WO}$), as the case may be. An application of Zorn's Lemma appears sometimes to be more convenient and less technical than using ($\mathbb{WO}$).

1.5.1 Compactness for Propositional Logic

We will show that a set of propositional formulas is satisfiable iff each finite subset is satisfiable. This is usually called the *Compactness Theorem for Propositional Logic*.

Fix a set $V \neq \emptyset$ of variables. A propositional formula φ is given through this grammar

$$\varphi ::= x \mid \varphi \wedge \varphi \mid \neg\varphi$$

with $x \in V$. Hence a formula is either a variable, the conjunction of two formulas, or the negation of a formula. The disjunction $\varphi \vee \psi$ is defined through $\neg(\neg\varphi \wedge \neg\psi)$, implication $\varphi \to \psi$ as $\neg\varphi \vee \psi$, finally $\varphi \leftrightarrow \psi$ is defined through $(\varphi \to \psi) \wedge (\psi \to \varphi)$. Denote by $\mathcal{F}$ the set of all *propositional formulas*—actually, the set of all formulas depends on the set of variables, so we ought to write $\mathcal{F}(V)$; since we fix V, however, we use this somewhat lighter notation.

A *valuation* v evaluates formulas. Instead of using `true` and `false`, we use the values 0 and 1; hence a valuation is a map $V \to \{0, 1\}$ which is extended in a straightforward manner to a map $\mathcal{F} \to \{0, 1\}$, which is again denoted by v:

$$\begin{aligned} v(\varphi_1 \wedge \varphi_2) &:= \min\{v(\varphi_1), v(\varphi_2)\}, \\ v(\neg\varphi) &:= 1 - v(\varphi). \end{aligned}$$

Then we have obviously, e.g.,

$$\begin{aligned} v(\varphi_1 \vee \varphi_2) &= \max\{v(\varphi_1), v(\varphi_2)\}, \\ v(\varphi_1 \to \varphi_2) &= 1 \text{ iff } v(\varphi_1) \leq v(\varphi_2), \\ v(\varphi_1 \leftrightarrow \varphi_2) &= 1 \text{ iff } v(\varphi_1) = v(\varphi_2). \end{aligned}$$

For example,

$$\begin{aligned} v(\varphi \to (\psi \to \gamma)) &= \max\{1 - v(\varphi), \max\{1 - v(\psi), v(\gamma)\}\} \\ &= \max\{1 - v(\varphi), 1 - v(\psi), v(\gamma)\} \\ &= \max\{1 - v(\psi), \max\{1 - v(\varphi), v(\gamma)\}\} \\ &= v(\psi \to (\varphi \to \gamma)). \end{aligned}$$

Hence

$$\big(\varphi \to (\psi \to \gamma)\big) \leftrightarrow \big(\psi \to (\varphi \to \gamma)\big) \leftrightarrow \big((\varphi \wedge \psi) \to \gamma\big).$$

A formula is true for a valuation iff this valuation gives it the value 1; a set $\mathcal{A}$ of formulas is satisfied by a valuation iff each formula in $\mathcal{A}$ is true under this valuation. Formally

Definition 1.5.5 *Let $v : \mathcal{F} \to \{0, 1\}$ be a valuation. Then formula φ is* true for v *(in symbols: $v \models \varphi$) iff $v(\varphi) = 1$. If $\mathcal{A} \subseteq \mathcal{F}$ is a set of propositional formulas, then $\mathcal{A}$ is said to be* satisfied by v *iff each formula in $\mathcal{A}$ is true for v, i.e., iff $v \models \varphi$ for all $\varphi \in \mathcal{A}$. This is written as $v \models \mathcal{A}$.* $v \models \varphi$ $v \models \mathcal{A}$

We are interested in the question whether or not we can find for a set of formulas a valuation satisfying it.

Definition 1.5.6 *$\mathcal{A} \subseteq \mathcal{F}$ is called* satisfiable *iff there exists a valuation $v : \mathcal{F} \to \{0, 1\}$ with $v \models \mathcal{A}$.*

Depending on the size of the set of variables, the set of formulas may be quite large. If V is countable, however, $\mathcal{F}$ is countable as well, so in this case the question may be easier to answer; this will be discussed briefly after giving the proof of the Compactness Theorem. We want to establish the general case.

Before we state and prove the result, we need a lemma which permits us to extend the range of our knowledge of satisfiability just by one formula.

Lemma 1.5.7 *Let $\mathcal{A} \subseteq \mathcal{F}$ be satisfiable and $\varphi \notin \mathcal{A}$ be a formula. Then one of $\mathcal{A} \cup \{\varphi\}$ and $\mathcal{A} \cup \{\neg\varphi\}$ is satisfiable.*

Proof If $\mathcal{A} \cup \{\varphi\}$ is not satisfiable, but $\mathcal{A}$ is, let v be the valuation for which $v \models \mathcal{A}$ holds. Because $v(\varphi) = 0$, we conclude $v(\neg\varphi) = 1$, so that $v \models \mathcal{A} \cup \{\neg\varphi\}$. $\dashv$

We establish now the *Compactness Theorem* for propositional logic. It permits reducing the question of satisfiability of a set $\mathcal{A}$ of formulas to finite subsets of $\mathcal{A}$.

Theorem 1.5.8 *Let $\mathcal{A} \subseteq \mathcal{F}$ be a set of propositional formulas. Then $\mathcal{A}$ is satisfiable iff each finite subset of $\mathcal{A}$ is satisfiable.*

Outline for the proof

Proof 0. We will focus on satisfiability of $\mathcal{A}$ provided each finite subset of $\mathcal{A}$ is satisfiable, because the other half of the assertion is trivial . The idea is to apply $(\mathbb{ZL})$, so that we have to construct an ordered set which satisfies the chain condition. This set will consist of pairs $\langle \mathcal{B}, v \rangle$ with $\mathcal{B} \subseteq \mathcal{A}$ and $v \models \mathcal{B}$. We know from the assumption that we have plenty of these pairs. The order is straightforward, and we will establish the chain condition easily. The maximal element will have $\mathcal{A}$ as its first component.

1. Let

$$\mathcal{C} := \{\langle \mathcal{B}, v \rangle \mid \mathcal{B} \subseteq \mathcal{A}, v \models \mathcal{B}\},$$

and define $\langle \mathcal{B}_1, v_1 \rangle \leq \langle \mathcal{B}_2, v_2 \rangle$ iff $\mathcal{B}_1 \subseteq \mathcal{B}_2$ and $v_1(\varphi) = v_2(\varphi)$ for all $\varphi \in \mathcal{B}_1$, so that $\langle \mathcal{B}_1, v_1 \rangle \leq \langle \mathcal{B}_2, v_2 \rangle$ holds iff $\mathcal{B}_1$ is contained in $\mathcal{B}_2$ and if the valuations coincide on the smaller set. This is a partial order. If $\mathcal{D} \subseteq \mathcal{C}$ is a chain, then put $\mathcal{B} := \bigcup \mathcal{D}$, and define $v(\varphi) := v'(\varphi)$, if $\varphi \in \mathcal{B}'$ with $\langle \mathcal{B}', v' \rangle \in \mathcal{D}$. Since $\mathcal{D}$ is a chain, v is well defined. Moreover, $v \models \mathcal{B}$: Let $\varphi \in \mathcal{B}$, then $\varphi \in \mathcal{B}'$ for some $\langle \mathcal{B}', v' \rangle \in \mathcal{D}$, since $v' \models \mathcal{B}'$, we have $v(\varphi) = v'(\varphi) = 1$. Hence by Zorn's Lemma there exists a maximal element $\langle \mathcal{M}, w \rangle$, in particular $w \models \mathcal{M}$.

We claim that $\mathcal{M} = \mathcal{A}$. Suppose this is not the case; then there exists $\varphi \in \mathcal{A}$ with $\varphi \notin \mathcal{M}$. But either $\mathcal{M} \cup \{\varphi\}$ or $\mathcal{M} \cup \{\neg\varphi\}$ is satisfiable by Lemma 1.5.7; hence $\langle \mathcal{M}, w \rangle$ is not maximal. This is a contradiction.

But this means that $\mathcal{M} = \mathcal{A}$; hence $\mathcal{A}$ is satisfiable. $\dashv$

Suppose that V is countable, then we know that $\mathcal{F}$ is countable as well. Then another proof for Theorem 1.5.8 can be given; this will be sketched now. Enumerate $\mathcal{F}$ as $\{\varphi_1, \varphi_2, \ldots\}$. Call—just temporarily—$\mathcal{A} \subseteq \mathcal{F}$ *finitely satisfiable* iff each finite subset of $\mathcal{A}$ is satisfiable. Let $\mathcal{A}$ be such a finitely satisfiable set. We construct a sequence $\mathcal{M}_0, \mathcal{M}_1, \ldots$ of finitely satisfiable sets, starting from $\mathcal{M}_0 := \mathcal{A}$. If $\mathcal{M}_n$ is defined, put

$$\mathcal{M}_{n+1} := \begin{cases} \mathcal{M}_n \cup \{\varphi_{n+1}\} & \text{if } \mathcal{M}_n \cup \{\varphi_{n+1}\} \text{ is finitely satisfiable,} \\ \mathcal{M}_n \cup \{\neg\varphi_{n+1}\} & \text{otherwise.} \end{cases}$$

This will give a finitely satisfiable set $\mathcal{M}^* := \bigcup_{n \geq 0} \mathcal{M}_n$. Now define $v^*(\varphi) := 1$ iff $\varphi \in \mathcal{M}^*$. We claim that $v^* \models \varphi$ iff $\varphi \in \mathcal{M}^*$. This is proved by a straightforward induction on φ. Because $\mathcal{A} \subseteq \mathcal{M}^*$, we know that $v^* \models \mathcal{A}$. This approach could be modified for the general case well ordering $\mathcal{F}$.

The approach used for the general proof can be extended from propositional logic to first-order logic by introducing suitable constants (they are called *Henkin's constants*). We refer the reader to [Bar77, Chap. 1], since we are concentrating presently on applications of Zorn's Lemma; see, however, the discussion in Sect. 3.6.1, where additional references can be found.

1.5.2 Extending Orders

We will establish a generalization to the well-known fact that each finite graph $\mathcal{G}$ can be embedded into a linear order, provided the graph does not have any cycles. This is known as a *topological sort* of the graph [Knu73, Algorithm T, p. 262] or [CLR92, Sect. 23.4]. One notes first that $\mathcal{G}$ must have a node k which does not have any predecessor (hence there is no node ℓ which is connected to k through an edge $\ell \to k$). If such a node k would not exist, one could construct for each node a cycle on which it lies. The algorithm proceeds recursively.

If the graph contains at most one node, it returns either the empty list or the list containing the node. The general case constructs a list having k as its head and the list for $\mathcal{G} \setminus k$ as its tail; here $\mathcal{G} \setminus k$ is the graph with node k and all edges emanating from k are removed.

Finiteness plays a special role in the argument above, because it makes sure that we have a well order among the nodes, which in turn is needed for making sure that the algorithm terminates. Let us turn to the general case. Given a partial order $\leq$ on a set S, we show that $\leq$ can be extended to a strict order $\leq_s$ (hence $a \leq b$ implies $a \leq_s b$ for all $a, b \in S$).

This will be shown through Zorn's Lemma. Put

$$\mathcal{G} := \{R \mid R \text{ is a partial order on } S \text{ with } \leq \ \subseteq R\}$$

and order $\mathcal{G}$ by inclusion. Let $\mathcal{C} \subseteq \mathcal{G}$ be a chain, and then we claim that $R_0 := \bigcup \mathcal{C}$ is a partial order. It is obvious that R_0 is reflexive; if aR_0b and bR_0a, then there exist relations $R_1, R_2 \in \mathcal{C}$ with aR_1b and bR_2a. Since $\mathcal{C}$ is a chain, we know that $R_1 \subseteq R_2$ or $R_2 \subseteq R_1$ holds. Assume that the former holds, and then aR_2b follows, so that we may conclude $a = b$. Hence R_0 is antisymmetric. Transitivity is proved along the same lines, using that $\mathcal{C}$ is a chain. By Zorn's Lemma, $\mathcal{G}$ has a maximal element M; since $M \in \mathcal{G}$, M is a partial order which contains the given partial order $\leq$.

We have to show that M is linear. Assume that it is not, so that there exists $a, b \in S$ such that both aMb and bMa are false. Put

$$M' := M \cup \{\langle x, y\rangle \mid xMa \text{ and } bMy\}.$$

Plan

Then M' contains M properly. If we can show that M' is a partial order, we have shown that M is not maximal, which is a contradiction. Let us see:

- M' is reflexive: Since $M \subseteq M'$ and M is reflexive, xMx holds for all $x \in S$.
- M' is transitive: Let $xM'y$ and $yM'z$; then these cases are possible:
  1. xMy and yMz, hence xMz, thus $xM'z$.
  2. xMy and yMa and bMz, thus xMa and bMz, so that $xM'z$.

3. xMa and bMy and yMz, hence $xM'z$.

4. xMa and bMy and yMa and bMz, but then bMa contrary to our assumption. Hence this case cannot occur.

Thus we may conclude that M' is transitive.

- M' is antisymmetric. Assume that $xM'y$ and $yM'x$, and look at the cases above with $z = x$. Case 2 would imply xMa and bMx, so it is not possible; case 3 is excluded for the same reason, so only case 1 is left, which implies $x = y$.

Thus the assumption that there exists $a, b \in S$ such that both aMb and bMa are false leads to the conclusion that M is not maximal in $\mathcal{G}$, which is a contradiction.

Then $a <_s b$ iff $\langle a, b\rangle \in M$ defines the desired total order, and by construction it extends the given order.

Hence we have shown

Proposition 1.5.9 *Each partial order on a set can be extended to a total order.* $\dashv$

It is clear that this applies to acyclic graphs, so that we have here a very general version of topological sorting.

1.5.3 Bases in Vector Spaces

Fix a vector space V over field K. A set $B \subseteq V$ is called *linearly independent* iff $\sum_{b \in B_0} a_b \cdot b = 0$ implies $a_b = 0$ for all $b \in B_0$, whenever B_0 is a finite nonempty subset of B. Hence, e.g., a single vector v with $v \neq 0$ is linear independent. Linear independence

Example 1.5.10 The reals $\mathbb{R}$ form a vector space over the rationals $\mathbb{Q}$. Then $\sqrt{2}$ and $\sqrt{3}$ are linearly independent. In fact, assume that $q_1\sqrt{2} + q_2\sqrt{3} = 0$ with rational numbers $q_1 = r_1/s_1$ and $q_2 = r_2/s_2$. Then we can find integers t_1, t_2 such that $t_1\sqrt{2} = t_2\sqrt{3}$ so that t_1 and t_2 have no common divisors. But $2t_1^2 = 3t_2^2$ implies that 2 and 3 are both common divisors to t_1 and to t_2. ✌

Base

The linear independent set B is called a *base* for V iff B is linear independent and if each element $v \in V$ can be represented as

$$v = \sum_{i=1}^{n} a_i \cdot b_i$$

for some $a_1, \ldots, a_n \in K$ and $b_1, \ldots, b_n \in B$. This representation is unique.

Proposition 1.5.11 *Each vector space V has a base.*

Plan

Proof 0. We first find a maximal independent set through $(\mathbb{ZL})$ by considering the family of all independent sets, and then we show that this set is a base.

1. Let

$$\mathcal{V} := \{B \subseteq V \mid B \text{ is linear independent}\}.$$

Then $\mathcal{V}$ contains all singletons with non-null vectors; hence it is not empty. Order $\mathcal{V}$ by inclusion, and let $\mathcal{B}$ be a chain in $\mathcal{V}$. Then $B_0 := \bigcup \mathcal{B}$ is independent. In fact, if $\sum_{i=1}^{n} a_i \cdot b_i = 0$, let $b_i \in B_i \in \mathcal{B}$ for $1 \leq i \leq n$. Since $\mathcal{C}$ is linearly ordered, we find some k such that $b_i \in B_k$, and since B_k is independent, we may conclude $b_1 = \ldots = b_n = 0$. By Zorn's Lemma there exists a maximal independent set $B^* \in \mathcal{V}$.

2. If B^* is not a basis, then we find a vector x which cannot be represented as a finite linear combination of elements of B^*. Clearly $x \notin B^*$. But then $B^* \cup \{x\}$ is linear independent, for it could otherwise be represented by elements from B^*. This contradicts the maximality of B^*. $\dashv$

One notes that part 1 of the proof could as well argue with the Maximality Principle, because a set is linear independent iff each finite subset is linear independent. The set $\mathcal{V}$ constructed in the proof is of finite character and hence contains by $(\mathbb{MP})$ a maximal element. Then one argues exactly as in part 2 of the proof. This shows that $(\mathbb{ZL})$ and $(\mathbb{MP})$ are close relatives.

These proofs are not constructive, since they do not tell us how to construct a base for a given vector space, not even in the finite dimensional case.

1.5.4 Extending Linear Functionals

Sometimes one is given a linear map from a subspace of a vector space to the reals, and one wants to extend this map to a linear map on the whole space. Usually there is the constraint that both the given map and the extension should be dominated by a sublinear map.

Let V be a vector space over the reals. A map $f : V \to \mathbb{R}$ is said to be a *linear functional* (or a *linear map*) on V iff $f(\alpha \cdot x + \beta y) = \alpha \cdot f(x) + \beta \cdot f(y)$ holds for all $x, y \in V$ and $\alpha, \beta \in \mathbb{R}$. Thus a linear functional is compatible with the vector space structure of V. Call $p : V \to \mathbb{R}$ *sublinear* iff $p(x + y) \leq p(x) + p(y)$, and $p(\alpha \cdot x) = \alpha \cdot p(x)$ for all $x, y \in V$ and $\alpha \geq 0$.

Linear map

Sublinearity

We have a look at the situation in the finite dimensional case first. This will permit us to isolate the central argument easily, which then will be applied to the general situation.

Proposition 1.5.12 *Let V be a finite dimensional real vector space with a sublinear functional $p : V \to \mathbb{R}$. Given a subspace V_0 and a linear map $f_0 : V_0 \to \mathbb{R}$ such that $f_0(x) \leq p(x)$ for all $x \in V_0$, then there exists a linear functional $f : V \to \mathbb{R}$ which extends f_0 such that $f(x) \leq p(x)$ for all $x \in V$.*

Proof 1. It is enough to show that f_0 can be extended to a linear functional dominated by p to the vector space generated by $V_0 \cup \{z\}$ with $z \notin V_0$. In fact, we can then repeat this procedure a finite number of times, in each step adding a new basis vector not contained in the previous subspace. Since V is finite dimensional, this will eventually give us V as the domain for the linear functional.

Line of attack

2. Let $z \notin V_0$, and then $\{v + \alpha \cdot z \mid v \in V_0, \alpha \in \mathbb{R}\}$ is the vector space generated by V_0 and z, because this is clearly a vector space containing $V_0 \cup \{z\}$, and each vector space containing $V_0 \cup \{z\}$ must also contain linear combinations of the form $v + \alpha \cdot z$ with $v \in V_0$ and $\alpha \in \mathbb{R}$. The representation of an element in this vector space is unique: Assume $v + \alpha \cdot z = v' + \alpha' \cdot z$, then $v - v' = (\alpha - \alpha') \cdot z$, and because $z \notin V_0$, this implies $v - v' = 0$, and hence also $\alpha = \alpha'$.

3. Now set

$$f(v + \alpha \cdot z) := f_0(v) + \alpha \cdot c$$

with a value c which will have to be determined. Consider $v, v' \in V_0$; then we have

$$f_0(v) - f_0(v') = f_0(v - v') \leq p(v - v') \leq p(v + z) + p(-z - v')$$

for an arbitrary $v_1 \in V$. Thus we obtain $-p(z - v') - f_0(v') \leq p(v + z) - f_0(v)$. Note that the left-hand side of this inequality is independent of v and that the right-hand side is independent of v', which means that we can find c with

i $c \leq p(v + z) - f_0(v)$ for all $v \in V_0$,

ii $c \geq -p(-z - v') - f_0(v')$ for all $v \in V_0$.

Now let us see what happens. Fix α. If $\alpha = 0$, we have $f(v + 0 \cdot z) = f_0(v) \leq p(v + 0 \cdot z)$. If $\alpha > 0$, we have

$$\begin{aligned} f(v + \alpha \cdot z) &= \alpha \cdot f(v/\alpha + z) = \alpha \cdot (f_0)(v/\alpha) + c) \\ &\leq \alpha \cdot (f_0(v) + p(v/\alpha + z) - f_0(v/\alpha) \\ &= p(v + \alpha \cdot z) \end{aligned}$$

by i and sublinearity. If, however, $\alpha < 0$, we use the inequality ii and sublinearity of p; note that the coefficient $-z/\alpha$ of z is positive in this case.

Summarizing, we have $f(v + \alpha \cdot z) \leq p(v + \alpha \cdot z)$ for all $v \in V_0$ and $\alpha \in \mathbb{R}$. $\dashv$

When having a closer look at the proof, we see that the assumption on working in a finite dimensional vector space is only important for making sure that the extension process terminates in a finite number of steps. The core of this proof, however, consists in the observation that we can extend a linear functional from a vector space V_0 to a vector space $\{v + \alpha \cdot z \mid v \in V_0, \alpha \in \mathbb{R}\}$ with $z \notin V_0$ without losing domination by the sublinear functional p. Let us record this important intermediate result.

Corollary 1.5.13 *Let V_0 be a vector space, $V_0 \subseteq V$, $p : V \to \mathbb{R}$ be a sublinear functional, and $z \notin V_0$. Then each linear functional $f_0 : V_0 \to \mathbb{R}$ which is dominated by p can be extended to a linear functional f on the vector space generated by V_0 and z such that f is also dominated by p.* $\dashv$

Now we are in a position to formulate and prove the *Hahn–Banach Theorem*. We will use Zorn's Lemma for the proof by setting up a partial order such that each chain has an upper bound. The elements of this ordered set will be pairs $\langle V', f' \rangle$ such that V' is a subspace of V with $V_0 \subseteq V$, and f' will be a linear map extending f_0 and being dominated by p, the order being straightforward, induced by the extension condition. We may conclude then that there exists a maximal element. By the "dimension free" version of the extension just stated, we will then show that the assumption that we did not capture the whole vector space through our maximal element will yield a contradiction. Outline

Theorem 1.5.14 *Let V be a real vector space with a sublinear functional $p : V \to \mathbb{R}$. Given a subspace V_0 and a linear map $f_0 : V_0 \to \mathbb{R}$ such that $f_0(x) \leq p(x)$ for all $x \in V_0$, then there exists a linear functional $f : V \to \mathbb{R}$ which extends f_0 such that $f(x) \leq p(x)$ for all $x \in V$.*

Proof 1. Define $\langle V', f' \rangle \in \mathcal{W}$ iff V' is a vector space with $V_0 \subseteq V' \subseteq V$, and $f' : V' \to \mathbb{R}$ extends f_0 and is dominated by p. Define $\langle V', f' \rangle \leq \langle V'', f'' \rangle$ iff V' is a subspace of V'' and f'' is an extension to f' for $\langle V', f' \rangle, \langle V'', f'' \rangle \in \mathcal{W}$. Then $\leq$ is a partial order on $\mathcal{W}$. Let $\big(\langle V_i, f_i \rangle\big)_{i \in I}$ be a chain in $\mathcal{W}$, and then $V' := \bigcup_{i \in I} V_i$ is a subspace of V. In fact, let $x, x' \in V'$, and then $x \in V_i$ and $x' \in V_{i'}$. Then either $V_i \subseteq V_{i'}$ or $V_{i'} \subseteq V_i$. Assume the former, hence $x, x' \in V_{i'}$; thus $\alpha \cdot x + \beta \cdot x' \in V_{i'} \subseteq V'$ for all $\alpha, \beta \in \mathbb{R}$. Put $f'(x) := f_i(x)$; if $x \in V_i$ for some $i \in I$, then $f' : V' \to \mathbb{R}$ is well defined, linear, and dominated by p; moreover, f' extends every f_i, hence, by transitivity, f_0. This implies $\langle V', f' \rangle \in \mathcal{W}$, and this is obviously an upper bound for the chain.

2. Hence each chain has an upper bound in $\mathcal{W}$, so that Zorn's Lemma implies the existence of a maximal element $\langle V^+, f^+ \rangle \in \mathcal{W}$. Assume that $V^+ \neq V$; then there exists $z \in V$ with $z \notin V^+$. Then the vector space V^* generated by $V^+ \cup \{z\}$ contains V^+ properly, and f^+ has a linear extension f^* to V^* which is dominated by p by Corollary 1.5.13. But this means $\langle V^+, f^+ \rangle$ is strictly smaller than $\langle V^*, f^* \rangle \in \mathcal{W}$, a contradiction. Hence $V^+ = V$, and f^+ is the desired extension. $\dashv$

The Hahn–Banach Theorem is sometimes considered as one of the cornerstones in functional analysis, because it permits to construct linear functionals with given properties. A first idea why this might be important is hinted at in Exercise 4.32.

1.5.5 Maximal Filters

Fix a set S. The power set $\mathcal{P}(S)$ is ordered by inclusion $\subseteq$, exhibiting some interesting properties. We single out subsets of $\mathcal{P}(S)$ which are called filters. These filters will be discussed in subsequent sections, and then the aspect that a filter lives in an ordered environment becomes dominant. But here is the definition of a filter of subsets.

Definition 1.5.15 *A nonempty subset $\mathcal{F} \subseteq \mathcal{P}(S)$ is called a* filter *iff*

1. $\emptyset \notin \mathcal{F}$,
2. *if* $F_1, F_2 \in \mathcal{F}$, *then* $F_1 \cap F_2 \in \mathcal{F}$,
3. *if* $F \in \mathcal{F}$ *and* $F \subseteq F'$, *then* $F' \in \mathcal{F}$.

Thus a filter is closed under finite intersections and closed with respect to super sets, and it must not contain the empty set.

Principal ultrafilter

Example 1.5.16 Given $s \in S$, the set $\mathcal{F}_s := \{A \subseteq S \mid s \in A\}$ is a filter, which is called the *principal ultrafilter* associated with x. Let M be an infinite set. Then $\mathcal{F} := \{A \subseteq M \mid M \setminus A \text{ is finite}\}$ is a filter, the filter of cofinite sets. ✌

The filter $\mathcal{F}_s$ from Example 1.5.16 is special because it is maximal; we cannot find a filter $\mathcal{G}$ which properly contains $\mathcal{F}_s$. Let us try: Take $G \in \mathcal{G}$ with $G \notin \mathcal{F}_s$, then $s \notin G$, and hence $s \in S \setminus G$, so that both $G \in \mathcal{G}$ and $S \setminus G \in \mathcal{G}$, the latter one via $\mathcal{F}_s$. This implies $\emptyset \in \mathcal{G}$, since a filter is closed under finite intersections. We have arrived at a contradiction, giving rise to the definition of a maximal filter (Definition 1.5.20) in a moment.

Before stating it, we will introduce filter bases. Sometimes we are not presented with a filter proper, but rather with a family of sets which generates one.

Definition 1.5.17 *A subset $\mathcal{B} \subseteq \mathcal{P}(S)$ is called a* filter base *iff no intersection of a finite collection of elements of $\mathcal{B}$ is empty and thus iff* $\emptyset \notin \{B_1 \cap \ldots \cap B_n \mid B_1, \ldots, B_n \in \mathcal{B}\}$.

Example 1.5.18 . Fix $x \in \mathbb{R}$; then the set $\mathcal{B} := \big\{\,]a, b[\, \mid a < x < b\big\}$ of all open intervals containing x is a filter base. Let $(a_n)_{n \in \mathbb{N}}$ be a sequence in $\mathbb{R}$; then the set $\mathcal{E} := \big\{\{a_k \mid k \geq n\} \mid n \in \mathbb{N}\big\}$ of infinite tails of the sequence is a filter base as well. ✌

Clearly, if $\mathcal{B}$ is to be contained in a filter $\mathcal{F}$, then it must not have the empty sets among its finite intersections, because all these finite intersections are elements of $\mathcal{F}$. It is easy to characterize the filter generated by a base.

Lemma 1.5.19 *Let $\mathcal{B} \subseteq \mathcal{P}(S)$ be a filter base; then*

$$\mathcal{F} := \{B \subseteq S \mid B \supseteq B_1 \cap \ldots \cap B_n \text{ for some } B_1, \ldots, B_n \in \mathcal{B}\}$$

is the smallest filter containing $\mathcal{B}$.

Proof It is clear that $\mathcal{F}$ is a filter, because it cannot contain the empty set, it is closed under finite intersections, and it is closed under super sets. Let $\mathcal{G}$ be a filter containing $\mathcal{B}$, and let $B \supseteq B_1 \cap \ldots \cap B_n$ for some $B_1, \ldots, B_n \in \mathcal{B} \subseteq \mathcal{G}$; hence $B \in \mathcal{G}$. Thus $\mathcal{F} \subseteq \mathcal{G}$, so that $\mathcal{F}$ is in fact the smallest filter containing $\mathcal{B}$. $\dashv$

Let us return to the properties of the filter which is defined in Example 1.5.16.

Definition 1.5.20 *A filter is called* maximal *iff it is not properly contained in another filter. Maximal filters are also called* ultrafilters.

This is an easy characterization of maximal filters.

Lemma 1.5.21 *These conditions are equivalent for a filter $\mathcal{F}$:*

1. *$\mathcal{F}$ is maximal.*
2. *For each subset $A \subseteq S$, either $A \in \mathcal{F}$ or $S \setminus A \in \mathcal{F}$.*

Proof 1 $\Rightarrow$ 2: Assume there is a set $A \subseteq S$ such that both $A \notin \mathcal{F}$ and $S \setminus A \notin \mathcal{F}$ hold. Then

$$\mathcal{G}_0 := \{F \cap A \mid F \in \mathcal{F}\}$$

is a filter base, because $F \cap A = \emptyset$ for some $F \in \mathcal{F}$ would imply $F \subseteq S \setminus A$; thus $S \setminus A \in \mathcal{F}$. Because $F \cap A \notin \mathcal{F}$ for all $F \in \mathcal{F}$, we conclude that the filter $\mathcal{G}$ generated by $\mathcal{G}_0$ contains $\mathcal{F}$ properly. Thus $\mathcal{F}$ is not maximal.

2 $\Rightarrow$ 1: A filter $\mathcal{G}$ which contains $\mathcal{F}$ properly will contain a set $A \notin \mathcal{F}$. By assumption, $S \setminus A \in \mathcal{F} \subseteq \mathcal{G}$, so that $\emptyset \in \mathcal{G}$. Thus $\mathcal{G}$ is not a filter. $\dashv$

Example 1.5.22 The filter $\mathcal{F}$ of cofinite sets from Example 1.5.15 for an infinite set M is not an ultrafilter. In fact, decompose $M = M_0 \cup M_1$ into disjoint sets M_0 and M_1 which are both infinite. Then neither M_0 nor its complement is contained in $\mathcal{F}$. ✌

The existence of ultrafilters is trivial by Example 1.5.16, but we do not know whether each filter is actually contained in an ultrafilter. The answer is in the affirmative.

Theorem 1.5.23 *Each filter can be extended to a maximal filter.*

Proof Let $\mathcal{F}$ be a filter on S, and define

$$\mathcal{V} := \{\mathcal{G} \mid \mathcal{G} \text{ is a filter with } \mathcal{F} \subseteq \mathcal{G}\}.$$

Order $\mathcal{V}$ by inclusion. Then each chain $\mathcal{C}$ in $\mathcal{V}$ has an upper bound in $\mathcal{V}$. In fact, let $\mathcal{H} := \bigcup \mathcal{C}$. If $A \in \mathcal{H}$ and $A \subseteq B$, there exists a filter $\mathcal{G} \in \mathcal{C}$ with $A \in \mathcal{G}$; hence $B \in \mathcal{G}$, so that $B \in \mathcal{H}$. If $A, B \in \mathcal{H}$, we find $\mathcal{G}_A, \mathcal{G}_B \in \mathcal{H}$ with either $\mathcal{G}_A \subseteq \mathcal{G}_B$ or $\mathcal{G}_B \subseteq \mathcal{G}_A$, because $\mathcal{C}$ is linearly ordered. Assume the former, hence $A, B \in \mathcal{C}_B$, hence $A \cap B \in \mathcal{G}_B \subseteq \mathcal{H}$. So $\mathcal{H}$ is a filter in $\mathcal{V}$.

Thus there exists in $\mathcal{V}$ a maximal element $\mathcal{F}^*$ which is a maximal filter (just repeat the argument in the proof of 2 $\Rightarrow$ 1 for Lemma 1.5.21). $\mathcal{F}^*$ contains $\mathcal{F}$. ⊣

Corollary 1.5.24 *Let $\emptyset \neq A \subseteq X$ be a nonempty subset of a set X. Then there exists an ultrafilter containing A.*

Proof Using Theorem 1.5.23, extend the filter $\{B \subseteq X \mid A \subseteq B\}$ to an ultrafilter. ⊣

1.5.6 Ideals and Filters

We will now translate some of the arguments above from the power set of a set to a partially ordered set which has at least part of the algebraic properties of the power set.

Recall that a *lattice* $(L, \leq)$ is a set L with an order relation $\leq$ such that each nonempty finite subset has a lower bound and an upper bound. Put $a \wedge b := \inf\{a, b\}$ (this is the *meet of a and b*) and $a \vee b := \sup\{a, b\}$ (this is the *join of a and b*). In a similar way, we put for $\bigvee S := \sup S$

Lattice, join, meet

and $\bigwedge S := \inf S$ for $S \subseteq L$, provided $\sup S$ resp. $\inf S$ exists in L.

We note these properties ($a, b \in L$):

Impotency $a \wedge a = a \vee a = a$.

Commutativity $a \wedge b = b \wedge a$ and $a \vee b = b \vee a$.

Absorption $a \wedge (a \vee b) = a$ and $a \vee (a \wedge b) = a$. In fact, $a \leq a \vee b$, and thus $a = a \wedge a \leq a \wedge (a \vee b)$; on the other hand $a \wedge (a \vee b) \leq a$. The second equality is proved similarly.

For simplicity we assume that the lattice is *bounded*, i.e., that it has a smallest element $\perp$ and a largest element $\top$, so that we can put $\perp := \sup \emptyset$ and $\top := \inf \emptyset$, resp.

That is, the generalization of the properties of a power set is clear.

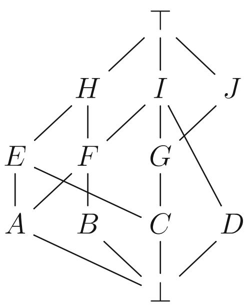
Example 1.5.25 The power set $\mathcal{P}(S)$ of a set S is a lattice, where $A \leq B$ iff $A \subseteq B$, so that

$$\begin{aligned} A \cap B &= \inf\{A, B\}, \\ A \cup B &= \sup\{A, B\}. \end{aligned}$$

✌

But there are lattices which do not derive from a power set, as the following example shows:

Example 1.5.26 Look at this example



Then $\{B, C\}$ has these upper bounds, $\{\top, H, I\}$, and thus has no smallest upper bound, so that—probably contrary to the first view—$B \vee C$ does not exist. Trying to determine $A \vee B$, we see that the set of upper bounds to $\{A, B\}$ is just $\{\top, F, H, I\}$: hence $A \vee B = F$.

✌

This is another example of a lattice. It indicates that we have to carefully look at the context, when discussing joins and meets.

Example 1.5.27 Consider the set $\mathcal{J}$ of all open intervals $]a, b[$ with $a, b \in \mathbb{R}$, and take the order inherited from $\mathcal{P}(\mathbb{R})$; then $\mathcal{J}$ is closed under taking the infimum of two elements (since the intersection of two open intervals is again an open interval), but $\mathcal{J}$ is not closed under taking the supremum of two elements in $\mathcal{P}(\mathbb{R})$, since the union of two open intervals is not necessarily an open interval. Nevertheless, $\mathcal{J}$ is a lattice in its own right, because we have

$$]a_1, b_1[\ \vee\ ]a_2, b_2[\ =\ ]\min\{a_1, a_2\}, \max\{b_1, b_2\}[$$

in $\mathcal{J}$. Hence we have to make sure that we look for the supremum in the proper set. ✌

The next example also asks for a cautionary approach.

Example 1.5.28 Similarly, consider the set $\mathcal{R}$ of all closed rectangles in the plane $\mathbb{R} \times \mathbb{R}$, again with the order inherited from $\mathcal{P}(\mathbb{R} \times \mathbb{R})$. The intersection $R_1 \cap R_2$ of two closed rectangles $R_1, R_2 \in \mathcal{R}$ is an element of $\mathcal{R}$ and is indeed the infimum of R_1 and R_2. But what do we take as the supremum in $\mathcal{R}$ if it exists at all? From the definition of the supremum, we have

$$R_1 \vee R_2 = \bigcap\{R \in \mathcal{R} \mid R_1 \subseteq R \text{ and } R_2 \subseteq R\},$$

in plain words, the smallest closed rectangle which encloses both R_1 and R_2. Hence, e.g.,

$$[0, 1] \times [0, 1]\ \vee\ [5, 6] \times [8, 9] = [0, 6] \times [0, 9].$$

This renders $\mathcal{R}$ a lattice indeed. ✌

A lattice is called *distributive* iff

$$\begin{aligned} a \wedge (b \vee c) &= (a \wedge b) \vee (a \wedge c), \\ a \vee (b \wedge c) &= (a \vee b) \wedge (a \vee c) \end{aligned}$$

holds (both equations are actually equivalent; see Exercise 1.23).

Example 1.5.29 The power set lattice $\mathcal{P}(S)$ is a distributive lattice, because unions and intersections are distributive.

But BEWARE! Distributivity is not necessarily inherited. Consider the lattice $\mathcal{J}$ of closed intervals of the real line, as in Example 1.5.27; then

$$\begin{aligned} I_1 \wedge I_2 &= I_1 \cap I_2, \\ I_1 \vee I_2 &= [\min I_1 \cup I_2, \max I_1 \cup I_2], \end{aligned}$$

as above. Put $A := [-3, -2], B := [-1, 1], C := [2, 3]$; then

$$\begin{aligned} (A \wedge B) \vee (B \wedge C) &= \emptyset, \\ B \wedge (A \vee C) &= [-1, 1]. \end{aligned}$$

Thus $\mathcal{J}$ is not distributive, although the order has been inherited from the power set. ✌

Example 1.5.30 Let P be a set with a partial order $\leq$. A set $D \subseteq P$ is called a *down set* iff $t \in D$ and $s \leq t$ imply $s \in D$. Hence a down set is downward closed in the sense that all elements below an element of the set belong to the set as well. A generic example for a down set is $\{s \in P \mid s \leq t\}$ with $t \in P$. Down sets of this shape are called *principal down sets*. The intersection and the union of two down sets are down sets again. For example, let D_1 and D_2 be down sets, let $t \in D_1 \cup D_2$, and assume $s \leq t$. Because $t \in D_1$ or $t \in D_2$, we may conclude that $s \in D_1$ or $s \in D_2$; hence $s \in D_1 \cup D_2$. Let $\mathcal{D}(P)$ be the set of all down sets of P; then $\mathcal{D}(P)$ is a distributive lattice; this is so because the infimum and the supremum of two elements in $\mathcal{D}(P)$ are the same as in $\mathcal{P}(P)$.

Define

$$\Psi : \begin{cases} P & \to \mathcal{D}(P) \\ t & \mapsto \{s \in P \mid s \leq t\}. \end{cases}$$

Then $t_1 \leq t_2$ implies $\Psi(t_1) \subseteq \Psi(t_2)$; hence the order structure carries over from P to $\mathcal{D}(P)$. Moreover $\Psi(t_1) = \Psi(t_2)$ implies $t_1 = t_2$, so that Ψ is injective. Hence we have embedded the partially ordered set P into a distributive lattice. ✌

Filters and ideals are important structures in a lattice.

Definition 1.5.31 *Let L be a lattice.*

$J \subseteq L$ is called an ideal *iff*

- *$\emptyset \neq J \neq L$.*
- *If $a, b \in J$, then $a \vee b \in J$.*
- *If $a \in J$ and $b \in L$ with $b \leq a$, then $b \in J$.*

The ideal J is called prime *iff $a \wedge b \in J$ implies $a \in J$ or $b \in J$, and it is called* maximal *iff it is not properly contained in another ideal.*

$F \subseteq L$ is called a filter *iff*

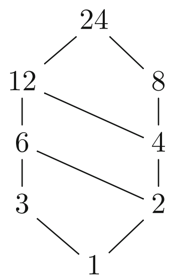
- *$\emptyset \neq F \neq L$.*
- *If $a, b \in F$, then $a \wedge b \in F$.*
- *If $a \in F$ and $b \in L$ with $b \geq a$, then $b \in F$.*

The filter F is called prime *iff $a \vee b \in F$ implies $a \in F$ or $b \in F$, and it is called* maximal *iff it is not properly contained in another filter.*

Maximal filters are also called *ultrafilters*. Recall the definition of an ultrafilter of sets in Definition 1.5.20; we have defined already ultrafilters for the special case that the underlying lattice is the power set of a given set. The notion of a filter base fortunately carries over directly from Definition 1.5.17, so that we may use Lemma 1.5.19 in the present context as well. We will talk in this section in a more general context, but first some simple examples.

Example 1.5.32 $I := \{F \subseteq \mathbb{N} \mid F$ is finite $\}$ is an ideal in $\mathcal{P}(\mathbb{N})$ with set inclusion as the partial order. This is so since the intersection of two finite sets is finite again and because subsets of finite sets are finite again. Also $\emptyset \neq I \neq \mathcal{P}(\mathbb{N})$. This ideal is not prime. ✌

Example 1.5.33 Consider all divisors of 24.



$\{1, 2, 3, 6\}$ is an ideal, and $\{1, 2, 3, 4, 6\}$ is not. ✌

Example 1.5.34 Let $S \neq \emptyset$ be a set, $a \in S$. Then $\mathcal{P}(S \setminus \{a\})$ is a prime ideal in $\mathcal{P}(S)$ (with set inclusion as the partial order). In fact, $\emptyset \neq \mathcal{P}(S \setminus \{a\}) \neq \mathcal{P}(S)$, and if $a \notin A$ and $a \notin B$, then $a \notin A \cup B$. On the other hand, if $a \notin A \cap B$, then $a \notin A$ or $a \notin B$. ✌

Lemma 1.5.35 *Let L be a lattice and $\emptyset, \neq F \neq L$ be a proper nonempty subset of L.*

- *These conditions are equivalent*
  1. *F is a filter.*
  2. *$\top \in F$ and $(a \wedge b \in F \Leftrightarrow a \in F$ and $b \in F)$.*
- *If filter F is maximal and L is distributive, then F is a prime filter*

Proof 1. The implication 1 $\Rightarrow$ 2 in the first part is trivial, for 2 $\Rightarrow$ 1 one notes that $a \leq b$ is equivalent to $a \wedge b = a$.

2. In order to show that the maximal filter F is prime, we show that $a \vee b \in F$ implies $a \in F$ or $b \in F$. Assume that $a \vee b \in F$ with $a \notin F$. Consider $B := \{f \wedge b \mid f \in F\}$; then $\bot \notin B$. In fact, assume that $f \wedge b = \bot$ for some $f \in F$; then we could write $a = (f \wedge b) \vee a = (f \vee a) \wedge (b \vee a)$ by distributivity. Since $f \in F$ and F are a filter, $f \vee a \in F$ follows, and since $b \vee a \in F$, we obtain $a \in F$, contradicting the assumption. Thus B is a filter base, and because F is maximal, we may conclude that $B \subseteq F$, which in turn implies $b \in F$. $\dashv$

Hence maximal filters are prime in a distributive lattice. If the lattice is not distributive, this may not be true. Look at this example:

$$\begin{tikzcd} & \top \arrow[dl, dash] \arrow[d, dash] \arrow[dr, dash] & \\ a \arrow[dr, dash] & b \arrow[d, dash] & c \arrow[dl, dash] \\ & \bot & \end{tikzcd}$$

The lattice is not distributive, because $(a \wedge b) \vee c = c \neq \top = (a \vee c) \wedge (b \vee c)$. Then $\{\top\}$, $\{\top, a\}$, $\{\top, b\}$, $\{\top, c\}$ and $\{\top, d\}$ are filters, $\{\top, a\}$, $\{\top, b\}$, $\{\top, c\}$ are maximal, but none of them is prime.

Prime ideals and prime filters are not only dual notions, they are also complementary concepts.

Lemma 1.5.36 *In a lattice L a subset F is a prime filter iff its complement $L \setminus F$ is a prime ideal.*

Proof Exercise 1.15. $\dashv$

When defining a lattice as a generalization of the power set construct, we restricted the attention to joins and meets of elements but neglected the observation that in a power set each set has a complement. The corresponding abstraction is a Boolean algebra. Such a *Boolean algebra* B is a distributive lattice such that there exists a unary operation $- : B \to B$ such that

Boolean algebra

$$a \vee -a = \top$$
$$a \wedge -a = \bot$$

$-a$ is called the complement of a. We assume that $\wedge$ binds stronger than $\vee$ and that complementation binds stronger than the binary operations.

The power of complementation shows already in the next lemma, which relates prime ideals to maximal ideals and prime filters to maximal filters.

Lemma 1.5.37 *Let B be a Boolean algebra. Then an ideal is maximal iff it is prime, and a filter is maximal iff it is prime.*

Proof Note that a Boolean algebra is a distributive lattice with more than one element (viz., $\bot$ and $\top$). We prove the assertion only for filters. That a maximal filter is prime has been shown in Lemma 1.5.35. If $\mathcal{F}$ is not maximal, there exists a with $a \notin \mathcal{F}$ and $-a \notin \mathcal{F}$ by Lemma 1.5.21. But $\top = a \vee -a \in \mathcal{F}$; hence $\mathcal{F}$ is not prime. $\dashv$

This is another and probably surprising equivalent to $(\mathbb{AC})$.

$$\boxed{\begin{array}{c}(\mathbb{MI})\text{ Each lattice with more than one}\\ \text{element contains a maximal ideal.}\end{array}}$$

Theorem 1.5.38 $(\mathbb{MI})$ *is equivalent to* $(\mathbb{AC})$.

Proof 1. $(\mathbb{MI}) \Rightarrow (\mathbb{AC})$: We show actually that $(\mathbb{MI})$ implies $(\mathbb{MP})$; an application of Theorem 1.5.4 will then establish the claim. Let $\mathcal{F} \subseteq \mathcal{P}(S)$ be a family of finite character. In order to apply $(\mathbb{MI})$, we need a lattice, which we will define now. Define $\mathcal{L} := \mathcal{F} \cup \{S\}$, and put for $X, Y \in \mathcal{F}$

$$X \wedge Y := X \cap Y,$$
$$X \vee Y := \begin{cases} X \cup Y, & \text{if } X \cup Y \in \mathcal{F}, \\ S, & \text{otherwise.} \end{cases}$$

Then $\mathcal{L}$ is a lattice with top element S and bottom element $\emptyset$. Let $\mathcal{M}$ be a maximal ideal in $\mathcal{L}$; then we assert that $M^* := \bigcup \mathcal{M}$ is a maximal element of $\mathcal{F}$. Then $M^* \neq S$.

First we show that $M^* \in \mathcal{F}$. If $\{a_1, \ldots, a_k\} \in M^*$, then we can find $M_i \in \mathcal{M}$ such that $m_i \in M_i$ for $1 \leq i \leq n$. Since $\mathcal{M}$ is an ideal in $\mathcal{L}$, we know that $M_1 \vee \ldots \vee M_n \in \mathcal{M}$, so that $\{a_1, \ldots, a_k\} \in \mathcal{F}$; hence $M^* \in \mathcal{F}$.

Now assume that M^* is not maximal, then we can find $N \in \mathcal{F}$ such that M^* is a proper subset of N, and hence there exists $t \in N$ such that $t \notin M^*$. Because $N \in \mathcal{F}$ and $\mathcal{F}$ are of finite character, $\{t\} \in \mathcal{F}$. Now put $\mathcal{M}' := \mathcal{M} \cup \{M \vee \{t\} \mid M \in \mathcal{M}\} = \mathcal{M} \cup \{M \cup \{t\} \mid M \in \mathcal{M}\}$; then $\mathcal{M}'$ is an ideal in $\mathcal{L}$ which properly contains $\mathcal{M}$. This is a contradiction; hence we have found a maximal element of $\mathcal{F}$.

2. $(\mathbb{AC}) \Rightarrow (\mathbb{MI})$: Again, we use the equivalences in Theorem 1.5.4, because we actually show $(\mathbb{ZL}) \Rightarrow (\mathbb{MI})$. Let L be a lattice with at least two elements, and order

$$\mathcal{I} := \{I \subseteq L \mid I \text{ is an ideal in } L\}$$

by inclusion. Because $\{b \in L \mid b \leq a\} \in \mathcal{I}$ for $a \in L, a \neq \top$ (by assumption, such an element exists), we know that $\mathcal{I} \neq \emptyset$. If $\mathcal{C} \subseteq \mathcal{I}$ is a chain, then $I := \bigcup \mathcal{C} \in \mathcal{I}$. In fact, $\emptyset \neq I \neq L$, because $\top \notin I$, and if $a, b \in I$, we find I_1, I_2 with $a \in I_1, b \in I_2$, because $\mathcal{C}$ is a chain; we may assume that $I_1 \subseteq I_2$, and hence $a, b \in I_2$, so that $a \vee b \in I_2 \subseteq I$. If $a \leq b$ and $b \in I$, then $a \in I$, because $b \in I_1$ for some $I_1 \in \mathcal{I}$. Hence each chain has an upper bound in $\mathcal{I}$. $(\mathbb{ZL})$ implies the existence of a maximal element $M \in \mathcal{I}$. $\dashv$

Since each Boolean algebra is a lattice with more than the top element, the following corollary is a consequence of Theorem 1.5.38. It is known under the name *Prime Ideal Theorem*. We know from Lemma 1.5.37 that prime ideals and maximal ideals are really the same.

Theorem 1.5.39 $(\mathbb{AC})$ *implies the existence of a prime ideal in a Boolean algebra.* $\dashv$

The converse does not hold—it can be shown that the Prime Ideal Theorem is strictly weaker than $(\mathbb{AC})$ [Jec06, p. 81].

1.5.7 The Stone Representation Theorem

Let us stick for a moment to Boolean algebras and discuss the famous Stone Representation Theorem, which requires the Prime Ideal Theorem at a crucial point.

Fix a Boolean algebra B and define for two elements $a, b \in B$ their *symmetric difference* $a \ominus b$ through

$$a \ominus b := (a \wedge -b) \vee (-a \wedge b).$$

If $B = \mathcal{P}(S)$ for some set S and if $\wedge, \vee, -$ are the respective set operations $\cap, \cup, S \setminus \cdot$, then $A \ominus B$ is in fact equal to the symmetric difference $A \Delta B := (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (B \cap A)$.

$\sim_I$ Fix an ideal I of B, and define

$$a \sim_I b \Leftrightarrow a \ominus b \in I.$$

Then $\sim_I$ is a congruence, i.e., an equivalence relation which is compatible with the operations on the Boolean algebra. This will be shown now through a sequence of statements.

We state some helpful properties.

Lemma 1.5.40 *Let B be a Boolean algebra; then*

1. $a \ominus a = \bot$, $a \ominus b = b \ominus a$ *and* $a \ominus b = (-a) \ominus (-b)$.
2. $a \ominus b = (a \vee b) \wedge -(a \wedge b)$.
3. $(a \ominus b) \wedge c = (a \wedge c) \ominus (b \wedge c)$ *and* $c \wedge (a \ominus b) = (c \wedge a) \ominus (c \wedge b)$.

Proof The properties under 1 are fairly obvious, 2 is calculated directly using distributivity, and finally the first part of 3 follows; thus

$$\begin{aligned}(a \wedge c) \ominus (b \wedge c) &= (a \wedge c \wedge -(b \wedge c)) \vee (b \wedge c \wedge -(a \wedge c)) \\ &= (a \wedge c \wedge (-b \vee -c)) \vee (b \wedge c \wedge (-a \vee -c)) \\ &= (a \wedge -b \wedge c) \vee (b \wedge -a \wedge c) \\ &= (a \ominus b) \wedge b,\end{aligned}$$

because $a \wedge c \wedge -c = \bot = b \wedge c \wedge -c$. ⊣

Lemma 1.5.41 *$\sim_I$ is an equivalence relation on B with these properties:*

1. *$a \sim_I a'$ and $b \sim_I b'$ imply $a \wedge b \sim_I a' \wedge b'$ and $a \vee b \sim_I a' \vee b'$.*
2. *$a \sim_I a'$ implies $-a \sim_I -a'$.*

Proof Because $a \ominus a' \in I$ and $b \ominus b' \in I$, we conclude that $(a \ominus a') \vee (b \ominus b') \in I$; thus

$$\begin{aligned}(a \vee a') \ominus (b \vee b') &\le ((a \vee b) \wedge -(a \wedge b)) \vee ((a' \vee b') \wedge -(a' \wedge b')) \\ &= (a \ominus a') \vee (b \ominus b') \in I.\end{aligned}$$

Since I is an ideal, we conclude $(a \vee a') \ominus (b \vee b') \in I$.

From Lemma 1.5.40, we conclude that $a \wedge b \sim_I a' \wedge b \sim_I a' \wedge b'$. The assertion about complementation follows from Lemma 1.5.40 as well. $\dashv$

Denote by $[x]_{\sim_I}$ the equivalence class of $x \in B$, and let $\eta_{\sim_I} : x \mapsto [x]_{\sim_I}$ be the associated factor map. Define on the factor space $B/I := \{[x]_{\sim_I} \mid x \in B\}$ the operations $B/I, \eta_{\sim_I}$

$$\begin{aligned}[a]_{\sim_I} \wedge [b]_{\sim_I} &:= [a \wedge b]_{\sim_I}, \\ [a]_{\sim_I} \vee [b]_{\sim_I} &:= [a \vee b]_{\sim_I}, \\ -[a]_{\sim_I} &:= [-a]_{\sim_I}.\end{aligned}$$

We have also

$$\begin{aligned}[a]_{\sim_I} \le [b]_{\sim_I} &\Leftrightarrow a \ominus (a \wedge b) \in I \Leftrightarrow b \ominus (a \vee b) \in I, \\ a \in I &\Leftrightarrow a \sim_I \perp.\end{aligned}$$

The following statement is now fairly easy to prove. Recall that a homomorphism $f : (B, \wedge, \vee, -) \to (B', \wedge', \vee', -')$ is a map $f : B \to B'$ such that

$$f(a \wedge b) = f(a) \wedge' f(b),\ f(a \vee b) = f(a) \vee' f(b), and\ f(-a) = -' f(a)$$

for all $a, b \in B$ are valid.

Proposition 1.5.42 *The factor space B/I is a Boolean algebra, and $\eta_{\sim_I}$ is a homomorphism of Boolean algebras.*

Proof The operations on B/I are well defined by Lemma 1.5.41 and yield a lattice with $[\top]_{\sim_I}$ as the largest and and $[\bot]_{\sim_I}$ as the smallest element, resp. Hence $-$ is a complementation operator on B/I because

$$[a]_{\sim_I} \wedge [-a]_{\sim_I} = [\bot]_{\sim_I},$$
$$[a]_{\sim_I} \vee [-a]_{\sim_I} = [\top]_{\sim_I}.$$

It is evident from the construction that $\eta_{\sim_I}$ is a homomorphism. ⊣

The Prime Ideal Theorem implies that the Boolean algebra B/I has a prime ideal J by Corollary 1.5.39. This observation leads to a stronger version of this theorem for the given Boolean algebra.

Theorem 1.5.43 *Let I be an ideal in a Boolean algebra. Then* ($\mathbb{AC}$) *implies that there exists a prime ideal K which contains I.*

Plan

Proof 0. The plan of the proof is fairly straightforward: We know that B/I has a maximal ideal by the Prime Ideal Theorem. This prime ideal is lifted to the given Boolean algebra B, and then we claim that this is the prime ideal on B we are looking for.

1. Construct the factor algebra B/I; then ($\mathbb{AC}$) implies that this Boolean algebra has a prime ideal J. We claim that

$$K := \{x \in B \mid [x]_{\sim_I} \in J\}$$

is the desired prime ideal. Since $I = [\bot]_{\sim_I} \in J$, we see that $I \subseteq K$ holds; thus $K \neq \emptyset$.

2. K is an ideal. If $K = B$, then $\top \in K$ which would mean $[\top]_{\sim_I} \in J$, but this is impossible. Let $a \leq b$ with $b \in K$; hence $a = a \wedge b$, so that $[a]_{\sim_I} = [a \wedge b]_{\sim_I}$. Because $b \in K$, we infer $[a \wedge b]_{\sim_I} \in J$; hence $[a]_{\sim_I} \in J$, so that $a \in K$. If $a, b \in K$, then $a \vee b \in K$, because J is an ideal.

3. K is prime. In fact, we have

$$\begin{array}{rcl} a \wedge b \in K & \Leftrightarrow & [a \wedge b]_{\sim_I} \in J \quad \Leftrightarrow \quad [a]_{\sim_I} \wedge [b]_{\sim_I} \in J \\ & \Rightarrow & [a]_{\sim_I} \in J \text{ or } [b]_{\sim_I} \in J \quad \Leftrightarrow \quad a \in K \text{ or } b \in K. \end{array}$$

⊣

As a consequence, we can find in a Boolean algebra for any given element $a \neq \top$ a prime ideal which does contain it.

Corollary 1.5.44 *Let B be a Boolean algebra and assume that* $(\mathbb{AC})$ *holds.*

1. *Given $a \neq \top$, there exists a prime ideal which contains a.*
2. *Given $a, b \in B$ with $a \neq b$, there exists a prime ideal which contains a but not b.*
3. *Given $a, b \in B$ with $a \neq b$, there exists an ultrafilter which contains a but not b.*

Proof We find a prime ideal K which extends the ideal $\{x \in B \mid x \leq a\}$. This establishes the first part.

If $a \neq b$, we have $a \ominus b \neq \bot$, so $a \wedge -b \neq \bot$ or $-a \wedge b \neq \bot$. Assume the former; then there exists a prime ideal K with $-(a \wedge -b) \in K$, so that both $b \in K$ and $-a \in K$ hold. Since $-a \in K$ implies $a \notin K$, we are done with the second part. The third part follows through Lemma 1.5.36. $\dashv$

This yields one of the classics, the Stone Representation Theorem. It states that each Boolean algebra is essentially a set algebra, i.e., a Boolean algebra comprised of sets.

Theorem 1.5.45 *Let B be a Boolean algebra, and assume that* $(\mathbb{AC})$ *holds. Then there exists a set S_0 and a Boolean set algebra $S \subseteq \mathcal{P}(S_0)$ such that B is isomorphic to S.*

Proof 0. We map each element of the Boolean algebra to the ultrafilters in which it is contained as an element. This yields a map which is compatible with the Boolean structure, from which we obtain the objects we are looking for. Outline

1. Define

$$\mathcal{S}_0 := \{U \mid U \text{ is an ultrafilter on } B\},$$
$$\psi(b) := \{U \in \mathcal{S}_0 \mid b \in U\}.$$

Then these properties are easily established:

$$\psi(b_1 \wedge b_2) = \psi(b_1) \cap \psi(b_2),$$
$$\psi(b_1 \vee b_2) = \psi(b_1) \cup \psi(b_2),$$
$$\psi(-b) = \mathcal{S}_0 \setminus \psi(b).$$

For example, we obtain from Lemma 1.5.35 that

$$\begin{aligned} U \in \psi(b_1 \wedge b_2) &\Leftrightarrow b_1 \wedge b_2 \in U \\ &\Leftrightarrow b_1 \in U \text{ and } b_2 \in U \\ &\Leftrightarrow U \in \psi(b_1) \text{ and } U \in \psi(b_2) \\ &\Leftrightarrow U \in \psi(b_1) \cap \psi(b_2). \end{aligned}$$

Similarly, $U \in \psi(-b) \Leftrightarrow -b \in U \Leftrightarrow b \notin U \Leftrightarrow U \notin \psi(b)$ by Lemma 1.5.21, because U is an ultrafilter.

3. Because we can find by Corollary 1.5.44 for $b_1 \neq b_2$ an ultrafilter which contains b_1, but not b_2, we conclude that ψ is injective (this is actually the place where $(\mathbb{AC})$ is used). Thus the Boolean algebras B and $\psi\left[B\right]$ are isomorphic, and the latter one is comprised of sets. $\dashv$

The Stone Representation Theorem gives a representation of a Boolean algebra as a set algebra. So, at first sight, the effort of introducing the additional abstraction looks futile. But it is not. First, it does not say that a Boolean algebra is always a full power set. Second, it is sometimes much easier in an application to work with an abstract tool like a Boolean algebra than with a set algebra (because you then have to cater for a carrier set, after all). A third reason is a classification issue—Boolean algebras are isomorphic to set algebras, are lattices then always isomorphic to set lattices? The representation through down sets from Example 1.5.30 seems to suggest just that. But looking at this representation more closely, one sees that additional properties are probably not preserved; for example, a Boolean algebra may be represented as a lattice through its down sets, but it is far from clear that the representation preserves also complements. Thus we do not have in general such a clear-cut picture for general lattices, as we have for Boolean algebras.

1.5.8 Compactness and Alexander's Subbase Theorem

We will prove in this section Alexander's Subbase Theorem as an application for Zorn's Lemma. The theorem states that when proving a topological space compact, one may restrict one's attention to a particular subclass of open sets, a class which is usually easier to handle than the full family of open sets. This application of Zorn's Lemma is interesting because it shows in which way a maximality argument can be used for establishing a property through a subclass (rather than extending a property until maximality puts a stop to it, as we did in showing that each vector space has a basis). Alexander's Theorem is also a very practical tool, as we will see later.

This section assumes that $(\mathbb{AC})$ holds.

We start with the closed interval $[u, v]$ with $-\infty < u < v < +\infty$ as an important example of a compact space. It has the following property: Each cover through a countable number of open intervals contains a finite subcover which already cover the interval. This is what the famous *Heine–Borel Theorem* states. We give below Borel's proof [Fic64, vol. I, p. 163].

Theorem 1.5.46 *Let an interval* $[u, v]$ *with* $-\infty < u < v < +\infty$ *be given. Then each cover* $\{]x_n, y_n[\mid n \in \mathbb{N}\}$ *of* $[u, v]$ *through a countable number of open intervals contains a finite cover* $]x_{n_1}, y_{n_1}[, \dots,]x_{n_k}, y_{n_k}[$.

Proof Suppose the assertion is false; then either $[u, 1/2(u + v)]$ or $[1/2(u + v), v]$ is not covered by finitely many of those intervals; select the corresponding one, and call it $[a_1, b_1]$. This interval can be halved; let $[a_2, b_2]$ be the half which cannot be covered by finitely many intervals. Repeating this process, one obtains a sequence $\{[a_n, b_n] \mid n \in \mathbb{N}\}$ of closed intervals, each having half of the length of its predecessor and each one not being covered by an finite number of intervals from $\{]x_n, y_n[\mid n \in \mathbb{N}\}$. Because the lengths of the intervals shrink to zero, there exists $c \in [u, v]$ with $\lim_{n\to\infty} a_n = c = \lim_{n\to\infty} b_n$; hence $c \in]x_m, y_m[$ for some m. But there is some $n_0 \in \mathbb{N}$ with $[a_n, b_n] \subseteq]x_m, y_m[$ for $n \geq n_0$, contradicting the assumption that $[a_n, b_n]$ cannot be covered by a finite number of those intervals. $\dashv$

Although the proof is given for a countable cover, its analysis shows that it goes through for an arbitrary cover of open intervals. This is so because each cover induces a partition of the interval considered into two parts, so a sequence of intervals will result in any case.

This section will discuss compact spaces which have the property that an arbitrary cover contains a finite one. To be on firm ground, we first introduce topological spaces as the kind of objects to be discussed here.

Definition 1.5.47 *Given a set X, a subset $\tau \subseteq \mathcal{P}(X)$ is called a* topology *iff these conditions are satisfied:*

- *$\emptyset, X \in \tau$.*
- *If $G_1, \ldots, G_k \in \tau$, then $G_1 \cap \ldots \cap G_k \in \tau$, and thus τ is closed under finite intersections.*
- *If $\tau_0 \subseteq \tau$, then $\bigcup \tau_0 \in \tau$, and thus τ is closed under arbitrary unions.*

The pair (X, τ) is then called a topological space*, and the elements of τ are called* open sets*. An* open neighborhood *U of an element $x \in X$ is an open set U with $x \in U$, a* neighborhood *of x is a set which contains an open neighborhood of x.*

These are the topologies one can always find on a set X.

Example 1.5.48 $\mathcal{P}(X)$ and $\{\emptyset, X\}$ are always topologies; the former one is called the *discrete topology*, the latter one is called *indiscrete*. ✌

The topology one deals with usually on the reals is given by intervals, and the plane is topologically described by open balls (well, they really are disks, but they are given through measuring a distance, and in this case the name "ball" sticks).

Example 1.5.49 Call a set $G \subseteq \mathbb{R}$ open iff for each $x \in G$ there exists $a, b \in \mathbb{R}$ with $a < b$ such that $x \in \,]a, b[\, \subseteq G$; note that $\emptyset$ is open. Then the open sets form a topology on the reals, which is also called the *interval topology*.

Clearly, G is open iff, given $x \in G$, there exists $\epsilon > 0$ with $]x - \epsilon, x + \epsilon[\subseteq G$. Call a subset $G \subseteq R^2$ of the Euclidean plane open iff, given $x \in G$, there exists $\epsilon > 0$ such that $B_\epsilon(x) \subseteq G$, where

$$B(\langle x_1, x_2\rangle, r) := \{\langle y_1, y_2\rangle \mid \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2} < r\}$$

is the open ball centered at $\langle x_1, x_2\rangle$ with radius r. ✌

τ_Y

Let (X, τ) be a topological space. If $Y \subseteq X$, then the trace of τ on Y gives a topology τ_Y on Y, formally, $\tau_Y := \{G \cap Y \mid G \in \tau\}$, the *subspace topology*. This permits sometimes to transfer a property from the space to its subsets.

Closed set

A set $F \subseteq X$ is called *closed* iff its complement $X \setminus F$ is open. Then both $\emptyset$ and X are closed, and the closed sets are closed (no pun intended) under arbitrary intersections and finite unions. We associate with each set an open set and a closed set.

Definition 1.5.50 *Let $M \subseteq X$; then*

- $M^o := \bigcup\{G \in \tau \mid G \subseteq M\}$ *is called the* interior *of* M.
- $M^a := \bigcap\{F \subseteq X \mid M \subseteq F$ *and* F *is closed*$\}$ *is called the* closure *of* M.
- $\partial M := M^a \setminus M^o$ *is called the* boundary *of* M. $M^o, M^a, \partial M$

We have always $M^o \subseteq M \subseteq M^a$; this is apparent from the definition. Clearly, M^o is an open set, and it is the largest open set which is contained in M, so that M is open iff $M = M^o$. Similarly, M^a is a closed set, and it is the smallest closed set which contains M. We also have M is closed iff $M = M^a$. The boundary ∂M is also a closed set, because it is the intersection of two closed sets, and we have $\partial M = \partial(X \setminus M)$. M is closed iff $\partial M \subseteq M$. All this is easily established through the definitions.

Look at the indiscrete topology: Here we have $\{x\}^o = \emptyset$ and $\{x\}^a = X$ for each $x \in X$. For the discrete topology, one sees $A^o = A^a = A$ for each $A \subseteq X$.

Example 1.5.51 In the Euclidean topology on $\mathbb{R}^2$ of Example 1.5.49, we have

$$B_r(x_1, x_2)^a = \{\langle y_1, y_2\rangle \mid \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2} \leq r\},$$

$$\partial B_r(x_1, x_2) = \{\langle y_1, y_2\rangle \mid \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2} = r\}.$$

✌

Just to get familiar with boundaries

Lemma 1.5.52 *Let (X, τ) be a topological space, $A \subseteq X$. Then $x \in \partial A$ iff each open neighborhood of x has a nonempty intersection with A and with $X \setminus A$. In particular $\partial A = \partial(X \setminus A)$ and $\partial(A \cup B) \subseteq (\partial A) \cup (\partial B)$.*

Proof Let $x \in \partial A$ and U an open neighborhood of x. If $A \cap U = \emptyset$, then $A \subseteq X \setminus U$, so $x \notin A^a$, and if $U \cap X \setminus A = \emptyset$, it follows $x \in A^o$. So we arrive at a contradiction. Assume that $x \in \bigcap\{U \in \tau \mid x \in U, U \cap A \neq \emptyset, U \cap X \setminus A \neq \emptyset\}$; then $x \notin A^o$; similarly, $x \notin X \setminus A^o = X \setminus (A^a)$.
$\dashv$

Clopen set

A set without a boundary is both closed and open, so it is called *clopen*. The clopen sets of a topological space form a Boolean algebra.

Sometimes it is sufficient to describe the topology in terms of some special sets, like the open balls for the Euclidean topology. These balls form a base in the following sense:

Base, subbase

Definition 1.5.53 *A subset $\beta \subseteq \tau$ of the open sets is called a* base for the topology *iff for each open set $G \in \tau$ and for each $x \in G$, there exists $B \in \beta$ such that $x \in B \subseteq G$ and thus iff each open set is the union of all base elements contained in it.*

A subset $\sigma \subseteq \tau$ is called a subbase *for τ iff the set of finite intersections of elements of σ forms a base for τ.*

Then the open intervals are a base for the interval topology, and the open balls are a base for the Euclidean topology (actually, we did introduce the respective topologies through their bases). A subbase for the interval topology is given by the sets $\{]-\infty, a[\mid a \in \mathbb{R}\}$, because the set of finite intersections includes all open intervals, which in turn form a base. Bases and subbases are not uniquely determined, for example, $\{]r, s[\mid r < s, r, s \in \mathbb{Q}\}$ is a base for the interval topology.

Let us return to the problem discussed in the opening of this section. We have seen that bounded closed intervals have the remarkable property that, whenever we cover them by an arbitrary number of open intervals, we can find a finite collection among these intervals which already cover the interval. This property can be generalized to arbitrary topological spaces; subsets with this property are called compact, formally:

Definition 1.5.54 *The topological space (X, τ) is called* compact *iff each cover of X by open sets contains a finite subcover.*

Thus X is compact iff, whenever $(G_i)_{i \in I}$ is a collection of open sets with $X = \bigcup_{i \in I} G_i$, there exists $I_0 \subseteq I$ finite such that $C \subseteq \bigcup_{i \in I_0} G_i$. It is apparent that compactness is a generalization of finiteness, so that compact sets are somewhat small, measured in terms of open sets. Consider as a trivial example the discrete topology. Then X is compact precisely when X is finite.

This is an easy consequence of the definition.

Lemma 1.5.55 *Let (X, τ) be a compact topological space and $F \subseteq X$ closed. Then (F, τ_F) is a compact topological space.* $\dashv$

The following example shows a close connection of Boolean algebras to compact topological spaces; this is the famous *Stone Duality Theorem*.

Example 1.5.56 Let B be a Boolean algebra with $\wp_B$ as the set of all prime ideals of B. Define

$$X_a := \{I \in \wp_B \mid a \notin I\}.$$

Then we have these properties:

- $X_\top = \wp_B$, since an ideal does not contain $\top$.
- $X_{-a} = \wp_B \setminus X_a$. To see this, let I be a prime filter, and then $I \in X_{-a}$ iff $-a \notin I$; this is the case iff $-a \in B \setminus I$ and hence iff $a \notin B \setminus I$, since $B \setminus I$ is a maximal filter by Lemmas 1.5.36 and 1.5.37; the latter condition is equivalent to $a \in I$ and hence to $I \notin X_a$.
- $X_{a \wedge b} = X_a \cap X_b$ and $X_{a \vee b} = X_a \cup X_b$. This follows similarly from Lemma 1.5.35.

Define a topology τ on $\wp_B$ by taking the sets X_a as a base, formally

$$\beta := \{X_a \mid a \in B\}.$$

We claim that $(\wp_B, \tau)$ is compact. In fact, let $\mathcal{U}$ be a cover of $\wp_B$ with open sets. Because each $U \in \mathcal{U}$ can be written as a union of elements of β, we may and do assume that $\mathcal{U} \subseteq \beta$, so that $\mathcal{U} = \{X_a \mid a \in A\}$ for some $A \subseteq B$. Now let J be the ideal generated by A, so that J can be written as

$$J := \{b \in B \mid b \leq a_1 \vee \ldots \vee a_k \text{ for some } a_1, \ldots, a_k \in A\}.$$

We distinguish these cases:

$\top \in J$**:** In this case we have $\top = a_1 \vee \ldots \vee a_k$ for some $a_1, \ldots, a_k \in A$, which means

$$\wp_B = X_\top = X_{a_1 \vee \ldots \vee a_k} = X_{a_1} \cup \ldots \cup X_{a_k}$$

with $X_{a_1}, \ldots, X_{a_k} \in \mathcal{U}$, so we have found a finite subcover in $\mathcal{U}$.

$\top \notin J$**:** Then J is a proper ideal, so by Corollary 1.5.39 there exists a prime ideal K with $J \subseteq K$. But we cannot find $a \in A$ such that $K \in X_a$: Assume on the contrary that $K \in X_a$ for some $a \in A$, then $a \notin K$, and hence $-a \in K$, since K is prime (Lemma 1.5.36). But by construction $a \in J$, since $a \in A$, which implies $a \in K$; hence $\top \in K$, a contradiction. Thus $K \in \wp_B$, but K fails to be covered by $\mathcal{U}$, which is a contradiction.

Thus $(\wp_B, \tau)$ is a compact space, which is sometimes called the *prime ideal space* of the Boolean algebra.

We conclude that the sets X_a are clopen, since $X_{-a} = \wp_B \setminus X_a$. Moreover, each clopen set in this space can be represented in this way. In fact, let U be clopen, and thus $U = \bigcup\{X_a \mid a \in A\}$ for some $A \subseteq B$. Since U is closed, it is compact by Lemma 1.5.55, so there exist $a_1, \ldots, a_n \in A$ such that $U = X_{a_1} \cup \ldots \cup X_{a_n} = X_{a_1 \vee \ldots \vee a_n}$. ✌

Compactness is formulated in terms of a cover through arbitrary open sets. Alexander's Theorem states that it is sufficient to consider covers which come from a subbase for the topology. This is usually quite a considerable help, since subbases are mostly easier to handle than the collection of all open sets; Example 1.5.58 confirms this impression. The proof comes as an application of Zorn's Lemma. The proof follows essentially the one given in [HS65, Theorem 6.40].

Theorem 1.5.57 *Let (X, τ) be a topological space with a subbase σ. Then the following statements are equivalent:*

1. *X is compact.*
2. *Each cover of X by elements of σ contains a finite subcover.*

Proof 0. Because the elements of a subbase are open, the implication 1 $\Rightarrow$ 2 is trivial; hence we have to show 2 $\Rightarrow$ 1. The idea of Proof outline the proof goes as follows: If the assertion is false, there exists a cover which does not have a finite subcover. Take the set of all these covers, and order them by inclusion. It is not difficult to see that each chain has an upper bound in this set, so Zorn's Lemma gives a maximal element. Maximality is somewhat fragile here, because adding something to this maximal element will break it. This will permit us to derive a contradiction.

1. Assume that the assertion is false, and define

$$\mathfrak{Z} := \{\mathcal{C} \mid \mathcal{C} \text{ is on open cover of } X \text{ without a finite subcover}\}.$$

Order $\mathfrak{Z}$ by inclusion, and let $\mathfrak{Z}_0 \subseteq \mathfrak{Z}$ be a chain; then $\mathcal{C} := \bigcup \mathfrak{Z}_0 \in \mathfrak{Z}$. In fact, it is clear that $\mathcal{C}$ is a cover, and assume that $\mathcal{C}$ has a finite subcover, say $\{E_1, \ldots, E_k\}$. Then $E_j \in \mathcal{C}_j \in \mathfrak{Z}_0$, and since $\mathfrak{Z}_0$ is a chain with respect to inclusion, we find some $\mathcal{C}_i \in \mathfrak{Z}_0$ with $\{E_1, \ldots, E_k\} \subseteq \mathcal{C}_i$, which is a contradiction. By Zorn's Lemma, $\mathfrak{Z}$ has a maximal element $\mathcal{V}$. This means that

- $\mathcal{V}$ is an open cover of X.
- $\mathcal{V}$ does not contain a finite subcover.
- If $U \in \tau$ is open with $U \notin \mathcal{V}$, then $\mathcal{V} \cup \{U\}$ contains a finite subcover.

Let $\mathcal{W} := \mathcal{V} \cap \sigma$, and hence $\mathcal{W}$ contains all elements of $\mathcal{V}$ which are taken from the subbase. By assumption, no finite subfamily of $\mathcal{W}$ covers X; hence $\mathcal{W}$ is not a cover for X, which implies that $R := X \setminus \bigcup \mathcal{W} \neq \emptyset$. Let $x \in R$; then there exists $V \in \mathcal{V}$ such that $x \in V$, because $\mathcal{V}$ is a cover for X. Since V is open and σ is a subbase, we find $S_1, \ldots, S_k \in \sigma$ with $x \in S_1 \cap \ldots \cap S_n \subseteq V$. Because $x \notin \bigcup \mathcal{W}$, we conclude that no S_j is an element of $\mathcal{V}$ (otherwise $S_j \in \mathcal{V} \cap \sigma = \mathcal{W}$, a contradiction). $\mathcal{V}$ is maximal, each S_j is open, and thus $\mathcal{V} \cup \{S_j\}$ contains a finite cover of X. Hence we can find for each j some open set A_j which is a finite union of elements in $\mathcal{V}$ such that $A_j \cup S_j = X$. But this means

$$V \cup \bigcup_{j=1}^{k} A_j \supseteq (\bigcap_{j=1}^{k} S_j) \cup (\bigcup_{j=1}^{k} A_j) = X.$$

Hence X can be covered through a finite number of elements in $\mathcal{V}$; this is a contradiction to the maximality of $\mathcal{V}$. $\dashv$

Observe how $(\mathbb{ZL})$ enters the argument precisely when we need a maximal cover with no finite subcover.

The Priestley topology as discussed by Goldblatt [Gol12] provides a first example for the use of Alexander's Theorem. It illustrates that using a cover comprised of elements coming from a subbase simplifies the argumentation considerably.

Example 1.5.58 Given $x \in X$, define

$$\|x\| := \{A \subseteq X \mid x \in A\},$$
$$-\|x\| := \{A \subseteq X \mid x \notin A\}.$$

The *Priestley topology* on $\mathcal{P}(X)$ is defined as the topology which is generated by the subbase

$$\sigma := \{\|x\| \mid x \in X\} \cup \{-\|x\| \mid x \in X\}.$$

Hence the basic sets of this topology have the form

$$\|x_1\| \cap \ldots \cap \|x_k\| \cap -\|y_1\| \cap \ldots \cap -\|y_n\|$$

for $x_1, \ldots, x_k, y_1, \ldots, y_n \in X$ and some $n, k \in \mathbb{N}$.

We claim that $\mathcal{P}(X)$ is compact in the Priestley topology. In fact, let $\mathcal{C}$ be a cover of $\mathcal{P}(X)$ with elements from the subbase σ. Put $P := \{x \in X \mid -\|x\| \in \mathcal{C}\}$. Then $P \in \mathcal{P}(X)$, so we must find some element from $\mathcal{C}$ which contains P. If $P \in -\|x\| \in \mathcal{C}$ for some $x \in X$, this means $x \notin P$, so by definition $-\|x\| \notin \mathcal{C}$, which is a contradiction. Thus there exists $x \in X$ such that $P \in \|x\| \in \mathcal{C}$. But this means $x \in P$; hence $-\|x\| \in \mathcal{C}$, so $\{\|x\|, -\|x\|\} \subseteq \mathcal{C}$ is a cover of $\mathcal{P}(X)$. Thus $\mathcal{P}(X)$ is compact by Alexander's Theorem 1.5.57. ✌

Alexander's Subbase Theorem will be of considerable help, e.g., when characterizing compactness through ultrafilters in Theorem 3.2.11 and in establishing Tihonov's Theorem 3.2.12 on product compactness. It also emphasizes the close connection of compactness and $(\mathbb{AC})$.

1.6 Boolean σ-Algebras

We generalize the notion of a Boolean algebra by introducing countable operations, leading to Boolean σ-algebras. This extension becomes important, e.g., when working with probabilities or, more generally, with measures. For example, one of the fundamental probability laws states that the probability of a disjoint union of countable events equals the infinite sum of the events' probabilities. In order to express this adequately, the domain of the probability must be closed under countable unions.

We assume in this section that $(\mathbb{AC})$ holds.

Given a Boolean algebra B, we associate as above with the lattice operations on B an order relation $\leq$ by

$$a \leq b \iff a \wedge b = a \ (\iff a \vee b = b).$$

We will switch in the discussion below between the order and the use of the algebraic operations.

Definition 1.6.1 *A Boolean algebra B is called a* Boolean σ-algebra *iff it is closed under countable suprema and infima.*

Example 1.6.2 The power set of each set is a Boolean σ-algebra. Consider

$$\mathcal{A} := \{A \subseteq \mathbb{R} \mid A \text{ is countable or } \mathbb{R} \setminus A \text{ is countable}\}.$$

Then $\mathcal{A}$ is a Boolean σ-algebra (we use here that the countable union of countable set is countable again, hence $(\mathbb{AC})$). This is sometimes called the *countable–cocountable σ-algebra.* On the other hand, her little sister,

$$\mathcal{D} := \{A \subseteq \mathbb{R} \mid A \text{ is finite or } \mathbb{R} \setminus A \text{ is finite}\},$$

the finite–cofinite algebra, is a Boolean algebra, but evidently not σ-algebra. ✌

We define for the countable subset $A = \{a_n \mid n \in \mathbb{N}\}$ of a Boolean algebra B

$$\bigwedge A := \bigwedge_{n \in \mathbb{N}} a_n := \inf\{a_n \mid n \in \mathbb{N}\},$$
$$\bigvee A := \bigvee_{n \in \mathbb{N}} a_n := \sup\{a_n \mid n \in \mathbb{N}\}$$

as its infimum resp. its supremum, in which both exist, since B is closed under countable infima and suprema. In addition, we note that

$$\inf \emptyset = \top,$$
$$\sup \emptyset = \bot.$$

We know that a Boolean algebra is a distributive lattice, in addition to that a stronger infinite distributive law holds for a Boolean σ-algebra.

Lemma 1.6.3 *Let B be a Boolean σ-algebra and $(a_n)_{n \in \mathbb{N}}$ be a sequence of elements in B; then*

$$b \wedge \bigvee_{n \in \mathbb{N}} a_n = \bigvee_{n \in \mathbb{N}} (b \wedge a_n),$$
$$b \vee \bigwedge_{n \in \mathbb{N}} a_n = \bigwedge_{n \in \mathbb{N}} (b \vee a_n)$$

holds for any $b \in B$.

Proof We establish the first equality, and the second one follows by duality. Since $b \wedge a_n \leq b$ and $b \wedge a_n \leq a_n$, we see that $\bigvee_{n\in\mathbb{N}}(b \wedge a_n) \leq b \wedge \bigvee_{n\in\mathbb{N}} a_n$. For establishing the reverse inequality, assume that s is an upper bound to $\{b \wedge a_n \mid n \in \mathbb{N}\}$; hence $b \wedge a_n \leq s$ for all $n \in \mathbb{N}$, and consequently, $a_n = (b \wedge a_n) \vee (-b \wedge a_n) \leq s \vee (-b \wedge a_n) \leq s \vee -b$. Thus

$$b \wedge \bigvee_{n\in\mathbb{N}} a_n \leq b \wedge (s \vee -b) = (b \wedge s) \vee (b \wedge -b) \leq b \wedge s \leq s.$$

Hence s is an upper bound to $b \wedge \bigvee_{n\in\mathbb{N}} a_n$ as well. Now apply this to the upper bound $s := \bigvee_{n\in\mathbb{N}}(b \wedge a_n)$. $\dashv$

Let A be a nonempty subset of a Boolean σ-algebra B, and then there exists a smallest σ-algebra C which contains A. In fact, this must be

$$C = \bigcap\{D \subseteq B \mid D \text{ is a } \sigma\text{-algebra with } A \subseteq D\}.$$

We first note that the intersection of a set of σ-algebras is a σ-algebra again. Moreover, there exists always a σ-algebra which contains A, viz., the superset B. Consequently, C, the object of our desire, is denoted by

$\sigma(A)$ $\sigma(A)$, so that $\sigma(A)$ denotes the smallest σ-algebra containing A. σ is an example for a *closure operator*: We have $A \subseteq \sigma(A)$, and $A_1 \subseteq A_2$ implies $\sigma(A_1) \subseteq \sigma(A_2)$; moreover, applying the operator twice does not yield anything new: $\sigma(\sigma(A)) = \sigma(A)$.

Example 1.6.4 Let $\mathcal{A} := \{[a, b] \mid a, b \in [0, 1]\}$ be the set of all closed intervals $[a, b] := \{x \in \mathbb{R} \mid a \leq x \leq b\}$ of the unit interval $[0, 1]$. Denote by $\mathcal{B} := \sigma(\mathcal{A})$ the σ-algebra generated by $\mathcal{A}$; the ele-

Borel sets ments of $\mathcal{B}$ are sometimes called the *Borel sets* of $[0, 1]$. Then the half open intervals $[a, b[$ and $]a, b]$ are members of B. We can write, e.g., $[a, b[= \bigcup_{n\in\mathbb{N}}[a, b - 1/n]$. Since $[a, b - 1/n] \in \mathcal{A} \subseteq \mathcal{B}$ for all $n \in \mathbb{N}$ and since $\mathcal{B}$ is closed under countable unions, the claim follows.

A more complicated Borel set is constructed in this way: Define $C_0 := [0, 1]$ and assume that C_n is defined already as a union of 2^n mutually disjoint closed intervals of length $1/3^n$ each, say $C_n = \bigcup_{1\leq j\leq 2^n} I_j$. Obtain C_{n+1} by removing the open middle third of each interval I_j. For example,

$$\begin{aligned}
C_0 =&[0,1],\\
C_1 =&[0,1/3] \cup [2/3,1],\\
C_2 =&[0,1/9] \cup [2/9,1/3] \cup [2/3,7/9] \cup [8/9,1],\\
C_3 =&[0,1/27] \cup [2/27,1/9] \cup [2/9,7/27] \cup [8/27] \cup [2/3,19/27]\\
&\cup [20/27,7/9] \cup [8/9,25/27] \cup [26/27,1]
\end{aligned}$$

and so on. Clearly $C_n \in \mathcal{B}$, because this set is the finite union of closed intervals. Now put

$$C := \bigcap_{n\in\mathbb{N}} C_n;$$

then $C \in \mathcal{B}$, because it is the countable intersection of sets in B. This set is known as the *Cantor ternary set*. ✌ Cantor set

The next two examples deal with σ-algebras of sets, each defined on the infinite product $\{0,1\}^{\mathbb{N}}$. It may be used as a model for an infinite sequence of flipping coins—0 denoting head and 1 denoting tail. But we can only observe a finite number of these events, probably as long as we want. So we cater for that by having a look at the σ-algebra which is defined by these finite observations.

Example 1.6.5 Let $X := \{0,1\}^{\mathbb{N}}$ be the set of all infinite binary sequences, and put $\mathcal{B} := \sigma(\{A_{k,i} \mid k \in \mathbb{N}, i = 0,1\})$ with $A_{k,i} := \{\langle x_1, x_2, \ldots\rangle \mid x_k = i\}$ as the set of all sequences, the kth component of which is i.

We claim that for $r \in \mathbb{N}_0$ both $S_{k,r} := \{\langle x_1, x_2, \ldots\rangle \in X \mid x_1 + \ldots + x_k = r\}$ and $T_r := \{\langle x_1, x_2, \ldots\rangle \in X \mid \sum_{i=0}^{\infty} x_i = r\}$ are elements of $\mathcal{B}$.

In fact, given a finite binary sequence $v := \langle v_1, \ldots, v_k\rangle$, the set

$$Q_v := \{x \in X \mid \langle x_1, \ldots x_k\rangle = v\} = \bigcap_{i=1}^{k} A_{i,v_i}$$

is a member of $\mathcal{B}$, and the set $L_{k,r}$ of binary sequences of length k which sum up to r is finite. Thus

$$S_{k,r} = \bigcup_{v\in L_{k,r}} Q_v \in \mathcal{B}.$$

Since
$$T_r = \bigcup_{n\in\mathbb{N}} S_{n,r} \cap \left(X^k \times \prod_{k>n}\{0\}\right),$$
the assertion follows also for T_r. ✌

We continue the example by looking at all sequences for which the average result of flipping a coin n times will converge as n tends to infinity. This is a bit more involved because we now have to take care of the limit.

Example 1.6.6 Let $X := \{0,1\}^{\mathbb{N}}$ be the set of all infinite binary sequences as in Example 1.6.5, and put
$$W := \{\langle x_1, x_2, \ldots\rangle \in X \mid \frac{1}{n}\sum_{i=1}^{n} x_i \text{ converges}\}.$$
We claim that $W \in \mathcal{B}$, noting that a real sequence $(y_n)_{n\in\mathbb{N}}$ converges iff it is a Cauchy sequence, i.e., iff given $0 < \epsilon \in \mathbb{Q}$ there exists $n_0 \in \mathbb{N}$ such that $|y_m - y_n| < \epsilon$ for all $n, m \geq n_0$.

Given $F \subseteq \mathbb{N}$ finite, the set
$$\begin{aligned} H_F &:= \{x \in X \mid x_j = 1 \text{ for all } j \in F \text{ and } x_i = 0 \text{ for all } i \notin F\} \\ &= \bigcap_{j\in F} A_{j,1} \cap \bigcap_{i\notin F} A_{i,0} \end{aligned}$$
is a member of $\mathcal{B}$; since there are countably many finite subsets of $\mathbb{N}$ which have exactly r elements, we obtain $T = \bigcup\{H_F \mid F \subseteq \mathbb{N} \text{ with } |F| = r\}$, which is a countable union of elements of $\mathcal{B}$, hence an element of $\mathcal{B}$.

The sequence $\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)_{n\in\mathbb{N}}$ converges iff
$$\forall\epsilon > 0, \epsilon \in \mathbb{Q}\exists n_0 \in \mathbb{N}\forall n \geq n_0 \forall m \geq n_0 : \left|\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{m}\sum_{i=1}^{m} x_i\right| < \epsilon \; and$$
thus iff
$$\langle x_1, x_2, \ldots\rangle \in \bigcap_{\epsilon>0,\epsilon\in\mathbb{Q}} \bigcup_{n_0\in\mathbb{N}} \bigcap_{\mathbb{N}\ni n\geq n_0} \bigcap_{\mathbb{N}\ni m\geq n_0} W_{n,m,\epsilon}$$
with
$$W_{n,m,\epsilon} := \{\langle x_1, x_2, \ldots\rangle \mid \left|\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{m}\sum_{i=1}^{m} x_i\right| < \epsilon\}.$$

Now $\langle x_1, x_2, \ldots\rangle \in W_{n,m,\epsilon}$ iff $\left|m \cdot \sum_{i=1}^{n} x_i - n \cdot \sum_{j=1}^{m} x_j\right| < n \cdot m \cdot \epsilon$. If $n < m$, this is equivalent to

$$-n \cdot m \cdot \epsilon < (m - n) \cdot \sum_{i=1}^{n} x_i - n \cdot \sum_{j=n+1}^{m} x_j < n \cdot m \cdot \epsilon;$$

hence $-n \cdot m \cdot \epsilon < (m - n) \cdot a - n \cdot b < n \cdot m \cdot \epsilon$ for $a = \sum_{i=1}^{n} x_i$ and $b = \sum_{j=n+1}^{m} x_j$; the same applies to the case $m < n$. Since there are only finitely many combinations of $\langle a, b\rangle$ satisfying these constraints, we conclude that $W_{n,m,\epsilon} \in \mathcal{B}$, so that the set W of all sequences for which the average sum converges is a member of $\mathcal{B}$ as well. ✌

1.6.1 Construction Through Transfinite Induction

We will in this section show that the σ-algebra generated by a subset of a Boolean σ-algebra can actually be constructed directly through transfinite induction. We have introduced $\sigma(H)$ as a closure operation, viz., the smallest σ-algebra containing H; this is an operation which works from outside H. In contrast, the inductive construction works from the inside, constructing $\sigma(H)$ through operations with the elements of H, the elements derived from it, etc. In addition, the description of $\sigma(A)$ given above is nonconstructive. Transfinite induction permits us to construct $\sigma(A)$ explicitly (if one dares to speak in these terms of a transfinite operation).

In order to describe it, we introduce two operators on the subsets of B as follows. Let $H \subseteq B$; then H_σ, H_δ

$$H_\sigma := \{\bigvee_{n\in\mathbb{N}} a_n \mid a_n \in H \text{ for all } n \in \mathbb{N}\},$$
$$H_\delta := \{\bigwedge_{n\in\mathbb{N}} a_n \mid a_n \in H \text{ for all } n \in \mathbb{N}\}.$$

Thus H_σ contains all countable suprema of elements of H, and H_δ contains all countable infima. Hence A is a Boolean sub σ-algebra of B iff

$$A_\sigma \subseteq A, A_\delta \subseteq A, \text{ and } \{-a \mid a \in A\} \subseteq A$$

hold.

So could we not, when constructing $\sigma(A)$, just take all complements, then all countable infima and suprema of elements in A, then their countable suprema and infima, and so on? This is the basic idea for the construction. But since the process indicated above is not guaranteed to terminate after a finite number of applications of the σ- and the δ-operations, we do a transfinite construction.

In order to implement this idea, we fix a set algebra $A \subseteq B$; hence A is a Boolean algebra. Thus, we have only to focus on the infinitary operations, and the union and intersection of two elements are special cases. Define by transfinite induction

$$\begin{aligned} A_0 &:= A, \\ A_\zeta &:= \bigcup_{\eta<\zeta} A_\eta, \text{ if } \zeta \text{ is a limit ordinal} \\ A_{\zeta+1} &:= (A_\zeta)_\sigma, \text{ if } \zeta \text{ is odd,} \\ A_{\zeta+1} &:= (A_\zeta)_\delta, \text{ if } \zeta \text{ is even,} \\ A_{\omega_1} &:= \bigcup_{\zeta<\omega_1} A_\zeta. \end{aligned}$$

It is clear that $A_\zeta \subseteq C$ holds for each σ-algebra C which contains A, so that $A_{\omega_1} \subseteq \sigma(A)$ is inferred.

Let us work on the other inclusion. It is sufficient to show that A_{ω_1} is a σ-algebra. This is so because $A \subseteq A_{\omega_1}$, so that in this case A_{ω_1} would contribute to the intersections defining $\sigma(A)$; hence we could infer $A_{\omega_1} \subseteq \sigma(A)$. We prove the assertion through a series of auxiliary statements, noting that $\langle A_\zeta \mid \zeta < \omega_1\rangle$ forms a chain with respect to set inclusion.

Lemma 1.6.7 *For each $\zeta < \omega_1$, if $a \in A_\zeta$, then $-a \in A_{\zeta+1}$.*

Plan

Proof 0. The proof proceeds by transfinite induction on ζ; it will have to discuss the case that ζ is a limit ordinal and distinguish whether ζ is even or odd.

1. The assertion is true for $\zeta = 0$. Assume for the induction step that it is true for all $\eta < \zeta$.

2. If ζ is a limit ordinal, we know that we can find for $a \in A_\zeta$ an ordinal $\eta < \zeta$ with $a \in A_\eta$, hence by induction hypothesis $-a \in A_{\eta+1} \subseteq A_\zeta$,

because $\eta + 1 < \zeta$ by the definition of a limit ordinal (see Definition 1.4.8 on page 22).

If ζ is even, but not a limit ordinal, we can write ζ as $\zeta = \xi + 1$. Then $A_\zeta = (A_\xi)_\sigma$, and hence $a = \bigvee_{n\in\mathbb{N}} a_n$ for some $a_n \in A_\xi \subseteq A_\zeta$, so that $-a = \bigwedge_{n\in\mathbb{N}}(-a_n) \in (A_\zeta)_\delta = A_{\zeta+1}$. The argumentation for ζ odd is exactly the same. $\dashv$

Thus A_{ω_1} is closed under complementation. Closure under countable infima and suprema is shown similarly, but we have to cater for a countable sequence of countable ordinals.

Lemma 1.6.8 *A_{ω_1} is closed under countable infima and countable suprema.*

The proof rests on the observation that the supremum of a countable number of countable ordinals is countable again (or, putting it differently, that ω_1 is not reachable by countable suprema of countable ordinals). Basic observation

Proof We focus on countable suprema; the proof for infima works exactly in the same way. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of elements in A_{ω_1}; then we find ordinal numbers $\zeta_n < \omega_1$ such that $a_n \in A_{\zeta_n}$. Because ζ_n is countable for each $n \in \mathbb{N}$, we conclude from Proposition 1.4.17 that $\zeta^* := \bigcup_{n\in\mathbb{N}} \zeta_n$ is a countable ordinal, so that $\zeta^* < \omega_1$. Because $\langle A_\zeta \mid \zeta < \omega_1\rangle$ forms a chain, we infer that $a_n \in A_{\zeta^*}$ for all $n \in \mathbb{N}$. Consequently, $\bigvee_{n\in\mathbb{N}} a_n \in (A_{\zeta^*})_\sigma \subseteq A_{\omega_1}$. $\dashv$

Thus we have shown

Proposition 1.6.9 $A_{\omega_1} = \sigma(A)$. $\dashv$

To summarize, we have two possibilities to construct the σ-algebra generated by an algebra A: We can use the closure operation suggested by the σ-operator, or we can go through the explicit construction using transfinite induction. One usually prefers the first way, since it is easier to handle, and, as we will see, information about the context will be easier to be factored in. But sometimes one does not have another choice but going the inductive way. An immediate use of this construction will be the Extension Theorem 1.6.29 for measures.

1.6.2 Factoring Through σ-Ideals

Factoring a Boolean σ-algebra through an ideal works as for general Boolean algebras, resulting in a Boolean algebra again. There is no reason why the factor algebra should be a σ-algebra, however, so if we want to obtain a σ-algebra, we have to make stronger assumptions on the object used for factoring.

Definition 1.6.10 *Let B be a Boolean algebra and $I \subseteq B$ an ideal. I is called a σ-ideal iff $\sup_{n\in\mathbb{N}} a_n \in I$, provided $a_n \in I$ for all $n \in \mathbb{N}$.*

Not every ideal is a σ-ideal: $\{F \subseteq \mathbb{N} \mid F \text{ is finite}\}$ is an ideal but certainly not a σ-ideal in $\mathcal{P}(\mathbb{N})$, even if $\mathcal{P}(\mathbb{N})$ is a Boolean σ-algebra.

The following statement is the σ-variant of Proposition 1.5.42; its proof follows [Aum54, p. 79] quite closely.

Proposition 1.6.11 *Let B be a Boolean σ-algebra and $I \subseteq B$ be a σ-ideal. Then B/I is a Boolean σ-algebra.*

Approach **Proof** 1. Because each Boolean σ-algebra is a Boolean algebra and each σ-ideal is an ideal, we may conclude from Proposition 1.5.42 that B/I is a Boolean algebra. Hence it remains to be shown that this Boolean algebra is closed under countable suprema; since B/I is closed under complementation, closedness under countable infima will follow.

2. Let $a_n \in B$, and then $a := \bigvee_{n\in\mathbb{N}} a_n \in B$. We claim that $[a]_{\sim_I} = \bigvee_{n\in\mathbb{N}} [a_n]_{\sim_I}$. Because $a_n \leq a$ for all $n \in \mathbb{N}$, we conclude that $[a_n]_{\sim_I} \leq [a]_{\sim_I}$ for all $n \in \mathbb{N}$; hence $\bigvee_{n\in\mathbb{N}} [a_n]_{\sim_I} \leq [a]_{\sim_I}$. Now let $[a_n]_{\sim_I} \leq [b]_{\sim_I}$ for all $n \in \mathbb{N}$; then we show that $[a]_{\sim_I} \leq [b]_{\sim_I}$. In fact, because $[a_n]_{\sim_I} \leq [b]_{\sim_I}$, we conclude that $c_n := a_n \ominus (a_n \wedge b) \in I$ and $b \wedge c_n = \bot$ for all $n \in \mathbb{N}$ (since $c_n = a_n \wedge -(a_n \wedge b)$). Thus $a_n = c_n \vee (a_n \wedge c_n)$, so that we have $\bigvee_{n\in\mathbb{N}} a_n = (a \wedge b) \vee \bigvee_{n\in\mathbb{N}} c_n$ by the infinite distributive law from Lemma 1.6.3. This implies $a \ominus (a \wedge b) = \bigvee_{n\in\mathbb{N}} c_n \in I$, or, equivalently, $[a]_{\sim_I} \leq [b]_{\sim_I}$. Consequently, $[a]_{\sim_I}$ is the smallest upper bound to $\{[a_n]_{\sim_I} \mid n \in \mathbb{N}\}$. $\dashv$

This construction will be put to use later on when we want to identify sets which differ only by a set of measure zero. Then it will turn out that this equivalence relation on subsets is based on a σ-ideal. But before we can do that, we have to know what a measure is. This is what we will discuss next.

1.6.3 Measures

Boolean σ-algebras model events. The top element $\top$ is interpreted as an event which can happen unconditionally and always; the bottom element $\perp$ is the impossible event. The complement of an event is an event, and if we have a countable sequence of events, then their supremum is an event, viz., the event that at least one of the events in the sequence happens.

To illustrate, suppose that we have a set T of traders which may form unions or coalitions; then T and $\emptyset$ are coalitions; if A is a coalition, then $T \setminus A$ is a coalition as well, and if A_n is a coalition for each $n \in \mathbb{N}$, then we want to be able to form the "big" coalition $\bigcup_{n\in\mathbb{N}} A_n$. Hence the set of all coalitions forms a σ-algebra.

We deal in the sequel with set-based σ-algebras, so we fix a set S of events.

Definition 1.6.12 *Let $\mathcal{C} \subseteq \mathcal{P}(S)$ be a family of sets with $\emptyset \in \mathcal{C}$. A map $\mu : \mathcal{C} \to [0, \infty]$ with $\mu(\emptyset) = 0$ is called*

1. monotone *iff $\mu(A) \leq \mu(B)$ for $A, B \in \mathcal{C}, A \subseteq B$,*
2. additive *iff $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \in \mathcal{C}$, with $A \cup B \in \mathcal{C}$ and $A \cap B = \emptyset$,*
3. countably subadditive *iff $\mu(\bigcup_{n\in\mathbb{N}} A_n) \leq \sum_{n\in\mathbb{N}} \mu(A_n)$, whenever $(A_n)_{n\in\mathbb{N}}$ is a sequence of sets in $\mathcal{C}$ with $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{C}$,*
4. countably additive *iff $\mu(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$, whenever $(A_n)_{n\in\mathbb{N}}$ is a mutually disjoint sequence of sets in $\mathcal{C}$ with $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{C}$,*

If $\mathcal{C}$ is a σ-algebra, then a map $\mu : \mathcal{C} \to [0, \infty]$ with $\mu(\emptyset) = 0$ is called a measure *iff μ is monotone and countably additive. If S can be written as $S = \bigcup_{n\in\mathbb{N}} S_n$ with $S_n \in \mathcal{C}$ and $\mu(S_n) < \infty$ for all $n \in \mathbb{N}$, then the measure is called σ-*finite.

We permit that μ assumes the value $+\infty$. Clearly, a countably additive set function is additive, and it is countably subadditive, provided it is monotone.

Example 1.6.13 Let S be a set, and define for $a \in S$, $A \subseteq S$

$$\delta_a(A) := \begin{cases} 1, & \text{if } a \in A \\ 0, & \text{otherwise.} \end{cases}$$

Dirac measure

Then δ_a is a measure on the power set of S. It is usually referred to as the *Dirac measure* on a. ✌

A slightly more complicated example indicates the connection to ultrafilters.

Example 1.6.14 Let $\mu : \mathcal{P}(S) \to \{0, 1\}$ be a binary-valued measure. Define

$$\mathcal{F} := \{A \subseteq S \mid \mu(A) = 1\}.$$

Then $\mathcal{F}$ is an ultrafilter on $\mathcal{P}(S)$. First, we check that $\mathcal{F}$ is a filter: $\emptyset \notin \mathcal{F}$ is obvious, and if $A \in \mathcal{F}$ with $A \subseteq B$, then certainly $B \in \mathcal{F}$. Let $A, B \in \mathcal{F}$, then $2 = \mu(A) + \mu(B) = \mu(A \cup B) + \mu(A \cap B)$, hence $\mu(A \cap B) = 1$, and thus $A \cap B \in \mathcal{F}$. Thus $\mathcal{F}$ is indeed a filter. It is also an ultrafilter by Lemma 1.5.21, because $A \notin \mathcal{F}$ implies $S \setminus A \in \mathcal{F}$.

The converse construction, viz., to generate a binary-valued measure from a filter, would require $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$ if and only if there exists $n \in \mathbb{N}$ with $A_n \in \mathcal{F}$ for any disjoint family $(A_n)_{n\in\mathbb{N}}$. This, however, leads to very deep questions on set theory; see [Jec06, Chap. 10] for a discussion. ✌

Let us have a look at an important example.

Example 1.6.15 Let $\mathcal{C} := \big\{]a, b] \mid a, b \in [0, 1]\big\}$ be all left open, right closed intervals of the unit interval. Put $\ell(]a, b]) := b - a$; hence $\ell(I)$ is the length of interval I. Note that $\ell(\emptyset) = \ell(]a, a]) = 0$. Certainly $\ell : \mathcal{C} \to \mathbb{R}_+$ is monotone and additive.

1. If $\bigcup_{i=1}^{k}]a_i, b_i] \subseteq]a, b]$ and the intervals are disjoint, then $\sum_{i=1}^{k} \ell(]a_i, b_i]) \leq \ell(]a, b])$. The proof proceeds by induction on the number k of intervals. For the induction step, we have mutually disjoint intervals with $\bigcup_{i=1}^{k+1}]a_k, b_k] \subseteq]a, b]$. Renumbering, if necessary, we may assume that $a_1 \leq b_1 \leq a_2 \leq b_2 \leq \ldots \leq a_k \leq b_k \leq a_{k+1} \leq b_{k+1}$. Then $\sum_{i=1}^{k} \ell(]a_i, b_i]) + \ell(]a_{k+1}, b_{k+1}]) \leq \ell(]a_1, b_k]) + \ell(]a_{k+1}, b_{k+1}]) \leq \ell(]a, b])$, because ℓ is monotone and additive.

2. If $\bigcup_{i=1}^{\infty}]a_i, b_i] \subseteq]a, b]$ and the intervals are disjoint, then $\sum_{i=1}^{k} \ell(]a_i, b_i]) \leq \ell(]a, b])$ for all k; hence
$$\sum_{i=1}^{\infty} \ell(]a_i, b_i]) = \sup_{k \in \mathbb{N}} \sum_{i=1}^{k} \ell(]a_i, b_i]) \leq \ell(]a, b]).$$

3. If $]a, b] \subseteq \bigcup_{i=1}^{k}]a_i, b_k]$, then $\ell(]a, b]) \leq \sum_{i=1}^{k} \ell(]a_i, b_i])$ with no necessarily disjoint intervals. This is established by induction on k. If $k = 1$, the assertion is obvious. The induction step proceeds as follows: Assume that $]a, b] \subseteq \bigcup_{i=1}^{k+1}]a_i, b_k]$. By renumbering, if necessary, we can assume that $a_{k+1} < b \leq b_{k+1}$. If $a_{k+1} \leq a$, the assertion follows, so let us assume that $a < a_{k+1}$. Then $]a, a_{k+1}] \subseteq \bigcup_{i=1}^{k}]a_i, b_i]$, so that by the induction hypothesis $a_{k+1} - a = \ell(]a, a_{k+1}]) \geq \sum_{i=1}^{k} \ell(]a_i, b_i])$. Thus
$$\ell(]b, a]) = b - a \leq (a_{k+1} - a) + (b_{k+1} - a_{k+1}) \leq \sum_{i=1}^{k+1} \ell(]a_i, b_i]).$$

4. Now assume that $]a, b] \subseteq \bigcup_{i=1}^{\infty}]a_i, b_i]$. This is a little bit more complicated since we do not know whether the interval $]a, b]$ is covered already by a finite number of intervals, so we have to resort to a little trick. The interval $[a+\epsilon, b]$ is closed and bounded, hence compact, for every fixed $\epsilon > 0$; we also know that for each $i \in \mathbb{N}$ the semi-open interval $]a_i, b_i]$ is contained in the open interval $]a_i, b_i + \epsilon/2^i[$, so that we have
$$[a + \epsilon, b] \subseteq \bigcup_{i=1}^{\infty}]a_i, b_i + \epsilon/2^i[.$$
By the Heine–Borel Theorem 1.5.46, we can find a finite subset of these intervals which cover $[a + \epsilon, b]$, say $[a + \epsilon, b] \subseteq \bigcup_{i \in K}]a_i, b_i + \epsilon/2^i[$, with $K \subseteq \mathbb{N}$ finite. Hence
$$]a + \epsilon, b] \subseteq \bigcup_{i \in K}]a_i, b_i + \epsilon/2^i],$$
and we conclude from the finite case that
$$\begin{aligned} b - (a + \epsilon) &= \ell([a + \epsilon, b]) \leq \sum_{i \in K} \ell(]a_i, b_i + \epsilon/2^i]) \\ &= \sum_{i=1}^{\infty} (b_i + \epsilon/2^i - a_i) < \sum_{i=1}^{\infty} \ell(]a_i, b_i]) + \epsilon. \end{aligned}$$

Since $\epsilon > 0$ was arbitrary, we have established

$$\ell(]a, b]) \leq \sum_{i=1}^{\infty} \ell(]a_i, b_i]).$$

✌

Thus we have shown

Proposition 1.6.16 *Let $\mathcal{C}$ be the set of all left open, right closed intervals of the unit interval, and denote by $\ell(]a, b]) := b - a$ the length of interval $]a, b] \in \mathcal{C}$. Then $\ell : \mathcal{C} \to \mathbb{R}_+$ is monotone and countably additive.* $\dashv$

When having a look at $\mathcal{C}$, we note that this family is not closed under complementation, but the complement of a set in $\mathcal{C}$ can be represented through elements of $\mathcal{C}$, e.g., $]0, 1] \setminus]1/3, 1/2] =]0, 1/3] \cup]1/2, 1]$. This is captured through the following definition:

Definition 1.6.17 *$\mathcal{R} \subseteq \mathcal{P}(S)$ is called a* semiring *iff*

1. *$\emptyset \in \mathcal{R}$,*
2. *$\mathcal{R}$ is closed under finite intersections,*
3. *If $B \in \mathcal{R}$, then there exists a finite family of mutually disjoint sets $C_1, \ldots, C_k \in \mathcal{R}$ with $S \setminus B = C_1 \cup \ldots \cup C_k$.*

Thus the complement of a set in $\mathcal{R}$ can be represented through a finite disjoint union of elements of $\mathcal{R}$.

Lebesgue measure

We want to extend $\ell : \mathcal{C} \to \mathbb{R}_+$ from the semiring of left open, right closed intervals to a measure λ on the σ-algebra $\sigma(\mathcal{C})$. This measure is fairly Important; it is called the *Lebesgue measure* on the unit interval.

A first step toward an extension of ℓ to the σ-algebra generated by the intervals is the extension to the algebra generated by them. This can be accomplished easily once this algebra has been identified.

Lemma 1.6.18 *Let $\mathcal{C}$ be the set of all left open, right closed intervals in* $]0, 1]$. *Then the algebra generated by $\mathcal{C}$ consists of all disjoint unions of elements of $\mathcal{C}$.*

Proof Denote by

$$\mathcal{D} := \{\bigcup_{i=1}^{n}]a_i, b_i] \mid n \in \mathbb{N}, a_1 \leq b_1 \leq a_2 \leq b_2 \ldots \leq a_n \leq b_n\}.$$

Then all elements of $\mathcal{D}$ are certainly contained in the algebra generated by $\mathcal{C}$. If we can show that $\mathcal{D}$ is an algebra itself, we are done, because then $\mathcal{D}$ is the smallest algebra containing $\mathcal{C}$.

$\mathcal{D}$ is certainly closed under finite unions and finite intersections, and $\emptyset \in \mathcal{D}$. Then

$$]0, 1] \setminus \bigcup_{i=1}^{n}]a_i, b_i] =]0, a_1] \cup]b_1, a_2] \cup \ldots \cup]b_n, 1],$$

which is a member of $\mathcal{D}$ as well. Thus $\mathcal{D}$ is also closed under complementation and hence is an algebra. $\dashv$

This permits us to extend ℓ to the algebra generated by the Intervals.

Corollary 1.6.19 *ℓ extends uniquely to the algebra generated by $\mathcal{C}$ such that the extension is monotone and countably additive.*

Proof Put

$$\ell\big(\bigcup_{i=1}^{n}]a_i, b_i]\big) := \sum_{i=1}^{n} \ell(]a_i, b_i]),$$

whenever $]a_i, b_i] \in \mathcal{C}$. This is well defined. Assume

$$\bigcup_{i=1}^{n}]a_i, b_i] = \bigcup_{j=1}^{m}]c_j, d_j];$$

then $]a_i, b_i]$ can be represented as a disjoint union of those intervals $]c_j, d_j]$ which it contains, so that we have

$$\begin{aligned}
\sum_{i=1}^{n} \ell(]a_i, b_i]) &= \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(]a_i, b_i] \cap]c_j, d_j]) \\
&= \sum_{j=1}^{m} \sum_{i=1}^{m} \ell(]c_j, d_j] \cap]a_i, b_i]) \\
&= \sum_{j=1}^{m} \ell(]c_j, d_j]).
\end{aligned}$$

We may conclude from Example 1.6.15 that ℓ is countably additive on the algebra. $\dashv$

For the sake of illustration, let us assume that we have Lebesgue measure constructed already, and let us compute $\lambda(C)$ where C is the Cantor ternary set constructed in Example 1.6.4 on page 66. The construction of the ternary set is done through sets C_n, each of which is the union of 2^n mutually disjoint intervals of length 3^{-n}. If I is an interval of length 3^{-n}, we know that $\lambda(I) = 3^{-n}$, so that $\lambda(C_n) = (2/3)^n$. We also know that $C_1 \supseteq C_2 \supseteq \ldots$, so that we have a descending chain of sets with $C = \bigcap_{n\in\mathbb{N}} C_n$ and $\inf_{n\in\mathbb{N}} \lambda(C_n) = 0$.

In order to compute $\lambda(C)$, we need so know something about the behavior of measures when monotone limits of sets are encountered.

Lemma 1.6.20 *Let $\mu : \mathcal{A} \to [0, \infty]$ be a measure on the σ-algebra $\mathcal{A}$.*

1. *If $A_n \in \mathcal{A}$ is a monotone increasing sequence of sets in $\mathcal{A}$ and $A = \bigcup_{n\in\mathbb{N}} A_n$, then $\mu(A) = \sup_{n\in\mathbb{N}} \mu(A_n)$.*
2. *If $A_n \in \mathcal{A}$ is a monotone decreasing sequence of sets in $\mathcal{A}$ and $A = \bigcap_{n\in\mathbb{N}} A_n$, then $\mu(A) = \inf_{n\in\mathbb{N}} \mu(A_n)$, provided $\mu(A_k) < \infty$ for some $k \in \mathbb{N}$.*

Proof 1. We can write $A_n = \bigcup_{i=1}^{n} B_i$ with $B_1 := A_1$ and $B_i := A_i \setminus A_{i-1}$. Because the A_n form an increasing sequence, the B_n are mutually disjoint. Assume without loss of generality that $\mu(A_n) < \infty$ for all $n \in \mathbb{N}$ (otherwise the assertion is trivial); then by countable additivity and through telescoping

$$\begin{aligned} \mu(A) &= \sum_{i=1}^{\infty} \mu(B_i) = \mu(A_1) + \sum_{i=1}^{\infty} \big(\mu(A_{i+1}) - \mu(A_i)\big) \\ &= \lim_{n\to\infty} \mu(A_n) = \sup_{n\in\mathbb{N}} \mu(A_n). \end{aligned}$$

2. Assume $\mu(A_1) < \infty$, and then the sequence $A_1 \setminus A_n$ is increasing toward $A_1 \setminus A$; hence

$$\mu(A) = \mu(A_1) - \mu(A_1 \setminus A) = \mu(A_1) - \sup_{n\in\mathbb{N}} \mu(A_1 \setminus A_n) = \inf_{n\in\mathbb{N}} \mu(A_n).$$

$\dashv$

Ok, so let us return to the discussion of Cantor's set. We know that $\lambda(C_n) = (2/3)^n$ and that $C_1 \supseteq C_2 \supseteq C_3 \ldots$, so we conclude

$$\lambda(C) = \inf_{n \in \mathbb{N}} \lambda(C_n) = 0.$$

We have identified a geometrically fairly complicated set which has measure zero. This set is not easy to visualize, since it does not contain an interval of positive length.

Now fix a semiring $\mathcal{C} \subseteq \mathcal{P}(S)$ and $\mu : \mathcal{C} \to [0, \infty]$ with $\mu(\emptyset) = 0$, which is monotone and countably subadditive. We will first compute an outer approximation for each subset of S by elements of $\mathcal{C}$. But since the subsets of S may be as a whole somewhat inaccessible and since $\mathcal{C}$ may be somewhat small, we try to cover the subsets of S by countable unions of elements of $\mathcal{C}$ and take the best approximation we can, i.e., we take the infimum. Define

$$\mu^*(A) := \inf\left\{\sum_{n \in \mathbb{N}} \mu(C_n) \mid A \subseteq \bigcup_{n \in \mathbb{N}} C_n, C_n \in \mathcal{C}\right\}$$

for $A \subseteq S$. This is the *outer measure* of A associated with μ.

These are some interesting (for us, that is) properties of μ^*.

Lemma 1.6.21 $\mu^* : \mathcal{P}(S) \to [0, \infty]$ *is monotone and countably subadditive,* $\mu^*(\emptyset) = 0$. *If* $A \in \mathcal{C}$, *then* $\mu^*(A) = \mu(A)$.

Proof 1. Let $(A_n)_{n \in \mathbb{N}}$ be a sequence of subset of S; put $A := \bigcup_{n \in \mathbb{N}} A_n$. If $\sum_{n \in \mathbb{N}} \mu^*(A_n) < \infty$, fixing n, we find a cover $\{C_{n,m} \mid m \in \mathbb{N}\} \subseteq \mathcal{C}$ for A_n with

$$\mu(A_n) \leq \sum_{m \in \mathbb{N}} \mu(C_{n,m}) \leq \mu^*(A_n) + \epsilon/2^n, and$$

thus $\{C_{n,m} \mid n, m \in \mathbb{N}\} \subseteq \mathcal{C}$ is a cover of A with

$$\mu(A) \leq \sum_{n,m \in \mathbb{N}} \mu(C_{n,m}) \leq \sum_{n \in \mathbb{N}} \mu(A_n) + \epsilon.$$

Since $\epsilon > 0$ was arbitrary, we conclude $\mu^*(A) \leq \sum_{n \in \mathbb{N}} \mu^*(A_n)$. If, however, $\sum_{n \in \mathbb{N}} \mu^*(A_n) = \infty$, the assertion is immediate.

2. The other properties are readily seen. $\dashv$

The next step is somewhat mysterious—it has been suggested by Carathéodory around 1914 for the construction of a measure extension. It splits a set $A = (A \cap X) \cup (A \cap S \setminus X)$ along an arbitrary other set X, and look what happens to the outer measure. If $\mu^*(A) = \mu^*(A \cap X) + \mu^*(A \cap S \setminus X)$, then A is considered well behaved. Those sets which are well behaved no matter what set X we use for splitting are considered next.

$\mathcal{C}_\mu$

Definition 1.6.22 *A set $A \subseteq S$ is called μ-*measurable *iff $\mu^*(X) = \mu^*(X \cap A) + \mu^*(X \cap S \setminus A)$ holds for all $X \subseteq S$. The set of all μ-measurable sets is denoted by $\mathcal{C}_\mu$.*

So take a μ-measurable set A and an arbitrary subset $X \subseteq S$; then X splits into a part $X \cap A$ which belongs to A and another one $X \cap S \setminus A$ outside of A. Measuring these pieces through μ^*, we demand that they add up to $\mu^*(X)$ again.

These properties are immediate.

Lemma 1.6.23 *The outer measure has these properties:*

1. $\mu^*(\emptyset) = 0$.
2. $\mu^*(A) \geq 0$ *for all* $A \subseteq S$.
3. μ^* *is monotone.*
4. μ^* *is countably subadditive.*

Proof We establish only the last property. Here we have to show that $\mu^*(\bigcup_{n\in\mathbb{N}} A_n) \leq \sum_{n\in\mathbb{N}} \mu^*(A_n)$. We may and do assume that all $\mu^*(A_n)$ are finite. Given $\epsilon > 0$, we find for each $n \in \mathbb{N}$ a sequence $B_{n,k} \in \mathcal{C}$ for A_n such that $A_n \subseteq \bigcup_{k\in\mathbb{N}} B_{n,k}$ and $\sum_{k\in\mathbb{N}} \mu(B_{n,k}) \leq \mu^*(A_n) + \epsilon/2^n$. Thus

$$\sum_{n,k\in\mathbb{N}} \mu(B_{n,k}) \leq \sum_{n\in\mathbb{N}} (\mu^*(A_n) + \epsilon/2^n) < \sum_{n\in\mathbb{N}} \mu^*(A_n) + \epsilon,$$

which implies $\sum_{n\in\mathbb{N}} \mu^*(A_n) \leq \mu^*(\bigcup_{n\in\mathbb{N}} A_n)$, because $\bigcup_{n\in\mathbb{N}} A_n \subseteq \bigcup_{n,k\in\mathbb{N}} B_{n,k}$ and because $\epsilon > 0$ was arbitrary. $\dashv$

Because of countably subadditivity, we conclude

Corollary 1.6.24 *$A \in \mathcal{C}_\mu$ iff $\mu^*(X \cap A) + \mu^*(X \cap S \setminus A) \leq \mu^*(X)$ for all $X \subseteq S$.* $\dashv$

Let us have a look at the set of all μ-measurable sets. It turns out that the originally given sets are all μ-measurable and that $\mathcal{C}_\mu$ is an algebra.

Proposition 1.6.25 *$\mathcal{C}_\mu$ is an algebra. Also if μ is additive, $\mathcal{C} \subseteq \mathcal{C}_\mu$ and $\mu(A) = \mu^*(A)$ for all $A \in \mathcal{C}$.*

Proof 1. $\mathcal{C}_\mu$ is closed under complementation; this is obvious from its definition, and $S \in \mathcal{C}_\mu$ is also clear. So we have only to show that $\mathcal{C}_\mu$ is closed under finite intersections. For simplicity, denote complementation by $\cdot^c$.

Now let $A, B \in \mathcal{C}_\mu$; we want to show

$$\mu^*(X) \geq \mu^*((A \cap B) \cap X) + \mu^*((A \cap B)^c \cap X),$$

for each $X \subseteq S$; from Corollary 1.6.24, we infer that this implies $A \cap B \in \mathcal{C}_\mu$. Since $B \in \mathcal{C}_\mu$ and then $A \in \mathcal{C}_\mu$, we know

$$\begin{aligned}
\mu^*(X) &= \mu^*(X \cap B) + \mu^*(X \cap B^c)\\
&= \mu^*\big(X \cap (A \cap B)\big) + \mu^*\big(X \cap (A^c \cap B)\big)\\
&\quad + \mu^*\big(X \cap (A \cap B^c)\big) + \mu^*\big(X \cap (A^c \cap B^c)\big)\\
&\geq \mu^*\big(X \cap (A \cap B)\big) + \mu^*\big(X \cap ((A^c \cap B) \cup (A \cap B^c)\\
&\quad \cup (A^c \cap B^c))\big)\\
&\overset{(\ddagger)}{=} \mu^*\big(X \cap (A \cap B)\big) + \mu^*\big(X \cap (A \cap B)^c\big).
\end{aligned}$$

Equality ($\ddagger$) uses

$$\begin{aligned}
(A^c \cap B) \cup (A \cap B^c) \cup (A^c \cap B^c) &= \big(A^c \cap (B \cup B^c)\big) \cup (A \cap B^c)\\
&= A^c \cup (A \cap B^c)\\
&= A^c \cup B^c.
\end{aligned}$$

Hence we see that $A \cap B$ satisfies the defining inequality.

2. We still have to show that $\mathcal{C} \subseteq \mathcal{C}_\mu$ and that μ^* extends μ. Let $A \in \mathcal{C}$, and then $S \setminus A = D_1 \cup \ldots \cup D_k$ for some mutually disjoint $D_1, \ldots, D_k \in \mathcal{C}$, because $\mathcal{C}$ is a semiring. Fix $X \subseteq S$, and assume that $\mu^*(S) < \infty$ (otherwise, the assertion is trivial). Given $\epsilon > 0$, there exists in $\mathcal{C}$ a cover $(A_n)_{n \in \mathbb{N}}$ of X with $\mu^*(X) < \sum_{n \in \mathbb{N}} \mu(A_n) + \epsilon$. Now put $B_n := A \cap A_n$ and $C_{i,n} := A_n \cap D_i$. Then $X \cap A \subseteq \bigcup_{n \in \mathbb{N}} B_n$ with

$B_n \in \mathcal{C}$ and $X \cap A^c \subseteq \bigcup_{n\in\mathbb{N},1\leq i\leq k} C_{i,n}$ with $C_{i,n} \in \mathcal{C}$. Hence

$$\begin{aligned}\mu^*(X \cap A) + \mu^*(X \cap A^C) &\leq \sum_{n\in\mathbb{N}} \mu(B_n) + \sum_{n\in\mathbb{N},1\leq i\leq k} \mu(C_{i,n}) \\ &\leq \sum_{n\in\mathbb{N}} \mu(A_n) \\ &< \mu^*(X) - \epsilon,\end{aligned}$$

because μ is (finitely) additive. Hence $A \in \mathcal{C}_\mu$. μ^* is an extension to μ by Lemma 1.6.21. $\dashv$

But we can in fact say more on the behavior of μ^* on $\mathcal{C}_\mu$: It turns out to be additive on the splitting parts.

Lemma 1.6.26 *Let $\mathcal{D} \subseteq \mathcal{C}_\mu$ be a finite or infinite family of mutually disjoint sets in $\mathcal{C}_\mu$; then*

$$\mu^*(X \cap \bigcup_{D\in\mathcal{D}} D) = \sum_{D\in\mathcal{D}} \mu^*(X \cap D)$$

holds for all $X \subseteq S$.

Plan

Proof 1. The proof goes like this: We establish the equality above for finite $\mathcal{D}$, say, $\mathcal{D} = \{A_1, \ldots, A_n\}$ with $A_n \in \mathcal{C}_\mu$ for $1 \leq j \leq n$. From this we obtain the equality for the countable case as well, because then

$$\mu^*(X \cap \bigcup_{i=1}^{\infty} A_i) \geq \mu^*(X \cap \bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu^*(X \cap A_i),$$

for all $n \in \mathbb{N}$, so that $\mu^*(X \cap \bigcup_{i=1}^{\infty} A_i) \geq \sum_{i=1}^{\infty} \mu^*(X \cap A_i)$, which together with countable subadditivity gives the desired result.

2. The proof for $\mu^*(X \cap \bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu^*(X \cap A_i)$ proceeds by induction on n, starting with $n = 2$. If $A_1 \cup A_2 = S$, this is just the definition that A_1 (or A_2) is μ-measurable, so the equality holds. If $A_1 \cup A_2 \neq S$, we note that

$$\mu^*(X) = \mu^*\big((X \cap (A_1 \cup A_2)) \cap A_1\big) + \mu^*\big((X \cap (A_1 \cup A_2)) \cap S \setminus A_1\big).$$

Evaluating the pieces, we see that

$$\begin{aligned}\big(X \cap (A_1 \cup A_2)\big) \cap A_1 &= X \cap A_1, \\ \big(X \cap (A_1 \cup A_2)\big) \cap S \setminus A_1 &= X \cap A_2,\end{aligned}$$

because $A_1 \cap A_2 = \emptyset$. The induction step is straightforward:

$$\begin{aligned}\mu^*(X \cap \bigcup_{i=1}^{n+1} A_i) &= \mu^*((X \cap \bigcup_{i=1}^{n} A_n) \cup (X \cap A_{n+1})) \\ &= \sum_{i=1}^{n} \mu^*(X \cap A_i) + \mu^*(X \cap A_{n+1}) \\ &= \sum_{i=1}^{n+1} \mu^*(X \cap A_i).\end{aligned}$$

⊣

We can relax the condition on a set being a member of $\mathcal{C}_\mu$ if we know that the domain $\mathcal{C}$ from which we started is an algebra and that μ is additive on $\mathcal{C}$. Then we do not have to test whether a μ-measurable set splits all the subsets of S, but it is rather sufficient that A splits S, to be specific

Proposition 1.6.27 *Let $\mathcal{C}$ be an algebra and $\mu : \mathcal{C} \to [0, +\infty]$ be additive. Then $A \in \mathcal{C}_\mu$ iff $\mu^*(A) + \mu^*(S \setminus A) = \mu^*(S)$.*

Proof This is a somewhat lengthy and laborious computation similarly to the one above; see [Bog07, 1.11.7, 1.11.8]. ⊣

Returning to the general discussion, we have

Proposition 1.6.28 *$\mathcal{C}_\mu$ is a σ-algebra, and μ^* is countably additive on $\mathcal{C}_\mu$.*

Proof 0. Let $(A_n)_{n\in\mathbb{N}}$ be a countable family of mutually disjoint sets in $\mathcal{C}_\mu$, then we have to show that $A := \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{C}_\mu$, and thus we have to show that

$$\mu^*(X \cap A) + \mu^*(X \cap A^c) \leq \mu^*(X)$$

for each $X \subseteq S$ (here $\cdot^c$ is complementation again). Fix X.

1. We know that $\mathcal{C}_\mu$ is closed under finite unions, so we have for each $n \in \mathbb{N}$

$$\begin{aligned}\mu^*(X) &\geq \sum_{i=1}^{n} \mu^*(X \cap A_i) + \mu^*(X \cap \bigcap_{i=1}^{n} A_i^c) \\ &\geq \sum_{i=1}^{n} \mu^*(X \cap A_i) + \mu^*(X \cap A),\end{aligned}$$

because $\bigcap_{i=1}^{n} A_i^c \supseteq A^c$. Letting $n \to \infty$ we obtain the desired inequality.

3. Thus $\mathcal{C}_\mu$ is closed under disjoint countable unions. Using the first entrance trick (Exercise 1.37) and the observation that $\mathcal{C}_\mu$ is an algebra by Proposition 1.6.25, we convert each countable union into a countable union of mutually disjoint sets, so we have shown that $\mathcal{C}_\mu$ is a σ-algebra. Countable additivity of μ^* on $\mathcal{C}_\mu$ follows from Lemma 1.6.26 when putting $X := S$. $\dashv$

Summarizing, we have demonstrated this Extension Theorem.

Extension Theorem

Theorem 1.6.29 *Let $\mathcal{C}$ be an algebra over a set S and $\mu : \mathcal{C} \to [0, \infty]$ monotone and countably additive.*

1. *There exists an extension of μ to a measure on the σ-algebra $\sigma(\mathcal{C})$ generated by $\mathcal{C}$.*

2. *If μ is σ-*finite*, the extension is uniquely determined.*

Steps in the proof

Proof 0. For establishing the existence of an extension, we simply collect the results obtained so far. For a finite measure, we obtain uniqueness by following the construction of $\sigma(\mathcal{C})$ through transfinite induction, as outlined in Sect. 1.6.1; note that finiteness is necessary here because we are bound by Lemma 1.6.20 to finite measures for establishing the limit of a decreasing sequence. Finally, we localize the measure in the σ-finite case to the countably many finite pieces which make up the entire space.

1. Proposition 1.6.28 shows that $\mathcal{C}_\mu$ is a σ-algebra containing $\mathcal{C}$ and that μ^* is a measure on $\mathcal{C}_\mu$. Hence $\sigma(\mathcal{C}) \subseteq \mathcal{C}_\mu$, and we can restrict μ^* to $\sigma(\mathcal{C})$. Denote this restriction also by μ; then μ is a measure on $\sigma(\mathcal{C})$.

2. In order to establish uniqueness, assume first that $\mu(S) < \infty$. Let ν be a measure which extends μ to $\sigma(\mathcal{C})$. Recall the construction of $\sigma(\mathcal{C})$ through transfinite induction on page 70. We claim that

$$\mu(A) = \nu(A) \text{ for all } A \in \mathcal{C}_\zeta$$

holds for all ordinals $\zeta < \omega_1$. Because $\mathcal{C}$ is an algebra, it is easy to see that for odd ordinals ζ a set $A \in (\mathcal{C}_\zeta)_\delta$ iff there exists a decreasing sequence $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{C}_\zeta$ with $A = \bigcap_{n\in\mathbb{N}} A_n$; similarly, each element of $(\mathcal{C}_\zeta)_\sigma$ can be represented as the union of an increasing sequence of elements of $\mathcal{A}_\xi$ if ζ is even. Assume for the induction step that ζ is odd, and let $A \in \mathcal{C}_{\zeta+1}$; thus $A = \bigcap_{n\in\mathbb{N}} A_n$ with $A_1 \supseteq A_2 \supseteq \ldots$ and $A_n \in \mathcal{C}_\zeta$. Hence by Lemma 1.6.20

$$\mu(A) = \mu(\bigcap_{n\in\mathbb{N}} A_n) = \inf_{n\in\mathbb{N}} \mu(A_n) = \inf_{n\in\mathbb{N}} \nu(A_n) = \nu(A).$$

Thus μ and ν coincide on $\mathcal{C}_{\zeta+1}$, if ζ is odd. One argues similarly, but with a monotone increasing sequence in the case that ζ is even. If μ and ν coincide on all $\mathcal{C}_\eta$ for all η with $\eta < \zeta$ for a limit number ζ, then it is clear that they also coincide on $\mathcal{C}_\zeta$ as well.

3. Assume that $\mu(S) = \infty$, but that there exists a sequence $(S_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ with $\mu(S_n) < \infty$ and $S = \bigcup_{n\in\mathbb{N}} S_n$. Because $\mu(S_1 \cup \ldots \cup S_n) \leq \mu(S_1) + \ldots + \mu(S_n) < \infty$, we may and do assume that the sequence is monotonically increasing. Let $\mu_n(A) := \mu(A \cap S_n)$ be the localization of μ to S_n. μ_n has a unique extension to $\sigma(\mathcal{C})$, and since we have $\mu(A) = \sup_{n\in\mathbb{N}} \mu_n(A)$ for all $A \in \sigma(\mathcal{C})$, the assertion follows. $\dashv$

But we are not quite done yet, witnessed by a glance at Lebesgue measure. There we started from the semiring of intervals, but our uniqueness theorem states only what happens when we carry out our extension process starting from an algebra.

It turns out to be most convenient to have a closer look at the construction of σ-algebras when the family of sets we start from has already some structure. This gives the occasion to introduce Dynkin's π-λ-Theorem. This is a very important tool, which sometimes simplifies the identification of the σ-algebra generated from some family of sets.

Theorem 1.6.30 ***(π-λ-Theorem)*** *Let $\mathcal{P}$ be a family of subsets of S that is closed under finite intersections (this is called a π-*class*). Then $\sigma(\mathcal{P})$ is the smallest λ-class containing $\mathcal{P}$, where a family $\mathcal{L}$ of subsets of S is called a λ-*class *iff it is closed under complements and countable disjoint unions.* π-λ-Theorem

Proof 1. Let $\mathcal{L}$ be the smallest λ-class containing P; then we show that $\mathcal{L}$ is a σ-algebra.

2. We show first that it is an algebra. Being a λ-class, $\mathcal{L}$ is closed under complementation. Let $A \subseteq S$; then $\mathcal{L}_A := \{B \subseteq S \mid A \cap B \in \mathcal{L}\}$ is a

λ-class again: If $A \cap B \in \mathcal{L}$, then

$$A \cap (S \setminus B) = A \setminus B = S \setminus ((A \cap B) \cup (S \setminus A)),$$

which is in $\mathcal{L}$, since $(A \cap B) \cap S \setminus A = \emptyset$ and since $\mathcal{L}$ is closed under disjoint unions.

If $A \in \mathcal{P}$, then $\mathcal{P} \subseteq \mathcal{L}_A$, because $\mathcal{P}$ is closed under intersections. Because $\mathcal{L}_A$ is a λ-system, this implies $\mathcal{L} \subseteq \mathcal{L}_A$ for all $A \in \mathcal{P}$. Now take $B \in \mathcal{L}$, then the preceding argument shows that $\mathcal{P} \subseteq \mathcal{L}_B$, and again we may conclude that $\mathcal{L} \subseteq \mathcal{L}_B$. Thus we have shown that $A \cap B \in \mathcal{L}$, provided $A, B \in \mathcal{L}$, so that $\mathcal{L}$ is closed under finite intersections. Thus $\mathcal{L}$ is a Boolean algebra.

3. $\mathcal{L}$ is a σ-algebra as well. It is enough to show that $\mathcal{L}$ is closed under countable unions. But since

$$\bigcup_{n \in \mathbb{N}} A_n = \bigcup_{n \in \mathbb{N}} \left(A_n \setminus \bigcup_{i=1}^{n-1} A_i \right),$$

this follows immediately. $\dashv$

Principle of good sets

Consider an immediate and fairly typical application. It states that two finite measures are equal on a σ-algebra, provided they are equal on a generator which is closed under finite intersections. The proof technique called the *principle of good sets* in [Els99] is worth noting: We collect all sets for which the assertion holds into one family of sets and investigate its properties, starting from an originally given set. If we find that the family has the desired property, then we look at the corresponding closure. With this in mind, we have a look at the proof of the following statement:

Lemma 1.6.31 *Let μ, ν be finite measures on a σ-algebra $\sigma(\mathcal{B})$, where $\mathcal{B}$ is a family of sets which is closed under finite intersections. Then $\mu(A) = \nu(A)$ for all $A \in \sigma(\mathcal{B})$, provided $\mu(B) = \nu(B)$ for all $B \in \mathcal{B}$.*

Proof We investigate the family of all sets for which the assertion is true. Put

$$\mathcal{G} := \{A \in \sigma(\mathcal{B}) \mid \mu(A) = \nu(A)\};$$

then $\mathcal{G}$ has these properties:

- $\mathcal{B} \subseteq \mathcal{G}$ by assumption.
- Since $\mathcal{B}$ is closed under finite intersections, $S \in \mathcal{B} \subseteq \mathcal{G}$.

- $\mathcal{G}$ is closed under complements.
- $\mathcal{G}$ is closed under countable disjoint unions; in fact, let $(A_n)_{n\in\mathbb{N}}$ be a sequence of mutually disjoint sets in $\mathcal{G}$ and $A := \bigcup_{n\in\mathbb{N}} A_n$; then
$$\mu(A) = \sum_{n\in\mathbb{N}} \mu(A_n) = \sum_{n\in\mathbb{N}} \nu(A_n) = \nu(A),$$
hence $A \in \mathcal{G}$.

But this means that $\mathcal{G}$ is a λ-class containing $\mathcal{B}$. But the smallest λ-class containing $\mathcal{G}$ is $\sigma(\mathcal{B})$ by Theorem 1.6.30, so that we have now

$$\sigma(\mathcal{B}) \subseteq \mathcal{G} \subseteq \sigma(\mathcal{B}),$$

the last inclusion coming from the definition of $\mathcal{G}$. Thus we may conclude that $\mathcal{G} = \sigma(\mathcal{B})$; hence all sets in $\sigma(\mathcal{B})$ have the desired property. $\dashv$

We obtain as a slight extension to Theorem 1.6.29 through Lemma 1.6.18.

Theorem 1.6.32 *Let $\mathcal{C}$ be a semiring over a set S and $\mu : \mathcal{C} \to [0, \infty]$ monotone and countably additive.*

1. *There exists an extension of μ to a measure on the σ-algebra $\sigma(\mathcal{C})$ generated by $\mathcal{C}$.*
2. *If μ is σ*-finite, *then the extension is uniquely determined.*

$\dashv$

The assumption on μ being σ-finite is in fact necessary.

Example 1.6.33 Let $\mathcal{S}$ be the semiring of all left open, right closed intervals on $\mathbb{R}$, and put

$$\mu(I) := \begin{cases} 0 & \text{if } I = \emptyset, \\ \infty, & \text{otherwise.} \end{cases}$$

Then μ has more than one extension to $\sigma(\mathcal{S})$. For example, let $c > 0$ and put $\nu_c(A) := c \cdot |A|$ with $|A|$ as the number of elements of A. Plainly, ν_c extends μ for every c. ✌

Consequently, the assumption that μ is σ-finite cannot be omitted in order to make sure that the extension is uniquely determined.

1.6.4 μ-Measurable Sets

Carathéodory's approach gives even more than an extension to the σ-algebra generated from a semiring. This is what we will discuss next in order to point out a connection with the discussion about the axiom of choice.

Fix for the time being an outer measure μ on $\mathcal{P}(S)$ which we assume to be finite. Call $A \subseteq S$ a *μ-null* set iff we can find a μ-measurable set A_1 with $A \subseteq A_1$ and $\mu(A_1) = 0$. Thus a μ-null set is a set which is covered by a measurable set with μ-measure 0. Because $\mu(X \cap S \setminus A) \leq \mu(X)$ for every $X \subseteq S$ and because an outer measure is monotone, we conclude that each μ-null set is itself μ-measurable. In the same way, we conclude that each set A which can be squeezed between two μ-measurable sets of the same measure (hence $A_1 \subseteq A \subseteq A_2$ with $\mu(A_1) = \mu(A_2)$) must be μ-measurable, because in this case $A \setminus A_1 \subseteq A \setminus A_2$ with $\mu(A \setminus A_2) = 0$. Hence $\mathcal{C}_\mu$ is *complete* in the sense that any A which can be sandwiched in this way is a member of $\mathcal{C}_\mu$.

Completeness

This is a characterization of $\mathcal{C}_\mu$ using these ideas.

Corollary 1.6.34 *Let $\mathcal{C}$ be an algebra over a set S and $\mu : \mathcal{C} \to \mathbb{R}_+$ monotone and countably additive with $\mu(\emptyset) = 0$. Then these statements are equivalent for $A \subseteq S$:*

1. *$A \in \mathcal{C}_\mu$.*
2. *There exists $A_1, A_2 \in \sigma(\mathcal{C})$ with $A_1 \subseteq A \subseteq A_2$ and $\mu(A_1) = \mu(A_2)$.*

Proof The implication 2 $\Rightarrow$ 1 follows from the discussion above, so we will look at 1 $\Rightarrow$ 2. But this is trivial. $\dashv$

Looking back at this development, we see that we can extend our measure far beyond the σ-algebra which is generated from the given semiring. One might even suspect that this extension process gives us the whole power set of the set we started from as the domain for the extended measure. That would of course be tremendously practical, because we then could assign a measure to each subset. But, alas, if the axiom of choice is assumed, these hopes are shattered. The following example demonstrates this. Before discussing it, however, we define and characterize μ-measurable sets on a σ-algebra.

If μ is a finite measure on σ-algebra $\mathcal{B}$, we can define the outer measure $\mu^*(A)$ for any subset $A \subseteq S$ in the same way as we did for functions on a semiring. But since the algebraic structure of a σ-algebra is richer, it is not difficult to see that

$$\mu^*(A) = \inf\{\mu(B) \mid B \in \mathcal{B}, A \subseteq B\}.$$

This is so because a cover of the set A through a countable union of elements on $\mathcal{B}$ is the same as the cover of A through an element of $\mathcal{B}$, because the σ-algebra $\mathcal{B}$ is closed under countable unions. In a similar way, we can try to approximate A from the inside, defining the *inner measure* through μ^*, μ_*

$$\mu_*(A) := \sup\{\mu(B) \mid B \in \mathcal{B}, A \supseteq B\}.$$

So $\mu_*(A)$ is the best approximation from the inside that is available to us. Of course, if $A \in \mathcal{B}$, we have $\mu^*(A) = \mu(A) = \mu_*(A)$, because apparently A is the best approximation to itself.

We can perform the approximation through a sequence of sets, so we are able to precisely fix the inner and the outer measure through elements of the σ-algebra.

Lemma 1.6.35 *Let $A \subseteq S$ and μ be a finite measure on the σ-algebra $\mathcal{B}$.*

1. *There exists $A^* \in \mathcal{B}$ such that $\mu^*(A) = \mu(A*)$.*
2. *There exists $A_* \in \mathcal{B}$ such that $\mu_*(A) = \mu(A_*)$.*

Proof We demonstrate only the first part. For each $n \in \mathbb{N}$, there exists $A_n \in \mathcal{B}$ such that $A \subseteq B_n$ and $\mu(B_n) < \mu(A) + 1/n$. Put $A_n := B_1 \cap \ldots \cap B_n \in \mathcal{B}$, then $A \subseteq A_n$, $\mu(A_n) < \mu(A) + 1/n$, and $(A_n)_{n\in\mathbb{N}}$ decreases. Let $A^* := \bigcap_{n\in\mathbb{N}} A_n \in \mathcal{B}$; then $\mu(A^*) = \inf_{n\in\mathbb{N}} \mu(A_n) = \mu^*(A)$ by the second part of Lemma 1.6.20, because $\mu(A_1) < \infty$. $\dashv$

The set A^* could be called the measurable closure of A; similarly, A_* is its measurable kernel. Using this terminology, we call a set μ-measurable iff its closure and its kernel give the same value.

Definition 1.6.36 *Let μ be a finite measure on the σ-algebra $\mathcal{B}$. $A \subseteq S$ is called μ-*measurable *iff $\mu_*(A) = \mu^*(A)$.*

Every set in $\mathcal{B}$ is μ-measurable, and $\mathcal{B}_\mu$ is the σ-algebra of all μ-measurable sets.

The example which has been announced above shows us that under the assumption of $(\mathbb{AC})$ not every subset of the unit interval is λ-measurable, where λ is Lebesgue measure. Hence we will present a set the inner and the outer measure of which are different.

Vitali's relation

Example 1.6.37 Define $x \ \alpha \ y$ iff $x-y$ is rational for $x, y \in [0,1]$. Then α is an equivalence relation, because the sum of two rational numbers is a rational number again. This is sometimes called *Vitali's equivalence relation*. The relation α partitions the interval $[0,1]$ into equivalence classes. Select from each equivalence class an element (which we can do by $(\mathbb{AC})$), and denote by V the set of selected elements. Hence $V \cap [x]_\alpha$ contains for each $x \in [0,1]$ exactly one element. We want to show that V is not λ-measurable, where λ is Lebesgue measure.

The set $P := \mathbb{Q} \cap [0,1]$ is countable. Define $V_p := \{v + p \mid v \in V\}$, for $p \in P$. If $p, q \in P$ are different, $V_p \cap V_q = \emptyset$. This is so because $v_1 + p = v_2 + q$ implies $v_1 - v_2 = q - p \in \mathbb{Q}$, and thus $v_1 \ \alpha \ v_2$, so v_1 and v_2 are in the same class; hence $v_1 = v_2$, and thus $p = q$, which is a contradiction.

Put $A := \bigcup_{p \in P} V_p$; then $[0,1] \subseteq A \subseteq [0,2]$: Take $x \in [0,1]$, then there exists $v \in V$ with $x \ \alpha \ v$, thus $r := x - v \in \mathbb{Q}$, and hence $x \in V_r$. On the other hand, if $x \in V_r$, then $x = v + r$, and hence $0 \leq x \leq 2$.

If A is λ-measurable, then $\lambda(A) = 0$ is impossible, because this would imply $\lambda([0,1]) = 0$, since λ is monotone. Thus $\lambda(A) > 0$. But $\lambda(V_p) = \lambda(V)$ for each p, so that $\lambda(A) = \infty$ by countable additivity. But this contradicts $\lambda([0,2]) = 2$. Hence A is not λ-measurable, which implies that V is not λ-measurable. ✌

So, let us record for later use: If $(\mathbb{AC})$ holds, then there exists a subset of the unit interval which is not Lebesgue measurable.

1.7 Banach–Mazur Games

We will now demonstrate that the games we are about to introduce lead to considerations replacing $(\mathbb{AC})$ by another axiom, which in turn will be the base for establishing that *every* subset of $[0,1]$ is Lebesgue measurable. This will be done through a suitable two-person game.

We have two players, *Angel* and *Demon*, playing against each other. For simplicity, we assume that *playing* means offering a natural number and that the game—like True Love—never ends. Let A be a set of infinite sequences of natural numbers; then the game G_A is played as follows. Angel starts with $a_0 \in \mathbb{N}$, and Demon answers with $b_0 \in \mathbb{N}$, taking Angel's move a_0 into account. Angel replies with a_1, taking the game's history $\langle a_0, b_0 \rangle$ into account, and then Demon answers with b_1, contingent upon $\langle a_0, b_0, a_1 \rangle$, and so on. Angel wins this game, if the sequence $\langle a_0, b_0, a_1, b_1, \ldots \rangle$ is a member of A; otherwise Demon wins.

Let us have a look at strategies. Define

$$\mathcal{N} := \mathbb{N}^\infty$$

as the set of all sequences of natural numbers, and let

$$\mathcal{S} := \bigcup \{\langle n_1, \ldots, n_k \rangle \mid k \geq 0, n_1, \ldots, n_k \in \mathbb{N}\}$$

be the set of all finite sequences of natural numbers. For easier notation later on, we define appending an element to a finite sequence by $\langle n_1, \ldots, n_k \rangle \frown n := \langle n_1, \ldots, n_k, n \rangle$. $\mathcal{S}_u$ and $\mathcal{S}_g$ denote all sequences of odd, resp., even and length, the empty sequence is denoted by ϵ, and we assume $\epsilon \in \mathcal{S}_g$. $\mathcal{S}, \mathcal{S}_u, \mathcal{S}_g$

A *strategy* $\mathfrak{d}$ for Angel is a map $\mathfrak{d} : \mathcal{S}_g \to \mathbb{N}$ which works in the following way: $a_0 := \mathfrak{d}(\epsilon)$ is the first move of Angel, Demon replies with b_0, then Angel answers with $a_1 := \mathfrak{d}(a_0, b_0)$, Demon reacts with b_1, which Angel answers with $a_2 := \mathfrak{d}(a_0, b_0, a_1, b_1)$, and so on. If Angel plays according to strategy $\mathfrak{d}$ and Demon's moves are given by $b := \langle b_0, b_1, \ldots \rangle \in \mathcal{N}$, then the game's events are collected in $\mathfrak{d} * b \in \mathcal{N}$, where we define $\mathfrak{d} * b := \langle a_0, b_0, a_1, b_1, \ldots \rangle$ with $a_{2k+1} = \mathfrak{d}(a_0, b_0, \ldots, a_{2k}, b_{2k} \rangle$ for $k \geq 0$ and $a_0 = \mathfrak{d}(\epsilon)$. Similarly, a strategy $\mathfrak{a}$ for Demon is a map $\mathfrak{a} : \mathcal{S}_u \to \mathbb{N}$, working in this manner: If Angel plays a_0, Demon answers with $b_0 := \mathfrak{a}(a_0)$; then Angel plays a_1, to which Demon replies with $b_1 := \mathfrak{a}(a_0, b_1, a_1)$; and so on. If Angel's moves are collected in $a := \langle a_0, a_1, \ldots \rangle$ and if Demon plays strategy $\mathfrak{a}$, then the entire game is recorded in the sequence $a * \mathfrak{a}$. Thus we define $a * \mathfrak{a} := \langle a_0, b_0, a_1, b_1, \ldots \rangle$ with $b_k = \mathfrak{a}(a_0, b_0, \ldots, a_k)$ for $k \geq 0$. $\mathfrak{d} * b, a * \mathfrak{a}$

Definition 1.7.1 $\mathfrak{d} : \mathcal{S}_g \to \mathbb{N}$ *is a* winning strategy for Angel *in game* G_A *iff*

$$\{\mathfrak{d} * b \mid b \in \mathcal{N}\} \subseteq A,$$

$\mathfrak{a} : \mathcal{S}_u \to \mathbb{N}$ *is a* winning strategy for Demon *in game* G_A *iff*

$$\{a * \mathfrak{a} \mid a \in \mathcal{N}\} \subseteq \mathcal{N} \setminus A.$$

It is clear that at most one of Angel and Demon can have a winning strategy. Suppose that in the contrary both have one, say, $\mathfrak{d}$ for Angel and $\mathfrak{a}$ for Demon. Then $\mathfrak{d} * \mathfrak{a} \in A$, since $\mathfrak{d}$ is winning for Angel, and $\mathfrak{d} * \mathfrak{a} \notin A$, since $\mathfrak{a}$ is winning for Demon. So this assumption does not make sense.

Banach-Mazur game

We have a look at *Banach–Mazur games*, another formulation of the games just introduced, which is sometimes more convenient. Each Banach–Mazur game can be transformed into a game which we have defined above.

Before discussing it, it will be convenient to introduce some notation. Let $a, b \in \mathcal{S}$; hence a and b are finite sequences of natural numbers. We say that $a \preceq b$ iff a is an *initial piece* of b (including $a = b$), so there exists $c \in \mathcal{S}$ with $b = ac$; c is denoted by b/a. If we want to exclude equality, we write $a \prec b$.

Example 1.7.2 The game is played over $\mathcal{S}$; a subset $B \subseteq \mathcal{N}$ indicates a winning situation. Angel plays $a_0 \in \mathcal{S}$, Demon plays b_0 with $a_0 \preceq b_0$, then Angel plays a_1 with $a_0 b_0 \preceq a_1$, etc. Angel wins this game iff the finite sequence $a_0 b_0 a_1 b_1 \ldots$ converges to an infinite sequence $x \in B$.

We encode this game in the following way. $\mathcal{S}$ is countable by Proposition 1.2.5, so write this set as $\mathcal{S} = \{r_n \mid n \in \mathbb{N}\}$. Put

$$A := \{\langle w_0, w_1, \ldots \rangle \mid r_{w_0} \preceq r_{w_1} \preceq r_{w_2} \ldots \text{ converges to a sequence in } B\}.$$

It is then immediate that Angel has a strategy for winning the Banach–Mazur game iff it has one for winning that game G_A. ✌

Consequently, this class of games will be called Banach–Mazur games throughout (we will encounter other games as well).

1.7.1 Determined Games

Games in which neither Angel nor Demon has a winning strategy are somewhat, well, indeterminate and might be avoided. We see some similarity between a strategy and the selection argument in $(\mathbb{AC})$, because a strategy selects an answer among several possible choices, while a choice function picks elements, each from a given set. This intuitive similarity will be investigated now.

Definition 1.7.3 *A game G_A is called* determined *iff either Angel or Demon has a winning strategy.*

Suppose that each game G_A is determined, no matter what set $A \subseteq \mathcal{N}$ we chose; then we can define a choice function for countable families of nonempty subsets of $\mathcal{N}$.

Theorem 1.7.4 *Assume that each game is determined. Then there exists a choice function for countable families of nonempty subsets of $\mathcal{N}$.*

Proof 1. Let $\mathcal{F} := \{X_n \mid n \in \mathbb{N}\}$ be a countable family with $\emptyset \neq X_n \subseteq \mathcal{N}$ for $n \in \mathbb{N}$. We will define a function $f : \mathcal{F} \to \mathcal{N}$ such that $f(X_n) \in X_n$ for all $n \in \mathbb{N}$. The idea is to play a game which Angel cannot win, hence for which Demon has a winning strategy. To be specific, if Angel plays $\langle a_0, a_1, \ldots \rangle$ and Demon plays $b := \langle b_0, b_1, \ldots \rangle$, then Demon wins iff $b \in X_{a_0}$. Since by assumption Demon has a winning strategy $\mathfrak{a}$, we then put

The idea is to play a game

$$f(X_n) := \langle n, 0, 0, \ldots \rangle * \mathfrak{a}.$$

2. Let us look at this idea. Put

$$A := \{\langle x_0, x_1, \ldots \rangle \in \mathcal{N} \mid \langle x_1, \ldots \rangle \notin X_{x_0}\}.$$

Suppose that Angel starts upon playing a_0. Since $X_{a_0} \neq \emptyset$, Demon can take an arbitrary $b \in X_{a_0}$ and plays $\langle b_0, b_1, \ldots \rangle$. Hence Angel cannot win, so B has a winning strategy $\mathfrak{a}$.

3. Now look at $\langle n, 0, 0, \ldots \rangle * \mathfrak{a} \notin A$, because $\mathfrak{a}$ is a winning strategy. From the definition of A, we see that this is an element of X_n, so we have found a choice function indeed. $\dashv$

The space $\mathcal{N}$ looks a bit artificial, just as a mathematical object to play around with. But this is not the case. It can be shown that there exists a bijection $\mathcal{N} \to \mathbb{R}$ with some desirable properties (we will not enter into this construction, however, but refer the reader to Sect. 4.4).

We state as a consequence of Theorem 1.7.4

Corollary 1.7.5 *Assume that each game is determined. Then there exists a choice function for countable families of nonempty subsets of $\mathbb{R}$.*
$\dashv$

Let us fix the assumption on the existence of a winning strategy for either Angel or Demon in an axiom, the *axiom of determinacy*.

$$\boxed{(\mathbb{AD})\text{ Each game is determined.}}$$

Given Corollary 1.7.5, the relationship of the axiom of determinacy to the axiom of choice is of interest.

Does $(\mathbb{AD})$ imply $(\mathbb{AC})$?

The hope of establishing this is shattered, however, by this observation.

Proposition 1.7.6 *If* $(\mathbb{AC})$ *holds, there exists* $A \subseteq \mathcal{N}$ *such that* G_A *is not determined.*

Before entering the proof, we observe that the set of all strategies S_A for Angel resp. S_D for Demon has the same cardinality as the power set $\mathcal{P}(\mathbb{N})$ of $\mathbb{N}$.

Proof 0. We have to find a set $A \subseteq \mathcal{N}$ such that neither Angel nor Demon has a winning strategy for the game G_A. By $(\mathbb{AC})$, the sets S_A resp. S_D can be well ordered; by the observation just made, we can write

$$S_A = \{\mathfrak{d}_\alpha \mid \alpha < \omega_1\},$$
$$S_D = \{\mathfrak{a}_\alpha \mid \alpha < \omega_1\}.$$

1. We will construct now disjoint sets $X = \{x_\alpha \mid \alpha < \omega_1\} \subseteq \mathcal{N}$ and $Y = \{y_\alpha \mid \alpha < \omega_1\} \subseteq \mathcal{N}$ indexed by $\{\alpha \mid \alpha < \omega_1\}$, which will help define the game. Suppose x_β and y_β are defined for all $\beta < \alpha$. Then, because α is countable, the sets $\{x_\beta \mid \beta < \alpha\}$ and $\{y_\beta \mid \beta < \alpha\}$ are countable as well, and there are uncountably many $b \in \mathcal{N}$ such that $\mathfrak{d}_\alpha * b \notin \{x_\beta \mid \beta < \alpha\}$. Take one of them and put $y_\alpha := \mathfrak{d}_\alpha * b$. For the same reason, there are uncountably many $a \in \mathcal{N}$ such that $a * \mathfrak{a}_\alpha \notin \{y_\beta \mid \beta \leq \alpha\}$; take one of them and put $x_\alpha := a * \mathfrak{a}_\alpha$.

2. Clearly, X and Y are disjoint. Angel does not have a strategy for winning game G_X. Suppose it has a winning strategy $\mathfrak{d}$, so that $\mathfrak{d} = \mathfrak{d}_\alpha$ for some $\alpha < \omega_1$. But $y_\alpha = \mathfrak{d}_\alpha * b \notin X$ by construction, which is a contradiction. One similarly shows that Demon cannot have a winning strategy for game G_X. Hence this game is not determined. $\dashv$

1.7.2 Proofs Through Games

Games are a tool for proofs. The basic idea is to attach a statement to a game, and if Angel has a strategy for winning the game, then the statement is established; otherwise it is not. Hence we have to encode the statement in such a way that this mechanism can be used, but we have also to establish a scenario in which to argue. The formulation chosen suggests that Angel has to have a winning strategy for winning a game, which in turn suggests that we assume a framework in which games are determined. But we have seen above that this is not without conflicts when considering ($\mathbb{AC}$).

This section is devoted to establish that every subset of the unit interval is Lebesgue measurable, provided each game is determined. We have seen in Example 1.6.37 that ($\mathbb{AC}$) implies that there exists a set which is not Lebesgue measurable. Hence "it is natural to postulate that Determinacy holds to the extent that it does not contradict the Axiom of Choice," as T. Jech writes in his massive treatise of set theory [Jec06, p. 628].

We will discuss another example for using games as tools for proofs when we formulate a Banach–Mazur game for establishing properties of a topological space in Sect. 3.5.2.

The Goal. We want to show that each subset of the unit interval is measurable, provided each game is determined. This is based on the observation that it is sufficient to establish that $\lambda_*(A) > 0$ or $\lambda_*([0, 1] \setminus A) > 0$ for each and every subset $A \subseteq [0, 1]$, where, as above, λ is Lebesgue measure on the unit interval. This is the reason, which follows from the construction for Vitali's equivalence relation; see Example 1.6.37.

Lemma 1.7.7 *Assume that there exists a subset of the unit interval which is not λ-measurable. Then there exists a subset $M \subseteq [0, 1]$ with $\lambda_*(M) = 0$ and $\lambda^*(M) = 1$.* $\dashv$

The Basic Approach. Given an arbitrary subset $X \subseteq [0, 1]$, we will define a game G_X such that if there exists a winning strategy for Angel, then we can find a measurable subset $A \subseteq X$ which has positive Lebesgue measure (hence $\lambda_*(X) > 0$). If there exists, however, a winning strategy for Demon, then we can find a measurable subset $A \subseteq [0, 1]$ with positive Lebesgue measure such that $A \cap X = \emptyset$ (hence $\lambda_*([0, 1] \setminus A) > 0$).

Little Helpers. We need some preparations before we start. So let us get on with it now as not to interrupt the flow of discussion later on.

Lemma 1.7.8 *Let $(F_n)_{n\in\mathbb{N}}$ be a sequence of nonempty subsets of the unit interval* $[0, 1]$ *such that*

1. *Each F_n is a finite union of closed intervals.*
2. *The sequence is monotonically decreasing; hence $F_1 \supseteq F_1 \supseteq \dots$.*
3. *The sequence of diameters* $\mathrm{diam}(F_n) := \sup_{x,y\in F_n} |x - y|$ *tends to zero.* Diameter

Then there exists a unique $p \in [0, 1]$ with $\{p\} = \bigcap_{n\in\mathbb{N}} F_n$.

Proof 1. It is clear from the last condition that there can be at most one point in the intersection of this sequence: Suppose there are two distinct points p, q in this intersection; then $\delta := |p - q| > 0$. But there exists some $n_0 \in \mathbb{N}$ with $\mathrm{diam}(F_m) < \delta$ for all $m \geq n_0$. This is a contradiction.

2. Assume that $\bigcap_{n\in\mathbb{N}} F_n = \emptyset$. Put $G_n := [0, 1] \setminus F_n$; then G_n is the union of a finite number of open intervals, say $G_n = H_{n,1} \cup \ldots \cup H_{n,k_n}$ and $[0, 1] \subseteq \bigcup_{n\in\mathbb{N}} G_n$. By the Heine–Borel Theorem 1.5.46, there exists a finite set of intervals H_{n_i,j_i} with $1 \leq i \leq r, 1 \leq j_i \leq k_{n_i}$ such that $[0, 1] \subseteq \bigcup_{i=1}^{r} H_{n_i,j_i}$. Because the sequence of the F_n decreases, the sequence $(G_n)_{n\in\mathbb{N}}$ is increasing, so we find an index N such that $H_{n_i,j_i} \subseteq G_N$ for $1 \leq i \leq r, 1 \leq j_i \leq k_{n_i}$. But this means $[0, 1] \subseteq G_N$; thus $F_N = \emptyset$, contradicting the assumption that all F_n are nonempty. $\dashv$

A more general version of this Lemma will be found in Propositions 3.5.25 and 3.6.67; interesting enough, Proposition 3.5.25 will also be an important tool in the investigation of the game in Sect. 3.5.2.

Another preparation concerns the convergence of an infinite product.

Lemma 1.7.9 *Let* $(a_n)_{n\in\mathbb{N}}$ *be a sequence of real numbers with* $0 < a_n < 1$ *for all* $n \in \mathbb{N}$. *Then the following statements are equivalent:*

1. $\prod_{i\in\mathbb{N}}(1-a_i) := \lim_{n\to\infty}\prod_{i=1}^{n}(1-a_i)$ *exists.*
2. $\sum_{n\in\mathbb{N}} a_n$ *converges.*

Proof One shows easily by induction on n that

$$\prod_{i=1}^{n}(1-a_i) > 1 - \big(\sum_{1=1}^{n} a_n\big)$$

for $n \geq 2$. Since $0 < a_n < 1$ for all $n \in \mathbb{N}$, this implies the equivalence. ⊣

This has an interesting consequence, viz., that we have a positive infinite product, provided the corresponding series converges. To be specific

Corollary 1.7.10 *Let* $(a_n)_{n\in\mathbb{N}}$ *be a sequence of real numbers with* $0 < a_n < 1$ *for all* $n \in \mathbb{N}$. *Then the following statements are equivalent:*

1. $\prod_{i\in\mathbb{N}}(1-a_i)$ *is positive.*
2. $\sum_{n\in\mathbb{N}} a_n$ *converges.*

Proof 1. Put $Q_k := \prod_{i=1}^{k} a_i, Q := \lim_{k\to\infty} Q_k$. Assume that $\sum_{n\in\mathbb{N}} a_n$ converges; then there exists $m \in \mathbb{N}$ such that $\mathfrak{d}_m := \sum_{i=m}^{\infty} < 1$. Hence we have

$$\frac{Q_n}{Q_m} > 1 - (a_{m+1} + \ldots + a_n) > 1 - \mathfrak{d}_m$$

for $n > m$, so that

$$Q = \lim_{k\to\infty} Q_k > Q_m \cdot (1 - \mathfrak{d}_m) > 0.$$

2. On the other hand, if the series diverges, then we can find an index m for $N \in \mathbb{N}$ such that $a_1 + \ldots + a_n > N$ whenever $n > m$. Hence

$$\prod_{n\in\mathbb{N}} \frac{1}{1-a_n} \leq \lim_{k\to\infty} \frac{1}{1-(a_1 + \ldots + a_k)} = 0.$$

⊣

This observation will be helpful when looking at our game.

The Game. Before discussing the game proper, we set its stage. Fix a sequence $(r_n)_{n\in\mathbb{N}}$ of positive reals such that $\sum_{n\in\mathbb{N}} r_n < 1$ and $1/2 > r_1 > r_2 > \dots$.

Let $k \in \mathbb{N}$ be a natural number, and define $\mathcal{J}_k$ as the collection of sets S with these properties:

- $S \subseteq [0, 1]$ is a finite union of closed intervals with rational end-points.
- The diameter $\mathrm{diam}(S) = \sup_{x,y\in S} |x - y|$ of S is smaller than $1/2^k$.
- The Lebesgue measure $\lambda(S)$ of S is $r_1 \cdot \dots \cdot r_k$.

Important note

Put $\mathcal{J}_0 := \{[0, 1]\}$ as the mandatory first draw of Angel. Note that $\mathcal{J}_k$ is countable for all $k \in \mathbb{N}$, so that $\bigcup_{k\geq 0} \mathcal{J}_k$ is countable as well by Proposition 1.2.6 (it is important to note this was proved without using $(\mathbb{AC})$).

The game

The game starts. We fix $X \subseteq [0, 1]$ as the Great Prize; this is the set we want to investigate. Angels starts with choosing the unit interval $S_0 := [0, 1]$, Demon chooses a set $S_1 \in \mathcal{J}_1$, then Angel chooses a set $S_2 \in \mathcal{J}_2$ with $S_2 \subset S_1 \subset S_0$, Demon chooses a set $S_3 \subset S_2$ with $S_3 \in \mathcal{J}_3$, and so on. In this way, the game defines a decreasing sequence $(S_n)_{n\in\mathbb{N}}$ of closed sets, the diameter of which tends to zero. By Lemma 1.7.8 there exists exactly one point p with $p \in \bigcap_{n\in\mathbb{N}} S_n$. If $p \in [0, 1] \setminus X$, then Angel wins, and if $p \in X$, then Demon wins.

Analysis of the Game. First note that we will not encode the game into a syntactic form according to the definition of G_A. This would require much encoding and decoding between the formal representation and the informal one, so that the basic ideas might get lost [Ven07]. Since life is difficult enough, we stick to the informal representation, trusting that the formal one is easily derived from it, and focus on the ideas behind the game. After all, we want to prove something through this game which is not entirely trivial.

The game spawns a tree rooted at $S_0 := [0, 1]$ with offsprings all those elements S_1 of $\mathcal{J}_1$ with $S_1 \subset S_0$. Continuing inductively, assume that we are at node $S_k \in \mathcal{J}_k$; then this node has all elements $S \in \mathcal{J}_{k+1}$ as offsprings for which $S \subset S_k$ holds. Consequently, the tree's depth will be infinite, because the game continues forever. The offsprings of a node will be investigated in a moment.

We define for easier discussion the sets

$$\mathcal{W}_k := \{\langle S_0, \dots, S_k\rangle \in \prod_{i=0}^{k} \mathcal{J}_i \mid S_0 \supset S_1 \supset \dots \supset S_k\},$$
$$\mathcal{W}^* := \bigcup_{k\geq 0} \mathcal{W}_k$$

as the set of all paths which are possible in this game. Hence Angel chooses initially $S_0 = [0,1]$, Demon chooses $S_1 \in \mathcal{J}_1$ with $S_1 \subset S_0$ (hence $\langle S_0, S_1\rangle \in \mathcal{W}_1$), so that $\langle S_0, S_1, S_2\rangle \in \mathcal{W}_2$, etc. $\mathcal{W}_{2n}$ is the set of all possible paths after the nth draw of Angel, and $\mathcal{W}_{2n+1}$ yields the state of affairs after the nth move of Demon.

For an analysis of strategies, we will fix now $k \in \mathbb{N}$ and a map $\Gamma : \mathcal{W}_k \to \mathcal{J}_{k+1}$ such that $\Gamma(S_0, \dots, S_k) \subset S_k$; hence $\langle S_0, \dots, S_k, \Gamma(S_0, \dots, S_k)\rangle = \langle S_0, \dots, S_k\rangle \frown \Gamma(S_0, \dots, S_k) \in \mathcal{W}_{k+1}$. Just to have a handy name for it, call such a map *admissible at* k. Γ admissible

Lemma 1.7.11 *Assume Γ is admissible at k. Given $\langle S_0, \dots, S_k\rangle \in \mathcal{W}_k$, there exists $m \in \mathbb{N}$ and a finite sequence $T_{k+1,i} \in \mathcal{J}_{k+1}$ for $1 \leq i \leq m$ such that*

1. *$T_{k+1,i} \subset S_k$ for all i,*
2. $\lambda\big(\bigcup_{i=1}^m \Gamma(S_0, \dots, S_k, T_{k+1,i})\big) \geq \lambda(S_k) \cdot (1 - 2 \cdot r_{k+1})$,
3. *The sets $\Gamma(S_0, \dots, S_k, T_{k+1,1}), \dots, \Gamma(S_0, \dots, S_k, T_{k+1,m})$ are mutually disjoint.*

Proof The sets $T_{k+1,i}$ are defined by induction. Assume that $T_{k+1,1}, \dots, T_{k+1,j}$ is already defined for $j \geq 0$; put

$$R_j := S_k \setminus \bigcup_{i=1}^{j} \Gamma(S_0, S_1, \dots, S_k, T_{k+1,i}).$$

Now we have two possible cases: either

$$(\ddagger)\ \lambda(R_j) > 2 \cdot \lambda(S_k) \cdot r_{k+1}$$

or this inequality is false. Note that $\lambda(S_k) = r_1 \cdot \ldots \cdot r_k$, and $1/2 > r_k > r_{k+1}$, so that initially $\lambda(R_0) = \lambda(S_k) > 2 \cdot \lambda(S_k) \cdot r_{k+1}$. Now assume that (‡) holds. Because S_k is the union of a finite number of closed intervals and because R_j does not exhaust S_k, we conclude that R_j contains a subset P with diameter $\text{diam}(P) \leq \text{diam}(R_j) \leq 2^{-(k+1)}$ such that $\lambda(P) > \lambda(S_k)$. We can select P in such a way that it is a finite union of intervals. Then there exists $T_{k+1,j+1} \subseteq P$ which belongs to $\mathcal{J}_{k+1}$. Take it. Then the first property is satisfied.

This process continues until inequality (‡) becomes false, which gives the second property. Because

$$\Gamma(S_0, \ldots, S_k, T_{k+1,i}) \subset T_{k+1,i} \subset S_k \setminus \bigcup_{j=1}^{i-1} \Gamma(S_0, \ldots, S_k, T_{k+1,j}),$$

we conclude that the sets $\Gamma(S_0, \ldots, S_k, T_{k+1,1}), \ldots, \Gamma(S_0, \ldots, S_k, T_{k+1,m})$ are mutually disjoint. ⊣

Now let $\mathfrak{a}$ be a strategy for Demon; hence $\mathfrak{a} : \bigcup_{k\geq 0} \mathcal{W}_{2k+1} \to \bigcup_{k\leq 0} \mathcal{J}_{2k}$ is a map such that $\mathfrak{a}(S_0, \ldots, S_{2k}) \subset S_{2k}$. If the game's history at time k is given by the path $\langle S_0, \ldots, S_{2k}\rangle$ with Angels having played S_{2k} as a last move, then the game continues with $\mathfrak{a}(S_0, \ldots, S_{2k})$ as Demon's next move, so that the new path is just $\langle S_0, \ldots, S_{2k}\rangle \frown \mathfrak{a}(S_0, \ldots, S_{2k})$.

Let us see what happens if Angel selects the next move according to Lemma 1.7.11. Initially, Angels plays S_0, then Demon plays $\mathfrak{a}(S_0)$, so that the game's history is now $\langle S_0\rangle \frown \mathfrak{a}(S_0)$; let $T_{0,1}, \ldots, T_{0,m_0}$ be the sets selected according to Lemma 1.7.11 for this history; then the possible continuations in the game are $t_i := \langle S_0\rangle \frown \mathfrak{a}(S_0) \frown T_{0,i}$ for $1 \leq i \leq m_0$, so that Demon's next move is $t_i \frown \mathfrak{a}(t_i)$, and thus

$$\begin{aligned}&K_{\mathfrak{a}}\big(\langle S_0\rangle \frown \mathfrak{a}(S_0)\big)\\ &\quad := \{\langle S_0\rangle \frown \mathfrak{a}(S_0) \frown T_{0,i} \frown \mathfrak{a}\big(\langle S_0\rangle \frown \mathfrak{a}(S_0) \frown T_{0,i}\big) \mid 1 \leq i \leq m_0\} \in \mathcal{W}_3\end{aligned}$$

describes all possible moves for Demon in this scenario. Given $\mathfrak{a}$, this depends on S_0 as the history up to that moment and on the choice to Angel's moves according to Lemma 1.7.11. To see the pattern, consider Demon's next move. Take $t = \langle t_0, t_1, t_2, t_3\rangle \in K_{\mathfrak{a}}(S_0 \frown \mathfrak{a}(S_0))$, then $\mathfrak{a}(t) \in \mathcal{J}_4$ with $\mathfrak{a}(t) \subset t_3$, and choose $T_{1,1}, \ldots, T_{1,m_1}$ according to Lemma 1.7.11 as possible next moves for Angel, so that the set of all possible moves for Demon given this history is an element of the set

$$
\begin{aligned}
K_{\mathfrak{a}}(t) &= K_{\mathfrak{a}}(\langle t_1, t_2, t_3\rangle \frown \mathfrak{a}(t_1, t_2, t_3)) \\
&:= \{t \frown T_{1,i} \frown \mathfrak{a}(t \frown T_{1,i}) \mid 1 \le i \le m_1\} \in \mathcal{W}_5.
\end{aligned}
$$

This provides a window into what is happening. Now let us look at the broader picture. Denote for $t \in \mathcal{W}_n$ by $J_{\mathfrak{a}}(t)$ the set $\{t \frown T_{n,1}, \dots, t \frown T_{n,m}\}$, where $T_{n,1}, \dots, T_{n,m}$ are determined for t and $\mathfrak{a}$ according to Lemma 1.7.11 as the set of all possible moves for Angel. Hence given history t, $J_{\mathfrak{a}}(t)$ is the set of all possible paths for which Demon has to provide the next move. Then put

$$
J_{\mathfrak{a}}^n := \bigcup_{s_2 \in J_{\mathfrak{a}}(\langle S_0\rangle \frown \mathfrak{a}(S_0))} \; \bigcup_{s_4 \in J_{\mathfrak{a}}(s_2 \frown \mathfrak{a}(s_2))} \dots \bigcup_{s_{2(n-1)} \in J_{\mathfrak{a}}(s_{2(n-2)} \frown \mathfrak{a}(s_{2(n-2)}))} J_{\mathfrak{a}}\big(s_{2(n-1)} \frown \mathfrak{a}(s_{2(n-1)})\big)
$$

with

$$
J_{\mathfrak{a}}^1 = J_{\mathfrak{a}}(\langle S_0\rangle \frown \mathfrak{a}(S_0)).
$$

Finally, define

$$
A_n := \bigcup \{\mathfrak{a}(s_{2n}) \mid s_{2n} \in J_{\mathfrak{a}}^n\}.
$$

Hence $J_{\mathfrak{a}}^n$ contains all possible moves of Angel at time $2n$, so that A_n tells us what Demon can do at time $2n + 1$. These are the important properties of $(A_n)_{n \in \mathbb{N}}$.

Lemma 1.7.12 *We have for all $n \in \mathbb{N}$*

1. $\lambda(A_n) \ge r_1 \cdot \prod_{i=1}^n (1 - 2 \cdot r_{2i})$
2. $A_{n+1} \subset A_n$

Proof 1. The second property follows immediately from Lemma 1.7.11, so we will focus on the first property. It will be proved by induction on n. We infer from Lemma 1.7.11 that the sets $\mathfrak{a}(s_{2n})$ are mutually disjoint, when s_{2n} runs through $J_{\mathfrak{a}}^n$.

2. The induction begins at $n = 1$. We obtain immediately from Lemma 1.7.11 that

$$
\begin{aligned}
\lambda(A_1) &= \lambda\big(\bigcup \{\mathfrak{a}(s_2) \mid s_2 \in J_{\mathfrak{a}}(\langle S_0\rangle \frown \mathfrak{a}(S_0))\}\big) \\
&\ge r_1 \cdot (1 - 2 \cdot r_2)
\end{aligned}
$$

(set $\Gamma := \mathfrak{a}$ and $k = 1$).

2. Induction step $n \to n+1$. We infer from Lemma 1.7.11 that

$$(\dagger)\ \lambda\Big(\bigcup\{\mathfrak{a}(s_{2(n+1)} \mid s_{2(n+1)} \in J_{\mathfrak{a}}(s_{2n})\}\Big) \\ \geq \lambda(\mathfrak{a}(s_{2n+1})) \cdot (1 - 2 \cdot r_{2(n+1)}).$$

Disjointness then implies

$$\begin{aligned}
\lambda(A_{n+1}) &= \sum_{s_{2n} \in J_{\mathfrak{a}}^n} \lambda\Big(\bigcup\{\mathfrak{a}(s_{2(n+1)} \mid s_{2(n+1)} \in J_{\mathfrak{a}}(s_{2n})\}\Big) \\
&\geq \sum_{s_{2n} \in J_{\mathfrak{a}}^n} \lambda(\mathfrak{a}(s_{2n})) \cdot (1 - 2 \cdot r_{2(n+1)}) && \text{(inequality } (\dagger)) \\
&= \lambda\Big(\bigcup_{s_{2n} \in J_{\mathfrak{a}}^n} \mathfrak{a}(s_{2n+1})\Big) \cdot (1 - 2 \cdot r_{2(n+1)}) && \text{(disjointness)} \\
&= \lambda(A_n) \cdot (1 - 2 \cdot r_{2(n+1)}) && \text{(induction hypothesis)} \\
&\geq r_1 \cdot \prod_{i=1}^{n+1} (1 - 2 \cdot r_{2i}).
\end{aligned}$$

$\dashv$

Now we are getting somewhere—we show that we can find for every element in $\bigcap_{n \in \mathbb{N}} A_n$ a strategy so that the moves of Angel and of Demon *converge* to this point. To be more specific

Lemma 1.7.13 *Assume that Demon adopts strategy $\mathfrak{a}$. For every point $p \in \bigcap_{n \in \mathbb{N}} A_n$, there exists for Angel a strategy $\mathfrak{d}_p$ with this property: If Angel plays $\mathfrak{d}_p$ and Demon plays $\mathfrak{a}$, then $\bigcap_{i=0}^{\infty} S_i = \{p\}$, where $S_0, S_1, \ldots$ are the consecutive moves of the players.*

Proof The sets $s_{2n} \in J_{\mathfrak{a}}^n$ are mutually disjoint for fixed n, so we find a unique sequence $s'_{2n} \in J_{\mathfrak{a}}^n$ for which $p \in \mathfrak{a}(s'_{2n})$. Represent $s'_{2n} = \langle S_0, \ldots, S_{2n} \rangle$, and let $\mathfrak{d}_p$ be a strategy for Angel such that $\mathfrak{d}_p(\langle S_0, \ldots, S_{2n-1} \rangle \frown \mathfrak{a}(S_0, \ldots, S_{2n-1})) = S_{2n}$ holds. Thus $p \in \bigcap_{n \in \mathbb{N}} S_n$, if Angel plays $\mathfrak{d}_p$ and Demon plays $\mathfrak{a}$. $\dashv$

Now let $\mathfrak{a}$ be a winning strategy for Demon; then $A := \bigcap_{n \in \mathbb{N}} A_n \subseteq [0,1] \setminus X$; this is the outcome if Angel plays one of the strategies in $\{\mathfrak{d}_p \mid p \in A\}$. There may be other strategies for Angel than the one described above, but no matter how Angel plays the game, we will end up in an element not in X. This implies $\lambda(A) \leq \lambda_*([0,1] \setminus X)$. But we know from Lemma 1.7.12 that $\lambda(A) \geq r_1 \cdot \prod_{i=1}^{\infty}(1 - 2 \cdot r_{2i}) > 0$ by Lemma 1.7.9 and its corollary, consequently, $\lambda_*([0,1] \setminus X) > 0$.

If, however, Demon does not have a winning strategy, then Angel has one, if we assume that the game is determined. The argumentation is completely the same as above to show that $\lambda_*(X) > 0$.

Thus we have shown

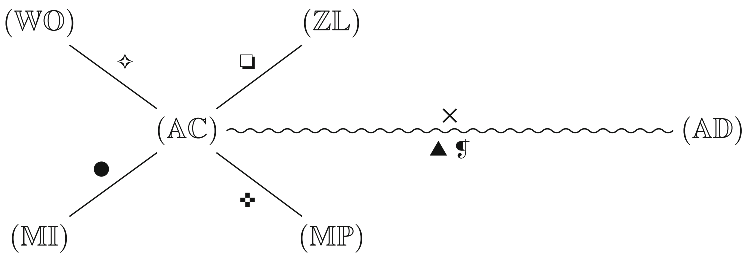
Theorem 1.7.14 *If each game is determined, then each subset of the unit interval is λ-measurable.* $\dashv$

We have seen that games are not only just for fun, but are a tool for investigating properties of sets. In fact, one can define games for investigating many topological properties, not all as laborious as the one we have defined above.

1.8 Wrapping It Up

This summarizes the discussion. Some hints for further information can be found in the Bibliographic Notes. The Lecture Note [Her06] by H. Herrlich and the list of its references contain a lot of suggestions for further reading. The discussion in P. Taylor's book [Tay99, p. 58] ("Although we, at the cusp of the century, now reject Choice . . .") is also worth looking at, albeit from a completely different angle.

This is a small diagram indicating the dependencies discussed here:



The symbols provide a guide to the corresponding statements:

✧	Theorem 1.4.20
❏	Proposition 1.5.1
●	Theorem 1.5.38
✜	Proposition 1.5.3
×	Existence of a nondetermined game under $(\mathbb{AC})$, Proposition 1.7.6
▲	Choice function for countable families under $(\mathbb{AD})$, Theorem 1.7.4
❡	Measurability of every subset of $[0, 1]$ under $(\mathbb{AD})$, Theorem 1.7.14

1.9 Bibliographic Notes

This chapter contains mostly classical topics. The proof of Cantor's enumeration and its consequences for enumerating the set of all finite sequences of natural numbers is taken from [KM76], so is the discussion of ordinals. Jech's representation [Jec06] has been helpful as well, so was [Gol96]. The books by Davis [Dav00] and by Aczel [Acz00] contain some gripping historical information on the subject of early set theory; the monograph [COP01] discusses implications for computing when the axiom of foundations (p. 7) is weakened.

Term rewriting is discussed in [BN98]; reduction systems (Example 1.3.7) are central to it. Aumanns's classic [Aum54], unfortunately not as frequently used as this valuable book should be, helped in discussing Boolean algebras, and the proof for the general distributive law in Boolean algebras as well as some exercises has been taken from [Bir67] and from [DP02]; see also [Sta97] for finite lattices. The discussion on measure extension follows quite closely the representation given in the first three chapters of [Bil95] with an occasional glimpse at [Els99] and the awesome [Bog07]. Finally, games are introduced as in [Jec06, Chap. 33]; see also [Jec73]; the game-theoretic proofs on measurability are taken from [MS64]. Infinite products are discussed at length in the delightful textbook [Bro08]; see also [Chr64]. A general source for this chapter was the exposition by Herrlich [Her06], providing a *tour d'horizon*. A graphic view of the foundational crisis in mathematics at the turn of the twentieth century and B. Russell's attempts to solve it can be found in [DPP09].

1.10 Exercises

Exercise 1.1 The axiom of pairs defines $\langle a, b\rangle := \big\{\{a\}, \{a, b\}\big\}$; see page 6. Using the axioms of ZFC, show that $\langle a, b\rangle = \langle a', b'\rangle$ iff $a = a'$ and $b = b'$.

Exercise 1.2 Show that $f : A \to B$ is injective iff $f^{-1} : \mathcal{P}(B) \to \mathcal{P}(A)$ is surjective; f is surjective iff f^{-1} is injective.

Exercise 1.3 Define $\leq_d$ on $\mathbb{N}$ as in Example 1.3.2. Show that p is prime iff p is a minimal element of $\mathbb{N} \setminus \{1\}$.

Exercise 1.4 Order $S := \mathcal{P}(\mathbb{N}) \setminus \{\mathbb{N}\}$ by inclusion as in Example 1.4. Show that the set $A := \{\{2 \cdot n, 2 \cdot n + 1\} \mid n \in \mathbb{N}\}$ is bounded in S; does A have a smallest lower bound?

Exercise 1.5 Let S be a set and $H : \mathcal{P}(S) \to \mathcal{P}(S)$ be an order preserving map. Show that $A := \bigcup\{X \in \mathcal{P}(S) \mid X \subseteq H(X)\}$ is a fixed point of H, i.e., that satisfies $H(A) = A$. Moreover, A is the greatest fixed point of H ,i.e., if $H(Y) = Y$, then $Y \subseteq A$.

Exercise 1.6 Let $f : X \to Y$ and $g : Y \to X$ be maps. Using Exercise 1.5, show that there exist disjoint subsets X_1 and X_2 of X and disjoint subsets Y_1 and Y_2 of Y such that $X = X_1 \cup X_2$, $Y = Y_1 \cup Y_2$ and $f[X_1] = Y_1$, $g[Y_2] = X_2$. The map $A \mapsto X \setminus g[Y \setminus f[A]]$ might be helpful.

This decomposition is attributed to *S. Banach*.

Exercise 1.7 Use Exercise 1.6 for a proof of the Schröder–Bernstein Theorem 1.2.1.

Exercise 1.8 Show that there exist for the bijection J from Proposition 1.2.3 surjective maps $K : \mathbb{N}_0 \to \mathbb{N}_0$ and $L : \mathbb{N}_0 \to \mathbb{N}_0$ such that $J\big(K(x), L(x)\big) = x$, $K(x) \leq x$ and $L(x) \leq x$ for all $x \in \mathbb{N}_0$.

Exercise 1.9 Construct a bijection from the power set $\mathcal{P}(\mathbb{N})$ to $\mathbb{R}$ using the Schröder–Bernstein Theorem 1.2.1.

Exercise 1.10 Show using the Schröder–Bernstein Theorem 1.2.1 that the set of all subsets of $\mathbb{N}$ of size exactly 2 is countable. Extend this result by showing that the set of all subsets of $\mathbb{N}$ of size exactly k is countable. Can you show without $(\mathbb{AC})$ that the set of all finite subsets of $\mathbb{N}$ is countable?

Exercise 1.11 Show that $\omega_1 := \{\alpha \mid \alpha \text{ is a countable ordinal}\}$ is an ordinal. Show that ω_1 is not countable.

Exercise 1.12 An undirected graph $\mathcal{G} = (V, E)$ has nodes V and (undirected) edges E. An edge connecting nodes x and y should be written as $\{x, y\}$; note $x \neq y$. A subgraph $\mathcal{G}' = (G', E')$ of $\mathcal{G}$ is a graph with $G' \subseteq G$ and $E' \subseteq E$. $\mathcal{G}$ is *k-colorable* iff there exists a map $c : V \to \{1, \ldots, k\}$ such that $c(x) \neq c(y)$, whenever $\{x, y\} \in E$ is an edge in $\mathcal{G}$. Show that $\mathcal{G}$ is k-colorable iff each of its finite subgraphs is k-colorable.

Exercise 1.13 Let B be a Boolean algebra, and define $a \ominus b := (a \vee b) \wedge -(a \wedge b)$, as in Sect. 1.5.6. Show that $(B, \ominus, \wedge)$ is a commutative ring.

Exercise 1.14 Complete the proof of Proposition 1.5.3 by proving that $(\mathbb{AC}) \Rightarrow (\mathbb{MP})$.

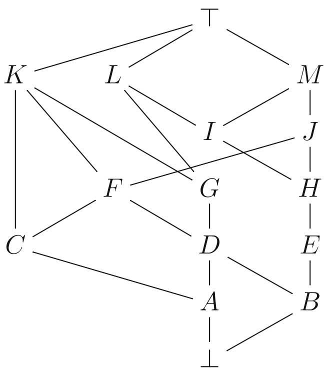
Exercise 1.15 Complete the proof of Lemma 1.5.36.

Exercise 1.16 Using the notation of Sect. 1.5.1, show that $v^* \models \varphi$ iff $\varphi \in \mathcal{M}^*$ using induction on the structure of φ.

Exercise 1.17 Do Exercise 1.12 again, using the Compactness Theorem 1.5.8.

Exercise 1.18 Let $\mathcal{F}$ be an ultrafilter over an infinite set S. Show that if $\mathcal{F}$ contains a finite set, then there exists $s \in S$ such that $\mathcal{F} = \mathcal{F}_s$, the ultrafilter defined by s; see Example 1.5.16.

Exercise 1.19 Consider this ordered set:



Discuss whether these values exist, and determine their values, if they do:

$D \wedge E$;	$D \vee E$;	$(J \wedge L) \wedge K$
$(L \wedge E) \wedge K$;	$L \wedge (E \wedge K)$;	$C \wedge E$
$(C \vee D) \vee E$;	$\sup\{C, D, E\}$;	$C \vee (D \vee E)$
$J \wedge (L \vee K)$;	$(J \wedge L)$ $\vee (J \wedge K)$;	$C \vee G$.

Exercise 1.20 Let L be a lattice. An element $s \in L$ with $s \neq \perp$ is called *join irreducible* iff $s = r \vee t$ implies $s = r$ or $s = t$. Element t *covers* element s iff $s < t$ and if $s < v < t$ for no element v. Show that if L is a finite distributive lattice, then s is join irreducible iff s covers exactly one element.

Exercise 1.21 Let P be a finite partially ordered set. Show that the down set $I \in \mathcal{D}(P)$ is join irreducible iff I is a principal down set.

Exercise 1.22 Identify the join-irreducible elements in $\mathcal{P}(S)$ for $S \neq \emptyset$ and in the lattice of all open intervals $\{]a, b[\mid a \leq b\}$, both ordered by inclusion.

Exercise 1.23 Show that in a lattice one distributive law implies the other one.

Exercise 1.24 Give an example for a down set in a lattice which is not an ideal.

Exercise 1.25 Show that in a distributive lattice $c \wedge x = c \wedge y$ and $c \vee x = x \vee y$ for some c implies $x = y$.

Exercise 1.26 Let $(G, +)$ be a commutative group. Show that the subgroups form a lattice under the subset relation.

Exercise 1.27 Assume that in lattice L there exists for each $a, b \in L$ the relative *pseudo-complement* $a \to b$ of a in b; this is the largest element $x \in L$ such that $a \wedge x \leq b$. Show that a pseudo-complemented lattice is distributive. Furthermore, show that each Boolean algebra is pseudo-complemented. Lattices with pseudo-complements are called *Brouwerian lattices* or *Heyting algebras*.

Exercise 1.28 A lattice is called *complete* iff it contains suprema and infima for arbitrary subsets. Show that a bounded partially ordered set $(L, \leq)$ is a complete lattice if the infimum for arbitrary sets exists. Conclude that the set of all equivalence relations on a set forms a complete lattice under inclusion.

Exercise 1.29 Let L be a complete lattice and $f : L \to L$ monotone. Then the set $\{x \in L \mid f(x) = x\}$ of fix points of f is not empty, and a complete lattice is in the induced order. This is *Tarski's Fixpoint Theorem*.

Exercise 1.30 Let $S \neq \emptyset$ be a set and $a \in S$. Compute for the Boolean algebra $B := \mathcal{P}(S)$ and the ideal $I := \mathcal{P}(S \setminus \{a\})$ the factor algebra B/I.

Exercise 1.31 Show that a topology τ forms a complete Brouwerian lattice under inclusion.

Exercise 1.32 Given a topological space (X, τ), the following conditions are equivalent for all $x, y \in X$:

1. $\{x\}^a \subseteq \{y\}^a$.
2. $x \in \{y\}^a$.
3. $x \in U$ implies $y \in U$ for all open sets U.

Exercise 1.33 Characterize those ideals I in a Boolean algebra B for which the factor algebra B/I consists of exactly two elements.

Exercise 1.34 Let $\emptyset \neq \mathcal{A} \subseteq \mathcal{P}(S)$ be a finite family of sets with $S \in \mathcal{A}$, say $\mathcal{A} = \{A_1, \ldots, A_n\}$. Define $A_T := \bigcap_{i \in T} A_i \cap \bigcap_{i \notin T} S \setminus A_i$ for $T \subseteq \{1, \ldots, n\}$.

1. $\mathcal{P} := \{A_T \mid \emptyset \neq T \subseteq \{1, \ldots, n\}, A_T \neq \emptyset\}$ forms a partition of S.

2. $\{\bigcup \mathcal{P}_0 \mid \mathcal{P}_0 \subseteq \mathcal{P}\}$ is the smallest set algebra over S which contains $\mathcal{A}$.

Exercise 1.35 As in Example 1.6.5 on page 67, let $X := \{0, 1\}^{\mathbb{N}}$ be the space of all infinite sequences. Put

$$\mathcal{C} := \{A \times \prod_{j>k} \{0, 1\} \mid k \in \mathbb{N}, A \in \mathcal{P}\left(\{0, 1\}^k\right)\}.$$

Show that $\mathcal{C}$ is an algebra.

Exercise 1.36 Let X and $\mathcal{C}$ be as in Exercise 1.35. Show that

$$\mu\big(A \times \prod_{j>k} \{0, 1\}\big) := \frac{|A|}{2^k}$$

defines a monotone and countably additive map $\mu : \mathcal{C} \to [0, 1]$ with $\mu(\emptyset) = 0$.

Exercise 1.37 Show that a countably subadditive and monotone set function on a set algebra is additive.

Chapter 2

Categories

Many areas of mathematics show surprising structural similarities, which suggests that it might be interesting and helpful to focus on an abstract view, hereby unifying concepts. This view looks at the mathematical objects from the outside and studies the relationship between them, for example, groups and homomorphisms, or topological spaces together with continuous maps, or ordered sets with monotone maps. The list could be extended. It leads to the general notion of a category. A category is based on a class of objects together with morphisms for each pair of objects. Morphisms can be composed; the composition follows laws which are considered evident and natural.

This approach has considerable appeal to a software engineer as well. In software engineering, the implementation details of a software system are usually not particularly important from an architectural point of view; they are encapsulated in a component. In contrast, the relationship of components with each other is of interest because this knowledge is necessary for composing a system from its components. Roughly speaking, the architecture of a software system is characterized both by its components and their interaction, the static part of which can be described by what we may perceive as morphisms.

This has been recognized fairly early in the software architecture community, witnessed by the April 1995 issue of the *IEEE Transactions on Software Engineering*, which was devoted to software architecture and which introduced some categorical language in discussing architectures.

E.-E. Doberkat, *Special Topics in Mathematics for Computer Scientists*, DOI 10.1007/978-3-319-22750-4_2

So the language of categories offers some attractions to software engineers, as can also be seen from, e.g., [Bar01, Fia05, Dob03]. We will also see that the tool set of modal logics, another area which is important to software construction, profits substantially from constructions which are firmly grounded in categories.

We will discuss categories here and introduce the reader to the basic constructions. The world of categories is too rich to be captured in these few pages, so we have made an attempt to provide the reader with some basic proficiency in categories, helping her or him to get a grasp of the basic techniques. This modest goal is attained by blending the abstract mathematical development with a plethora of examples. We give a brief overview over the contents.

Overview The definition of a category and a discussion of its most elementary properties are found in Sect. 2.1; examples show that categories are indeed a very versatile and general instrument for mathematical modeling. Section 2.2 discusses constructions like products and coproducts, which are familiar from other contexts, in this new language, and we look at pushouts and pullbacks, first in the familiar context of sets and then in a more general setting. Functors are introduced in Sect. 2.3 for relating categories to each other, and natural transformations permit functors to enter into a relationship. We show also that set-valued functors play a special rôle, which provides an opportunity to investigate more deeply the hom sets of a category. Products and coproducts have an appearance again, but this time as instances of the more general concept of limits resp. colimits.

Monads and Kleisli tripels are introduced as very special functors and discussed in Sect. 2.4, their relationship is investigated, and some examples are given, which provide an idea about the usefulness of this concept; a small section on monads in the programming language `Haskell` provides a pointer to the practical use of monads. Next, we show that monads are generated from adjunctions. This important concept is introduced and discussed in Sect. 2.5; we define adjunctions, show by examples that adjunctions are a colorfully blooming and nourished flower in the garden of mathematics, and give an alternative formulation in terms of units and counits; we then show that each adjunction gives us a monad and that each monad also generates an adjunction. The latter part is interesting since it provides an opportunity of introducing the algebras for a monad; we discuss two examples fairly extensively, indicating what such algebras might look like.

While an algebra provides a morphism $\boldsymbol{F}a \to a$, a coalgebra provides a morphism $a \to \boldsymbol{F}a$. This is introduced and discussed in Sect. 2.6; many examples show that this concept models a broad variety of applications in the area of systems. Coalgebras and their properties are studied, among them bisimulations, a concept which originates from the theory of concurrent systems and which is captured now coalgebraically. The Kripke models for modal logics provide an excellent playground for coalgebras, so they are introduced in Sect. 2.7; examples show the broad applicability of this concept (but neighborhood models as a generalization are introduced as well). We go a bit beyond a mere application of coalgebras and give also the construction of the canonical model through Lindenbaum's construction of maximally consistent sets, which, by the way, provide an application of transfinite induction as well. We finally show that coalgebras may be put to use when constructing coalgebraic logics, a very fruitful and general approach to modal logics and their generalizations.

2.1 Basic Definitions

We will define what a category is and give some examples for categories. It shows that this is a very general notion, covering also many formal structures that are studied in theoretical computer science. A very rough description would be to say that a category is a bunch of objects which are related to each other, the relationships being called morphisms. This gives already the gist of the definition—objects which are related to each other. But the relationship has to be made a bit more precise to be amenable for further investigation. So here is the definition of a category.

Definition 2.1.1 *A category $\boldsymbol{K}$ consists of a class $|\boldsymbol{K}|$ of* objects *and for any objects a, b in $|\boldsymbol{K}|$ of a set* $\text{hom}_{\boldsymbol{K}}(a, b)$ *of* morphisms *with a* composition operation $\circ$*, mapping* $\text{hom}_{\boldsymbol{K}}(b, c) \times \text{hom}_{\boldsymbol{K}}(a, b)$ *to* $\text{hom}_{\boldsymbol{K}}(a, c)$ *with the following properties:*

Identity *For every object a in $|\boldsymbol{K}|$, there exists a morphism $id_a \in$* $\text{hom}_{\boldsymbol{K}}(a, a)$ *with $f \circ id_a = f = id_b \circ f$, whenever $f \in$* $\text{hom}_{\boldsymbol{K}}(a, b)$.

Associativity *If* $f \in \mathrm{hom}_{\boldsymbol{K}}(a, b), g \in \mathrm{hom}_{\boldsymbol{K}}(b, c)$*, and* $h \in \mathrm{hom}_{\boldsymbol{K}}(c, d)$*, then* $h \circ (g \circ f) = (h \circ g) \circ f$*.*

Note that we do not think that a category is based on a set of objects (which would yield difficulties) but rather on a class. In fact, if we would insist on having a set of objects, we would not be able to talk about the category of sets, which is an important species for a category. We insist, however, on having *sets* of morphisms, because we want morphisms to be somewhat clearly represented. Usually we write for $f \in \mathrm{hom}_{\boldsymbol{K}}(a, b)$ also $f : a \to b$, if the context is clear. Thus if $f : a \to b$ and $g : b \to c$, then $g \circ f : a \to c$; one may think that first f is applied (or executed), and then g is applied to the result of f. Note the order in which the application is written down: $g \circ f$ means that g is applied to the result of f. The first postulate says that there is an *identity morphism* $id_a : a \to a$ for each object a of $\boldsymbol{K}$ which does not have an effect on the other morphisms upon composition, so no matter if you do id_a first and then morphism $f : a \to b$ or if you do f first and then id_b, you end up with the same result as if doing only f. Associativity is depicted through this diagram:

$$\begin{tikzcd}
a \arrow[r, "f"] \arrow[rr, bend left, dashed, "g\circ f"] & b \arrow[r, "g"] \arrow[rr, bend right, dashed, "h\circ g"'] & c \arrow[r, "h"] & d
\end{tikzcd}$$

Hence if you take the fast train $g \circ f$ from a to c first (no stop at b) and then switch to train h or if you travel first with f from a to b and then change to the fast train $h \circ g$ (no stop at c), you will end up with the same result.

Given $f \in \mathrm{hom}_{\boldsymbol{K}}(a, b)$, we call object a the *domain* and object b the *codomain* of morphism f.

Let us have a look at some examples.

Set **Example 2.1.2** The category ***Set*** is the most important of them all. It has sets as its class of objects, and the morphisms $\mathrm{hom}_{\boldsymbol{Set}}(a, b)$ are just the maps from set a to set b. The identity map $id_a : a \to a$ maps each element to itself, and composition is just composition of maps, which is associative:

$$\big(f \circ (g \circ h)\big)(x) = f\big(g \circ h(x)\big) \quad = f\big(g(h(x))\big)$$
$$= \big(f \circ g\big)(h(x)) = \big((f \circ g) \circ h\big)(x)$$

✌

The next example shows that one class of objects can carry more than one kind of morphisms.

Example 2.1.3 The category ***Rel*** has sets as its class of objects. Given sets a and b, $f \in \text{hom}_{\boldsymbol{Rel}}(a, b)$ is a morphism from a to b iff $f \subseteq a \times b$ is a relation. Given set a, define *Rel*

$$id_a := \{\langle x, x\rangle \mid x \in a\}$$

as the identity relation and define for $f \in \text{hom}_{\boldsymbol{Rel}}(a, b)$, $g \in \text{hom}_{\boldsymbol{Rel}}(b, c)$ the composition as

$$g \circ f := \{\langle x, z\rangle \mid \text{ there exists } y \in b \text{ with } \langle x, y\rangle \in f \text{ and } \langle y, z\rangle \in g\}$$

Because existential quantifiers can be interchanged, composition is associative, and id_a serves in fact as the identity element for composition. ✌

But morphisms do not need to be maps or relations.

Example 2.1.4 Let $(P, \leq)$ be a partially ordered set. Define $\boldsymbol{P}$ by taking the class $|\boldsymbol{P}|$ of objects as P, and put

$$\text{hom}_{\boldsymbol{P}}(p, q) := \begin{cases} \{\langle p, q\rangle\}, & \text{if } p \leq q \\ \emptyset, & \text{otherwise.} \end{cases}$$

Then id_p is $\langle p, p\rangle$, the only element of $\text{hom}_{\boldsymbol{P}}(p, p)$, and if $f : p \to q, g : q \to r$, thus $p \leq q$ and $q \leq r$; hence by transitivity $p \leq r$, so that we put $g \circ f := \langle p, r\rangle$. Let $h : r \to s$, then

$$h \circ (g \circ f) = h \circ \langle p, r\rangle = \langle p, s\rangle = \langle q, s\rangle \circ f = (h \circ g) \circ f$$

It is clear that $id_p = \langle p, p\rangle$ serves as a neutral element. ✌

A directed graph generates a category through all its finite paths. Composition of two paths is then just their combination, indicating movement from one node to another, possibly via intermediate nodes. But we also have to cater to the situation that we want to stay in a node.

Example 2.1.5 Let $\mathcal{G} = (V, E)$ be a directed graph. Recall that a *path* $\langle p_0, \ldots, p_n \rangle$ is a finite sequence of nodes such that adjacent nodes form an edge, i.e., $\langle p_i, p_{i+1} \rangle \in E$ for $0 \leq i < n$; each node a has an empty path $\langle a, a \rangle$ attached to it, which may or may not be an edge in the graph. The objects of the category $F(\mathcal{G})$ are the nodes V of $\mathcal{G}$, and a morphism $a \to b$ in $F(\mathcal{G})$ is a path connecting a with b in $\mathcal{G}$, hence a path $\langle p_0, \ldots, p_n \rangle$ with $p_0 = a$ and $p_n = b$. The empty path serves as the identity Morphism; the composition of morphism is just their concatenation; this is plainly associative. This category is called the *free category generated by graph* $\mathcal{G}$. ✌

Free category

These two examples base categories on a set of objects; they are instances of small categories. A category is called *small* iff the objects form a set (rather than a class).

The discrete category is a trivial but helpful example.

Example 2.1.6 Let $X \neq \emptyset$ be a set, and define a category $\boldsymbol{K}$ through $|\boldsymbol{K}| := X$ with

$$\hom_{\boldsymbol{K}}(x, y) := \begin{cases} \{id_x\}, & x = y \\ \emptyset, & \text{otherwise} \end{cases}$$

This is the *discrete category* on X. ✌

Algebraic structures furnish a plentiful source for examples. Let us have a look at groups and at Boolean algebras.

Example 2.1.7 The category of groups has as objects all groups $(G, \cdot)$ and as morphisms $f : (G, \cdot) \to (H, *)$ all maps $f : G \to H$ which are group homomorphisms, i.e., for which $f(1_G) = 1_H$ (with $1_G, 1_H$ as the respective neutral elements), for which $f(a^{-1}) = (f(a))^{-1}$ and $f(a \cdot b) = f(a) * f(b)$ always hold. The identity morphism $id_{(G,\cdot)}$ is the identity map, and composition of homomorphisms is composition of maps. Because composition is inherited from category ***Set***, we do not have to check for associativity or for identity.

Category of groups

We do not associate the category with a particular name; it is simply referred to as the *category of groups*. ✌

Example 2.1.8 Similarly, the category of Boolean algebras has Boolean algebras as objects, and a morphism $f : G \to H$ for the Boolean algebras G and H is a map f between the carrier sets with these properties:

$$\begin{aligned} f(-a) &= -f(a) \\ f(a \wedge b) &= f(a) \wedge f(b) \\ f(\top) &= \top \end{aligned}$$

(hence also $f(\bot) = \bot$, and $f(a \vee b) = f(a) \vee f(b)$). Again, composition of morphisms is composition of maps, and the identity morphism is just the identity map. ✌

The next example deals with transition systems. Formally, a transition system is a directed graph. But whereas discussing a graph puts the emphasis usually on its paths, a transition system is concerned more with the study of, well, the transition between two states; hence the focus is more strongly localized. This is reflected when defining morphisms, which, as we will see, come in two flavors.

Example 2.1.9 A *transition system* $(S, \rightsquigarrow_S)$ is a set S of states together $(S, \rightsquigarrow_S)$
with a transition relation $\rightsquigarrow_S \subseteq S \times S$. Intuitively, $s \rightsquigarrow_S s'$ iff there is a transition from s to s'. Transition systems form a category: the objects are transition systems, and a morphism $f : (S, \rightsquigarrow_S) \to (T, \rightsquigarrow_T)$ is a map $f : S \to T$ such that $s \rightsquigarrow_S s'$ implies $f(s) \rightsquigarrow_T f(s')$. This means that a transition from s to s' in $(S, \rightsquigarrow_S)$ entails a transition from $f(s)$ to $f(s')$ in the transition system $(T, \rightsquigarrow_T)$. Note that the defining condition for f can be written as $\rightsquigarrow_S \subseteq (f \times f)^{-1}\left[\rightsquigarrow_T\right]$ with $f \times f : \langle s, s'\rangle \mapsto \langle f(s), f(s')\rangle$. ✌

The morphisms in Example 2.1.9 are interesting from a relational point of view. We will require an additional property which, roughly speaking, makes sure that we not only transport transitions through morphisms but that we are also able to capture transitions which emanate from the image of a state. So we want to be sure that we obtain a transition $f(s) \rightsquigarrow_T t$ from a transition arising from s in the original system. This idea is formulated in the next example; it will arise again in a very natural manner in Example 2.6.12 in the context of coalgebras.

Example 2.1.10 We continue with transition systems, so we define a category which has transition systems as objects. A *morphism* $f : (S, \rightsquigarrow_S) \to (T, \rightsquigarrow_T)$ in the present category is a map $f : S \to T$ such that for all $s, s' \in S, t \in T$.

Forward: $s \rightsquigarrow_S s'$ implies $f(s) \rightsquigarrow_T f(s')$.

***Backward*:** if $f(s) \rightsquigarrow_T t'$, then there exists $s' \in S$ with $f(s') = t'$ and $s \rightsquigarrow_S s'$.

The forward condition is already known from Example 2.1.9; the backward condition is new. It states that if we start a transition from some $f(s)$ in T, then this transition originates from some transition starting from s in S; to distinguish these morphisms from the ones considered in Example 2.1.9, they are called *bounded* morphisms. The identity map $S \to S$ yields a bounded morphism, and the composition of bounded morphisms is a bounded morphism again. In fact, let $f : (S, \rightsquigarrow_S) \to (T, \rightsquigarrow_T)$, $g : (T, \rightsquigarrow_T) \to (U, \rightsquigarrow_U)$ be bounded morphisms, and assume that $g(f(s)) \rightsquigarrow_U u'$. Then we can find $t' \in T$ with $g(t') = u'$ and $f(s) \rightsquigarrow_T t'$; hence we find $s' \in S$ with $f(s') = t'$ and $s \rightsquigarrow_S s'$.

Bounded morphisms

Bounded morphisms are of interest in the study of models for modal logics [BdRV01]; see Lemma 2.7.25. ✌

The next examples reverse arrows when it comes to defining morphisms. The examples so far observed the effects of maps in the direction in which the maps were defined. We will, however, also have an opportunity to operate in the backward direction and to see what properties the inverse image of a map is supposed to have.

We study this in the context of topological and of measurable spaces.

Example 2.1.11 Let (S,τ) be a topological space; see Definition 1.5.47; hence $\tau \subseteq \mathcal{P}(S)$ with $\emptyset, S \in \tau$, τ is closed under finite intersections and arbitrary unions. Given another topological space (T, ϑ), a map $f : S \to T$ is called *τ-ϑ-continuous* iff the inverse image of an open set is open again, i.e., iff $f^{-1}\left[G\right] \in \vartheta$ for all $G \in \tau$; this will be discussed in greater depth in Sect. 3.1.1. Category ***Top*** of topological spaces has all topological spaces as objects and continuous maps as morphisms. The identity $(S, \tau) \to (S, \tau)$ is certainly continuous. Again, it follows that the composition of morphisms yields a morphism and that their composition is associative. ✌

Top

The next example deals with σ-algebras, which are of course also sets of subsets. Measurability is formulated similar to continuity in terms of the inverse rather than the direct image.

Example 2.1.12 Let S be a set, and assume that $\mathcal{A}$ is a σ-algebra on S. Then the pair $(S, \mathcal{A})$ is called a *measurable space*; the elements of the σ-algebra are sometimes called $\mathcal{A}$-measurable sets. The category ***Meas*** has as objects all measurable spaces. ***Meas***

Given two measurable spaces $(S, \mathcal{A})$ and $(T, \mathcal{B})$, a map $f : S \to T$ is called a *morphism of measurable spaces* iff f is $\mathcal{A}$-$\mathcal{B}$*-measurable*. This means that $f^{-1}\left[B\right] \in \mathcal{A}$ for all $B \in \mathcal{B}$; hence the set $\{s \in S \mid f(s) \in B\}$ is an $\mathcal{A}$-measurable set for each $\mathcal{B}$-measurable set B. Each σ-algebra is a Boolean algebra, but the definition of a morphism of measurable spaces does not entail that such a morphism induces a morphism of Boolean algebras, as defined in Example 2.1.8. Measurable maps are rather modeled on the prototype of continuous maps, for which the inverse image of an open set is open again. Replace "open" by r"measurable"; then you obtain the definition of a measurable map. Consequently the behavior of f^{-1} rather than the one of f determines whether f belongs to this distinguished set of morphisms.

Thus the $\mathcal{A}$-$\mathcal{B}$-measurable maps $f : S \to T$ are the morphisms $f : (S, \mathcal{A}) \to (T, \mathcal{B})$ in category ***Meas***. The identity morphism on $(S, \mathcal{A})$ is the identity map (this map is measurable because $id^{-1}\left[A\right] = A \in \mathcal{A}$ for each $A \in \mathcal{A}$). Composition of measurable maps yields a measurable map again: let $f : (S, \mathcal{A}) \to (T, \mathcal{B})$ and $g : (T, \mathcal{B}) \to (U, \mathcal{C})$, then $(g \circ f)^{-1}\left[D\right] = f^{-1}\left[g^{-1}\left[D\right]\right] \in \mathcal{A}$, for $D \in \mathcal{C}$, because $g^{-1}\left[D\right] \in \mathcal{B}$. It is clear that composition is associative, since it is based on composition of ordinary maps. ✌

Now that we know what a category is, we begin constructing new categories from given ones. We start by building on category ***Meas*** another interesting category, indicating that a category can be used as a building block for another one.

Example 2.1.13 A measurable space $(S, \mathcal{A})$ together with a probability measure μ on $\mathcal{A}$ (see Definition 1.6.12) is called a *probability space* and written as $(S, \mathcal{A}, \mu)$. The category ***Prob*** of all probability spaces has—you guessed it—as objects all probability spaces; a morphism $f : (S, \mathcal{A}, \mu) \to (T, \mathcal{B}, \nu)$ is a morphism $f : (S, \mathcal{A}) \to (T, \mathcal{B})$ in ***Meas*** for the underlying measurable spaces such that $\nu(B) = \mu(f^{-1}\left[B\right])$ holds for all $B \in \mathcal{B}$. Thus the ν-probability for event $B \in \mathcal{B}$ is the same as the μ-probability for all those $s \in S$, the image of which is in B. Note that ***Prob***

$f^{-1}\left[B\right] \in \mathcal{A}$ due to f being a morphism in ***Meas***, so that $\mu(f^{-1}\left[B\right])$ is in fact defined. ✌

We go a bit further and combine two measurable spaces into a third one; this requires adjusting the notion of a morphism, which are in this new category basically pairs of morphisms from the underlying category. This shows the flexibility with which we may—and do—manipulate morphisms.

$\mathbb{P}(S, \mathcal{A})$

Example 2.1.14 Denote for the measurable space $(S, \mathcal{A})$ by $\mathbb{P}(S, \mathcal{A})$ the set of all subprobability measures. Define

$$\begin{aligned} \boldsymbol{\beta}_{\mathcal{A}}(A, r) &:= \{\mu \in \mathbb{P}(S, \mathcal{A}) \mid \mu(A) \geq r\}, \\ \wp(X, \mathcal{A}) &:= \wp(\mathcal{A}) := \sigma\big(\{\boldsymbol{\beta}_{\mathcal{A}}(A, r) \mid A \in \mathcal{A}, 0 \leq r \leq 1\}\big). \end{aligned}$$

$\beta_S(A, r)$, $\wp(\mathcal{A})$

Thus $\boldsymbol{\beta}_{\mathcal{A}}(A, r)$ denotes all probability measures which evaluate set A not smaller than r, and $\wp(X, \mathcal{A})$ collects all these sets into a σ-algebra; $\wp(X, \mathcal{A})$ is called the *weak σ-algebra* associated with $\mathcal{A}$ as the σ-algebra generated by the family of sets; its elements are sometimes called *weakly measurable sets*. We will usually omit the carrier set from the notation. This renders $\big(\mathbb{P}(S, \mathcal{A}), \wp(\mathcal{A})\big)$ a measurable space.

Let $(T, \mathcal{B})$ be another measurable space. A map $K : S \to \mathbb{P}(T, \mathcal{B})$ is $\mathcal{A}$-$\wp(\mathcal{B})$-measurable iff $\{s \in S \mid K(s)(B) \geq r\} \in \mathcal{A}$ for all $B \in \mathcal{B}$; this follows from Exercise 2.7. We take as objects for our category the triplets $\big((S, \mathcal{A}), (T, \mathcal{B}), K\big)$, where $(S, \mathcal{A})$ and $(T, \mathcal{B})$ are measurable spaces and $K : S \to \mathbb{P}(T, \mathcal{B})$ is $\mathcal{A}$-$\wp(\mathcal{B})$-measurable. A morphism

$$(f, g) : \big((S, \mathcal{A}), (T, \mathcal{B}), K\big) \to \big((S', \mathcal{A}'), (T', \mathcal{B}'), K'\big)$$

is a pair of morphisms

$$f : (S, \mathcal{A}) \to (S', \mathcal{A}') \text{ and } g : (T, \mathcal{B}) \to (T', \mathcal{B}')$$

such that

$$K(s)(g^{-1}\left[B'\right]) = K'(f(s))(B')$$

holds for all $s \in S$ and for all $B' \in \mathcal{B}'$.

The composition of morphisms is defined component wise:

$$(f', g') \circ (f, g) := (f' \circ f, g' \circ g).$$

Note that $f' \circ f$ and $g' \circ g$ refer to the composition of maps, while $(f', g') \circ (f, g)$ refers to the newly defined composition in our new spic-and-span category (we should probably use another symbol, but no confusion can arise, since the new composition operates on pairs). The identity morphism for $\big((S, \mathcal{A}), (T, \mathcal{B}), K\big)$ is just the pair (id_S, id_T). Because the composition of maps is associative, composition in our new category is associative as well, and because (id_S, id_T) is composed from identities, it is also an identity.

Category of stochastic relations

This category is sometimes called the *category of stochastic relation*s. ✌

Before continuing, we introduce commutative diagrams. Suppose that we have in a category $\boldsymbol{K}$ morphisms $f : a \to b$ and $g : b \to c$. The combined morphism $g \circ f$ is represented graphically as

$$\begin{tikzcd}
a \arrow[rr, "f"] \arrow[dr, "g\circ f"'] & & b \arrow[dl, "g"] \\
& c &
\end{tikzcd}$$

If the morphisms $h : a \to d$ and $\ell : d \to c$ satisfy $g \circ f = \ell \circ h$, we have a *commutative diagram*; in this case, we do not draw out the morphism in the diagonal.

$$\begin{tikzcd}
a \arrow[r, "f"] \arrow[d, "h"'] \arrow[dr, dotted, "g\circ f", "\ell\circ h"'] & b \arrow[d, "g"] \\
d \arrow[r, "\ell"'] & c
\end{tikzcd}$$

We consider automata next, to get some feeling for the handling of commutative diagrams and as an illustration for an important formalism looked at through the glasses of categories.

Example 2.1.15 Given sets X and S of inputs and states, respectively, an *automaton* (X, S, δ) is defined by a map $\delta : X \times S \to S$. The interpretation is that $\delta(x, s)$ is the new state after input $x \in X$ in state $s \in S$. Reformulating, $\delta(x) : s \mapsto \delta(x, s)$ is perceived as a map $S \to S$ for each $x \in X$, so that the new state now is written as $\delta(x)(s)$; manipulating a map with two arguments in this way is called *currying* and will be examined in Example 2.5.2. The objects of our category of automata are the automata, and an *automaton morphism* $f : (X, S, \delta) \to (X, S', \delta')$ is a map $f : S \to S'$ such that this diagram commutes for all $x \in X$:

$$\begin{tikzcd} S \arrow[r, "f"] \arrow[d, "\delta(x)"'] & S' \arrow[d, "\delta'(x)"] \\ S \arrow[r, "f"'] & S' \end{tikzcd}$$

Hence we have $f(\delta(x)(s)) = \delta'(x)(f(s))$ for each $x \in X$ and $s \in S$ (or, in the old notation, $f\big(\delta(x, s)\big) = \delta'\big(x, f(s)\big)$); this means that computing the new state and mapping it through f yields the same result as computing the new state for the mapped one. The identity map $S \to S$ yields a morphism; hence automata with these morphisms form a category.

Note that morphisms are defined only for automata with the same input alphabet. This reflects the observation that the input alphabet is usually given by the environment, while the set of states represents a model about the automata's behavior and, hence, is at our disposal for manipulation. ✌

Whereas we constructed the above new categories from the given one in an ad hoc manner, categories also yield new categories systematically. This is a simple example.

Example 2.1.16 Let $\boldsymbol{K}$ be a category; fix an object x on $\boldsymbol{K}$. The objects of our new category are the morphisms $f \in \mathrm{hom}_{\boldsymbol{K}}(a, x)$ for an object a. Given objects $f \in \mathrm{hom}_{\boldsymbol{K}}(a, x)$ and $g \in \mathrm{hom}_{\boldsymbol{K}}(b, x)$ in the new category, a morphism $\varphi : f \to g$ is a morphism $\varphi \in \mathrm{hom}_{\boldsymbol{K}}(a, b)$ with $f = g \circ \varphi$, so that this diagram commutes

$$\begin{tikzcd} a \arrow[rr, "\varphi"] \arrow[dr, "f"'] & & b \arrow[dl, "g"] \\ & x & \end{tikzcd}$$

Composition is inherited from $\boldsymbol{K}$. The identity $id_f : f \to f$ is $id_a \in \mathrm{hom}_{\boldsymbol{K}}(a, a)$, provided $f \in \mathrm{hom}_{\boldsymbol{K}}(a, x)$. Since the composition in $\boldsymbol{K}$ is associative, we have only to make sure that the composition of two morphisms is a morphism again. This can be read off the following diagram: $(\varphi \circ \psi) \circ h = \varphi \circ (\psi \circ h) = \varphi \circ g = f$.

$$\begin{tikzcd}
a \arrow[r, "\varphi"] \arrow[rr, bend left, dashed, "\psi\circ\varphi"] \arrow[dr, "f"'] & b \arrow[r, "\psi"] \arrow[d, "g"] & c \arrow[dl, "h"] \\
& x &
\end{tikzcd}$$

This category is sometimes called the *slice category* $\boldsymbol{K}/x$; the object x is interpreted as an index, so that a morphism $f : a \to x$ serves as an indexing function. A morphism $\varphi : a \to b$ in $\boldsymbol{K}/x$ is then compatible with the index operation. ✌ $\boldsymbol{K}/x$

The next example reverses arrows while at the same time maintaining the same class of objects.

Example 2.1.17 Let $\boldsymbol{K}$ be a category. We define $\boldsymbol{K}^{op}$, the category *dual* to $\boldsymbol{K}$, in the following way: the objects are the same as for the original category; hence $|\boldsymbol{K}^{op}| = |\boldsymbol{K}|$, and the arrows are reversed; hence we put $\text{hom}_{\boldsymbol{K}^{\text{op}}}(a, b) := \text{hom}_{\boldsymbol{K}}(b, a)$ for the objects a, b; the identity remains the same. We have to define composition in this new category. Let $f \in \text{hom}_{\boldsymbol{K}^{\text{op}}}(a, b)$ and $g \in \text{hom}_{\boldsymbol{K}^{\text{op}}}(b, c)$; then $g * f := f \circ g \in \text{hom}_{\boldsymbol{K}^{\text{op}}}(a, c)$. It is readily verified that $*$ satisfies all the laws for composition from Definition 2.1.1. $\boldsymbol{K}^{op}$

The dual category is sometimes helpful because it permits to cast notions into a uniform framework. ✌

Example 2.1.18 Let us look at $\boldsymbol{Rel}$ again. The morphisms $\text{hom}_{\boldsymbol{Rel}^{\text{op}}}(S, T)$ from S to T in $\boldsymbol{Rel}^{op}$ are just the morphisms $\text{hom}_{\boldsymbol{Rel}}(T, S)$ in $\boldsymbol{Rel}$. Take $f \in \text{hom}_{\boldsymbol{Rel}^{\text{op}}}(S, T)$ and then $f \subseteq T \times S$; hence $f^t \subseteq S \times T$, where relation

$$f^t := \{\langle s, t\rangle \mid \langle t, s\rangle \in f\}$$

is the transposed of relation f. The map $f \mapsto f^t$ is injective and compatible with composition; moreover, it maps $\text{hom}_{\boldsymbol{Rel}^{\text{op}}}(S, T)$ onto $\text{hom}_{\boldsymbol{Rel}}(T, S)$. But this means that $\boldsymbol{Rel}^{op}$ is essentially the same as $\boldsymbol{Rel}$. ✌

It is sometimes helpful to combine two categories into a product:

Lemma 2.1.19 *Given categories* $\boldsymbol{K}$ *and* $\boldsymbol{L}$*, define the objects of* $\boldsymbol{K} \times \boldsymbol{L}$ *as pairs* $\langle a, b\rangle$*, where* a *is an object in* $\boldsymbol{K}$ *and* b *is an object in* $\boldsymbol{L}$*. A* $\boldsymbol{K} \times \boldsymbol{L}$

morphism $\langle a, b\rangle \to \langle a', b'\rangle$ in $\mathbf{K}\times\mathbf{L}$ is comprised of morphisms $a \to a'$ in $\mathbf{K}$ and $b \to b'$ in $\mathbf{L}$. Then $\mathbf{K}\times\mathbf{L}$ is a category. $\dashv$

We have a closer look at morphisms now. Experience tells us that injective and surjective maps are important, so a characterization in a category might be desirable. There is a small but not insignificant catch, however. We have seen that morphisms are not always maps, so that we are forced to find a characterization purely in terms of composition and equality, because this is all we have in a category. The following characterization of injective maps provides a clue for a more general definition.

Proposition 2.1.20 *Let $f : X \to Y$ be a map; then these statements are equivalent.*

1. *f is injective.*
2. *If A is an arbitrary set, $g_1, g_2 : A \to X$ are maps with $f \circ g_1 = f \circ g_2$, then $g_1 = g_2$.*

Proof 1 $\Rightarrow$ 2: Assume f is injective and $f \circ g_1 = f \circ g_2$, but $g_1 \neq g_2$. Thus there exists $x \in A$ with $g_1(x) \neq g_2(x)$. But $f(g_1(x)) = f(g_2(x))$, and since f is injective, $g_1(x) = g_2(x)$. This is a contradiction.

2 $\Rightarrow$ 1: Assume the condition holds, but f is not injective. Then there exists $x_1 \neq x_2$ with $f(x_1) = f(x_2)$. Let $A := \{*\}$ and put $g_1(*) := x_1, g_2(*) := x_2$; thus $f(x_1) = (f \circ g_1)(*) = (f \circ g_2)(*) = f(x_2)$. By the condition $g_1 = g_2$, thus $x_1 = x_2$. Another contradiction. $\dashv$

This leads to a definition of the category version of injectivity as a morphism which is cancelable on the left.

Definition 2.1.21 *Let $\mathbf{K}$ be a category, a, b objects in $\mathbf{K}$. Then $f : a \to b$ is called a* monomorphism *(or a* mono*) iff whenever $g_1, g_2 : x \to a$ are morphisms with $f \circ g_1 = f \circ g_2$; then $g_1 = g_2$.*

Mono

These are some simple properties of monomorphisms, which are also sometimes called monos.

Lemma 2.1.22 *In a category $\mathbf{K}$:*

1. *The identity is a monomorphism.*
2. *The composition of two monomorphisms is a monomorphism again.*

3. *If $k \circ f$ is a monomorphism for some morphism k, then f is a monomorphism.*

Proof The first part is trivial. Let $f : a \to b$ and $g : b \to c$ both monos. Assume $h_1, h_2 : x \to a$ with $h_1 \circ (g \circ f) = h_2 \circ (g \circ f)$. We want to show $h_1 = h_2$. By associativity $(h_1 \circ g) \circ f = (h_2 \circ g) \circ f$. Because f is a mono, we conclude $h_1 \circ g = h_2 \circ g$; because g is a mono, we see $h_1 = h_2$.

Finally, let $f : a \to b$ and $k : b \to c$. Assume $h_1, h_2 : x \to a$ with $f \circ h_1 = f \circ h_2$. We claim $h_1 = h_2$. Now $f \circ h_1 = f \circ h_2$ implies $k \circ f \circ h_1 = k \circ f \circ h_2$. Thus $h_1 = h_2$. ⊣

In the same way, we characterize surjectivity purely in terms of composition, exhibiting a nice symmetry between the two notions.

Proposition 2.1.23 *Let $f : X \to Y$ be a map; then these statements are equivalent.*

1. *f is surjective.*
2. *If B is an arbitrary set, $g_1, g_2 : Y \to B$ are maps with $g_1 \circ f = g_2 \circ f$, then $g_1 = g_2$.*

Proof 1 $\Rightarrow$ 2: Assume f is surjective, $g_1 \circ f = g_2 \circ f$, but $g_1(y) \neq g_2(y)$ for some y. If we can find $x \in X$ with $f(x) = y$, then $g_1(y) = (g_1 \circ f)(x) = (g_2 \circ f)(x) = g_2(y)$, which would be a contradiction. Thus $y \notin f[X]$; hence f is not onto.

2 $\Rightarrow$ 1: Assume that there exists $y \in Y$ with $y \notin f[X]$. Define $g_1, g_2 : Y \to \{0, 1, 2\}$ through

$$g_1(y) := \begin{cases} 0, & \text{if } y \in f[X], \\ 1, & \text{otherwise.} \end{cases} \qquad g_2(y) := \begin{cases} 0, & \text{if } y \in f[X], \\ 2, & \text{otherwise.} \end{cases}$$

Then $g_1 \circ f = g_2 \circ f$, but $g_1 \neq g_2$. This is a contradiction. ⊣

This suggests a definition of surjectivity through a morphism which is right cancelable.

Definition 2.1.24 *Let $\mathbf{K}$ be a category, a, b objects in $\mathbf{K}$. Then $f : a \to b$ is called a* epimorphism *(or an* epi*) iff whenever $g_1, g_2 : b \to c$ are morphisms with $g_1 \circ f = g_2 \circ f$; then $g_1 = g_2$.* Epi

These are some important properties of epimorphisms, which are sometimes called epis:

Lemma 2.1.25 *In a category* $\boldsymbol{K}$*:*

1. *The identity is an epimorphism.*
2. *The composition of two epimorphisms is an epimorphism again.*
3. *If* $f \circ k$ *is an epimorphism for some morphism* k*, then* f *is an epimorphism.*

Proof We sketch the proof only for the third part:

$$g_1 \circ f = g_2 \circ f \Rightarrow g_1 \circ f \circ k = g_2 \circ f \circ k \Rightarrow g_1 = g_2.$$

$\dashv$

This is a small application of the decomposition of a map into an epimorphism and a monomorphism.

Proposition 2.1.26 *Let* $f : X \to Y$ *be a map. Then there exists a factorization of* f *into* $m \circ e$ *with* e *an epimorphism and* m *a monomorphism.*

$$\begin{tikzcd} X \arrow[rr, "f"] \arrow[dr, "e"'] & & Y \\ & M \arrow[ur, "m"'] & \end{tikzcd}$$

The idea of the proof may best be described in terms of X as inputs and Y as outputs of system f. We collect all inputs with the same functionality, and assign each collection the functionality through which it is defined.

Proof Define

$$\ker(f) := \{\langle x_1, x_2 \rangle \mid f(x_1) = f(x_2)\}$$

$\ker(f)$ (the *kernel* $\ker(f)$ of f). This is an equivalence relation on X:

- *reflexivity*: $\langle x, x \rangle \in \ker(f)$ for all x,
- *symmetry*: if $\langle x_1, x_2 \rangle \in \ker(f)$, then $\langle x_2, x_1 \rangle \in \ker(f)$;
- *transitivity*: $\langle x_1, x_2 \rangle \in \ker(f)$ and $\langle x_2, x_3 \rangle \in \ker(f)$ together imply $\langle x_1, x_3 \rangle \in \ker(f)$.

Define

$$e : \begin{cases} X & \rightarrow X/\ker(f), \\ x & \mapsto [x]_{\ker(f)} \end{cases}$$

then e is an epimorphism. In fact, if $g_1 \circ e = g_2 \circ e$ for $g_1, g_2 : X/\ker(f) \rightarrow B$ for some set B, then $g_1(t) = g_2(t)$ for all $t \in X/\ker(f)$; hence $g_1 = g_2$.

Moreover,

$$m : \begin{cases} X/\ker(f) & \rightarrow Y \\ [x]_{\ker(f)} & \mapsto f(x) \end{cases}$$

is well defined, since if $[x]_{\ker(f)} = [x']_{\ker(f)}$, then $f(x) = f(x')$. It is also a monomorphism. In fact, if $m \circ g_1 = m \circ g_2$ for arbitrary $g_1, g_2 : A \rightarrow X/\ker(f)$ for some set A, then $f(g_1(a)) = f(g_2(a))$ for all a; hence $\langle g_1(a), g_2(a)\rangle \in \ker(f)$. But this means $[g_1(a)]_{\ker(f)} = [g_2(a)]_{\ker(f)}$ for all $a \in A$, so $g_1 = g_2$. Evidently, $f = m \circ e$. $\dashv$

Looking a bit harder at the diagram, we find that we can say even more, viz., that the decomposition is unique up to isomorphism.

Corollary 2.1.27 *If the map $f : X \rightarrow Y$ can be written as $f = e \circ m = e' \circ m'$ with epimorphisms e, e' and monomorphisms m, m', then there is a bijection b with $e' = b \circ e$ and $m = m' \circ b$.*

Proof Since the composition of bijections is a bijection again, we may and do assume without loss of generality that $e : X \rightarrow X/\ker(f)$ maps x to its class $[x]_{\ker(f)}$ and that $m : X/\ker(f) \rightarrow Y$ maps $[x]_{\ker(f)}$ to $f(x)$. Then we have this diagram for the primed factorization $e' : X \rightarrow Z$ and $m' : Z \rightarrow Y$:

$$\begin{tikzcd}
X \arrow[rr, "f"] \arrow[dr, "e"] \arrow[ddr, bend right, "e'"'] & & Y \\
& X/\ker(f) \arrow[ur, "m"] \arrow[d, dotted, leftrightarrow, "b"] & \\
& Z \arrow[uur, bend right, "m'"'] &
\end{tikzcd}$$

Note that

$$\begin{aligned}
[x]_{\ker(f)} \neq [x']_{\ker(f)} &\Leftrightarrow f(x) \neq f(x') \\
&\Leftrightarrow m'(e'(x)) \neq m'(e'(x')) \\
&\Leftrightarrow e(x) \neq e(x')
\end{aligned}$$

Thus defining $b([x]_{\ker(f)}) := e'(x)$ gives an injective map $X/\ker(f) \to Z$. Given $z \in Z$, there exists $x \in X$ with $e'(x) = z$; hence $b([x]_{\ker(f)}) = z$; thus b is onto. Finally, $m'(b([x]_{\ker(f)})) = m'(e'(x)) = f(x) = m([x]_{\ker(f)})$. $\dashv$

This factorization of a morphism is called an *epi/mono factorization*, and we just have shown that such a factorization is unique up to isomorphisms (a.k.a. bijections in ***Set***).

The following example shows that epimorphisms are not necessarily surjective, even if they are maps.

Example 2.1.28 Recall that $(M, *)$ is a *monoid* iff $* : M \times M \to M$ is associative with a neutral element 0_M. For example, $(\mathbb{Z}, +)$ and $(\mathbb{N}, \cdot)$ are monoids, so is the set X^* of all strings over alphabet X with concatenation as composition and the empty string as neutral element. A *morphism* $f : (M, *) \to (N, \ddagger)$ is a map $f : M \to N$ such that $f(a * b) = f(a) \ddagger f(b)$, and $f(0_M) = 0_N$.

Now let $f : (\mathbb{Z}, +) \to (N, \ddagger)$ be a morphism; then f is uniquely determined by the value $f(1)$. This is so since $m = 1 + \ldots + 1$ (m times) for $m > 0$; thus $f(m) = f(1 + \ldots + 1) = f(1) \ddagger \ldots \ddagger f(1)$. Also $f(-1) \ddagger f(1) = f(-1+1) = f(0)$, so $f(-1)$ is inverse to $f(1)$; hence $f(-m)$ is inverse to $f(m)$. Consequently, if two morphisms map 1 to the same value, then the morphisms are identical.

Note that the inclusion $i : x \mapsto x$ is a morphism $i : (\mathbb{N}_0, +) \to (\mathbb{Z}, +)$. We claim that i is an epimorphism. Let $g_1 \circ i = g_2 \circ i$ for some morphisms $g_1, g_2 : (\mathbb{Z}, +) \to (M, *)$. Then $g_1(1) = (g_1 \circ i)(1) = (g_2 \circ i)(1) = g_2(1)$. Hence $g_1 = g_2$. Thus epimorphisms are not necessarily surjective. ✌

Composition induces maps between the hom sets of a category, which we are going to study now. Specifically, let $\boldsymbol{K}$ be a fixed category, take objects a and b, and fix for the moment a morphism $f : a \to b$. Then $g \mapsto f \circ g$ maps $\hom_{\boldsymbol{K}}(x, a)$ to $\hom_{\boldsymbol{K}}(x, b)$, and $h \mapsto h \circ f$ maps $\hom_{\boldsymbol{K}}(b, x)$ to $\hom_{\boldsymbol{K}}(a, x)$ for each object x. We investigate $g \mapsto f \circ g$ first. Define for an object x of $\boldsymbol{K}$ the map

$\hom_{\boldsymbol{K}}(x, \cdot)$

$$\hom_{\boldsymbol{K}}(x, f) : \begin{cases} \hom_{\boldsymbol{K}}(x, a) & \to \hom_{\boldsymbol{K}}(x, b) \\ g & \mapsto f \circ g \end{cases}$$

Then $\hom_{\mathbf{K}}(x, f)$ defines a map between morphisms, and we can determine through this map whether or not f is a monomorphism.

Lemma 2.1.29 *$f : a \to b$ is a monomorphism iff $\hom_{\mathbf{K}}(x, f)$ is injective for all objects x.*

Proof This follows immediately from the observation

$$f \circ g_1 = f \circ g_2 \Leftrightarrow \hom_{\mathbf{K}}(x, f)(g_1) = \hom_{\mathbf{K}}(x, f)(g_2).$$

$\dashv$

Dually, define for an object x of $\mathbf{K}$ the map $\hom_{\mathbf{K}}(\cdot, x)$

$$\hom_{\mathbf{K}}(f, x) : \begin{cases} \hom_{\mathbf{K}}(b, x) & \to \hom_{\mathbf{K}}(a, x) \\ g & \mapsto g \circ f \end{cases}$$

Note that we change directions here: $f : a \to b$ corresponds to $\hom_{\mathbf{K}}(b, x) \to \hom_{\mathbf{K}}(a, x)$. Note also that we did reuse the name $\hom_{\mathbf{K}}(\cdot, \cdot)$; but no confusion should arise, because the signature tells us which map we specifically have in mind. Lemma 2.1.29 seems to suggest that surjectivity of $\hom_{\mathbf{K}}(f, x)$ and f being an epimorphism are related. This, however, is not the case. But try this:

Lemma 2.1.30 *$f : a \to b$ is an epimorphism iff $\hom_{\mathbf{K}}(f, x)$ is injective for each object x.*

Proof $\hom_{\mathbf{K}}(f, x)(g_1) = \hom_{\mathbf{K}}(f, x)(g_2)$ is equivalent to $g_1 \circ f = g_2 \circ f$. $\dashv$

Not surprisingly, an isomorphism is an invertible morphism; this is described in our scenario as follows.

Definition 2.1.31 *$f : a \to b$ is called an* isomorphism *iff there exists a morphism $g : b \to a$ such that $g \circ f = id_a$ and $f \circ g = id_b$.*

It is clear that morphism g is in this case uniquely determined: let g and g' be morphisms with the property above; then we obtain $g = g \circ id_b = g \circ (f \circ g') = (g \circ f) \circ g' = id_a \circ g' = g'$.

When we are in the category ***Set*** of sets with maps, an isomorphism f is bijective. In fact, let g be chosen to f according to Definition 2.1.31, then

$$\begin{aligned} h_1 \circ f = h_2 \circ f \quad &\Rightarrow h_1 \circ f \circ g = h_2 \circ f \circ g \quad &&\Rightarrow h_1 = h_2, \\ f \circ g_1 = f \circ g_2 \quad &\Rightarrow g \circ f \circ g_1 = g \circ f \circ g_2 \quad &&\Rightarrow g_1 = g_2, \end{aligned}$$

so that the first line makes f an epimorphism, and the second one a monomorphism.

The following lemma is often helpful (and serves as an example of the popular art of *diagram chasing*).

Lemma 2.1.32 *Assume that in this diagram*

$$\begin{tikzcd}
a \arrow[r, "f"] \arrow[d, "k"'] & b \arrow[r, "g"] \arrow[d, "\ell"'] & c \arrow[d, "m"] \\
x \arrow[r, "r"'] & y \arrow[r, "s"'] & z
\end{tikzcd}$$

the outer diagram commutes, that the leftmost diagram commutes, and that f is an epimorphism. Then the rightmost diagram commutes as well.

Proof In order to show that $m \circ g = s \circ \ell$, it is enough to show that $m \circ g \circ f = s \circ \ell \circ f$, because we then can cancel f, since f is an epi. But now

$$\begin{aligned}
(m \circ g) \circ f &= m \circ (g \circ f) \\
&= (s \circ r) \circ k && \text{(commutativity of the outer diagram)} \\
&= s \circ (r \circ k) \\
&= s \circ (\ell \circ f) && \text{(commutativity of the leftmost diagram)} \\
&= (s \circ \ell) \circ f
\end{aligned}$$

Now cancel f. $\dashv$

2.2 Elementary Constructions

In this section, we deal with some elementary constructions, showing mainly how some important constructions for sets can be carried over to categories, hence are available in more general structures. Specifically, we will study products and sums (coproducts) as well as pullbacks and pushouts. We will not study more general constructs at present; in particular we will not have a look at limits and colimits. Once products and pullbacks are understood, the step to limits should not be too complicated, similarly for colimits, as the reader can see in the brief discussion in Sect. 2.3.3.

We fix a category $\boldsymbol{K}$.

2.2.1 Products and Coproducts

The Cartesian product of two sets is just the set of pairs. In a general category, we do not have a characterization through sets and their elements at our disposal, so we have to fill this gap by going back to morphisms. Thus we require a characterization of the product through morphisms. The first thought is using the projections $\langle x, y\rangle \mapsto x$ and $\langle x, y\rangle \mapsto y$, since a pair can be reconstructed through its projections. But this is not specific enough. An additional characterization of the projections is obtained through factoring: if there is another pair of maps pretending to be projections, they better be related to the "genuine" projections. This is what the next definition expresses.

Definition 2.2.1 *Given objects a and b in $\boldsymbol{K}$. An object c is called the* product *of a and b iff:*

1. *there exist morphisms $\pi_a : c \to a$ and $\pi_b : c \to b$,*
2. *for each object d and morphisms $\alpha_a : d \to a$ and $\alpha_b : d \to b$, there exists a unique morphism $\rho : d \to c$ such that $\alpha_a = \pi_a \circ \rho$ and $\alpha_b = \pi_b \circ \rho$.*

Morphisms π_a and π_b are called projections *to a resp. b.*

Thus α_a and α_b factor uniquely through π_a and π_b. Note that we insist on having a unique factor and that the factor should be the same for both pretenders. We will see in a minute why this is a sensible assumption. If it exists, the product of objects a and b is denoted by $a \times b$; the projections π_a and π_b are usually understood and not mentioned explicitly.

This diagram depicts the situation:

$$\begin{tikzcd} & d \arrow[dl, "\alpha_a"'] \arrow[d, dotted, "\rho"', "!"] \arrow[dr, "\alpha_b"] & \\ a & c \arrow[l, "\pi_a"] \arrow[r, "\pi_b"'] & b \end{tikzcd} \qquad \text{(PRODUCT)}$$

Lemma 2.2.2 *If the product of two objects exists, it is unique up to isomorphism.*

Proof Let a and b be the objects in question; also assume that c_1 and c_2 are products with morphisms $\pi_{i,a} \to a$ and $\pi_{i,b} \to b$ as the corresponding morphisms, $i = 1, 2$.

Because c_1 together with $\pi_{1,a}$ and $\pi_{1,b}$ is a product, we find a unique morphism $\xi : c_2 \to c_1$ with $\pi_{2,a} = \pi_{1,a} \circ \xi$ and $\pi_{2,b} = \pi_{1,b} \circ \xi$; similarly, we find a unique morphism $\zeta : c_1 \to c_2$ with $\pi_{1,a} = \pi_{2,a} \circ \zeta$ and $\pi_{1,b} = \pi_{2,b} \circ \zeta$.

$$\begin{tikzcd}
& c_1 \arrow[dl, "\pi_{1,a}"'] \arrow[dr, "\pi_{1,b}"] \arrow[dd, dotted, bend left, "\zeta"] & \\
a & & b \\
& c_2 \arrow[ul, "\pi_{2,a}"] \arrow[ur, "\pi_{2,b}"'] \arrow[uu, dotted, bend left, "\xi"] &
\end{tikzcd}$$

Now look at $\xi \circ \zeta$: We obtain

$$\pi_{1,a} \circ \xi \circ \zeta = \pi_{2,a} \circ \zeta = \pi_{1,a}$$
$$\pi_{1,b} \circ \xi \circ \zeta = \pi_{2,b} \circ \zeta = \pi_{1,b}$$

Then uniqueness of the factorization implies that $\xi \circ \zeta = id_{c_1}$; similarly, $\zeta \circ \xi = id_{c_2}$. Thus ξ and ζ are isomorphisms. $\dashv$

Let us have a look at some examples, first and foremost sets.

Example 2.2.3 Consider the category ***Set*** with maps as morphisms. Given sets A and B, we claim that $A \times B$ together with the projections $\pi_A : \langle a, b\rangle \mapsto a$ and $\pi_B : \langle a, b\rangle \mapsto b$ constitute the product of A and B in ***Set***. In fact, if $\vartheta_A : D \to A$ and $\vartheta_B : D \to B$ are maps for some set D, then $\rho : d \mapsto \langle \vartheta_A(d), \vartheta_B(d)\rangle$ satisfies the equations $\vartheta_A = \pi_A \circ \rho$, $\vartheta_B = \pi_B \circ \rho$, and it is clear that this is the only way to factor, so ρ is uniquely determined. ✌

If sets carry an additional structure, this demands additional attention.

Example 2.2.4 Let $(S, \mathcal{A})$ and $(T, \mathcal{B})$ be measurable spaces, so we are now in the category ***Meas*** of measurable spaces with measurable maps as morphisms; see Example 2.1.12. For constructing a product, one is tempted to take the product $S \times T$ as ***Set*** and to find a suitable σ-algebra $\mathcal{C}$ on $S \times T$ such that the projections π_S and π_T become measurable. Thus $\mathcal{C}$ would have to contain $\pi_S^{-1}\big[A\big] = A \times T$ and $\pi_T^{-1}\big[B\big] = S \times B$ for each $A \in \mathcal{A}$ and each $B \in \mathcal{B}$. Because a σ-algebra is closed under intersections, $\mathcal{C}$ would have to contain all measurable rectangles $A \times B$ with sides in $\mathcal{A}$ and $\mathcal{B}$. So let us try this:

$$\mathcal{C} := \sigma\big(\{A \times B \mid A \in \mathcal{A}, B \in \mathcal{B}\}\big).$$

Then clearly, $\pi_S : (S \times T, \mathcal{C}) \to (S, \mathcal{A})$ and $\pi_T : (S \times T, \mathcal{C}) \to (T, \mathcal{B})$ are morphisms in ***Meas***. Now let $(D, \mathcal{D})$ be a measurable space with morphisms $\vartheta_S : D \to S$ and $\vartheta_T : D \to T$, and define ρ as above through $\rho(d) := \langle \vartheta_S(d), \vartheta_T(d) \rangle$. We claim that ρ is a morphism in ***Meas***. It has to be shown that $\rho^{-1}\left[C\right] \in \mathcal{D}$ for all $C \in \mathcal{C}$. We have a look at all elements of $\mathcal{C}$ for which this is true, and we define

$$\mathcal{G} := \{C \in \mathcal{C} \mid \zeta^{-1}\left[C\right] \in \mathcal{D}\}.$$

We plan to use the principle of good sets from page 86. If we can show that $\mathcal{G} = \mathcal{C}$, we are done. It is evident that $\mathcal{G}$ is a σ-algebra, because the inverse image of a map respects countable Boolean operations. Moreover, if $A \in \mathcal{A}$ and $B \in \mathcal{B}$, then $\rho^{-1}\left[A \times B\right] = \vartheta_S^{-1}\left[A\right] \cap \vartheta_T^{-1}\left[B\right] \in \mathcal{D}$, so that $A \times B \in \mathcal{G}$, provided $A \in \mathcal{A}$, $B \in \mathcal{B}$. But now we have

$$\mathcal{C} = \sigma(\{A \times B \mid A \in \mathcal{A}, B \in \mathcal{B}\}) \subseteq \mathcal{G} \subseteq \mathcal{C}.$$

Hence each element of $\mathcal{C}$ is a member of $\mathcal{G}$; thus ρ is $\mathcal{D}$-$\mathcal{C}$-measurable. Again, the construction shows that there is no other possibility for defining ρ. Hence we have shown that two objects in the category ***Meas*** of measurable spaces with measurable maps have a product.

The σ-algebra $\mathcal{C}$ which is constructed above is usually denoted by $\mathcal{A} \otimes \mathcal{B}$ $\mathcal{A} \otimes \mathcal{B}$
and called the *product σ-algebra* of $\mathcal{A}$ and $\mathcal{B}$. ✌

The next example requires a forward reference to the construction of the product measure in Sect. 4.9. I suggest that you skip them on the first reading; just to make things easier, I have marked them with a special symbol.

Example 2.2.5 Ⓢ While the category ***Meas*** has products, the situation changes when taking probability measures into account; hence when changing to the category ***Prob*** of probability spaces, see Example 2.1.13. The product measure $\mu \otimes \nu$ of two probability measures μ on σ-algebra $\mathcal{A}$ resp. ν on $\mathcal{B}$ is the unique probability measure on the product σ-algebra $\mathcal{A} \otimes \mathcal{B}$ with $(\mu \otimes \nu)(A \times B) = \mu(A) \cdot \nu(B)$ for $A \in \mathcal{A}$ and $B \in \mathcal{B}$, in particular

$$\pi_S : (S \times T, \mathcal{A} \otimes \mathcal{B}, \mu \otimes \nu) \to (S, \mathcal{A}, \mu),$$
$$\pi_T : (S \times T, \mathcal{A} \otimes \mathcal{B}, \mu \otimes \nu) \to (T, \mathcal{B}, \nu)$$

are morphisms in ***Prob***.

Now define $S := T := [0,1]$ and take in each case the smallest σ-algebra which is generated by the open intervals as a σ-algebra; hence put $\mathcal{A} := \mathcal{B} := \mathcal{B}([0,1])$; λ is Lebesgue measure on $\mathcal{B}([0,1])$. Define

$$\kappa(E) := \lambda(\{x \in [0,1] \mid \langle x, x\rangle \in E\})$$

for $E \in \mathcal{A} \otimes \mathcal{B}$ (well, we have to show that $\{x \in [0,1] \mid \langle x, x\rangle \in E\} \in \mathcal{B}([0,1])$, whenever $E \in \mathcal{A} \otimes \mathcal{B}$. This is relegated to Exercise 2.10). Then

$$\pi_S : (S \times T, \mathcal{A} \otimes \mathcal{B}, \kappa) \to (S, \mathcal{A}, \lambda)$$
$$\pi_T : (S \times T, \mathcal{A} \otimes \mathcal{B}, \kappa) \to (T, \mathcal{B}, \lambda)$$

are morphisms in ***Prob***, because

$$\kappa(\pi_S^{-1}\left[G\right]) = \kappa(G \times T) = \lambda(\{x \in [0,1] \mid \langle x, x\rangle \in G \times T\}) = \lambda(G)$$

for $G \in \mathcal{B}(S)$. If we could find a morphism $f : (S \times T, \mathcal{A} \otimes \mathcal{B}, \kappa) \to (S \times T, \mathcal{A} \otimes \mathcal{B}, \lambda \otimes \lambda)$ factoring through the projections, f would have to be the identity; thus it would imply that $\kappa = \lambda \otimes \lambda$, but this is not the case: take $E := [1/2, 1] \times [0, 1/3]$; then $\kappa(E) = 0$, but $(\lambda \otimes \lambda)(E) = 1/6$.

Thus we conclude that the category ***Prob*** of probability spaces does not have products. ✌

The product topology on the Cartesian product of the carrier sets of topological spaces is familiar, open sets in the product just contain open rectangles. The categorical view is that of a product in the category of topological spaces.

Example 2.2.6 Let (T, τ) and (S, ϑ) be topological spaces, and equip the Cartesian product $S \times T$ with the product topology $\tau \times \vartheta$. This is the smallest topology on $S \times T$ which contains all the open rectangles $G \times H$ with $G \in \tau$ and $H \in \vartheta$. We claim that this is a product in the category ***Top*** of topological spaces. In fact, the projections $\pi_S : S \times T \to S$ and $\pi_T : S \times T \to T$ are continuous, because, e.g, $\pi_S^{-1}\left[G\right] = G \times T \in \tau \times \vartheta$. Now let (D, ρ) be a topological space with continuous maps $\xi_S : D \to S$ and $\xi_T : D \to T$, and define $\zeta : D \to S \times T$ through $\zeta : d \mapsto \langle \xi_S(d), \xi_T(d)\rangle$. Then $\zeta^{-1}\left[G \times H\right] = \xi_S^{-1}\left[G\right] \cap \xi_T^{-1}\left[H\right] \in \rho$, and since the inverse image of a topology under a map is a topology again, $\zeta : (D, \mathcal{D}) \to (S \times T, \tau \times \vartheta)$ is continuous. Again, this is the only way to define a morphism ζ so that $\xi_S = \pi_S \circ \zeta$ and $\xi_T = \pi_T \circ \zeta$. ✌

The category coming from a partially ordered set from Example 2.1.4 is investigated next.

Example 2.2.7 Let $(P, \leq)$ be a partially ordered set, considered as a category $\boldsymbol{P}$. Let $a, b \in P$, and assume that a and b have a product x in $\boldsymbol{P}$. Thus there exist morphisms $\pi_a : x \to a$ and $\pi_b : x \to b$, which means by the definition of this category that $x \leq a$ and $x \leq b$ hold and hence that x is a lower bound to $\{a, b\}$. Moreover, if y is such that there are morphisms $\tau_a : y \to a$ and $\tau_b : y \to b$, then there exists a unique $\sigma : y \to x$ with $\tau_a = \pi_a \circ \sigma$ and $\tau_b : \pi_b \circ \sigma$. Translated into $(P, \leq)$, this means that if $y \leq a$ and $y \leq b$, then $y \leq x$ (morphisms in $\boldsymbol{P}$ are unique, if they exist). Hence the product x is just the greatest lower bound of $\{a, b\}$.

So the product corresponds to the infimum. This example demonstrates again that products do not necessarily exist in a category and, if they exist, are not always what one would expect. ✌

Given morphisms $f : x \to a$ and $g : x \to b$, and assuming that the product $a \times b$ exists, we want to "lift" f and g to the product; i.e., we want to find a morphism $h : x \to a \times b$ with $f = \pi_a \circ h$ and $g = \pi_b \circ h$. Let us see how this is done in ***Set***: Here, $f : X \to A$ and $g : X \to B$ are maps, and one defines the lifted map $h : X \to A \times B$ through $h : x \mapsto \langle f(x), g(x) \rangle$, so that the conditions on the projections are satisfied. The next lemma states that this is always possible in a unique way.

Lemma 2.2.8 *Assume that the product $a \times b$ exists for the objects a and b. Let $f : x \to a$ and $g : x \to b$ be morphisms. Then there exists a unique morphism $q : x \to a \times b$ such that $f = \pi_a \circ q$ and $g = \pi_b \circ q$. Morphism q is denoted by $f \times g$.*

Proof The diagram looks like this:

$$\begin{tikzcd}
 & x \arrow[dl, "f"'] \arrow[d, "q"', "!"] \arrow[dr, "g"] & \\
a & a \times b \arrow[l, "\pi_a"] \arrow[r, "\pi_b"'] & b
\end{tikzcd}$$

Because $f : x \to a$ and $g : x \to b$, there exists a unique $q : x \to a \times b$ with $f = \pi_a \circ q$ and $g = \pi_b \circ q$. This follows from the definition of the product. ⊣

Let us look at the product through our $\mathrm{hom}_{\boldsymbol{K}}$-glasses. If $a \times b$ exists, and if $\zeta_a : d \to a$ and $\zeta_b : d \to b$ are morphisms, we know that there is a *unique* $\xi : d \to a \times b$ rendering this diagram commutative

$$\begin{tikzcd} & d \arrow[dl, "\zeta_a"'] \arrow[d, "\xi"', "!"] \arrow[dr, "\zeta_b"] & \\ a & a \times b \arrow[l, "\pi_a"] \arrow[r, "\pi_b"'] & b \end{tikzcd}$$

Thus the map

$$p_d : \begin{cases} \hom_{\boldsymbol{K}}(d, a) \times \hom_{\boldsymbol{K}}(d, b) & \to \hom_{\boldsymbol{K}}(d, a \times b) \\ \langle \zeta_a, \zeta_b \rangle & \mapsto \xi \end{cases}$$

is well defined. In fact, we can say more.

Proposition 2.2.9 *p_d is a bijection.*

Proof Assume $\xi = p_d(f, g) = p_d(f', g')$. Then $f = \pi_a \circ \xi = f'$ and $g = \pi_b \circ \xi = g'$. Thus $\langle f, g \rangle = \langle f', g' \rangle$. Hence p_d is injective. Similarly, one shows that p_d is surjective: Let $h \in \hom_{\boldsymbol{K}}(d, a \times b)$, then $\pi_a \circ h : d \to a$ and $\pi_b \circ h : d \to b$ are morphisms, so there exists a unique $h' : d \to a \times b$ with $\pi_a \circ h' = \pi_a \circ h$ and $\pi_b \circ h' = \pi_b \circ h$. Uniqueness implies that $h = h'$, so h occurs in the image of p_d. $\dashv$

Let us consider the construction dual to the product.

Definition 2.2.10 *Given objects a and b in category $\boldsymbol{K}$, the object s together with morphisms $i_a : a \to s$ and $i_b : b \to s$ is called the* coproduct*(or the* sum*) of a and b iff for each object t with morphisms $j_a : a \to t$ and $j_b : b \to t$ there exists a unique morphism $r : s \to t$ such that $j_a = r \circ i_a$ and $j_b = r \circ i_b$. Morphisms i_a and i_b are called* injections*; the coproduct of a and b is denoted by $a + b$.*

This is the corresponding diagram:

$$\begin{tikzcd} & t & \\ a \arrow[ur, "j_a"] \arrow[r, "i_a"'] & s \arrow[u, dotted, "r", "!"'] & b \arrow[ul, "j_b"'] \arrow[l, "i_b"] \end{tikzcd} \qquad \text{(COPRODUCT)}$$

These are again some simple examples.

Example 2.2.11 Let $(P, \leq)$ be a partially ordered set, and consider category $\boldsymbol{P}$, as in Example 2.2.7. The coproduct of the elements a and b is just the supremum $\sup\{a, b\}$. This is shown with exactly the same arguments which have been used in Example 2.2.7 for showing that the product of two elements corresponds to their infimum. ✌

And then there is of course the category ***Set***.

Example 2.2.12 Let A and B be disjoint sets. Then $S := A \cup B$ together with

$$i_A : \begin{cases} A & \to S \\ a & \mapsto a \end{cases} \qquad\qquad i_B : \begin{cases} B & \to S \\ b & \mapsto b \end{cases}$$

form the coproduct of A and B. In fact, if T is a set with maps $j_A : A \to T$ and $j_B : B \to T$, then define

$$r : \begin{cases} S & \to T \\ s & \mapsto j_A(a), \text{ if } s = i_A(a), \\ s & \mapsto j_B(b), \text{ if } s = i_B(b) \end{cases}$$

Then $j_A = r \circ i_A$ and $j_B = r \circ i_B$, and these definitions are the only possible ones.

Note that we needed for this construction to work disjointness of the participating sets. Consider, for example, $A := \{-1, 0\}$, $B := \{0, 1\}$, and let $T := \{-1, 0, 1\}$ with $j_A(x) := -1$, $j_B(x) := +1$. No matter where we embed A and B, we cannot factor j_A and j_B uniquely.

If the sets are not disjoint, we first do a preprocessing step and embed them, so that the embedded sets are disjoint. The injections have to be adjusted accordingly. So the following construction would work: Given sets A and B, define $S := \{\langle a, 1\rangle \mid a \in A\} \cup \{\langle b, 2\rangle \mid b \in B\}$ with $i_A : a \mapsto \langle a, 1\rangle$ and $b \mapsto \langle b, 2\rangle$. Note that we do not take a product like $S \times \{1\}$, but rather use a very specific construction; this is so since the product is determined uniquely only by isomorphism, so we might not have gained anything by using that product. Of course, one has to be sure that the sum is not dependent in an essential way on this embedding. ✌

The question of uniqueness is answered through this observation. It relates the coproduct in $\boldsymbol{K}$ to the product in the dual category $\boldsymbol{K}^{op}$ (see Example 2.1.17).

Proposition 2.2.13 *The coproduct s of objects a and b with injections $i_a : a \to s$ and $i_b : b \to s$ in category $\mathbf{K}$ is the product in category $\mathbf{K}^{op}$ with projections $i_a : s \to^{op} a$ and $i_s : s \to^{op} b$.*

Proof Revert in diagram (COPRODUCT) on page 134 to obtain diagram (PRODUCT) on page 129. ⊣

Corollary 2.2.14 *If the coproduct of two objects in a category exists, it is unique up to isomorphisms.*

Proof Proposition 2.2.13 together with Lemma 2.2.2. ⊣

Let us have a look at the coproduct for topological spaces.

Example 2.2.15 Given topological spaces (S, τ) and (T, ϑ), we may and do assume that S and T are disjoint. Otherwise, wrap the elements of the sets accordingly; put

$$A^\dagger := \{\langle a, 1\rangle \mid a \in A\},$$
$$B^\ddagger := \{\langle b, 2\rangle \mid b \in B\},$$

and consider the topological spaces $(S^\dagger, \{G^\dagger \mid G \in \tau\})$ and $(T^\ddagger, \{H^\ddagger \mid H \in \vartheta\})$ instead of (S, τ) and (T, ϑ). Define on the coproduct $S + T$ of S and T in $\mathbf{Set}$ with injections i_S and i_T the topology

$$\tau + \vartheta := \{W \subseteq S + T \mid i_S^{-1}\big[W\big] \in \tau \text{ and } i_T^{-1}\big[W\big] \in \vartheta\}.$$

This is a topology: Both $\emptyset$ and $S + T$ are members of $\tau + \vartheta$, and since τ and ϑ are topologies, $\tau + \vartheta$ is closed under finite intersections and arbitrary unions. Moreover, both $i_S : (S, \tau) \to (S + T, \tau + \vartheta)$ and $i_T : (T, \vartheta) \to (S + T, \tau + \mathcal{H})$ are continuous; in fact, $\tau + \vartheta$ is the smallest topology on $S + T$ with this property.

Now assume that $j_S : (S, \tau) \to (R, \rho)$ and $j_T : (T, \vartheta) \to (R, \rho)$ are continuous maps, and let $r : S + T \to R$ be the unique map determined by the coproduct in $\mathbf{Set}$. Would it not be nice if r would be continuous? Actually, it is. Let $W \in \rho$ be open in R, then $i_S^{-1}\big[r^{-1}\big[W\big]\big] = (r \circ i_S)^{-1}\big[W\big] = j_S^{-1}\big[W\big] \in \tau$; similarly, $i_S^{-1}\big[r^{-1}\big[W\big]\big] \in \vartheta$; thus by definition, $r^{-1}\big[W\big] \in \tau + \vartheta$. Hence we have found the factorization $j_S = r \circ i_S$ and $j_T = r \circ i_T$ in the category $\mathbf{Top}$. This factorization is unique, because it is inherited from the unique factorization in $\mathbf{Set}$. Hence we have shown that $\mathbf{Top}$ has finite coproducts. ✌

A similar construction applies to the category of measurable spaces.

Example 2.2.16 Let $(S, \mathcal{A})$ and $(T, \mathcal{B})$ be measurable spaces; we may assume again that the carrier sets S and T are disjoint. Take the injections $i_S : S \to S + T$ and $i_T : T \to S + T$ from ***Set***. Then

$$\mathcal{A} + \mathcal{B} := \{W \subseteq S + T \mid i_S^{-1}\big[W\big] \in \mathcal{A} \text{ and } i_T^{-1}\big[W\big] \in \mathcal{B}\}$$

is a σ-algebra and $i_S : (S, \mathcal{A}) \to (S + T, \mathcal{A} + \mathcal{B})$ and $i_T : (T, \mathcal{B}) \to (S + T, \mathcal{A} + \mathcal{B})$ are measurable. The unique factorization property is established in exactly the same way as in ***Top***. ✌

Example 2.2.17 Let us consider the category ***Rel*** of relations, which is based on sets as objects. If S and T are sets, we again may and do assume that they are disjoint. Then $S + T = S \cup T$ together with the injections

$$\begin{aligned} I_S &:= \{\langle s, i_S(s)\rangle \mid s \in S\}, \\ I_T &:= \{\langle t, i_T(t)\rangle \mid t \in T\} \end{aligned}$$

form the coproduct, where i_S and i_T are the injections into $S + T$ from ***Set***. In fact, we have to show that we can find for given relations $q_S \subseteq S \times D$ and $q_T \subseteq T \times D$ a unique relation $Q \subseteq (S + T) \times D$ with $q_S = I_S \circ Q$ and $q_T = I_T \circ Q$. The choice is fairly straightforward: Define

$$Q := \{\langle i_S(s), q\rangle \mid \langle s, q\rangle \in q_S\} \cup \{\langle i_T(t), q\rangle \mid \langle t, q\rangle \in q_T\}.$$

Thus

$$\begin{aligned} &\langle s, q\rangle \in I_S \circ Q \Leftrightarrow \text{there exists } x \text{ with } \langle s, x\rangle \in I_S \text{ and } \langle x, q\rangle \in Q \\ &\Leftrightarrow \langle s, q\rangle \in q_S. \end{aligned}$$

Hence $q_S = I_S \circ Q$, similarly, $q_T = I_T \circ Q$. It is clear that no other choice is possible.

Consequently, the coproduct is the same as in ***Set***. ✌

We have just seen in a simple example that dualizing, i.e., going to the dual category, is very helpful. Instead of proving directly that the coproduct is uniquely determined up to isomorphism, if it exists, we turned to the dual category and reused the already established result that the product is uniquely determined, casting it into a new context. The duality, however, is a purely structural property; it usually does not help us with specific constructions. This became apparent when we constructed the coproduct of two sets; it did not help here at all that we knew how to construct the product of two sets, even though product and coproduct are intimately related through dualization. We will make the same observation when we deal with pullbacks and pushouts.

2.2.2 Pullbacks and Pushouts

Sometimes, one wants to complete the square, as in the diagram below on the left-hand side:

$$
\begin{tikzcd}
 & b \arrow[d, "g"] \\
a \arrow[r, "f"'] & c
\end{tikzcd}
\qquad
\begin{tikzcd}
d \arrow[r, dashed, "i_2"] \arrow[d, dashed, "i_1"'] & b \arrow[d, "g"] \\
a \arrow[r, "f"'] & c
\end{tikzcd}
$$

Hence one wants to find an object d together with morphisms $i_1 : d \to a$ and $i_2 : d \to b$ rendering the diagram on the right-hand side commutative. This completion should be as coarse as possible in the following sense. If we have another object, say, e with morphisms $j_1 : e \to a$ and $j_2 : e \to b$ such that $f \circ j_1 = g \circ j_2$, then we want to be able to uniquely factor through i_1 and i_2.

This is captured in the following definition.

Definition 2.2.18 *Let $f : a \to c$ and $g : b \to c$ be morphisms in **K** with the same codomain. An object d together with morphisms $i_1 : d \to a$ and $i_2 : d \to b$ is called a* pullback *of f and g iff:*

1. $f \circ i_1 = g \circ i_2$,
2. *If e is an object with morphisms $j_1 : e \to a$ and $j_2 : e \to b$ such that $f \circ j_1 = g \circ j_2$, then there exists a unique morphism $h : e \to d$ such that $j_1 = i_1 \circ h$ and $j_2 = i_2 \circ h$.*

If we postulate the existence of the morphism $h : e \to d$, but do not insist on its uniqueness, then d with i_1 and i_2 is called a weak pullback.

A diagram for a pullback looks like this:

$$\begin{tikzcd}
e \arrow[dr, dotted, "h", "!"'] \arrow[drr, dashed, bend left, "j_2"] \arrow[ddr, dashed, bend right, "j_1"'] & & \\
& d \arrow[r, dashed, "i_2"] \arrow[d, dashed, "i_1"'] & b \arrow[d, "g"] \\
& a \arrow[r, "f"'] & c
\end{tikzcd}$$

It is clear that a pullback is unique up to isomorphism; this is shown in exactly the same way as in Lemma 2.2.2. Let us have a look at ***Set*** as an important example to get a first impression on the inner workings of a pullback.

Example 2.2.19 Let $f : X \to Z$ and $g : Y \to Z$ be maps. We claim that

$$P := \{\langle x, y\rangle \in X \times Y \mid f(x) = g(y)\}$$

together with the projections $\pi_X : \langle x, y\rangle \mapsto x$, and $\pi_Y : \langle x, y\rangle \mapsto y$ is a pullback for f and g.

Let $\langle x, y\rangle \in P$, then

$$(f \circ \pi_X)(x, y) = f(x) = g(y) = (g \circ \pi_Y)(x, y),$$

so that the first condition is satisfied. Now assume that $j_X : T \to X$ and $j_Y : T \to Y$ satisfy $f(j_X(t)) = g(j_Y(t))$ for all $t \in T$. Thus $\langle j_X(t), j_Y(t)\rangle \in P$ for all t, and defining $r(t) := \langle j_X(t), j_Y(t)\rangle$, we obtain $j_X = \pi_X \circ r$ and $j_Y = \pi_Y \circ r$. Moreover, this is the only possibility to define a factor map with the desired property.

An interesting special case occurs for $X = Y$ and $f = g$. Then $P = \ker(f)$, so that the kernel of a map occurs as a pullback in category ***Set***.
✌

As an illustration for the use of a pullback construction, look at this simple statement.

Lemma 2.2.20 *Assume that d with morphisms $i_a : d \to a$ and $i_b : d \to b$ is a pullback for $f : a \to c$ and $g : b \to c$. If g is a mono, so is i_a.*

Proof Let $g_1, g_2 : e \to d$ be morphisms with $i_a \circ g_1 = i_a \circ g_2$. We have to show that $g_1 = g_2$ holds. If we know that $i_b \circ g_1 = i_b \circ g_2$, we may use the definition of a pullback and capitalize on the uniqueness of the factorization. But let us see.

From $i_a \circ g_1 = i_b \circ g_2$, we conclude $f \circ i_a \circ g_1 = f \circ i_b \circ g_2$, and because $f \circ i_a = g \circ i_b$, we obtain $g \circ i_b \circ g_1 = g \circ i_b \circ g_2$. Since g is a mono, we may cancel on the left of this equation, and we obtain, as desired, $i_b \circ g_1 = i_b \circ g_2$.

But since we have a pullback, there exists a unique $h : e \to d$ with $i_a \circ g_1 = i_a \circ h \ (= i_a \circ g_2)$ and $i_b \circ g_1 = i_b \circ h \ (= i_b \circ g_2)$. We see that the morphisms g_1, g_2, and h have the same properties with respect to factoring, so they must be identical by uniqueness. Hence $g_1 = h = g_2$, and we are done. ⊣

This is another simple example for the use of a pullback in ***Set***.

Example 2.2.21 Let R be an equivalence relation on a set X with projections $\pi_1 : \langle x_1, x_2 \rangle \mapsto x_1$; the second projection $\pi_2 : R \to X$ is defined similarly. Then

$$\begin{tikzcd} R \arrow[r, "\pi_2"] \arrow[d, "\pi_1"'] & X \arrow[d, "\eta_R"] \\ X \arrow[r, "\eta_R"'] & X/R \end{tikzcd}$$

(with $\eta_R : x \mapsto [x]_R$) is a pullback diagram. In fact, the diagram commutes. Let $\alpha, \beta : M \to X$ be maps with $\alpha \circ \eta_R = \beta \circ \eta_R$; thus $[\alpha(m)]_R = [\beta(m)]_R$ for all $m \in M$; hence $\langle \alpha(m), \beta(m) \rangle \in R$ for all m. The only map $\vartheta : M \to R$ with $\alpha = \pi_1 \circ \vartheta$ and $\beta = \pi_2 \circ \vartheta$ is $\vartheta(m) := \langle \alpha(m), \beta(m) \rangle$. ✌

Pullbacks are compatible with products in a sense which we will make precise in a moment. Before we do that, however, we need an auxiliary statement:

Lemma 2.2.22 *Assume that the products $a \times a'$ and $b \times b'$ exist in category **K**. Given morphisms $f : a \to b$ and $f' : a' \to b'$, there exists a unique morphism $f \times f' : a \times a' \to b \times b'$ such that*

$$\pi_b \circ f \times f' = f \circ \pi_a$$
$$\pi_{b'} \circ f \times f' = f' \circ \pi_{a'}$$

Proof Apply the definition of a product to the morphisms $f \circ \pi_a : a \times a' \to b$ and $f' \circ \pi_{a'} : a \times a' \to b'$. $\dashv$

The morphism $f \times f'$ constructed in the lemma renders both parts of this diagram commutative.

$$\begin{tikzcd} a \arrow[d, "f"'] & a \times a' \arrow[l, "\pi_a"'] \arrow[r, "\pi_{a'}"] \arrow[d, "f \times f'"'] & a' \arrow[d, "f'"] \\ b & b \times b' \arrow[l, "\pi_b"] \arrow[r, "\pi_{b'}"'] & b' \end{tikzcd}$$

Denoting this morphism as $f \times f'$, we note that $\times$ is overloaded for morphisms; a look at domains and codomains indicates without ambiguity, however, which version is intended.

Quite apart from its general interest, this is what we need Lemma 2.2.22 for.

Lemma 2.2.23 *Assume that we have these pullbacks:*

$$\begin{tikzcd} a \arrow[r, "f"] \arrow[d, "g"'] & b \arrow[d, "h"] \\ c \arrow[r, "k"'] & d \end{tikzcd} \qquad \begin{tikzcd} a' \arrow[r, "f'"] \arrow[d, "g'"'] & b' \arrow[d, "h'"] \\ c' \arrow[r, "k'"'] & d' \end{tikzcd}$$

Then this is a pullback diagram as well:

$$\begin{tikzcd} a \times a' \arrow[r, "f \times f'"] \arrow[d, "g \times g'"'] & b \times b' \arrow[d, "h \times h'"] \\ c \times c' \arrow[r, "k \times k'"'] & d \times d' \end{tikzcd}$$

Proof 1. We show first that the diagram commutes. It is sufficient to compute the projections. From uniqueness equality will follow. Allora:

$$\begin{aligned} \pi_d \circ (h \times h') \circ (f \times f') &= (h \circ \pi_a) \circ (f \times f') = h \circ f \circ \pi_a \\ \pi_d \circ (k \times k') \circ (g \times g') &= k \circ \pi_c \circ (g \times g') = k \circ g \circ \pi_a \\ &= h \circ f \circ \pi_a. \end{aligned}$$

A similar computation is carried out for $\pi_{a'}$.

2. Let $j : t \to c \times c'$ and $\ell : t \to b \times b'$ be morphisms such that $(k \times k') \circ j = (h \times h') \circ \ell$; then we claim that there exists a unique morphism $r : t \to a \times a'$ such that $j = (g \times g') \circ r$ and $\ell = (f \times f') \circ r$.

The plan is to obtain r from the projections and then show that this morphism is unique.

3. We show that this diagram commutes

$$
\begin{tikzcd}
t \arrow[drr, "\pi_b \circ \ell"] \arrow[ddr, "\pi_c \circ j"'] & & \\
& a \arrow[r, dotted, no head] \arrow[d, dotted, no head] & b \arrow[d, "h"] \\
& c \arrow[r, "k"'] & d
\end{tikzcd}
$$

We have

$$
\begin{aligned}
k \circ (\pi_c \circ j) = (k \circ \pi_c) \circ j \qquad & = (\pi_d \circ k \times k') \circ j \\
= \pi_d \circ (k \times k' \circ j) & \overset{(\ddagger)}{=} \pi_d \circ (h \times h' \circ \ell) \\
= (\pi_d \circ h \times h') \circ \ell & = (h \circ \pi_b) \circ \ell \\
= h \circ (\pi_b \circ \ell) &
\end{aligned}
$$

In $(\ddagger)$, we use Lemma 2.2.22. Using the primed part of Lemma 2.2.22, we obtain $k' \circ (\pi_{c'} \circ j) = h' \circ (\pi_{b'} \circ \ell)$.

Because the left-hand side in the assumption is a pullback diagram, there exists a unique morphism $\rho : t \to a$ with $\pi_c \circ j = g \circ \rho, \pi_b \circ \ell = f \circ \rho$. Similarly, there exists a unique morphism $\rho' : t \to a'$ with $\pi_{c'} \circ j = g' \circ \rho', \pi_{b'} \circ \ell = f' \circ \rho'$.

4. Put $r := \rho \times \rho'$, then $r : t \to a \times a'$, and we have this diagram:

$$
\begin{tikzcd}
& a' \arrow[r, "g'"] & c' \\
t \arrow[ur, "\rho'"] \arrow[r, "\rho \times \rho'"] \arrow[dr, "\rho"'] & a \times a' \arrow[u, "\pi_{a'}"] \arrow[r, "g \times g'"] \arrow[d, "\pi_a"] & c \times c' \arrow[u, "\pi_{c'}"] \arrow[d, "\pi_c"] \\
& a \arrow[r, "g"'] & c
\end{tikzcd}
$$

Hence

$$
\begin{aligned}
\pi_c \circ (g \times g') \circ (\rho \times \rho') &= g \circ \pi_a \circ (\rho \times \rho') = g \circ \rho = \pi_c \circ j, \\
\pi_{c'} \circ (g \times g') \circ (\rho \times \rho') &= g' \circ \pi_{a'} \circ (\rho \times \rho') = g' \circ \rho' = \pi_{c'} \circ j.
\end{aligned}
$$

Because a morphism into a product is uniquely determined by its projections, we conclude that $(g \times g') \circ (\rho \times \rho') = j$. Similarly, we obtain $(f \times f') \circ (\rho \times \rho') = \ell$.

5. Thus $r = \rho \times \rho'$ can be used for factoring; it remains to be shown that this is the only possible choice. In fact, let $\sigma : t \to a \times a'$ be a morphism with $(g \times g') \circ \sigma = j$ and $(f \times f') \circ \sigma = \ell$; then it is enough to show that $\pi_a \circ \sigma$ has the same properties as ρ and that $\pi_{a'} \circ \sigma$ has the same properties as ρ'. Calculating the composition with g resp. f, we obtain

$$\begin{aligned} g \circ \pi_a \circ \sigma &= \pi_c \circ (g \times g') \circ \sigma &&= \pi_c \circ j, \\ f \circ \pi_a \circ \sigma &= \pi_d \circ (f \times f') \circ \sigma &&= \pi_b \circ \ell. \end{aligned}$$

This implies $\pi_a \circ \sigma = \rho$ by uniqueness of ρ; the same argument implies $\pi_{a'} \circ \sigma = \rho'$. But this means $\sigma = \rho \times \rho'$, and uniqueness is established. $\dashv$

Let us dualize. The pullback was defined so that the upper left corner of a diagram is filled in an essentially unique way; the dual construction will have to fill the lower right corner of a diagram in the same way. But by reversing arrows, we convert a diagram in which the lower right corner is missing into a diagram without an upper left corner:

$$\begin{tikzcd} a \arrow[r] \arrow[d] & c & \Rightarrow & & c \arrow[d] \\ b & & & b \arrow[r] & a \end{tikzcd}$$

The corresponding construction is called a pushout.

Definition 2.2.24 *Let $f : a \to b$ and $g : a \to c$ be morphisms in category $\boldsymbol{K}$ with the same domain. An object d together with morphisms $p_b : b \to d$ and $p_c : c \to d$ is called the* pushout *of f and g iff these conditions are satisfied:*

1. $p_b \circ f = p_c \circ g$
2. *if $q_b : b \to e$ and $q_c : c \to e$ are morphisms such that $q_b \circ f = q_c \circ g$, then there exists a unique morphism $h : d \to e$ such that $q_b = h \circ p_b$ and $q_c = h \circ q_c$.*

This diagram obviously looks like this:

$$\begin{tikzcd} a \arrow[r, "g"] \arrow[d, "f"'] & b \arrow[d, "p_b"] \arrow[ddr, dashed, bend left, "q_b"] & \\ c \arrow[r, "p_c"] \arrow[drr, dashed, bend right, "q_c"'] & d \arrow[dr, dotted, "!", "h"'] & \\ & & e \end{tikzcd}$$

It is clear that the pushout of $f \in \text{hom}_{\boldsymbol{K}}(a, b)$ and $g \in \text{hom}_{\boldsymbol{K}}(a, c)$ is the pullback of $f \in \text{hom}_{\boldsymbol{K}^{\text{op}}}(b, a)$ and of $g \in \text{hom}_{\boldsymbol{K}^{\text{op}}}(c, a)$ in the dual category. This, however, does not really provide assistance when constructing a pushout. Let us consider specifically the category ***Set*** of sets with maps as morphisms. We know that dualizing a product yields a sum, but it is not quite clear how to proceed further. The next example tells us what to do.

Example 2.2.25 We are in the category ***Set*** of sets with maps as morphisms now. Consider maps $f : A \to B$ and $g : A \to C$. Construct on the sum $B + C$ the smallest equivalence relation R which contains $R_0 := \{\langle (i_B \circ f)(a), (i_C \circ g)(a) \rangle \mid a \in A\}$. Here, i_B and i_C are the injections of B resp. C into the sum. Let D be the factor space $(A + B)/R$ with $p_B : b \mapsto [i_B(b)]_R$ and $p_C : c \mapsto [i_C(C)]_R$. The construction yields $p_B \circ f = p_C \circ g$, because R identifies the embedded elements $f(a)$ and $g(a)$ for any $a \in A$.

Now assume that $q_B : B \to E$ and $q_C : C \to E$ are maps with $q_B \circ f = q_C \circ g$. Let $q : D \to E$ be the unique map with $q \circ i_B = q_B$ and $q \circ i_C = q_C$ (Lemma 2.2.8 together with Proposition 2.2.13). Then $R_0 \subseteq \ker(q)$: Let $a \in A$, then

$$q(i_B(f(a))) = q_B(f(a)) = q_C(g(a)) = q(i_C(f(a)),$$

so that $\langle i_B(f(a)), i_C(f(a)) \rangle \in \ker(q)$. Because $\ker(q)$ is an equivalence relation on D, we conclude $R \subseteq \ker(q)$. Thus $h([x]_R) := q(x)$ defines a map $D/R \to E$ with

$$\begin{aligned} h(p_B(b)) &= h([i_B(b)]_R) &&= q(i_B(b)) &&= q_B(b), \\ h(p_C(c)) &= h([i_C(c)]_R) &&= q(i_C(c)) &&= q_C(c) \end{aligned}$$

for $b \in B$ and $c \in C$. It is clear that there is no other way to define a map h with the desired properties. ✌

So we have shown that the pushout in the category ***Set*** of sets with maps exists. To illustrate the construction, consider the pushout of two factor maps. In this example, $\rho \vee \tau$ denotes the smallest equivalence relation which contains the equivalence relations ρ and τ.

Example 2.2.26 Let ρ and ϑ be equivalence relations on a set X with factor maps $\eta_\rho : X \to X/\rho$ and $\eta_\vartheta : X \to X/\vartheta$. Then the pushout of these maps is $X/(\rho \vee \vartheta)$ with $\zeta_\rho : [x]_\rho \mapsto [x]_{\rho\vee\vartheta}$ and $\zeta_\vartheta : [x]_\vartheta \mapsto [x]_{\rho\vee\vartheta}$ as the associated maps. In fact, we have $\eta_\rho \circ \zeta_\rho = \eta_\vartheta \circ \zeta_\vartheta$, so the first property is satisfied. Now let $t_\rho : X/\rho \to E$ and $t_\vartheta : X/\vartheta \to E$ be maps with $t_\rho \circ \eta_\rho = t_\vartheta \circ \eta_\vartheta$ for a set E, then $h : [x]_{\rho\vee\vartheta} \mapsto t_\rho([x]_\rho)$ maps $X/(\rho \vee \vartheta)$ to E with plainly $t_\rho = h \circ \zeta_\rho$ and $t_\vartheta = h \circ \zeta_\vartheta$; moreover, h is uniquely determined by this property. Because the pushout is uniquely determined up to isomorphism by Lemma 2.2.2 and Proposition 2.2.13, we have shown that the supremum of two equivalence relations in the lattice of equivalence relations can be computed through the pushout of its components. ✌

2.3 Functors and Natural Transformations

We introduce functors which help in transporting information between categories in a way similar to morphisms, which are thought to transport information between objects. Of course, we will have to observe some properties in order to capture the intuitive understanding of a functor as a structure-preserving element in a formal way. Functors themselves can be related, leading to the notion of a natural transformation. Given a category, there is a plethora of functors and natural transformations provided by the hom sets; this is studied in some detail, first, because it is a built-in for every category and second because the Yoneda lemma relates this rich structure to set-based functors, which in turn will be used when studying adjunctions.

2.3.1 Functors

Loosely speaking, a functor is a pair of structure-preserving maps between categories: it maps one category to another one in a compatible way. A bit more precise, a functor $\boldsymbol{F}$ between categories $\boldsymbol{K}$ and $\boldsymbol{L}$ assigns to each object a in category $\boldsymbol{K}$ an object $\boldsymbol{F}(a)$ in $\boldsymbol{L}$, and it assigns each morphism $f : a \to b$ in $\boldsymbol{K}$ a morphism $\boldsymbol{F}(f) : \boldsymbol{F}(a) \to \boldsymbol{F}(b)$ in $\boldsymbol{L}$; some obvious properties have to be observed. To be more specific:

Definition 2.3.1 *A* functor $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ *assigns to each object a in category $\boldsymbol{K}$ an object $\boldsymbol{F}(a)$ in category $\boldsymbol{L}$ and maps each hom set* $\text{hom}_{\boldsymbol{K}}(a, b)$

of $\boldsymbol{K}$ *to the hom set* $\hom_{\boldsymbol{L}}(\boldsymbol{F}(a), \boldsymbol{F}(b))$ *of* $\boldsymbol{L}$ *subject to these conditions:*

- $\boldsymbol{F}(id_a) = id_{\boldsymbol{F}(a)}$ *for each object* a *of* $\boldsymbol{K}$*,*
- *if* $f : a \to b$ *and* $g : b \to c$ *are morphisms in* $\boldsymbol{K}$*, then* $\boldsymbol{F}(g \circ f) = \boldsymbol{F}(g) \circ \boldsymbol{F}(f)$.

A functor $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{K}$ *is called an* endofunctor *on* $\boldsymbol{K}$.

The first condition says that the identity morphisms in $\boldsymbol{K}$ are mapped to the identity morphisms in $\boldsymbol{L}$, and the second condition tells us that $\boldsymbol{F}$ has to be compatible with composition in the respective categories. Note that for specifying a functor, we have to say what the functor does with objects and how the functor transforms morphisms. By the way, we often write $\boldsymbol{F}(a)$ as $\boldsymbol{F}a$ and $\boldsymbol{F}(f)$ as $\boldsymbol{F}f$.

Let us have a look at some examples. Trivial examples for functors include the *identity functor* $Id_{\boldsymbol{K}}$, which maps objects resp. morphisms to itself, and the *constant functor* Δ_x for an object x, which maps every object to x and every morphism to id_x.

Example 2.3.2 Consider the category $\boldsymbol{Set}$ of sets with maps as morphisms. Given set X, $\mathcal{P}X$ is a set again; define $\mathcal{P}(f)(A) := f\big[A\big]$ for the map $f : X \to Y$ and for $A \subseteq X$, then $\mathcal{P}f : \mathcal{P}X \to \mathcal{P}Y$. We check the laws for a functor:

- $\mathcal{P}(id_X)(A) = id_X\big[A\big] = A = id_{\mathcal{P}X}(A)$, so that $\mathcal{P}id_X = id_{\mathcal{P}X}$.
- let $f : X \to Y$ and $g : Y \to Z$, then $\mathcal{P}f : \mathcal{P}X \to \mathcal{P}Y$ and $\mathcal{P}g : \mathcal{P}Y \to \mathcal{P}Z$ with

$$\begin{aligned}(\mathcal{P}(g) \circ \mathcal{P}(f))(A) &= \mathcal{P}(g)(\mathcal{P}(f)(A)) &&= g\big[f\big[A\big]\big]\\ &= \{g(f(a)) \mid a \in A\} &&= (g \circ f)\big[A\big]\\ &= \mathcal{P}(g \circ f)(A)\end{aligned}$$

  for $A \subseteq X$. Thus the *power set functor* $\mathcal{P}$ is compatible with composition of maps.

✌

Example 2.3.3 Given a category $\boldsymbol{K}$ and an object a of $\boldsymbol{K}$, associate

$$a_+ : x \mapsto \hom_{\boldsymbol{K}}(a, x)$$
$$a^+ : x \mapsto \hom_{\boldsymbol{K}}(x, a)$$

with a together with the maps on hom sets $\hom_{\boldsymbol{K}}(a, \cdot)$ resp. $\hom_{\boldsymbol{K}}(\cdot, a)$. Then a^+ is a functor $\boldsymbol{K} \to \boldsymbol{Set}$. hom functors

In fact, given morphism $f : x \to y$, we have $a_+ f : \hom_{\boldsymbol{K}}(a, x) \to \hom_{\boldsymbol{K}}(a, y)$, taking g into $f \circ g$. Plainly, $a_+(id_x) = id_{\hom_{\boldsymbol{K}}(a,x)} = id_{a_+(x)}$, and

$$a_+(g \circ f)(h) = (g \circ f) \circ h = g \circ (f \circ h) = a_+(g)(a_+(f)(h)),$$

if $f : x \to y, g : y \to z$ and $h : a \to x$. ✌

Functors come in handy when we want to forget part of the structure.

Example 2.3.4 Let ***Meas*** be the category of measurable spaces. Assign to each measurable space $(X, \mathcal{C})$ its carrier set X and to each morphism $f : (X, \mathcal{C}) \to (Y, \mathcal{D})$ the corresponding map $f : X \to Y$. It is immediately checked that this constitutes a functor $\boldsymbol{Meas} \to \boldsymbol{Set}$. Similarly, we might forget the topological structure by assigning each topological space its carrier set, and assign each continuous map to itself. These functors are sometimes called *forgetful functors*. ✌

The following example twists Example 2.3.4 a little bit.

Example 2.3.5 Assign to each measurable space $(X, \mathcal{C})$ its σ-algebra $\boldsymbol{B}(X, \mathcal{C}) := \mathcal{C}$. Let $f : (X, \mathcal{C}) \to (Y, \mathcal{D})$ be a morphism in ***Meas***; put $\boldsymbol{B}(f) := f^{-1}$, then $\boldsymbol{B}(f) : \boldsymbol{B}(Y, \mathcal{D}) \to \boldsymbol{B}(X, \mathcal{C})$, because f is $\mathcal{C}$-$\mathcal{D}$-measurable. We plainly have $\boldsymbol{B}(id_{X,\mathcal{C}}) = id_{\boldsymbol{B}(X,\mathcal{C})}$ and $\boldsymbol{B}(g \circ f) = (g \circ f)^{-1} = f^{-1} \circ g^{-1} = \boldsymbol{B}(f) \circ \boldsymbol{B}(g)$, so $\boldsymbol{B} : \boldsymbol{Meas} \to \boldsymbol{Set}$ is no functor, although it behaves like one. DO NOT PANIC! If we reverse arrows, things work out properly: $\boldsymbol{B} : \boldsymbol{Meas} \to \boldsymbol{Set}^{op}$ is, as we have just shown, a functor (the dual $\boldsymbol{K}^{op}$ of a category $\boldsymbol{K}$ has been introduced in Example 2.1.17).

This functor could be called the *Borel functor* (the measurable sets are sometimes called the Borel sets). ✌

Definition 2.3.6 *A functor $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}^{op}$ is called a* contravariant functor *between* **K** *and* **L***; in contrast, a functor according to Definition 2.3.1 is called* covariant.

If we talk about functors, we always mean the covariant flavor; contravariance is mentioned explicitly.

Let us complete the discussion from Example 2.3.3 by considering a^+, which takes $f : x \to y$ to $a^+ f : \hom_{\boldsymbol{K}}(y, a) \to \hom_{\boldsymbol{K}}(x, a)$ through

$g \mapsto g \circ f$. a^+ maps the identity on x to the identity on $\hom_{\boldsymbol{K}}(x, a)$. If $g : y \to z$, we have

$$a^+(f)(a^+(g)(h)) = a^+(f)(h \circ g) = h \circ g \circ f = a^+(g \circ f)(h)$$

for $h : z \to a$. Thus a^+ is a contravariant functor $\boldsymbol{K} \to \boldsymbol{Set}$, while its cousin a_+ is a covariant.

Functors may also be used to model structures.

Example 2.3.7 Consider this functor $\boldsymbol{S} : \boldsymbol{Set} \to \boldsymbol{Set}$ which assigns each set X the set $X^{\mathbb{N}}$ of all sequences over X; the map $f : X \to Y$ is assigned the map $\boldsymbol{S} : (x_n)_{n \in \mathbb{N}} \mapsto \big(f(x_n)\big)_{n \in \mathbb{N}}$. Evidently, id_X is mapped to $id_{X^{\mathbb{N}}}$, and it is easily checked that $\boldsymbol{S}(g \circ f) = \boldsymbol{S}(g) \circ \boldsymbol{S}(f)$. Hence $\boldsymbol{S}$ constitutes an endofunctor on $\boldsymbol{Set}$. ✌

Example 2.3.8 Similarly, define the endofunctor $\boldsymbol{F}$ on $\boldsymbol{Set}$ by assigning X to $X^{\mathbb{N}} \cup X^*$ with X^* as the set of all finite sequences over X. Then $\boldsymbol{F}X$ has all finite or infinite sequences over the set X. Let $f : X \to Y$ be a map, and let $(x_i)_{i \in I} \in \boldsymbol{F}X$ be a finite or infinite sequence; then put $(\boldsymbol{F}f)(x_i)_{i \in I} := \big(f(x_i)\big)_{i \in I} \in \boldsymbol{F}Y$. It is not difficult to see that $\boldsymbol{F}$ satisfies the laws for a functor. ✌

The next example deals with automata which produce an output (in contrast to Example 2.1.15 where we mainly had state transitions in view).

Example 2.3.9 An *automaton with output* (A, B, X, δ) has an input alphabet A, an output alphabet B, and a set X of states with a map $\delta : X \times A \to X \times B$; $\delta(x, a) = \langle x', b \rangle$ yields the next state x' and the output b, if the input is a in state x. A morphism $f : (X, A, B, \delta) \to (Y, A, B, \vartheta)$ of automata is a map $f : X \to Y$ such that $\vartheta(f(x), a) = (f \times id_B)(\delta(x, a))$ for all $x \in X, a \in A$; thus $(f \times id_B) \circ \delta = \vartheta \circ (f \times id_A)$. This yields apparently a category $\boldsymbol{AutO}$, the category of automata with output.

We want to expose the state space X in order to make it a parameter to an automata, because input and output alphabets are given from the outside; so for modeling purposes, only states are at our disposal. Hence we reformulate δ and take it as a map $\delta_* : X \to (X \times B)^A$ with $\delta_*(x)(a) := \delta(x, a)$. Now $f : (X, A, B, \delta) \to (Y, A, B, \vartheta)$ is a morphism iff this diagram commutes:

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "\delta_*"'] & Y \arrow[d, "\vartheta_*"] \\
(X \times B)^A \arrow[r, "f^\bullet"'] & (Y \times B)^A
\end{tikzcd}$$

with $f^\bullet(t)(a) := (f \times id_B)(t(a))$. Let us see why this is the case. Given $x \in X, a \in A$, we have

$$\begin{aligned}
f^\bullet(\delta_*(x))(a) &= (f \times id_B)(\delta_*(x)(a)) = (f \times id_B)(\delta(x,a)) \\
&= \vartheta(f(x), a) = \vartheta_*(f(x))(a);
\end{aligned}$$

thus $f^\bullet(\delta_*(x)) = \vartheta_*(f(x))$ for all $x \in X$; hence $f^\bullet \circ \delta_* = \vartheta_* \circ f$, so the diagram is commutative indeed. Define $\boldsymbol{F}(X) := (X \times A)^{\boldsymbol{B}}$, for an object (A, B, X, δ) in category ***AutO***, and put $\boldsymbol{F}(f) := f^\bullet$ for the automaton morphism $f : (X, A, B, \delta) \to (Y, A, B, \vartheta)$; thus $\boldsymbol{F}(f)$ renders this diagram commutative:

$$\begin{tikzcd}
& A \arrow[dl, "t"'] \arrow[dr, "\boldsymbol{F}(f)(t)"] & \\
X \times B \arrow[rr, "f \times id_B"'] & & Y \times B
\end{tikzcd}$$

We claim that $\boldsymbol{F}$: ***AutO*** $\to$ ***Set*** is a functor. Let $g : (Y, A, B, \vartheta) \to (Z, A, B, \zeta)$ be a morphism; then $\boldsymbol{F}(g)$ makes this diagram commutative for all $s \in (Y \times B)^A$:

$$\begin{tikzcd}
& A \arrow[dl, "s"'] \arrow[dr, "\boldsymbol{F}(g)(s)"] & \\
X \times B \arrow[rr, "g \times id_B"'] & & Y \times B
\end{tikzcd}$$

In particular, we have for $s := \boldsymbol{F}(f)(t)$ with an arbitrary $t \in (X \times B)^A$ this commutative diagram:

$$\begin{tikzcd}
& A \arrow[dl, "\boldsymbol{F}(f)(t)"'] \arrow[dr, "\boldsymbol{F}(g)(\boldsymbol{F}(f)(t))"] & \\
Y \times B \arrow[rr, "g \times id_B"'] & & Z \times B
\end{tikzcd}$$

Thus the outer diagram commutes

$$\begin{tikzcd}
& A \arrow[dl, "t"'] \arrow[d, "\boldsymbol{F}(f)(t)"'] \arrow[dr, "\boldsymbol{F}(g)(\boldsymbol{F}(f)(t))"] & \\
X \times B \arrow[r, "f \times id_B"'] & Y \times B \arrow[r, "g \times id_B"'] & Z \times B
\end{tikzcd}$$

Consequently, we have

$$\begin{aligned}\boldsymbol{F}(g)(\boldsymbol{F}(f)(t)) &= (g \times id_B) \circ (f \times id_B) \circ t\\ &= \big((g \circ f) \times id_B\big) \circ t\\ &= \boldsymbol{F}(g \circ f)(t).\end{aligned}$$

Now $\boldsymbol{F}(id_X) = id_{(X \times B)^A}$ is trivial, so that we have established that $\boldsymbol{F} : \boldsymbol{AutO} \to \boldsymbol{Set}$ is indeed a functor, assigning states to possible state transitions. ✌

The next example shows that we may perceive labeled transition systems as functors based on the power set functor.

Example 2.3.10 A *labeled transition system* is a collection of transitions indexed by a set of actions. Formally, given a set A of actions, $\big(S, (\rightsquigarrow_a)_{a \in A}\big)$ is a labeled transition system iff $\rightsquigarrow_a \subseteq S \times S$ for all $a \in A$. Thus state s may go into state s' after action $a \in A$; this is written as $s \rightsquigarrow_a s'$. A morphism $f : \big(S, (\rightsquigarrow_{S,a})_{a \in A}\big) \to \big(T, (\rightsquigarrow_{T,a})_{a \in A}\big)$ of transition systems is a map $f : S \to T$ such that $s \rightsquigarrow_{S,a} s'$ implies $f(s) \rightsquigarrow_{T,a} f(s')$ for all actions a, cp. Example 2.1.9.

We model a transition system $\big(S, (\rightsquigarrow_a)_{a \in A}\big)$ as a map $F : S \to \mathcal{P}(A \times S)$ by defining $F(s) := \{\langle a, s'\rangle \mid s \rightsquigarrow_a s'\}$; thus $F(s) \subseteq A \times S$ collects actions and new states; conversely, we may recover $\rightsquigarrow_a$ from F: $\rightsquigarrow_a = \{\langle s, s'\rangle \mid \langle a, s'\rangle \in F(s)\}$. This suggests defining $\boldsymbol{F}(S) := \mathcal{P}(A \times S)$ which can be made a functor once we have decided what to do with morphisms

$$f : \big(S, (\rightsquigarrow_{S,a})_{a \in A}\big) \to \big(T, (\rightsquigarrow_{T,a})_{a \in A}\big).$$

Take $V \subseteq A \times S$ and define $\boldsymbol{F}(f)(V) := \{\langle a, f(s)\rangle \mid \langle a, s\rangle \in V\}$ (clearly we want to leave the actions alone). Then we have

$$\begin{aligned}\boldsymbol{F}(g \circ f)(V) &= \{\langle a, g(f(s))\rangle \mid \langle a, s\rangle \in V\}\\ &= \{\langle a, g(y)\rangle \mid \langle a, y\rangle \in \boldsymbol{F}(f)(V)\}\\ &= \boldsymbol{F}(g)(\boldsymbol{F}(f)(V))\end{aligned}$$

for a morphism $g : \big(T, (\rightsquigarrow_{T,a})_{a \in A}\big) \to \big(U, (\rightsquigarrow_{U,a})_{a \in A}\big)$. Thus we have shown that $\boldsymbol{F}(g \circ f) = \boldsymbol{F}(g) \circ \boldsymbol{F}(f)$ holds. Because $\boldsymbol{F}$ maps the identity to the identity, $\boldsymbol{F}$ is a functor from the category of labeled transition systems to $\boldsymbol{Set}$. ✌

The next examples deal with functors induced by probabilities.

Example 2.3.11 Given a set X, define the support $\mathrm{supp}(p)$ for a map $p : X \to [0,1]$ as $\mathrm{supp}(p) := \{x \in X \mid p(x) \neq 0\}$. A *discrete probability* p on X is a map $p : X \to [0,1]$ with finite support such that

Support supp

$$\sum_{x \in X} p(x) := \sum_{x \in \mathrm{supp}(p)} p(x) = 1.$$

Denote by

$$\boldsymbol{D}(X) := \{p : X \to [0,1] \mid p \text{ is a discrete probability}\}$$

the set of all discrete probabilities. Let $f : X \to Y$ be a map, and define

$$\boldsymbol{D}(f)(p)(y) := \sum_{\{x \in X \mid f(x) = y\}} p(x).$$

Because $\boldsymbol{D}(f)(p)(y) > 0$ iff $y \in f\left[\mathrm{supp}(p)\right]$, $\boldsymbol{D}(f)(p) : Y \to [0,1]$ has finite support, and

$$\sum_{y \in Y} \boldsymbol{D}(f)(p)(y) = \sum_{y \in Y} \sum_{\{x \in X \mid f(x) = y\}} p(x) = \sum_{x \in X} p(x) = 1.$$

It is clear that $\boldsymbol{D}(id_X)(p) = p$, so we have to check whether $\boldsymbol{D}(g \circ f) = \boldsymbol{D}(g) \circ \boldsymbol{D}(f)$ holds.

We use a little trick for this, which will turn out to be helpful later as well. Define

$$p(A) := \sum_{x \in A \cap \mathrm{supp}(p)} p(x)$$

for $p \in \boldsymbol{D}(X)$ and $A \subseteq X$; then p is a probability measure on $\mathcal{P}X$. A direct calculation shows that $\boldsymbol{D}(f)(p)(y) = p(f^{-1}\left[\{y\}\right])$ and $\boldsymbol{D}(f)(B) = p(f^{-1}\left[B\right])$ for $B \subseteq Y$ hold. Thus we obtain for the maps $f : X \to Y$ and $g : Y \to Z$

$$\begin{aligned} \boldsymbol{D}(g \circ f)(p)(z) &= p\big((g \circ f)^{-1}\left[\{z\}\right]\big) &&= p\big(f^{-1}\left[g^{-1}\left[\{z\}\right]\right]\big) \\ &= \boldsymbol{D}(f)(p)(g^{-1}\left[\{z\}\right]) &&= \boldsymbol{D}(f)(\boldsymbol{D}(g)(p))(z). \end{aligned}$$

Thus $\boldsymbol{D}(g \circ f) = \boldsymbol{D}(g) \circ \boldsymbol{D}(f)$, as claimed.

Discrete probability functor $\boldsymbol{D}$

Hence $\boldsymbol{D}$ is an endofunctor on ***Set***, the *discrete probability functor*. It is immediate that all the arguments above hold also for probabilities, the support of which is countable; but since we will discuss an interesting example on page 192 which deals with the finite case, we stick to that here. ✌

There is a continuous version of this functor as well. We generalize things a bit and formulate the example for subprobabilities.

Example 2.3.12 We are now working in the category ***Meas*** of measurable spaces with measurable maps as morphisms. Given a measurable space $(X, \mathcal{A})$, the set $\mathbb{S}(X, \mathcal{A})$ of all subprobability measures is a measurable space with the weak σ-algebra $\wp(\mathcal{A})$ associated with $\mathcal{A}$; see Example 2.1.14. Hence $\mathbb{S}$ maps measurable spaces to measurable spaces. Define for a morphism $f : (X, \mathcal{A}) \to (Y, \mathcal{B})$

$$\mathbb{S}(f)(\mu)(B) := \mu(f^{-1}\left[B\right])$$

for $B \in \mathcal{B}$. Then $\mathbb{S}(f) : \mathbb{S}(X, \mathcal{A}) \to \mathbb{S}(Y, \mathcal{B})$ is $\wp(\mathcal{A})$-$\wp(\mathcal{B})$-measurable by Exercise 2.8. Now let $g : (Y, \mathcal{B}) \to (Z, \mathcal{C})$ be a morphism in ***Meas***; then we show as in Example 2.3.11 that

$$\mathbb{S}(g \circ f)(\mu)(C) = \mu\big(f^{-1}\left[g^{-1}\left[C\right]\right]\big) = \mathbb{S}(f)\big(\mathbb{S}(g)(\mu)\big)(C),$$

Subprobability functor $\mathbb{S}$

for $C \in \mathcal{C}$; thus $\mathbb{S}(g \circ f) = \mathbb{S}(g) \circ \mathbb{S}(f)$. Since $\mathbb{S}$ preserves the identity, $\mathbb{S}$: ***Meas*** $\to$ ***Meas*** is an endofunctor, the (continuous space) *subprobability functor*. ✌

The next two examples deal with upper closed sets, the first one with these sets proper and the second one with a more refined version, viz., with ultrafilters. Upper closed sets are used, e.g., for the interpretation of game logic, a variant of modal logics; see Example 2.7.22 and Sect. 4.1.3.

Example 2.3.13 Call a subset $V \subseteq \mathcal{P}S$ *upper closed* iff $A \in V$ and $A \subseteq B$ together imply $B \in V$; for example, each filter is upper closed. Denote by

$$\boldsymbol{V}S := \{V \subseteq \mathcal{P}S \mid V \text{ is upper closed}\}$$

$\boldsymbol{V}$, Upper closed sets

the set of all upper closed subsets of $\mathcal{P}S$. Given $f : S \to T$, define

$$(\boldsymbol{V}f)(V) := \{W \subseteq \mathcal{P}T \mid f^{-1}\left[W\right] \in V\}$$

for $V \in \boldsymbol{V}S$. Let $W \in \boldsymbol{V}(V)$ and $W_0 \supseteq W$, then $f^{-1}\left[W\right] \subseteq f^{-1}\left[W_0\right]$, so that $f^{-1}\left[W_0\right] \in V$; hence $\boldsymbol{V}f : \boldsymbol{V}S \to \boldsymbol{V}W$. It is easy to see that $\boldsymbol{V}(g \circ f) = \boldsymbol{V}(g) \circ \boldsymbol{V}(f)$, provided $f : S \to T$ and $g : T \to V$. Moreover, $\boldsymbol{V}(id_S) = id_{\boldsymbol{V}(S)}$. Hence $\boldsymbol{V}$ is an endofunctor on the category ***Set*** of sets with maps as morphisms. ✌

Ultrafilters are upper closed, but are much more complex than plain upper closed sets, since they are filters and they are maximal. Thus we have to look a bit closer at the properties which the functor is to represent.

Example 2.3.14 Let

$$\boldsymbol{U}S := \{q \mid q \text{ is an ultrafilter over } S\}$$

assign to each set S its ultrafilters, to be more precise, all ultrafilters of the power set of S. This is the object part of an endofunctor over the category ***Set*** with maps as morphisms. Given a map $f : S \to T$, we have to define $\boldsymbol{U}f : \boldsymbol{U}S \to \boldsymbol{U}T$. Before doing so, a preliminary consideration will help.

$\boldsymbol{U}$, Ultrafilters

One first notes that, given two Boolean algebras B and B' and a Boolean algebra morphism $\gamma : B \to B'$, γ^{-1} maps ultrafilters over B' to ultrafilters over B. In fact, let w be an ultrafilter over B'; put $v := \gamma^{-1}\big[w\big]$; we go quickly over the properties of an ultrafilter should have. First, v does not contain the bottom element $\perp_{\boldsymbol{B}}$ of B, for otherwise, $\perp_{\boldsymbol{B}'} = \gamma(\perp_{\boldsymbol{B}}) \in w$. If $a \in v$ and $b \geq a$, then $\gamma(b) \geq \gamma(a) \in w$; hence $\gamma(b) \in w$; thus $b \in v$; plainly, v is closed under $\wedge$. Now assume $a \notin v$, then $\gamma(a) \notin w$; hence $\gamma(-a) = -\gamma(a) \in w$, since w is an ultrafilter. Consequently, $-a \in v$. This establishes the claim.

Given a map $f : S \to T$, define $F_f : \mathcal{P}T \to \mathcal{P}S$ through $F_f := f^{-1}$. This is a homomorphism of the Boolean algebras $\mathcal{P}T$ and $\mathcal{P}S$; thus F_f^{-1} maps $\boldsymbol{U}S$ to $\boldsymbol{U}T$. Put $\boldsymbol{U}(f) := F_f^{-1}$; note that we reverse the arrows' directions twice. It is clear that $\boldsymbol{U}(id_S) = id_{\boldsymbol{U}(S)}$, and if $g : T \to Z$, then

$$\boldsymbol{U}(g \circ f) = F_{g \circ f}^{-1} = (F_f \circ F_g)^{-1} = F_g^{-1} \circ F_f^{-1} = \boldsymbol{U}(g) \circ \boldsymbol{U}(f).$$

This shows that $\boldsymbol{U}$ is an endofunctor on the category ***Set*** of sets with maps as morphisms ($\boldsymbol{U}$ is sometimes denoted by β). ✌

We can use functors for constructing new categories from given ones. As an example, we define the comma category associated with two functors.

Definition 2.3.15 *Let $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ and $\boldsymbol{G} : \boldsymbol{M} \to \boldsymbol{L}$ be functors. The* comma category *$(\boldsymbol{F}, \boldsymbol{G})$ associated with $\boldsymbol{F}$ and $\boldsymbol{G}$ has as objects the triplets $\langle a, f, b\rangle$ with objects a from $\boldsymbol{K}$, b from $\boldsymbol{M}$, and morphisms*

$f : \boldsymbol{F}a \to \boldsymbol{G}b$. A morphism $(\varphi, \psi) : \langle a, f, b\rangle \to \langle a', f', b'\rangle$ is a pair of morphisms $\varphi : a \to a'$ of $\boldsymbol{K}$ and $\psi : b \to b'$ of $\boldsymbol{M}$ such that this diagram commutes:

$$\begin{tikzcd}
\boldsymbol{F}a \arrow[r, "\boldsymbol{F}\varphi"] \arrow[d, "f"'] & \boldsymbol{F}a' \arrow[d, "f'"] \\
\boldsymbol{G}b \arrow[r, "\boldsymbol{G}\psi"'] & \boldsymbol{G}b'
\end{tikzcd}$$

Composition of morphism is component wise.

The slice category $\boldsymbol{K}/x$ defined in Example 2.1.16 is apparently the comma category $(Id_{\boldsymbol{K}}, \Delta_x)$.

Functors can be composed, yielding a new functor. The proof for this statement is straightforward.

Proposition 2.3.16 *Let $\boldsymbol{F} : \boldsymbol{C} \to \boldsymbol{D}$ and $\boldsymbol{G} : \boldsymbol{D} \to \boldsymbol{E}$ be functors. Define $(\boldsymbol{G} \circ \boldsymbol{F})a := \boldsymbol{G}(\boldsymbol{F}a)$ for an object a of $\boldsymbol{C}$ and $(\boldsymbol{G} \circ \boldsymbol{F})f := \boldsymbol{G}(\boldsymbol{F}f)$ for a morphism $f : a \to b$ in $\boldsymbol{C}$; then $\boldsymbol{G} \circ \boldsymbol{F} : \boldsymbol{C} \to \boldsymbol{E}$ is a functor.* $\dashv$

2.3.2 Natural Transformations

We see that we can compose functors in an obvious way. This raises the question whether or not functors themselves form a category. But we do not yet have morphisms between functors at out disposal. Natural transformations will assume this rôle. Nevertheless, the question remains, but it will not be answered in the positive; this is so because morphisms between objects should form a set, and it will be clear that this is not the case. Pumplün [Pum99] points at some difficulties that might arise and arrives at the pragmatic view that for practical problems this question is not particularly relevant.

But let us introduce natural transformations between functors $\boldsymbol{F}, \boldsymbol{G}$ now. The basic idea is that for each object a, $\boldsymbol{F}a$ is transformed into $\boldsymbol{G}a$ in a way which is compatible with the structure of the participating categories.

$\eta_a : \boldsymbol{F}a \to \boldsymbol{G}a$

Definition 2.3.17 *Let $\boldsymbol{F}, \boldsymbol{G} : \boldsymbol{K} \to \boldsymbol{L}$ be covariant functors. A family $\eta = (\eta_a)_{a \in |\boldsymbol{K}|}$ is called a* natural transformation *$\eta : \boldsymbol{F} \to \boldsymbol{G}$ iff $\eta_a : \boldsymbol{F}a \to \boldsymbol{G}a$ is a morphism in $\boldsymbol{L}$ for all objects a in $\boldsymbol{K}$ such that this* *diagram commutes for any morphism $f : a \to b$ in $\boldsymbol{K}$:*

$$
\begin{tikzcd}
a \arrow[d, "f"] \\
b
\end{tikzcd}
\qquad
\begin{tikzcd}
\boldsymbol{F}a \arrow[r, "\eta_a"] \arrow[d, "\boldsymbol{F}f"'] & \boldsymbol{G}a \arrow[d, "\boldsymbol{G}f"] \\
\boldsymbol{F}b \arrow[r, "\eta_b"'] & \boldsymbol{G}b
\end{tikzcd}
$$

Thus a natural transformation $\eta : \boldsymbol{F} \to \boldsymbol{G}$ is a family of morphisms, indexed by the objects of the common domain of $\boldsymbol{F}$ and $\boldsymbol{G}$; η_a is called the *component of* η *at* a.

If $\boldsymbol{F}$ and $\boldsymbol{G}$ are both contravariant functors $\boldsymbol{K} \to \boldsymbol{L}$, we may perceive them as covariant functors $\boldsymbol{K} \to \boldsymbol{L}^{op}$, so that we get for the contravariant case this diagram:

$$
\begin{tikzcd}
a \arrow[d, "f"] \\
b
\end{tikzcd}
\qquad
\begin{tikzcd}
\boldsymbol{F}a \arrow[r, "\eta_a"] & \boldsymbol{G}a \\
\boldsymbol{F}b \arrow[r, "\eta_b"'] \arrow[u, "\boldsymbol{F}f"] & \boldsymbol{G}b \arrow[u, "\boldsymbol{G}f"']
\end{tikzcd}
$$

Let us have a look at some examples.

Example 2.3.18 $a_+ : x \mapsto \hom_{\boldsymbol{K}}(a, x)$ yields a (covariant) functor $\boldsymbol{K} \to \boldsymbol{Set}$ for each object a in $\boldsymbol{K}$; see Example 2.3.3 (just for simplifying notation, we use again a_+ rather than $\hom_{\boldsymbol{K}}(a, -)$; see page 126). Let $\zeta : b \to a$ be a morphism in $\boldsymbol{K}$; then this induces a natural transformation $\eta_\zeta : a_+ \to b_+$ with

$$
\eta_{\zeta,x} : \begin{cases} a_+(x) & \to b_+(x) \\ g & \mapsto g \circ \zeta \end{cases}
$$

In fact, look at this diagram with a $\boldsymbol{K}$-morphism $f : x \to y$:

$$
\begin{tikzcd}
x \arrow[d, "f"'] \\
y
\end{tikzcd}
\qquad
\begin{tikzcd}
a_+(x) \arrow[r, "\eta_{\zeta,x}"] \arrow[d, "a_+(f)"'] & b_+(x) \arrow[d, "b_+(f)"] \\
a_+(y) \arrow[r, "\eta_{\zeta,y}"'] & b_+(y)
\end{tikzcd}
$$

Then we have for $h \in a_+(x) = \hom_{\boldsymbol{K}}(a, x)$

$$
\begin{aligned}
(\eta_{\zeta,y} \circ a_+(f))(h) &= \eta_{\zeta,y}(f \circ h) && = (f \circ h) \circ \zeta \\
&= f \circ (h \circ \zeta) && = b_+(f)(\eta_{\zeta,x}(h)) \\
&= (b_+(f) \circ \eta_{\zeta,x})(h)
\end{aligned}
$$

Hence η_ζ is in fact a natural transformation. ✌

This is an example in the category of groups:

Example 2.3.19 Let $\boldsymbol{K}$ be the category of groups (see Example 2.1.7). It is not difficult to see that $\boldsymbol{K}$ has products. Define for a group H the map $\boldsymbol{F}_H(G) := H \times G$ on objects, and if $f : G \to G'$ is a morphism in $\boldsymbol{K}$, define $\boldsymbol{F}_H(f) : H \times G \to H \times G'$ through $\boldsymbol{F}_H(f) : \langle h, g\rangle \mapsto \langle h, f(g)\rangle$. Then $\boldsymbol{F}_H$ is an endofunctor on $\boldsymbol{K}$. Now let $\varphi : H \to K$ be a morphism. Then φ induces a natural transformation η_φ upon setting

$$\eta_{\varphi,G} : \begin{cases} \boldsymbol{F}_H & \to \boldsymbol{F}_K \\ \langle h, g\rangle & \mapsto \langle \varphi(h), g\rangle. \end{cases}$$

In fact, let $\psi : L \to L'$ be a group homomorphism, then this diagram commutes:

$$\begin{tikzcd} L \arrow[d, "\psi"'] & & \boldsymbol{F}_H\ L \arrow[r, "\eta_{\varphi,L}"] \arrow[d, "\boldsymbol{F}_H\ \psi"'] & \boldsymbol{F}_K\ L \arrow[d, "\boldsymbol{F}_K\ \psi"] \\ L' & & \boldsymbol{F}_H\ L' \arrow[r, "\eta_{\varphi,L'}"'] & \boldsymbol{F}_K\ L' \end{tikzcd}$$

To see this, take $\langle h, \ell\rangle \in \boldsymbol{F}_H\ L = H \times L$, and chase it through the diagram:

$$(\eta_{\varphi,L'} \circ \boldsymbol{F}_H\ \psi)(h, \ell) = \langle \varphi(h), \varphi(\ell)\rangle = (\boldsymbol{F}_K(\psi) \circ \eta_{\varphi,L})(h, \ell).$$

✌

Consider as another example a comma category $(\boldsymbol{F}, \boldsymbol{G})$ (Definition 2.3.15). There are functors akin to a projection which permit to recover the original functors and which are connected through a natural transformation. To be specific:

Proposition 2.3.20 *Let $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ and $\boldsymbol{G} : \boldsymbol{M} \to \boldsymbol{L}$ be functors. Then there are functors $\boldsymbol{S} : (\boldsymbol{F}, \boldsymbol{G}) \to \boldsymbol{K}$ and $\boldsymbol{R} : (\boldsymbol{F}, \boldsymbol{G}) \to \boldsymbol{L}$ rendering this diagram commutative:*

$$\begin{tikzcd} (\boldsymbol{F}, \boldsymbol{G}) \arrow[r, "\boldsymbol{R}"] \arrow[d, "\boldsymbol{S}"'] & \boldsymbol{M} \arrow[d, "\boldsymbol{G}"] \\ \boldsymbol{K} \arrow[r, "\boldsymbol{F}"'] & \boldsymbol{L} \end{tikzcd}$$

There exists a natural transformation $\eta : \boldsymbol{F} \circ \boldsymbol{S} \to \boldsymbol{G} \circ \boldsymbol{R}$.

Proof Put for the object $\langle a, f, b\rangle$ of $(\boldsymbol{F}, \boldsymbol{G})$ and the morphism (φ, ψ)

$$\begin{aligned} \boldsymbol{S}\langle a, f, b\rangle &:= a, \quad \boldsymbol{S}(\varphi, \psi) := \varphi, \\ \boldsymbol{R}\langle a, f, b\rangle &:= b, \quad \boldsymbol{R}(\varphi, \psi) := \psi. \end{aligned}$$

Then it is clear that the desired equality holds. Moreover, $\eta_{\langle a,f,b\rangle} := f$ is the desired natural transformation. The crucial diagram commutes by the definition of morphisms in the comma category. $\dashv$

Example 2.3.21 Assume that the product $a \times b$ for the objects a and b in category $\boldsymbol{K}$ exists; then Proposition 2.2.9 tells us that we have for each object d a bijection $p_d : \mathrm{hom}_{\boldsymbol{K}}(d,a) \times \mathrm{hom}_{\boldsymbol{K}}(d,b) \to \mathrm{hom}_{\boldsymbol{K}}(d, a \times b)$. Thus $(\pi_a \circ p_d)(f,g) = f$ and $(\pi_b \circ p_d)(f,g) = g$ for every morphism $f : d \to a$ and $g : d \to b$. Actually, p_d is the component of a natural transformation $p : \boldsymbol{F} \to \boldsymbol{G}$ with $\boldsymbol{F} := \mathrm{hom}_{\boldsymbol{K}}(-,a) \times \mathrm{hom}_{\boldsymbol{K}}(-,b)$ and $\boldsymbol{G} := \mathrm{hom}_{\boldsymbol{K}}(-, a \times b)$ (note that this is shorthand for the obvious assignments to objects and functors). Both $\boldsymbol{F}$ and $\boldsymbol{G}$ are *contra*variant functors from $\boldsymbol{K}$ to $\boldsymbol{Set}$. So in order to establish naturalness, we have to establish that the following diagram commutes:

$$\begin{tikzcd}
c \arrow[d, "f"'] & \mathrm{hom}_{\boldsymbol{K}}(c,a) \times \mathrm{hom}_{\boldsymbol{K}}(c,b) \arrow[rr, "p_c"] & & \mathrm{hom}_{\boldsymbol{K}}(c, a \times b) \\
d & \mathrm{hom}_{\boldsymbol{K}}(d,a) \times \mathrm{hom}_{\boldsymbol{K}}(d,b) \arrow[rr, "p_d"'] \arrow[u, "\boldsymbol{F}\, f"] & & \mathrm{hom}_{\boldsymbol{K}}(d, a \times b) \arrow[u, "\boldsymbol{G}\, f"']
\end{tikzcd}$$

Now take $\langle g, h\rangle \in \mathrm{hom}_{\boldsymbol{K}}(d,a) \times \mathrm{hom}_{\boldsymbol{K}}(d,b)$; then

$$\pi_a\big((p_c \circ \boldsymbol{F}\, f)(g,h)\big) = g \circ f = \pi_a\big((\boldsymbol{G}\, f) \circ p_d\big)(g,h),$$
$$\pi_b\big((p_c \circ \boldsymbol{F}\, f)(g,h)\big) = h \circ f = \pi_b\big((\boldsymbol{G}\, f) \circ p_d\big)(g,h).$$

From this, commutativity follows. ✌

We will—for the sake of illustration—define two ways of composing natural transformations. One is somewhat canonical, since it is based on the composition of morphisms; the other one is a bit tricky, since it involves the functors directly. Let us have a look at the direct one first.

Lemma 2.3.22 *Let $\eta : \boldsymbol{F} \to \boldsymbol{G}$ and $\zeta : \boldsymbol{G} \to \boldsymbol{H}$ be natural transformations. Then*

$$(\vartheta \circ \zeta)_a := \vartheta_a \circ \zeta_a$$

defines a natural transformation $\vartheta \circ \zeta : \boldsymbol{F} \to \boldsymbol{H}$. $\vartheta \circ \zeta$

Proof Let $\boldsymbol{K}$ be the domain of functor $\boldsymbol{F}$, and assume that $f : a \to b$ is a morphism in $\boldsymbol{K}$. Then we have this diagram:

$$\begin{tikzcd} a \arrow[d, "f"'] & \boldsymbol{F}a \arrow[r, "\eta_a"] \arrow[d, "\boldsymbol{F}f"'] & \boldsymbol{G}a \arrow[r, "\zeta_a"] \arrow[d, "\boldsymbol{G}g"'] & \boldsymbol{H}a \arrow[d, "\boldsymbol{H}f"] \\ b & \boldsymbol{F}b \arrow[r, "\eta_b"'] & \boldsymbol{G}b \arrow[r, "\zeta_b"'] & \boldsymbol{H}b \end{tikzcd}$$

Then

$$\begin{aligned} \boldsymbol{H}(f) \circ (\vartheta \circ \zeta)_a &= \boldsymbol{H}(f) \circ \vartheta_a \circ \zeta_a = \vartheta_b \circ \boldsymbol{G}(f) \circ \zeta_a \\ &= \vartheta_b \circ \zeta_b \circ \boldsymbol{F}(f) = (\vartheta \circ \zeta)_b \circ \boldsymbol{F}(f). \end{aligned}$$

Hence the outer diagram commutes. ⊣

The next composition is slightly more involved.

Proposition 2.3.23 *Given natural transformations $\eta : \boldsymbol{F} \to \boldsymbol{G}$ and $\vartheta : \boldsymbol{S} \to \boldsymbol{R}$ for functors $\boldsymbol{F}, \boldsymbol{G} : \boldsymbol{K} \to \boldsymbol{L}$ and $\boldsymbol{S}, \boldsymbol{R} : \boldsymbol{L} \to \boldsymbol{M}$. Then $\vartheta_{\boldsymbol{G}a} \circ \boldsymbol{S}(\eta_a) = \boldsymbol{R}(\eta_a) \circ \vartheta_{\boldsymbol{F}a}$ always holds. Put*

$$(\vartheta * \eta)_a := \vartheta_{\boldsymbol{G}a} \circ \boldsymbol{S}(\eta_a).$$

$\vartheta * \eta$ *Then $\vartheta * \eta$ defines a natural transformation $\boldsymbol{S} \circ \boldsymbol{F} \to \boldsymbol{R} \circ \boldsymbol{G}$. $\vartheta * \eta$ is called the* Godement product *of η and ϑ.*

Proof 1. Because $\eta_a : \boldsymbol{F}a \to \boldsymbol{G}a$, this diagram commutes by naturality of ϑ:

$$\begin{tikzcd} \boldsymbol{S}(\boldsymbol{F}a) \arrow[r, "\vartheta_{\boldsymbol{F}a}"] \arrow[d, "\boldsymbol{S}(\eta_a)"'] & \boldsymbol{R}(\boldsymbol{F}a) \arrow[d, "\boldsymbol{R}(\eta_a)"] \\ \boldsymbol{S}(\boldsymbol{G}a) \arrow[r, "\vartheta_{\boldsymbol{G}a}"'] & \boldsymbol{R}(\boldsymbol{G}a) \end{tikzcd}$$

This establishes the first claim.

2. Now let $f : a \to b$ be a morphism in $\boldsymbol{K}$, then the outer diagram commutes, since $\boldsymbol{S}$ is a functor and since ϑ is a natural transformation.

$$\begin{tikzcd} \boldsymbol{S}(\boldsymbol{F}a) \arrow[r, "\boldsymbol{F}(\eta_a)"] \arrow[d, "\boldsymbol{S}(\boldsymbol{F}f)"'] \arrow[rr, dashed, bend left, "(\vartheta*\eta)_a"] & \boldsymbol{S}(\boldsymbol{G}a) \arrow[r, "\vartheta_{\boldsymbol{G}a}"] \arrow[d, "\boldsymbol{S}(\boldsymbol{G}f)"'] & \boldsymbol{R}(\boldsymbol{G}a) \arrow[d, "\boldsymbol{R}(\boldsymbol{G}f)"] \\ \boldsymbol{S}(\boldsymbol{F}b) \arrow[r, "\boldsymbol{F}(\eta_b)"'] \arrow[rr, dashed, bend right, "(\vartheta*\eta)_b"'] & \boldsymbol{S}(\boldsymbol{G}b) \arrow[r, "\vartheta_{\boldsymbol{G}b}"'] & \boldsymbol{R}(\boldsymbol{G}b) \end{tikzcd}$$

Hence $\vartheta * \eta : \boldsymbol{S} \circ \boldsymbol{F} \to \boldsymbol{R} \circ \boldsymbol{G}$ is natural indeed. ⊣

In [ML97], $\eta \circ \vartheta$ is called the *vertical* and $\eta * \vartheta$ the *horizontal* composition of the natural transformations η and ϑ. If $\eta : \boldsymbol{F} \to \boldsymbol{G}$ is a natural transformation, then the morphisms $(\boldsymbol{F}\eta)(a) := \boldsymbol{F}\eta_a : (\boldsymbol{F} \circ \boldsymbol{F})(a) \to (\boldsymbol{F} \circ \boldsymbol{G})(a)$ and $(\eta\boldsymbol{F})(a) := \eta_{\boldsymbol{F}a} : (\boldsymbol{F} \circ \boldsymbol{F})(a) \to (\boldsymbol{G} \circ \boldsymbol{F})(a)$ are available.

We know from Example 2.3.3 that $\text{hom}_{\boldsymbol{K}}(a, -)$ defines a covariant set-valued functor; suppose we have another set- valued functor $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{Set}$. Can we somehow compare these functors? This question looks on first sight quite strange, because we do not have any yardstick to compare these functors against. On second thought, we might use natural transformations for such an endeavor. It turns out that for any object a of $\boldsymbol{K}$, the set $\boldsymbol{F}a$ is essentially given by the natural transformations $\eta : \text{hom}_{\boldsymbol{K}}(a, -) \to \boldsymbol{F}$. We will show now that there exists a bijective assignment between $\boldsymbol{F}a$ and these natural transformations. The reader might wonder about this somewhat intricate formulation; it is due to the observation that these natural transformations in general do not form a set but rather a class, so that we cannot set up a proper bijection (which would require sets as the basic scenario).

Lemma 2.3.24 *Let* $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{Set}$ *be a functor; given the object* a *of* $\boldsymbol{K}$ *and a natural transformation* $\eta : \text{hom}_{\boldsymbol{K}}(a, -) \to F$*, define the* Yoneda isomorphism

$$y_{a,\boldsymbol{F}}(\eta) := \eta(a)(id_a) \in \boldsymbol{F}\, a$$

Then $y_{a,\boldsymbol{F}}$ *is bijective (i.e., onto and one to one).*

Proof 0. The assertion is established by defining for each $t \in \boldsymbol{F}a$ a natural transformation $\sigma_{a,\boldsymbol{F}}(t) : \text{hom}_{\boldsymbol{K}}(a, -) \to F$ which is inverse to $y_{a,\boldsymbol{F}}$. This is also the outline for the proof: we define a map, establish that it is a natural transformation, and show that it is inverse to $y_{a,\boldsymbol{F}}$. Outline for the proof

1. Given an object b of $\boldsymbol{K}$ and $t \in \boldsymbol{F}\, a$, put

$$\big(\sigma_{a,\boldsymbol{F}}(t)\big)_b := \sigma_{a,\boldsymbol{F}}(t)(b) : \begin{cases} \text{hom}_{\boldsymbol{K}}(a, b) & \to \boldsymbol{F}\, b \\ f & \mapsto (\boldsymbol{F}\, f)(t) \end{cases}$$

(note that $\boldsymbol{F}\, f : \boldsymbol{F}\, a \to \boldsymbol{F}\, b$ for $f : a \to b$; hence $(\boldsymbol{F}\, f)(t) \in \boldsymbol{F}\, b$). This defines a natural transformation $\sigma_{a,\boldsymbol{F}}(t) : \hom_{\boldsymbol{K}}(a, -) \to \boldsymbol{F}$. In fact, if $f : b \to b'$, then

$$\begin{aligned}
&\sigma_{a,\boldsymbol{F}}(t)(b')(\hom_{\boldsymbol{K}}(a, f)g) = \sigma_{a,\boldsymbol{F}}(t)(b')(f \circ g) = \boldsymbol{F}(f \circ g)(t) \\
&\quad = (\boldsymbol{F}\, f)(\boldsymbol{F}(g)(t)) = (\boldsymbol{F}\, f)(\sigma_{a,\boldsymbol{F}}(t)(b)(g)).
\end{aligned}$$

Hence $\sigma_{a,\boldsymbol{F}}(t)(b') \circ \hom_{\boldsymbol{K}}(a, f) = (\boldsymbol{F}\, f) \circ \sigma_{a,\boldsymbol{F}}(t)(b)$.

2. We obtain

$$\begin{aligned}
&(y_{a,\boldsymbol{F}} \circ \sigma_{a,\boldsymbol{F}})(t) = y_{a,\boldsymbol{F}}(\sigma_{a,\boldsymbol{F}}(t)) = \sigma_{a,\boldsymbol{F}}(t)(a)(id_a) \\
&\quad = (\boldsymbol{F}\, id_a)(t) = id_{\boldsymbol{F}a}(t) = t
\end{aligned}$$

That is not too bad; so let us try to establish that $\sigma_{a,\boldsymbol{F}} \circ y_{a,\boldsymbol{F}}$ is the identity as well. Given a natural transformation $\eta : \hom_{\boldsymbol{K}}(a, -) \to F$, we obtain

$$(\sigma_{a,\boldsymbol{F}} \circ y_{a,\boldsymbol{F}})(\eta) = \sigma_{a,\boldsymbol{F}}(y_{a,\boldsymbol{F}}(\eta)) = \sigma_{a,\boldsymbol{F}}(\eta_a(id_a)).$$

Thus we have to evaluate $\sigma_{a,\boldsymbol{F}}(\eta_a(id_a))$. Take an object b and a morphism $f : a \to b$; then

$$\begin{aligned}
\sigma_{a,\boldsymbol{F}}(\eta_a(id_a)(b)(f)) &= (\boldsymbol{F} f)(\eta_a(id_a)) \\
&= (\boldsymbol{F}(f) \circ \eta_a)(id_a) \\
&= (\eta_b \circ \hom_{\boldsymbol{K}}(a, f)) && (\eta \text{ is natural}) \\
&\quad (id_a) \\
&= \eta_b(f) && (\text{since } \hom_{\boldsymbol{K}}(a, f) \\
& && \quad \circ id_a = f \circ id_a = f)
\end{aligned}$$

Thus $\sigma_{a,\boldsymbol{F}}(\eta_a(id_a)) = \eta$. Consequently we have shown that $y_{a,\boldsymbol{F}}$ is left and right invertible, hence is a bijection. $\dashv$

Now consider the set-valued functor $\hom_{\boldsymbol{K}}(b, -)$; then the Yoneda embedding says that $\hom_{\boldsymbol{K}}(b, a)$ can be mapped bijectively to the natural transformations from $\hom_{\boldsymbol{K}}(a, -)$ to $\hom_{\boldsymbol{K}}(b, -)$. This means that these natural transformations are essentially the morphisms $b \to a$, and, conversely, each morphism $b \to a$ yields a natural transformation $\hom_{\boldsymbol{K}}(a, -) \to \hom_{\boldsymbol{K}}(b, -)$. The following statement makes this observation precise.

Proposition 2.3.25 *Given a natural transformation* $\eta : \mathrm{hom}_{\boldsymbol{K}}(a, -) \to \mathrm{hom}_{\boldsymbol{K}}(b, -)$*, there exists a unique morphism* $g : b \to a$ *such that* $\eta_c(h) = h \circ g$ *for every object* c *and every morphism* $h : a \to c$ *(thus* $\eta = \mathrm{hom}_{\boldsymbol{K}}(g, -)$*).*

Proof 0. Let $y := y_{a,\mathrm{hom}_{\boldsymbol{K}}(a,-)}$ and $\sigma := \sigma_{a,\mathrm{hom}_{\boldsymbol{K}}(a,-)}$. Then y is a bijection with $(y \circ \sigma)(\eta) = \eta$ and $(\sigma \circ y)(h) = h$.

1. Put $g := \eta_a(id_a)$, then $g \in \mathrm{hom}_{\boldsymbol{K}}(b, a)$, since $\eta_a : \mathrm{hom}_{\boldsymbol{K}}(a, a) \to \mathrm{hom}_{\boldsymbol{K}}(b, a)$ and $id_a \in \mathrm{hom}_{\boldsymbol{K}}(a, a)$. Now let $h \in \mathrm{hom}_{\boldsymbol{K}}(a, c)$, then

$$\begin{aligned}
\eta_c(h) &= \sigma\big(\eta_a(id_a)\big)(c)(h) && (\text{since } \eta = y \circ \sigma)\\
&= \sigma(g)(c)(h) && (\text{Definition of } \sigma)\\
&= \mathrm{hom}_{\boldsymbol{K}}(b, g)(h) && (\mathrm{hom}_{\boldsymbol{K}}(b, -) \text{ is the target functor})\\
&= h \circ g
\end{aligned}$$

2. If $\eta = \mathrm{hom}_{\boldsymbol{K}}(g, -)$, then $\eta_a(id_a) = \mathrm{hom}_{\boldsymbol{K}}(g, id_a) = id_a \circ g = g$, so $g : b \to a$ is uniquely determined. $\dashv$

A final example for natural transformations comes from measurable spaces, dealing with the weak σ-algebra. We consider in Example 2.3.5 the contravariant functor which assigns to each measurable space its σ-algebra, and we have defined in Example 2.1.14 the weak σ-algebra on its set of probability measures together with a set of generators. We show that this set of generators yields a family of natural transformations between the two contravariant functors involved.

Example 2.3.26 The contravariant functor $\boldsymbol{B} : \boldsymbol{Meas} \to \boldsymbol{Set}$ assigns to each measurable space its σ-algebra and to each measurable map its inverse. Denote by $\boldsymbol{W} := \mathbb{P} \circ \boldsymbol{B}$ the functor that assigns to each measurable space the weak σ-algebra on its probability measures; $\boldsymbol{W} : \boldsymbol{Meas} \to \boldsymbol{Set}$ is contravariant as well. Recall from Example 2.1.14 that the set

$$\boldsymbol{\beta}_{\mathcal{A}}(A, r) := \{\mu \in \mathbb{P}(S, \mathcal{A}) \mid \mu(A) \geq r\}$$

denotes the set of all probability measures which evaluate the measurable set A not smaller than a given r and that the weak σ-algebra on $\mathbb{P}(S, \mathcal{A})$ is generated by all these sets. We claim that $\beta(\cdot, r)$ is a natural transformation $\boldsymbol{B} \to \boldsymbol{W}$. Thus we have to show that this diagram commutes

$$
\begin{tikzcd}
(S,\mathcal{A}) \arrow[d, "f"'] \\
(T,\mathcal{B})
\end{tikzcd}
\qquad
\begin{tikzcd}
\boldsymbol{B}(S,\mathcal{A}) \arrow[r, "\beta_{\mathcal{A}}(\cdot,r)"] & \boldsymbol{W}(S,\mathcal{A}) \\
\boldsymbol{B}(T,\mathcal{B}) \arrow[u, "\boldsymbol{B}f"] \arrow[r, "\beta_{\mathcal{B}}(\cdot,r)"'] & \boldsymbol{W}(T,\mathcal{B}) \arrow[u, "\boldsymbol{W}f"']
\end{tikzcd}
$$

Recall that we have $\boldsymbol{B}(f)(C) = f^{-1}\big[C\big]$ for $C \in \mathcal{B}$ and that $\boldsymbol{W}(f)(D) = \mathbb{P}(f)^{-1}\big[D\big]$, if $D \subseteq \mathbb{P}(T,\mathcal{B})$ is measurable. Now, given $C \in \mathcal{B}$, by expanding definitions we obtain

$$
\begin{aligned}
\mu \in \boldsymbol{W}(f)(\boldsymbol{\beta}_{\mathcal{B}}(C,r)) &\Leftrightarrow \mu \in \mathbb{P}(f)^{-1}\big[\boldsymbol{\beta}_{\mathcal{B}}(C,r)\big] \\
&\Leftrightarrow \mathbb{P}(f)(\mu) \in \boldsymbol{\beta}_{\mathcal{B}}(C,r) \\
&\Leftrightarrow \mathbb{P}(f)(\mu)(C) \geq r \\
&\Leftrightarrow \mu(f^{-1}\big[C\big]) \geq r \\
&\Leftrightarrow \mu \in \boldsymbol{\beta}_{\mathcal{A}}(\boldsymbol{B}(f)(C), r)
\end{aligned}
$$

Thus the diagram commutes in fact, and we have established that the generators for the weak σ-algebra come from a natural transformation. ✌

2.3.3 Limits and Colimits

We define above some constructions which permit to build new objects in a category from given ones, e.g., the product from two objects or the pushout. Each time we had some universal condition which had to be satisfied.

We will discuss these general constructions very briefly and refer the reader to [ML97, BW99, Pum99], where they are studied in great detail.

Definition 2.3.27 *Given a functor $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$, a* cone on *$\boldsymbol{F}$ consists of an object c in $\boldsymbol{L}$ and of a family of morphisms $p_d : c \to \boldsymbol{F}d$ in $\boldsymbol{L}$ for each object d in $\boldsymbol{K}$ such that $p_{d'} = (\boldsymbol{F}g) \circ p_d$ for each morphism $g : d \to d'$ in $\boldsymbol{K}$.*

So a cone $(c, (p_d)_{d \in |\boldsymbol{K}|})$ on $\boldsymbol{F}$ looks like, well, a cone:

$$\begin{tikzcd}
& c \arrow[dl, "p_d"'] \arrow[dr, "p_{d'}"] & \\
\boldsymbol{F}d \arrow[rr, "\boldsymbol{F}g"'] & & \boldsymbol{F}d' \\
d \arrow[rr, dashed, "g"'] & & d'
\end{tikzcd}$$

A limiting cone provides a factorization for each other cone, to be specific:

Definition 2.3.28 *Let $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ be a functor. The cone $(c, (p_d)_{d \in |\boldsymbol{K}|})$ is a* limit of $\boldsymbol{F}$ *iff for every cone $(e, (q_d)_{d \in |\boldsymbol{K}|})$ on $\boldsymbol{F}$ there exists a unique morphism $f : e \to c$ such that $q_d = p_d \circ f$ for each object d in $\boldsymbol{K}$.*

Thus we have locally this situation for each morphism $g : d \to d'$ in $\boldsymbol{K}$:

$$\begin{tikzcd}
& e \arrow[d, dotted, "f"', "!"] \arrow[ddl, "q_d"'] \arrow[ddr, "q_{d'}"] & \\
& c \arrow[dl, "p_d"] \arrow[dr, "p_{d'}"'] & \\
\boldsymbol{F}d \arrow[rr, "\boldsymbol{F}g"'] & & \boldsymbol{F}d' \\
d \arrow[rr, dashed, "g"'] & & d'
\end{tikzcd}$$

The unique factorization probably gives already a clue for the application of this concept. Let us interpret two known examples in the light of this concept.

Example 2.3.29 Let $X := \{1, 2\}$ and $\boldsymbol{K}$ be the discrete category on X (see Example 2.1.6). Put $\boldsymbol{F}1 := a$ and $\boldsymbol{F}2 := b$ for the objects $a, b \in |\boldsymbol{L}|$. Assume that the product $a \times b$ with projections π_a and π_b exists in $\boldsymbol{L}$, and put $p_1 := \pi_a$, $p_2 := \pi_b$. Then $(a \times b, p_1, p_2)$ is a limit of $\boldsymbol{F}$. Clearly, this is a cone on $\boldsymbol{F}$, and if $q_1 : e \to a$ and $q_2 : e \to b$ are morphisms, there exists a unique morphism $f : e \to a \times b$ with $q_1 = p_1 \circ f$ and $q_2 = p_2 \circ f$ by the definition of a product. ✌

The next example shows that a pullback can be interpreted as a limit.

Example 2.3.30 Let a, b, c objects in category $\boldsymbol{L}$ with morphisms $f : a \to c$ and $g : b \to c$. Define category $\boldsymbol{K}$ by $|\boldsymbol{K}| := \{a, b, c\}$; the hom sets are defined as follows:

$$\hom_{\boldsymbol{K}}(x, y) := \begin{cases} \{id_x\}, & x = y \\ \{f\}, & x = a, y = c \\ \{g\}, & x = b, y = c \\ \emptyset, & \text{otherwise} \end{cases}$$

Let $\boldsymbol{F}$ be the identity on $|\boldsymbol{K}|$ with $\boldsymbol{F}f := f, \boldsymbol{F}g := g$, and $\boldsymbol{F}id_x := id_x$ for $x \in |\boldsymbol{K}|$. If object p together with morphisms $t_a : p \to a$ and $t_b : p \to b$ is a pullback for f and g, then it is immediate that (p, t_a, t_b, t_c) is a limit cone for $\boldsymbol{F}$, where $t_c := f \circ t_a = g \circ t_b$. ✌

Dualizing the concept of a cone, we obtain cocones.

Definition 2.3.31 *Given a functor $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$, an object $c \in |\boldsymbol{L}|$ together with morphisms $s_d : \boldsymbol{F}D \to c$ for each object d of $\boldsymbol{K}$ such that $s_d = s_{d'} \circ \boldsymbol{F}g$ for each morphism $g : d \to d'$ is called a* cocone on $\boldsymbol{F}$.

Thus we have this situation

$$\begin{tikzcd}
d \arrow[rr, dashed, "g"] & & d' \\
\boldsymbol{F}d \arrow[rr, "\boldsymbol{F}g"] \arrow[dr, "s_d"'] & & \boldsymbol{F}d' \arrow[dl, "s_{d'}"] \\
& c &
\end{tikzcd}$$

A colimit is then defined for a cocone.

Definition 2.3.32 *A cocone $(c, (s_d)_{d \in |\boldsymbol{K}|})$ is called a* colimit *for the functor $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ iff for every cocone $(e, (t_d)_{d \in |\boldsymbol{K}|})$ for $\boldsymbol{F}$ there exists a unique morphism $f : c \to e$ such that $t_d = f \circ s_d$ for every object $d \in |\boldsymbol{K}|$.*

So this yields

$$\begin{tikzcd}
d \arrow[rr, dashed, "g"] & & d' \\
\boldsymbol{F}d \arrow[rr, "\boldsymbol{F}g"] \arrow[dr, "s_d"] \arrow[ddr, "t_d"'] & & \boldsymbol{F}d' \arrow[dl, "s_{d'}"'] \arrow[ddl, "t_{d'}"] \\
& c \arrow[d, dotted, "f"', "!"] & \\
& e &
\end{tikzcd}$$

Coproducts are examples of cocones.

Example 2.3.33 Let a and b be objects in category $\boldsymbol{L}$ and assume that their coproduct $a + b$ with injections j_a and j_b exists in $\boldsymbol{L}$. Take again $I := \{1, 2\}$, and let $\boldsymbol{K}$ be the discrete category over I. Put $\boldsymbol{F}1 := a$ and $\boldsymbol{F}2 := b$; then it follows from the definition of the coproduct that the cocone $(a + b, j_a, j_b)$ is a colimit for $\boldsymbol{F}$. ✌

One shows that the pushout can be represented as a colimit in the same way as in Example 2.3.30 for the representation of the pullback as a limit.

Both limits and colimits are powerful general concepts for representing important constructions with and on categories. We will encounter them later on, albeit mostly indirectly.

2.4 Monads and Kleisli Tripels

We have now functors and natural transformations at our disposal, and we will put them to work. The first application we will tackle concerns monads. Moggi's work [Mog91, Mog89] shows a connection between monads and computation which we will discuss now. Kleisli tripels as a practical disguise for monads are introduced first, and it will be shown through Manes' Theorem that they are equivalent in the sense that each Kleisli tripel generates a monad, and vice versa, in a reversible construction. Some examples for monads follow, and we will finally have a brief look at the monadic construction in the programming language `Haskell`.

2.4.1 Kleisli Tripels

Assume that we work in a category $\boldsymbol{K}$ and interpret values and computations of a programming language in $\boldsymbol{K}$. We need to distinguish between the values of type a and the computations of type a, which are of type $\boldsymbol{T}a$. For example: Values vs. computations

Nondeterministic computations Taking the values from set A yields computations of type $\boldsymbol{T}A = \mathcal{P}_f(A)$, where the latter denotes all finite subsets of A.

Probabilistic computations Taking values from set A will give computations in the set $\boldsymbol{T}A = \boldsymbol{D}A$ of all discrete probabilities on A; see Example 2.3.11.

Exceptions Here, values of type A will result in values taken from $\boldsymbol{T}A = A + E$ with E as the set of *exceptions*.

Side effects Let L be the set of addresses in the store and U the set of all storage cells; a computation of type A will assign each element of U^L an element of A or another element of U^L; thus we have $\boldsymbol{T}A = (A + U^L)^{U^L}$.

Interactive input Let U be the set of characters; then $\boldsymbol{T}A$ is the set of all trees with finite fan out, so that the internal nodes have labels coming from U and the leaves have labels taken from A.

In order to model this, we require an embedding of the values taken from a into the computations of type $\boldsymbol{T}a$, which is represented as a morphism $\eta_a : a \to \boldsymbol{T}a$. Moreover, we want to be able to "lift" values to computations in this sense: if $f : a \to \boldsymbol{T}b$ is a map from values to computations, we want to extend f to a map $f^* : \boldsymbol{T}a \to \boldsymbol{T}b$ from computations to computations (thus we will be able to combine computations in a modular fashion). Understanding a morphism $a \to \boldsymbol{T}b$ as a program performing computations of type b on values of type a, this lifting will then permit performing computations of type b depending on computations of type a.

This leads to the definition of a Kleisli tripel.

Kleisli tripel

Definition 2.4.1 *Let $\boldsymbol{K}$ be a category. A* Kleisli tripel $(\boldsymbol{T}, \eta, -^*)$ *over $\boldsymbol{K}$ consists of a map $\boldsymbol{T} : |\boldsymbol{K}| \to |\boldsymbol{K}|$ on objects, a morphism $\eta_a : a \to \boldsymbol{T}a$ for each object a, and an operation * such that $f^* : \boldsymbol{T}a \to \boldsymbol{T}b$, if $f : a \to \boldsymbol{T}b$ with the following properties:*

① $\eta_a^* = id_{\boldsymbol{T}a}$.

② $f^* \circ \eta_a = f$*, provided* $f : a \to \boldsymbol{T}b$.

③ $g^* \circ f^* = (g^* \circ f)^*$ *for* $f : a \to \boldsymbol{T}b$ *and* $g : b \to \boldsymbol{T}c$.

The first property says that lifting the embedding $\eta_a : a \to \boldsymbol{T}a$ will give the identity on $\boldsymbol{T}a$. The second condition says that applying the lifted morphism f^* to an embedded value η_a will yield the same value as the given f. The third condition says that combining lifted morphisms is the same as lifting the lifted second morphism applied to the value of the first morphism.

The category associated with a Kleisli tripel has the same objects as the originally given category (which is not too much of a surprise), but morphisms will correspond to programs: a program which performs a computation of type b on values of type a. Hence a morphism in this new category is of type $a \to \boldsymbol{T}b$ (this is a morphism on $\boldsymbol{K}$).

Definition 2.4.2 *Given a Kleisli tripel* $(\boldsymbol{T}, \eta, -^*)$ *over category* $\boldsymbol{K}$*, the* Kleisli category $\boldsymbol{K_T}$ *is defined as follows:* $\boldsymbol{K_T}$

- $|\boldsymbol{K_T}| = |\boldsymbol{K}|$*; thus* $\boldsymbol{K_T}$ *has the same objects as* $\boldsymbol{K}$*.*
- $\mathrm{hom}_{\boldsymbol{K_T}}(a, b) = \mathrm{hom}_{\boldsymbol{K}}(a, \boldsymbol{T}b)$*; hence* f *is a morphism* $a \to b$ *in* $\boldsymbol{K_T}$ *iff* $f : a \to \boldsymbol{T}b$ *is a morphism in* $\boldsymbol{K}$*.*
- *The identity for* a *in* $\boldsymbol{K_T}$ *is* $\eta_a : a \to \boldsymbol{T}a$*.*
- *The composition* $g * f$ *of* $f \in \mathrm{hom}_{\boldsymbol{K_T}}(a, b)$ *and* $g \in \mathrm{hom}_{\boldsymbol{K_T}}(b, c)$ *is defined through* $g * f := g^* \circ f$*.*

We have to show that Kleisli composition is associative: in fact, we have

$$\begin{aligned}
(h * g) * f &= (h * g)^* \circ f \\
&= (h^* \circ g)^* \circ f && \text{(definition of } h * g\text{)} \\
&= h^* \circ g^* \circ f && \text{(property ③)} \\
&= h^* \circ (g * f) && \text{(definition of } g * f\text{)} \\
&= h * (f * g)
\end{aligned}$$

Thus $\boldsymbol{K_T}$ is indeed a category. The map on objects in a Kleisli category extends to a functor (note that we did not postulate for a Kleisli tripel that $\boldsymbol{T}f$ is defined for morphisms). This functor is associated with two natural transformations which together form a monad. We will first define what a monad formally is and then discuss the construction in some detail.

2.4.2 Monads

Definition 2.4.3 *A* monad *over a category* $\boldsymbol{K}$ *is a triple* $(\boldsymbol{T}, \eta, \mu)$ *with these properties:*

❶ *$\boldsymbol{T}$ is an endofunctor on $\boldsymbol{K}$.*

❷ *$\eta : Id_{\boldsymbol{K}} \to \boldsymbol{T}$ and $\mu : \boldsymbol{T}^2 \to \boldsymbol{T}$ are natural transformations. η is called the* unit *and μ the* multiplication *of the monad.*

❸ *These diagrams commute*

$$\begin{tikzcd}
\boldsymbol{T}^3\, a \arrow[r, "\mu_{\boldsymbol{T}a}"] \arrow[d, "\boldsymbol{T}\mu_a"'] & \boldsymbol{T}^2\, a \arrow[d, "\mu_a"] \\
\boldsymbol{T}^2\, a \arrow[r, "\mu_a"'] & \boldsymbol{T}\, a
\end{tikzcd}
\qquad
\begin{tikzcd}
\boldsymbol{T}\, a \arrow[r, "\eta_{\boldsymbol{T}a}"] \arrow[dr, "id_{\boldsymbol{T}a}"'] & \boldsymbol{T}^2\, a \arrow[d, "\mu_a"] & \boldsymbol{T}\, a \arrow[l, "\boldsymbol{T}\eta_a"'] \arrow[dl, "id_{\boldsymbol{T}a}"] \\
& \boldsymbol{T}\, a &
\end{tikzcd}$$

Each Kleisli tripel generates a monad, and vice versa. This is what Manes' Theorem says:

Theorem 2.4.4 *Given a category $\boldsymbol{K}$, there is a one-to-one correspondence between Kleisli tripels and monads.*

Outline and strategy

Proof 0. The proof will be somewhat longish, because so many properties have to be established or checked. In the first part, $\boldsymbol{T}$ will be extended to a functor, and a multiplication will be defined (the unit remains what it is in the Kleisli tripel), and the laws for a monad will be established. The second part will define the $-^*$ operation and establish the corresponding properties of a Kleisli tripel; again, the unit remains what it is in the monad. The proof as a whole demonstrates the interaction of the concepts in a very clean way; this is why I did not put it into separate pieces.

1. Let $(\boldsymbol{T}, \eta, -^*)$ be a Kleisli tripel. We will extend $\boldsymbol{T}$ to a functor $\boldsymbol{K} \to \boldsymbol{K}$ and define the multiplication; the monad's unit will be η. Define

$$\begin{aligned}
\boldsymbol{T}f &:= (\eta_b \circ f)^*, \text{ if } f : a \to b,\\
\mu_a &:= (id_{\boldsymbol{T}a})^*.
\end{aligned}$$

Then μ is a natural transformation $\boldsymbol{T}^2 \to \boldsymbol{T}$. Clearly, $\mu_a : \boldsymbol{T}^2 a \to \boldsymbol{T}a$ is a morphism. Let $f : a \to b$ be a morphism in $\boldsymbol{K}$; then we have

$$\begin{aligned}
\mu_b \circ \boldsymbol{T}^2 f &= id_{\boldsymbol{T}b}^* \circ (\eta_{\boldsymbol{T}b} \circ (\eta_b \circ f)^*)^* \\
&= \left(id_{\boldsymbol{T}b}^* \circ (\eta_{\boldsymbol{T}b} \circ (\eta_b \circ f)^*)\right)^* && \text{(by ③)} \\
&= \left(id_{\boldsymbol{T}b} \circ (\eta_b \circ f)^*\right)^* && \text{(since } id_{\boldsymbol{T}b}^* \circ \eta_{\boldsymbol{T}b} = id_{\boldsymbol{T}b}\text{)} \\
&= (\eta_b \circ f)^{**}.
\end{aligned}$$

Similarly, we obtain

$$(\boldsymbol{T}f) \circ \mu_a = (\eta_b \circ f)^* \circ id^*_{\boldsymbol{T}a} = \big((\eta_b \circ f)^* \circ id_{\boldsymbol{T}a}\big)^* = (\eta_b \circ f)^{**}.$$

Hence $\mu : \boldsymbol{T}^2 \to \boldsymbol{T}$ is natural. Because we obtain for the morphisms $f : a \to b$ and $g : b \to c$ the identity

$$\begin{aligned}(\boldsymbol{T}g) \circ (\boldsymbol{T}f) &= (\eta_c \circ g)^* \circ (\eta_b \circ f)^* = \big((\eta_c \circ g)^* \circ \eta_b \circ f\big)^* \\ &= (\eta_c \circ g \circ f)^* = \boldsymbol{T}(g \circ f),\end{aligned}$$

and since by ①

$$\boldsymbol{T}\, id_a = (\eta_a \circ id_{\boldsymbol{T}a})^* = \eta^*_a = id_{\boldsymbol{T}a},$$

we conclude that $\boldsymbol{T}$ is an endofunctor on $\boldsymbol{K}$.

We check the laws for unit and multiplication according to ❸. One notes first that

$$\mu_a \circ \eta_{\boldsymbol{T}a} = id^*_{\boldsymbol{T}a} \circ \eta_{\boldsymbol{T}a} \overset{(\ddagger)}{=} id_{\boldsymbol{T}a}$$

(in equation (‡) we use ②) and that

$$\begin{aligned}\mu_a \circ \boldsymbol{T}a &= id^*_{\boldsymbol{T}a}(\eta_{\boldsymbol{T}a} \circ \eta_a)^* = (id^*_{\boldsymbol{T}a} \circ \eta_{\boldsymbol{T}a} \circ \eta_a)^* \overset{(\dagger)}{=} \eta^*_a \\ &= (\eta_a \circ id_a)^* = T(id_a)\end{aligned}$$

(in equation (†) we use ② again). Hence the rightmost diagram in ❸ commutes. Turning to the leftmost diagram, we note that

$$\mu_a \circ \mu_{\boldsymbol{T}a} = id^*_{\boldsymbol{T}a} \circ id^*_{\boldsymbol{T}^2a} = (id^*_{\boldsymbol{T}a} \circ id_{\boldsymbol{T}^2a})^* \overset{(\%)}{=} \mu^*_a,$$

using ③ in equation (%). On the other hand,

$$\begin{aligned}\mu_a \circ (\boldsymbol{T}\,\mu_a) &= id^*_{\boldsymbol{T}a} \circ (\boldsymbol{T}\, id^*_{\boldsymbol{T}a}) = id^*_{\boldsymbol{T}a} \circ (\eta_{\boldsymbol{T}a} \circ id^*_{\boldsymbol{T}a})^* \\ &= {id^*_{\boldsymbol{T}a}}^* = \mu^*_a,\end{aligned}$$

because $id^*_{\boldsymbol{T}a} \circ \eta_{\boldsymbol{T}a} = id_{\boldsymbol{T}a}$ by ②. Hence the leftmost diagram commutes as well, and we have indeed defined a monad.

2. To establish the converse, define $f^* := \mu_b \circ (\boldsymbol{T}f)$ for the morphism $f : a \to \boldsymbol{T}b$. We obtain from the right-hand triangle $\eta^*_a = \mu_a \circ (\boldsymbol{T}\eta_a) = id_{\boldsymbol{T}a}$; thus ① holds. Since $\eta : Id_{\boldsymbol{K}} \to \boldsymbol{T}$ is natural, we have $(\boldsymbol{T}f) \circ \eta_a = \eta_{\boldsymbol{T}b} \circ f$ for $f : a \to \boldsymbol{T}b$. Hence

$$f^* \circ \eta_a = \mu_b \circ (\boldsymbol{T}f) \circ \eta_a = \mu_b \circ \eta_{\boldsymbol{T}b} \circ f = f$$

by the left-hand side of the right triangle, giving ②. Finally, note that due to $\mu : \boldsymbol{T}^2 \to \boldsymbol{T}$ being natural, we have for $g : b \to \boldsymbol{T}c$ the commutative diagram

$$\begin{tikzcd} \boldsymbol{T}^2 b \arrow[r, "\mu_b"] \arrow[d, "\boldsymbol{T}^2 g"'] & \boldsymbol{T} b \arrow[d, "\boldsymbol{T} g"] \\ \boldsymbol{T}^3 c \arrow[r, "\mu_{\boldsymbol{T}c}"'] & \boldsymbol{T}^2 c \end{tikzcd}$$

Then

$$\begin{aligned} g^* \circ f^* &= \mu_c \circ (\boldsymbol{T}g) \circ \mu_b \circ (\boldsymbol{T}f) \\ &= \mu_c \circ \mu_{\boldsymbol{T}c} \circ (\boldsymbol{T}^2 g) \circ (\boldsymbol{T}f) && (\text{since } (\boldsymbol{T}g) \circ \mu_b \\ & && = \mu_{\boldsymbol{T}c} \circ \boldsymbol{T}^2 g) \\ &= \mu_c \circ (\boldsymbol{T}\mu_c) \circ \boldsymbol{T}(\boldsymbol{T}(g) \circ f) && (\text{since } \mu_c \circ \mu_{\boldsymbol{T}c} \\ & && = \mu_c \circ (\boldsymbol{T}\mu_c)) \\ &= \mu_c \circ \boldsymbol{T}(\mu_c \circ \boldsymbol{T}(g) \circ f) \\ &= \mu_c \circ \boldsymbol{T}(g^* \circ f) \\ &= (g^* \circ f)^* \end{aligned}$$

This establishes ③ and shows that this defines a Kleisli tripel. ⊣

Taking a Kleisli tripel and producing a monad from it, one suspects that one might end up with a different Kleisli tripel for the generated monad. But this is not the case; just for the record:

Corollary 2.4.5 *If the monad is given by a Kleisli tripel, then the Kleisli tripel defined by the monad coincides with the given one. Similarly, if the Kleisli tripel is given by the monad, then the monad defined by the Kleisli tripel coincides with the given one.*

Proof We use the notation from above. Given the monad, put $f^+ := id_{\boldsymbol{T}b} \circ (\eta_b \circ f)^*$; then

$$f^+ = \mu_{\boldsymbol{T}b} \circ (\eta_b \circ f)^* = (id_{\boldsymbol{T}a} \circ f)^* = f^*.$$

On the other hand, given the Kleisli tripel, put $\boldsymbol{T}_0 f := (\eta_b \circ f)^*$; then

$$\boldsymbol{T}_0 f = \mu_b \circ \boldsymbol{T}(\eta_b \circ f) = \mu_b \circ \boldsymbol{T}(\eta_b) = \boldsymbol{T}f.$$

⊣

Some examples explain this development. Theorem 2.4.4 tells us that the specification of a Kleisli tripel will give us the monad, and vice versa. Thus we are free to specify one or the other; usually the specification of the Kleisli tripel is shorter and more concise.

Example 2.4.6 Nondeterministic computations may be modeled through a map $f : S \to \mathcal{P}(T)$: given a state (or an input, or whatever) from set S, the set $f(s)$ describes the set of all possible outcomes. Thus we work in category ***Set*** with maps as morphisms and take the power set functor $\mathcal{P}$ as the functor. Define

$$\begin{aligned} \eta_S(x) &:= \{x\}, \\ f^*(B) &:= \bigcup_{x \in B} f(x) \end{aligned}$$

for the set S, for $B \subseteq S$ and the map $f : S \to \mathcal{P}(T)$. Then clearly $\eta_S : S \to \mathcal{P}(S)$, and $f^* : \mathcal{P}(S) \to \mathcal{P}(T)$. We check the laws for a Kleisli tripel:

① Since $\eta_S^*(B) = \bigcup_{x \in B} \eta_S(x) = B$, we see that $\eta_S^* = id_{\mathcal{P}(S)}$.

② It is clear that $f^* \circ \eta_a = f$ holds for $f : S \to \mathcal{P}(S)$.

③ Let $f : S \to \mathcal{P}(T)$ and $g : T \to \mathcal{P}(U)$, then

$$\begin{aligned} u \in (g^* \circ f^*)(B) &\Leftrightarrow u \in g(y) \text{ for some } x \in B \\ &\qquad \text{and some } y \in f(x) \\ &\Leftrightarrow u \in g^*(f(x)) \text{ for some } x \in B \end{aligned}$$

Thus $(g^* \circ f^*)(B) = (g^* \circ f)^*(B)$.

Hence the laws for a Kleisli tripel are satisfied. Let us just compute $\mu_S = id^*_{\mathcal{P}(S)}$: Given $\beta \in \mathcal{P}(\mathcal{P}(S))$, we obtain

$$\mu_S(\beta) = id^*_{\mathcal{P}(S)} = \bigcup_{B \in \beta} B = \bigcup \beta.$$

The same argumentation can be carried out when the power set functor is replaced by the finite power set functor $\mathcal{P}_f : S \mapsto \{A \subseteq S \mid A \text{ is finite}\}$ with the obvious definition of $\mathcal{P}_f$ on maps. ✌

In contrast to nondeterministic computations, probabilistic ones argue with probability distributions. We consider the discrete case first, and here we focus on probabilities with finite support.

Example 2.4.7 We work in the category ***Set*** of sets with maps as morphisms and consider the discrete probability functor $\boldsymbol{D}S := \{p : S \to [0, 1] \mid p \text{ is a discrete probability}\}$; see Example 2.3.11. Let $f : S \to \boldsymbol{D}S$ be a map and $p \in \boldsymbol{D}S$, put

$$f^*(p)(s) := \sum_{t \in S} f(t)(s) \cdot p(t).$$

Then

$$\begin{aligned}\sum_{s \in S} f^*(p)(s) &= \sum_s \sum_t f(t)(s) \cdot p(t) = \sum_t \sum_s f(t)(s) \cdot p(t) \\ &= \sum_t p(t) = 1;\end{aligned}$$

hence $f^* : \boldsymbol{D}S \to \boldsymbol{D}S$. Note that the set $\{\langle s, t\rangle \in S \times T \mid f(s)(t) \cdot p(s) > 0\}$ is finite, because p has finite support and because each $f(s)$ has finite support as well. Since each of the summands is nonnegative, we may reorder the summations at our convenience. Define moreover

$$\eta_S(s)(s') := d_S(s)(s') := \begin{cases} 1, & s = s' \\ 0, & \text{otherwise,}\end{cases}$$

so that $\eta_S(s)$ is the discrete *Dirac measure* on s. Then:

① $\eta_S^*(p)(s) = \sum_{s'} d_S(s)(s') \cdot p(s') = p(s)$; hence we may conclude that $\eta_S^* \circ p = p$.

② $f^*(\eta_S)(s) = f(s)$ is immediate.

③ Let $f : S \to \boldsymbol{D}T$ and $g : T \to \boldsymbol{D}U$; then we have for $p \in \boldsymbol{D}S$ and $u \in U$

$$\begin{aligned}(g^* \circ f^*)(p)(u) &= \textstyle\sum_{t \in T} g(t)(u) \cdot f^*(p)(u) \\ &= \textstyle\sum_{t \in T} \sum_{s \in S} g(t)(u) \cdot f(s)(t) \cdot p(s) \\ &= \textstyle\sum_{\langle s,t\rangle \in S \times T} g(t)(u) \cdot f(s)(t) \cdot p(s) \\ &= \textstyle\sum_{s \in S} \big[\sum_{t \in T} g(t)(u) \cdot f(s)(t)\big] \cdot p(s) \\ &= \textstyle\sum_{s \in S} g^*(f(s))(u) \cdot p(s) \\ &= (g^* \circ f)^*(p)(u)\end{aligned}$$

Again, we are not bound to any particular order of summation.

We obtain for $M \in (\boldsymbol{D} \circ \boldsymbol{D})S$

$$\mu_S(M)(s) = id^*_{\boldsymbol{D}S}(M)(s) = \sum_{q \in \boldsymbol{D}(S)} M(q) \cdot q(s).$$

The last sum extends over a finite set, because the support of M is finite. ✌

Since programs may fail to halt, one works sometimes in models which are formulated in terms of subprobabilities rather than probabilities. This is what we consider next, extending the previous example to the case of general measurable spaces. Recall that examples requiring techniques from Chap. 4 are marked.

Example 2.4.8 Ⓢ We work in the category of measurable spaces with measurable maps as morphisms; see Example 2.1.12. In Example 2.3.12, the subprobability functor was introduced, and it was shown that for a measurable space S, the set $\mathbb{S}S$ of all subprobabilities is a measurable space again (we omit in this example the σ-algebra from notation, a measurable space is for the time being a pair consisting of a carrier set and a σ-algebra on it). A probabilistic computation f on the measurable spaces S and T produces from an input of an element of S a subprobability distribution $f(s)$ on T, hence an element of $\mathbb{S}T$. We want f to be a morphism in ***Meas***, so $f : S \to \mathbb{S}T$ is assumed to be measurable.

We know from Example 2.1.14 and Exercise 2.7 that $f : S \to \mathbb{S}T$ is measurable iff these conditions are satisfied:

1. $f(s) \in \mathbb{S}(T)$ for all $s \in S$; thus $f(s)$ is a subprobability on (the measurable sets of) T.
2. For each measurable set D in T, the map $s \mapsto f(s)(D)$ is measurable.

Returning to the definition of a Kleisli tripel, we define for the measurable space S, $f : S \to \mathbb{S}T$,

$$\begin{aligned} e_S &:= \delta_S, \\ f^*(\mu)(B) &:= \int_S f(s)(B)\, \mu(ds) \quad (\mu \in \mathbb{S}S, B \subseteq T \text{ measurable}). \end{aligned}$$

Thus $e_S(x) = \delta_S(x)$, the Dirac measure associated with x, and $f^* : \mathbb{S}S \to \mathbb{S}S$ is a morphism (in this example, we write e for the unit and m

for the multiplication). Note that $f^*(\mu) \in \mathbb{S}(T)$ in the scenario above; in order to see whether the properties of a Kleisli tripel are satisfied, we need to know how to integrate with this measure. Standard arguments like Levi's Theorem 4.8.2 show that

$$\int_T h \, df^*(\mu) = \int_S \int_T h(t) \, f(s)(dt) \, \mu(ds), \tag{2.1}$$

whenever $h : T \to \mathbb{R}_+$ is measurable and bounded; see also the discussion leading to Eq. (4.20) on page 633.

Let us again check the properties of a Kleisli tripel. Fix B as a measurable subset of S, $f : S \to \mathbb{S}S$ and $g : T \to \mathbb{S}U$ as morphisms in ***Meas***.

① Let $\mu \in \mathbb{S}S$; then

$$e_S^*(\mu)(B) = \int_S \delta_S(x)(B) \, \mu(dx) = \mu(B);$$

hence $e_S^* = id_{\mathbb{S}S}$.

② If $x \in S$, then

$$f^*(e_S(x))(B) = \int_S f(s)(B) \, \delta_S(x)(ds) = f(x)(B),$$

since $\int_S h \, d\delta_S(x) = h(x)$ for every measurable map h. Thus $f^* \circ e_S = f$.

③ Given $\mu \in \mathbb{S}S$, we have

$$\begin{aligned}
(g^* \circ f^*)(\mu)(B) &= g^*(f^*(\mu))(B) \\
&= \int_T g(t)(B) \, f^*(\mu)(dt) \\
&\overset{(2.1)}{=} \int_S \int_T g(t)(B) \, f(s)(dt) \, \mu(ds) \\
&= \int_S g^*(f(s))(B) \, \mu(ds) \\
&= (g^* \circ f)^*(\mu)(B)
\end{aligned}$$

Thus $g^* \circ f^* = (g^* \circ f)^*$.

Hence $(\mathbb{S}, e, -^*)$ forms a Kleisli tripel over the category ***Meas*** of measurable spaces.

Let us finally determine the monad's multiplication. We have for $M \in (\mathbb{S} \circ \mathbb{S})S$ and the measurable set $B \subseteq S$

$$m_S(M)(B) = id^*_{\mathbb{S}(S)}(M)(B) = \int_{\mathbb{S}(S)} \tau(B)\ M(d\tau).$$

✌

The underlying monad has been investigated by M. Giry, so it is called in her honor the *Giry monad*, and $\mathbb{S}$ is called the *Giry functor*. Both are used extensively as the machinery on which Markov transition systems are based. Giry monad

The next example shows that ultrafilter defines a monad as well.

Example 2.4.9 Let $\boldsymbol{U}$ be the ultrafilter functor on ***Set***; see Example 2.3.14. Define for the set S and the map $f : S \to \boldsymbol{U}T$

$$\begin{aligned} \eta_S(s) &:= \{A \subseteq S \mid s \in A\}, \\ f^*(U) &:= \big\{B \subseteq T \mid \{s \in S \mid B \in f(s)\} \in U\big\}, \end{aligned}$$

provided $U \in \boldsymbol{U}S$ is an ultrafilter. Then $\emptyset \notin f^*(U)$, since $\emptyset \notin U$. $\eta_S(s)$ is the principal ultrafilter associated with $s \in S$ (see page 42); hence $\eta_S : S \to \boldsymbol{U}S$. Because the intersection of two sets is a member of an ultrafilter iff both sets are elements of it,

$$\{s \in S \mid B_1 \cap B_2 \in f(s)\} = \{s \in S \mid B_1 \in f(s)\} \cap \{s \in S \mid B_2 \in f(s)\},$$

$f^*(U)$ is closed under intersections; moreover, $B \subseteq C$ and $B \in f^*(U)$ imply $C \in f^*(U)$. If $B \notin f^*(U)$, then $\{s \in S \mid f(s) \in B\} \notin U$; hence $\{s \in S \mid B \notin f(s)\} \in U$; thus $S \setminus B \in f^*(U)$, and vice versa. Hence $f^*(U)$ is an ultrafilter; thus $f^* : \boldsymbol{U}S \to \boldsymbol{U}T$.

We check whether $(\boldsymbol{U}, \eta, -^*)$ is a Kleisli tripel:

① Since $B \in \eta^*_S(U)$ iff $B = \{s \in S \mid s \in B\} \in U$, we conclude that $\eta^*_S = id_{\boldsymbol{U}S}$.

② Similarly, if $f : S \to \boldsymbol{U}T$ and $s \in S$, then $B \in (f^* \circ \eta_S)(s)$ iff $B \in f(s)$; hence $f^* \circ \eta_S = f$.

③ Let $f : S \to \boldsymbol{U}T$ and $g : T \to \boldsymbol{U}W$. Then

$$B \in (g^* \circ f^*)(U) \Leftrightarrow \bigl\{s \in S \mid \{t \in T \mid B \in g(t)\} \in f(s)\bigr\} \in U$$
$$\Leftrightarrow B \in (g^* \circ f)^*(U)$$

for $U \in \boldsymbol{U}S$. Consequently, $g^* \circ f^* = (g^* \circ f)^*$.

Let us compute the monad's multiplication. Define for $B \subseteq S$ the set

$$[B] := \{C \in \boldsymbol{U}S \mid B \in C\}$$

as the set of all ultrafilters on S which contain B as an element; then an easy computation shows

$$\mu_S(V) = id^*_{\boldsymbol{U}S}(V) = \{B \subseteq S \mid [B] \in V\}$$

for $V \in (\boldsymbol{U} \circ \boldsymbol{U})S$. ✌

Example 2.4.10 This example deals with upper closed subsets of the power set of a set; see Example 2.3.13. Let again

$$\boldsymbol{V}S := \{V \subseteq \mathcal{P}S \mid V \text{ is upper closed}\}$$

be the endofunctor on ***Set*** which assigns to set S all upper closed subsets of $\mathcal{P}S$. We define the components of a Kleisli tripel as follows: $\eta_S(s)$ is the principal ultrafilter generated by $s \in S$, which is upper closed, and if $f : S \to \boldsymbol{V}T$ is a map, we put

$$f^*(V) := \bigl\{B \subseteq T \mid \{s \in S \mid B \in f(s)\} \in V\bigr\}$$

for $V \in \boldsymbol{V}T$; see in Example 2.4.9.

The argumentation in Example 2.4.9 carries over and shows that this defines a Kleisli tripel. ✌

These examples show that monads and Kleisli tripels are constructions which model many computationally interesting subjects. After looking at the practical side of this, we return to the discussion of the relationship of monads with adjunctions, another important concept.

2.4.3 Monads in `Haskell`

The functional programming language `Haskell` thrives on the construction of monads. We have a brief look.

`Haskell` permits the definition of type classes; the definition of a type class requires the specification of the types on which the class is based and the signature of the functions defined by this class. The definition of class `Monad` is given below (actually, it is rather a specification of Kleisli tripels).

```
class Monad m where
  (>>=)  :: m a -> (a -> m b) -> m b
  return :: a -> m a
  (>>)   :: m a -> m b -> m b
  fail   :: String -> m a
```

Thus class `Monad` is based on type constructor `m`; it specifies four functions of which `>>=` and `return` are the most interesting. The first one is called *bind* and used as an infix operator: given `x` of type `m a` and a function `f` of type `a -> m b`, the evaluation of `x >>= f` will yield a result of type `m b`. This corresponds to f^*. The function `return` takes a value of type `a` and evaluates to a value of type `m a`; hence it corresponds to η_a (the name `return` has probably not been a fortunate choice). The function `>>`, usually used as an infix operator as well, is defined by default in terms of `>>=`, and function `fail` serves to handling exceptions; both functions will not concern us here.

Not every conceivable definition of the functions `return` and the bind function `>>=` is suitable for the definition of a monad. These are the laws the `Haskell` programmer has to enforce, and it becomes evident that these are just the laws for a Kleisli tripel from Definition 2.4.1:

```
return x >>= f               == f x
p >>= return                 == p
p >>= (\x -> (f x >>= g))    == (p >>= (\x ->
                                f x)) >>= g
```

(here, `x` is not free in `g`; `\x -> f x` is `Haskell`'s way of expressing the anonymous function $\lambda x.fx$). The compiler for Haskell cannot check these laws, so the programmer has to make sure that they hold.

We demonstrate the concept with a simple example. Lists are a popular data structure. They are declared as a monad in this way:

```
instance Monad [] where
        return t = [t]
        x >>= f  = concat (map f x)
```

This makes the polymorphic-type constructor `[]` for lists a monad; it specifies essentially that `f` is mapped over the list `x` (`x` has to be a list, and `f` a function defined on the list's base type yielding a list as a value); this results in a list of lists which then will be flattened through an application of function `concat`. This example should clarify things:

```
>>> q = (\w -> [0 .. w])
>>> [0 .. 2] >>= q
[0,0,1,0,1,2]
```

The effect is explained in the following way: The definition of the bind operation `>>=` requires the computation of

```
concat (map q [0 .. 2])
= concat [(q 0), (q 1), (q 2)]
== concat [[0], [0, 1], [0, 1, 2]]
== [0,0,1,0,1,2].
```

We check the laws of a monad.

- We have `return x == [x]`; hence

  ```
  return x >>= f == [x] >>= f
                 == concat (map f [x])
                 == concat [f x]
                 == f x
  ```

- Similarly, if `p` is a given list, then

  ```
  p >>= return == concat (map return p)
               == concat [[x] | x <- p]
               == [x | x <- p]
               == p
  ```

- For the third law, if `p` is the empty list, then the left- and the right-hand side are empty as well. Hence let us assume that `p = [x1, .., xn]`. We obtain for the left-hand side

  ```
  p >>= (\x -> (f x >>= g))
  == concat (map (\x -> (f x >>= g)) p)
  == concat (concat [map g (f x) | x <-p]),
  ```

  and for the right-hand side

  ```
  (concat [f x | x <- p]) >>= g
  == ((f x1) ++ (f x2) ++ .. ++ (f xn))
     >>= g
  == concat (map g ((f x1) ++ (f x2) ++ ..
  ```

```
        ++ (f xn)))
    == concat (concat [map g (f x) | x <- p])
```

(this argumentation could of course be made more precise through a proof by induction on the length of list `p`, but this would lead us too far from the present discussion).

Kleisli composition `>=>` can be defined in a monad as follows:

```
(>=>) :: Monad m => (a -> m b) -> (b -> mc)
-> (a -> m c) f >=> g = \x -> (f x) >>= g
```

This gives in the first line a type declaration for operation `>=>` by indicating that the infix operator `>=>` takes two arguments, viz., a function with the signature `a -> m b` and a second one with the signature `b -> m c`, and that the result will be a function of type `a -> m c`, as expected. The precondition to this type declaration is that `m` is a monad. The body of the function will use the bind operator for binding `f x` to `g`; this results in a function depending on `x`. It can be shown that this composition is associative.

2.5 Adjunctions and Algebras

An adjunction relates two functors $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ and $\boldsymbol{G} : \boldsymbol{L} \to \boldsymbol{K}$ in a systematic way. We define this formally and investigate some examples in order to show that this is a natural concept which arises in a variety of situations. In fact, we will show that monads are closely related to adjunctions via algebras, so we will study algebras as well and provide the corresponding constructions.

2.5.1 Adjunctions

We define the basic notion of an adjunction and show that an adjunction defines a pair of natural transformations through universal arrows (which is sometimes taken as the basis for adjunctions).

Definition 2.5.1 *Let $\boldsymbol{K}$ and $\boldsymbol{L}$ be categories. Then $(\boldsymbol{F}, \boldsymbol{G}, \varphi)$ is called an* adjunction *iff:*

1. *$\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ and $\boldsymbol{G} : \boldsymbol{L} \to \boldsymbol{K}$ are functors,*

2. *for each object a in **L** and x in **K** there is a bijection*

$$\varphi_{x,a} : \hom_{\boldsymbol{L}}(\boldsymbol{F}x, a) \to \hom_{\boldsymbol{K}}(x, \boldsymbol{G}a)$$

which is natural in x and a.

F** is called the* left adjoint to ***G**; **G** is called the* right adjoint to ***F.

That $\varphi_{x,a}$ is natural for each x, a means that for all morphisms $f : a \to b$ in $\boldsymbol{L}$ and $g : y \to x$ in $\boldsymbol{K}$, both diagrams commute:

$$\begin{tikzcd} \hom_{\boldsymbol{L}}(\boldsymbol{F}x, a) \arrow[r, "\varphi_{x,a}"] \arrow[d, "f_*"'] & \hom_{\boldsymbol{K}}(x, \boldsymbol{G}a) \arrow[d, "(\boldsymbol{G}f)_*"] \\ \hom_{\boldsymbol{L}}(\boldsymbol{F}x, b) \arrow[r, "\varphi_{x,b}"'] & \hom_{\boldsymbol{K}}(x, \boldsymbol{G}b) \end{tikzcd} \qquad \begin{tikzcd} \hom_{\boldsymbol{L}}(\boldsymbol{F}x, a) \arrow[r, "\varphi_{x,a}"] \arrow[d, "(\boldsymbol{F}g)^*"'] & \hom_{\boldsymbol{K}}(x, \boldsymbol{G}a) \arrow[d, "g^*"] \\ \hom_{\boldsymbol{L}}(\boldsymbol{F}y, a) \arrow[r, "\varphi_{y,a}"'] & \hom_{\boldsymbol{K}}(y, \boldsymbol{G}a) \end{tikzcd}$$

Here, $f_* := \hom_{\boldsymbol{L}}(\boldsymbol{F}x, f)$ and $g^* := \hom_{\boldsymbol{K}}(g, \boldsymbol{G}a)$ are the hom set functors associated with f resp. g, similar for $(\boldsymbol{G}f)_*$ and for $(\boldsymbol{F}g)^*$; for the hom set functors, see Example 2.3.3.

Let us have a look at currying as a simple example.

Example 2.5.2 A map $f : X \times Y \to Z$ is sometimes considered as a map $f : X \to (Y \to Z)$, so that $f(x, y)$ is considered as the value $F(x)(y)$ at y for the "higher order" map $F(x) := \lambda b.f(x, b)$. This technique is popular in functional programming; it is called *currying* and will be discussed now.

Currying

Fix a set E and define the endofunctors $\boldsymbol{F}, \boldsymbol{G} : \boldsymbol{Set} \to \boldsymbol{Set}$ by $\boldsymbol{F} := - \times E$ resp. $\boldsymbol{G} := -^E$. Thus we have in particular $(\boldsymbol{F}f)(x, e) := \langle f(x), e\rangle$ and $(\boldsymbol{G}f)(g)(e) := f(g(e))$, whenever $f : X \to Y$ is a map.

Define the map $\varphi_{X,A} : \hom_{\boldsymbol{Set}}(\boldsymbol{F}X, A) \to \hom_{\boldsymbol{Set}}(X, \boldsymbol{G}A)$ by $\varphi_{X,A}(k)(x)(e) := k(x, e)$. Then $\varphi_{X,A}$ is a bijection. In fact, let $k_1, k_2 : \boldsymbol{F}X \to A$ be different maps, then $k_1(x, e) \neq k_2(x, e)$ for some $\langle x, e\rangle \in X \times E$; hence $\varphi_{X,A}(k_1)(x)(e) \neq \varphi_{X,A}(k_2)(x)(e)$, so that $\varphi_{X,A}$ is one to one. Let $\ell : X \to \boldsymbol{G}A$ be a map, then $\ell = \varphi_{X,A}(k)$ with $k(x, e) := \ell(x)(e)$. Thus $\varphi_{X,A}$ is onto.

In order to show that φ is natural both in X and in A, take maps $f : A \to B$ and $g : Y \to X$ and trace $k \in \hom_{\boldsymbol{Set}}(\boldsymbol{F}X, A)$ through the diagrams in Definition 2.5.1. We have

$$\begin{aligned}\varphi_{X,B}(f_*(k))(x)(e) &= f_*(k)(x, e) = f(k(x, e)) = f(\varphi_{X,A}(k)(x, e)) \\ &= (\boldsymbol{G}f)_*(\varphi_{X,A}(k)(x)(e).\end{aligned}$$

Similarly,

$$\begin{aligned} g^*(\varphi_{X,A}(k))(y)(e) &= k(g(y), e) = (\boldsymbol{F}g)^*(k)(y, e) \\ &= \varphi_{Y,A}((\boldsymbol{F}g)^*(k))(y)(e). \end{aligned}$$

This shows that $(\boldsymbol{F}, \boldsymbol{G}, \varphi)$ with φ as the currying function is an adjunction. ✌

Another popular example is furnished through the diagonal functor.

Example 2.5.3 Let $\boldsymbol{K}$ be a category such that for any two objects a and b, their product $a \times b$ exists. Recall the definition of the Cartesian product of categories from Lemma 2.1.19. Define the diagonal functor $\Delta : \boldsymbol{K} \to \boldsymbol{K} \times \boldsymbol{K}$ through $\Delta a := \langle a, a\rangle$ for objects and $\Delta f := \langle f, f\rangle$ for morphism f. Conversely, define $\boldsymbol{T} : \boldsymbol{K} \times \boldsymbol{K} \to \boldsymbol{K}$ by putting $\boldsymbol{T}(a, b) := a \times b$ for objects and $\boldsymbol{T}\langle f, g\rangle := f \times g$ for morphism $\langle f, g\rangle$.

Let $\langle k_1, k_2\rangle \in \hom_{\boldsymbol{K}\times\boldsymbol{K}}(\Delta a, \langle b_1, b_2\rangle)$; hence we have morphisms $k_1 : a \to b_1$ and $k_2 : a \to b_2$. By the definition of the product, there exists a unique morphism $k : a \to b_1 \times b_2$ with $k_1 = \pi_1 \circ k$ and $k_2 = \pi_2 \circ k$, where $\pi_i : b_1 \times b_2 \to b_i$ are the projections, $i = 1, 2$. Define $\varphi_{a,b_1\times b_2}(k_1, k_2) := k$; then it is immediate that $\varphi_{a,b_1\times b_2} : \hom_{\boldsymbol{K}\times\boldsymbol{K}}(\Delta a, \langle b_1, b_2\rangle) \to \hom_{\boldsymbol{K}}(a, \boldsymbol{T}(b_1, b_2)$ is a bijection.

Let $\langle f_1, f_2\rangle : \langle a_1, a_2\rangle \to \langle b_1, b_2\rangle$ be a morphism; then the diagram

$$\begin{tikzcd} \hom_{\boldsymbol{K}\times\boldsymbol{K}}(\Delta x, \langle a_1, a_2\rangle) \arrow[r, "\varphi_{x,\langle a_1,a_2\rangle}"] \arrow[d, "\langle f_1,f_2\rangle_*"'] & \hom_{\boldsymbol{K}}(x, a_1 \times a_2) \arrow[d, "(\boldsymbol{T}(f_1,f_2))_*"] \\ \hom_{\boldsymbol{K}\times\boldsymbol{K}}(\Delta x, \langle b_1, b_2\rangle) \arrow[r, "\varphi_{x,\langle b_1,b_2\rangle}"'] & \hom_{\boldsymbol{K}}(x, b_1 \times b_2) \end{tikzcd}$$

splits into the two commutative diagrams

$$\begin{tikzcd} \hom_{\boldsymbol{K}}(x, a_i) \arrow[r, "\pi_i\circ\varphi_{x,\langle a_1,a_2\rangle}"] \arrow[d, "f_{i,*}"'] & \hom_{\boldsymbol{K}}(x, a_i) \arrow[d, "(\pi_i\circ(\boldsymbol{T}(f_1,f_2)))_*"] \\ \hom_{\boldsymbol{K}}(x, b_i) \arrow[r, "\pi_i\circ\varphi_{x,\langle b_1,b_2\rangle}"'] & \hom_{\boldsymbol{K}}(x, b_i) \end{tikzcd}$$

for $i = 1, 2$, hence is commutative itself. One argues similarly for a morphism $g : b \to a$. Thus the bijection φ is natural.

Hence we have found out that $(\Delta, \boldsymbol{T}, \varphi)$ is an adjunction, so that the diagonal functor has the product functor as an adjoint. ✌

A map $f : X \to Y$ between sets provides us with another example, which is the special case of a Galois connection: A pair $f : P \to Q$ and $g : Q \to P$ of monotone maps between the partially ordered sets P and Q form a *Galois connection* iff $f(p) \geq q \Leftrightarrow p \geq g(q)$ for all $p \in P, q \in Q$.

Example 2.5.4 Let X and Y be sets, then the inclusion on $\mathcal{P}X$ resp. $\mathcal{P}Y$ makes these sets categories of their own; see Example 2.1.4. Given a map $f : X \to Y$, define $f_? : \mathcal{P}X \to \mathcal{P}Y$ as the direct image $f_?(A) := f\big[A\big]$ and $f_! : \mathcal{P}Y \to \mathcal{P}X$ as the inverse image $f_!(B) := f^{-1}\big[B\big]$. Now we have for $A \subseteq X$ and $B \subseteq Y$

$$B \subseteq f_?(A) \Leftrightarrow B \subseteq f\big[A\big] \Leftrightarrow f^{-1}\big[B\big] \subseteq A \Leftrightarrow f_!(B) \subseteq A.$$

This means in terms of the hom sets that $\hom_{\mathcal{P}\mathrm{Y}}(B, f_?(A)) \neq \emptyset$ iff $\hom_{\mathcal{P}\mathrm{X}}(f_!(B), A) \neq \emptyset$. Hence this gives an adjunction $(f_!, f_?, \varphi)$. ✌

Back to the general development. This auxiliary statement will help in some computations.

Lemma 2.5.5 *Let* $(\boldsymbol{F}, \boldsymbol{G}, \varphi)$ *be an adjunction and* $f : a \to b$ *and* $g : y \to x$ *be morphisms in* $\boldsymbol{L}$ *resp.* $\boldsymbol{K}$*. Then we have*

$$(\boldsymbol{G}f) \circ \varphi_{x,a}(t) = \varphi_{x,b}(f \circ t),$$
$$\varphi_{x,a}(t) \circ g = \varphi_{y,a}(t \circ \boldsymbol{F}g)$$

for each morphism $t : \boldsymbol{F}x \to a$ *in* $\boldsymbol{L}$*.*

Proof Chase t through the left-hand diagram of Definition 2.5.1 to obtain

$$((\boldsymbol{G}f)_* \circ \varphi_{x,a})(t) = (\boldsymbol{G}f) \circ \varphi_{x,a}(t) = \varphi_{x,b}(f_*(t)) = \varphi_{x,b}(f \circ t).$$

This yields the first equation; the second is obtained from tracing t through the diagram on the right-hand side. ⊣

An adjunction induces natural transformations which make this important construction easier to handle and which helps indicating connections of adjunctions to monads and Eilenberg–Moore algebras. Before entering the discussion, universal arrows are introduced.

Definition 2.5.6 *Let* $\boldsymbol{S} : \boldsymbol{C} \to \boldsymbol{D}$ *be a functor and* c *an object in* $\boldsymbol{C}$*.*

1. *the pair* $\langle r, u\rangle$ *is called a* universal arrow from c to $\boldsymbol{S}$ *iff* r *is an object in* $\boldsymbol{C}$ *and* $u : c \to \boldsymbol{S}r$ *is a morphism in* $\boldsymbol{D}$ *such that for any*

arrow $f : c \to \boldsymbol{S}d$ *there exists a unique arrow* $f' : r \to d$ *in* $\boldsymbol{C}$ *such that* $f = (\boldsymbol{S}f') \circ u$.

2. *the pair* $\langle r, v\rangle$ *is called a* universal arrow from $\boldsymbol{S}$ to c *iff* r *is an object in* $\boldsymbol{C}$ *and* $v : \boldsymbol{S}r \to c$ *is a morphism in* $\boldsymbol{D}$ *such that for any arrow* $f : \boldsymbol{S}d \to c$ *there exists a unique arrow* $f' : d \to r$ *in* $\boldsymbol{C}$ *such that* $f = v \circ (\boldsymbol{S}f')$.

Thus, if the pair $\langle r, u\rangle$ is universal from c to $\boldsymbol{S}$, then each arrow $c \to \boldsymbol{S}d$ in $\boldsymbol{C}$ factors uniquely through the $\boldsymbol{S}$-image of an arrow $r \to d$ in $\boldsymbol{C}$. Similarly, if the pair $\langle r, v\rangle$ is universal from $\boldsymbol{S}$ to c, then each $\boldsymbol{D}$-arrow $\boldsymbol{S}d \to c$ factors uniquely through the $\boldsymbol{S}$-image of an $\boldsymbol{C}$-arrow $d \to r$. These diagrams depict the situation for a universal arrow $u : c \to \boldsymbol{S}r$ resp. a universal arrow $v : \boldsymbol{S}r \to c$.

$$
\begin{tikzcd}
c \arrow[r, "u"] \arrow[dr, "f"'] & \boldsymbol{S}r \arrow[d, "\boldsymbol{S}f'"] & r \arrow[d, dotted, "!\ f'"] \\
& \boldsymbol{S}d & d
\end{tikzcd}
\qquad
\begin{tikzcd}
\boldsymbol{S}r \arrow[r, "v"] & c & r \\
\boldsymbol{S}d \arrow[u, "\boldsymbol{S}f'"] \arrow[ur, "f"'] & & d \arrow[u, dotted, "!\ f'"']
\end{tikzcd}
$$

This is a characterization of a universal arrow from c to $\boldsymbol{S}$.

Lemma 2.5.7 *Let* $\boldsymbol{S} : \boldsymbol{C} \to \boldsymbol{D}$ *be a functor. Then* $\langle r, u\rangle$ *is a universal arrow from* c *to* $\boldsymbol{S}$ *iff the function* ψ_d *which maps each morphism* $f' : r \to d$ *to the morphism* $(\boldsymbol{S}f') \circ u$ *is a natural bijection* $\hom_{\boldsymbol{C}}(r, d) \to \hom_{\boldsymbol{D}}(c, \boldsymbol{S}d)$.

Proof 1. If $\langle r, u\rangle$ is a universal arrow, then bijectivity of ψ_d is just a reformulation of the definition. It is also clear that ψ_d is natural in d, because if $g : d \to d'$ is a morphism, then $\boldsymbol{S}(g' \circ f') \circ u = (\boldsymbol{S}g') \circ (\boldsymbol{S}g) \circ u$.

2. Now assume that $\psi_d : \hom_{\boldsymbol{C}}(r, d) \to \hom_{\boldsymbol{D}}(c, \boldsymbol{S}d)$ is a bijection for each d, and choose in particular $r = d$. Define $u := \psi_r(id_r)$; then $u : c \to \boldsymbol{S}r$ is a morphism in $\boldsymbol{D}$. Consider this diagram for an arbitrary $f' : r \to d$

$$
\begin{tikzcd}
\hom_{\boldsymbol{C}}(r, r) \arrow[r, "\psi_r"] \arrow[d, "\hom_{\boldsymbol{C}}(r, f')"'] & \hom_{\boldsymbol{D}}(c, \boldsymbol{S}r) \arrow[d, "\hom_{\boldsymbol{D}}(s, \boldsymbol{S}f')"] \\
\hom_{\boldsymbol{C}}(r, d) \arrow[r, "\psi_d"'] & \hom_{\boldsymbol{D}}(c, \boldsymbol{S}d)
\end{tikzcd}
$$

Given a morphism $f : c \to \boldsymbol{S}d$ in $\boldsymbol{D}$, there exists a unique morphism $f' : r \to d$ such that $f = \psi_d(f')$, because ψ_d is a bijection. Then we have

$$\begin{aligned} f &= \psi_d(f') \\ &= \big(\psi_d \circ \hom_{\boldsymbol{C}}(r, f')\big)(id_r) \\ &= \big(\hom_{\boldsymbol{D}}(c, \boldsymbol{S}f') \circ \psi_r\big)(id_r) && \text{(commutativity)} \\ &= \hom_{\boldsymbol{D}}(c, \boldsymbol{S}f') \circ u && (u = \psi_r(id_r)) \\ &= (\boldsymbol{S}f') \circ u. \end{aligned}$$

⊣

Universal arrows will be used now for a characterization of adjunctions in terms of natural transformations (we will sometimes omit the indices for the natural transformation φ that comes with an adjunction).

Theorem 2.5.8 *Let $(\boldsymbol{F}, \boldsymbol{G}, \varphi)$ be an adjunction for the functors $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ and $\boldsymbol{G} : \boldsymbol{L} \to \boldsymbol{K}$. Then there exist natural transformations $\eta : Id_{\boldsymbol{K}} \to \boldsymbol{G} \circ \boldsymbol{F}$ and $\varepsilon : \boldsymbol{F} \circ \boldsymbol{G} \to Id_{\boldsymbol{L}}$ with these properties:*

1. *the pair $\langle \boldsymbol{F}x, \eta_x \rangle$ is a universal arrow from x to $\boldsymbol{G}$ for each x in $\boldsymbol{K}$, and $\varphi(f) = \boldsymbol{G}f \circ \eta_x$ holds for each $f : \boldsymbol{F}x \to a$,*

2. *the pair $\langle \boldsymbol{G}a, \varepsilon_a \rangle$ is universal from $\boldsymbol{F}$ to a for each a in $\boldsymbol{L}$, and $\varphi^{-1}(g) = \varepsilon_a \circ \boldsymbol{F}g$ holds for each $g : x \to \boldsymbol{G}a$,*

3. *the composites*

$$\begin{tikzcd} \boldsymbol{G} \arrow[r, "\eta\boldsymbol{G}"] \arrow[dr, "id_{\boldsymbol{G}}"'] & \boldsymbol{G} \circ \boldsymbol{F} \circ \boldsymbol{G} \arrow[d, "\boldsymbol{G}\varepsilon"] \\ & \boldsymbol{G} \end{tikzcd} \qquad \begin{tikzcd} \boldsymbol{F} \arrow[r, "\boldsymbol{F}\eta"] \arrow[dr, "id_{\boldsymbol{F}}"'] & \boldsymbol{F} \circ \boldsymbol{G} \circ \boldsymbol{F} \arrow[d, "\varepsilon\boldsymbol{F}"] \\ & \boldsymbol{F} \end{tikzcd}$$

are the identities for $\boldsymbol{G}$ resp. $\boldsymbol{F}$.

Proof 1. Put $\eta_x := \varphi_{x,\boldsymbol{F}x}(id_{\boldsymbol{F}x})$; then $\eta_x : x \to \boldsymbol{GF}x$. In order to show that $\langle \boldsymbol{F}x, \eta_x \rangle$ is a universal arrow from x to $\boldsymbol{G}$, we take a morphism $f : x \to \boldsymbol{G}a$ for some object a in $\boldsymbol{L}$. Since $(\boldsymbol{F}, \boldsymbol{G}, \varphi)$ is an adjunction, we know that there exists a unique morphism $f' : \boldsymbol{F}x \to a$ such that $\varphi_{x,a}(f') = f$. We have also this commutative diagram

$$\begin{tikzcd} \hom_{\boldsymbol{K}}(\boldsymbol{F}x, \boldsymbol{F}x) \arrow[r, "\varphi_{x,\boldsymbol{F}x}"] \arrow[d, "\hom(\boldsymbol{F}x, f')"'] & \hom_{\boldsymbol{L}}(x, \boldsymbol{GF}x) \arrow[d, "\hom_{\boldsymbol{L}}(x, \boldsymbol{G}f')"] \\ \hom_{\boldsymbol{K}}(\boldsymbol{F}x, a) \arrow[r, "\varphi_{x,a}"'] & \hom_{\boldsymbol{L}}(x, \boldsymbol{G}a) \end{tikzcd}$$

Thus

$$\begin{aligned}(\boldsymbol{G}f')\circ\eta_x &= \big(\hom_{\boldsymbol{L}}(x,\boldsymbol{G}f')\circ\varphi_{x,\boldsymbol{F}x}\big)(id_{\boldsymbol{F}x})\\ &= \big(\varphi_{x,a}\circ\hom_{\boldsymbol{K}}(\boldsymbol{F}x,f')\big)(id_{\boldsymbol{F}x})\\ &= \varphi_{x,a}(f')\\ &= f\end{aligned}$$

2. $\eta : Id_{\boldsymbol{K}} \to \boldsymbol{G}\circ\boldsymbol{F}$ is a natural transformation. Let $h : x \to y$ be a morphism in $\boldsymbol{K}$, then we have by Lemma 2.5.5

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{F}h)\circ\eta_x &= \boldsymbol{G}(\boldsymbol{F}h)\circ\varphi_{x,\boldsymbol{F}x}(id_{\boldsymbol{F}x}) &&= \varphi_{x,\boldsymbol{F}y}(\boldsymbol{F}h\circ id_{\boldsymbol{F}x})\\ &= \varphi_{x,\boldsymbol{F}y}(id_{\boldsymbol{F}y}\circ\boldsymbol{F}h) &&= \varphi_{y,\boldsymbol{F}y}(id_{\boldsymbol{F}y})\circ h\\ &= \eta_y\circ h.\end{aligned}$$

3. Put $\varepsilon_a := \varphi^{-1}_{\boldsymbol{G}a,a}(id_{\boldsymbol{G}a})$ for the object a in $\boldsymbol{L}$; then the properties for ε are proved in exactly the same way as for those of η.

4. From $\varphi_{x,a}(f) = \boldsymbol{G}f\circ\eta_x$, we obtain

$$id_{\boldsymbol{G}a} = \varphi(\varepsilon_a) = \boldsymbol{G}\varepsilon_a\circ\eta_{\boldsymbol{G}a} = (\boldsymbol{G}\varepsilon\circ\eta\boldsymbol{G})(a),$$

so that $\boldsymbol{G}\varepsilon\circ\eta\boldsymbol{G}$ is the identity transformation on $\boldsymbol{G}$. Similarly, $\eta\boldsymbol{F}\circ\boldsymbol{F}\varepsilon$ is the identity for $\boldsymbol{F}$. ⊣

The transformation η is sometimes called the *unit* of the adjunction, whereas ε is called its *counit*. The converse to Theorem 2.5.8 holds as well: from two transformations η and ε with the signatures as above, one can construct an adjunction. The proof is a fairly straightforward verification. Unit, counit

Proposition 2.5.9 *Let $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ and $\boldsymbol{G} : \boldsymbol{L} \to \boldsymbol{K}$ be functors, and assume that natural transformations $\eta : Id_{\boldsymbol{K}} \to \boldsymbol{G}\circ\boldsymbol{F}$ and $\varepsilon : \boldsymbol{F}\circ\boldsymbol{G} \to Id_{\boldsymbol{L}}$ are given so that $(\boldsymbol{G}\varepsilon)\circ(\eta\boldsymbol{G})$ is the identity of $\boldsymbol{G}$ and $(\varepsilon\boldsymbol{F})\circ(\boldsymbol{F}\eta)$ is the identity of $\boldsymbol{F}$. Define $\varphi_{x,a}(k) := (\boldsymbol{G}k)\circ\eta_x$, whenever $k : \boldsymbol{F}x \to a$ is a morphism in $\mathbf{L}$. Then $(\boldsymbol{F},\boldsymbol{G},\varphi)$ defines an adjunction.*

Proof 1. Define $\vartheta_{x,a}(\ell) := \varepsilon_a \circ \boldsymbol{F}g$ for $\ell : x \to \boldsymbol{G}a$; then we have

$$\begin{aligned}
\varphi_{x,a}(\vartheta_{x,a}(g)) &= \boldsymbol{G}(\varepsilon_a \circ \boldsymbol{F}g) \circ \eta_x \\
&= (\boldsymbol{G}\varepsilon_a) \circ (\boldsymbol{GF}g) \circ \eta_x \\
&= (\boldsymbol{G}\varepsilon_a) \circ \eta_{\boldsymbol{G}a} \circ g && (\eta \text{ is natural}) \\
&= \big((\boldsymbol{G}\varepsilon \circ \eta\boldsymbol{G})a\big) \circ g \\
&= id_{\boldsymbol{G}a} g \\
&= g
\end{aligned}$$

Thus $\varphi_{x,a} \circ \vartheta_{x,a} = id_{\hom_{\boldsymbol{L}}(x,\boldsymbol{G}a)}$. Similarly, one shows that $\vartheta_{x,a} \circ \varphi_{x,a} = id_{\hom_{\boldsymbol{K}}(\boldsymbol{F}x,a)}$, so that $\varphi_{x,a}$ is a bijection.

2. We have to show that $\varphi_{x,a}$ is natural for each x, a, so take a morphism $f : a \to b$ in $\boldsymbol{L}$ and chase $k : \boldsymbol{F}x \to a$ through this diagram.

$$\begin{tikzcd}
\hom_{\boldsymbol{L}}(\boldsymbol{F}x, a) \arrow[r, "\varphi_{x,a}"] \arrow[d, "f_*"'] & \hom_{\boldsymbol{K}}(x, \boldsymbol{G}a) \arrow[d, "(\boldsymbol{G}f)_*"] \\
\hom_{\boldsymbol{L}}(\boldsymbol{F}x, b) \arrow[r, "\varphi_{x,b}"'] & \hom_{\boldsymbol{K}}(x, \boldsymbol{G}b)
\end{tikzcd}$$

Then $\big((\boldsymbol{G}f)_* \circ \varphi_{x,a}\big)(k) = (\boldsymbol{G}f \circ \boldsymbol{G}k) \circ \eta_x = \boldsymbol{G}(f \circ k) \circ \eta_x = \varphi_{x,b}(f_* \circ k)$. $\dashv$

Thus it is sufficient to identify its unit and its counit for identifying an adjunction. This includes verifying the identity laws of the functors for the corresponding compositions. The following example has another look at currying (Example 2.5.2), demonstrating the approach and suggesting that identifying unit and counit is sometimes easier than working with the originally given definition.

Example 2.5.10 Continuing Example 2.5.2, we take the definitions of the endofunctors $\boldsymbol{F}$ and $\boldsymbol{G}$ from there. Define for the set X the natural transformations $\eta : Id_{\boldsymbol{Set}} \to \boldsymbol{G} \circ \boldsymbol{F}$ and $\varepsilon : \boldsymbol{F} \circ \boldsymbol{G} \to Id_{\boldsymbol{Set}}$ through

$$\eta_X : \begin{cases} X & \to (X \times E)^E \\ x & \mapsto \lambda e.\langle x, e\rangle \end{cases}$$

and

$$\varepsilon_X : \begin{cases} (X \times E)^E \times E & \to X \\ \langle g, e\rangle & \mapsto g(e) \end{cases}$$

Note that we have $(\boldsymbol{G}f)(h) = f \circ h$ for $f : X^E \to Y^E$ and $h \in X^E$, so that we obtain

$$\begin{aligned}
\boldsymbol{G}\varepsilon_X(\eta_{\boldsymbol{G}X}(g))(e) &= (\varepsilon_X \circ \eta_{\boldsymbol{G}X}(g))(e) &&= \varepsilon_X(\eta_{\boldsymbol{G}X}(g))(e)\\
&= \varepsilon_X(\eta_{\boldsymbol{G}X}(g)(e)) &&= \varepsilon_X(g, e)\\
&= g(e),
\end{aligned}$$

whenever $e \in E$ and $g \in \boldsymbol{G}X = X^E$; hence $(\boldsymbol{G}\varepsilon) \circ (\eta\boldsymbol{G}) = id_{\boldsymbol{G}}$. One shows similarly that $(\varepsilon\boldsymbol{F}) \circ (\boldsymbol{F}\eta) = id_{\boldsymbol{F}}$ through

$$\varepsilon_{\boldsymbol{F}X}(\boldsymbol{F}\eta_X(x, e)) = \eta_X(x)(e) = \langle x, e\rangle.$$

✌

Now let $(\boldsymbol{F}, \boldsymbol{G}, \varphi)$ be an adjunction with functors $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ and $\boldsymbol{G} : \boldsymbol{L} \to \boldsymbol{K}$, the unit η, and the counit ε. Define the functor $\boldsymbol{T}$ through $\boldsymbol{T} := \boldsymbol{G} \circ \boldsymbol{F}$. Then $\boldsymbol{T} : \boldsymbol{K} \to \boldsymbol{K}$ defines an endofunctor on category $\boldsymbol{K}$ with $\mu_a := (\boldsymbol{G}\varepsilon\boldsymbol{F})(a) = \boldsymbol{G}\varepsilon_{\boldsymbol{F}a}$ as a morphism $\mu_a : \boldsymbol{T}^2(a) \to \boldsymbol{T}a$. Because $\varepsilon_a : \boldsymbol{F}\boldsymbol{G}a \to a$ is a morphism in $\boldsymbol{L}$, and because $\varepsilon : \boldsymbol{F} \circ \boldsymbol{G} \to Id_{\boldsymbol{L}}$ is natural, the diagram

$$\begin{tikzcd}
(\boldsymbol{F} \circ \boldsymbol{G} \circ \boldsymbol{F} \circ \boldsymbol{G})a \arrow[r, "\varepsilon_{(\boldsymbol{F}\circ\boldsymbol{G})a}"] \arrow[d, "(\boldsymbol{F}\circ\boldsymbol{G})\varepsilon_a"'] & (\boldsymbol{F} \circ \boldsymbol{G})a \arrow[d, "\varepsilon_a"] \\
(\boldsymbol{F} \circ \boldsymbol{G})a \arrow[r, "\varepsilon_a"'] & a
\end{tikzcd}$$

is commutative. This means that this diagram

$$\begin{tikzcd}
\boldsymbol{F} \circ \boldsymbol{G} \circ \boldsymbol{F} \circ \boldsymbol{G} \arrow[r, "\varepsilon(\boldsymbol{F}\circ\boldsymbol{G})"] \arrow[d, "(\boldsymbol{F}\circ\boldsymbol{G})\varepsilon"'] & \boldsymbol{F} \circ \boldsymbol{G} \arrow[d, "\varepsilon"] \\
\boldsymbol{F} \circ \boldsymbol{G} \arrow[r, "\varepsilon"'] & Id_{\boldsymbol{K}}
\end{tikzcd}$$

of functors and natural transformations commutes. Multiplying from the left with $\boldsymbol{G}$ and from the right with $\boldsymbol{F}$ gives this diagram.

$$\begin{tikzcd}
\boldsymbol{G} \circ \boldsymbol{F} \circ \boldsymbol{G} \circ \boldsymbol{F} \circ \boldsymbol{G} \circ \boldsymbol{F} \arrow[r, "\boldsymbol{G}\varepsilon(\boldsymbol{F}\circ\boldsymbol{G}\circ\boldsymbol{F})"] \arrow[d, "(\boldsymbol{G}\circ\boldsymbol{F}\circ\boldsymbol{G})\varepsilon\boldsymbol{F}"'] & \boldsymbol{G} \circ \boldsymbol{F} \circ \boldsymbol{G} \circ \boldsymbol{F} \arrow[d, "\boldsymbol{G}\varepsilon\boldsymbol{F}"] \\
\boldsymbol{G} \circ \boldsymbol{F} \circ \boldsymbol{G} \circ \boldsymbol{F} \arrow[r, "\boldsymbol{G}\varepsilon\boldsymbol{F}"'] & \boldsymbol{G} \circ \boldsymbol{F}
\end{tikzcd}$$

Because $\boldsymbol{T}\mu = (\boldsymbol{G} \circ \boldsymbol{F} \circ \boldsymbol{G})\varepsilon\boldsymbol{F}$, and $\boldsymbol{G}\varepsilon(\boldsymbol{F} \circ \boldsymbol{G} \circ \boldsymbol{F}) = \mu\boldsymbol{T}$, this diagram can be written as

$$\begin{tikzcd}
\boldsymbol{T}^3 \arrow[r, "\boldsymbol{T}\mu"] \arrow[d, "\mu\boldsymbol{T}"'] & \boldsymbol{T}^2 \arrow[d, "\mu"] \\
\boldsymbol{T}^2 \arrow[r, "\mu"'] & \boldsymbol{T}
\end{tikzcd}$$

Lo and behold!

This gives the commutativity of the left-hand diagram in Definition 2.4.3 for a monad. Because $\boldsymbol{G}\varepsilon \circ \eta\boldsymbol{G}$ is the identity on $\boldsymbol{G}$, we obtain

$$\boldsymbol{G}\varepsilon_{\boldsymbol{F}a} \circ \eta_{\boldsymbol{GF}a} = (\boldsymbol{G}\varepsilon \circ \eta\boldsymbol{G})(\boldsymbol{F}a) = \boldsymbol{GF}a,$$

which implies that the diagram

$$\begin{tikzcd} \boldsymbol{T}a \arrow[r, "\mu_{\boldsymbol{T}a}"] \arrow[dr, "id_{\boldsymbol{T}a}"'] & \boldsymbol{T}^2a \arrow[d, "\mu_a"] \\ & \boldsymbol{T}a \end{tikzcd}$$

commutes. On the other hand, we know that $\varepsilon\boldsymbol{F} \circ \boldsymbol{F}\eta$ is the identity on $\boldsymbol{F}$; this yields

$$\boldsymbol{G}\varepsilon_{\boldsymbol{F}a} \circ \boldsymbol{GF}\eta_a = \boldsymbol{G}(\varepsilon\boldsymbol{F} \circ \boldsymbol{F}\eta)a = \boldsymbol{GF}a.$$

Hence we may complement the last diagram:

$$\begin{tikzcd} \boldsymbol{T}a \arrow[r, "\mu_{\boldsymbol{T}a}"] \arrow[dr, "id_{\boldsymbol{T}a}"'] & \boldsymbol{T}^2a \arrow[d, "\mu_a"] & \boldsymbol{T}a \arrow[l, "\boldsymbol{T}\eta_a"'] \arrow[dl, "id_{\boldsymbol{T}a}"] \\ & \boldsymbol{T}a & \end{tikzcd}$$

This gives the right-hand-side diagram in Definition 2.4.3 for a monad. We have shown

Proposition 2.5.11 *Each adjunction defines a monad.* ⊣

It turns out that we not only may proceed from an adjunction to a monad, but that it is also possible to traverse this path in the other direction.

2.5.2 Eilenberg–Moore Algebras

We will show that a monad defines an adjunction. In order to do that, we have to represent the functorial part of a monad as the composition of two other functors, so we need a second category for this. The algebras which are defined for a monad provide us with this category. So we will define algebras (and in a later chapter, their counterparts, coalgebras), and we will study them. This will help us in showing that each monad defines an adjunction. Finally, we will have a look at two examples for algebras, in order to illuminate this concept.

Given a monad $(\boldsymbol{T}, \eta, \mu)$ in a category $\boldsymbol{K}$, a pair $\langle x, h\rangle$ consisting of an object x and a morphism $h : \boldsymbol{T}x \to x$ in $\boldsymbol{K}$ is called an *Eilenberg–Moore algebra* for the monad iff the following diagrams commute

$$\begin{tikzcd} \boldsymbol{T}^2x \arrow[r, "\boldsymbol{T}h"] \arrow[d, "\mu_x"'] & \boldsymbol{T}x \arrow[d, "h"] \\ \boldsymbol{T}x \arrow[r, "h"'] & x \end{tikzcd} \qquad \begin{tikzcd} x \arrow[r, "\eta_x"] \arrow[dr, "id_x"'] & \boldsymbol{T}x \arrow[d, "h"] \\ & x \end{tikzcd}$$

Eilenberg-Moore algebra

The morphism h is called the *structure morphism* of the algebra, x its carrier.

An *algebra morphism* $f : \langle x, h\rangle \to \langle x', h'\rangle$ between the algebras $\langle x, h\rangle$ and $\langle x', h'\rangle$ is a morphism $f : x \to x'$ in $\boldsymbol{K}$ which renders the diagram

$$\begin{tikzcd} \boldsymbol{T}x \arrow[r, "h"] \arrow[d, "\boldsymbol{T}f"'] & x \arrow[d, "f"] \\ \boldsymbol{T}x' \arrow[r, "h'"'] & x' \end{tikzcd}$$

commutative. Eilenberg–Moore algebras together with their morphisms form a category $\boldsymbol{Alg}_{(\boldsymbol{T},\eta,\mu)}$. We will usually omit the reference to the monad. Fix for the moment $(\boldsymbol{T}, \eta, \mu)$ as a monad in category $\boldsymbol{K}$, and let $\boldsymbol{Alg} := \boldsymbol{Alg}_{(\boldsymbol{T},\eta,\mu)}$ be the associated category of Eilenberg–Moore algebras.

$\boldsymbol{Alg}_{(\boldsymbol{T},\eta,\mu)}$

We give some simple examples.

Lemma 2.5.12 *The pair $\langle \boldsymbol{T}x, \mu_x\rangle$ is a $\boldsymbol{T}$-algebra for each x in $\boldsymbol{K}$.*

Proof This is immediate from the laws for η and μ in a monad. $\dashv$

These algebras are usually called the *free algebras* for the monad. Morphisms in the base category $\boldsymbol{K}$ translate into morphisms in $\boldsymbol{Alg}$ through functor $\boldsymbol{T}$.

Free algebra

Lemma 2.5.13 *If $f : x \to y$ is a morphism in $\boldsymbol{K}$, then $\boldsymbol{T}f : \langle \boldsymbol{T}x, \mu_x\rangle \to \langle \boldsymbol{T}y, \mu_y\rangle$ is a morphism in $\boldsymbol{Alg}$. If $\langle x, h\rangle$ is an algebra, then $h : \langle \boldsymbol{T}x, \mu_x\rangle \to \langle x, h\rangle$ is a morphism in $\boldsymbol{Alg}$.*

Proof Because $\mu : \boldsymbol{T}^2 \to \boldsymbol{T}$ is a natural transformation, we see $\mu_y \circ \boldsymbol{T}^2 f = (\boldsymbol{T}f) \circ \mu_x$. This is just the defining equation for a morphism in $\boldsymbol{Alg}$. The second assertion follows also from the defining equation of an algebra morphism. $\dashv$

We will identify the algebras for the power set monad now, which are closely connected to semi-lattices. Recall that an ordered set $(X, \leq)$ is a sup *semi-lattice* iff each subset has its supremum in X.

Manes monad

Example 2.5.14 The algebras for the monad $(\mathcal{P}, \eta, \mu)$ in the category ***Set*** of sets with maps (sometimes called the *Manes monad*) may be identified with the complete sup semi-lattices. We will show this now.

Assume first that $\leq$ is a partial order on a set X that is sup-complete, so that $\sup A$ exists for each $A \subseteq X$. Define $h(A) := \sup A$; then we have for each $\mathcal{A} \in \mathcal{P}(\mathcal{P}(X))$ from the familiar properties of the supremum

$$\sup(\bigcup \mathcal{A}) = \sup \{\sup a \mid a \in A\}.$$

This translates into $\big(h \circ \mu_X\big)(\mathcal{A}) = \big(h \circ (\mathcal{P}h)\big)(\mathcal{A})$. Because $x = \sup\{x\}$ holds for each $x \in X$, we see that $\langle X, h\rangle$ defines an algebra.

Assume on the other hand that $\langle X, h\rangle$ is an algebra, and put

$$x \leq x' \Leftrightarrow h(\{x, x'\}) = x'$$

for $x, x' \in X$. This defines a partial order: reflexivity and antisymmetry are obvious. Transitivity is seen as follows: assume $x \leq x'$ and $x' \leq x''$; then

$$\begin{array}{lll} h(\{x, x''\}) & = h(\{h(\{x\}), h(\{x', x''\})) & = \big(h \circ (\mathcal{P}h)\big)(\{\{x\}, \{x', x''\}\}) \\ & = \big(h \circ \mu_X\big)(\{\{x\}, \{x', x''\}\}) & = h(\{x, x', x''\}) \\ & = \big(h \circ \mu_X\big)(\{\{x, x'\}, \{x', x''\}\}) & = \big(h \circ (\mathcal{P}h)\big)(\{\{x, x'\}, \{x', x''\}\}) \\ & = h(\{x', x''\}) & = x''. \end{array}$$

It is clear from $\{x\} \cup \emptyset = \{x\}$ for every $x \in X$ that $h(\emptyset)$ is the smallest element. Finally, it has to be shown that $h(A)$ is the smallest upper bound for $A \subseteq X$ in the order $\leq$. We may assume that $A \neq \emptyset$. Suppose that $x \leq t$ holds for all $x \in A$, then

$$\begin{array}{lll} h(A \cup \{t\}) & = h\big(\bigcup_{x \in A}\{x, t\}\big) & = \big(h \circ \mu_X\big)(\{\{x, t\} \mid x \in A\}) \\ & = \big(h \circ (\mathcal{P}h)\big)(\{\{x, t\} \mid x \in A\}) & = h\big(\{h(\{x, t\})\ x \in A\}\big) \\ & = h(\{t\}) & = t. \end{array}$$

Thus, if $x \leq t$ for all $x \in A$, hence $h(A) \leq t$, thus $h(A)$ is an upper bound to A, and similarly, $h(A)$ is the smallest upper bound. ✌

We have shown that each adjunction defines a monad, and—as announced above—now turn to the converse. In fact, we will show that each monad defines an adjunction, the monad of which is the given monad.

Fix the monad $(\boldsymbol{T}, \eta, \mu)$ over category $\boldsymbol{K}$, and define as above $\boldsymbol{Alg} := \boldsymbol{Alg}_{(\boldsymbol{T},\eta,\mu)}$ as the category of Eilenberg–Moore algebras. We intend to define an adjunction, so by Proposition 2.5.9, it will be the most convenient approach to solve the problem by defining unit and counit, after the corresponding functors have been identified.

Lemma 2.5.15 *Define $\boldsymbol{F}a := \langle \boldsymbol{T}a, \mu_a \rangle$ for the object $a \in |\boldsymbol{K}|$, and if $f : a \to b$ is a morphism if $\boldsymbol{K}$, define $\boldsymbol{F}f := \boldsymbol{T}f$. Then $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{Alg}$ is a functor.*

Proof We have to show that $\boldsymbol{F}f : \langle \boldsymbol{T}a, \mu_a \rangle \to \langle \boldsymbol{T}b, \mu_b \rangle$ is an algebra morphism. Since $\mu : \boldsymbol{T}^2 \to \boldsymbol{T}$ is natural, we obtain this commutative diagram:

$$\begin{tikzcd}
\boldsymbol{T}a \arrow[r, "\mu_a"] \arrow[d, "\boldsymbol{T}f"'] & \boldsymbol{T}^2 a \arrow[d, "\boldsymbol{T}^2 f"] \\
\boldsymbol{T}b \arrow[r, "\mu_b"'] & \boldsymbol{T}^2 b
\end{tikzcd}$$

But this is just the defining condition for an algebra morphism. $\dashv$

This statement is trivial:

Lemma 2.5.16 *Given an Eilenberg–Moore algebra $\langle x, h \rangle \in |\boldsymbol{Alg}|$, define $\boldsymbol{G}(x, h) := x$; if $f : \langle x, h \rangle \to \langle x', h' \rangle$ is a morphism in $\boldsymbol{Alg}$, put $\boldsymbol{G}f := f$. Then $\boldsymbol{G} : \boldsymbol{Alg} \to \boldsymbol{K}$ is a functor. Moreover, we have $\boldsymbol{G} \circ \boldsymbol{F} = \boldsymbol{T}$.* $\dashv$

We require two natural transformations, which are defined now and which are intended to serve as the unit and as the counit, respectively, for the adjunction. We define for the unit η the originally given η, so that $\eta : Id_{\boldsymbol{K}} \to \boldsymbol{G} \circ \boldsymbol{F}$ is a natural transformation. The counit ε is defined through $\varepsilon_{\langle x,h \rangle} := h$, so that $\varepsilon_{\langle x,h \rangle} : (\boldsymbol{F} \circ \boldsymbol{G})(x, h) \to Id_{\boldsymbol{Alg}}(x, h)$. This defines a natural transformation $\varepsilon : \boldsymbol{F} \circ \boldsymbol{G} \to Id_{\boldsymbol{Alg}}$. In fact, let $f : \langle x, h \rangle \to \langle x', h' \rangle$ be a morphism in $\boldsymbol{Alg}$; then—by expanding definitions—the diagram on the left-hand side translates to the one on the right-hand side, which commutes:

$$\begin{tikzcd}
(\boldsymbol{F} \circ \boldsymbol{G})(x, h) \arrow[r, "\varepsilon_{\langle x,h \rangle}"] \arrow[d, "(\boldsymbol{F}\circ\boldsymbol{G})f"'] & \langle x, h \rangle \arrow[d, "f"] \\
(\boldsymbol{F} \circ \boldsymbol{G})(x', h') \arrow[r, "\varepsilon_{\langle x',h' \rangle}"'] & \langle x', h' \rangle
\end{tikzcd}
\qquad
\begin{tikzcd}
\langle \boldsymbol{T}x, \mu_x \rangle \arrow[r, "h"] \arrow[d, "\boldsymbol{T}f"'] & \langle x, h \rangle \arrow[d, "f"] \\
\langle \boldsymbol{T}x', \mu_{x'} \rangle \arrow[r, "h'"'] & \langle x', h' \rangle
\end{tikzcd}$$

Now take an object $a \in |\boldsymbol{K}|$; then

$$(\varepsilon \boldsymbol{F} \circ \boldsymbol{F}\eta)(a) = \varepsilon_{\boldsymbol{F}a}(\boldsymbol{F}\eta_a) = \varepsilon_{\langle \boldsymbol{T}a, \mu_a\rangle}(\boldsymbol{T}\eta_a) = \mu_a(\boldsymbol{T}\eta_a) = id_{\boldsymbol{F}a}.$$

On the other hand, we have for the algebra $\langle x, h\rangle$

$$\begin{aligned}(\boldsymbol{G}\varepsilon \circ \eta \boldsymbol{G})(x, h) &= \boldsymbol{G}\varepsilon_{\langle x,h\rangle}(\eta_{\boldsymbol{G}\langle x,h\rangle}) = \boldsymbol{G}\varepsilon_{\langle x,h\rangle}(\eta_x) = \varepsilon_{\langle x,h\rangle}(\eta_x) \\ &= h\eta_x \overset{(*)}{=} id_x = id_{\boldsymbol{G}\langle x,h\rangle}\end{aligned}$$

where $(*)$ uses that $h : \boldsymbol{T}x \to x$ is the structure morphism of an algebra. Taken together, we see that η and ε satisfy the requirements of unit and counit for an adjunction according to Proposition 2.5.9.

Hence we have nearly established:

Proposition 2.5.17 *Every monad defines an adjunction. The monad defined by the adjunction is the original one.*

Proof We have only to prove the last assertion. But this is trivial, because $(\boldsymbol{G}\varepsilon\boldsymbol{F})a = (\boldsymbol{G}\varepsilon)\langle \boldsymbol{T}a, \mu_a\rangle = \boldsymbol{G}\mu_a = \mu_a$. $\dashv$

Algebras for Discrete Probabilities We identify now the Eilenberg–Moore algebras for the functor $\boldsymbol{D}$, which assigns to each set its discrete subprobabilities with finite support; see Example 2.3.11. Some preliminary and motivating observations are made first.

Put

$$\Omega := \{\langle \alpha_1, \dots, \alpha_k\rangle \mid k \in \mathbb{N}, \alpha_i \geq 0, \sum_{i=1}^k \alpha_i \leq 1\}$$

as the set of all positive convex coefficients, and call a subset V of a real vector space *positive convex* iff $\sum_{i=1}^k \alpha_i \cdot x_i \in V$. for $x_1, \dots, x_k \in V$, $\langle \alpha_1, \dots, \alpha_k\rangle \in \Omega$. Positive convexity is to be related to subprobabilities: if $\sum_{i=1}^k \alpha_i \cdot x_i$ is perceived as an observation in which item x_i is assigned probability α_i, then clearly $\sum_{i=1}^k \alpha_i \leq 1$ under the assumption that the observation is incomplete, i.e., that not every possible case has been observed.

Suppose a set X over which we formulate subprobabilities is embedded as a positive convex set into a linear space V over $\mathbb{R}$. In this case we could read off a positive convex combination for an element the probabilities with which the respective components occur.

These observations meet the intuition about positive convexity, but it has the drawback that we have to look for a linear space V into which X to embed. It has the additional shortcoming that once we did identify V, the positive convex structure on X is fixed through the vector space, but we will see soon that we need flexibility. Consequently, we propose an abstract description of positive convexity, much in the spirit of Pumplün's approach [Pum03]. Thus the essential properties (for us, that is) of positive convexity are described *intrinsically* for X without having to resort to a vector space, leading to the definition of a positive convex structure.

Definition 2.5.18 *A* positive convex structure $\mathfrak{p}$ *on a set X has for each $\alpha = \langle\alpha_1, \dots, \alpha_n\rangle \in \Omega$ a map $\alpha_{\mathfrak{p}} : X^n \to X$ which we write as* $\sum^{\mathfrak{p}}$

$$\alpha_{\mathfrak{p}}(x_1, \dots, x_n) = \textstyle\sum_{1\le i\le n}^{\mathfrak{p}} \alpha_i \cdot x_i,$$

such that

☆ $\sum_{1\le i\le n}^{\mathfrak{p}} \delta_{i,k} \cdot x_i = x_k$*, where $\delta_{i,j}$ is Kronecker's δ (thus $\delta_{i,j} = 1$ if $i = j$, and $\delta_{i,j} = 0$, otherwise),*

✪ *the identity*

$$\textstyle\sum_{1\le i\le n}^{\mathfrak{p}} \alpha_i \cdot \left(\sum_{1\le k\le m}^{\mathfrak{p}} \beta_{i,k} \cdot x_k\right) = \sum_{1\le k\le m}^{\mathfrak{p}} \left(\sum_{1\le i\le n}^{\mathfrak{p}} \alpha_i \beta_{i,k}\right) \cdot x_k$$

holds whenever $\langle\alpha_1, \dots, \alpha_n\rangle, \langle\beta_{i,1}, \dots, \beta_{i,m}\rangle \in \Omega$, $1 \le i \le n$.

Property ☆ looks quite trivial, when written down this way. Rephrasing, it states that the map

$$\langle\delta_{1,k}, \dots, \delta_{n,k}\rangle_{\mathfrak{p}} : T^n \to T,$$

which is assigned to the n-tuple $\langle\delta_{1,k}, \dots, \delta_{n,k}\rangle$ through $\mathfrak{p}$ acts as the projection to the k^{th} component for $1 \le k \le n$. Similarly, property ✪ may be recoded in a formal but less concise way. Thus we will use freely the notation from vector spaces, omitting in particular the explicit reference to the structure whenever possible. Hence simple addition $\alpha_1 \cdot x_1 + \alpha_2 \cdot x_2$ will be written rather than $\sum_{1\le i\le 2}^{\mathfrak{p}} \alpha_i \cdot x_i$, with the understanding that it refers to a given positive convex structure $\mathfrak{p}$ on X.

It is an easy exercise to establish that for a positive convex structure, the usual rules for manipulating sums in vector spaces apply, e.g., $1 \cdot$

$x = x$, $\sum_{i=1}^{n} \alpha_i \cdot x_i = \sum_{i=1, \alpha_i \neq 0}^{n} \alpha_i \cdot x_i$ or the law of associativity, $(\alpha_1 \cdot x_1 + \alpha_2 \cdot x_2) + \alpha_3 \cdot x_3 = \alpha_1 \cdot x_1 + (\alpha_2 \cdot x_2 + \alpha_3 \cdot x_3)$. Nevertheless, care should be observed, for of course not all rules apply: we cannot in general conclude $x = x'$ from $\alpha \cdot x = \alpha \cdot x'$, even if $\alpha \neq 0$.

A morphism $\vartheta : \langle X_1, \mathfrak{p}_1 \rangle \to \langle X_2, \mathfrak{p}_2 \rangle$ between positive convex structures is a map $\vartheta : X_1 \to X_2$ such that

$$\vartheta\left(\sum_{1 \leq i \leq n}^{\mathfrak{p}_1} \alpha_i \cdot x_i\right) = \sum_{1 \leq i \leq n}^{\mathfrak{p}_2} \alpha_i \cdot \vartheta(x_i)$$

holds for $x_1, \dots, x_n \in X$ and $\langle \alpha_1, \dots, \alpha_n \rangle \in \Omega$. In analogy to linear algebra, ϑ will be called an *affine* map. Positive convex structures with their morphisms form a category ***StrConv***.

We need some technical preparations, which are collected in the following:

Lemma 2.5.19 *Let X and Y be sets.*

1. *Given a map $f : X \to Y$, let $p = \alpha_1 \cdot \delta_{a_1} + \ldots + \alpha_n \cdot \delta_{a_n}$ be the linear combination of Dirac measures for $x_1, \dots, x_n \in X$ with positive convex $\langle \alpha_1, \dots, \alpha_n \rangle \in \Omega$. Then $\boldsymbol{D}(f)(p) = \alpha_1 \cdot \delta_{f(x_1)} + \ldots + \alpha_n \cdot \delta_{f(x_n)}$.*
2. *Let $p_1, \dots, p_n$ be discrete subprobabilities X, and let $M = \alpha_1 \cdot \delta_{p_1} + \ldots + \alpha_n \cdot \delta_{p_n}$ be the linear combination of the corresponding Dirac measures in $(\boldsymbol{D} \circ \boldsymbol{D})X$ with positive convex coefficients $\langle \alpha_1, \dots, \alpha_n \rangle \in \Omega$. Then $\mu_X(M) = \alpha_1 \cdot p_1 + \ldots + \alpha_n \cdot p_n$.*

Proof The first part follows directly from the observation $\boldsymbol{D}(f)(\delta_x)(B) = \delta_x(f^{-1}\left[B\right]) = \delta_{f(x)}(B)$, and the second one is easily inferred from the formula for μ in Example 2.4.7. $\dashv$

The algebras are described now without having to resort to $\boldsymbol{D}X$ through an intrinsic characterization using positive convex structures with affine maps. This characterization is comparable to the one given by Manes for the power set monad (which also does not resort explicitly to the underlying monad or its functor); see Example 2.5.14.

Lemma 2.5.20 *Given an algebra $\langle X, h \rangle$ for $\boldsymbol{D}$, define for $x_1, \dots, x_n \in X$ and the positive convex coefficients $\langle \alpha_1, \dots, \alpha_n \rangle \in \Omega$; put*

$$\langle \alpha_1, \dots, \alpha_n \rangle_{\mathfrak{p}}(x_1, \dots, x_n) := h\left(\sum_{i=1}^{n} \alpha_i \cdot \delta_{x_i}\right)$$

This defines a positive convex structure $\mathfrak{p}$ on X.

Proof 1. Write $\sum_{i=1}^{n} \alpha_i \cdot x_i \;:=\; h(\sum_{i=1}^{n} \alpha_i \cdot \delta_{x_i})$ for convenience. Because $h\big(\sum_{i=1}^{n} \delta_{i,j} \cdot \delta_{x_i}\big) = h(\delta_{x_j}) = x_j$, property ☆ in Definition 2.5.18 is satisfied.

2. Proving property ✪, we resort to the properties of an algebra and a monad:

$$
\begin{aligned}
\sum_{i=1}^{n} \alpha_i \cdot (\sum_{k=1}^{m} \beta_{i,k} \cdot x_k) &= h(\sum_{i=1}^{n} \alpha_i \cdot \delta_{\sum_{k=1}^{m} \beta_{i,k} \cdot x_k}) && (2.2)\\
&= h(\sum_{i=1}^{n} \alpha_i \cdot \delta_{h(\sum_{k=1}^{m} \beta_{i,k} \cdot \delta_{x_k})}) && (2.3)\\
&= h(\sum_{i=1}^{n} \alpha_i \cdot \mathbb{S}(h)(\delta_{\sum_{k=1}^{m} \beta_{i,k} \cdot \delta_{x_k}})) && (2.4)\\
&= (h \circ \mathbb{S}(h))(\sum_{i=1}^{n} \alpha_i \cdot \delta_{\sum_{k=1}^{m} \beta_{i,k} \cdot \delta_{x_k}}) && (2.5)\\
&= (h \circ \mu_X)(\sum_{i=1}^{n} \alpha_i \cdot \delta_{\sum_{k=1}^{m} \beta_{i,k} \cdot \delta_{x_k}}) && (2.6)\\
&= h(\sum_{i=1}^{n} \alpha_i \cdot \mu_X(\delta_{\sum_{k=1}^{m} \beta_{i,k} \cdot \delta_{x_k}})) && (2.7)\\
&= h(\sum_{i=1}^{n} \alpha_i \cdot (\sum_{k=1}^{m} \beta_{i,k} \cdot \delta_{x_k})) && (2.8)\\
&= h(\sum_{k=1}^{m} (\sum_{i=1}^{n} \alpha_i \cdot \beta_{i,k}) \delta_{x_k}) && (2.9)\\
&= \sum_{k=1}^{m} (\sum_{i=1}^{n} \alpha_i \cdot \beta_{i,k}) x_k. && (2.10)
\end{aligned}
$$

Equations (2.2) and (2.3) reflect the definition of the structure, Eq. (2.4) applies $\delta_{h(\tau)} = \mathbb{S}(h)(\delta_\tau)$, Eq. (2.5) uses the linearity of $\mathbb{S}(h)$ according to Lemma 2.5.19, and Eq. (2.6) is due to h being an algebra. Winding down, Eq. (2.7) uses Lemma 2.5.19 again; this time for μ_X, Eq. (2.8) uses that $\mu_X \circ \delta_\tau = \tau$; Eq. (2.9) is just rearranging terms; and Eq. (2.10) is the definition again. ⊣

The converse holds as well, as we will show now.

Lemma 2.5.21 *Let $\mathfrak{p}$ be a positive convex structure on X. Put*

$$h\big(\sum_{i=1}^{n} \alpha_i \cdot \delta_{x_i}\big) := \sum_{1 \le i \le n}^{\mathfrak{p}} \alpha_i \cdot x_i$$

for $\langle \alpha_1, \ldots, \alpha_n \rangle \in \Omega$ and $x_1, \ldots, x_n \in X$. Then $\langle X, h \rangle$ is an algebra.

Proof 0. We show first that h is well defined, and then we establish that h is an affine map, so that we may interchange the application of h with summation. Then we apply the elementary properties established in Lemma 2.5.19 for $\mu_X : (\boldsymbol{D} \circ \boldsymbol{D})X \to \boldsymbol{D}X$ to show that the equation $h \circ \mu_X = id_X$ holds. Outline

1. We first check that h is well defined: This is so since

$$\sum_{i=1}^{n} \alpha_i \cdot \delta_{x_i} = \sum_{j=1}^{m} \alpha'_j \cdot \delta_{x'_j}$$

implies that

$$\sum_{i=1,\alpha_i\neq 0}^{n} \alpha_i \cdot \delta_{x_i} = \sum_{j=1,\alpha'_j\neq 0}^{m} \alpha'_j \cdot \delta_{x'_j};$$

hence given i with $\alpha_i \neq 0$, there exists j with $\alpha'_j \neq 0$ such that $x_i = x'_j$ with $\alpha_i = \alpha'_j$ and vice versa. Consequently,

$$\begin{aligned}\sum\nolimits_{1\leq i\leq n}^{\mathfrak{p}} \alpha_i \cdot x_i &= \sum_{1\leq i\leq n,\alpha_i\neq 0}^{\mathfrak{p}} \alpha_i \cdot x_i = \sum_{1\leq j\leq n,\alpha'_j\neq 0}^{\mathfrak{p}} \alpha'_j \cdot x'_j \\ &= \sum_{1\leq j\leq n}^{\mathfrak{p}} \alpha'_j \cdot x'_j\end{aligned}$$

is inferred from the properties of positive convex structures. Thus $h : \boldsymbol{D}X \to X$.

An easy induction using property ✪ shows that h is an affine map, i.e., that we have

$$h\big(\sum\nolimits_{i=1}^{n} \alpha_i \cdot \tau_i\big) = \sum\nolimits_{1\leq i\leq n}^{\mathfrak{p}} \alpha_i \cdot h(\tau_i) \tag{2.11}$$

for $\langle \alpha_1, \ldots, \alpha_n\rangle \in \Omega$ and $\tau_1, \ldots, \tau_n \in \boldsymbol{D}X$.

Now let $f = \sum_{i=1}^{n} \alpha_i \cdot \delta_{\tau_i} \in \boldsymbol{D}^2X$ with $\tau_1, \ldots, \tau_n \in \boldsymbol{D}X$. Then we obtain from Lemma 2.5.19 that $\mu_X f = \sum_{i=1}^{n} \alpha_i \cdot \tau_i$. Consequently, we obtain from (2.11) that $h(\mu_X f) = \sum_{1\leq i\leq n}^{\mathfrak{p}} \alpha_i \cdot h(\tau_i)$. On the other hand, Lemma 2.5.19 implies together with (2.11)

$$\begin{aligned}(h \circ \boldsymbol{D}h) f &= h\big(\sum\nolimits_{1\leq i\leq n}^{\mathfrak{p}} \alpha_i \cdot (\boldsymbol{D}h)(\tau_i)\big) \\ &= \sum\nolimits_{1\leq i\leq n}^{\mathfrak{p}} \alpha_i \cdot h\big((\boldsymbol{D}h)(\tau_i)\big) \\ &= \sum\nolimits_{1\leq i\leq n}^{\mathfrak{p}} \alpha_i \cdot h(\delta_{h(\tau_i)}) \\ &= \sum\nolimits_{1\leq i\leq n}^{\mathfrak{p}} \alpha_i \cdot h(\tau_i),\end{aligned}$$

because $h(\delta_{h(\tau_i)}) = h(\tau_i)$. We infer from ☆ that $h \circ \mu_X = id_X$. ⊣

Hence we have established:

Proposition 2.5.22 *Each positive convex structure on X induces an algebra for $\boldsymbol{D}X$.* ⊣

Summarizing, we obtain a complete characterization of the Eilenberg–Moore algebras for this monad.

Theorem 2.5.23 *The Eilenberg–Moore algebras for the discrete probability monad are exactly the positive convex structures.* $\dashv$

This characterization carries over to the probabilistic version of the monad; we leave the simple formulation to the reader. A similar characterization is possible for the continuous version of this functor, at least in Polish spaces. This requires a continuity condition, however, and is further discussed in Sect. 4.10.2.

2.6 Coalgebras

A coalgebra for a functor $\boldsymbol{F}$ is characterized by a carrier object c and by a morphism $c \to \boldsymbol{F}c$. This fairly general structure can be found in many applications, as we will see. So we will first define formally what a coalgebra is and then provide a gallery of examples, some of them already discussed in another disguise, some of them new. The common thread is their formulation as a coalgebra. The fundamental notion of bisimilarity is introduced, and bisimilar coalgebras will be discussed, indicating some interesting facets of the possibilities to describe behavioral equivalence of some sorts.

Definition 2.6.1 *Given the endofunctor* $\boldsymbol{F}$ *on category* $\boldsymbol{K}$*, an object* a *on* $\boldsymbol{K}$ *together with a morphism* $f : a \to \boldsymbol{F}a$ *is a* coalgebra (a, f) *for* $\boldsymbol{K}$*. Morphism* f *is sometimes called the* dynamics *of the coalgebra,* a *its* carrier.

Comparing the definitions of an algebra and a coalgebra, we see that for a coalgebra, the functor $\boldsymbol{F}$ is an arbitrary endofunctor on $\boldsymbol{K}$, while an algebra requires a monad and compatibility with unit and multiplication. Thus coalgebras are conceptually simpler by imposing less constraints.

We are going to enter now the gallery of examples and start with coalgebras for the power set functor. This example will be with us for quite some time, in particular when we will interpret modal logics. A refinement of this example will be provided by labeled transition systems.

Example 2.6.2 We consider the category $\boldsymbol{Set}$ of sets with maps as morphisms and the power set functor $\mathcal{P}$. An $\mathcal{P}$-coalgebra consists of a set A and a map $f : A \to \mathcal{P}(A)$. Hence we have $f(a) \subseteq A$ for all $a \in A$, so that a ***Set*** coalgebra can be represented as a relation $\{\langle a, b\rangle \mid b \in f(a), a \in A\}$ over A. If, conversely, $R \subseteq A \times A$ is a relation, then $f(a) := \{b \in A \mid \langle a, b\rangle \in R\}$ is a map $f : A \to \mathcal{P}(A)$. ✌

A slight extension is to be observed when we introduce actions, formally captured as labels for our transitions. Here a transition is dependent on an action, which then serves as a label to the corresponding relation.

Example 2.6.3 Let us interpret a labeled transition system $\big(S, (\rightsquigarrow_a)_{a\in A}\big)$ over state space S with set A of actions; see Example 2.3.10. Then $\rightsquigarrow_a \subseteq S \times S$ for all actions $a \in A$.

Working again in ***Set***, we define for the set S and for the map $f : S \to T$

$$\begin{aligned} \boldsymbol{T}S &:= \mathcal{P}(A \times S), \\ (\boldsymbol{T}f)(B) &:= \{\langle a, f(x)\rangle \mid \langle a, x\rangle \in B\} \end{aligned}$$

(hence $\boldsymbol{T} = \mathcal{P}(A \times -)$). Define $f(s) := \{\langle a, s'\rangle \mid s \rightsquigarrow_a s'\}$; thus $f : S \to \boldsymbol{T}S$ is a morphism in ***Set***. Consequently, a labeled transition system is interpreted as a coalgebra for the functor $\mathcal{P}(A \times -)$. ✌

Example 2.6.4 Let A be the inputs, B the outputs, and X the states of an automaton with output; see Example 2.3.9. Put $\boldsymbol{F} := (- \times B)^A$. For $f : X \to Y$, we have this commutative diagram:

$$\begin{tikzcd} & A \arrow[dl, "t"'] \arrow[dr, "(\boldsymbol{F}f)t"] & \\ X \times B \arrow[rr, "f \times id_B"'] & & Y \times B \end{tikzcd}$$

Let (S, f) be an $\boldsymbol{F}$-coalgebra; thus $f : S \to \boldsymbol{F}S = (S \times B)^A$. Input $a \in A$ in state $s \in S$ yields $f(s)(a) = \langle s', b\rangle$, so that s' is the new state and b is the output. Hence automata with output are perceived as coalgebras, in this case for the functor $(- \times B)^A$. ✌

While the automata in Example 2.6.4 are deterministic (and completely specified), we can also use a similar approach to modeling nondeterministic automata.

Example 2.6.5 Let A, B, X be as in Example 2.6.4, but take this time $\boldsymbol{F} := \mathcal{P}(- \times B)^A$ as a functor, so that this diagram commutes for $f : X \to Y$:

$$\begin{tikzcd}
& A \arrow[dl, "t"'] \arrow[dr, "(\boldsymbol{F}f)t"] & \\
\mathcal{P}(X \times B) \arrow[rr, "\mathcal{P}(f \times id_B)"'] & & \mathcal{P}(Y \times B)
\end{tikzcd}$$

Thus $\mathcal{P}(f \times B)(D) = \{\langle f(x), b\rangle \in Y \times B \mid \langle x, b\rangle \in B\}$. Then (S, g) is an $\boldsymbol{F}$ coalgebra iff input $a \in A$ in state $s \in S$ gives $g(s)(a) \in S \times B$ as the set of possible new states and outputs.

As a variant, we can replace $\mathcal{P}(- \times B)$ by $\mathcal{P}_f(- \times B)$, so that the automaton presents only a finite number of alternatives. ✌

Binary trees may be modeled through coalgebras as well:

Example 2.6.6 Put $\boldsymbol{F}X := \{*\} + X \times X$, where $*$ is a new symbol. If $f : X \to Y$, put

$$\boldsymbol{F}(f)(t) := \begin{cases} *, & \text{if } t = * \\ \langle x_1, x_2\rangle, & \text{if } t = \langle x_1, x_2\rangle. \end{cases}$$

Then $\boldsymbol{F}$ is an endofunctor on ***Set***. Let (S, f) be an $\boldsymbol{F}$-coalgebra, then $f(s) \in \{*\} + S \times S$. This is interpreted that s is a leaf iff $f(s) = *$ and an inner node with offsprings $\langle s_1, s_2\rangle$, if $f(s) = \langle s_1, s_2\rangle$. Thus such a coalgebra represents a binary tree (which may be of infinite depth). ✌

The following example shows that probabilistic transitions may also be modeled as coalgebras.

Example 2.6.7 Working in the category ***Meas*** of measurable spaces with measurable maps, we have introduced in Example 2.4.8 the subprobability functor $\mathbb{S}$ as an endofunctor on ***Meas***. Let (X, K) be a coalgebra for $\mathbb{S}$ (we omit here the σ-algebra from the notation); then $K : X \to \mathbb{S}X$ is measurable, so that:

1. $K(x)$ is a subprobability on (the measurable sets of) X,
2. for each measurable set $D \subseteq X$, the map $x \mapsto K(x)(D)$ is measurable,

See Example 2.1.14 and Exercise 2.7. Thus K is a subprobabilistic transition kernel (or a stochastic relation) on X. ✌

Let us have a look at the upper closed sets introduced in Example 2.3.13. Coalgebras for this functor will be used for an interpretation of games; see Example 2.7.22.

Example 2.6.8 Let $\mathbf{V}S := \{V \subseteq \mathcal{P}S \mid V \text{ is upper closed}\}$. This functor has been studied in Example 2.3.13. A coalgebra (S, f) for $\mathbf{V}$ is a map $f : S \to \mathbf{V}S$, so that $f(s) \subseteq \mathcal{P}(S)$ is upper closed; hence $A \in f(s)$ and $B \supseteq A$ imply $B \in f(s)$ for each $s \in S$. We interpret $f(s)$ as the collection of all sets of states a player has a strategy to reach in state s, so that if the player can reach A and $A \subseteq B$, then the player certainly can reach B.

$\mathbf{V}$ is the basis for neighborhood models in modal logics; see, e.g., [Che89, Ven07] and page 232. ✌

It is natural to ask for morphisms of coalgebras, which relate coalgebras to each other. This is a fairly straightforward definition.

Definition 2.6.9 *Let $\mathbf{F}$ be an endofunctor on category $\mathbf{K}$, then $t : (a, f) \to (b, g)$ is a* coalgebra morphism *for the $\mathbf{F}$-coalgebras (a, f) and (b, g) iff $t : a \to b$ is a morphism in $\mathbf{K}$ such that $g \circ t = \mathbf{F}(t) \circ f$.*

Thus $t : (a, f) \to (b, g)$ is a coalgebra morphism iff $t : a \to b$ is a morphism so that this diagram commutes:

$$\begin{tikzcd} a \arrow[r, "t"] \arrow[d, "f"'] & b \arrow[d, "g"] \\ \mathbf{F}a \arrow[r, "\mathbf{F}t"'] & \mathbf{F}b \end{tikzcd}$$

It is clear that $\mathbf{F}$-coalgebras form a category with coalgebra morphisms as morphisms. We reconsider some previously discussed examples and shed some light on the morphisms for these coalgebras.

Example 2.6.10 Continuing Example 2.6.6 on binary trees, let $r : (S, f) \to (T, g)$ be a morphism for the $\mathbf{F}$-coalgebras (S, f) and (T, g). Thus $g \circ r = \mathbf{F}(r) \circ f$. This entails:

1. $f(s) = *$, then $g(r(s)) = (\mathbf{F}r)(f(s)) = *$ (thus s is a leaf iff $r(s)$ is one),
2. $f(s) = \langle s_1, s_2\rangle$, then $g(r(s)) = \langle t_1, t_2\rangle$ with $t_1 = r(s_1)$ and $t_2 = r(s_2)$ (thus $r(s)$ branches out to $\langle r(s_1), r(s_2)\rangle$, provided s branches out to $\langle s_1, s_2\rangle$).

A coalgebra morphism preserves the tree structure. ✌

Example 2.6.11 Continuing the discussion of deterministic automata with output from Example 2.6.4, let (S, f) and (T, g) be $\boldsymbol{F}$-coalgebras and $r : (S, F) \rightarrow (T, g)$ be a morphism. Given state $s \in S$, let $f(s)(a) = \langle s', b\rangle$ be the new state and the output, respectively, after input $a \in A$ for automaton (S, f). Then $g(r(s))(a) = \langle r(s'), b\rangle$, so after input $a \in A$, the automaton (T, g) will be in state $r(s)$ and give the output b, as expected. Hence coalgebra morphisms preserve the working of the automata. ✌

Example 2.6.12 Continuing the discussion of transition systems from Example 2.6.3, let (S, f) and (T, g) be labeled transition systems with A as the set of actions. Thus a transition from s to s' on action a is given in (S, f) iff $\langle a, s'\rangle \in f(s)$. Let us just for convenience write $s \rightsquigarrow_{a,S} s'$ iff this is the case; similarly, we write $t \rightsquigarrow_{a,T} t'$ iff $t, t' \in T$ with $\langle a, t'\rangle \in g(t)$.

Now let $r : (S, f) \rightarrow (T, g)$ be a coalgebra morphism. We claim that for given $s \in S$, we have a transition $r(s) \rightsquigarrow_{a,T} t_0$ for some t_0 iff we can find s_0 such that $s \rightsquigarrow_{a,S} s_0$ and $r(s_0) = t_0$. Because $r : (S, f) \rightarrow (T, g)$ is a coalgebra morphism, we have $g \circ r = (\boldsymbol{T}r) \circ f$ with $\boldsymbol{T} = \mathcal{P}(A \times -)$. Thus

$$g(r(s)) = \mathcal{P}(A \times r)(s) = \{\langle a, r(s')\rangle \mid \langle a, s'\rangle \in f(s)\}.$$

Consequently,

$$\begin{aligned} r(s) \rightsquigarrow_{a,T} t_0 &\Leftrightarrow \langle a, t_0\rangle \in g(r(s)) \\ &\Leftrightarrow \langle a, t_0\rangle = \langle a, r(s_0)\rangle \text{ for some } \langle a, s_0\rangle \in f(s) \\ &\Leftrightarrow s \rightsquigarrow_{a,S} s_0 \text{ for some } s_0 \text{ with } r(s_0) = t_0 \end{aligned}$$

This means that the transitions in (T, g) are essentially controlled by the morphism r and the transitions in (S, f). Hence a coalgebra morphism between transition systems is a bounded morphism in the sense of Example 2.1.10. ✌

Example 2.6.13 We continue the discussion of upper closed sets from Example 2.6.8. Let (S, f) and (T, g) be $\boldsymbol{V}$-coalgebras, so this diagram is commutative for morphism $r : (S, F) \rightarrow (T, g)$:

$$\begin{tikzcd} S \arrow[r, "r"] \arrow[d, "f"'] & T \arrow[d, "g"] \\ \boldsymbol{V}S \arrow[r, "\boldsymbol{V}r"'] & \boldsymbol{V}T \end{tikzcd}$$

Consequently, $W \in g(r(s))$ iff $r^{-1}\left[W\right] \in f(s)$. Taking up the interpretation of sets of states which may be achieved by a player, we see that it[1] may achieve W in state $r(s)$ in (T, g) iff it may achieve in (S, f) the set $r^{-1}\left[W\right]$ in state s. ✌

2.6.1 Bisimulations

The notion of bisimilarity is fundamental for the application of coalgebras to system modeling. Bisimilar coalgebras behave in a similar fashion, witnessed by a mediating system.

Definition 2.6.14 *Let $\boldsymbol{F}$ be an endofunctor on a category $\boldsymbol{K}$. The $\boldsymbol{F}$-coalgebras (S, f) and (T, g) are said to be* bisimilar *iff there exists a coalgebra (M, m) and coalgebra morphisms* $(S, f) \longleftarrow (M, m) \longrightarrow (T, g)$. *The coalgebra (M, m) is called* mediating.

Thus we obtain this characteristic diagram with ℓ and r as the corresponding morphisms.

$$\begin{tikzcd}
S \arrow[d, "f"'] & M \arrow[l, "\ell"'] \arrow[r, "r"] \arrow[d, "m"] & T \arrow[d, "g"] \\
\boldsymbol{F}F & \boldsymbol{F}M \arrow[l, "\boldsymbol{F}\ell"] \arrow[r, "\boldsymbol{F}r"'] & \boldsymbol{F}T
\end{tikzcd}$$

This gives us $f \circ \ell = (\boldsymbol{F}\ell) \circ m$ and together with $g \circ r = (\boldsymbol{F}r) \circ m$. It is easy to see why (M, m) is called mediating.

Bisimilarity was originally investigated when concurrent systems became of interest. The original formulation, however, was not coalgebraic but rather relational. Here it is:

Definition 2.6.15 *Let $(S, \rightsquigarrow_S)$ and $(T, \rightsquigarrow_T)$ be transition systems. Then $B \subseteq S \times T$ is called a* bisimulation *iff for all $\langle s, t\rangle \in B$ these conditions are satisfied:*

1. *if $s \rightsquigarrow_S s'$, then there is a $t' \in T$ such that $t \rightsquigarrow_T t'$ and $\langle s', t'\rangle \in B$,*

[1] The present author is not really sure about the players' gender—players are considered female in the overwhelming majority of papers in the literature, but addressed as *Angel* or *Demon*. This may be politically correct, but does not seem to be biblically so with a view toward Matthew 22:30. To be on the safe side, players are neutral in the present treatise.

2. *if $t \rightsquigarrow_T t'$, then there is a $s' \in S$ such that $s \rightsquigarrow_S s'$ and $\langle s', t'\rangle \in B$.*

Hence a bisimulation simulates transitions in one system through the other one. On first sight, these notions of bisimilarity are not related to each other. Recall that transition systems are coalgebras for the power set functor $\mathcal{P}$. This is the connection:

Theorem 2.6.16 *Given the transition systems $(S, \rightsquigarrow_S)$ and $(T, \rightsquigarrow_T)$ with the associated $\mathcal{P}$-coalgebras (S, f) and (T, g), then these statements are equivalent for $B \subseteq S \times T$:*

1. *B is a bisimulation.*
2. *There exists a $\mathcal{P}$-coalgebra structure h on B such that $(S, f) \longleftarrow (B, h) \longrightarrow (T, g)$ with the projections as morphisms is mediating.*

Proof That $(S, f) \xleftarrow{\pi_S} (B, h) \xrightarrow{\pi_T} (T, g)$ is mediating follows from commutativity of this diagram:

$$\begin{tikzcd}
S \arrow[d, "f"'] & B \arrow[l, "\pi_S"'] \arrow[r, "\pi_T"] \arrow[d, "h"'] & T \arrow[d, "g"] \\
\mathcal{P}(S) & \mathcal{P}(B) \arrow[l, "\mathcal{P}(\pi_S)"] \arrow[r, "\mathcal{P}(\pi_T)"'] & \mathcal{P}(T)
\end{tikzcd}$$

1 $\Rightarrow$ 2: We have to construct a map $h : B \to \mathcal{P}(B)$ such that $f(\pi_S(s, t)) = \mathcal{P}(\pi_S)(h(s, t))$ and $f(\pi_T(s, t)) = \mathcal{P}(\pi_T)(h(s, t))$ for all $\langle s, t\rangle \in B$. The choice is somewhat obvious: put for $\langle s, t\rangle \in B$

$$h(s, t) := \{\langle s', t'\rangle \in B \mid s \rightsquigarrow_S s', t \rightsquigarrow_T t'\}.$$

Thus $h : B \to \mathcal{P}(B)$ is a map; hence (B, h) is a $\mathcal{P}$-coalgebra.

Now fix $\langle s, t\rangle \in B$; then we claim that $f(s) = \mathcal{P}(\pi_S)(h(s, t))$.

"$\subseteq$": Let $s' \in f(s)$; hence $s \rightsquigarrow_S s'$; thus there exists t' with $\langle s', t'\rangle \in B$ such that $t \rightsquigarrow_T t'$; hence

$$\begin{aligned}
s' &\in \{\pi_S(s_0, t_0) \mid \langle s_0, t_0\rangle \in h(s, t)\} \\
&= \{s_0 \mid \langle s_0, t_0\rangle \in h(s, t) \text{ for some } t_0\} \\
&= \mathcal{P}(\pi_S)(h(s, t)).
\end{aligned}$$

"$\supseteq$": If $s' \in \mathcal{P}(\pi_S)(h(s, t))$, then in particular $s \rightsquigarrow_S s'$; thus $s' \in f(s)$.

Thus we have shown that $\mathcal{P}(\pi_S)(h(s,t)) = f(s) = f(\pi_S(s,t))$. One shows $\mathcal{P}(\pi_T)(h(s,t)) = g(t) = f(\pi_T(s,t))$ in exactly the same way. We have constructed h such that (B, h) is a $\mathcal{P}$-coalgebra and such that the diagrams above commute.

2 $\Rightarrow$ 1: Assume that h exists with the properties described in the assertion; then we have to show that B is a bisimulation. Now let $\langle s, t\rangle \in B$ and $s \rightsquigarrow_S s'$; hence $s' \in f(s) = f(\pi_S(s,t)) = \mathcal{P}(\pi_S)(h(s,t))$. Thus there exists t' with $\langle s', t'\rangle \in h(s,t) \subseteq B$, and hence $\langle s', t'\rangle \in B$. We claim that $t \rightsquigarrow_T t'$, which is tantamount to saying $t' \in g(t)$. But $g(t) = \mathcal{P}(\pi_T)(h(s,t))$, and $\langle s', t'\rangle \in h(s,t)$; hence $t' \in \mathcal{P}(\pi_T)(h(s,t)) = g(t)$. This establishes $t \rightsquigarrow_T t'$. A similar argument finds s' with $s \rightsquigarrow_S s'$ with $\langle s', t'\rangle \in B$ in case $t \rightsquigarrow_T t'$.

This completes the proof. $\dashv$

Thus we may use bisimulations for transition systems as relations and bisimulations as coalgebras interchangeably, and this characterization suggests a definition in purely coalgebraic terms for those cases in which a set-theoretic relation is not available or not adequate. The connection to $\mathcal{P}$-coalgebra morphisms and bisimulations is further strengthened by investigating the graph of a morphism (recall that the graph of a map $r : S \to T$ is the relation $\text{graph}(r) := \{\langle s, r(s)\rangle \mid s \in S\}$).

graph(r)

Proposition 2.6.17 *Given coalgebras (S, f) and (T, g) for the power set functor $\mathcal{P}$, $r : (S, f) \to (T, g)$ is a morphism iff* $\text{graph}(r)$ *is a bisimulation for (S, f) and (T, g).*

Proof 1. Assume that $r : (S, f) \to (T, g)$ is a morphism, so that $g \circ r = \mathcal{P}(r) \circ f$. Now define

$$h(s,t) := \{\langle s', r(s')\rangle \mid s' \in f(s)\} \subseteq \text{graph}(r)$$

for $\langle s, t\rangle \in \text{graph}(r)$. Then $g(\pi_T(s,t)) = g(t) = \mathcal{P}(\pi_T)(h(s,t))$ for $t = r(s)$.

"$\subseteq$": If $t' \in g(t)$ for $t = r(s)$, then

$$\begin{aligned} t' \in g(r(s)) \quad &= \mathcal{P}(r)(f(s)) = \{r(s') \mid s' \in f(s)\} \\ &= \mathcal{P}(\pi_T)(\{\langle s', r(s')\rangle \mid s' \in f(s)\}) \\ &= \mathcal{P}(\pi_T)(h(s,t)) \end{aligned}$$

"$\supseteq$": If $\langle s', t'\rangle \in h(s,t)$, then $s' \in f(s)$ and $t' = r(s')$, but this implies $t' \in \mathcal{P}(r)(f(s)) = g(r(s))$.

Thus $g \circ \pi_T = \mathcal{P}(\pi_T) \circ h$. The equation $f \circ \pi_S = \mathcal{P}(\pi_S) \circ h$ is established similarly.

Hence we have found a coalgebra structure h on graph(r) such that

$$(S, f) \xleftarrow{\pi_S} (\mathrm{graph}(r), h) \xrightarrow{\pi_T} (T, g)$$

are coalgebra morphisms, so that (graph$(r), h$) is now officially a bisimulation.

2. If, conversely, (graph$(r), h$) is a bisimulation with the projections as morphisms, then we have $r = \pi_T \circ \pi_S^{-1}$. Then π_T is a morphism, and π_S^{-1} is a morphism as well (note that we work on the graph of r). So r is a morphism. $\dashv$

Let us have a look at upper closed sets from Example 2.4.10. There we find a comparable situation. We cannot, however, translate the definition directly, because we do not have access to the transitions proper, but rather to the sets from which the next state may come from. Let (S, f) and (T, g) be $\mathbf{V}$-coalgebras, and assume that $\langle s, t\rangle \in B$. Assume $X \in f(s)$; then we want to find $Y \in g(t)$ such that, when we take $t' \in Y$, we find a state $s' \in X$ with s' being related via B to s', and vice versa. Formally:

Definition 2.6.18 *Let*

$$\mathbf{V}S := \{V \subseteq \mathcal{P}(S) \mid V \text{ is upper closed}\}$$

be the endofunctor on **Set** *which assigns to set S all upper closed subsets of $\mathcal{P}S$. Given $\mathbf{V}$-coalgebras (S, f) and (T, g), a subset $B \subseteq S \times T$ is called a* bisimulation *of (S, f) and (T, g) iff for each $\langle s, t\rangle \in B$*

1. *for all $X \in f(s)$, there exists $Y \in g(t)$ such that for each $t' \in Y$, there exists $s' \in X$ with $\langle s', t'\rangle \in B$,*

2. *for all $Y \in g(t)$, there exists $X \in f(s)$ such that for each $s' \in X$, there exists $t' \in Y$ with $\langle s', t'\rangle \in B$.*

We have then a comparable characterization of bisimilar coalgebras.

Proposition 2.6.19 *Let (S, f) and (T, g) be coalgebras for $\mathbf{V}$. Then the following statements are equivalent for $B \subseteq S \times T$ with $\pi_S\big[B\big] = S$ and $\pi_T\big[B\big] = T$*

1. *B is a bisimulation of (S, f) and (T, g).*

2. *There exists a coalgebra structure h on B so that the projections $\pi_S : B \to S, \pi_T : B \to T$ are morphisms $(S, f) \xleftarrow{\pi_S} (B, h) \xrightarrow{\pi_T} (T, g)$.*

Proof 1 $\Rightarrow$ 2: Define $\langle s, t\rangle \in B$

$$h(s,t) := \{D \subseteq B \mid \pi_S\big[D\big] \in f(s) \text{ and } \pi_T\big[D\big] \in f(t)\}.$$

Hence $h(s,t) \subseteq \mathcal{P}(S)$, and because both $f(s)$ and $g(t)$ are upper closed, so is $h(s,t)$.

Now fix $\langle s,t\rangle \in B$. We show first that $f(s) = \{\pi_S\big[Z\big] \mid Z \in h(s,t)\}$. From the definition of $h(s,t)$, it follows that $\pi_S\big[Z\big] \in f(s)$ for each $Z \in h(s,t)$. So we have to establish the other inclusion. Let $X \in f(s)$; then $X = \pi_S\big[\pi_S^{-1}\big[X\big]\big]$, because $\pi_S : B \to S$ is onto, so it suffices to show that $\pi_S^{-1}\big[X\big] \in h(s,t)$ hence that $\pi_T\big[\pi_S^{-1}\big[X\big]\big] \in g(t)$. Given X, there exists $Y \in g(t)$ so that for each $t' \in Y$, there exists $s' \in X$ such that $\langle s', t'\rangle \in B$. Thus $Y = \pi_T\big[(X \times Y) \cap B\big]$. But this implies $Y \subseteq \pi_T\big[\pi_S^{-1}\big[X\big]\big]$; hence $Y \subseteq \pi_T\big[\pi_S^{-1}\big[X\big]\big] \in g(t)$. One similarly shows that $g(t) = \{\pi_T\big[Z\big] \mid Z \in h(s,t)\}$.

In a second step, we show that

$$\{\pi_S\big[Z\big] \mid Z \in h(s,t)\} = \{C \mid \pi_S^{-1}\big[C\big] \in h(s,t)\}.$$

In fact, if $C = \pi_S\big[Z\big]$ for some $Z \in h(s,t)$, then $Z \subseteq \pi_S^{-1}\big[C\big] = \pi_S^{-1}\big[\pi_S\big[Z\big]\big]$; hence $\pi_S^{-1}\big[C\big] \in h(s,t)$. If, conversely, $Z := \pi_S^{-1}\big[C\big] \in h(s,t)$, then $C = \pi_S\big[Z\big]$. Thus we obtain

$$\begin{aligned} f(s) &= \{\pi_S\big[Z\big] \mid Z \in h(s,t)\} = \{C \mid \pi_S^{-1}\big[C\big] \in h(s,t)\} \\ &= (\mathbf{V}\pi_S)(h(s,t)) \end{aligned}$$

for $\langle s,t\rangle \in B$. Summarizing, this means that $\pi_S : (B,h) \to (S, f)$ is a morphism. A very similar argumentation shows that $\pi_T : (B,h) \to (T, g)$ is a morphism as well.

2 $\Rightarrow$ 1: Assume, conversely, that the projections are coalgebra morphisms, and let $\langle s,t\rangle \in B$. Given $X \in f(s)$, we know that $X = \pi_S\big[Z\big]$ for some $Z \in h(s,t)$. Thus we find for any $t' \in Y$ some $s' \in X$ with $\langle s', t'\rangle \in B$. The symmetric property of a bisimulation is established exactly in the same way. Hence B is a bisimulation for (S, f) and (T, g). $\dashv$

Encouraged by these observations, we define bisimulations for set- based functors, i.e., for endofunctors on the category ***Set*** of sets with maps as morphisms. This is nothing but a specialization of the general notion of bisimilarity, taking specifically into account that in ***Set*** we may consider subsets of the Cartesian product and that we have projections at our disposal.

Definition 2.6.20 *Let $\boldsymbol{F}$ be an endofunctor on* **Set**. *Then $R \subseteq S \times T$ is called a* bisimulation *for the $\boldsymbol{F}$-coalgebras (S, f) and (T, g) iff there exists a map $h : R \to \boldsymbol{F}(R)$ rendering this diagram commutative:*

$$\begin{tikzcd} S \arrow[d, "f"'] & R \arrow[l, "\pi_S"'] \arrow[r, "\pi_T"] \arrow[d, "h"'] & T \arrow[d, "g"] \\ \boldsymbol{F}(S) & \boldsymbol{F}(R) \arrow[l, "\boldsymbol{F}\pi_S"] \arrow[r, "\boldsymbol{F}\pi_T"'] & \boldsymbol{F}(T) \end{tikzcd}$$

These are immediate consequences:

Lemma 2.6.21 *$\Delta_S := \{\langle s, s\rangle \mid s \in S\}$ is a bisimulation for every $\boldsymbol{F}$-coalgebra (S, f). If R is a bisimulation for the $\boldsymbol{F}$-coalgebras (S, f) and (T, g), then R^{-1} is a bisimulation for (T, g) and (S, f).* $\dashv$

It is instructive to look back and investigate again the graph of a morphism $r : (S, f) \to (T, g)$, where this time we do not have the power set functor—as in Proposition 2.6.17—but a general endofunctor $\boldsymbol{F}$ on ***Set***.

Corollary 2.6.22 *Given coalgebras (S, f) and (T, g) for the endofunctor $\boldsymbol{F}$ on* **Set**, *$r : (S, f) \to (T, g)$ is a morphism iff* graph(r) *is a bisimulation for (S, f) and (T, g).*

Proof 0. The proof for Proposition 2.6.17 needs some small adjustments, because we do not know how exactly functor $\boldsymbol{F}$ is operating on maps.

1. If $r : (S, f) \to (T, g)$ is a morphism, we know that $g \circ r = \boldsymbol{F}(r) \circ f$. Consider the map $\tau : S \to S \times T$ which is defined as $s \mapsto \langle s, r(s)\rangle$, thus $\boldsymbol{F}(\tau) : \boldsymbol{F}(S) \to \boldsymbol{F}(S \times T)$. Define

$$h : \begin{cases} \text{graph}(r) & \to \boldsymbol{F}(\text{graph}(r)) \\ \langle s, r(s)\rangle & \mapsto \boldsymbol{F}(\tau)(f(s)) \end{cases}$$

Then it is not difficult to see that both $g \circ \pi_T = \boldsymbol{F}(\pi_T) \circ h$ and $f \circ \pi_S = \boldsymbol{F}(\pi_S) \circ h$ hold. Hence $(\text{graph}(r), h)$ is an $\boldsymbol{F}$-coalgebra mediating between (S, f) and (T, g).

2. Assume that graph(r) is a bisimulation for (S, f) and (T, g), then both π_T and π_S^{-1} are morphisms for the $\boldsymbol{F}$-coalgebras, so the proof proceeds exactly as the corresponding one for Proposition 2.6.17. $\dashv$

We will study some properties of bisimulations now, including a preservation property of functor $\boldsymbol{F} : \boldsymbol{Set} \to \boldsymbol{Set}$. This functor will be fixed for the time being.

We may construct bisimulations from morphisms. This is what we show first.

Lemma 2.6.23 *Let (S, f), (T, g), and (U, h) be $\boldsymbol{F}$-coalgebras with morphisms $\varphi : (S, f) \to (T, g)$ and $\psi : (S, f) \to (U, h)$. Then the image of S under $\varphi \times \psi$,*

$$\langle \varphi, \psi \rangle [S] := \{\langle \varphi(s), \psi(s) \rangle \mid s \in S\}$$

is a bisimulation for (T, g) and (U, h).

Proof 1. Look at this diagram:

$$\begin{tikzcd}
& \langle \varphi, \psi \rangle [S] \arrow[dl, "\pi_T"'] \arrow[dr, "\pi_U"] \arrow[d, bend left, "i"] & \\
T & S \arrow[l, "\varphi"] \arrow[r, "\psi"'] \arrow[u, bend left, "j"] & U
\end{tikzcd}$$

Here $j(s) := \langle \varphi(s), \psi(s) \rangle$; hence $j : S \to \langle \varphi, \psi \rangle [S]$ is surjective. We can find a map $i : \langle \varphi, \psi \rangle [S] \to S$ so that $j \circ i = id_{\langle \varphi, \psi \rangle [S]}$ using the Axiom of Choice: For each $r \in \langle \varphi, \psi \rangle [S]$, there exists at least one $s \in S$ with $r = \langle \varphi(s), \psi(s) \rangle$. Pick for each r such an s and call it $i(r)$; thus $r = \langle \varphi(i(r)), \psi(i(r)) \rangle$. So we have a left inverse to j, which will help us in the construction below.

2. We want to define a coalgebra structure for $\langle \varphi, \psi \rangle [S]$ such that the diagram below commutes, i.e., forms a bisimulation diagram. Put $k := \boldsymbol{F}(j) \circ f \circ i$; then we have

$$\begin{tikzcd}
T \arrow[d, "g"'] & \langle \varphi, \psi \rangle [S] \arrow[l, "\pi_T"'] \arrow[r, "\pi_U"] \arrow[d, "k"] & U \arrow[d, "h"] \\
\boldsymbol{F}T & \boldsymbol{F}(\langle \varphi, \psi \rangle [S]) \arrow[l, "\boldsymbol{F}\pi_T"] \arrow[r, "\boldsymbol{F}\pi_U"'] & \boldsymbol{F}U
\end{tikzcd}$$

Now

$$
\begin{aligned}
F(\pi_T) \circ k &= F(\pi_T) \circ F(j) \circ f \circ i \\
&= F(\pi_T \circ j) \circ f \circ i \\
&= F(\varphi) \circ f \circ i && \text{(since } \pi_T \circ j = \varphi\text{)} \\
&= g \circ \varphi \circ i && \text{(since } F(\varphi) \circ f = g \circ \varphi\text{)} \\
&= g \circ \pi_T
\end{aligned}
$$

Hence the left-hand diagram commutes. Similarly

$$F(\pi_U) \circ k = F(\pi_U \circ j) \circ f \circ i = F(\psi) \circ f \circ i = h \circ \psi \circ i = h \circ \pi_U$$

Thus we obtain a commutative diagram on the right-hand side as well. ⊣

This technical result is applied to the composition of relations:

Lemma 2.6.24 *Let $R \subseteq S \times T$ and $Q \subseteq T \times U$ be relations, and put $X := \{\langle s, t, u\rangle \mid \langle s, t\rangle \in R, \langle t, u\rangle \in Q\}$. Then*

$$R \circ Q = \langle \pi_S \circ \pi_R, \pi_U \circ \pi_Q\rangle[X].$$

Proof Simply trace an element of $R \circ Q$ through this construction:

$$
\begin{aligned}
\langle s, u\rangle \in R &\Leftrightarrow \exists t \in T : \langle s, t\rangle \in R, \langle t, u\rangle \in Q \\
&\Leftrightarrow \exists t \in T : \langle s, t, u\rangle \in X \\
&\Leftrightarrow \exists t \in T : s = (\pi_S \circ \pi_R)(s, t, u) \\
&\text{and } u = (\pi_U \circ \pi_Q)(s, t, u).
\end{aligned}
$$

⊣

Looking at X in its relation to the projections, we see that X is actually a weak pullback (Definition 2.2.18), to be precise:

Lemma 2.6.25 *Let R, Q, X be as above; then X is a weak pullback of $\pi_T^R : R \to T$ and $\pi_T^Q : Q \to T$, so that in particular $\pi_T^Q \circ \pi_Q = \pi_T^R \circ \pi_R$.*

Proof 1. We establish first that this diagram

$$
\begin{tikzcd}
X \arrow[r, "\pi_Q"] \arrow[d, "\pi_R"'] & Q \arrow[d, "\pi_Q^T"] \\
R \arrow[r, "\pi_T^R"'] & T
\end{tikzcd}
$$

commutes. In fact, given $\langle s, t, u\rangle \in X$, we know that $\langle s, t\rangle \in R$ and $\langle t, u\rangle \in Q$; hence $(\pi_T^Q \circ \pi_Q)(s, t, u) = \pi_T^Q(t, u) = t$ and $(\pi_T^R \circ \pi_R)(s, t, u) = \pi_T^R(s, t) = t$.

2. Now for the pullback property in its weak form. If $f_1 : Y \to R$ and $f_2 : Y \to Q$ are maps for some set Y such that $\pi_T^R \circ f_1 = \pi_T^R \circ f_2$, we can write $f_1(y) = \langle f_1^S(y), f_2^T(y)\rangle \in R$ and $f_2(y) = \langle f_2^T(y), f_2^U(y)\rangle \in Q$. Put $\sigma(y) := \langle f_1^S(y), f_2^T(y), f_2^U(y)\rangle$, then $\sigma : Y \to X$ with $f_1 = \pi_R \circ \sigma$ and $f_2 = \pi_q \circ \sigma$. Thus X is a weak pullback. $\dashv$

It will turn out that the functor should preserve the pullback property. Preserving the uniqueness property of a pullback will be too strong a requirement, but preserving weak pullbacks will be helpful and not too restrictive.

Definition 2.6.26 *Functor $\boldsymbol{F}$* preserves weak pullbacks *iff $\boldsymbol{F}$ maps weak pullbacks to weak pullbacks.*

Take a weak pullback diagram; then

$$\begin{tikzcd} H \arrow[dr, dashed] \arrow[drr, "g'"] \arrow[ddr, "f'"'] & & \\ & P \arrow[r, "g"] \arrow[d, "f"] & X \arrow[d, "h"] \\ & Y \arrow[r, "i"] & Z \end{tikzcd} \quad \text{translates under } \boldsymbol{F} \text{ to} \quad \begin{tikzcd} H \arrow[dr, dashed] \arrow[drr, "g'"] \arrow[ddr, "f'"'] & & \\ & \boldsymbol{F}P \arrow[r, "\boldsymbol{F}g"] \arrow[d, "\boldsymbol{F}f"] & \boldsymbol{F}X \arrow[d, "\boldsymbol{F}h"] \\ & \boldsymbol{F}Y \arrow[r, "\boldsymbol{F}i"] & \boldsymbol{F}Z \end{tikzcd}$$

We want to show that the composition of bisimulations is a bisimulation again: this requires that the functor preserves weak pullbacks. Before we state and prove a corresponding property, we need an auxiliary statement which is of independent interest, viz., that the weak pullback of bisimulations forms a bisimulation again. To be specific:

Lemma 2.6.27 *Assume that functor $\boldsymbol{F}$ preserves weak pullbacks, and let $r : (S, f) \to (T, g)$ and $s : (U, h) \to (T, g)$ be morphisms for the $\boldsymbol{F}$-coalgebras (S, f), (T, g), and (U, h). Then there exists a coalgebra structure $p : P \to \boldsymbol{F}P$ for the weak pullback P of r and s with projections π_S and π_T such that (P, p) is a bisimulation for (S, f) and (U, h).*

Proof We will need these diagrams:

$$\begin{tikzcd}
S \arrow[r, "r"] \arrow[d, "f"'] & R \arrow[d, "g"] & U \arrow[l, "s"'] \arrow[d, "h"] \\
\boldsymbol{F}S \arrow[r, "\boldsymbol{F}r"'] & \boldsymbol{F}T & \boldsymbol{F}U \arrow[l, "\boldsymbol{F}s"]
\end{tikzcd} \tag{2.12}$$

$$\begin{tikzcd}
P \arrow[r, "\pi_U"] \arrow[d, "\pi_S"'] & U \arrow[d, "s"] \\
S \arrow[r, "r"'] & T
\end{tikzcd} \tag{2.13}$$

$$\begin{tikzcd}
S \arrow[d, "f"'] & P \arrow[l, "\pi_S"'] \arrow[r, "\pi_U"] \arrow[d, dashed, "?!?"] & U \arrow[d, "h"] \\
\boldsymbol{F}S & \boldsymbol{F}P \arrow[l, "\boldsymbol{F}\pi_S"] \arrow[r, "\boldsymbol{F}\pi_U"'] & \boldsymbol{F}U
\end{tikzcd} \tag{2.14}$$

While the first two diagrams are helping with the proof's argument, the third diagram has a gap in the middle; this is this gap which we have to fill. We want to find an arrow $P \to \boldsymbol{F}P$ so that the diagrams will commute. Actually, the weak pullback will help us obtain this information. Plan

Because

$$\begin{aligned}
\boldsymbol{F}(r) \circ f \circ \pi_S &= g \circ r \circ \pi_S && \text{(diagram 2.12, left)} \\
&= g \circ s \circ \pi_U && \text{(diagram 2.13)} \\
&= \boldsymbol{F}(s) \circ h \circ \pi_U && \text{(diagram 2.12, right)}
\end{aligned}$$

we may conclude that $\boldsymbol{F}(r) \circ f \circ \pi_S = \boldsymbol{F}(s) \circ h \circ \pi_U$. Diagram 2.13 is a pullback diagram. Because $\boldsymbol{F}$ preserves weak pullbacks, this diagram can be complemented by an arrow $P \to \boldsymbol{F}P$ rendering the upper triangles commutative.

$$\begin{tikzcd}
P \arrow[dr, dashed] \arrow[drr, "h \circ \pi_U"] \arrow[ddr, "f \circ \pi_S"'] & & \\
& \boldsymbol{F}P \arrow[r, "\boldsymbol{F}\pi_U"] \arrow[d, "\boldsymbol{F}\pi_S"] & \boldsymbol{F}U \arrow[d, "\boldsymbol{F}s"'] \\
& \boldsymbol{F}S \arrow[r, "\boldsymbol{F}r"'] & \boldsymbol{F}T
\end{tikzcd}$$

Hence there exists $p : P \to \boldsymbol{F}P$ with $\boldsymbol{F}(\pi_S) \circ p = f \circ \pi_S$ and $\boldsymbol{F}(\pi_U) \circ p = h \circ \pi_U$. Thus p makes diagram (2.14) a bisimulation diagram. $\dashv$

Now we are in a position to show that the composition of bisimulations is a bisimulation again, provided the functor $\boldsymbol{F}$ behaves decently.

Proposition 2.6.28 *Given the $\boldsymbol{F}$-coalgebras (S, f), (T, g), and (U, h), assume that R is a bisimulation of (S, f) and (T, g) and Q is a bisimulation of (T, g) and (U, h), and assume moreover that $\boldsymbol{F}$ preserves weak pullbacks. Then $R \circ Q$ is a bisimulation of (S, f) and (U, h).*

Proof We can write $R \circ Q = \langle \pi_S \circ \pi_R, \pi_U \circ \pi_Q \rangle[X]$ with $X := \{\langle s, t, u\rangle \mid \langle s, t\rangle \in R, \langle t, u\rangle \in Q\}$. Since X is a weak pullback of π_T^R and π_T^Q by Lemma 2.6.25, we know that X is a bisimulation of (R, r) and (Q, q), with r and q as the dynamics of the corresponding $\boldsymbol{F}$-coalgebras. $\pi_S \circ \pi_R : X \to S$ and $\pi_U \circ \pi_Q : X \to U$ are morphisms; thus $\langle \pi_S \circ \pi_R, \pi_U \circ \pi_Q \rangle[X]$ is a bisimulation, since X is a weak pullback. Thus the assertion follows from Lemma 2.6.24. $\dashv$

The proof shows in which way the existence of the morphism $P \to \boldsymbol{F}P$ is used for achieving the desired properties.

Bisimulation equivalence

Bisimulations on a single coalgebra may have an additional structure, viz., they may be equivalence relations as well. Accordingly, we call these bisimulations *bisimulation equivalences*. Hence given a coalgebra (S, f), a bisimulation equivalence α for (S, f) is a bisimulation for (S, f) which is also an equivalence relation. While bisimulations carry properties which are concerned with the coalgebraic structure, an equivalence relation is purely related to the set structure. It is, however, fairly natural to ask in view of the properties which we did explore so far (Lemma 2.6.21, Proposition 2.6.28) whether or not we can take a bisimulation and turn it into an equivalence relation, or at least do so under favorable conditions on functor $\boldsymbol{F}$. We will deal with this question and some of its cousins now.

Observe first that the factor space of a bisimulation equivalence can be turned into a coalgebra.

Lemma 2.6.29 *Let (S, f) be an $\boldsymbol{F}$-coalgebra and α be a bisimulation equivalence on (S, f). Then there exists a unique dynamics $\alpha_R : S/\alpha \to \boldsymbol{F}(S/\alpha)$ with $\boldsymbol{F}(\eta_\alpha) \circ f = \alpha_R \circ \eta_\alpha$.*

Proof Because α is in particular a bisimulation, we know that there exists by Theorem 2.6.16 a dynamics $\rho : \alpha \to \boldsymbol{F}(\alpha)$ rendering this diagram commutative:

$$\begin{tikzcd}
S \arrow[d, "f"'] & \alpha \arrow[l, "\pi_S^{(1)}"'] \arrow[r, "\pi_S^{(2)}"] \arrow[d, "\rho"] & S \arrow[d, "f"] \\
\boldsymbol{F}S & \boldsymbol{F}\alpha \arrow[l, "\boldsymbol{F}\pi_S^{(1)}"] \arrow[r, "\boldsymbol{F}\pi_S^{(2)}"'] & \boldsymbol{F}S
\end{tikzcd}$$

The obvious choice for the dynamics α_R would be to define $\alpha_R([s]_\alpha) :=$ $(\boldsymbol{F}(\eta_\alpha) \circ f)(s)$, but this is only possible if we know that the map is well Outline defined, so we have to check whether $(\boldsymbol{F}(\eta_\alpha) \circ f)(s_1) = (\boldsymbol{F}(\eta_\alpha) \circ f)(s_2)$ holds, whenever $s_1 \ \alpha \ s_2$.

But this holds indeed, for $s_1 \ \alpha \ s_2$ means $\langle s_1, s_2\rangle \in \alpha$, so that $f(s_1) =$ $f(\pi_S^{(1)}(s_1, s_2)) = (\boldsymbol{F}(\pi_S^{(1)}) \circ \rho)(s_1, s_2)$, similarly for $f(s_2)$. Because α is an equivalence relation, we have $\eta_\alpha \circ \pi_S^{(1)} = \eta_\alpha \circ \pi_S^{(2)}$. Thus

$$\begin{aligned}
\boldsymbol{F}(\eta_\alpha)(f(s_1)) &= \big(\boldsymbol{F}(\eta_\alpha \circ \pi_S^{(1)}) \circ \rho\big)(s_1, s_2) \\
&= \big(\boldsymbol{F}(\eta_\alpha \circ \pi_S^{(2)}) \circ \rho\big)(s_1, s_2) \\
&= \boldsymbol{F}(\eta_\alpha)(f(s_2))
\end{aligned}$$

This means that α_R is in fact well defined and that η_α is a morphism. Hence the dynamics α_R exists and renders η_α a morphism.

Now assume that $\beta_R : S/\alpha \to \boldsymbol{F}(S/\alpha)$ satisfies also $\boldsymbol{F}(\eta_\alpha) \circ f =$ $\beta_R \circ \eta_\alpha$. But then $\beta_R \circ \eta_\alpha = \boldsymbol{F}(\eta_\alpha) \circ f = \alpha_R \circ \eta_\alpha$, and, since η_α is onto, it is an epi, so that we may conclude $\beta_R = \alpha_R$. Hence α_R is uniquely determined. $\dashv$

Bisimulations can be transported along morphisms, if the functor preserves weak pullbacks.

Proposition 2.6.30 *Assume that $\boldsymbol{F}$ preserves weak pullbacks, and let $r : (S, f) \to (T, g)$ be a morphism. Then:*

1. *If R is a bisimulation on (S, f), then $(r \times r)\big[R\big] = \{\langle r(s), r(s')\rangle \mid \langle s, s'\rangle \in R\}$ is a bisimulation on (T, g).*
2. *If Q is a bisimulation on (T, g), then $(r \times r)^{-1}\big[Q\big] = \{\langle s, s'\rangle \mid \langle r(r), r(s')\rangle \in Q\}$ is a bisimulation on (S, f).*

Proof 0. Note that $\mathrm{graph}(r)$ is a bisimulation by Corollary 2.6.22, because r is a morphism.

1. We claim that

$$(r \times r)\bigl[R\bigr] = \bigl(\mathrm{graph}(r)\bigr)^{-1} \circ R \circ \mathrm{graph}(r)$$

holds. Granted that, we can apply Proposition 2.6.28 together with Lemma 2.6.21 for establishing the first property. But $\langle t, t' \rangle \in (r \times r)\bigl[R\bigr]$ iff we can find $\langle s, s' \rangle \in R$ with $\langle t, t' \rangle = \langle r(s), r(s') \rangle$; hence $\langle r(s), s \rangle \in \mathrm{graph}(r)^{-1}$, $\langle s, s' \rangle \in R$ and $\langle s', r(s') \rangle \in \mathrm{graph}(r)$, hence iff $\langle t, t' \rangle \in \mathrm{graph}(r)^{-1} \circ R \circ \mathrm{graph}(r)$. Thus the equality holds indeed.

2. Similarly, we show that $(r \times r)^{-1}\bigl[Q\bigr] = \mathrm{graph}(r) \circ R \circ \mathrm{graph}(r)^{-1}$. This is left to the reader. ⊣

For investigating further structural properties, we need:

Lemma 2.6.31 *If (S, f) and (T, g) are $\boldsymbol{F}$-coalgebras, then there exists a unique coalgebraic structure on $S + T$ such that the injections i_S and i_T are morphisms.*

Proof We have to find a morphism $S + T \to \boldsymbol{F}(S + T)$ such that this diagram is commutative

$$\begin{tikzcd}
S \arrow[r, "i_S"] \arrow[d, "f"'] & S + T \arrow[d, dashed] & T \arrow[l, "i_T"'] \arrow[d, "g"] \\
\boldsymbol{F}S \arrow[r, "\boldsymbol{F}i_S"'] & \boldsymbol{F}(S + T) & \boldsymbol{F}T \arrow[l, "\boldsymbol{F}i_T"]
\end{tikzcd}$$

Because $\boldsymbol{F}(i_S) \circ f : S \to \boldsymbol{F}(S + T)$ and $\boldsymbol{F}(i_T) \circ g : T \to \boldsymbol{F}(S + T)$ are morphisms, there exists a unique morphism $h : S + T \to \boldsymbol{F}(S + T)$ with $h \circ i_S = \boldsymbol{F}(i_S) \circ f$ and $h \circ i_T = \boldsymbol{F}(i_T) \circ g$. Thus $(S + T, h)$ is a coalgebra, and i_S and i_T are morphisms. ⊣

The attempt to establish a comparable property for the product could not work with the universal property for products, as a look at the diagram for the universal property of the product will show.

We obtain as a consequence that bisimulations are closed under finite unions.

Lemma 2.6.32 *Let (S, f) and (T, g) be coalgebras with bisimulations R_1 and R_2. Then $R_1 \cup R_2$ is a bisimulation.*

Proof 1. We can find morphisms $r_i : R_i \to \boldsymbol{F}R_i$ for $i = 1, 2$ rendering the corresponding bisimulation diagrams commutative. Then $R_1 + R_2$

is an $\boldsymbol{F}$-coalgebra with

$$
\begin{tikzcd}
R_1 \arrow[r, "j_1"] \arrow[d, "r_1"'] & R_1 + R_2 \arrow[d, "r"'] & R_2 \arrow[l, "j_2"'] \arrow[d, "r_2"] \\
\boldsymbol{F}R_1 \arrow[r, "\boldsymbol{F}j_1"'] & \boldsymbol{F}(R_1 + R_2) & \boldsymbol{F}R_2 \arrow[l, "\boldsymbol{F}j_2"]
\end{tikzcd}
$$

as commuting diagram, where $j_i : R_i \to R_1 + R_2$ is the respective embedding, $i = 1, 2$.

2. We claim that the projections $\pi'_S : R_1 + R_2 \to S$ and $\pi'_T : R_1 + R_2 \to T$ are morphisms. We establish this property only for π'_S. First note that $\pi'_S \circ j_1 = \pi_S^{R_1}$, so that we have $f \circ \pi'_S \circ j_1 = \boldsymbol{F}(\pi'_S) \circ \boldsymbol{F}(j_1) \circ r_1 = \boldsymbol{F}(\pi'_S) \circ r \circ j_1$, similarly, $f \circ \pi'_S \circ j_2 = \boldsymbol{F}(\pi'_S) \circ r \circ j_2$. Thus we may conclude that $f \circ \pi'_S = \boldsymbol{F}(\pi'_S) \circ r$, so that indeed $\pi'_S : R_1 + R_2 \to S$ is a morphism.

3. Since $R_1 + R_2$ is a coalgebra, we know from Lemma 2.6.23 that $\langle \pi'_S, \pi'_T \rangle [R_1 + R_2]$ is a bisimulation. But this equals $R_1 \cup R_2$. $\dashv$

We briefly explore lattice properties for bisimulations on a coalgebra. For this, we investigate the union of an arbitrary family of bisimulations. Looking back at the union of two bisimulations, we used their sum as an intermediate construction. A more general consideration requires the sum of an arbitrary family. The following definition describes the coproduct as a specific form of a colimit; see Definition 2.3.32.

Definition 2.6.33 *Let $(s_k)_{k \in I}$ be an arbitrary nonempty family of objects on a category $\mathbf{K}$. The object s together with morphisms $i_k : s_k \to s$ is called the* coproduct *of $(s_k)_{k \in I}$ iff given morphisms $j_k : s_k \to t$ for an object t there exists a unique morphism $j : s \to t$ with $j_k = j \circ i_k$ for all $k \in I$. s is denoted as $\sum_{k \in I} s_k$.*

Taking $I = \{1, 2\}$, one sees that the coproduct of two objects is in fact a special case of the coproduct just defined. The following diagram gives a general idea:

$$
\begin{tikzcd}
\ldots & s_{r_1} \arrow[dr, "i_{r_1}"] \arrow[ddr, "j_{r_1}"'] & \ldots & s_{r_k} \arrow[dl, "i_{r_k}"'] \arrow[ddl, "j_{r_k}"] & \ldots \\
& & s \arrow[d, dashed, "j"'] & & \\
& & t & &
\end{tikzcd}
$$

The coproduct is uniquely determined up to isomorphisms.

Example 2.6.34 Consider the category ***Set*** of sets with maps as morphisms, and let $(S_k)_{k\in I}$ be a family of sets. Then

$$S := \bigcup_{k\in I}\{\langle s,k\rangle \mid s \in S_k\}$$

is a coproduct. In fact, $i_k : s \mapsto \langle s,k\rangle$ maps S_k to S, and if $j_k : S_k \to T$, put $j : S \to T$ with $j(s,k) := j_k(s)$; then $j_k = j \circ i_k$ for all k. ✌

We put this new machinery to use right away, returning to our scenario given by functor $\boldsymbol{F}$.

Proposition 2.6.35 *Assume that* $\boldsymbol{F}$ *preserves weak pullbacks. Let* $(R_k)_{k\in I}$ *be a family of bisimulations for coalgebras* (S,f) *and* (T,g)*. Then* $\bigcup_{k\in I} R_k$ *is a bisimulation for these coalgebras.*

Proof 1. Given $k \in I$, let $r_k : R_k \to \boldsymbol{F}R_k$ be the morphism on R_k such that $\pi_{S,k} : (S,f) \to (R_k,r_k)$ and $\pi_{T,k} : (T,g) \to (R_k,r_k)$ are morphisms for the coalgebras involved. Then there exists a unique coalgebra structure r on $\sum_{k\in I} R_k$ such that $i_\ell : (R_\ell, r_\ell) \to (\sum_{k\in I} R_k, r)$ is a coalgebra structure for all $\ell \in I$. This is shown exactly through the same argument as in the proof of Lemma 2.6.32 (*mutatis mutandis*: replace the coproduct of two bisimulations by the general coproduct).

2. The projections $\pi'_S : \sum_{k\in I} R_k \to S$ and $\pi'_T : \sum_{k\in I} R_k \to T$ are morphisms, and one shows exactly as in the proof of Lemma 2.6.32 that

$$\bigcup_{k\in I} R_k = \langle \pi'_S, \pi'_T\rangle[\sum_{k\in I} R_k].$$

An application of Lemma 2.6.23 now establishes the claim. $\dashv$

This is applied to an investigation of the lattice structure on the set of all bisimulations between coalgebras.

Proposition 2.6.36 *Assume that* $\boldsymbol{F}$ *preserves weak pullbacks. Let* $(R_k)_{k\in I}$ *be a nonempty family of bisimulations for coalgebras* (S,f) *and* (T,g)*. Then:*

1. *There exists a smallest bisimulation* R^* *with* $R_k \subseteq R^*$ *for all* k.
2. *There exists a largest bisimulation* R_* *with* $R_k \supseteq R_*$ *for all* k.

Proof 1. We claim that $R^* = \bigcup_{k\in I} R_k$. It is clear that $R_k \subseteq R^*$ for all $k \in I$. If R' is a bisimulation on (S,f) and (T,g) with $R_k \subseteq R'$ for all

k, then $\bigcup_k R_k \subseteq R'$; thus $R^* \subseteq R'$. In addition, R^* is a bisimulation by Proposition 2.6.35. This establishes part 1.

2. Put

$$\mathcal{R} := \{R \mid R \text{ is a bisimulation for } (S, f) \text{ and } (T, g) \\ \text{with } R \subseteq R_k \text{ for all } k\}$$

If $\mathcal{R} = \emptyset$, we put $R_* := \emptyset$, so we may assume that $\mathcal{R} \neq \emptyset$. Put $R_* := \bigcup \mathcal{R}$. By Proposition 2.6.35, this is a bisimulation for (S, f) and (T, g) with $R_k \subseteq R_*$ for all k. Assume that R' is a bisimulation for (S, f) and (T, g) with $R' \subseteq R_k$ for all k; then $R' \in \mathcal{R}$; hence $R' \subseteq R_*$, so R_* is the largest one. This settles part 2. $\dashv$

Looking a bit harder at bisimulations for (S, f) alone, we find that the largest bisimulation is actually an equivalence relation. But we have to make sure first that a largest bisimulation exists at all.

Proposition 2.6.37 *If functor $\mathbf{F}$ preserves weak pullbacks, then there exists a largest bisimulation R^* on coalgebra (S, f). R^* is an equivalence relation.*

Proof 1. Let

$$\mathcal{R} := \{R \mid R \text{ is a bisimulation on } (S, f)\}.$$

Then $\Delta_S \in \mathcal{R}$; hence $\mathcal{R} \neq \emptyset$. We know from Lemma 2.6.21 that $R \in \mathcal{R}$ entails $R^{-1} \in \mathcal{R}$, and from Proposition 2.6.35, we infer that $R^* := \bigcup \mathcal{R} \in \mathcal{R}$. Hence R^* is a bisimulation on (S, f).

2. We use the characterization of the infimum as the supremum of the lower bounds. R^* is even an equivalence relation. Idea

- Since $\Delta_S \in \mathcal{R}$, we know that $\Delta_S \subseteq R^*$; thus R^* is reflexive.
- Because $R^* \in \mathcal{R}$, we conclude that $(R^*)^{-1} \in \mathcal{R}$; thus $(R^*)^{-1} \subseteq R^*$. Hence R^* is symmetric.
- Since $R^* \in \mathcal{R}$, we conclude from Proposition 2.6.28 that $R^* \circ R^* \in \mathcal{R}$; hence $R^* \circ R^* \subseteq R^*$. This means that R^* is transitive.

$\dashv$

This has an interesting consequence. Given a bisimulation equivalence on a coalgebra, we do not only find a larger one which contains it, but we can also find a morphism between the corresponding factor spaces. This is what I mean:

Corollary 2.6.38 *Assume that functor $\boldsymbol{F}$ preserves weak pullbacks and that α is a bisimulation equivalence on (S, f); then there exists a unique morphism $\vartheta_\alpha : (S/\alpha, f_\alpha) \to (S/R^*, f_{R^*})$, where $f_\alpha : S/\alpha \to \boldsymbol{F}(S/\alpha)$ and $f_{R^*} : S/R^* \to \boldsymbol{F}(S/R^*)$ are the induced dynamics.*

Proof 0. The dynamics $f_\alpha : S/\alpha \to \boldsymbol{F}(S/\alpha)$ and $f_{R^*} : S/R^* \to \boldsymbol{F}(S/R^*)$ exist by the definition of a bisimulation.

1. Define

$$\vartheta([s]_\alpha) := [s]_{R^*}$$

for $s \in S$. This is well defined. In fact, if $s\ \alpha\ s'$, we conclude by the maximality of R^* that $s\ R^*\ s'$, so $[s]_\alpha = [s']_\alpha$ implies $[s]_{R^*} = [s']_{R^*}$.

2. We claim that ϑ_α is a morphism, hence that the right-hand side of this diagram commutes; the left-hand side of the diagram is just for nostalgia.

$$\begin{tikzcd}
S \arrow[r, "\eta_\alpha"] \arrow[d, "f"'] & S/\alpha \arrow[r, "\vartheta_\alpha"] \arrow[d, "f_\alpha"'] & S/R^* \arrow[d, "f_{R^*}"] \\
\boldsymbol{F}S \arrow[r, "\boldsymbol{F}\eta_\alpha"'] & \boldsymbol{F}(S/\alpha) \arrow[r, "\boldsymbol{F}\vartheta_\alpha"'] & \boldsymbol{F}(S/R^*)
\end{tikzcd}$$

Now $\vartheta_\alpha \circ \eta_\alpha = \eta_{R^*}$, and the outer diagram commutes. The left diagram commutes because $\eta_\alpha : (S, f) \to S/f_\alpha$ is a morphism; moreover, η_α is a surjective map. Hence the claim follows from Lemma 2.1.32, so that ϑ_α is a morphism indeed.

3. If ϑ'_α is another morphism with these properties, then we have $\vartheta'_\alpha \circ \eta_\alpha = \eta_{R^*} = \vartheta_\alpha \circ \eta_\alpha$, and since η_α is surjective, it is an epi by Proposition 2.1.23, which implies $\vartheta_\alpha = \vartheta'_\alpha$. $\dashv$

This is all very well, but where do we get bisimulation equivalences from? If we cannot find examples for them, the efforts just spent may run dry. Fortunately, we are provided with ample bisimulation equivalences through coalgebra morphisms, specifically through their kernel (for a definition, see page 124). It will turn out that all such equivalences can be generated in this way.

Proposition 2.6.39 *Assume that* $\boldsymbol{F}$ *preserves weak pullbacks and that* $\varphi : (S, f) \to (T, g)$ *is a coalgebra morphism. Then* $\ker(\varphi)$ *is a bisimulation equivalence on* (S, f)*. Conversely, if* α *is a bisimulation equivalence on* (S, f)*, then there exists a coalgebra* (T, g) *and a coalgebra morphism* $\varphi : (S, f) \to (T, g)$ *with* $\alpha = \ker(\varphi)$.

Proof 1. We know that $\ker(\varphi)$ is an equivalence relation; since $\ker(\varphi) = \mathrm{graph}(\varphi) \circ \mathrm{graph}(\varphi)^{-1}$, we conclude from Corollary 2.6.22 that $\ker(\varphi)$ is a bisimulation.

2. Let α be a bisimulation equivalence on (S, f); then the factor map $\eta_\alpha : (S, f) \to (S/\alpha, f_\alpha)$ is a morphism by Lemma 2.6.29, and $\ker(\eta_\alpha) = \{\langle s, s'\rangle \mid [s]_\alpha = [s']_\alpha\} = \alpha$. ⊣

We know what morphisms are, and usually morphisms are studied also in the context of congruences as those equivalence relations which respect the underlying structure. This is what we will do next.

2.6.2 Congruences

Bisimulations compare two systems with each other, while a congruence permits to talk about elements in a coalgebra which behave in a similar manner. Let us have a look at Abelian groups. An equivalence relation α on an Abelian group $(G, +)$ is a congruence iff $g \; \alpha \; h$ and $g' \; \alpha \; h'$ together imply $(g + g') \; \alpha \; (h + h')$. This means that α is compatible with the group structure; an equivalent formulation says that there exists a group structure on G/α such that the factor map $\eta_\alpha : G \to G/\alpha$ is a group morphism. Thus the factor map is the harbinger of the good news. We translate this observation now into the language of coalgebras.

Definition 2.6.40 *Let* (S, f) *be an* $\boldsymbol{F}$*-coalgebra for the endofunctor* $\boldsymbol{F}$ *on the category* **Set** *of sets. An equivalence relation* α *on* S *is called an* $\boldsymbol{F}$-congruence *iff there exists a coalgebra structure* f_α *on* S/α *such that* $\eta_\alpha : (S, f) \to (S/\alpha, f_\alpha)$ *is a coalgebra morphism.*

Thus we want that this diagram

$$\begin{tikzcd} S \arrow[r, "\eta_\alpha"] \arrow[d, "f"'] & S/\alpha \arrow[d, "f_\alpha"] \\ \boldsymbol{F}S \arrow[r, "\boldsymbol{F}\eta_\alpha"'] & \boldsymbol{F}(S/\alpha) \end{tikzcd}$$

is commutative, so that we have

$$f_\alpha([s]_\alpha) = (\boldsymbol{F}\eta_\alpha)(f(s))$$

for each $s \in S$. A brief look at Lemma 2.6.29 shows that bisimulation equivalences are congruences, and we see from Proposition 2.6.39 that the kernels of coalgebra morphisms are congruences, provided the functor $\boldsymbol{F}$ preserves weak pullbacks.

Hence congruences and bisimulations on a coalgebra are actually very closely related. They are, however, not the same, because we have:

Proposition 2.6.41 *Let $\varphi : (S, f) \to (T, g)$ be a morphism for the $\boldsymbol{F}$-coalgebras (S, f) and (T, g). Assume that* $\ker(\boldsymbol{F}\varphi) \subseteq \ker\left(\boldsymbol{F}\eta_{\ker(\varphi)}\right)$. *Then* $\ker(\varphi)$ *is a congruence for* (S, f).

Proof Define $f_{\ker(\varphi)}([s]_{\ker(\varphi)}) := \boldsymbol{F}(\eta_{\ker(\varphi)})(f(s))$ for $s \in S$. Then $f_{\ker(\varphi)} : S/\ker(\varphi) \to \boldsymbol{F}(S/\ker(\varphi))$ is well defined. In fact, assume that $[s]_{\ker(\varphi)} = [s']_{\ker(\varphi)}$, then $g(\varphi(s)) = g(\varphi(s'))$, so that $(\boldsymbol{F}\varphi)(f(s)) = (\boldsymbol{F}\varphi)(f(s'))$, consequently $\langle f(s), f(s')\rangle \in \ker(\boldsymbol{F}\varphi)$. By assumption, $(\boldsymbol{F}\eta_{\ker(\varphi)})(f(s)) = (\boldsymbol{F}\eta_{\ker(\varphi)})(f(s'))$, so that $f_{\ker(\varphi)}([s]_{\ker(\varphi)}) = f_{\ker(\varphi)}([s']_{\ker(\varphi)})$. It is clear that η_α is a coalgebra morphism. $\dashv$

The definition of a congruence is not tied to functors which operate on the category of sets. The next example leaves this category and considers the category of measurable spaces, introduced in Example 2.1.12. The subprobability functor $\mathbb{S}$ from Example 2.3.12 is an endofunctor on ***Meas***, and we know that the coalgebras for this functor are just the subprobabilistic transition kernels $K : (S, \mathcal{A}) \to (S, \mathcal{A})$; see Example 2.6.7.

Final measurable map

Definition 2.6.42 *A measurable map $f : (S, \mathcal{A}) \to (T, \mathcal{B})$ measurable spaces $(S, \mathcal{A})$ and $(T, \mathcal{B})$ is called* final *iff $\mathcal{B}$ is the largest σ-algebra on T which renders f measurable.*

Thus, if f is onto, we conclude from $f^{-1}\left[B\right] \in \mathcal{A}$ that $B \in \mathcal{B}$, because f^{-1} is injective. Given an equivalence relation α on S, we can make the factor space S/α a measurable space by endowing it with the final σ-algebra $\mathcal{A}/\alpha$ with respect to η_α; compare Exercise 2.25.

This, then, is the definition of a morphism for coalgebras for the Giry functor (see Example 2.4.8).

Definition 2.6.43 *Let $(S, \mathcal{A}, K)$ and $(T, \mathcal{B}, L)$ be coalgebras for the subprobability Functor; then $\varphi : (S, \mathcal{A}, K) \to (T, \mathcal{B}, L)$ is a* coalgebra morphism *iff* $\varphi : (S, \mathcal{A}) \to (T, \mathcal{B})$ *is a measurable map such that this diagram commutes:*

$$\begin{tikzcd} S \arrow[r, "\varphi"] \arrow[d, "K"'] & T \arrow[d, "L"] \\ \mathbb{S}(S, \mathcal{A}) \arrow[r, "\mathbb{S}\varphi"'] & \mathbb{S}(T, \mathcal{B}) \end{tikzcd}$$

Thus we have

$$L(\varphi(s))(B) = \mathbb{S}(\varphi)(K(s))(B) = K(s)(\varphi^{-1}\left[B\right])$$

for each $s \in S$ and for each measurable set $B \in \mathcal{B}$. We will investigate the kernel of a morphism now in order to obtain a result similar to the one reported in Proposition 2.6.41. The crucial property in that development has been the comparison of the kernel $\ker(\boldsymbol{F}\varphi)$ with $\ker\left(\boldsymbol{F}\eta_{\ker(\varphi)}\right)$. We will concentrate on this property now.

Call a morphism φ *strong* iff φ is surjective and final. Now fix a strong morphism $\varphi : K \to L$. A measurable subset $A \in \mathcal{A}$ is called φ-*invariant* iff $a \in A$ and $\varphi(a) = \varphi(a')$ together imply $a' \in A$, so that $A \in \mathcal{A}$ is φ invariant iff A is the union of $\ker(\varphi)$-equivalence classes. Strong morphism

We obtain this characterization of φ-invariant sets (note that this is an intrinsic property of the map proper):

Lemma 2.6.44 *Let $\varphi : (S, \mathcal{A}) \to (T, \mathcal{B})$ be a strong morphism, and define*

$$\Sigma_\varphi := \{A \in \mathcal{A} \mid A \text{ is } \varphi\text{-invariant}\}.$$

Then:

1. *Σ_φ is a σ-algebra.*
2. *Σ_φ is isomorphic to $\{\varphi^{-1}\left[B\right] \mid B \in \mathcal{B}\}$ as a Boolean σ-algebra.*

Proof 1. Clearly, both $\emptyset$ and S are φ-invariant, and the complement of an invariant set is invariant again. Invariant sets are closed under countable unions. Hence Σ_φ is a σ-algebra.

2. Given $B \in \mathcal{B}$, it is clear that $\varphi^{-1}\left[B\right]$ is φ-invariant; since the latter is also a measurable subset of S, we conclude that $\{\varphi^{-1}\left[B\right] \mid B \in \mathcal{B}\} \subseteq \Sigma_\varphi$. Now let $A \in \Sigma_\varphi$; we claim that $A = \varphi^{-1}\left[\varphi\left[A\right]\right]$. In fact, since $\varphi(a) \in \varphi\left[A\right]$ for $a \in A$, the inclusion $A \subseteq \varphi^{-1}\left[\varphi\left[A\right]\right]$ is trivial. Let $a \in \varphi^{-1}\left[\varphi\left[A\right]\right]$, so that there exists $a' \in A$

with $\varphi(a) = \varphi(a')$. Since A is φ-invariant, we conclude $a \in A$, establishing the other inclusion. Because φ is final and surjective, we infer from this representation that $\varphi\big[A\big] \in \mathcal{B}$, whenever $A \in \Sigma_\varphi$, and that $\varphi^{-1} : \mathcal{B} \to \Sigma_\varphi$ is surjective. Since φ is surjective and φ^{-1} is injective, hence φ^{-1} yields a bijection. The latter map is compatible with the operations of a Boolean σ-algebra, so it is an isomorphism. $\dashv$

This result is usually the unbeknownst and deeper reason why the construction of various kinds of bisimulations for Markov transition systems work. In our context, it helps in establishing the crucial property for kernels.

Corollary 2.6.45 *Let $\varphi : K \to L$ be a strong morphism; then* $\ker(\mathbb{S}\varphi) \subseteq \ker\big(\mathbb{S}\eta_{\ker(\varphi)}\big)$.

Proof Let $\langle \mu, \mu' \rangle \in \ker(\mathbb{S}\varphi)$; thus $(\mathbb{S}\varphi)(\mu)(B) = (\mathbb{S}\varphi)(\mu')(B)$ for all $B \in \mathcal{B}$. Now let $C \in \mathcal{A}/\ker(\varphi)$, then $\eta^{-1}_{\ker(\varphi)}\big[C\big] \in \Sigma_\varphi$, so that there exists by Lemma 2.6.44 some $B \in \mathcal{B}$ such that $\eta^{-1}_{\ker(\varphi)}\big[C\big] = \varphi^{-1}\big[B\big]$. This is the central property. Hence

$$\begin{aligned}(\mathbb{S}\eta_{\ker(\varphi)})(\mu)(C) &= \mu(\eta^{-1}_{\ker(\varphi)}\big[C\big]) &&= \mu(\varphi^{-1}\big[B\big]) \\ &= (\mathbb{S}\varphi)(\mu)(B) &&= (\mathbb{S}\varphi)(\mu')(B) \\ &= (\mathbb{S}\eta_{\ker(\varphi)})(\mu')(C),\end{aligned}$$

so that $\langle \mu, \mu' \rangle \in \ker\big(\mathbb{S}\eta_{\ker(\varphi)}\big)$. $\dashv$

Now everything is in place to show that the kernel of a strong morphism is a congruence for the $\mathbb{S}$-coalgebra $(S, \mathcal{A}, K)$.

Proposition 2.6.46 *Let $\varphi : K \to L$ be a strong morphism for the $\mathbb{S}$-coalgebras $(S, \mathcal{A}, K)$ and $(T, \mathcal{B}, L)$. Then* $\ker(\varphi)$ *is a congruence for* $(S, \mathcal{A}, K)$.

Plan

Proof 0. We have to find a coalgebra structure on the measurable space $(S/\ker(\varphi), \mathcal{A}/\ker(\varphi))$ first; the candidate is obvious. After having established that this is possible indeed, we check the condition for a congruence. This invites an application of Proposition 2.6.41 through Corollary 2.6.45.

1. We want to define the coalgebra $K_{\ker(\varphi)}$ on $(S/\ker(\varphi), \mathcal{A}/\ker(\varphi))$ upon setting

$$\begin{aligned}K_{\ker(\varphi)}([s]_{\ker(\varphi)})(C) \quad &:= \quad (\mathbb{S}\eta_{\ker(\varphi)})(K(s))(C) \\ & \qquad \big(= K(s)(\eta^{-1}_{\ker(\varphi)}\big[C\big])\big)\end{aligned}$$

for $C \in \mathcal{A}/\ker(\varphi)$, but we have to be sure first that this is well defined. In fact, let $[s]_{\ker(\varphi)} = [s']_{\ker(\varphi)}$, which means $\varphi(s) = \varphi(s')$; hence $L(\varphi(s)) = L(\varphi(s'))$, so that $(\mathbb{S}\varphi)K(s) = (\mathbb{S}\varphi)K(s')$, because $\varphi : K \to L$ is a morphism. But the latter equality implies $\langle K(s), K(s')\rangle \in \ker(\mathbb{S}\varphi) \subseteq \ker\left(\mathbb{S}\eta_{\ker(\varphi)}\right)$, the inclusion holding by Corollary 2.6.45. Thus we conclude

$$(\mathbb{S}\eta_{\ker(\varphi)})(K(s)) = (\mathbb{S}\eta_{\ker(\varphi)})(K(s')),$$

so that $K_{\ker(\varphi)}$ is well defined indeed.

2. It is immediate that $C \mapsto K_{\ker(\varphi)}([s]_{\ker(\varphi)})(C)$ is a subprobability on $\mathcal{A}/\ker(\varphi)$ for fixed $s \in S$, so it remains to show that $t \mapsto K_{\ker(\varphi)}(t)(C)$ is a measurable map on the factor space $(S/\ker(\varphi), \mathcal{A}/\ker(\varphi))$. Let $q \in [0, 1]$, and consider for $C \in \mathcal{A}/\ker(\varphi)$ the set

$$G := \{t \in S/\ker(\varphi) \mid K_{\ker(\varphi)}(t)(C) < q\}.$$

We have to show that $G \in \mathcal{A}/\ker(\varphi)$. Because $C \in \mathcal{A}/\ker(\varphi)$, we know that $A := \eta^{-1}_{\ker(\varphi)}\left[C\right] \in \Sigma_\varphi$; hence it is sufficient to show that the set $H := \{s \in S \mid K(s)(A) < q\}$ is a member of Σ_f. Since K is the dynamics of a $\mathbb{S}$-coalgebra, we know that $H \in \mathcal{A}$, so it remains to show that H is φ-invariant. Because $A \in \Sigma_f$, we infer from Lemma 2.6.44 that $A = \varphi^{-1}\left[B\right]$ for some $B \in \mathcal{B}$. Now take $s \in H$ and assume $\varphi(s) = \varphi(s')$. Thus

$$\begin{aligned} K(s')(A) &= K(s')(\varphi^{-1}\left[B\right]) && = (\mathbb{S}\varphi)(K(s'))(B) && = L(\varphi(s'))(B) \\ &= L(\varphi(s))(B) && = K(s)(A) \\ &< q, \end{aligned}$$

so that $H \in \Sigma_\varphi$ indeed. Because $H = \eta^{-1}_{\ker(\varphi)}\left[G\right]$, it follows that $G \in \mathcal{A}/\ker(f)$, and we are done. $\dashv$

We will deal with coalgebras now when interpreting modal logics. This connection between modal logics and coalgebras is at first sight fairly surprising but becomes at second sight interesting because modal logics are firmly tied to transition systems, and we have seen that transition systems can be interpreted as coalgebras. So one wants to know what coalgebraic properties are reflected in the relational interpretation of modal logics, in particular the relation to the coalgebraic reading of bisimilarity is interesting and may promise new insights.

2.7 Modal Logics

This section will discuss modal logics and provide a closer look at the interface between models for this logics and coalgebras. Thus the topics of this section may be seen as an application and illustration of coalgebras.

We will define the language for the formulas of modal logics, first for the conventional logics which permits expressing sentences like "it is possible that formula φ holds" or "formula φ holds necessarily" and then for an extended version, allowing for modal operators that govern more than one formula. The interpretation through Kripke models is discussed, and it becomes clear that at least elementary elements of the language of categories are helpful in investigating these logics. For completeness, we also give the construction for the canonical model, displaying the elegant construction through the Lindenbaum Lemma.

It shows that coalgebras can be used directly in the interpretation of modal logics. We demonstrate that a set of predicate liftings define a modal logics, discuss briefly expressivity for these modal logics, and display an interpretation of `CTL*`, one of the basic logics for model checking, through coalgebras.

Φ

We fix a set Φ of *propositional letters*.

$\mathcal{L}(\Phi)$

Definition 2.7.1 *The* basic modal language $\mathcal{L}(\Phi)$ *over* Φ *is given by this grammar:*

$$\varphi ::= \bot \mid p \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \Diamond\varphi$$

with $p \in \Phi$.

We introduce additional operators:

$$\begin{aligned} \top &\text{ denotes } \neg\bot, \\ \varphi_1 \vee \varphi_2 &\text{ denotes } \neg(\neg\varphi_1 \wedge \neg\varphi_2), \\ \varphi_1 \rightarrow \varphi_2 &\text{ denotes } \neg\varphi_1 \vee \varphi_2, \\ \Box\varphi &\text{ denotes } \neg\Diamond\neg\varphi. \end{aligned}$$

The constant $\bot$ denotes falsehood, consequently, $\top = \neg\bot$ denotes truth, and negation $\neg$ as well as conjunction $\wedge$ should not come as a surprise; informally, $\Diamond\varphi$ means that it is possible that formula φ holds, while $\Box\varphi$ expresses that φ holds necessarily. Syntactically, this looks like propositional logic, extended by the modal operators $\Diamond$ and $\Box$.

Before we have a look at the semantics of modal logic, we indicate that this logic is syntactically sometimes a bit too restricted; after all, the modal operators operate only on one argument at a time. The extension we want should offer modal operators with more arguments. For this, we introduce the notion of a *modal similarity type* $\mathfrak{t} = (O, \rho)$, which is a set O of operators; each operator $\Delta \in O$ has an arity $\rho(\Delta) \in \mathbb{N}_0$. Note that $\rho(\Delta) = 0$ is not excluded; these modal constants will not play a distinguished rôle; however, they are sometimes nice to have. (O, ρ)

Clearly, the set $\{\Diamond\}$ together with $\rho(\Diamond) = 1$ is an example for such a modal similarity type.

Definition 2.7.2 *Given a modal similarity type $\mathfrak{t} = (O, \rho)$ and the set Φ of propositional letters, the* extended modal language $\mathcal{L}(\mathfrak{t}, \Phi)$ *is given by this grammar:*

$$\varphi ::= \bot \mid p \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \Delta(\varphi_1, \ldots, \varphi_k)$$

with $p \in \Phi$ and $\Delta \in O$ such[2] *that $\rho(\Delta) = k$.*

We also introduce for the general case operators which are called *nablas*. The nabla ∇ of Δ is defined through $(\Delta \in O, \rho(\Delta) = k)$

$$\nabla(\varphi_1, \ldots, \varphi_k) := \neg\Delta(\neg\varphi_1, \ldots, \neg\varphi_k)$$

$\Box$ is the nabla of $\Diamond$, so ∇ generalizes a well-known operation.

It is time to have a look at some examples.

Example 2.7.3 Let $O = \{\boldsymbol{F}, \boldsymbol{P}\}$ with $\rho(\boldsymbol{F}) = \rho(\boldsymbol{P}) = 1$; the operator $\boldsymbol{F}$ looks into the future and $\boldsymbol{P}$ into the past. This may be useful, e.g., when you are traversing a tree and are visiting an inner node. The future may then look at all nodes in its subtree, the past at all nodes on a path from the root to this tree.

Then $\mathfrak{t}_{Fut} := (O, \rho)$ is a modal similarity type. If φ is a formula in $\mathcal{L}(\mathfrak{t}_{Fut}, \Phi)$, formula $\boldsymbol{F}\varphi$ is true iff φ will hold in the future, and $\boldsymbol{P}\varphi$ is

[2] In this section, Δ does not denote the diagonal of a set (as elsewhere in this book); we could have used another letter, but the symmetry of Δ and ∇ is too good to be missed.

true iff φ did hold in the past. The nablas are defined as

$$\begin{aligned} \boldsymbol{G}\varphi &:= \neg\boldsymbol{F}\neg\varphi && (\varphi \text{ will always be the case}) \\ \boldsymbol{H}\varphi &:= \neg\boldsymbol{P}\neg\varphi && (\varphi \text{ has always been the case}). \end{aligned}$$

Look at some formulas:

$\boldsymbol{P}\varphi \to \boldsymbol{GP}\varphi$: If something has happened, it will always have happened.

$\boldsymbol{F}\varphi \to \boldsymbol{FF}\varphi$: If φ will be true in the future, then it will be true in the future that φ will be true.

$\boldsymbol{GF}\varphi \to \boldsymbol{FG}\varphi$: If φ will be true in the future, then it will at some point be always true.

✌

The next example deals with a simple model for sequential programs.

Example 2.7.4 Take Ψ as a set of atomic programs (think of elements of Ψ as executable program components). The set of programs is defined through this grammar:

$$t ::= \psi \mid t_1 \cup t_2 \mid t_1; t_2 \mid t^* \mid \varphi?$$

with $\psi \in \Psi$ and φ a formula of the underlying modal logic.

Here $t_1 \cup t_2$ denotes the nondeterministic choice between programs t_1 and t_2, $t_1; t_2$ is the sequential execution of t_1 and t_2 in that order, and t^* is iteration of program t a finite number of times (including zero). The program $\varphi?$ tests whether or not formula φ holds; $\varphi?$ serves as a guard: $(\varphi?; t_1) \cup (\neg\varphi?; t_2)$ tests whether φ holds, if it does t_1 is executed; otherwise, t_2 is. So the informal meaning of $\langle t\rangle\varphi$ is that formula φ holds after program t is executed (we use here and later an expression like $\langle t\rangle\varphi$ rather than the functional notation or just juxtaposition).

So, formally we have the modal similarity type $\mathfrak{t}_{PDL} := (O, \rho)$ with $O := \{\langle t\rangle \mid t \text{ is a program}\}$. This logic is known as PDL—propositional dynamic logic. ✌

The next example deals with games and a syntax very similar to the one just explored for PDL.

Example 2.7.5 We introduce two players, Angel and Demon, playing against each other, taking turns. So Angel starts, then Demon makes the next move, then Angel replies, etc.

For modeling game logic, we assume that we have a set Γ of simple games; the syntax for games looks like this:

$$g ::= \gamma \mid g_1 \cup g_2 \mid g_1 \cap g_2 \mid g_1 ; g_2 \mid g^d \mid g^* \mid g^\times \mid \varphi ?$$

with $\gamma \in \Gamma$ and φ a formula of the underlying logic. The informal interpretation of $g_1 \cup g_2$, $g_1 ; g_2$, g^* and φ? are as in PDL (Example 2.7.4) but as actions of player Angel. The actions of player Demon are indicated by:

$g_1 \cap g_2$: Demon chooses between games g_1 and g_2; this is called *demonic choice* (in contrast to *angelic choice* $g_1 \cup g_2$).

$g^\times$: Demon decides to play game g a finite number of times (including not at all).

g^d: Angel and Demon change places.

Again, we indicate through $\langle g \rangle \varphi$ that formula φ holds after game g. We obtain the similarity type $\mathfrak{t}_{GL} := (O, \rho)$ with $O := \{\langle g \rangle \mid g \text{ is a game}\}$ and $\rho = 1$. The corresponding logic is called *game logic* ✌

Another example is given by arrow logic. Assume that you have arrows in the plane; you can compose them, i.e., place the beginning of one arrow at the end of the first one, and you can reverse them. Finally, you can leave them alone, i.e., do nothing with an arrow.

Example 2.7.6 The set O of operators for arrow logic is given by $\{\circ, \otimes, \texttt{skip}\}$ with $\rho(\circ) = 2$, $\rho(\otimes) = 1$ and $\rho(\texttt{skip}) = 0$. The arrow composed from arrows a_1 and a_2 is arrow $a_1 \circ a_2$, $\otimes a_1$ is the reversed arrow a_1, and `skip` does nothing. ✌

2.7.1 Frames and Models

For interpreting the basic modal language, we introduce frames. A frame models transitions, which are at the very heart of modal logics. Let us have a brief look at a modal formula like $\Box p$ for some propositional letter $p \in \Phi$. This formula models "p always holds," which implies a transition from the current state to another one, in which p is assumed to hold always; without a transition, we would not have to think whether p always holds—it would just hold or not. Hence we need to have transitions at our disposal, thus a transition system, as in Example 2.1.9. In the current context, we take the disguise of a transition system as a relation. All this is captured in the notion of a frame.

Definition 2.7.7 *A* Kripke frame $\mathfrak{F} := (W, R)$ *for the basic modal language is a set* $W \neq \emptyset$ *of states together with a relation* $R \subseteq W \times W$. W *is sometimes called the* set of worlds*,* R *the* accessibility relation*.*

The accessibility relation of a Kripke frame does not yet carry enough information about the meaning of a modal formula, since the propositional letters are not captured by the frame. This is the case, however, in a Kripke model.

Definition 2.7.8 *A* Kripke model *(or simply a* model*, for the time being)* $\mathfrak{M} = (W, R, V)$ *for the basic modal language consists of a Kripke frame* (W, R) *together with a map* $V : \Phi \to \mathcal{P}(W)$*.*

So, roughly speaking, the frame part of a Kripke model caters to the propositional and the modal part of the logic, whereas the map V takes care of the propositional letters. This will enable us to define the meaning of the formulas for the basic modal language. We state the conditions under which a formula φ is true in a world $w \in W$; this is
$\mathfrak{M}, w \models \varphi$ expressed through $\mathfrak{M}, w \models \varphi$; note that this will depend on the model $\mathfrak{M}$; hence, we incorporate it usually into the notation. Here we go.

$\mathfrak{M}, w \models \bot$ is always false.

$\mathfrak{M}, w \models p \Leftrightarrow w \in V(p)$, if $p \in \Phi$.

$\mathfrak{M}, w \models \varphi_1 \wedge \varphi_2 \Leftrightarrow \mathfrak{M}, w \models \varphi_1$ and $\mathfrak{M}, w \models \varphi_2$.

$\mathfrak{M}, w \models \neg\varphi \Leftrightarrow \mathfrak{M}, w \models \varphi$ is false.

$\mathfrak{M}, w \models \Diamond\varphi \Leftrightarrow$ there exists v with $\langle w, v\rangle \in R$ and $\mathfrak{M}, v \models \varphi$.

The interesting part is of course the last line. We want $\Diamond\varphi$ to hold in state w; by our informal understanding, this means that a transition into a state such that φ holds in this state is possible. But this means that there exists some state v with $\langle w, v\rangle \in R$ such that φ holds in v. This is just the formulation we did use above. Look at $\Box\varphi$; an easy calculation shows that $\mathfrak{M}, w \models \Box\varphi$ iff $\mathfrak{M}, w \models \varphi$ for all v with $\langle w, v\rangle \in R$; thus, no matter what transition from world w to another world v we make, and $\mathfrak{M}, v \models \varphi$ holds, then $\mathfrak{M}, w \models \Box\varphi$. But we want to emphasize that for $\mathfrak{M}, w \models \Diamond\varphi$ to hold, we infer that w has at least one successor under relation R.

We define $[\![\varphi]\!]_{\mathfrak{M}}$ as the set of all states in which formula φ holds. Formally,

$$[\![\varphi]\!]_{\mathfrak{M}} := \{w \in W \mid \mathfrak{M}, w \models \varphi\}.$$

Let us look at some examples.

Example 2.7.9 Put $\Phi := \{p, q, r\}$ as the set of propositional letters and $W := \{1, 2, 3, 4, 5\}$ as the set of states; relation R is given through

$$1 \longrightarrow 2 \longrightarrow 3 \longrightarrow 4 \longrightarrow 5$$

Finally, put

$$V(\ell) := \begin{cases} \{2, 3\}, & \ell = p \\ \{1, 2, 3, 4, 5\}, & \ell = q \\ \emptyset, & \ell = r \end{cases}$$

Then we have for the Kripke model $\mathfrak{M} := (W, R, V)$, for example:

$\mathfrak{M}, 1 \models \Diamond\Box p$: This is so since $\mathfrak{M}, 3 \models p$ (because $3 \in V(p)$); thus $\mathfrak{M}, 2 \models \Box p$; hence $\mathfrak{M}, 1 \models \Diamond\Box p$.

$\mathfrak{M}, 1 \not\models \Diamond\Box p \to p$: Since $1 \notin V(p)$, we have $\mathfrak{M}, 1 \not\models p$.

$\mathfrak{M}, 2 \models \Diamond(p \wedge \neg r)$: The only successor to 2 in R is state 3, and we see that $3 \in V(p)$ and $3 \notin V(r)$.

$\mathfrak{M}, 1 \models q \wedge \Diamond(q \wedge \Diamond(q \wedge \Diamond(q \wedge \Diamond q)))$: Because $1 \in V(q)$ and 2 is the successor to 1, we investigate whether $\mathfrak{M}, 2 \models q \wedge \Diamond(q \wedge \Diamond(q \wedge \Diamond q))$ holds. Since $2 \in V(q)$ and $\langle 2, 3\rangle \in R$, we look at $\mathfrak{M}, 3 \models q \wedge \Diamond(q \wedge \Diamond q)$; now $\langle 3, 4\rangle \in R$ and $\mathfrak{M}, 3 \models q$, so we investigate $\mathfrak{M}, 4 \models q \wedge \Diamond q$. Since $4 \in V(q)$ and $\langle 4, 5\rangle \in R$, we find that this is true. Let φ denote the formula $q \wedge \Diamond(q \wedge \Diamond(q \wedge \Diamond(q \wedge \Diamond q)))$; then this peeling-off layers of parentheses shows that $\mathfrak{M}, 2 \not\models \varphi$, because $\mathfrak{M}, 5 \models \Diamond p$ does not hold.

$\mathfrak{M}, 1 \not\models \Diamond\varphi \wedge q$: Since $\mathfrak{M}, 2 \not\models \varphi$, and since state 2 is the only successor to 1, we see that $\mathfrak{M}, 1 \not\models \varphi$.

$\mathfrak{M}, w \models \Box q$: This is true for all worlds w, because $w' \in V(q)$ for all w' which are successors to some $w \in W$.

✌

Example 2.7.10 We have two propositional letters p and q; as set of states, we put $W := \{1, 2, 3, 4, 6, 8, 12, 24\}$; and we say

$$x \; R \; y \Leftrightarrow x \neq y \text{ and } x \text{ divides } y.$$

This is what R looks like without transitive arrows:

$$\begin{tikzcd}
1 \arrow[r] \arrow[dr] & 2 \arrow[r] \arrow[dr] & 4 \arrow[r] \arrow[dr] & 8 \arrow[dr] & \\
& 3 \arrow[r] & 6 \arrow[r] & 12 \arrow[r] & 24
\end{tikzcd}$$

Put $V(p) := \{4, 8, 12, 24\}$ and $V(q) := \{6\}$. Define the Kripke model $\mathfrak{M} := (W, R, V)$. We obtain for example:

$\mathfrak{M}, 4 \models \Box p$**:** The set of successor to state 4 is just $\{8, 12, 24\}$ which is a subset of $V(p)$.

$\mathfrak{M}, 6 \models \Box p$**:** Here we may reason in the same way.

$\mathfrak{M}, 2 \not\models \Box p$**:** State 6 is a successor to 2, but $6 \notin V(p)$.

$\mathfrak{M}, 2 \models \Diamond(q \wedge \Box p) \wedge \Diamond(\neg q \wedge \Box p)$**:** State 6 is a successor to state 2 with $\mathfrak{M}, 6 \models q \wedge \Box p$, and state 4 is a successor to state 2 with $\mathfrak{M}, 4 \models \neg q \wedge \Box p$

✌

Satisfiability etc

Let us introduce some terminology which will be needed later. We say that a formula φ is *globally true* in a Kripke model $\mathfrak{M}$ with state space W iff $[\![\varphi]\!]_{\mathfrak{M}} = W$, hence iff $\mathfrak{M}, w \models \varphi$ for all states $w \in W$; this is indicated by $\mathfrak{M} \models \varphi$. If $[\![\varphi]\!]_{\mathfrak{M}} \neq \emptyset$, thus if there exists $w \in W$ with $\mathfrak{M}, w \models \varphi$, we say that formula φ is *satisfiable*; φ is said to be *refutable* or *falsifiable* iff $\neg\varphi$ is satisfiable. A set Σ of formulas is said to be *globally true* iff $\mathfrak{M}, w \models \Sigma$ for all $w \in W$ (where we put $\mathfrak{M}, w \models \Sigma$ iff $\mathfrak{M}, w \models \varphi$ for all $\varphi \in \Sigma$). Σ is *satisfiable* iff $\mathfrak{M}, w \models \Sigma$ for some $w \in W$.

Kripke models are but one approach for interpreting modal logics. We observe that for a given transition system $(S, \rightsquigarrow)$, the set $N(s) := \{s' \in S \mid s \rightsquigarrow s'\}$ may consist of more than one state; one may consider $N(s)$ as the neighborhood of state s. An external observer may not be able to observe $N(s)$ exactly, but may determine that $N(s) \subseteq A$ for subsets $A \subseteq S$. Obviously, $N(s) \subseteq A$ and $A \subseteq B$ imply $N(s) \subseteq B$, so that the sets defined by containing the neighborhood $N(s)$ of a state s forms an upper closed set. This leads to the definition of neighborhood frames.

Definition 2.7.11 *Given a set S of states, a* neighborhood frame $\mathfrak{N} := (S, N)$ *is defined by a map* $N : S \to \mathbf{V}(S) := \{V \subseteq \mathcal{P}(S) \mid V$ *is upper closed*$\}$. N *is called an* effectivity function *on* S.

Effectivity function

The set $\mathbf{V}(S)$ of all upper closed families of subsets of S was introduced in Example 2.3.13.

So if we consider state $s \in S$ in a neighborhood frame, then $N(s)$ is an upper closed set which gives all sets the next state may be a member of. These frames occur in a natural way in topological spaces.

Example 2.7.12 Let T be a topological space, then

$$V(t) := \{A \subseteq T \mid U \subseteq A \text{ for some open neighborhood } U \text{ of } t\}$$

defines a neighborhood frame (T, V). ✌

Another straightforward example is given by ultrafilters.

Example 2.7.13 Given a set S, define

$$U(x) := \{U \subseteq S \mid x \in U\},$$

the ultrafilter associated with x. Then (S, U) is a neighborhood frame. ✌

Each Kripke frame gives rise to neighborhood frames in this way:

Example 2.7.14 Let (W, R) be a Kripke frame, and define for the world $w \in W$ the sets

$$V_R(w) := \{A \in \mathcal{P}(W) \mid R(w) \subseteq A\},$$
$$V'_R(w) := \{A \in \mathcal{P}(W) \mid R(w) \cap A \neq \emptyset\}$$

(with $R(w) := \{v \in W \mid \langle w, v\rangle \in R\}$); then both (W, V_R) and (W, V'_R) are neighborhood frames. ✌

A neighborhood frame induces a map on the power set of the state space into this power set. This map is used sometimes for an interpretation in lieu of the neighborhood function. Fix a map $P : S \to \mathbf{V}S$ for illustrating this. Given a subset $A \subseteq S$, we determine those states $\vartheta_P(A)$ which can achieve a state in A through P; hence $\vartheta_P(A) := \{s \in S \mid A \in P(s)\}$. This yields a map $\vartheta_P : \mathcal{P}(S) \to \mathcal{P}(S)$, which is monotone since $P(s)$ is upper closed for each s. Conversely, given a monotone map $\vartheta : \mathcal{P}(S) \to \mathcal{P}(S)$, we define $P_\vartheta : S \to \mathbf{V}(S)$

through $R_\vartheta(s) := \{A \subseteq S \mid s \in \vartheta(A)\}$. It is plain that $\vartheta_{R_\vartheta} = \vartheta$ and $R_{\vartheta_P} = P$.

Definition 2.7.15 *Given a set S of states, a neighborhood frame (S, N), and a map $V : \Phi \to \mathcal{P}(S)$, associating each propositional letter with a set of states. Then $\mathcal{N} := (S, N, V)$ is called a* neighborhood model

We define validity in a neighborhood model by induction on the structure of a formula, this time through the validity sets:

$$\begin{aligned}
[\![\top]\!]_{\mathcal{N}} &:= S, \\
[\![p]\!]_{\mathcal{N}} &:= V(p), \text{ if } p \in \Phi, \\
[\![\varphi_1 \wedge \varphi_2]\!]_{\mathcal{N}} &:= [\![\varphi_1]\!]_{\mathcal{N}} \cap [\![\varphi_2]\!]_{\mathcal{N}}, \\
[\![\neg\varphi]\!]_{\mathcal{N}} &:= S \setminus [\![\varphi]\!]_{\mathcal{N}}, \\
[\![\Box\varphi]\!]_{\mathcal{N}} &:= \{s \in S \mid [\![\varphi]\!]_{\mathcal{N}} \in N(s)\}.
\end{aligned}$$

$\mathcal{N}, s \models \varphi$

In addition, we put $\mathcal{N}, s \models \varphi$ iff $s \in [\![\varphi]\!]_{\mathcal{N}}$. Consider the last line and assume that the neighborhood frame underlying the model is generated by a Kripke frame (W, R), so that $A \in N(w)$ iff $R(w) \subseteq A$. Then $\mathcal{N}, w' \models \Box\varphi$ translates into $w' \in \{w \in S \mid R(w) \subseteq [\![\varphi]\!]_{\mathcal{N}}\}$, so that $\mathcal{N}, w \models \Box\varphi$ iff each world which is accessible from world w satisfies φ; this is what we want. Extending the definition above, we put

$$[\![\Diamond\varphi]\!]_{\mathcal{N}} := \{s \in S \mid S \setminus [\![\varphi]\!]_{\mathcal{N}} \notin N(s)\},$$

so that $\mathcal{N}, s \models \Diamond\varphi$ iff $\mathcal{N}, s \models \neg\Box\neg\varphi$.

Back to the general discussion. We generalize the notion of a Kripke model for capturing extended modal languages. The idea for an extension is straightforward—for interpreting a modal formula given by a modal operator of arity n, we require a subset of W^{n+1}. This leads to the definition of a frame, adapted to this purpose.

Definition 2.7.16 *Given a similarity type $\mathfrak{t} = (O, \rho)$, $\mathfrak{F} = (W, (R_\Delta)_{\Delta \in O})$ is said to be a* $\mathfrak{t}$-frame *iff $W \neq \emptyset$ is a set of states, and $R_\Delta \subseteq W^{\rho(\Delta)+1}$ for each $\Delta \in O$. A* $\mathfrak{t}$-model $\mathfrak{M} = (\mathfrak{F}, V)$ *is a $\mathfrak{t}$-frame $\mathfrak{F}$ with a map $V : \Phi \to \mathcal{P}(W)$.*

Given a $\mathfrak{t}$-model $\mathfrak{M}$, we define the interpretation of formulas like $\Delta(\varphi_1, \ldots, \varphi_n)$ and its nabla-cousin $\nabla(\varphi_1, \ldots, \varphi_n)$ in this way:

- $\mathfrak{M}, w \models \Delta(\varphi_1, \ldots, \varphi_n)$ iff there exist $w_1, \ldots, w_n$ with:

  1. $\mathfrak{M}, w_i \models \varphi_i$ for $1 \leq i \leq n$,
  2. $\langle w, w_1, \ldots, w_n \rangle \in R_\Delta$,

  if $n > 0$,
- $\mathfrak{M}, w \models \Delta$ iff $w \in R_\Delta$ for $n = 0$,
- $\mathfrak{M}, w \models \nabla(\varphi_1, \ldots, \varphi_n)$ iff $\big(\langle w, w_1, \ldots, w_n \rangle \in R_\Delta$ implies $\mathfrak{M}, w_i \models \varphi_i$ for all $i \in \{1, \ldots, n\}\big)$ for all $w_1, \ldots, w_n \in W$, if $n > 0$,
- $\mathfrak{M}, w \models \nabla$ iff $w \notin R_\Delta$, if $n = 0$.

In the last two cases, ∇ is the nabla for modal operator Δ.

Just in order to get a grip on these definitions, let us have a look at some examples.

Example 2.7.17 The set O of modal operators consists just of the unary operators $\{\langle a \rangle, \langle b \rangle, \langle c \rangle\}$, the relations on the set $W := \{w_1, w_2, w_3, w_4\}$ of worlds are given by:

$$
\begin{aligned}
R_a &:= \{\langle w_1, w_2 \rangle, \langle w_4, w_4 \rangle\}, \\
R_b &:= \{\langle w_2, w_3 \rangle\}, \\
R_c &:= \{\langle w_3, w_4 \rangle\}.
\end{aligned}
$$

There is only one propositional letter p, and put $V(p) := \{w_2\}$. This comprises a $\mathfrak{t}$-model $\mathfrak{M}$. We want to check whether $\mathfrak{M}, w_1 \models \langle a \rangle p \to \langle b \rangle p$ holds. Allora: In order to establish whether or not $\mathfrak{M}, w_1 \models \langle a \rangle p$ holds, we have to find a state v such that $\langle w_1, v \rangle \in R_a$ and $\mathfrak{M}, v \models p$; state w_2 is the only possible choice. But $\mathfrak{M}, w_1 \not\models p$, because $w_1 \notin V(p)$. Hence $\mathfrak{M}, w_1 \not\models \langle a \rangle p \to \langle b \rangle p$. ✌

Example 2.7.18 Let $W = \{u, v, w, s\}$ be the set of worlds; we take $O := \{\Diamond, \clubsuit\}$ with $\rho(\Diamond) = 2$ and $\rho(\clubsuit) = 3$. Put $R_\Diamond := \{\langle u, v, w \rangle\}$ and $R_\clubsuit := \{\langle u, v, w, s \rangle\}$. The set Φ of propositional letters is $\{p_0, p_1, p_2\}$ with $V(p_0) := \{v\}$, $V(p_1) := \{w\}$ and $V(p_2) := \{s\}$. This yields a model $\mathfrak{M}$.

1. We want to determine $[\![\Diamond(p_0, p_1)]\!]_{\mathfrak{M}}$. From the definition of $\models$, we see that

$$\mathfrak{M}, x \models \Diamond(p_0, p_1) \text{ iff } \exists x_0, x_1 : \mathfrak{M}, x_0 \models p_0 \text{ and } \mathfrak{M}, x_1 \models p_1 \\ \text{and } \langle x, x_0, x_1 \rangle \in R_\Diamond.$$

   We obtain by inspection $[\![\Diamond(p_0, p_1)]\!]_{\mathfrak{M}} = \{u\}$.

2. We have $\mathfrak{M}, u \models \clubsuit(p_0, p_1, p_2)$. This is so since $\mathfrak{M}, v \models p_0$, $\mathfrak{M}, w \models p_1$, and $\mathfrak{M}, s \models p_2$ together with $\langle u, v, w, s\rangle \in R_\clubsuit$.

3. Consequently, we have $[\![\Diamond(p_0, p_1) \to \clubsuit(p_0, p_1, p_2)]\!]_{\mathfrak{M}} = \{u\}$.

✌

Example 2.7.19 Let us look into the future and into the past. We are given the unary operators $O = \{\boldsymbol{F}, \boldsymbol{P}\}$ as in Example 2.7.3. The interpretation requires two binary relations $R_{\boldsymbol{F}}$ and $R_{\boldsymbol{P}}$; we have defined the corresponding nablas $\boldsymbol{G}$ resp $\boldsymbol{H}$. Unless we want to change the past, we assume that $R_{\boldsymbol{P}} = R_{\boldsymbol{F}}^{-1}$, so just one relation $R := R_{\boldsymbol{F}}$ suffices for interpreting this logic. Hence:

$\mathfrak{M}, x \models \boldsymbol{F}\varphi$: This is the case iff there exists $z \in W$ such that $\langle x, z\rangle \in R$ and $\mathfrak{M}, z \models \varphi$.

$\mathfrak{M}, x \models \boldsymbol{P}\varphi$: This is true iff there exists $z \in W$ with $\langle v, x\rangle \in R$ and $\mathfrak{M}, v \models \varphi$.

$\mathfrak{M}, x \models \boldsymbol{G}\varphi$: This holds iff we have $\mathfrak{M}, y \models \varphi$ for all y with $\langle x, y\rangle \in R$.

$\mathfrak{M}, x \models \boldsymbol{H}\varphi$: Similarly, for all y with $\langle y, x\rangle \in R$, we have $\mathfrak{M}, y \models \varphi$.

✌

The next case is a little more involved since we have to construct the relations from the information that is available. In the case of PDL (see Example 2.7.4), we have only information about the behavior of atomic programs, and we construct from it the relations for compound programs, since, after all, a compound program is composed from the atomic programs according to the rules laid down in Example 2.7.4.

Example 2.7.20 Let Ψ be the set of all atomic programs, and assume that we have for each $t \in \Psi$ a relation $R_t \subseteq W \times W$; so if atomic program t is executed in state s, then $R_t(s)$ yields the set of all possible successor states after execution. Now we define by induction on the structure of the programs these relations:

$$\begin{aligned} R_{\pi_1 \cup \pi_2} &:= R_{\pi_1} \cup R_{\pi_2}, \\ R_{\pi_1;\pi_2} &:= R_{\pi_1} \circ R_{\pi_2}, \\ R_{\pi^*} &:= \bigcup_{n \geq 0} R_{\pi^n}. \end{aligned}$$

(here we put $R_{\pi_0} := \{\langle w, w\rangle \mid w \in W\}$, and $R_{\pi^{n+1}} := R_\pi \circ R_{\pi^n}$). Then, if $\langle x, y\rangle \in R_{\pi_1 \cup \pi_2}$, we know that $\langle x, y\rangle \in R_{\pi_1}$ or $\langle x, y\rangle \in R_{\pi_2}$, which reflects the observation that we can enter a new state y upon choosing between π_1 and π_2. Hence executing $\pi_1 \cup \pi_2$ in state x, we should be able to enter this state upon executing one of the programs. Similarly, if in state x we execute first π_1 in x and then π_2, we should enter an intermediate state z after executing π_1 and then execute π_2 in state z, yielding the resulting state. Executing π^* means that we execute π^n a finite number of times (probably not at all). This explains the definition for R_{π^*}.

Finally, we should define $R_{\varphi?}$ for a formula φ. The intuitive meaning of a program like $\varphi?; \pi$ is that we want to execute π, provided formula φ holds. This suggests defining

$$R_{\varphi?} := \{\langle w, w\rangle \mid \mathfrak{M}, w \models \varphi\}.$$

Note that we rely here on a model $\mathfrak{M}$ which is already defined.

Just to get familiar with these definitions, let us have a look at the composition operator.

$$\begin{aligned} \mathfrak{M}, x \models \langle \pi_1; \pi_2\rangle\varphi &\Leftrightarrow \exists v : \mathfrak{M}, v \models \varphi \text{ and } \langle x, v\rangle \in R_{\pi_1;\pi_2} \\ &\Leftrightarrow \exists w \in W \exists v \in [\![\varphi]\!]_{\mathfrak{M}} : \langle x, w\rangle \in R_{\pi_1} \text{ and } \langle w, v\rangle \in R_{\pi_2} \\ &\Leftrightarrow \exists w \in [\![\langle \pi_2\rangle\varphi]\!]_{\mathfrak{M}} : \langle x, w\rangle \in R_{\pi_1} \\ &\Leftrightarrow \mathfrak{M}, x \models \langle \pi_1\rangle\langle \pi_2\rangle\varphi \end{aligned}$$

This means that $\langle \pi_1; \pi_2\rangle\varphi$ and $\langle \pi_1\rangle\langle \pi_2\rangle\varphi$ are semantically equivalent, which is intuitively quite clear.

The test operator is examined in the next formula. We have

$$R_{\varphi?;\pi} = R_{\varphi?} \circ R_\pi = \{\langle x, y\rangle \mid \mathfrak{M}, x \models x \text{ and } \langle x, y\rangle \in R_\pi\} = R_{\varphi?} \cap R_\pi.$$

hence $\mathfrak{M}, y \models \langle \varphi?; \pi\rangle\psi$ iff $\mathfrak{M}, y \models \varphi$ and $\mathfrak{M}, y \models \langle\pi\rangle\psi$, so that

$$\mathfrak{M}, y \models (\langle\varphi?; \pi_1\rangle \cup \langle\neg\varphi?; \pi_2\rangle)\varphi \text{ iff } \begin{cases} \mathfrak{M}, y \models \langle\pi_1\rangle\varphi, & \text{if } \mathfrak{M}, y \models \varphi \\ \mathfrak{M}, y \models \langle\pi_2\rangle\varphi, & \text{otherwise} \end{cases}$$

✌

The next example shows that we can interpret PDL in a neighborhood model as well.

Example 2.7.21 We associate with each atomic program $t \in \Psi$ of PDL an effectivity function $E_t : W \to \mathbf{V}(W)$ on the state space W. Hence if we execute t in state w, then $E_t(w)$ is the set of all subsets A of the states so that the next state is a member of A (we say that the program t can *achieve* a state in A). Hence $\big(W, (E_t)_{t\in\Psi}\big)$ is a neighborhood frame. We have indicated on page 231 that we can construct from a neighborhood function a monotone map from the power set of W into itself; see also Exercise 2.43; so we define

$$R'_t(A) := \{w \in W \mid A \in R_t(w)\},$$

giving a function $R'_t : \mathcal{P}(W) \to \mathcal{P}(W)$. R'_t is monotone, since R_t is an effectivity function: $A \subseteq B$, and $w \in R'_t(A)$, then $A \in R_t(w)$; hence $B \in R_t(w)$, and thus $w \in R'_t(B)$. These maps can be extended to programs along their syntax in the following way, which is very similar to the one for relations:

$$\begin{aligned} R'_{\pi_1\cup\pi_2} &:= R'_{\pi_1} \cup R'_{\pi_2}, \\ R'_{\pi_1;\pi_2} &:= R'_{\pi_1} \circ R'_{\pi_2} \\ R'_{\pi^*} &:= \bigcup_{n\geq 0} R'_{\pi^n} \end{aligned}$$

with R'_{π^0} and R'_{π^n} defined as above.

Assume that we have again a function $V : \Phi \to \mathcal{P}(W)$, yielding a neighborhood model $\mathcal{N}$. The definition above are used now for the interpretation of formulas $\langle\pi\rangle\varphi$ through

$$[\![\langle\pi\rangle\varphi]\!]_{\mathcal{N}} := R'_\pi([\![\varphi]\!]_{\mathcal{N}}).$$

Interpreting $R_\pi(w) := \{A \subseteq W \mid w \in R'_\pi(A)\}$ as the sets which can be achieved by the execution of program π, we have $[\![\langle\pi\rangle\varphi]\!]_{\mathcal{N}} = \{w \in W \mid [\![\varphi]\!]_{\mathcal{N}} \in R_\pi(w)\}$, so that $[\![\langle\pi\rangle\varphi]\!]_{\mathcal{N}}$ describes the set of all states for which $[\![\varphi]\!]_{\mathcal{N}}$ can be achieved upon execution of π. The definition of $R_{\varphi?}$ carries over, so that this yields an interpretation of PDL. ✌

Turning to game logic from Example 2.7.5, we note that neighborhood models are suited to interpret this logic as well. Assign for each atomic game $\gamma \in \Gamma$ to Angel the effectivity function P_γ, then $P_\gamma(s)$ indicates what Angel can achieve when playing γ in state s. Specifically, $A \in P_\gamma(s)$ indicates that Angel has a strategy for achieving by playing γ in state s that the next state of the game is a member of A. We will not formalize the notion of a strategy here but appeal rather to an informal understanding. The dual operator permit converting a game into its dual, where players change rôles: the moves of Angel become moves of Demon, and vice versa.

We should note that the Banach–Mazur games modeled in Sect. 1.7 and the games considered here display significant differences. First, Banach–Mazur games are played essentially over the playground $\mathbb{N}^\mathbb{N}$, which means that it should always be possible that such a game, if it is played over another domain, can be mapped to this urform. By construction, those games continue infinitely, and they have a well-defined notion of strategy, which permits to define what a winning strategy is. As pointed just out, the games we are about to consider do not have a formal definition of a strategy, we work rather with the informal notion that, e.g., Angel has a strategy to achieve something. When discussing games, we will be careful to distinguish both varieties.

Let us just indicate informally by $\langle\gamma\rangle\varphi$ that Angel has a strategy in game γ which makes sure that game γ results in a state which satisfies formula φ. We assume the game to be *determined*: if one player does not have a winning strategy, then the other one has. Thus if Angle does not have a $\neg\varphi$-strategy, then Demon has a φ-strategy, and vice versa. This means that we can derive the way Demon plays the game from the way Angel does, and vice versa.

Example 2.7.22 As in Example 2.7.5, we assume that games are given through this grammar

$$g ::= \gamma \mid g_1 \cup g_2 \mid g_1 \cap g_2 \mid g_1; g_2 \mid g^d \mid g^* \mid g^\times \mid \varphi?$$

with $\gamma \in \Gamma$, the set of atomic games. We assume that the game is determined; hence we may express demonic choice $g_1 \cap g_2$ through $(g_1^d \cup g_2^d)^d$ and demonic iteration $g^\times$ through angelic iteration $\big((g^d)^*\big)^d$).

Assign to each $\gamma \in \Gamma$ an effectivity function P_γ on the set W of worlds, and put

$$P'_\gamma(A) := \{w \in W \mid A \in P_\gamma(w)\}.$$

Hence $w \in P'_\gamma(A)$ indicates that Angel has a strategy to achieve A by playing game γ in state w. We extend P' to games along the lines of the games' syntax, and put

$$[\![\langle g\rangle\varphi]\!]_{\mathcal{N}} := P'_g([\![\varphi]\!]_{\mathcal{N}}).$$

for neighborhood model $\mathcal{N}$, the game g and formula φ (see the construction after Definition 2.7.15). This is the extension:

$$\begin{aligned}
P'_{g_1\cup g_2}(A) &:= P'_{g_1}(A) \cup P'_{g_2}(A), & P'_{g^d}(A) &:= W \setminus P'_g(W \setminus A),\\
P'_{g_1;g_2}(A) &:= P'_{g_1}(P'_{g_2}(A)), & P'_{g_1\cap g_2}(A) &:= P'_{(g_1^d\cup g_2^d)^d}(A),\\
P'_{g^*}(A) &:= \bigcup_{n\geq 0} P'_{g^n}(A), & P'_{g^\times}(A) &:= P'_{\big((g^d)^*\big)^d}(A),\\
P'_{\varphi ?}(A) &:= [\![\varphi]\!]_{\mathcal{N}} \cap A.
\end{aligned}$$

A straightforward suggestion for the interpretation of game logic is the approach through Kripke models, very much in line with the interpretation of modal logics in general. There are, however, some difficulties associated with this idea, which are discussed in [PP03, Sect. 3], in particular [PP03, Theorem 1]. This is the reason why Kripke models are not adequate for interpreting game logics: If games are interpreted through Kripke models, the interpretation turns out to be *disjunctive*. This means that $\langle g_1; (g_2 \cup g_3)\rangle\varphi$ is semantically equivalent to $\langle g_1; g_2 \cup g_1; g_3\rangle\varphi$ for all games g_1, g_2, g_3. This, however, is not desirable: Angle's decision after playing g_1 whether to play g_2 or g_3 should not be equivalent to decide whether to play $g_1; g_2$ or $g_1; g_3$. Neighborhood models in their greater generality do not display this equivalence, so they are more general. ✌

In this little gallery of examples, let us finally have a look at arrow logic; see Example 2.7.6.

Example 2.7.23 Arrows are interpreted as vectors, hence, e.g., as pairs. Let W be a set of states; then we take $W \times W$ as the domain of our interpretation. We have three modal operators:

- The nullary operator `skip` is interpreted through $R_{\texttt{skip}} := \{\langle w, w\rangle \mid w \in W\}$.
- The unary operator $\otimes$ is interpreted through $R_\otimes := \big\{\langle\langle a, b\rangle, \langle b, a\rangle\rangle \mid a, b \in W\big\}$.
- The binary operator is intended to model composition; thus one end of the first arrow should be the be other end of the second arrow; hence $R_\circ := \big\{\langle\langle a, b\rangle, \langle b, c\rangle, \langle a, c\rangle\rangle \mid a, b, c \in W\big\}$.

With this, we obtain, for example, $\mathfrak{M}, \langle w_1, w_2\rangle \models \psi_1 \circ \psi_2$ iff there exists v such that $\mathfrak{M}, \langle w_1, v\rangle \models \psi_1$ and $\mathfrak{M}, \langle v, w_2\rangle \models \psi_2$. ✌

Frames are related through frame morphisms. Take a frame (W, R) for the basic modal language; then $R : W \to \mathcal{P}(W)$ is perceived as a coalgebra for the power set functor. This helps in defining morphisms.

Definition 2.7.24 *Let $\mathfrak{F} = (W, R)$ and $\mathfrak{G} = (X, S)$ be Kripke frames. A* frame morphism *$f : \mathfrak{F} \to \mathfrak{G}$ is a map $f : W \to X$ which makes this diagram commutative:* f frame morphism

$$\begin{tikzcd} W \arrow[r, "f"] \arrow[d, "R"'] & X \arrow[d, "S"] \\ \mathcal{P}(W) \arrow[r, "\mathcal{P}f"'] & \mathcal{P}(X) \end{tikzcd}$$

Hence we have the condition for a frame morphism $f : \mathfrak{F} \to \mathfrak{G}$

$$S(f(w)) = (\mathcal{P}f)(R(w)) = f\big[R(w)\big] = \{f(w') \mid w' \in R(w)\}.$$

for all $w \in W$.

This is a characterization of frame morphisms.

Lemma 2.7.25 *Let $\mathfrak{F}$ and $\mathfrak{G}$ be frames, as above. Then $f : \mathfrak{F} \to \mathfrak{G}$ is a frame morphism iff these conditions hold:*

1. *$w\ R\ w'$ implies $f(w)\ S\ f(w')$.*
2. *If $f(w)\ S\ z$, then there exists $w' \in W$ with $z = f(w')$ and $w\ R\ w'$.*

Proof 1. These conditions are necessary. In fact, if $\langle w, w'\rangle \in R$, then $f(w') \in f\big[R(w)\big] = S(f(w))$, so that $\langle f(w), f(w')\rangle \in S$. Similarly, assume that $f(w)\ S\ z$; thus $z \in S(f(w)) = \mathcal{P}(f)$

$(R(w)) = f\big[R(w)\big]$. Hence there exists w' with $\langle w, w'\rangle \in R$ and $z = f(w')$.

2. The conditions are sufficient. The first condition implies $f\big[R(w)\big] \subseteq S(f(w))$. Now assume $z \in S(f(w))$; hence $f(w)\ S\ z$, and thus there exists $w' \in R(w)$ with $f(w') = z$, consequently, $z = f(w') \in f\big[R(w)\big]$. $\dashv$

We see that the bounded morphisms from Example 2.1.10 appear here again in a natural context.

If we want to compare models for the basic modal language, then we certainly should be able to compare the underlying frames. But this is not yet enough, because the interpretation for atomic propositions has to be taken care of.

f model morphism

Definition 2.7.26 *Let $\mathfrak{M} = (W, R, V)$ and $\mathfrak{N} = (X, S, Y)$ be models for the basic modal language and $f : (W, R) \to (X, S)$ be a frame morphism. Then $f : \mathfrak{M} \to \mathfrak{N}$ is said to be a* model morphism *iff $f^{-1} \circ Y = V$.*

Hence $f^{-1}\big[Y(p)\big] = V(p)$ for a model morphism f and for each atomic proposition p; thus $\mathfrak{M}, w \models p$ iff $\mathfrak{N}, f(w) \models p$ for each atomic proposition. This extends to all formulas of the basic modal language.

Proposition 2.7.27 *Assume $\mathfrak{M}$ and $\mathfrak{N}$ are models as above, and $f : \mathfrak{M} \to \mathfrak{N}$ is a model morphism. Then*

$$\mathfrak{M}, w \models \varphi \textit{ iff } \mathfrak{N}, f(w) \models \varphi$$

for all worlds w of $\mathfrak{M}$ and for all formulas φ.

Proof 0. The assertion is equivalent to

$$[\![\varphi]\!]_{\mathfrak{M}} = f^{-1}\big[[\![\varphi]\!]_{\mathfrak{N}}\big]$$

Line of attack

for all formulas φ. This is the claim which will be established by induction on the structure of a formula now.

1. If p is an atomic proposition, then this is just the definition of a frame morphism to be a model morphism:

$$[\![p]\!]_{\mathfrak{M}} = V(p) = f^{-1}\big[Y(p)\big] = [\![p]\!]_{\mathfrak{N}}.$$

Assume that the assertion holds for φ_1 and φ_2; then

$$\begin{aligned}[\![\varphi_1 \wedge \varphi_2]\!]_{\mathfrak{M}} &= [\![\varphi_1]\!]_{\mathfrak{M}} \cap [\![\varphi_2]\!]_{\mathfrak{M}} & &= f^{-1}\big[[\![\varphi_1]\!]_{\mathfrak{N}}\big] \cap f^{-1}\big[[\![\varphi_2]\!]_{\mathfrak{N}}\big] \\ &= f^{-1}\big[[\![\varphi_1]\!]_{\mathfrak{N}} \cap [\![\varphi_2]\!]_{\mathfrak{N}}\big] & &= f^{-1}\big[[\![\varphi_1 \wedge \varphi_2]\!]_{\mathfrak{N}}\big].\end{aligned}$$

Similarly, one shows that $[\![\neg\varphi]\!]_{\mathfrak{M}} = f^{-1}\big[[\![\neg\varphi]\!]_{\mathfrak{N}}\big]$.

2. Now consider $\Diamond\varphi$; assume that the hypothesis holds for formula φ, then we have

$$\begin{aligned}[\![\Diamond\varphi]\!]_{\mathfrak{M}} &= \{w \mid \exists w' \in R(w) : w' \in [\![\varphi]\!]_{\mathfrak{M}}\} \\ &= \{w \mid \exists w' \in R(w) : f(w') \in [\![\varphi]\!]_{\mathfrak{N}}\} && \text{(by hypothesis)} \\ &= \{w \mid \exists w' : f(w') \in S(f(w)), f(w') \in [\![\varphi]\!]_{\mathfrak{N}}\} && \text{(by Lemma 2.7.25)} \\ &= f^{-1}\big[\{x \mid \exists x' \in S(x) : x' \in [\![\varphi]\!]_{\mathfrak{N}}\}\big] \\ &= f^{-1}\big[[\![\Diamond\varphi]\!]_{\mathfrak{N}}\big].\end{aligned}$$

Thus the assertion holds for all formulas φ. $\dashv$

The observation from Proposition 2.7.27 permits comparing worlds which are given through two models. Two worlds are said to be equivalent iff they cannot be separated by a formula, i.e., iff they satisfy exactly the same formulas.

Definition 2.7.28 *Let $\mathfrak{M}$ and $\mathfrak{N}$ be models with state spaces W resp. X. States $w \in W$ and $x \in X$ are called* modally equivalent *iff we have*

$$\mathfrak{M}, w \models \varphi \textit{ iff } \mathfrak{N}, x \models \varphi$$

for all formulas φ

Hence if $f : \mathfrak{M} \to \mathfrak{N}$ is a model morphism, then w and $f(w)$ are modally equivalent for each world w of $\mathfrak{M}$. One might be tempted to compare models with respect to their transition behavior; after all, underlying a model is a transition system, a.k.a. a frame. This leads directly to this notion of bisimilarity for models—note that we have to take the atomic propositions into account.

Definition 2.7.29 *Let $\mathfrak{M} = (W, R, V)$ and $\mathfrak{N} = (X, S, Y)$ be models for the basic modal language; then a relation $B \subseteq W \times X$ is called a* bisimulation *iff:*

1. *If $w \; B \; x$, then w and x satisfy the same propositional letters ("atomic harmony").*

2. *If $w\ B\ x$ and $w\ R\ w'$, then there exists x' with $x\ S\ x'$ and $w'\ B\ x'$ (forth condition).*

Atomic harmony, forth, back

3. *If $w\ B\ x$ and $x\ S\ x'$, then there exists w' with $w\ R\ w'$ and $w'\ B\ x'$ (back condition).*

States w and x are called bisimilar *iff there exists a bisimulation B with $\langle w, x\rangle \in B$.*

Hence the forth condition says for a pair of worlds $\langle w, x\rangle \in B$ that if $w \rightsquigarrow_R w'$, there exists x' with $\langle w', x'\rangle \in B$ such that $x \rightsquigarrow_S x'$, similarly for the back condition. So this rings a bell: we did discuss this in Definition 2.6.15. Consequently, if models $\mathfrak{M}$ and $\mathfrak{N}$ are bisimilar, then the underlying frames are bisimilar coalgebras.

Consider this example for bisimilar states.

Example 2.7.30 Let relation B be defined through

$$B := \{\langle 1, a\rangle, \langle 2, b\rangle, \langle 2, c\rangle, \langle 3, d\rangle, \langle 4, e\rangle, \langle 5, e\rangle\}$$

with $V(p) := \{a, d\}, V(q) := \{b, c, e\}$.

The transitions for $\mathfrak{M}$ are given through

$$\begin{tikzcd} 1 \arrow[r] & 2 \arrow[r] & 3 \arrow[dl] \arrow[d] \\ & 5 & 4 \end{tikzcd}$$

$\mathfrak{N}$ is given through

$$\begin{tikzcd} & b \arrow[dr] & & \\ a \arrow[ur] \arrow[dr] & & d \arrow[r] & e \\ & c \arrow[ur] & & \end{tikzcd}$$

Then B is a bisimulation ✌

The first result relating bisimulation and modal equivalence is intuitively quite clear. Since a bisimulation reflects the structural similarity of the transition structure of the underlying transition systems, and since the validity of modal formulas is determined through this transition structure (and the behavior of the atomic propositional formulas), it does not come as a surprise that bisimilar states are modally equivalent.

Proposition 2.7.31 *Let $\mathfrak{M}$ and $\mathfrak{N}$ be models with states w and x. If w and x are bisimilar, then they are modally equivalent.*

Proof 0. Let B be the bisimulation for which we know that $\langle w, x\rangle \in B$. We have to show that

$$\mathfrak{M}, w \models \varphi \Leftrightarrow \mathfrak{N}, x \models \varphi$$

for all formulas φ. This is done by induction on the formula.

1. Because of atomic harmony, the equivalence holds for propositional formulas. It is also clear that conjunction and negation are preserved under this equivalence, so that the case of proving the equivalence for a formula $\Diamond\varphi$ under the assumption that it holds for φ remains to be taken care of.

"⇒": Assume that $\mathfrak{M}, w \models \Diamond\varphi$ holds. Thus there exists a world w' in $\mathfrak{M}$ with $w\ R\ w'$ and $\mathfrak{M}, w' \models \varphi$. Hence there exists by the forward condition a world x' in $\mathfrak{N}$ with $x\ S\ x'$ and $\langle w', x'\rangle \in B$ such that $\mathfrak{N}, x' \models \varphi$ by the induction hypothesis. Because x' is a successor to x, we conclude $\mathfrak{N}, x \models \Diamond\varphi$.

"⇐": This is shown in the same way, using the back condition for B.

$\dashv$

The converse holds only under the restrictive condition that the models are image finite. Thus each state has only a finite number of successor states; formally, model (W, R, V) is called *image finite* iff for each world w the set $R(w)$ is finite. Then the famous Hennessy–Milner Theorem says: Image finite

Theorem 2.7.32 *If the models $\mathfrak{M}$ and $\mathfrak{N}$ are image finite, then modal equivalent states are bisimilar.*

Proof 1. Given two modal equivalent states w^* and x^*, we have to find a bisimulation B with $\langle w^*, x^*\rangle \in B$. The only thing we know about the states is that they are modally equivalent, hence that they satisfy exactly the same formulas. This suggests to define Plan

$$B := \{\langle w', x'\rangle \mid w' \text{ and } x' \text{ are modally equivalent}\}$$

and to establish B as a bisimulation. Since by assumption $\langle w^*, x^*\rangle \in B$, this will then prove the claim.

2. If $\langle w, x\rangle \in B$, then both satisfy the same atomic propositions by the definition of modal equivalence. Now let $\langle w, x\rangle \in B$ and $w\ R\ w'$. Assume that we cannot find x' with $x\ S\ x'$ and $\langle w', x'\rangle \in B$. We

know that $\mathfrak{M}, w \models \Diamond\top$, because this says that there exists a successor to w, viz., w'. Since w and x satisfy the same formulas, $\mathfrak{N}, x \models \Diamond\top$ follows; hence $S(x) \neq \emptyset$. Let $S(x) = \{x_1, \ldots, x_k\}$. Then, since w and x_i are not modally equivalent, we can find for each $x_i \in S(x)$ a formula ψ_i such that $\mathfrak{M}, w' \models \psi_i$, but $\mathfrak{N}, x_i \not\models \psi_i$,. Hence $\mathfrak{M}, w \models \Diamond(\psi_1 \wedge \ldots \wedge \psi_k)$, but $\mathfrak{N}, w \not\models \Diamond(\psi_1 \wedge \ldots \wedge \psi_k)$. This is a contradiction, so the assumption is false, and we can find x' with $x \; S \; x'$ and $\langle w', x' \rangle \in B$.

The other conditions for a bisimulation are shown in exactly the same way. $\dashv$

Neighborhood models can be compared through morphisms as well. Recall that the functor $\mathbf{V}$ underlies a neighborhood frame; see Example 2.3.13.

f neighborhood morphism

Definition 2.7.33 *Let $\mathcal{N} = (W, N, V)$ and $\mathcal{M} = (X, M, Y)$ be neighborhood models for the basic modal language. A map $f : W \to X$ is called a* neighborhood morphism $f : \mathcal{N} \to \mathcal{M}$ *iff:*

- $N \circ f = (\mathbf{V} f) \circ M$,
- $V = f^{-1} \circ Y$.

A neighborhood morphism is a morphism for the neighborhood frame (the definition of which is straightforward), respecting the validity of atomic propositions. In this way, the definition follows the pattern laid out for morphisms of Kripke models. Expanding the definition above, $f : \mathcal{N} \to \mathcal{M}$ is a neighborhood morphism iff these conditions hold: $B \in N(f(w))$ iff $f^{-1}\big[B\big] \in M(w)$ for all $B \subseteq X$ and all worlds $w \in W$, and $V(p) = f^{-1}\big[Y(p)\big]$ for all atomic sentences $p \in \Phi$. Morphisms for neighborhood models preserve validity in the same way as morphisms for Kripke models do:

Proposition 2.7.34 *Let $f : \mathcal{N} \to \mathcal{M}$ be a neighborhood morphism for the neighborhood models $\mathcal{N} = (W, N, V)$ and $\mathcal{M} = (X, M, Y)$. Then*

$$\mathcal{N}, w \models \varphi \Leftrightarrow \mathcal{M}, f(w) \models \varphi$$

for all formulas φ and for all states $w \in W$.

Proof The proof proceeds by induction on the structure of formula φ. The induction starts with φ an atomic proposition. The assertion is true in this case because of atomic harmony; see the proof of Proposition 2.7.27. We pick only the interesting modal case for the induction step. Hence assume the assertion is established for formula φ; then

$$\begin{aligned}\mathcal{M}, f(w) \models \Box\varphi &\Leftrightarrow [\![\varphi]\!]_{\mathcal{M}} \in M(f(w)) && \text{(by definition)}\\ &\Leftrightarrow f^{-1}\big[[\![\varphi]\!]_{\mathcal{M}}\big] \in N(w) && \text{(f is a morphism)}\\ &\Leftrightarrow [\![\varphi]\!]_{\mathcal{N}} \in N(w) && \text{(by induction hypothesis)}\\ &\Leftrightarrow \mathcal{N}, w \models \Box\varphi\end{aligned}$$

$\dashv$

We will not pursue this observation further at this point, but rather turn to the construction of a canonic model. When we will discuss coalgebraic logics, however, this striking structural similarity of models and their morphisms will be shown to be an instance of more general pattern.

Before proceeding, we introduce the notion of a *substitution*, which is a map $\sigma : \Phi \to \mathcal{L}(\mathrm{t}, \Phi)$. We extend a substitution in a natural way to formulas. Define by induction on the structure of a formula Substitution

$$\begin{aligned}p^\sigma &:= \sigma(p), \text{ if } p \in \Phi,\\ (\neg\varphi)^\sigma &:= \neg(\varphi^\sigma),\\ (\varphi_1 \wedge \varphi_2)^\sigma &:= \varphi_1^\sigma \wedge \varphi_2^\sigma,\\ \big(\Delta(\varphi_1, \dots, \varphi_k)\big)^\sigma &:= \Delta(\varphi_1^\sigma, \dots, \varphi_k^\sigma), \text{ if } \Delta \in O \text{ with } \rho(\Delta) = k.\end{aligned}$$

2.7.2 The Lindenbaum Construction

We will show now how we obtain from a set of formulas a model which satisfies exactly these formulas. The scenario is the basic modal language, and it is clear that not every set of formulas is in a position to generate such a model.

Let Λ be a set of formulas; then we say that:

- Λ is *closed under modus ponens* iff $\varphi \in \Lambda$ and $\varphi \to \psi$ together imply $\psi \in \Lambda$;
- Λ is *closed under uniform substitution* iff given $\varphi \in \Lambda$ we may conclude that $\varphi^\sigma \in \Lambda$ for all substitutions σ.

These two closure properties turn out to be crucial for the generation of a model from a set of formulas. Those sets which satisfy them will be called modal logics, to be precise:

Definition 2.7.35 *Let Λ be a set of formulas of the basic modal language. Λ is called a* modal logic *iff these conditions are satisfied:*

1. *Λ contains all propositional tautologies.*
2. *Λ is closed under modus ponens and under uniform substitution.*

$\vdash_\Lambda \varphi$

If formula $\varphi \in \Lambda$, then φ is called a theorem of Λ; we write this as $\vdash_\Lambda \varphi$.

Example 2.7.36 These are some instances of elementary properties for modal logics.

1. If Λ_i is a modal logic for each $i \in I \neq \emptyset$, then $\bigcap_{i \in I} \Lambda_i$ is a modal logic. This is fairly easy to check.

$\mathfrak{F} \models \varphi$

2. We say for a formula φ and a frame $\mathfrak{F}$ over W as a set of states that φ *holds in this frame* (in symbols $\mathfrak{F} \models \varphi$) iff $\mathfrak{M}, w \models \varphi$ for each $w \in W$ and each model $\mathfrak{M}$ which is based on $\mathfrak{F}$. Let $\mathbb{C}$ be a class of frames, then
$$\Lambda_{\mathbb{C}} := \bigcap_{\mathfrak{F} \in \mathbb{C}} \{\varphi \mid \mathfrak{F} \models \varphi\}$$
is a modal logic. We abbreviate $\varphi \in \Lambda_{\mathbb{C}}$ by $\mathbb{C} \models \varphi$.

$\mathfrak{M} \models \varphi$

3. Define similarly $\mathfrak{M} \models \varphi$ for a model $\mathfrak{M}$ iff $\mathfrak{M}, w \models \varphi$ for each world w of $\mathfrak{M}$. Then put for a class $\mathbb{M}$ of models
$$\Lambda_{\mathbb{M}} := \bigcap_{\mathfrak{M} \in \mathbb{M}} \{\varphi \mid \mathfrak{M} \models \varphi\}.$$
There are sets $\mathbb{M}$ for which $\Lambda_{\mathbb{M}}$ is not a modal language. In fact, take a model $\mathfrak{M}$ with world W and two propositional letters p, q with $V(p) = W$ and $V(q) \neq W$, then $\mathfrak{M}, w \models p$ for all w; hence $\mathfrak{M} \models p$, but $\mathfrak{M} \not\models q$. On the other hand, $q = p^\sigma$ under the substitution $\sigma : p \mapsto q$. Hence $\Lambda_{\{\mathfrak{M}\}}$ is not closed under uniform substitution.

✌

This formalizes the notion of deduction:

Definition 2.7.37 *Let Λ be a logic, and $\Gamma \cup \{\varphi\}$ a set of modal formulas.*

- φ *is* deducible *in Λ from Γ iff either $\vdash_\Lambda$ or if there exist formulas $\psi_1, \ldots, \psi_k \in \Gamma$ such that $\vdash_\Lambda (\psi_1 \wedge \ldots \wedge \psi_k) \to \varphi$. We write this down as $\Gamma \vdash_\Lambda \varphi$.* $\Gamma \vdash_\Lambda \varphi$
- Γ *is* Λ-consistent *iff $\Gamma \nvdash_\Lambda \perp$; otherwise, Γ is called* Λ-inconsistent.
- *φ is called Λ-consistent iff $\{\varphi\}$ is Λ-consistent.*

This is a simple and intuitive criterion for inconsistency. We fix for the discussions below a modal logic Λ.

Lemma 2.7.38 *Let Γ be a set of formulas. Then these statements are equivalent:*

1. *Γ is Λ-inconsistent.*
2. *$\Gamma \vdash_\Lambda \varphi \wedge \neg\varphi$ for some formula φ.*
3. *$\Gamma \vdash_\Lambda \psi$ for all formulas ψ.*

Proof 1 $\Rightarrow$ 2: Because $\Gamma \vdash_\Lambda \perp$, we know that $\psi_1 \wedge \ldots \wedge \psi_k \to \perp$ is in Λ for some formulas $\psi_1, \ldots, \psi_k \in \Gamma$. But $\perp \to \varphi \wedge \neg\varphi$ is a tautology; hence $\Gamma \vdash_\Lambda \varphi \wedge \neg\varphi$.

2 $\Rightarrow$ 3: By assumption there exists $\psi_1, \ldots, \psi_k \in \Gamma$ such that $\vdash_\Lambda \psi_1 \wedge \ldots \wedge \psi_k \to \varphi \wedge \neg\varphi$, and $\varphi \wedge \neg\varphi \to \psi$ is a tautology for an arbitrary formula ψ; hence $\vdash_\Lambda \varphi \wedge \neg\varphi \to \psi$. Thus $\Gamma \vdash_\Lambda \psi$.

3 $\Rightarrow$ 1: We have in particular $\Gamma \vdash_\Lambda \perp$. $\dashv$

Λ-consistent sets have an interesting compactness property.

Lemma 2.7.39 *A set Γ of formulas is Λ-consistent iff each finite subset in Γ is Λ-consistent.*

Proof If Γ is Λ-consistent, then certainly each finite subset is. If, on the other hand, each finite subset is Λ-consistent, then the whole set must be consistent, since consistency is tested with finite witness sets. $\dashv$

Proceeding on our path to finding a model for a modal logic, we define normal logics. These logics are closed under some properties which appear as fairly, well, normal, so it is not surprising that they will play an important rôle.

Normal

Definition 2.7.40 *Modal logic Λ is called* normal *iff it satisfies these conditions for all propositional letters $p, q \in \Phi$ and all formulas φ:*

(K) $\vdash_\Lambda \Box(p \to q) \to (\Box p \to \Box q)$,

(D) $\vdash_\Lambda \Diamond p \leftrightarrow \neg\Box\neg p$,

(G) *If* $\vdash_\Lambda \varphi$, *then* $\vdash_\Lambda \Box\varphi$.

Property **(K)** states that if it is necessary that p implies q, then the fact that p is necessary will imply that q is necessary. Note that the formulas in Λ do not have a semantics yet; they are for the time being just syntactic entities. Property **(D)** connects the constructors $\Diamond$ and $\Box$ in the desired manner. Finally, **(G)** states that, loosely speaking, if something is the case, then it is necessarily the case. We should finally note that **(K)** and **(D)** are both formulated for propositional letters only. This, however, is sufficient for modal logics, since they are closed under uniform substitution.

In a normal logic, the equivalence of formulas is preserved by the diamond.

Lemma 2.7.41 *Let Λ be a normal modal logic, then $\vdash_\Lambda \varphi \leftrightarrow \psi$ implies $\vdash_\Lambda \Diamond\varphi \leftrightarrow \Diamond\psi$.*

Proof We show that $\vdash_\Lambda \varphi \to \psi$ implies $\vdash_\Lambda \Diamond\varphi \to \Diamond\psi$, the rest will follow in the same way.

$$\begin{aligned}
\vdash_\Lambda \varphi \to \psi &\Rightarrow \vdash_\Lambda \neg\psi \to \neg\varphi && \text{(contraposition)}\\
&\Rightarrow \vdash_\Lambda \Box(\neg\psi \to \neg\varphi) && \text{(by } \mathbf{(G)}\text{)}\\
&\Rightarrow \vdash_\Lambda (\Box(\neg\psi \to \neg\varphi)) && \text{(uniform substitution, } \mathbf{(K)}\text{)}\\
&\to (\Box\neg\psi \to \Box\neg\varphi) && \\
&\Rightarrow \vdash_\Lambda \Box\neg\psi \to \Box\neg\varphi && \text{(modus ponens)}\\
&\Rightarrow \vdash_\Lambda \neg\Box\neg\varphi \to \neg\Box\neg\psi && \text{(contraposition)}\\
&\Rightarrow \vdash_\Lambda \Diamond\varphi \to \Diamond\psi && \text{(by } \mathbf{(D)}\text{)}
\end{aligned}$$

$\dashv$

$\mathfrak{F} \models \Gamma$

We define a semantic counterpart to $\Gamma \vdash_\Lambda$ now. Let $\mathfrak{F}$ be a frame and Γ be a set of formulas; then we say that Γ holds on $\mathfrak{F}$ (written as $\mathfrak{F} \models \Gamma$) iff each formula in Γ holds in each model which is based on frame $\mathfrak{F}$ (see Example 2.7.36). We say that Γ entails formula φ ($\Gamma \models_\mathfrak{F} \varphi$) iff $\mathfrak{F} \models \Gamma$ implies $\mathfrak{F} \models \varphi$. This carries over to classes of frames in an obvious way. Let $\mathbb{C}$ be a class of frames; then $\Gamma \models_{\mathbb{C}} \varphi$ iff we have $\Gamma \models_{\mathfrak{F}} \varphi$ for all frames $\mathfrak{F} \in \mathbb{C}$.

Definition 2.7.42 *Let $\mathbb{C}$ be a class of frames, then the normal logic Λ is called* $\mathbb{C}$-sound *iff $\Lambda \subseteq \Lambda_{\mathbb{C}}$. If Λ is $\mathbb{C}$-sound, then $\mathbb{C}$ is called a* class of frames for Λ.

Note that $\mathbb{C}$-soundness indicates that $\vdash_{\Lambda} \varphi$ implies $\mathfrak{F} \models \varphi$ for all frames $\mathfrak{F} \in \mathbb{C}$ and for all formulas φ.

This example dwells on traditional names.

Example 2.7.43 Let Λ_4 be the smallest modal logic which contains $\Diamond \Diamond p \rightarrow \Diamond p$ (if it is possible that p is possible, then p is possible), and let $K4$ be the class of transitive frames. Then Λ_4 is $K4$-sound. In fact, it is easy to see that $\mathfrak{M}, w \models \Diamond \Diamond p \rightarrow \Diamond p$ for all worlds w, whenever $\mathfrak{M}$ is a model, the frame of which carries a transitive relation. ✌

Thus $\mathbb{C}$-soundness permits us to conclude that a formula which is deducible from Γ holds also in all frames from $\mathbb{C}$. Completeness goes the other way: roughly, if we know that a formula holds in a class of frames, then it is deducible. To be more precise:

Definition 2.7.44 *Let $\mathbb{C}$ be a class of frames and Λ a normal modal logic.*

1. *Λ is* strongly $\mathbb{C}$-complete *iff for any set $\Gamma \cup \{\varphi\}$ of formulas $\Gamma \models_{\mathbb{C}} \varphi$ implies $\Gamma \vdash_{\Lambda} \varphi$.*
2. *Λ is* weakly $\mathbb{C}$-complete *iff $\mathbb{C} \models \varphi$ implies $\vdash_{\Lambda} \varphi$ for any formula φ.*

This is a characterization of completeness.

Proposition 2.7.45 *Let Λ and $\mathbb{C}$ be as above.*

1. *Λ is strongly $\mathbb{C}$-complete iff every Λ-consistent set of formulas is satisfiable for some $\mathfrak{F} \in \mathbb{C}$.*
2. *Λ is weakly $\mathbb{C}$-complete iff every Λ-consistent formula is satisfiable for some $\mathfrak{F} \in \mathbb{C}$.*

Proof 1. If Λ is not strongly $\mathbb{C}$-complete, then we can find a set Γ of formulas and a formula φ with $\Gamma \models_{\mathbb{C}} \varphi$, but $\Gamma \not\vdash_{\Lambda} \varphi$. Then $\Gamma \cup \{\neg\varphi\}$ is Λ-consistent, but this set cannot be satisfied on $\mathbb{C}$. So the condition for strong completeness is sufficient. It is also necessary. In fact, we may assume by compactness that Γ is finite. Thus by consistency $\Gamma \nvdash_\Lambda \bot$; hence $\Gamma \not\models_{\mathbb{C}} \bot$ by completeness, and thus there exists a frame $\mathfrak{F} \in \mathbb{C}$ with $\mathfrak{F} \models \Gamma$ but $\mathfrak{F} \not\models \bot$.

2. This is but a special case of cardinality 1. $\dashv$

Consistent sets are not yet sufficient for the construction of a model, as we will see soon. We need consistent sets which cannot be extended further without jeopardizing their consistency. To be specific:

Definition 2.7.46 *The set Γ of formulas is* maximally Λ-consistent *iff Γ is Λ-consistent, and it is not properly contained in a Λ-consistent set.*

Thus if we have a maximal Λ-consistent set Γ, and if we know that $\Gamma \subset \Gamma_0$ with $\Gamma \neq \Gamma_0$, then we know that Γ_0 is not Λ-consistent. This criterion is sometimes a bit unpractical, but we have:

Lemma 2.7.47 *Let Λ be a normal logic and Γ be a maximally Λ-consistent set of formulas. Then:*

1. *Γ is closed under modus ponens.*
2. *$\Lambda \subseteq \Gamma$.*
3. *$\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$ for all formulas φ.*
4. *$\varphi \vee \psi \in \Gamma$ iff $\varphi \in \Gamma$ or $\psi \in \Gamma$ for all formulas φ, ψ.*
5. *$\varphi_1 \wedge \varphi_2 \in \Gamma$ if $\varphi_1, \varphi_2 \in \Gamma$.*

Proof 1. Assume that $\varphi \in \Gamma$ and $\varphi \to \psi \in \Gamma$, but $\psi \notin \Gamma$. Then $\Gamma \cup \{\psi\}$ is inconsistent; hence $\Gamma \cup \{\psi\} \vdash_\Lambda \bot$ by Lemma 2.7.38. Thus we can find formulas $\psi_1, \ldots, \psi_k \in \Gamma$ such that $\vdash_\Lambda \psi \wedge \psi_1 \wedge \ldots \wedge \psi_k \to \bot$. Because $\vdash_\Lambda \varphi \wedge \psi_1 \wedge \ldots \wedge \psi_k \to \psi \wedge \psi_1 \wedge \ldots \wedge \psi_k$, we conclude $\Gamma \vdash_\Lambda \bot$. This contradicts Λ-consistency by Lemma 2.7.38.

2. In order to show that $\Lambda \subseteq \Gamma$, we assume that there exists $\psi \in \Lambda$ such that $\psi \notin \Gamma$, then $\Gamma \cup \{\psi\}$ is inconsistent; hence $\vdash_\Lambda \psi_1 \wedge \ldots \wedge \psi_k \to \neg\psi$ for some $\psi_1, \ldots, \psi_k \in \Lambda$ (here we use $\Gamma \cup \{\psi\} \vdash_\Lambda \psi$ and Lemma 2.7.38). By propositional logic, $\vdash_\Lambda \psi \to \neg(\psi_1 \wedge \ldots \wedge \psi_k)$; hence $\psi \in \Lambda$ implies $\Gamma \vdash_\Lambda \neg(\psi_1 \wedge \ldots \wedge \psi_k)$. But $\Gamma \vdash_\Lambda \psi_1 \wedge \ldots \wedge \psi_k$, consequently, Γ is Λ-inconsistent.

3. If both $\varphi \notin \Gamma$ and $\neg\varphi \notin \Gamma$, Γ is Λ-inconsistent.

4. Assume first that $\varphi \vee \psi \in \Gamma$, but $\varphi \notin \Gamma$ and $\psi \notin \Gamma$; hence both $\Gamma \cup \{\varphi\}$ and $\Gamma \cup \{\psi\}$ are inconsistent. Thus we can find $\psi_1, \ldots, \psi_k, \varphi_1, \ldots, \varphi_n \in \Gamma$ with $\vdash_\Lambda \psi_1 \wedge \ldots \wedge \psi_k \to \neg\psi$ and $\vdash_\Lambda \varphi_1 \wedge \ldots \wedge \varphi_n \to \neg\varphi$. This implies $\vdash_\Lambda \psi_1 \wedge \ldots \wedge \psi_k \wedge \varphi_1 \wedge \ldots \wedge \varphi_n \to \neg\psi \wedge \neg\varphi$, and by arguing propositionally, $\vdash_\Lambda (\psi \vee \varphi) \wedge \psi_1 \wedge \ldots \wedge \psi_k \wedge \varphi_1 \wedge \ldots \wedge \varphi_n \to \bot$, which contradicts Λ-consistency of Γ. For the converse, assume that $\varphi \in \Gamma$. Since $\varphi \to \varphi \vee \psi$ is a tautology, we obtain $\varphi \vee \psi$ from modus ponens.

5. Assume $\varphi_1 \wedge \varphi_2 \notin \Gamma$, then $\neg\varphi_1 \vee \neg\varphi_2 \in \Gamma$ by part 3. Thus $\neg\varphi_1 \in \Gamma$ or $\neg\varphi_2 \in \Gamma$ by part 4; hence $\varphi_1 \notin \Gamma$ or $\varphi_2 \notin \Gamma$. $\dashv$

Hence consistent sets have somewhat convenient properties; they remind the reader probably of the properties of an ultrafilter in Lemma 1.5.35, and we will use this close relationship when establishing Gödel's Completeness Theorem in Sect. 3.6.1. But how do we construct them? The famous Lindenbaum Lemma states that we may obtain them by enlarging consistent sets.

From now on we fix a normal modal logic Λ.

Lemma 2.7.48 *If Γ is a Λ-consistent set, then there exists a maximal Λ-consistent set Γ^+ with $\Gamma \subseteq \Gamma^+$.*

We will give two proofs for the Lindenbaum Lemma, depending on the cardinality of the set of all formulas. If the set Φ of propositional letters is countable, the set of all formulas is countable as well, so the first proof may be applied. If, however, we have more than a countable number of formulas, then this proof will fail to exhaust all formulas, and we have to apply another method, in this case transfinite induction. The basic idea, however, is the same in each case. Based on the observation that either φ or $\neg\varphi$ is a member of a maximally consistent set, we take the consistent set presented to us and look at each formula. If adding the formula will leave the set consistent, then we add it; otherwise, we add its negation. If the set of all formulas is not countable, the process of adding formulas will be controlled by a well-ordering, i.e., through transfinite induction in the form of Tuckey's Lemma, if it is countable, however, life is easier and we may access the set of all formulas through enumerating them. Plan

Proof (First—countable case) Assume that the set of all formulas is countable, and let $\{\varphi_n \mid n \in \mathbb{N}\}$ be an enumeration of them. Define by induction

$$\begin{aligned} \Gamma_0 &:= \Gamma, \\ \Gamma_{n+1} &:= \Gamma_n \cup \{\psi_n\}, \end{aligned}$$

where

$$\psi_n := \begin{cases} \varphi_n, & \text{if } \Gamma_n \cup \{\varphi_n\} \text{ is consistent,} \\ \neg\varphi_n, & \text{otherwise.} \end{cases}$$

Put

$$\Gamma^+ := \bigcup_{n \in \mathbb{N}} \Gamma_n.$$

Then these properties are easily checked:

- Γ_n is consistent for all $n \in \mathbb{N}_0$.
- Either $\varphi \in \Gamma^+$ or $\neg\varphi \in \Gamma^+$ for all formulas φ.
- If $\Gamma^+ \vdash_\Lambda \varphi$, then $\varphi \in \Gamma^+$.
- Γ^+ is maximal.

$\dashv$

It may be noted that this proof is fairly similar to the second proof of the Compactness Theorem 1.5.8 for propositional logic in Sect. 1.5.1, suggested for the countable case.

Proof (Second—general case) Let

$$\mathbb{C} := \{\Gamma' \mid \Gamma' \text{ is } \Lambda\text{-consistent and } \Gamma \subseteq \Gamma'\}.$$

Then $\mathbb{C}$ contains Γ; hence $\mathbb{C} \neq \emptyset$, and $\mathbb{C}$ is ordered by inclusion. By Tuckey's Lemma, it contains a maximal chain $\mathbb{C}_0$. Let $\Gamma^+ := \bigcup \mathbb{C}_0$. Then Γ^+ is a Λ-consistent set which contains Γ as a subset. While the latter is evident, we have to take care of the former. Assume that Γ^+ is not Λ-consistent; hence $\Gamma^+ \vdash_\Lambda \varphi \wedge \neg\varphi$ for some formula φ. Thus we find $\psi_1, \dots, \psi_k \in \Gamma^+$ with $\vdash_\Lambda \psi_1 \wedge \dots \wedge \psi_k \rightarrow \varphi \wedge \neg\varphi$. Given $\psi_i \in \Gamma^+$, we find $\Gamma_i \in \mathbb{C}_0$ with $\psi_i \in \Gamma_i$. Since $\mathbb{C}_0$ is linearly ordered, we find some Γ' among them such that $\Gamma_i \subseteq \Gamma'$ for all i. Hence $\psi_1, \dots, \psi_k \in \Gamma'$, so that Γ' is not Λ-consistent. This is a contradiction. Now assume that Γ^+ is not maximal, then there exists φ such that

$\varphi \notin \Gamma^+$ and $\neg\varphi \notin \Gamma^+$. If $\Gamma^+ \cup \{\varphi\}$ is not consistent, $\Gamma^+ \cup \{\neg\varphi\}$ is, and vice versa, so either one of $\Gamma^+ \cup \{\varphi\}$ and $\Gamma^+ \cup \{\neg\varphi\}$ is consistent. But this means that $\mathbb{C}_0$ is not maximal. ⊣

We are in a position to construct a model now, specifically, we will define a set of states, a transition relation and the validity sets for the propositional letters. Put

$$\begin{aligned}
W^\sharp &:= \{\Sigma \mid \Sigma \text{ is } \Lambda\text{-consistent and maximal}\},\\
R^\sharp &:= \{\langle w, v\rangle \in W^\sharp \times W^\sharp \mid \text{for all formulas } \psi, \psi \in v\\
&\qquad \text{implies } \Diamond\,\psi \in w\},\\
V^\sharp(p) &:= \{w \in W^\sharp \mid p \in w\} \text{ for } p \in \Phi.
\end{aligned}$$

Then $\mathfrak{M}^\sharp := (W^\sharp, R^\sharp, V^\sharp)$ is called the *canonical model* for Λ.

This is another view of relation $R^\sharp$:

Lemma 2.7.49 *Let $v, w \in W^\sharp$, then $wR^\sharp v$ iff $\Box\psi \in w$ implies $\psi \in v$ for all formulas ψ.*

Proof 1. Assume that $\langle w, v\rangle \in R^\sharp$ but that $\psi \notin v$ for some formula ψ. Since v is maximal, we conclude from Lemma 2.7.47 that $\neg\psi \in v$; hence the definition of $R^\sharp$ tells us that $\Diamond\neg\psi \in w$, which in turn implies by the maximality of w that $\neg\Diamond\neg\psi \notin w$. Thus $\Box\psi \notin w$ follows.

2. If $\Diamond\psi \notin w$, then by maximality $\neg\Diamond\psi \in w$, so $\Box\neg\psi \in w$, which means by assumption that $\neg\psi \in v$. Hence $\psi \notin v$. ⊣

The next lemma gives a more detailed look at the transitions which are modeled by $R^\sharp$.

Lemma 2.7.50 *Let $w \in W^\sharp$ with $\Diamond\varphi \in w$. Then there exists a state $v \in W^\sharp$ such that $\varphi \in v$ and $w\ R^\sharp\ v$.*

Proof 0. Because we can extend Λ-consistent sets to maximal consistent ones by the Lindenbaum Lemma 2.7.48, it is enough to show that $v_0 := \{\varphi\} \cup \{\psi \mid \Box\psi \in w\}$ is Λ-consistent. If this succeeds, we extend the set v_0 to obtain v. Plan

1. Assume it is not. Then we have $\vdash_\Lambda (\psi_1 \wedge \ldots \wedge \psi_k) \rightarrow \neg\varphi$ for some $\psi_1, \ldots, \psi_k \in v_0$, from which we obtain with **(G)** and **(K)** that $\vdash_\Lambda \Box(\psi_1 \wedge \ldots \wedge \psi_k) \rightarrow \Box\neg\varphi$. Because $\Box\psi_1 \wedge \ldots \wedge \Box\psi_k \rightarrow \Box(\psi_1 \wedge \ldots \wedge \psi_k)$, this implies $\vdash_\Lambda \Box\psi_1 \wedge \ldots \wedge \Box\psi_k \rightarrow \Box\neg\varphi$. Since

$\Box\psi_1, \ldots, \Box\psi_k \in w$, we conclude from Lemma 2.7.47 that $\Box\psi_1 \wedge \ldots \wedge \Box\psi_k \in w$; thus we have $\Box\neg\varphi \in w$ by modus ponens; hence $\neg\Diamond\varphi \in w$. Since w is maximal, this implies $\Diamond\varphi \notin w$. But this is a contradiction. So v_0 is consistent; thus there exists by the Lindenbaum Lemma a maximal consistent set v with $v_0 \subseteq v$. We have in particular $\varphi \in v$, and we know that $\Box\psi \in w$ implies $\psi \in v$; hence $\langle w, v\rangle \in R^\sharp$. ⊣

This helps in characterizing the model, in particular the validity relation $\models$ by the well-known Truth Lemma.

Lemma 2.7.51 $\mathfrak{M}^\sharp, w \models \varphi$ *iff* $\varphi \in w$.

Plan

Proof The proof proceeds by induction on formula φ. The statement is trivially true if $\varphi = p \in \Phi$ is a propositional letter. The set of formulas for which the assertion holds is certainly closed under Boolean operations, so the only interesting case is the case that the formula in question has the shape $\Diamond\varphi$ and that the assertion is true for φ.

"⇒": If $\mathfrak{M}^\sharp, w \models \Diamond\varphi$, then we can find some v with $w\ R^\sharp\ v$ and $\mathfrak{M}^\sharp, v \models \varphi$. Thus there exists v with $\langle w, v\rangle \in R^\sharp$ such that $\varphi \in v$ by hypothesis, which in turn means $\Diamond\varphi \in w$.

"⇐": Assume $\Diamond\varphi \in w$; hence there exists $v \in W^\sharp$ with $w\ R^\sharp\ v$ and $\varphi \in v$; thus $\mathfrak{M}^\sharp, v \models \varphi$. But this means $\mathfrak{M}^\sharp, w \models \Diamond\varphi$.

⊣

Finally, we obtain:

Theorem 2.7.52 *Any normal logic is complete with respect to its canonical model.*

Proof Let Σ be a Λ-consistent set for the normal logic Λ. Then there exists by Lindenbaum's Lemma 2.7.48 a maximal Λ-consistent set Σ^+ with $\Sigma \subseteq \Sigma^+$. By the Truth Lemma, we have now $\mathfrak{M}^\sharp, \Sigma^+ \models \Sigma$. ⊣

We will generalize modal logics now to coalgebraic logics, which is particularly streamlined to be interpreted within the context of coalgebras and which contains the interpretation of modal logics as a special case. Neighborhood models may be captured in this realm as well, which indicates that the coalgebraic way of thinking about logics is a useful generalization of the relational way.

2.7.3 Coalgebraic Logics

We have seen several points where coalgebras and modal logics touch each other, for example, morphisms for Kripke models are based on morphisms for the underlying $\mathcal{P}$-coalgebra, as a comparison of Example 2.1.10 and Lemma 2.7.25 demonstrates. Let $\mathfrak{M} = (W, R, V)$ be a Kripke model; then the accessibility relation $R \subseteq W \times W$ can be perceived as a map, again denoted by R, with the signature $W \to \mathcal{P}(W)$. Map $V : \Phi \to \mathcal{P}(W)$, which indicates the validity of atomic propositions, can be encoded through a map $V_1 : W \to \mathcal{P}(\Phi)$ upon setting $V_1(w) := \{p \in \Phi \mid w \in V(p)\}$. Both V and V_1 describe the same relation $\{\langle p, w\rangle \in \Phi \times W \mid \mathfrak{M}, w \models p\}$, albeit from different angles. They are trivially interchangeable for each other. This new representation has the advantage of describing the model from the vantage point w.

Define $\boldsymbol{F}X := \mathcal{P}(X) \times \mathcal{P}(\Phi)$ for the set X, and put, given map $f : X \to Y$, $(\boldsymbol{F}f)(A, Q) := \langle f\left[A\right], Q\rangle = \langle (\mathcal{P}f)A, Q\rangle$ for $A \subseteq X, Q \subseteq \Phi$; then $\boldsymbol{F}$ is an endofunctor on $\boldsymbol{Set}$. Hence we obtain from the Kripke model $\mathfrak{M}$ the $\boldsymbol{F}$-coalgebra (W, γ) with $\gamma(w) := R(w) \times V_1(w)$. This construction can easily be reversed: given a $\boldsymbol{F}$-coalgebra (W, γ), we put $R(w) := \pi_1(\gamma(w))$ and $V_1(w) := \pi_2(\gamma(w))$ and construct V from V_1; then (W, R, V) is a Kripke model (here π_1, π_2 are the projections). Thus Kripke models and $\boldsymbol{F}$-coalgebras are in a one-to-one correspondence with each other. This correspondence goes a bit deeper, as can be seen when considering morphisms.

Proposition 2.7.53 *Let $\mathfrak{M} = (W, R, V)$ and $\mathfrak{N} = (X, S, Y)$ be Kripke models with associated $\mathfrak{F}$-coalgebras (W, γ) resp. (X, δ). Then these statements are equivalent for a map $f : W \to X$:*

1. *$f : (W, \gamma) \to (X, \delta)$ is a morphism of coalgebras.*
2. *$f : \mathfrak{M} \to \mathfrak{N}$ is a morphism of Kripke models.*

Proof 1 $\Rightarrow$ 2: We obtain for each $w \in W$ from the defining equation $(\boldsymbol{F}f) \circ \gamma = \delta \circ f$ the equalities $f\left[R(w)\right] = S(f(w))$, and $V_1(w) = Y_1(f(w))$. Since $f\left[R(w)\right] = (\mathcal{P}f)(R(w))$, we conclude that $(\mathcal{P}f) \circ R = S \circ f$, so f is a morphism of the $\mathcal{P}$-coalgebras. We have moreover for each atomic sentence $p \in \Phi$

$$w \in V(p) \Leftrightarrow p \in V_1(w) \Leftrightarrow p \in Y_1(f(w)) \Leftrightarrow f(w) \in Y(p).$$

This means $V = f^{-1} \circ Y$, so that $f : \mathfrak{M} \to \mathfrak{N}$ is a morphism.

2 $\Rightarrow$ 1: Because we know that $S \circ f = (\mathcal{P} f) \circ R$, and because one shows as above that $V_1 = Y_1 \circ f$, we obtain for $w \in W$

$$\begin{aligned}(\delta \circ f)(w) &= \langle S(f(w)), Y_1(f(w))\rangle = \langle (\mathcal{P} f)(R(w)), V_1(w)\rangle \\ &= \big((\boldsymbol{F} f) \circ \gamma\big)(w).\end{aligned}$$

Hence $f : (W, \gamma) \to (X, \delta)$ is a morphism for the $\boldsymbol{F}$-coalgebras. $\dashv$

Given a world w, the value of $\gamma(w)$ represents the worlds which are accessible from w, making sure that the validity of the atomic propositions is maintained; recall that they are not affected by a transition. This information is to be extracted in a variety of ways. We need predicate liftings for this. Before we define them, we observe that the same mechanism works for neighborhood models.

Example 2.7.54 Let $\mathcal{N} = (W, N, V)$ be a neighborhood model. Define functor $\boldsymbol{G}$ by putting $\boldsymbol{G}(X) := \boldsymbol{V}(X) \times \mathcal{P}(\Phi)$ for sets, and if $f : X \to Y$ is a map, put $(\boldsymbol{G} f)(U, Q) := \langle (\boldsymbol{V} f)U, Q\rangle$. Then $\boldsymbol{G}$ is an endofunctor on $\boldsymbol{Set}$. The $\boldsymbol{G}$-coalgebra (W, ν) associated with $\mathcal{N}$ is defined through $\nu(w) := \langle N(w), V_1(w)\rangle$ (with V_1 defined through V as above).

Let $\mathcal{M} = (X, M, Y)$ be another neighborhood model with associated coalgebra (X, μ). Exactly the same proof as the one for Proposition 2.7.53 shows that $f : \mathcal{N} \to \mathcal{M}$ is a neighborhood morphism iff $f : (W, \nu) \to (X, \mu)$ is a coalgebra morphism. ✌

Proceeding to define predicate liftings, let $\mathcal{P}^{op} : \boldsymbol{Set} \to \boldsymbol{Set}$ be the contravariant power set functor, i.e., given the set X, $\mathcal{P}^{op}(X)$ is the power set $\mathcal{P}(X)$ of X, and if $f : X \to Y$ is a map, then $(\mathcal{P}^{op} f) : \mathcal{P}^{op}(Y) \to \mathcal{P}^{op}(Y)$ works as $B \mapsto f^{-1}\big[B\big]$.

Definition 2.7.55 *Given a (covariant) endofunctor $\boldsymbol{T}$ on $\boldsymbol{Set}$, a* predicate lifting *λ for $\boldsymbol{T}$ is a monotone natural transformation $\lambda : \mathcal{P}^{op} \to \mathcal{P}^{op} \circ T$.*

Interpret $A \in \mathcal{P}^{op}(X)$ as a predicate on X, then $\lambda_X(A) \in \mathcal{P}^{op}(\boldsymbol{T}X)$ is a predicate on $\boldsymbol{T}X$; hence λ_X lifts the predicate into the realm of functor $\boldsymbol{T}$; the requirement of naturalness is intended to reflect compatibility with morphisms, as we will see below. Thus a predicate lifting helps in specifying a requirement on the level of sets, which it then transports onto the level of those sets that are controlled by functor $\boldsymbol{T}$. Technically, this requirement means that this diagram commutes, whenever

$f : X \to Y$ is a map:

$$\begin{tikzcd} \mathcal{P}X \arrow[r, "\lambda_X"] & \mathcal{P}(\boldsymbol{T}X) \\ \mathcal{P}Y \arrow[u, "f^{-1}"] \arrow[r, "\lambda_Y"'] & \mathcal{P}(\boldsymbol{T}Y) \arrow[u, "(\boldsymbol{T}f)^{-1}"'] \end{tikzcd}$$

Hence we have $\lambda_X(f^{-1}\big[G\big]) = (\boldsymbol{T}f)^{-1}\big[\lambda_Y(G)\big]$ for any $G \subseteq Y$.

Finally, monotonicity says that $\lambda_X(D) \subseteq \lambda_X(E)$, whenever $D \subseteq E \subseteq X$; this condition models the requirement that information about states should only depend on their precursors. Informally it is reflected in the rule $\vdash (\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$.

This example illuminates the idea.

Example 2.7.56 Let $\boldsymbol{F} = \mathcal{P}(-) \times \mathcal{P}\Phi$ be defined as above; put for the set X and for $D \subseteq X$

$$\lambda_X(D) := \{\langle D', Q\rangle \in \mathcal{P}(X) \times \mathcal{P}(\Phi) \mid D' \subseteq D\}.$$

This defines a predicate lifting $\lambda : \mathcal{P}^{op} \to \mathcal{P}^{op} \circ \boldsymbol{F}$. In fact, let $f : X \to Y$ be a map and $G \subseteq Y$, then

$$\begin{aligned} \lambda_X(f^{-1}\big[G\big]) &= \{\langle D', Q\rangle \mid D' \subseteq f^{-1}\big[G\big]\} \\ &= \{\langle D', Q\rangle \mid f\big[D'\big] \subseteq G\} \\ &= (\boldsymbol{F}f)^{-1}\big[\{\langle G', Q\rangle \in \mathcal{P}(Y) \times \mathcal{P}(\Phi) \mid G' \subseteq G\}\big] \\ &= (\boldsymbol{F}f)^{-1}\big[\lambda_Y(G)\big] \end{aligned}$$

(remember that $\boldsymbol{F}f$ leaves the second component of a pair alone). It is clear that λ_X is monotone for each set X.

Let $\gamma : W \to \boldsymbol{F}W$ be the coalgebra associated with Kripke model $\mathfrak{M} := (W, R, V)$, and look at this ($\varphi$ is a formula):

$$\begin{aligned} w \in \gamma^{-1}\big[\lambda_W([\![\varphi]\!]_{\mathfrak{M}})\big] &\Leftrightarrow \gamma(w) \in \lambda_W([\![\varphi]\!]_{\mathfrak{M}}) \\ &\Leftrightarrow \langle R(w), V_1(w)\rangle \in \lambda_W([\![\varphi]\!]_{\mathfrak{M}}) \\ &\Leftrightarrow R(w) \subseteq [\![\varphi]\!]_{\mathfrak{M}} \\ &\Leftrightarrow w \in [\![\Box\varphi]\!]_{\mathfrak{M}} \end{aligned}$$

This means that we can describe the semantics of the $\Box$-operator through a predicate lifting, which cooperates with the coalgebra's dynamics.

Note that it would be equally possible to do this for the $\Diamond$-operator: define the lifting through $D \mapsto \{\langle D', Q\rangle \mid D' \cap D \neq \emptyset\}$. But we will stick to the $\Box$-operator, keeping up with tradition. ✌

Example 2.7.57 The same technique works for neighborhood models. In fact, let (W, ν) be the $\boldsymbol{G}$-coalgebra associated with neighborhood model $\mathcal{N} = (W, N, V)$ as in Example 2.7.54, and define

$$\lambda_X(D) := \{\langle V, Q\rangle \in \boldsymbol{V}(X) \times \mathcal{P}(\Phi) \mid D \in V\}.$$

Then $\lambda_X : \mathcal{P}(X) \to \mathcal{P}(\boldsymbol{V}(X) \times \mathcal{P}(\Phi))$ is monotone, because the elements of $\boldsymbol{V}X$ are upper closed. If $f : (W, \nu) \to (X, \mu)$ is a $\boldsymbol{G}$-coalgebra morphism, we obtain for $D \subseteq X$

$$\begin{aligned}
\lambda_W(f^{-1}\big[D\big]) &= \{\langle V, Q\rangle \in \boldsymbol{V}(W) \times \mathcal{P}(\Phi) \mid f^{-1}\big[D\big] \subseteq V\} \\
&= \{\langle V, Q\rangle \in \boldsymbol{V}(W) \times \mathcal{P}(\Phi) \mid D \in (\boldsymbol{V}f)(V)\} \\
&= (\boldsymbol{G}f)^{-1}\big[\{\langle V', Q\rangle \in \boldsymbol{V}(X) \times \mathcal{P}(\Phi) \mid D \in V'\}\big] \\
&= (\boldsymbol{G}f)^{-1}\big[\lambda_X(D)\big]
\end{aligned}$$

Consequently, λ is a predicate lifting for $\boldsymbol{G}$. We see also for formula φ

$$\begin{aligned}
w \in \lambda_W([\![\varphi]\!]_{\mathcal{N}}) &\Leftrightarrow \langle [\![\varphi]\!]_{\mathcal{N}}, V_1(w)\rangle \in \lambda_X([\![\varphi]\!]_{\mathcal{N}}) \\
&\Leftrightarrow [\![\varphi]\!]_{\mathcal{N}} \in N(w) && \text{(by definition of } \nu) \\
&\Leftrightarrow w \in [\![\Box\varphi]\!]_{\mathcal{N}}
\end{aligned}$$

Hence we can define the semantics of the $\Box$-operator also in this case through a predicate lifting. ✌

There is a general mechanism permitting us to define predicate liftings, which is outlined in the next lemma.

Lemma 2.7.58 *Let $\eta : \boldsymbol{T} \to \mathcal{P}$ be a natural transformation, and define*

$$\lambda_X(D) := \{c \in \boldsymbol{T}X \mid \eta_X(c) \subseteq D\}$$

for $D \subseteq X$. Then λ defines a predicate lifting for $\boldsymbol{T}$.

Proof It is clear from the construction that $D \mapsto \lambda_X(D)$ defines a monotone map, so we have to show that the diagram below is commutative for $f : X \to Y$.

$$\begin{tikzcd}
\mathcal{P}X \arrow[r, "\lambda_X"] & \mathcal{P}\boldsymbol{T}X \\
\mathcal{P}Y \arrow[r, "\lambda_Y"'] \arrow[u, "f^{-1}"] & \mathcal{P}\boldsymbol{T}Y \arrow[u, "(\boldsymbol{T}f)^{-1}"']
\end{tikzcd}$$

We note that

$$\eta_X(c) \subseteq f^{-1}\big[E\big] \Leftrightarrow f\big[\eta_X(c)\big] \subseteq E \Leftrightarrow (\mathcal{P}f)(\eta_X(c)) \subseteq E$$

and

$$(\mathcal{P}f) \circ \eta_X = \eta_Y \circ (\boldsymbol{T}f),$$

because η is natural. Hence we obtain for $E \subseteq Y$:

$$\begin{aligned}
\eta_X(f^{-1}\big[E\big]) &= \{c \in \boldsymbol{T}X \mid \eta_X(c) \subseteq f^{-1}\big[E\big]\} \\
&= \{c \in \boldsymbol{T}X \mid \big((\mathcal{P}f) \circ \eta_X\big)(c) \subseteq E\} \\
&= \{c \in \boldsymbol{T}X \mid (\eta_Y \circ \boldsymbol{T}f)(c) \subseteq E\} \\
&= (\boldsymbol{T}f)^{-1}\big[\{d \in \boldsymbol{T}Y \mid \eta_Y(d) \subseteq E\}\big] \\
&= \big((\boldsymbol{T}f)^{-1} \circ \eta_Y\big)(E).
\end{aligned}$$

$\dashv$

Let us return to the endofunctor $\boldsymbol{F} = \mathcal{P}(-) \times \mathcal{P}(\Phi)$ and fix for the moment an atomic proposition $p \in \Phi$. Define the constant function

$$\lambda_{p,X}(D) := \{\langle D', Q\rangle \in \boldsymbol{F}X \mid p \in Q\}.$$

Then an easy calculation shows that $\lambda_p : \mathcal{P}^{op} \to \mathcal{P}^{op} \circ \boldsymbol{F}$ is a natural transformation, hence a predicate lifting for $\boldsymbol{F}$. Let $\gamma : W \to \boldsymbol{F}W$ be a coalgebra with carrier W which corresponds to the Kripke model $\mathfrak{M} = (W, R, V)$; then

$$\begin{aligned}
w \in (\gamma^{-1} \circ \lambda_{p,W})(D) \quad &\Leftrightarrow \quad \gamma(w) \in \lambda_{p,W}(D) \Leftrightarrow p \in \pi_2(\gamma(w)) \\
&\Leftrightarrow \quad w \in V(p),
\end{aligned}$$

which means that we can use λ_p for expressing the meaning of formula $p \in \Phi$. A very similar construction can be made for functor $\boldsymbol{G}$, leading to the same conclusion.

We cast this into a more general framework now. Let $\ell_X : X \to \{0\}$ be the unique map from set X to the singleton set $\{0\}$. Given $A \subseteq \boldsymbol{T}(\{0\})$, define $\lambda_{A,X}(D) := \{c \in \boldsymbol{T}X \mid (\boldsymbol{T}\ell_X)(c) \in A\} = (\boldsymbol{T}\ell_X)^{-1}\big[A\big]$. This defines a predicate lifting for $\boldsymbol{T}$. In fact, let $f : X \to Y$ be a map; then $\ell_X = \ell_Y \circ f$, so

$$(\boldsymbol{T}f)^{-1} \circ (\boldsymbol{T}\ell_Y)^{-1} = \big((\boldsymbol{T}\ell_Y) \circ (\boldsymbol{T}f)\big)^{-1} = \big(\boldsymbol{T}(\ell_Y \circ f)\big)^{-1} = (\boldsymbol{T}\ell_X)^{-1},$$

hence

$$\lambda_{A,X}(f^{-1}[B]) = (\boldsymbol{T} f)^{-1}[\lambda_{A,Y}(B)].$$

As we have seen, this construction is helpful for capturing the semantics of atomic propositions.

Negation can be treated as well in this framework. Given a predicate lifting λ for $\boldsymbol{T}$, we define for the set X and $A \subseteq X$ the set

$$\lambda_X^{\neg}(A) := (\boldsymbol{T} X) \setminus \lambda_X(X \setminus A);$$

then this defines a predicate lifting for $\boldsymbol{T}$. This is easily checked: monotonicity of $\lambda^{\neg}$ follows from λ being monotone, and since f^{-1} is compatible with the Boolean operations, naturality follows.

Summarizing, those operations which are dear to us when interpreting modal logics through a Kripke model or through a neighborhood model can also be represented using predicate liftings.

We now take a family $\mathbb{L}$ of predicate liftings and define a logic for it.

$\mathcal{L}(\mathbb{L})$ **Definition 2.7.59** *Let $\boldsymbol{T}$ be an endofunctor on the category* ***Set*** *of sets with maps, and let $\mathbb{L}$ be a set of predicate listings for $\boldsymbol{T}$. The formulas for the language $\mathcal{L}(\mathbb{L})$ are defined through*

$$\varphi ::= \bot \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid [\lambda]\varphi$$

with $\lambda \in \mathbb{L}$.

The semantics of a formula in $\mathcal{L}(\mathbb{L})$ in a $\boldsymbol{T}$-coalgebra (W, γ) is defined recursively by describing the sets of worlds $[\![\varphi]\!]_\gamma$ in which formula φ holds (with $w \models_\gamma \varphi$ iff $w \in [\![\varphi]\!]_\gamma$): $w \models_\gamma \varphi$

$$\begin{aligned}
[\![\bot]\!]_\gamma &:= \emptyset \\
[\![\varphi_1 \wedge \varphi_2]\!]_\gamma &:= [\![\varphi_1]\!]_\gamma \cap [\![\varphi_2]\!]_\gamma \\
[\![\neg\varphi]\!]_\gamma &:= W \setminus [\![\varphi]\!]_\gamma \\
[\![[\lambda]\varphi]\!]_\gamma &:= (\gamma^{-1} \circ \lambda_C)([\![\varphi]\!]_\gamma).
\end{aligned}$$

The most interesting definition is of course the last one. It is defined through a modality for the predicate lifting λ, and it says that formula $[\lambda]\varphi$ holds in world w iff the transition $\gamma(w)$ achieves a state which is lifted by λ from one in which φ holds. Hence each successor to w satisfies the predicate for φ lifted by λ.

Example 2.7.60 Continuing Example 2.7.56, we see that the simple modal logic can be defined as the modal logic for $\mathbb{L} = \{\lambda\} \cup \{\lambda_p \mid p \in \Phi\}$, where λ is defined in Example 2.7.56, and λ_p are the constant liftings associated with Φ. ✌

We obtain also in this case the invariance of validity under morphisms.

Proposition 2.7.61 *Let $f : (W, \gamma) \to (X, \delta)$ be a $\boldsymbol{T}$-coalgebra morphism. Then*

$$w \models_\gamma \varphi \Leftrightarrow f(w) \models_\delta \varphi$$

holds for all formulas $\varphi \in \mathcal{L}(\mathbb{L})$ and all worlds $w \in W$.

Proof The proof proceeds by induction on φ, the interesting case occurring for a modal formula $[\lambda]\varphi$ with $\lambda \in \mathbb{L}$. So assume that the hypothesis is true for φ; then we have

$$\begin{aligned}
f^{-1}\big[[\![[\lambda]\varphi]\!]_\delta\big] &= \big((\delta \circ f)^{-1} \circ \lambda_D\big)([\![\varphi]\!]_\delta) \\
&= \big((\boldsymbol{T}(f) \circ \gamma)^{-1} \circ \lambda_D\big)([\![\varphi]\!]_\delta) && (f \text{ is a morphism}) \\
&= \big(\gamma^{-1} \circ (\boldsymbol{T} f)^{-1} \circ \lambda_D\big)([\![\varphi]\!]_\delta) \\
&= \big(\gamma^{-1} \circ \lambda_C \circ f^{-1}\big)([\![\varphi]\!]_\delta) && (\lambda \text{ is natural}) \\
&= \big(\gamma^{-1} \circ \lambda_C\big)([\![\varphi]\!]_\gamma) && (\text{by hypothesis}) \\
&= [\![[\lambda]\varphi]\!]_\gamma
\end{aligned}$$

$\dashv$

Let (C, γ) be a $\boldsymbol{T}$-coalgebra, then we define the *theory of c*: $Th_\gamma(c)$

$$Th_\gamma(c) := \{\varphi \in \mathcal{L}(\mathbb{L}) \mid c \models_\gamma \varphi\}$$

for $c \in C$. Two worlds which have the same theory cannot be distinguished through formulas of the logic $\mathcal{L}(\mathbb{L})$.

Definition 2.7.62 *Let (C, γ) and (D, δ) be $\boldsymbol{T}$-coalgebras, $c \in C$ and $d \in D$.*

- *We call the states c and d* logically equivalent *iff $Th_\gamma(c) = Th_\delta(d)$.*
- *The states c and d are called* behaviorally equivalent *iff there exists a $\boldsymbol{T}$-coalgebra (E, ϵ) and morphisms $(C, \gamma) \xrightarrow{f} (E, \epsilon) \xleftarrow{g} (D, \delta)$ such that $f(c) = g(d)$.* Logical, behavioral equivalence

Thus, logical equivalence looks locally at all the formulas which are true in a state and then compares two states with each other. Behavioral equivalence looks for an external instance, viz., a mediating coalgebra, and at morphisms; whenever we find states the image of which coincide, we know that the states are behaviorally equivalent.

This implication is fairly easy to obtain.

Proposition 2.7.63 *Behaviorally equivalent states are logically equivalent.*

Proof Let $c \in C$ and $d \in D$ be behaviorally equivalent for the $\boldsymbol{T}$-coalgebras (C, γ) and (D, δ), and assume that we have a mediating $\boldsymbol{T}$-coalgebra (E, ϵ) with morphisms

$$(C, \gamma) \xrightarrow{f} (E, \epsilon) \xleftarrow{g} (D, \delta).$$

and $f(c) = g(d)$. Then we obtain

$$\begin{aligned} \varphi \in Th_\gamma(c) \quad &\Leftrightarrow \quad c \models_\gamma \varphi \Leftrightarrow f(c) \models_\epsilon \varphi \Leftrightarrow g(d) \models_\epsilon \varphi \Leftrightarrow d \models_\delta \varphi \\ &\Leftrightarrow \quad \varphi \in Th_\delta(d) \end{aligned}$$

from Proposition 2.7.61. $\dashv$

We have seen that coalgebras are useful when it comes to generalize modal logics to coalgebraic logics. Morphisms arise in a fairly natural way in this context, giving rise to behaviorally equivalent coalgebras. It is quite clear that bisimilarity can be treated on this level as well, by introducing a mediating coalgebra and morphisms from it; bisimilar states are logically equivalent; the argument to show this is exactly as in the case above through Proposition 2.7.61. In each case, the question arises whether the implications can be reversed—are logically equivalent states behaviorally equivalent? Bisimilar? Answering this question requires a fairly elaborate machinery and depends strongly on the underlying functor. We will not discuss this question here but rather point to the literature, e.g., to [Pat04]. For the subprobability functor, some answers and some techniques can be found in [DS11].

The following example discusses the basic modal language with no atomic propositions.

Example 2.7.64 We interpret $\mathcal{L}(\{\Diamond\})$ with $\Phi = \emptyset$ through $\mathcal{P}$-coalgebras, i.e., through transition systems. Given a transition system (S, R), denote by $\sim$ the equivalence provided by logical equivalence, so that $s \sim s'$ iff states s and s' cannot be separated through a formula in the logic, i.e., iff $Th_R(s) = Th_R(s')$. Then $\eta_\sim : (S, R) \to (S/\sim, R/\sim)$ is a coalgebra morphism. Here

$$R/\sim := \{\langle [s_1], [s_2] \rangle \mid \langle s_1, s_2 \rangle \in R\}.$$

In fact, look at this diagram:

$$\begin{tikzcd} S \arrow[r, "\eta_\sim"] \arrow[d, "R"'] & S/\sim \arrow[d, "R/\sim"] \\ \mathcal{P}(S) \arrow[r, "\mathcal{P}(\eta_\sim)"'] & \mathcal{P}(S/\sim) \end{tikzcd}$$

Then

$$[s_2] \in R/\sim([s_1]) \Leftrightarrow [s_2] \in \big\{[s] \mid s \in [s_1]\big\} = \mathcal{P}(\eta_\sim)\big(R/\sim([s_1])\big),$$

which means that the diagram commutes. We denote the factor model $(S/\sim, R/\sim)$ by (S', R') and denote the class of an element without an indication of the equivalence relation. It will be clear from the context from which set of worlds a state will be taken.

Call the transition systems (S, R) and (T, L) *logically equivalent* iff for each state in one system there exists a logically equivalent state in the other one. We carry over behavioral equivalence and bisimilarity from individual states to systems, taking the discussion for coalgebras in Sect. 2.6.1 into account. Call the transition systems (S, R) and (T, L) *behaviorally equivalent* iff there exists a transition system (U, M) with surjective morphisms

$$(S, R) \xrightarrow{f} (U, M) \xleftarrow{g} (T, L).$$

Finally, they are called *bisimilar* iff there exists a transition system (U, M) with surjective morphisms Bisimilar

$$(S, R) \xleftarrow{f} (U, M) \xrightarrow{g} (T, L).$$

We claim that logical equivalent transition systems have isomorphic factor spaces under the equivalence induced by the logic, provided both are image finite. Consider this diagram:

$$\begin{tikzcd} S' \arrow[r, "\zeta"] \arrow[d, "R'"'] & T' \arrow[d, "L'"] \\ \mathcal{P}(S') \arrow[r, "\mathcal{P}(\zeta)"'] & \mathcal{P}(T') \end{tikzcd}$$

with $\zeta([s]) := [t]$ iff $Th_R(s) = Th_L(t)$. Thus ζ preserves classes of logically equivalent states. It is immediate that $\zeta : S' \to T'$ is a bijection, so commutativity has to be established.

Before working on the diagram, we show first that for any $\langle t, t'\rangle \in L$ and for any $s \in S$ with $Th_R(s) = Th_L(t)$, there exists $s' \in S$ with $\langle s, s'\rangle \in R$ and $Th_R(s') = Th_L(t')$ (written more graphically in terms of arrows, we claim that $t \to_L t'$ and $Th_R(s) = Th_L(t)$ together imply the existence of s' with $s \to_R s'$ and $Th_R(s') = Th_L(t')$). This is established by adapting the idea from the proof of the Hennessy–Milner Theorem 2.7.32 to the situation at hand. Well, then: Assume that such a state s' cannot be found. Since $L, t' \models \top$, we know that $L, t \models \Diamond\top$; thus $Th_L(t) = Th_R(s) \neq \emptyset$. Let $R(s) = \{s_1, \ldots, s_k\}$ for some $k \geq 1$, then we can find for each s_i a formula ψ_i with $L, t' \models \psi_i$ and $R, s_i \not\models \psi_i$. Thus $L, t \models \Diamond(\psi_1 \wedge \ldots \wedge \psi_k)$, but $R, s \not\models \Diamond(\psi_1 \wedge \ldots \wedge \psi_k)$, which contradicts the assumption that $Th_R(s) = Th_L(t)$. This uses only image finiteness of (S, R), by the way.

Now let $s \in S$ with $[t_1] \in L'\big(\zeta([s])\big) = L'\big([t]\big)$ for some $t \in T$. Thus $\langle t, t_1\rangle \in L$, so we find $s_1 \in S$ with $Th_R(s_1) = Th_L(t_1)$ and $\langle s, s_1\rangle \in R$. Consequently, $[t_1] = \zeta([s_1]) \in \mathcal{P}(\zeta)\big(R'([s])\big)$. Hence $L'\big(\zeta([s])\big) \subseteq \mathcal{P}(\zeta)\left(R'([s])\right)$.

Working on the other inclusion, we take $[t_1] \in \mathcal{P}(\zeta)\left(R'([s]\right)$, and we want to show that $[t_1] \in L'\big(\zeta([s])\big)$. Now $[t_1] = \zeta([s_1])$ for some $s_1 \in S$ with $\langle s, s_1\rangle \in R$; hence $Th_R(s_1) = Th_L(t_1)$. Put $[t] = \zeta([s])$; thus $Th_R(s) = Th_L(t)$. Because (T, L) is image finite as well, we may conclude from the Hennessy–Milner argument above—by interchanging the rôles of the transition systems—that we can find $t_2 \in T$ with $\langle t, t_2\rangle \in L$ so that $Th_L(t_2) = Th_R(s_1) = Th_L(t_1)$. This implies $[t_2] = [t_1]$ and $[t_1] \in L'([t]) = L'\big(\zeta([s])\big)$. Hence $L'\big(\zeta([s])\big) \supseteq \mathcal{P}(\zeta)\left(R'([s])\right)$.

Thus the diagram above commutes, and we have shown that the factor models are isomorphic. Consequently, two image finite transition systems which are logically equivalent are behaviorally equivalent with one of the factors acting as a mediating system. ✌

Clearly, behaviorally equivalent systems are bisimilar, so that we obtain these relationships:

$$\begin{tikzcd} & \text{bisimilarity} \arrow[dl, "\text{Proposition 2.7.27}"'] & \\ \text{logical equivalence} \arrow[rr, "\textit{image finiteness}"'] & & \text{behavioral equivalence} \arrow[ul, "\text{Theorem 2.6.16}"'] \end{tikzcd}$$

We finally give an idea of modeling `CTL⋆` as a popular logic for model checking coalgebraically. This shows how this modeling technique is applied, and it shows also that some additional steps become necessary, since things are not always straightforward.

Example 2.7.65 The logic `CTL⋆` is used for model checking [CGP99]. The abbreviation `CTL` stands for *computational tree logic*. `CTL⋆` is actually one of the simpler members of this family of tree logics used for this purpose, some of which involve continuous time [BHHK03, Dob07]. The logic has state formulas and path formulas; the former ones are used to describe a particular state in the system, and the latter ones express dynamic properties. Hence `CTL⋆` operates on two levels.

These operators are used:

State operators They include the operators $\boldsymbol{A}$ and $\boldsymbol{E}$, indicating that a property holds in a state iff it holds on all paths resp. on at least one path emanating from it.

Path operators They include the operators:

- $\boldsymbol{X}$ for *next time*—a property holds in the next, i.e., second state of a path,
- $\boldsymbol{F}$ for *in the future*—the specified property holds for some state on the path,
- $\boldsymbol{G}$ for *globally*—the property holds always on a path,
- $\boldsymbol{U}$ for *until*—this requires two properties as arguments; it holds on a path if there exists a state on the path for which the second property holds, and the first one holds on each preceding state.

State formulas are given through this syntax:

$$\varphi ::= \bot \mid p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \boldsymbol{E}\psi \mid \boldsymbol{A}\psi$$

with $p \in \Phi$ an atomic proposition and ψ a path formula. Path formulas are given through

$$\psi ::= \varphi \mid \neg\psi \mid \psi_1 \wedge \psi_2 \mid \boldsymbol{X}\psi \mid \boldsymbol{F}\psi \mid \boldsymbol{G}\psi \mid \psi_1 \boldsymbol{U} \psi_2$$

with φ a state formula. So both state and path formulas are closed under the usual Boolean operations; each atomic proposition is a state formula, and state formulas are also path formulas. Path formulas are closed under the operators $\boldsymbol{X}, \boldsymbol{F}, \boldsymbol{G}, \boldsymbol{U}$, and the operators $\boldsymbol{A}$ and $\boldsymbol{E}$ convert a path formula to a state formula.

Let W be the set of all states, and assume that $V : \Phi \to \mathcal{P}(W)$ assigns to each atomic formula the states for which it is valid. We assume also that we are given a transition relation $R \subseteq W \times W$; it is sometimes assumed [CGP99] that R is left total, but this is mostly for computational purposes, so we will not make this assumption here. Put

$$S := \{\langle w_1, w_2, \ldots\rangle \in W^{\mathbb{N}} \mid w_i \; R \; w_{i+1} \text{ for all } i \in \mathbb{N}\}$$

as the set of all infinite R-paths over W. The interpretation of formulas is then defined as follows:

State formulas Let $w \in W$, $\varphi, \varphi_1, \varphi_2$ be state formulas and ψ be a path formula, then $w \models \top$ holds always, and proceeding inductively, we put

$$\begin{aligned} s \models p &\Leftrightarrow w \in V(p) \\ w \models \neg\varphi &\Leftrightarrow w \models \varphi \text{ is false} \\ w \models \varphi_1 \wedge \varphi_2 &\Leftrightarrow w \models \varphi_1 \text{ and } w \models \varphi_2 \\ w \models \boldsymbol{E}\psi &\Leftrightarrow \sigma \models \psi \text{ for some path } \sigma \text{ starting from } w \\ w \models \boldsymbol{A}\psi &\Leftrightarrow \sigma \models \psi \text{ for all paths } \sigma \text{ starting from } w \end{aligned}$$

Path formulas Let $\sigma \in S$ be an infinite path with the first node σ_1; σ^k is the path with the first k nodes deleted; ψ is a path formula and φ a state formula; then

$$\begin{aligned} \sigma \models \varphi &\Leftrightarrow \sigma_1 \models \varphi \\ \sigma \models \neg\psi &\Leftrightarrow \sigma \models \psi \text{ is false} \\ \sigma \models \psi_1 \wedge \psi_2 &\Leftrightarrow \sigma \models \psi_1 \text{ and } \sigma \models \psi_2 \\ \sigma \models \boldsymbol{X}\psi &\Leftrightarrow \sigma^1 \models \psi \\ \sigma \models \boldsymbol{F}\psi &\Leftrightarrow \sigma^k \models \psi \text{ for some } k \geq 0 \end{aligned}$$

$$\sigma \models G\psi \Leftrightarrow \sigma^k \models \psi \text{ for all } k \geq 0$$
$$\sigma \models \psi_1 U \psi_2 \Leftrightarrow \exists k \geq 0 : \sigma^k \models \psi_2 \text{ and } \forall 0 \leq j < k : \sigma^j \models \psi_1.$$

Thus a state formula holds on a path iff it holds on the first node, $\boldsymbol{X}\psi$ holds on path σ iff ψ holds on σ with its first node deleted, and $\psi_1\boldsymbol{U}\psi_2$ holds on path σ iff ψ_2 holds on σ^k for some k, and iff ψ_1 holds on σ^i for all i preceding k.

We need to provide interpretations only for conjunction, negation, for $\boldsymbol{A}$, $\boldsymbol{X}$, and $\boldsymbol{U}$. This is so since $\boldsymbol{E}$ is the nabla of $\boldsymbol{A}$, $\boldsymbol{G}$ is the nabla of $\boldsymbol{F}$, and $\boldsymbol{F}\psi$ is equivalent to $(\neg\bot)\boldsymbol{U}\psi$. Conjunction and negation are easily interpreted, so we have to take care only of the temporal operators $\boldsymbol{A}$, $\boldsymbol{X}$, and $\boldsymbol{U}$.

A coalgebraic interpretation reads as follows. The $\mathcal{P}$-coalgebras together with their morphisms form a category $\boldsymbol{CoAlg}$. Let (X, R) be a $\mathcal{P}$-coalgebra, then

$$\boldsymbol{R}(X, R) := \{(x_n)_{n\in\mathbb{N}} \in X^\infty \mid x_n \; R \; x_{n+1} \text{ for all } n \in \mathbb{N}\}$$

is the object part of a functor, $(\boldsymbol{R}f)\big((x_n)_{n\in\mathbb{N}}\big) := (f(x_n)_{n\in\mathbb{N}})$ sends each coalgebra morphism $f : (X, R) \to (Y, S)$ to a map $\boldsymbol{R}f : \boldsymbol{R}(X, R) \to \boldsymbol{R}(Y, S)$, which maps $(x_n)_{n\in\mathbb{N}}$ to $f(x_n)_{n\in\mathbb{N}}$; recall that $x \; R \; x'$ implies $f(x) \; S \; f(x')$. Thus $\boldsymbol{R} : \boldsymbol{CoAlg} \to \boldsymbol{Set}$ is a functor. Note that the transition structure of the underlying Kripke model is already encoded through functor $\boldsymbol{R}$. This is reflected in the definition of the dynamics $\gamma : X \to \boldsymbol{R}(X, R) \times \mathcal{P}(\Phi)$ upon setting

$$\gamma(x) := \big\langle\{w \in \boldsymbol{R}(X, R) \mid w_1 = x\}, V_1(x)\big\rangle,$$

where $V_1 : X \to \mathcal{P}(\Phi)$ is defined according to $V : \Phi \to \mathcal{P}(X)$ as above. Define for the model $\mathfrak{M} := (W, R, V)$ the map $\lambda_{\boldsymbol{R}(W,R)} : C \mapsto \{\langle C', A\rangle \in \mathcal{P}(\boldsymbol{R}(W, R)) \times \mathcal{P}(\Phi) \mid C' \subseteq C\}$; then λ defines a natural transformation $\mathcal{P}^{op} \circ \boldsymbol{R} \to \mathcal{P}^{op} \circ \boldsymbol{F} \circ \boldsymbol{R}$ (the functor $\boldsymbol{F}$ has been defined in Example 2.7.56); note that we have to check naturality in terms of model morphisms, which are in particular morphisms for the underlying $\mathcal{P}$-coalgebra. Thus we can define for $w \in W$

$$w \models_{\mathfrak{M}} \boldsymbol{A}\psi \Leftrightarrow w \in \big(\gamma^{-1} \circ \lambda_{\boldsymbol{R}(W,R)}\big)(\llbracket\psi\rrbracket_{\mathfrak{M}})$$

In a similar way, we define $w \models_{\mathfrak{M}} p$ for atomic propositions $p \in \Phi$; this is left to the reader.

The interpretation of path formulas requires a slightly different approach. We define

$$\mu_{\boldsymbol{R}(X,R)}(A) := \{\sigma \in \boldsymbol{R}(X,R) \mid \sigma^1 \in A\},$$
$$\vartheta_{\boldsymbol{R}(X,R)}(A,B) := \bigcup_{k\in\mathbb{N}}\{\sigma \in \boldsymbol{R}(X,R) \mid \sigma^k \in B, \sigma^i \in A \text{ for } 0 \leq i < k\},$$

whenever $A, B \in \boldsymbol{R}(X,R)$. Then $\mu : \mathcal{P}^{op} \circ \boldsymbol{R} \to \mathcal{P}^{op} \circ \boldsymbol{R}$ and $\vartheta : (\mathcal{P}^{op} \circ \boldsymbol{R}) \times (\mathcal{P}^{op} \circ \boldsymbol{R}) \to \mathcal{P}^{op} \circ \boldsymbol{R}$ are natural transformations, and we put

$$[\![\boldsymbol{X}\psi]\!]_{\mathfrak{M}} := \mu_{\boldsymbol{R}(M,R)}([\![\psi]\!]_{\mathfrak{M}}),$$
$$[\![\psi_1\boldsymbol{U}\psi_2]\!]_{\mathfrak{M}} := \vartheta_{\boldsymbol{R}(X,R)}([\![\psi_1]\!]_{\mathfrak{M}}, [\![\psi_2]\!]_{\mathfrak{M}}).$$

The example shows that a two-level logic can be interpreted as well through a coalgebraic approach, provided the predicate liftings which characterize this approach are complemented by additional natural transformations (which are called *bridge operators* in [Dob09]). It indicates also that defining a coalgebraic logic requires first and foremost the definition of a functor and of natural transformations. Thus a certain overhead comes with it.

✌

2.8 Bibliographic Notes

The monograph by Mac Lane [ML97] discusses all the definitions and basic constructions; the text [BW99] takes much of its motivation for categorical constructions from applications in computer science. Monads are introduced following essentially Moggi's seminal paper [Mog91]. The textbook [Pum99] is an exposition fine-tuned toward students interested in categories; the proof of Lemma 2.3.24 and the discussion on Yoneda's construction follow its exposition rather closely. There are many other fine textbooks on categories available, catering also to the needs of computer scientists, among them [Awo10, Bor94a, Bor94b]; giving an exhaustive list is difficult.

The discrete probability functor has been studied extensively in [Sok05], its continuous step twin in [Gir81, Dob03]. The use of upper closed subsets for the interpretation of game logic is due to Parikh [Par85]; Pauly and Parikh [PP03] defines bisimilarity in this context. The coalgebraic interpretation is investigated in [Dob10]. Coalgebras are carefully discussed at length in [Rut00], to which the present discussion in Sect. 2.6 owes much of its structure.

The programming language `Haskell` is discussed in a growing number of accessible books - a personal selection includes [OGS09, Lip11]; the present short discussion is taken from [Dob12a]. The representation of modal logics draws substantially from [BdRV01], and the discussion on coalgebraic logic is strongly influenced by Pattinson [Pat04], the survey paper [DS11], and the monograph [Dob09].

2.9 Exercises

Exercise 2.1 The category ***uGraph*** has as objects undirected graphs. A morphism $f : (G, E) \to (H, F)$ is a map $f : G \to H$ such that $\{f(x), f(y)\} \in F$ whenever $\{x, y\} \in E$ (hence a morphism respects edges). Show that the laws of a category are satisfied.

Exercise 2.2 A morphism $f : a \to b$ in a category $\boldsymbol{K}$ is a *split monomorphism* iff it has a left inverse, i.e., there exists $g : b \to a$ such that $g \circ f = id_a$. Similarly, f is a *split epimorphism* iff it has a right inverse, i.e. there exists $g : b \to a$ such that $f \circ g = id_b$.

1. Show that every split monomorphism is monic and every split epimorphism is epic.

2. Show that a split epimorphism that is monic must be an isomorphism.

3. Show that for a morphism $f : a \to b$, it holds that:

   (i) f is a split monomorphism $\Leftrightarrow$ $\hom_{\boldsymbol{K}}(f, x)$ is surjective for every object x,

   (ii) f is a split epimorphism $\Leftrightarrow$ $\hom_{\boldsymbol{K}}(x, f)$ is surjective for every object x,

4. Characterize the split monomorphisms in ***Set***. What can you say about split epimorphisms in ***Set***?

Exercise 2.3 The category ***Par*** of sets and partial maps is defined as follows:

1. Objects are sets.
2. A morphism in $\hom_{\boldsymbol{Par}}(A, B)$ is a partial map $f : A \rightharpoonup B$, i.e., it is a set-theoretic map $f : \mathrm{car}(f) \to B$ from a subset $\mathrm{car}(f) \subseteq A$ into B. $\mathrm{car}(f)$ is called the *carrier* of f.
3. The identity $\mathrm{id}_A : A \rightharpoonup A$ is the usual identity map with $\mathrm{car}(\mathrm{id}_A) = A$.
4. For $f : A \rightharpoonup B$ and $g : B \rightharpoonup C$ the composition $g \circ f$ is defined as the usual composition $g(f(x))$ on the carrier:
$$\mathrm{car}(g \circ f) := \{x \in \mathrm{car}(f) \mid f(x) \in \mathrm{car}(g)\}.$$

1. Show that ***Par*** is a category and characterize its monomorphisms and epimorphisms.
2. Show that the usual set-theoretic Cartesian product you know is not the categorical product in ***Par***. Characterize binary products in ***Par***.

Exercise 2.4 Define the category ***Pos*** of ordered sets and monotone maps. The objects are ordered sets $(P, \leq)$; morphisms are monotone maps $f : (P, \leq) \to (Q, \sqsubseteq)$, i.e., maps $f : P \to Q$ such that $x \leq y$ implies $f(x) \sqsubseteq f(y)$. Composition and identities are inherited from ***Set***.

1. Show that under this definition ***Pos*** is a category.
2. Characterize monomorphisms and epimorphisms in ***Pos***.
3. Give an example of an ordered set $(P, \leq)$ which is isomorphic (in ***Pos***) to $(P, \leq)^{op}$ but $(P, \leq) \neq (P, \leq)^{op}$.

Show that if $(P, \leq)$ is isomorphic (in ***Pos***) to a totally ordered set $(Q, \sqsubseteq)$, then $(P, \leq)$ is also totally ordered. Use this result to give an example of a monotone map $f : (P, \leq) \to (Q, \sqsubseteq)$ that is monic and epic but not an isomorphism.

Exercise 2.5 Given a set X, the set of (finite) strings of elements of X is again denoted by X^*.

1. Show that X^* forms a monoid under concatenation, the *free monoid* over X.

2. Given a map $f : X \to Y$, extend it uniquely to a monoid morphism $f^* : X^* \to Y^*$. In particular for all $x \in X$, it should hold that $f^*(\langle x \rangle) = \langle f(x) \rangle$, where $\langle x \rangle$ denotes the string consisting only of the character x.

3. Under what conditions on X is X^* a commutative monoid, i.e., has a commutative operation?

Exercise 2.6 Let $(M, *)$ be a monoid. We define a category $\boldsymbol{M}$ as follows: it has only one object $*$, $hom_{\boldsymbol{M}}(*, *) = M$ with id_* as the unit of the monoid, and composition is defined through $m_2 \circ m_1 := m_2 * m_1$.

1. Show that $\boldsymbol{M}$ indeed forms a category.

2. Characterize the dual category $\boldsymbol{M}^{op}$. When are $\boldsymbol{M}$ and $\boldsymbol{M}^{op}$ equal?

3. Characterize monomorphisms, epimorphisms, and isomorphisms for finite M. (What happens in the infinite case?)

Exercise 2.7 Let $(S, \mathcal{A})$ and $(T, \mathcal{B})$ be measurable spaces, and assume that the σ-algebra $\mathcal{B}$ is generated by $\mathcal{B}_0$. Show that a map $f : S \to T$ is $\mathcal{A}$-$\mathcal{B}$-measurable iff $f^{-1}\big[B_0\big] \in \mathcal{A}$ for all $B_0 \in \mathcal{B}_0$.

Exercise 2.8 Let $(S, \mathcal{A})$ and $(T, \mathcal{B})$ be measurable spaces and $f : S \to T$ be $\mathcal{A}$-$\mathcal{B}$-measurable. Define $f_*(\mu)(B) := \mu(f^{-1}\big[B\big])$ for $\mu \in \mathbb{P}(S, \mathcal{A})$, $B \in \mathcal{B}$, then $f_* : \mathbb{P}(S, \mathcal{A}) \to \mathbb{P}(T, \mathcal{B})$. Show that f_* is $\wp(\mathcal{A})$-$\wp(\mathcal{B})$-measurable. Hint: Use Exercise 2.7.

Exercise 2.9 Let S be a countable sets with $p : S \to [0, 1]$ as a *discrete probability distribution*; thus $\sum_{s \in S} p(s) = 1$; denote the corresponding probability measure on $\mathcal{P}(S)$ by μ_p; hence $\mu_p(A) = \sum_{s \in A} p(s)$. Let T be an at most countable set with a discrete probability distribution q. Show that a map $f : S \to T$ is a morphism for the probability spaces $(S, \mathcal{P}(S), \mu_p)$, and $(T, \mathcal{P}(T), \mu_q)$ iff $q(t) = \sum_{f(s)=t} p(s)$ holds for all $t \in T$.

Exercise 2.10 Show that $\{x \in [0, 1] \mid \langle x, x \rangle \in E\} \in \mathcal{B}([0, 1])$, whenever $E \in \mathcal{B}([0, 1]) \otimes \mathcal{B}([0, 1])$.

Exercise 2.11 Let us chase some objects through diagrams. Consider the following diagram in a category $\boldsymbol{K}$:

$$\begin{tikzcd}
a \arrow[r, "f"] \arrow[d, "k"'] & b \arrow[r, "g"] \arrow[d, "\ell"] & c \arrow[d, "m"] \\
x \arrow[r, "r"'] & y \arrow[r, "s"'] & z
\end{tikzcd}$$

1. Show that if the left inner and right inner diagrams commute, then the outer diagram commutes as well.

2. Show that if the outer and right inner diagrams commute and s is a monomorphism, then the left inner diagram commutes as well.

3. Give examples in ***Set*** such that:

   (a) the outer and left inner diagrams commute, but not the right inner diagram,

   (b) the outer and right inner diagrams commute, but not the left inner diagram.

Exercise 2.12 Give an example of a product in a category $\boldsymbol{K}$ such that one of the projections is not epic.

Exercise 2.13 What can you say about products and sums in the category $\boldsymbol{M}$ given by a finite monoid $(M, *)$, as defined in Exercise 2.5? (Consider the case that $(M, *)$ is commutative first.)

Exercise 2.14 Show that the product topology has this universal property: $f : (D, \mathcal{D}) \to (S \times T, \mathcal{G} \times \mathcal{H})$ is continuous iff $\pi_S \circ f : (D, \mathcal{D}) \to (S, \mathcal{G})$ and $\pi_T \circ f : (D, \mathcal{D}) \to (T, \mathcal{H})$ are continuous. Formulate and prove the corresponding property for morphisms in ***Meas***.

Exercise 2.15 A collection of morphisms $\left(f_i : a \to b_i\right)_{i \in I}$ with the same domain in category $\boldsymbol{K}$ is called *jointly monic* whenever the following holds: If $g_1 : x \to a$ and $g_2 : x \to a$ are morphisms such that $f_i \circ g_1 = f_i \circ g_2$ for all $i \in I$, then $g_1 = g_2$. Dually one defines a collection of morphisms to be *jointly epic*.

Show that the projections from a categorical product are jointly monic and the injections into a categorical sum are jointly epic.

Exercise 2.16 Assume the following diagram in a category $\boldsymbol{K}$ commutes:

$$\begin{tikzcd}
a \arrow[r, "f"] \arrow[d, "k"'] & b \arrow[r, "g"] \arrow[d, "\ell"] & c \arrow[d, "m"] \\
x \arrow[r, "r"'] & y \arrow[r, "s"'] & z
\end{tikzcd}$$

Prove or disprove: if the outer diagram is a pullback, one of the inner diagrams is a pullback as well. Which inner diagram has to be a pullback for the outer one to be also a pullback?

Exercise 2.17 Suppose $f, g : a \to b$ are morphisms in a category $\boldsymbol{C}$. An *equalizer* of f and g is a morphism $e : x \to a$ such that $f \circ e = g \circ e$, and whenever $h : y \to a$ is a morphism with $f \circ h = g \circ h$, then there exists a unique $j : y \to x$ such that $h = e \circ j$.

This is the diagram:

$$\begin{tikzcd} x \arrow[r, "e"] & a \arrow[r, shift left, "f"] \arrow[r, shift right, "g"'] & b \\ y \arrow[u, dotted, "j"', "!"] \arrow[ur, "h"'] & & \end{tikzcd}$$

1. Show that equalizers are uniquely determined up to isomorphism.
2. Show that the morphism $e : x \to a$ is a monomorphism.
3. Show that a category has pullbacks if it has products and equalizers.

Exercise 2.18 A *terminal object* in category $\boldsymbol{K}$ is an object $\mathbf{1}$ such that for every object a, there exists a unique morphism $! : a \to \mathbf{1}$.

1. Show that terminal objects are uniquely determined up to isomorphism.
2. Show that a category has (binary) products and equalizers if it has pullbacks and a terminal object.

Exercise 2.19 Show that the coproduct σ-algebra has this universal property: $f : (S + T, \mathcal{A} + \mathcal{B}) \to (R, \mathcal{X})$ is $\mathcal{A} + \mathcal{B}$-$\mathcal{X}$-measurable iff $f \circ i_S$ and $f \circ i_T$ are $\mathcal{A}$-$\mathcal{X}$- resp. $\mathcal{B}$-$\mathcal{X}$-measurable. Formulate and prove the corresponding property for morphisms in ***Top***.

Exercise 2.20 Assume that in category $\boldsymbol{K}$, any two elements have a product. Show that $a \times (b \times c)$ and $(a \times b) \times c$ are isomorphic.

Exercise 2.21 Prove Lemma 2.2.22.

Exercise 2.22 Assume that the coproducts $a + a'$ and $b + b'$ exist in category $\boldsymbol{K}$. Given morphisms $f : a \to b$ and $f' : a' \to b'$, show that there exists a unique morphism $q : a + a' \to b + b'$ such that this diagram commutes:

$$\begin{tikzcd} a \arrow[d, "f"'] & a + a' \arrow[l, "i_a"'] \arrow[r, "i_{a'}"] \arrow[d, "q"] & a' \arrow[d, "f'"] \\ b & b + b' \arrow[l, "i_b"] \arrow[r, "i_{b'}"'] & b' \end{tikzcd}$$

Exercise 2.23 Show that the category ***Prob*** has no coproducts (Hint: Considering $(S, \mathcal{C}) + (T, \mathcal{D})$, show that, e.g., $i_S^{-1}\left[i_S\left[A\right]\right]$ equals A for $A \subseteq S$).

Exercise 2.24 Identify the product of two objects in the category ***Rel*** of relations.

Exercise 2.25 We investigate the epi-mono factorization in the category ***Meas*** of measurable spaces. Fix two measurable spaces $(S, \mathcal{A})$ and $(T, \mathcal{B})$ and a morphism $f : (S, \mathcal{A}) \to (T, \mathcal{B})$.

1. Let $\mathcal{A}/\ker(f)$ be the largest σ-algebra $\mathcal{X}$ on $S/\ker(f)$ rendering the factor map $\eta_{\ker(f)} : S \to S/\ker(f)$ $\mathcal{A}$-$\mathcal{X}$-measurable. Show that $\mathcal{A}/\ker(f) = \{C \subseteq S/\ker(f) \mid \eta_{\ker(f)}^{-1}\left[C\right] \in \mathcal{A}\}$, and show that $\mathcal{A}/\ker(f)$ has this universal property: given a measurable space $(Z, \mathcal{C})$, a map $g : S/\ker(f) \to Z$ is $\mathcal{A}/\ker(f)$-$\mathcal{C}$ measurable iff $g \circ \eta_{\ker(f)} : S \to Z$ is $\mathcal{A}$-$\mathcal{C}$-measurable.
2. Show that $\eta_{\ker(f)}$ is an epimorphism in ***Meas*** and that $f_{\bullet} : [x]_{\ker(f)} \mapsto f(x)$ is a monomorphism in ***Meas***.
3. Let $f = m \circ e$ with an epimorphism $e : (S, \mathcal{A}) \to (Z, \mathcal{C})$ and a monomorphism $m : (Z, \mathcal{C}) \to (T, \mathcal{B})$, and define $b : S/\ker(f) \to Z$ through $[s]_{\ker(f)} \mapsto e(s)$; see Corollary 2.1.27. Show that b is $\mathcal{A}/\ker(f)$-$\mathcal{C}$-measurable, and prove or disprove measurability of b^{-1}.

Exercise 2.26 Let ***AbGroup*** be the category of Abelian groups. Its objects are commutative groups; a morphism $\varphi : (G, +) \to (H, *)$ is a map $\varphi : G \to H$ with $\varphi(a + b) = \varphi(a) * \varphi(b)$ and $\varphi(-a) = -\varphi(a)$. Each subgroup V of an Abelian group $(G, *)$ defines an equivalence relation ρ_V through $a\ \rho_V\ b$ iff $a - b \in V$. Characterize the pushout of η_{ρ_V} and η_{ρ_W} for subgroups V and W in ***AbGroup***.

Exercise 2.27 Given a set X, define $\boldsymbol{F}(X) := X \times X$, for a map $f : X \to Y$, $\boldsymbol{F}(f)(x_1, x_2) := \langle f(x_1), f(x_2)\rangle$ is defined. Show that $\boldsymbol{F}$ is an endofunctor on ***Set***.

Exercise 2.28 Let (X, τ) be a topological space, the closed sets of which are denoted just by F for this exercise. Define $f : \mathcal{P}(X) \to F$ by $f(A) := A^a$ and by i the embedding $i : F \to \mathcal{P}(X)$. Then i and f are a Galois connection. Similarly, defining $g : \mathcal{P}(X) \to \tau$ as $g(A) := A^o$ and $j : \tau \to \mathcal{P}(X)$ as the embedding, show that g and j form a Galois connection.

Exercise 2.29 Fix a set A of labels; define $\boldsymbol{F}(X) := \{*\} \cup A \times X$ for the set X, if $f : X \to Y$ is a map; put $\boldsymbol{F}(f)(*) := *$ and $\boldsymbol{F}(f)(a, x) := \langle a, f(x) \rangle$. Show that $\boldsymbol{F} : \boldsymbol{Set} \to \boldsymbol{Set}$ defines an endofunctor.

This endofunctor models termination or labeled output.

Exercise 2.30 Fix a set A of labels, and put for the set X

$$\boldsymbol{F}(X) := \mathcal{P}_f(A \times X),$$

where $\mathcal{P}_f$ denotes all finite subsets of its argument. Thus, $G \subseteq \boldsymbol{F}(X)$ is a finite subset of $A \times X$, which models finite branching, with $\langle a, x \rangle \in G$ as one of the possible branches, which is in this case labeled by $a \in A$. Define

$$\boldsymbol{F}(f)(B) := \big\{\langle a, f(x) \rangle \mid \langle a, x \rangle \in B\big\}$$

for the map $f : X \to Y$ and $B \subseteq A \times X$. Show that $\boldsymbol{F} : \boldsymbol{Set} \to \boldsymbol{Set}$ is an endofunctor.

Exercise 2.31 Show that the limit cone for a functor $\boldsymbol{F} : \boldsymbol{K} \to \boldsymbol{L}$ is unique up to isomorphisms, provided it exists.

Exercise 2.32 Let $I \neq \emptyset$ be an arbitrary index set, and let $\boldsymbol{K}$ be the discrete category over I. Given a family $(X_i)_{i \in I}$, define $\boldsymbol{F} : I \to \boldsymbol{Set}$ by $\boldsymbol{F}i := X_i$. Show that

$$X := \prod_{i \in I} X_i := \{x : I \to \bigcup_{i \in I} X_i \mid x(i) \in X_i \text{ for all } i \in I\}$$

with $\pi_i : x \mapsto x(i)$ is a limit $(X, (\pi_i)_{i \in I})$ of $\boldsymbol{F}$.

Exercise 2.33 Formulate the equalizer of two morphisms (cp. Exercise 2.17) as a limit.

Exercise 2.34 Define for the set X the free monoid X^* generated by X through

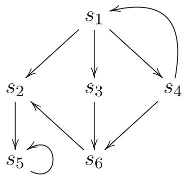
$$X^* := \{\langle x_1, \ldots, x_k \rangle \mid x_i \in X, k \geq 0\}$$

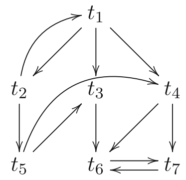
with juxtaposition as multiplication, i.e., $\langle x_1, \ldots, x_k \rangle * \langle x'_1, \ldots, x'_r \rangle := \langle x_1, \ldots, x_k, x'_1, \ldots, x'_r \rangle$; the neutral element ϵ is $\langle x_1, \ldots, x_k \rangle$ with $k = 0$; see Exercise 2.5. Define

$$f^*(x_1 * \ldots * x_k) := f(x_1) * \ldots * f(x_k)$$
$$\eta_X(x) := \langle x \rangle$$

for the map $f : X \to Y^*$ and $x \in X$. Put $\boldsymbol{F}X := X^*$. Show that $(\boldsymbol{F}, \eta, -^*)$ is a Kleisli tripel, and compare it with the `list` monad; see page 177. Compute μ_X for this monad.

Exercise 2.35 Given are the systems S and T.





1. Consider the transition systems S and T as coalgebras for a suitable functor $\boldsymbol{F} : \boldsymbol{Set} \to \boldsymbol{Set}$, $X \mapsto \mathcal{P}(X)$. Determine the dynamics of the respective coalgebras.

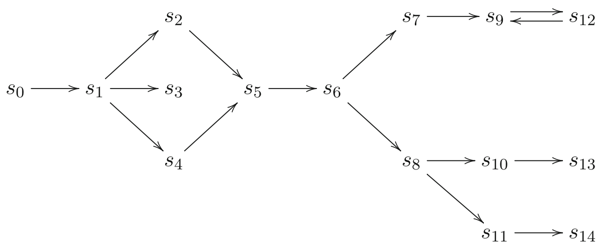
2. Show that there is no coalgebra morphism $S \to T$.

3. Construct a coalgebra morphism $T \to S$.

4. Construct a bisimulation between S and T as a coalgebra on the carrier

$$\{\langle s_2, t_3\rangle,\ \langle s_2, t_4\rangle,\ \langle s_4, t_2\rangle,\ \langle s_5, t_6\rangle,\ \langle s_5, t_7\rangle,\ \langle s_6, t_5\rangle\}.$$

Exercise 2.36 Characterize this nondeterministic transition system S as a coalgebra for a suitable functor $\boldsymbol{F} : \boldsymbol{Set} \to \boldsymbol{Set}$.



Show that

$$\begin{aligned}\alpha \quad := \quad & \{\langle s_i, s_i\rangle \mid 0 \le i \le 12\} \cup \{\langle s_2, s_4\rangle, \langle s_4, s_2\rangle, \langle s_9, s_{12}\rangle, \langle s_{12}, s_9\rangle, \\ & \langle s_{13}, s_{14}\rangle, \langle s_{14}, s_{13}\rangle\}\end{aligned}$$

is a bisimulation equivalence on S. Simplify S by giving a coalgebraic characterization of the factor system S/α. Furthermore, determine whether α is the largest bisimulation equivalence on S.

Exercise 2.37 The deterministic finite automata A_1, A_2 with input and output alphabet $\{0, 1\}$ and the following transition tables are given:

$\boldsymbol{A_1}$	state	input	output	next state		$\boldsymbol{A_2}$	state	input	output	next state
	s_0	0	0	s_1			s'_0	0	0	s'_0
	s_0	1	1	s_0			s'_0	1	1	s'_1
	s_1	0	0	s_2			s'_1	0	0	s'_0
	s_1	1	1	s_3			s'_1	1	1	s'_2
	s_2	0	1	s_4			s'_2	0	1	s'_3
	s_2	1	0	s_2			s'_2	1	0	s'_2
	s_3	0	0	s_1			s'_3	0	1	s'_4
	s_3	1	1	s_3			s'_3	1	0	s'_2
	s_4	0	1	s_3			s'_4	0	0	s'_5
	s_4	1	0	s_2			s'_4	1	1	s'_4
							s'_5	0	0	s'_2
							s'_5	1	1	s'_4

1. Formalize the automata as coalgebras for a suitable functor $\boldsymbol{F} : \boldsymbol{Set} \to \boldsymbol{Set}, \boldsymbol{F}(X) = (X \times O)^I$. You have to choose I and O first.

2. Construct a coalgebra morphism from A_1 to A_2, and use this to find a bisimulation R between A_1 and A_2. Describe the dynamics of R coalgebraically.

Exercise 2.38 Let P be an effectivity function on X, and define $\partial P(A) := X \setminus P(X \setminus A)$. Show that ∂P defines an effectivity function on X. Given an effectivity function Q on Y and a morphism $f : P \to Q$, show that $f : \partial P \to \partial Q$ is a morphism as well.

Exercise 2.39 Show that the power set functor $\mathcal{P} : \boldsymbol{Set} \to \boldsymbol{Set}$ does not preserve pullbacks. (Hint: You can use the fact that in $\boldsymbol{Set}$, the pullback of the left diagram is explicitly given as $P := \{\langle x, y\rangle \mid f(x) = g(y)\}$ with π_X and π_Y being the usual projections.)

Exercise 2.40 Suppose $\boldsymbol{F}, \boldsymbol{G} : \boldsymbol{Set} \to \boldsymbol{Set}$ are functors.

1. Show that if $\boldsymbol{F}$ and $\boldsymbol{G}$ both preserve weak pullbacks, then also the product functor $\boldsymbol{F} \times \boldsymbol{G} : \boldsymbol{Set} \to \boldsymbol{Set}$, defined as $(\boldsymbol{F} \times \boldsymbol{G})(X) = \boldsymbol{F}(X) \times \boldsymbol{G}(X)$ and $(\boldsymbol{F} \times \boldsymbol{G})(f) = \boldsymbol{F}(f) \times \boldsymbol{G}(f)$, preserves weak pullbacks.

2. Generalize to arbitrary products, i.e., show the following: If I is a set and for every $i \in I$, $\boldsymbol{F}_i : \boldsymbol{Set} \to \boldsymbol{Set}$ is a functor preserving weak pullbacks, then also the product functor $\prod_{i \in I} \boldsymbol{F}_i : \boldsymbol{Set} \to \boldsymbol{Set}$ preserves pullbacks.

Use this to show that the exponential functor $(-)^A : \boldsymbol{Set} \to \boldsymbol{Set}$, given by $X \mapsto X^A = \prod_{a \in A} X$ and $f \mapsto f^A = \prod_{a \in A} f$ preserves weak pullbacks.

3. Show that if $\boldsymbol{F}$ preserves weak pullbacks and there exist natural transformations $\eta : \boldsymbol{F} \to \boldsymbol{G}$ and $\nu : \boldsymbol{G} \to \boldsymbol{F}$, then also $\boldsymbol{G}$ preserves weak pullbacks.

4. Show that if both $\boldsymbol{F}$ and $\boldsymbol{G}$ preserve weak pullbacks, then also $\boldsymbol{F} + \boldsymbol{G} : \boldsymbol{Set} \to \boldsymbol{Set}$, defined as $X \mapsto \boldsymbol{F}(X) + \boldsymbol{G}(X)$ and $f \mapsto \boldsymbol{F}(f) + \boldsymbol{G}(f)$, preserves weak pullbacks. (Hint: Show first that for every morphism $f : X \to A + B$, one has a decomposition $X \cong X_A + X_B$ and $f_A : X_A \to A$, $f_B : X_B \to B$ such that $f \cong (f_A \circ i_A) + (f_B \circ i_B)$.)

Exercise 2.41 Consider the modal similarity type $\mathfrak{t} = (O, \rho)$, with $O := \{\langle a \rangle, \langle b \rangle\}$ and $\rho(\langle a \rangle) = \rho(\langle b \rangle) = 1$, over the propositional letters $\{p, q\}$. Let furthermore $[a]$, $[b]$ denote the nablas of $\langle a \rangle$ and $\langle b \rangle$.

Show that the following formula is a tautology, i.e., it holds in every possible $\mathfrak{t}$-model:

$$(\langle a \rangle p \vee \langle a \rangle q \vee [b](\neg p \vee q)) \to (\langle a \rangle (p \vee q) \vee \neg [b] p \vee [b] q)$$

A *frame morphism* between frames $(X, (R_{\langle a \rangle}, R_{\langle b \rangle}))$ and $(Y, (S_{\langle a \rangle}, S_{\langle b \rangle}))$ is given for this modal similarity type by a map $f : X \to Y$ which satisfies the following properties:

- If $\langle x, x_1 \rangle \in R_{\langle a \rangle}$, then $\langle f(x), f(x_1) \rangle \in S_{\langle a \rangle}$. Moreover, if $\langle f(x), y_1 \rangle \in S_{\langle a \rangle}$, then there exists $x_1 \in X$ with $\langle x, x_1 \rangle \in R_{\langle a \rangle}$ and $y_1 = f(x_1)$.
- If $\langle x, x_1 \rangle \in R_{\langle b \rangle}$, then $\langle f(x), f(x_1) \rangle \in S_{\langle b \rangle}$. Moreover, if $\langle f(x), y_1 \rangle \in S_{\langle b \rangle}$, then there exists $x_1 \in X$ with $\langle x, x_1 \rangle \in R_{\langle b \rangle}$ and $y_1 = f(x_1)$.

Give a coalgebraic definition of frame morphisms for this modal similarity type, i.e., find a functor $\boldsymbol{F} : \boldsymbol{Set} \to \boldsymbol{Set}$ such that frame morphisms correspond to $\boldsymbol{F}$-coalgebra morphisms.

Exercise 2.42 Consider the fragment of PDL defined mutually recursive by:

Formulas $\varphi ::= p \mid \varphi_1 \wedge \varphi_2 \mid \neg \varphi_1 \mid \langle \pi \rangle \varphi$ (where $p \in \Phi$ for a set of basic propositions Φ, and π is a program).

Programs $\pi ::= t \mid \pi_1;\pi_2 \mid \varphi?$ (where $t \in \text{Bas}$ for a set of basic programs Bas and φ is a formula).

Suppose you are given the set of basic programs $\text{Bas} := \{\text{init}, \text{run}, \text{print}\}$ and basic propositions $\Phi := \{is\_init, did\_print\}$.

We define a model $\mathfrak{M}$ for this language as follows:

- The basic set of $\mathfrak{M}$ is $X := \{-1, 0, 1\}$.
- The modal formulas for basic programs are interpreted by the relations

$$R_{\text{init}} := \{\langle -1, 0\rangle, \langle 0, 0\rangle, \langle 1, 1\rangle\},$$
$$R_{\text{run}} := \{\langle -1, -1\rangle, \langle 0, 0\rangle, \langle 1, 1\rangle\},$$
$$R_{\text{print}} := \{\langle -1, -1\rangle, \langle 0, 1\rangle, \langle 1, 1\rangle\}.$$

- The modal formulas for composite programs are defined by $R_{\pi_1;\pi_2} := R_{\pi_1} \circ R_{\pi_2}$ and $R_{\varphi?} := \{\langle x, x\rangle \mid \mathfrak{M}, x \models \varphi\}$, as usual.
- The valuation function is given by $V(is\_init) := \{0, 1\}$ and $V(did\_print) := \{1\}$.

Show the following:

1. $\mathfrak{M}, -1 \not\models \langle \text{run}; \text{print}\rangle did\_print$,
2. $\mathfrak{M}, x \models \langle \text{init}; \text{run}; \text{print}\rangle did\_print$ (for all $x \in X$),
3. $\mathfrak{M}, x \not\models \langle (\neg is\_init)?; \text{print}\rangle did\_print$ (for all $x \in X$).

Informally speaking, the model above allows one to determine whether a program composed of initialization (init), doing some kind of work (run), and printing (print) is initialized or has printed something.

Suppose we want to modify the logic by counting how often we have printed, i.e., we extend the set of basic propositional letters by $\{did\_print_n \mid n \in \mathbb{N}\}$. Give an appropriate model for the new logic.

Exercise 2.43 Let $\boldsymbol{E}$ be the monad which is given by all upper closed subsets of the power set of a set; see Example 2.4.10. Show that $\zeta_S : A \mapsto \{V \in \boldsymbol{E}(S) \mid A \in V\}$ defines a natural transformation $\mathcal{P} \to \boldsymbol{E}$. Compute the composition of two Kleisli morphisms for $\boldsymbol{E}$.

Chapter 3

Topological Spaces

A topology formalizes the notion of an open set; call a set open iff each of its members leaves a little room like a breathing space around it. This gives immediately a hint at the structure of the collection of open sets—they should be closed under finite intersections but under arbitrary unions, yielding the base for a calculus of observable properties, as outlined in [Smy92, Chap. 1] or in [Vic89]. This development makes use of properties of topological spaces but puts its emphasis subtly away from the classic approach, e.g., in mathematical analysis or probability theory, by stressing different properties of a space. The traditional approach, for example, stresses separation properties like being able to separate two distinct points through an open set. Such a strong emphasis is not necessarily observed in the computationally oriented use of topologies, where, for example, pseudometrics for measuring the conceptual distance between objects are important, when it comes to finding an approximation between Markov transition systems.

We give in this chapter a brief introduction to some of the main properties of topological spaces, given that we have touched upon topologies already in the context of the axiom of choice in Sect. 1.5.8. The objective is to provide the tools and methods offered by set-theoretic topology to an application-oriented reader. Thus we introduce the very basic notions of topology and hint at applications of these tools. Some connections to logic and set theory are indicated, but as Moschovakis writes

E.-E. Doberkat, *Special Topics in Mathematics for Computer Scientists*, DOI 10.1007/978-3-319-22750-4_3

"General (pointset) topology is to set theory like parsley to Greek food: some of it gets in almost every dish, but there are no 'parsley recipes' that the good Greek cook needs to know" [Mos06, 6.27, p. 79]. In this metaphor, we study the parsley here, so that it can get into the dishes which require it.

Basic notions of a topology and its construction, including bases and subbases, are already known from Chap. 1. Since compactness has been made available very early, compact spaces serve occasionally as an exercise ground. Continuity is an important topic in this context and the basic constructions like product or quotients which are enabled by it. Since some interesting and important topological constructions are tied to filters, we study *filters and convergence*, comparing in examples the sometimes more easily handled nets to the occasionally more cumbersome filters, which, however, offer some conceptual advantages. Talking about convergence, *separation properties* suggest themselves; they are studied in detail, providing some classic results like Urysohn's Theorem. It happens so often that one works with a powerful concept but that this concept requires assumptions which are too strong; hence, one has to weaken it in a sensible way. This is demonstrated in the transition from compactness to local compactness; we discuss locally compact spaces, and we give an example of a compactification. Quantitative aspects enter when one measures openness through a pseudometric; here many concepts are seen in a new, sharper light; in particular, the problem of completeness comes up—you have a sequence, the elements of which are eventually very close to each other, and you want to be sure that a limit exists. This is possible on complete spaces, and, even better, if a space is not complete, then you can complete it. Complete spaces have some very special properties, for example, the intersection of countably many open dense sets is dense again. This is Baire's Theorem. We show through a Banach–Mazur game played on a topological space that being of first category can be determined through Demon having a winning strategy.

This completes the round trip of the basic properties of topological spaces. We then present a small gallery in which topology is in action. The reason for singling out some topics is that we want to demonstrate the techniques developed with topological spaces for some interesting applications. For example, Gödel's Completeness Theorem for (countable) first-order logic has been proved by Rasiowa and Sikorski through a combination of Baire's Theorem and Stone's topological representation of Boolean algebras. This topic is discussed. The calculus of observations, which is mentioned above, leads to the notion of topological systems, as demonstrated by Vickers. This hints at an interplay of topology and order, since a topology is after all a complete Heyting algebra. Another important topic is the approximation of continuous functions by a given class of functions, like the polynomials on an interval, leading quickly to the Stone–Weierstraß Theorem on a compact topological space, a topic with a rich history. Finally, the relationship of pseudometric spaces to general topological spaces is reflected again; we introduce uniform spaces as an interesting class of spaces which is more general than pseudometric spaces but less general than their topological cousins. Here we find concepts like completeness or uniform continuity, which are formulated for metric spaces, but which cannot be realized in general topological ones. This gallery could be extended; for example, Polish spaces could be discussed here with considerable relish, but it seemed to be more adequate to discuss these spaces in the context of their measure theoretic use.

We assume throughout that the axiom of choice is valid.

3.1 Defining Topologies

Recall from Sect. 1.5.8 that a topology τ on a carrier set X is a collection of subsets which contains both $\emptyset$ and X and which is closed under finite intersections and arbitrary unions. The elements of τ are called the *open sets*. Usually, a topology is not written down as one set, but it is specified what an open set looks like. This is done through a base or a subbase. Recall that a *base* β for τ is a set of subsets of τ such that for any $x \in G$, there exists $B \in \beta$ with $x \in B \subseteq G$. A subbase is a family of sets for which the finite intersections form a base.

Base, subbase

Not every family of subsets qualifies as a subbase or a base. Kelley [Kel55, p. 47] gives the following example: Put $X := \{0, 1, 2\}$, $A := \{0, 1\}$ and $B := \{1, 2\}$; then $\beta := \{X, A, B, \emptyset\}$ cannot be the base for a topology. Assume it is, then the topology must be β itself, but $A \cap B \notin \beta$. We have the following characterization of a base.

Proposition 3.1.1 *A family β of sets is the base for a topology on $X = \bigcup \beta$ iff given $U, V \in \beta$ and $x \in U \cap V$, there exists $W \in \beta$ with $x \in W \subseteq U \cap V$, and if X.*

Thus we have to be a bit careful, in view of Kelley's example. Let us have a look at the proof.

Proof Checking the properties for a base shows that the condition is certainly necessary. Suppose that the condition holds, and define

$$\tau := \{\bigcup \beta_0 \mid \beta_0 \subseteq \beta\}.$$

Then $\emptyset, X \in \tau$, and τ is closed under arbitrary unions, so that we have to check whether τ is closed under finite intersections. In fact, let $x \in U \cap V$ with $U, V \in \tau$; then we can find $U_0, V_0 \in \beta$ with $x \in U_0 \cap V_0$. By assumption there exists $W \in \beta$ with $x \in W \subseteq U_0 \cap V_0 \subseteq U \cap V$, so that $U \cap V$ can be written as union of elements in β. $\dashv$

We perceive a base and a subbase, resp., relative to a topology, but it is usually clear what the topology looks like, once a basis is given. Let us have a look at some examples to clarify things.

Example 3.1.2 Consider the real numbers $\mathbb{R}$ with the Euclidean topology τ. We say that a set G is open iff given $x \in G$, there exists an open interval $]a, b[$ with $x \in \ ]a, b[\subseteq G$. Hence the set $\big\{]a, b[\mid a, b \in \mathbb{R}, a < b\big\}$ forms a base for τ; actually, we could have chosen a and b as rational numbers, so that we have even a countable base for τ. Note that although we can find a closed interval $[v, w]$ such that $x \in [v, w] \subseteq]a, b[\subseteq G$, we could not have used the closed intervals for a description of τ, since otherwise the singleton sets $\{x\} = [x, x]$ would be open as well. This is both undesirable and counterintuitive: in an open set, we expect each element to have some breathing space around it. ✌

The next example looks at higher dimensional Euclidean spaces; here we do not have intervals directly at our disposal, but we can measure distances as well, which is a suitable generalization, given that the interval $]x - r, x + r[$ equals $\{y \in \mathbb{R} \mid |x - y| < r\}$.

Example 3.1.3 Consider the three-dimensional space $\mathbb{R}^3$, and define for $x, y \in \mathbb{R}^3$ their distance

$$d(x, y) := \sum_{i=1}^{3} |x_i - y_i|.$$

Call $G \subseteq \mathbb{R}^3$ open iff given $x \in G$, there exists $r > 0$ such that $\{y \in \mathbb{R}^3 \mid d(x, y) < r\} \subseteq G$. Then it is clear that the set of all open sets form a topology:

- Both the empty set and $\mathbb{R}^3$ are open.
- The union of an arbitrary collection of open sets is open again.
- Let $G_1, \dots, G_k$ be open, and $x \in G_1 \cap \dots \cap G_k$. Take an index i; since $x \in G_i$, there exists $r_i > 0$ such that $K(d, x, r) := \{y \in \mathbb{R}^3 \mid d(x, y) < r_i\} \subseteq G_i$. Let $r := \min\{r_1, \dots, r_k\}$, then

$$\{y \in \mathbb{R}^3 \mid d(x, y) < r\} = \bigcap_{i=1}^{k} \{y \in \mathbb{R}^3 \mid d(x, y) < r_i\} \subseteq \bigcap_{i=1}^{k} G_i.$$

  Hence the intersection of a finite number of open sets is open again.

This argument would not work with a countable number of open sets, by the way.

We could have used other measures for the distance, e.g.,

$$d'(x, y) := \sqrt{\sum_i |x_i - y_i|^2},$$
$$d''(x, y) := \max_{1 \le i \le 3} |x_i - y_i|.$$

Then it is not difficult to see that all three describe the same collection of open sets. This is so because we can find for x and $r > 0$ some $r' > 0$ and $r'' > 0$ with $K(d', x, r') \subseteq K(d, x, r)$ and $K(d'', x, r'') \subseteq K(d, x, r)$, similarly for the other combinations.

It is noted that 3 is not a magical number here; we can safely replace it with any positive n, indicating an arbitrary finite dimension. Hence we have shown that $\mathbb{R}^n$ is a topological space in the Euclidean topology for each $n \in \mathbb{N}$. ✌

The next example uses also some notion of distance between two elements, which are given through evaluating real-valued functions. Think of $f(x)$ as the numerical value of attribute f for object x; then $|f(x) - f(y)|$ indicates how far apart x and y are with respect to their attribute values.

Example 3.1.4 Let X be an arbitrary nonempty set and $\mathcal{E}$ be a nonempty collection of functions $f : X \to \mathbb{R}$. Define for the finite collection

$\mathcal{F} \subseteq \mathcal{E}$, for $r > 0$, and for $x \in X$, the base set

$$W_{\mathcal{F};r}(x) := \{y \in X \mid |f(x) - f(y)| < r \text{ for all } f \in \mathcal{F}\}.$$

We define as a base $\beta := \{W_{\mathcal{F};r}(x) \mid x \in X, r > 0, \mathcal{F} \subseteq \mathcal{G} \text{ finite}\}$, and hence call $G \subseteq X$ open iff given $x \in G$, there exists $\mathcal{F} \subseteq \mathcal{E}$ finite and $r > 0$ such that $W_{\mathcal{F};r}(x) \subseteq G$.

Weak topology

It is immediate that the finite intersection of open sets is open again. Since the other properties are checked easily as well, we have defined a topology, which is sometimes called the *weak topology* on X induced by $\mathcal{E}$.

It is clear that in the last example, the argument would not work if we restrict ourselves to elements of $\mathcal{G}$ for defining the base, i.e., to sets of the form $W_{\{g\};r}$. These sets, however, have the property that they form a subbase, since finite intersections of these sets form a base. ✌

The next example shows that a topology may be defined on the set of all partial functions from some set to another one. In contrast to the previous example, we do without any numerical evaluations.

$A \rightharpoonup B$

Example 3.1.5 Let A and B be nonempty sets; define

$$A \rightharpoonup B := \{f \subseteq A \times B \mid f \text{ is a partial map}\};$$

see Exercise 2.3. A set $G \subseteq A \rightharpoonup B$ is called open iff given $f \in G$, there exists a finite $f_0 \in A \rightharpoonup B$ such that

$$f \in N(f_0) := \{g \in A \rightharpoonup B \mid f_0 \subseteq g\} \subseteq G.$$

Thus we can find for f a finite partial map f_0 which is extended by f, such that all extensions to f_0 are contained in G.

Then this is in fact a topology. The collection of open sets is certainly closed under arbitrary unions, and both the empty set and the whole set $A \rightharpoonup B$ are open. Let $G_1, \ldots, G_n$ be open, and $f \in G := G_1 \cap \ldots \cap G_n$; then we can find finite partial maps $f_1, \ldots, f_n$ which are extended by f such that $N(f_i) \subseteq G_i$ for $1 \leq i \leq n$. Since f extends all these maps, $f_0 := f_1 \cup \ldots \cup f_n$ is a well-defined finite partial map which is extended by f, and

$$f \in N(f_0) = N(f_1) \cap \ldots \cup N(f_n) \subseteq G.$$

Hence the finite intersection of open sets is open again.

A base for this topology is the set $\{N(f) \mid f \text{ is finite}\}$; a subbase is the set $\big\{N(\{\langle a, b\rangle\}) \mid a \in A, b \in B\big\}$. ✌

The next example deals with a topology which is induced by an order structure. Recall from Sect. 1.5 that a chain in a partially ordered set is a nonempty totally ordered subset and that in an inductively ordered set, each chain has an upper bound.

Example 3.1.6 Let $(P, \leq)$ be a inductively ordered set. Call $G \subseteq P$ *Scott open* iff:

1. G is upper closed (hence $x \in G$ and $x \leq y$ imply $y \in G$),
2. if $S \subseteq P$ is a chain with $\sup S \in G$, then $S \cap G \neq \emptyset$.

Again, this defines a topology on P. In fact, it is enough to show that $G_1 \cap G_2$ is open, if G_1 and G_2 are. Let S be a chain with $\sup S \in G_1 \cap G_2$; then we find $s_i \in S$ with $s_i \in G_i$. Since S is a chain, we may and do assume that $s_1 \leq s_2$; hence $s_2 \in G_1$, because G_1 is upper closed. Thus $s_2 \in S \cap (G_1 \cap G_2)$. Because G_1 and G_2 are upper closed, so is $G_1 \cap G_2$.

As an illustration, we show that the set $F := \{x \in P \mid x \leq t\}$ is Scott closed for each $t \in P$. Put $G := P \setminus F$. Let $x \in G$, and $x \leq y$; then obviously $y \notin F$, so $y \in G$. If S is a chain with $\sup S \in G$, then there exists $s \in S$ such that $s \notin F$; hence $S \cap G \neq \emptyset$. ✌

3.1.1 Continuous Functions

A continuous map between topological spaces is compatible with the topological structure. This is familiar from real functions, but we cannot copy the definition, since we have no means of measuring the distance between points in a topological space. All we have is the notion of an open set. So the basic idea is to say that given an open neighborhood U of the image, we want to be able to find an open neighborhood V of the inverse image so that all element of V are mapped to U. This is a direct translation of the familiar ϵ-δ-definition from calculus. Since we are concerned with continuity as a global concept (as opposed to one which focuses on a given point), we arrive at this definition and observe in the subsequent example that it is really a faithful translation.

Definition 3.1.7 *Let (X, τ) and (Y, ϑ) be topological spaces. A map $f : X \to Y$ is called τ-ϑ-*continuous *iff $f^{-1}\left[H\right] \in \tau$ for all $H \in \vartheta$ holds; we write this also as $f : (X, \tau) \to (Y, \vartheta)$.*

If the context is clear, we omit the reference to the topologies. Hence we say that the inverse image of an open set under a continuous map is an open set again; see Example 2.1.11.

Let us have a look at real functions.

Example 3.1.8 Endow the reals with the Euclidean topology, and let $f : \mathbb{R} \to \mathbb{R}$ be a map. Then the definition of continuity given above coincides with the usual ϵ-δ-definition.

ϵ-δ?

1. Assuming the ϵ-δ-definition, we show that the inverse image of an open set is open. In fact, let $G \subseteq \mathbb{R}$ be open, and pick $x \in f^{-1}\left[G\right]$. Since $f(x) \in G$, we can find $\epsilon > 0$ such that $]f(x) - \epsilon, f(x) + \epsilon[\subseteq G$. Pick $\delta > 0$ for this ϵ; hence $x' \in]x - \delta, x + \delta[$ implies $f(x') \in]f(x) - \epsilon, f(x) + \epsilon[\subseteq G$. Thus $x \in]x - \delta, x + \delta[\subseteq f^{-1}\left[G\right]$.

2. Assuming that the inverse image of an open set is open, we establish the ϵ-δ-definition. Given $x \in \mathbb{R}$, let $\epsilon > 0$ be arbitrary; we show that there exists $\delta > 0$ such that $|x - x'| < \delta$ implies $|f(x) - f(x')| < \epsilon$. Now $]f(x) - \epsilon, f(x') + \epsilon[$ is an open set; hence $H := f^{-1}\left[]f(x) - \epsilon, f(x') + \epsilon[\right]$ is open by assumption, and $x \in H$, Select $\delta > 0$ with $]x - \delta, x + \delta[\subseteq H$, then $|x - x'| < \delta$ implies $x' \in H$; hence $f(x') \in]f(x) - \epsilon, f(x') + \epsilon[$. ✌

Thus we work on familiar ground, when it comes to the reals. Continuity may be tested on a subbase:

Lemma 3.1.9 *Let (X, τ) and (Y, ϑ) be topological spaces and $f : X \to Y$ be a map. Then f is τ-ϑ-continuous iff $f^{-1}\left[S\right] \in \tau$ for each $S \in \sigma$ with $\sigma \subseteq \vartheta$ a subbase.*

Proof Clearly, the inverse image of a subbase element is open, whenever f is continuous. Assume, conversely, that the $f^{-1}\left[S\right] \in \tau$ for each $S \in \sigma$. Then $f^{-1}\left[B\right] \in \tau$ for each element B of the base β generated from σ, because B is the intersection of a finite number of subbase elements. Now, finally, if $H \in \vartheta$, then $H = \bigcup\{B \mid B \in \beta, B \subseteq H\}$, so that $f^{-1}\left[H\right] = \bigcup\{f^{-1}\left[B\right] \mid B \in \beta, B \subseteq H\} \in \tau$. Thus the inverse image of an open set is open. $\dashv$

Example 3.1.10 Take the topology from Example 3.1.5 on the space $A \rightharpoonup B$ of all partial maps. A map $q : (A \rightharpoonup B) \to (C \rightharpoonup D)$ is continuous in this topology iff the following condition holds: whenever $q(f)(c) = d$, then there exists $f_0 \subseteq f$ finite such that $q(f_0)(c) = d$.

In fact, let q be continuous, and $q(f)(c) = d$, then $G := q^{-1}\left[N(\{\langle c, d\rangle\})\right]$ is open and contains f; thus there exists $f_0 \subseteq f$ with $f \in N(f_0) \subseteq G$, in particular $q(f_0)(c) = d$. Conversely, assume that $H \subseteq C \rightharpoonup D$ is open, and we want to show that $G := q^{-1}\left[H\right] \subseteq A \rightharpoonup B$ is open. Let $f \in G$; thus $q(f) \in H$; hence there exists $g_0 \subseteq q(f)$ finite with $q(f) \in N(g_0) \subseteq H$. g_0 is finite, say $g_0 = \{\langle c_1, d_1\rangle, \ldots, \langle c_n, d_n\rangle\}$. By assumption, there exists $f_0 \in A \rightharpoonup B$ with $q(f_0)(c_i) = d_i$ for $1 \leq i \leq n$; then $f \in N(f_0) \subseteq G$, so that the latter set is open. ✌

This is an easy criterion for continuity with respect to the Scott topology.

Example 3.1.11 Let $(P, \leq)$ and $(Q, \leq)$ be inductively ordered sets, then $f : P \to Q$ is Scott continuous (i.e., continuous when both ordered sets carry their respective Scott topology) iff f is monotone and if $f(\sup S) = \sup f\left[S\right]$ holds for every chain S.

Assume that f is Scott continuous. If $x \leq x'$, then every open set which contains x also contains x'; so if $x \in f^{-1}\left[H\right]$, then $x' \in f^{-1}\left[H\right]$ for every Scott open $H \subseteq Q$; thus f is monotone. If $S \subseteq P$ is a chain, then $\sup S$ exists in P, and $f(s) \leq f(\sup S)$ for all $s \in S$, so that $\sup f\left[S\right] \leq f(\sup S)$. For the other inequality, assume that $f(\sup S) \not\leq \sup f\left[S\right]$. We note that $G := f^{-1}\left[\{q \in Q \mid q \not\leq \sup f\left[S\right]\}\right]$ is open with $\sup S \in G$; hence there exists $s \in S$ with $s \in G$. But this is impossible. On the other hand, assume that $H \subseteq Q$ is Scott open; we want to show that $G := f^{-1}\left[H\right] \subseteq P$ is Scott open. G is upper closed, since $x \in G$ and $x \leq x'$ imply $f(x) \in H$ and $f(x) \leq f(x')$; thus $f(x') \in H$, so that $x' \in G$. Let $S \subseteq P$ be a chain with $\sup S \in G$; hence $f(\sup S) \in H$. Since $f\left[S\right]$ is a chain, and $f(\sup S) = \sup f\left[S\right]$, we infer that there exists $s \in S$ with $f(s) \in H$; hence there is $s \in S$ with $s \in G$. Thus G is Scott open in P, and f is Scott continuous. ✌

The interpretation of modal logics in a topological space is interesting, when we interpret the transition which is associated with the diamond operator through a continuous map. Thus the next step of a transition is uniquely determined, and it depends continuously on its argument.

Example 3.1.12 The syntax of our modal logics is given through

$$\varphi ::= \top \mid p \mid \varphi_1 \vee \varphi_2 \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \Diamond\varphi$$

with $p \in \Phi$ an atomic proposition. The logic has the usual operators, viz., disjunction and negation, and $\Diamond$ as the modal operator.

For interpreting the logic, we take a topological state space (S, τ) and a continuous map $f : X \to X$, and we associate with each atomic proposition p an open set V_p as the set of all states in which p is true. We want the validity set $[\![\varphi]\!]$ of all those states in which formula φ holds to be open and define inductively the validity of a formula in a state in the following way:

$$\begin{aligned}
[\![\top]\!] &:= S\\
[\![p]\!] &:= V_p, \text{ if } p \text{ is atomic}\\
[\![\varphi_1 \vee \varphi_2]\!] &:= [\![\varphi_1]\!] \cup [\![\varphi_2]\!]\\
[\![\varphi_1 \wedge \varphi_2]\!] &:= [\![\varphi_1]\!] \cap [\![\varphi_2]\!]\\
[\![\neg\varphi]\!] &:= (S \setminus [\![\varphi]\!])^o\\
[\![\Diamond\varphi]\!] &:= f^{-1}\big[[\![\varphi]\!]\big]
\end{aligned}$$

All definitions but the last two are self-explanatory. The interpretation of $[\![\Diamond\gamma]\!]$ through $f^{-1}\big[[\![\varphi]\!]\big]$ suggests itself when considering the graph of f in the usual interpretation of the diamond in modal logics; see Definition 2.7.15.

Since we want $[\![\neg\varphi]\!]$ to be open, we cannot take the complement of $[\![\varphi]\!]$ and declare it as the validity set for φ, because the complement of an open set is not necessarily open. Instead, we take the largest open set which is contained in $S \setminus [\![\varphi]\!]$ (this is the best we can do) and assign it to $\neg\varphi$. One shows easily through induction on the structure of formula φ that $[\![\varphi]\!]$ is an open set.

But now look at this. Assume that $X := \mathbb{R}$ in the usual topology, $V_p = [\![p]\!] =]0, +\infty[$, then $[\![\neg p]\!] =]-\infty, 0]^o =]-\infty, 0[$; thus $[\![p \vee \neg p]\!] = \mathbb{R} \setminus \{0\} \neq [\![\top]\!]$. Thus the law of the excluded middle does not hold in this model. ✌

Returning to the general discussion, the following fundamental property is immediate; we state it here just for the record; see Example 2.1.11.

Proposition 3.1.13 *The identity $(X, \tau) \to (X, \tau)$ is continuous, and continuous maps are closed under composition. Consequently, topological spaces with continuous maps form a category.* $\dashv$

Continuous maps can be used to define topologies.

Definition 3.1.14 *Given a family $\mathcal{F}$ of maps $f : A \to X_f$, where (X_f, τ_f) is a topological space for each $f \in \mathcal{F}$, the* initial topology *$\tau_{in,\mathcal{F}}$ on A with respect to $\mathcal{F}$ is the smallest topology on A so that f is $\tau_{in,\mathcal{F}}$-τ_f-continuous for every $f \in \mathcal{F}$. Dually, given a family $\mathcal{G}$ of maps $g : X_g \to Z$, where (X_g, τ_g) is a topological space for each $g \in \mathcal{G}$, the* final topology *$\tau_{fi,\mathcal{G}}$ on Z is the largest topology on Z so that g is τ-$\tau_{fi,\mathcal{G}}$-continuous for every $g \in \mathcal{G}$.*

In the case of the initial topology for just one map $f : A \to X_f$, note that $\mathcal{P}(A)$ is a topology which renders f continuous, so there exists in fact a smallest topology on A with the desired property; because $\{f^{-1}\left[G\right] \mid G \in \tau_f\}$ is a topology that satisfies the requirement, and because each such topology must contain it, this is in fact the smallest one. If we have a family $\mathcal{F}$ of maps $A \to X_f$, then each topology making all $f \in \mathcal{F}$ continuous must contain $\xi := \bigcup_{f \in \mathcal{F}}\{f^{-1}\left[G\right] \mid G \in \tau_f\}$, so the initial topology with respect to $\mathcal{F}$ is just the smallest topology on A containing ξ. Similarly, being the largest topology rendering each $g \in \mathcal{G}$ continuous, the final topology with respect to $\mathcal{G}$ must contain the set $\bigcup_{g \in \mathcal{G}}\{H \mid g^{-1}\left[H\right] \in \tau_g\}$.

An easy characterization of the initial resp. the final topology is proposed here:

Proposition 3.1.15 *Let (Z, τ) be a topological space and $\mathcal{F}$ be a family of maps $A \to X_f$ with (X_f, τ_f) topological spaces; A is endowed with the initial topology $\tau_{in,\mathcal{F}}$ with respect to $\mathcal{F}$. A map $h : Z \to A$ is τ-$\tau_{in,\mathcal{F}}$-continuous iff $h \circ f : Z \to X_f$ is τ-τ_f-continuous for every $f \in \mathcal{F}$.*

Proof 1. Certainly, if $h : Z \to A$ is τ-$\tau_{in,\mathcal{F}}$ continuous, then $h \circ f : Z \to X_f$ is τ-τ_f-continuous for every $f \in \mathcal{F}$ by Proposition 3.1.13.

2. Assume, conversely, that $h \circ f$ is continuous for every $f \in \mathcal{F}$; we want to show that h is continuous. Consider

$$\zeta := \{G \subseteq A \mid h^{-1}\left[G\right] \in \tau\}.$$

Because τ is a topology, ζ is; because $h \circ f$ is continuous, ζ contains the sets $\{f^{-1}[H] \mid H \in \tau_f\}$ for every $f \in \mathcal{F}$. But this implies that ζ contains $\tau_{in,\mathcal{F}}$; hence $h^{-1}[G] \in \tau$ for every $G \in \tau_{in,\mathcal{F}}$. This establishes the assertion. $\dashv$

There is a dual characterization for the final topology; see Exercise 3.1.

These are the most popular examples for initial and final topologies:

Product

1. Given a family $(X_i, \tau_i)_{i \in I}$ of topological spaces, let $X := \prod_{i \in I} X_i$ be the Cartesian product of the carrier sets.[1] The *product topology* $\prod_{i \in I} \tau_i$ is the initial topology on X with respect to the projections $\pi_i : X \to X_i$. The product topology has as a base

   $$\{\prod_{i \in I} A_i \mid A_i \in \tau_i \text{ and } A_i \neq X_i \text{ only for finitely many indices}\}$$

Subspace

2. Let (X, τ) be a topological space, $A \subseteq X$. The *trace* $(A, \tau \cap A)$ of τ on A is the initial topology on A with respect to the embedding $i_A : A \to X$. It has the open sets $\{G \cap A \mid G \in \tau\}$; this is sometimes called the *subspace topology*; see page 58. We do not assume that A is open.

Sum

3. Given the family of spaces as above, let $X := \sum_{i \in I} X_i$ be the direct sum. The *sum topology* $\sum_{i \in I} \tau_i$ is the final topology on X with respect to the injections $\iota_i : X_i \to X$. Its open sets are described through

   $$\{\sum_{i \in I} \iota_i[G_i] \mid G_i \in \tau_i \text{ for all } i \in I\}.$$

Factor

4. Let ρ be an equivalence relation on X with τ a topology on the base space. The factor space X/ρ is equipped with the final topology τ/ρ with respect to the factor map η_ρ which sends each element to its ρ-class. This topology is called the *quotient topology* (with respect to τ and ρ). If a set $G \subseteq X/\rho$ is open, then its inverse image $\eta_\rho^{-1}[G] = \bigcup G \subseteq X$ is open in X. But the converse holds as well: assume that $\bigcup G$ is open in X for some $G \subseteq X/\rho$, then $G = \eta_\rho[\bigcup G]$, and, because $\bigcup G$ is the union if equivalence classes, one shows that $\eta_\rho^{-1}[G] = \eta_\rho^{-1}[\eta_\rho[\bigcup G]] = \bigcup G$. But this means that G is open in X/ρ.

[1] This works only if $X \neq \emptyset$; recall that we assume here that the axiom of choice is valid.

Just to gain some familiarity with the concepts involved, we deal with an induced map on a product space and with the subspace coming from the image of a map. The properties we find here will be useful later on as well.

The product space first. We will use that a map into a topological product is continuous iff all its projections are; this follows from the characterization of an initial topology. It goes like this:

Lemma 3.1.16 *Let M and N be nonempty sets and $f : M \to N$ be a map. Equip both $[0,1]^M$ and $[0,1]^N$ with the product topology. Then*

$$f^* : \begin{cases} [0,1]^N & \to [0,1]^M \\ g & \mapsto g \circ f \end{cases}$$

is continuous.

Proof Note the reversed order; we have $f^*(g)(m) = (g \circ f)(m) = g(f(m))$ for $g \in [0,1]^N$ and $m \in M$.

Because f^* maps $[0,1]^N$ into $[0,1]^M$, and the latter space carries the initial topology with respect to the projections $(\pi_{M,m})_{m \in N}$ with $\pi_{M,m} : q \mapsto q(m)$, it is by Proposition 3.1.15 sufficient to show that $\pi_{M,m} \circ f^* : [0,1]^N \to [0,1]$ is continuous for every $m \in M$. But $\pi_{M,m} \circ f^* = \pi_{N,f(m)}$; this is a projection, which is continuous by definition. Hence f^* is continuous. $\dashv$

Hence an application of the projection defuses a seemingly complicated map. Note in passing that neither M nor N are assumed to carry a topology; they are simply plain sets.

The next observation displays an example of a subspace topology. Each continuous map $f : X \to Y$ of one topological space to another one induces a subspace $f\left[X\right]$ of Y, which may or may not have interesting properties. In the case considered, it inherits compactness from its source.

Proposition 3.1.17 *Let (X, τ) and (Y, ϑ) be topological spaces and $f : X \to Y$ be τ-ϑ-continuous. If (X, τ) is compact, so is $(f\left[X\right], \vartheta \cap f\left[X\right])$, the subspace of (Y, ϑ) induced by f.*

Proof We take on open cover of $f\left[X\right]$ and show that it contains a finite cover of this space. So let $(H_i)_{i \in I}$ be an open cover of $f\left[X\right]$. There exists open sets $H_i' \in \vartheta$ such that $H_i' = H_i \cap f\left[X\right]$, since

$(f[X], \tau \cap f[X])$ carries the subspace topology. Then $(f^{-1}[H_i'])_{i \in I}$ is an open cover of X, so there exists a finite subset $J \subseteq I$ such that $X = \bigcup_{i \in J} f^{-1}[H_i']$, since X is compact. But then $(H_i' \cap f[X])_{i \in J}$ is an open cover of $f[X]$. Hence this space is compact. $\dashv$

Before proceeding further, we introduce the notion of homeomorphism (as an isomorphism in the category of topological spaces with continuous maps).

Definition 3.1.18 *Let X and Y be topological spaces. A bijection $f : X \to Y$ is called a* homeomorphism *iff both f and f^{-1} are continuous.*

It is clear that continuity and bijectivity alone do not make a homeomorphism. Take as a trivial example the identity $(\mathbb{R}, \mathcal{P}(\mathbb{R})) \to (\mathbb{R}, \tau)$ with τ as the Euclidean topology. It is continuous and bijective, but its inverse is not continuous.

Let us have a look at some examples, first one for the quotient topology.

Example 3.1.19 Let $U := [0, 2 \cdot \pi]$, and identify the endpoints of the interval, i.e., consider the equivalence relation

$$\rho := \{\langle x, x\rangle \mid x \in U\} \cup \{\langle 0, 2 \cdot \pi\rangle, \langle 2 \cdot \pi, 0\rangle\}.$$

Let $K := U/\rho$, and endow K with the quotient topology.

A set $G \subseteq K$ is open iff $\eta_\rho^{-1}[G] \subseteq U$ is open, thus iff we can find an open set $H \subseteq \mathbb{R}$ such that $\eta_\rho^{-1}[G] = H \cap U$, since U carries the trace of $\mathbb{R}$. Consequently, if $[0]_\rho \notin G$, we find that $\eta_\rho^{-1}[G] = \{x \in U \mid \{x\} \in G\}$, which is open by construction. If, however, $[0]_\rho \in G$, then $\eta_\rho^{-1}[G] = \{x \in U \mid \{x\} \in G\} \cup \{0, 2 \cdot \pi\}$, which is open in U.

We claim that K and the unit circle $S := \{\langle s, t\rangle \mid 0 \leq s, t \leq 1, s^2 + t^2 = 1\rangle\}$ are homeomorphic under the map $\psi : [x]_\rho \mapsto \langle \sin x, \cos x\rangle$. Because $\langle \sin 0, \cos 0\rangle = \langle \sin 2 \cdot \pi, \cos 2 \cdot \pi\rangle$, the map is well defined. Since we can write $S = \{\langle \sin x, \cos x\rangle \mid 0 \leq x \leq 2 \cdot \pi\}$, it is clear that ψ is onto. The topology on S is inherited from the Cartesian plane, so open arcs are a subbasis for it. Because the old Romans Sinus and Cosinus both are continuous, we find that $\psi \circ \eta_\rho$ is continuous. We infer from Exercise 3.1 that ψ is continuous, since K has the quotient topology, which is final.

We show now that ψ^{-1} is continuous. The argumentation is geometrical. Given an open arc on K, we may describe it through its endpoints (P_1, P_2) with a clockwise movement. If the arc does not contain the critical point $P := \langle 0, 1\rangle$, we find an open interval $I :=]a, b[$ with $0 < a < b < 2 \cdot \pi$ such that $\psi\big[(P_1, P_2)\big] = \{[x]_\rho \mid x \in I\}$, which is open in K. If, however, P is on this arc, we decompose it into two parts $(P_1, P) \cup (P, P_2)$. Then (P_1, P) is the image of some interval $]a, 2 \cdot \pi]$, and (P, P_2) is the image of an interval $[0, b[$, so that $\psi\big[(P_1, P_2)\big] = \eta_\rho\big[[0, b[\cup]a, 2 \cdot \pi]\big]$, which is open in K as well (note that $[0, b[$ and $]a, 2 \cdot \pi]$ are open in U). ✌

While we have described so far direct methods to describe a topology by saying under which conditions a set is open, we turn now to an observation due to Kuratowski which yields an indirect way. It describes axiomatically what properties the closure of a set should have. Assume that we have a *closure operator*, i.e., a map $A \mapsto A^{\mathbf{c}}$ on the powerset of a set X with these properties: Closure operator

1. $\emptyset^{\mathbf{c}} = \emptyset$ and $X^{\mathbf{c}} = X$.

2. $A \subseteq A^{\mathbf{c}}$, and $(A \cup B)^{\mathbf{c}} = A^{\mathbf{c}} \cup B^{\mathbf{c}}$.

3. $(A^{\mathbf{c}})^{\mathbf{c}} = A^{\mathbf{c}}$.

Thus the operator leaves the empty set and the whole set alone, the closure of the union is the union of the closures, and the operator is idempotent. One sees immediately that the operator which assigns to each set its closure with respect to a given topology is such a closure operator. It is also quite evident that a closure operator is monotone. Assume that $A \subseteq B$, then $B = A \cup (B \setminus A)$, so that $B^{\mathbf{c}} = A^{\mathbf{c}} \cup (B \setminus A)^{\mathbf{c}} \supseteq A^{\mathbf{c}}$.

Example 3.1.20 Let $(D, \leq)$ be a finite partially ordered set. We put $\emptyset^{\mathbf{c}} := \emptyset$ and $D^{\mathbf{c}} := D$; moreover,

$$\{x\}^{\mathbf{c}} := \{y \in D \mid y \leq x\}$$

is defined for $x \in D$ and $A^{\mathbf{c}} := \bigcup_{x \in X} \{x\}^{\mathbf{c}}$ for subsets A of D. Then this is a closure operator. It is enough to check whether $(\{x\}^{\mathbf{c}})^{\mathbf{c}} = \{x\}^{\mathbf{c}}$

holds. In fact, we have

$$\begin{aligned} z \in (\{x\}^{\mathbf{c}})^{\mathbf{c}} &\Leftrightarrow z \in \{y\}^{\mathbf{c}} \text{ for some } y \in \{x\}^{\mathbf{c}} \\ &\Leftrightarrow \text{there exists } y \leq x \text{ with } z \leq y \\ &\Leftrightarrow z \leq x \\ &\Leftrightarrow z \in \{x\}^{\mathbf{c}}. \end{aligned}$$

Thus we associate with each finite partially ordered set a closure operator, which assigns to each $A \subseteq D$ its down set. The map $x \mapsto \{x\}^{\mathbf{c}}$ embeds D into a distributive lattice; see the discussion in Sect. 1.5.6. ✌

We will show now that we can obtain a topology by calling open all those sets the complements of which remain fixed under the closure operator; in addition, it turns out that the topological closure and the one from the closure operator are the same.

Theorem 3.1.21 *Let $\cdot^{c}$ be a closure operator. Then:*

1. *The set $\tau := \{X \setminus F \mid F \subseteq X, F^{c} = F\}$ is a topology.*
2. *For each set $A^{a} = A^{c}$ with $\cdot^{a}$ as the closure in τ.*

Proof 1. For establishing that τ is a topology, it is enough to show that τ is closed under arbitrary unions, since the other properties are evident. Let $\mathcal{G} \subseteq \tau$, and put $G := \bigcup \mathcal{G}$, so we want to know whether $X \setminus G^{\mathbf{c}} = X \setminus G$. If $H \in \mathcal{G}$, then $X \setminus G \subseteq X \setminus H$, so $(X \setminus G)^{\mathbf{c}} \subseteq (X \setminus H)^{\mathbf{c}} = X \setminus H$; thus $(X \setminus G)^{\mathbf{c}} \subseteq X \setminus G$. Since the operator is monotone, it follows that $(X \setminus G)^{\mathbf{c}} = X \setminus G$; hence τ is in fact closed under arbitrary unions, and hence it is a topology.

2. Given $A \subseteq X$,

$$A^{a} = \bigcap \{F \subseteq X \mid F \text{ is closed, and } A \subseteq F\},$$

and $A^{\mathbf{c}}$ takes part in the intersection, so that $A^{a} \subseteq A^{\mathbf{c}}$. On the other hand, $A \subseteq A^{a}$, thus $A^{\mathbf{c}} \subseteq (A^{a})^{\mathbf{c}} = A^{a}$ by part 1. Consequently, A^{a} and $A^{\mathbf{c}}$ are the same. $\dashv$

It is on first sight a bit surprising that a topology can be described by finitary means, although arbitrary unions are involved for the topology. But we should not forget that we have also the subset relation at our disposal. Nevertheless, a rest of surprise remains.

3.1.2 Neighborhood Filters

The last method for describing a topology we are discussing here deals also with some order properties. Assume that we assign to each $x \in X$, where X is a given carrier set, a filter $\mathfrak{U}(x) \subseteq \mathcal{P}(X)$ with the property that $x \in U$ holds for each $U \in \mathfrak{U}(x)$. Thus $\mathfrak{U}(x)$ has these properties:

1. $x \in U$ for all $U \in \mathfrak{U}(x)$.
2. If $U, V \in \mathfrak{U}(x)$, then $U \cap V \in \mathfrak{U}(x)$.
3. If $U \in \mathfrak{U}(x)$ and $U \subseteq V$, then $V \in \mathfrak{U}(x)$.

It is fairly clear that, given a topology τ on X, the *neighborhood filter* $\mathfrak{U}_\tau(x)$

$$\mathfrak{U}_\tau(x) := \{V \subseteq X \mid \text{there exists } U \in \tau \text{ with } x \in U \text{ and } U \subseteq V\}$$

for x has these properties. It has also an additional property, which we will discuss shortly—for dramaturgical reasons.

Such a system of special filters defines a topology. We declare all those sets as open which belong to the neighborhoods of their elements. So if we take all balls in Euclidean $\mathbb{R}^3$ as the basis for a filter and assign to each point the balls which it centers, then the sphere of radius 1 around the origin would not be open (intuitively, it does not contain an open ball). So this appears to be an appealing idea. In fact:

Proposition 3.1.22 *Let* $\{\mathfrak{U}(x) \mid x \in X\}$ *be a family of filters such that* $x \in U$ *for all* $U \in \mathfrak{U}(x)$. *Then*

$$\tau := \{U \subseteq X \mid U \in \mathfrak{U}(x) \textit{ whenever } x \in U\}$$

defines a topology on X.

Proof We have to establish that τ is closed under finite intersections, since the other properties are fairly straightforward. Now, let U and V be open, and take $x \in U \cap V$. We know that $U \in \mathfrak{U}(x)$, since U is open, and we have $V \in \mathfrak{U}(x)$ for the same reason. Since $\mathfrak{U}(x)$ is a filter, it is closed under finite intersections; hence $U \cap V \in \mathfrak{U}(x)$, and thus $U \cap V$ is open. $\dashv$

We cannot, however, be sure that the neighborhood filter $\mathfrak{U}_\tau(x)$ for this new topology is the same as the given one. Intuitively, the reason is that we do not know if we can find for $U \in \mathfrak{U}(x)$ an open $V \in \mathfrak{U}(x)$ with $V \subseteq U$ such that $V \in \mathfrak{U}(y)$ for all $y \in V$. To illustrate, look at $\mathbb{R}^3$, and take the neighborhood filter for, say, 0 in the Euclidean topology. Put for simplicity

$$\|x\| := \sqrt{x_1^2 + x_2^2 + x_3^2}.$$

Let $U \in \mathfrak{U}(0)$; then we can find an open ball $V \in \mathfrak{U}(0)$ with $V \subseteq U$. In fact, assume $U = \{a \mid \|a\| < q\}$. Take $z \in U$; then we can find $r > 0$ such that the ball $V := \{y \mid \|y - z\| < r\}$ is entirely contained in U (select $\|z\| < r < q$); thus $V \in \mathfrak{U}(0)$. Now let $y \in V$, let $0 < t < r - \|z - y\|$, then $\{a \mid \|a - y\| < t\} \subseteq V$, since $\|a - z\| \leq \|a - y\| + \|z - y\| < r$. Hence $U \in \mathfrak{U}(y)$ for all $y \in V$.

We obtain now as a simple corollary:

Corollary 3.1.23 *Let $\{\mathfrak{U}(x) \mid x \in X\}$ be a family of filters such that $x \in U$ for all $U \in \mathfrak{U}(x)$, and assume that for any $U \in \mathfrak{U}(x)$, there exists $V \in \mathfrak{U}(x)$ with $V \subseteq U$ and $U \in \mathfrak{U}(y)$ for all $y \in V$. Then $\{\mathfrak{U}(x) \mid x \in X\}$ coincides with the neighborhood filter for the topology defined by this family.* $\dashv$

In what follows, unless otherwise stated, $\mathfrak{U}(x)$ will denote the neighborhood filter of a point x in a topological space X.

Example 3.1.24 Let $L := \{1, 2, 3, 6\}$ be the set of all divisors of 6, and define $x \leq y$ iff x divides y, so that we obtain

$$\begin{tikzcd} & 6 \arrow[dl] \arrow[dr] & \\ 2 \arrow[dr] & & 3 \arrow[dl] \\ & 1 & \end{tikzcd}$$

Let us compute—just for fun—the topology associated with this partial order and a basis for the neighborhood filters for each element. The topology can be seen from the table below (we have used that $A^o = X \setminus (X \setminus A)^a$; see page 59):

Set	Closure	Interior
$\{1\}$	$\{1\}$	$\emptyset$
$\{2\}$	$\{1,2\}$	$\emptyset$
$\{3\}$	$\{1,3\}$	$\emptyset$
$\{6\}$	$\{1,2,3,6\}$	$\{6\}$
$\{1,2\}$	$\{1,2\}$	$\emptyset$
$\{1,3\}$	$\{1,3\}$	$\emptyset$
$\{1,6\}$	$\{1,2,3,6\}$	$\{6\}$
$\{2,3\}$	$\{1,2,3,5\}$	$\emptyset$
$\{2,6\}$	$\{1,2,3,6\}$	$\{2,6\}$
$\{3,6\}$	$\{1,2,3,6\}$	$\emptyset$
$\{1,2,3\}$	$\{1,2,3\}$	$\emptyset$
$\{1,2,6\}$	$\{1,2,3,6\}$	$\{2,6\}$
$\{1,3,6\}$	$\{1,2,3,6\}$	$\{3,6\}$
$\{2,3,6\}$	$\{1,2,3,6\}$	$\{2,3,6\}$
$\{1,2,3,6\}$	$\{1,2,3,6\}$	$\{1,2,3,6\}$

This is the topology:

$$\tau = \big\{\emptyset, \{6\}, \{2,6\}, \{3,6\}, \{2,3,6\}, \{1,2,3,6\}\big\}.$$

A basis for the respective neighborhood filters is given in this table:

Element	Basis
1	$\big\{\{1,2,3,6\}\big\}$
2	$\big\{\{2,6\}, \{1,2,3,6\}\big\}$
3	$\big\{\{3,6\}, \{2,3,6\}, \{1,2,3,6\}\big\}$
6	$\big\{\{6\}, \{2,6\}, \{3,6\}, \{2,3,6\}, \{1,2,3,6\}\big\}$

✌

The next example deals with topological groups, i.e., topological spaces which have also a group structure rendering the group operations continuous. Here the neighborhood structure is fairly uniform—if you know the neighborhood filter of the neutral element, you know the neighborhood filter of each element, because you obtain them by a left shift or a right shift.

Example 3.1.25 Let $(G, \cdot)$ be a group and τ be a topology on G such that the map $\langle x, y\rangle \mapsto xy^{-1}$ is continuous. Then $(G, \cdot, \tau)$ is called a *topological group*. We will write down a topological group as G; the group operations and the topology will not be mentioned. The neutral element is denoted by e; multiplication will usually be omitted. Given a subset U of G, define $gU := \{gh \mid h \in U\}$ and $Ug := \{hg \mid h \in U\}$ for $g \in G$.

Let us examine the algebraic operations in a group. Put $\zeta(x, y) := xy^{-1}$, then the map $\xi : g \mapsto g^{-1}$ which maps each group element to its inverse is just $\zeta(e, g)$; hence the cut of a continuous map to it is continuous as well. ξ is a bijection with $\xi \circ \xi = id_G$, so it is in fact a homeomorphism. We obtain multiplication as $xy = \zeta(x, \xi(y))$, so multiplication is also continuous. Fix $g \in G$, then multiplication $\lambda_g : x \mapsto gx$ from the left and $\rho_g : x \mapsto xg$ from the right are continuous. Now both λ_g and ρ_g are bijections, and $\lambda_g \circ \lambda_{g^{-1}} = \lambda_{g^{-1}} \circ \lambda_g = id_G$, also $\rho_g \circ \rho_{g^{-1}} = \rho_{g^{-1}} \circ \rho_g = id_G$; thus λ_g and ρ_g are homeomorphisms for every $g \in G$.

Thus we have in a topological group this characterization of the neighborhood filter for every $g \in G$:

$$\mathfrak{U}(g) = \{gU \mid U \in \mathfrak{U}(e)\} = \{Ug \mid U \in \mathfrak{U}(e)\}.$$

In fact, let U be a neighborhood of g, then $\lambda_g^{-1}\big[U\big] = g^{-1}U$ is a neighborhood of e, so is $\rho_g^{-1}\big[U\big] = Ug^{-1}$. Conversely, a neighborhood V of e determines a neighborhood $\lambda_{g^{-1}}^{-1}\big[V\big] = gV$ resp. $\rho_{g^{-1}}^{-1}\big[V\big] = Vg$ of g. ✌

3.2 Filters and Convergence

The relationship between topologies and filters turns out to be fairly tight, as we saw when discussing the neighborhood filter of a point. We saw also that we can actually grow a topology from a suitable family of neighborhood filters. This relationship is even closer, as we will discuss now when having a look at convergence.

Let $(x_n)_{n \in \mathbb{N}}$ be a sequence in $\mathbb{R}$ which converges to $x \in \mathbb{R}$. This means that for any given open neighborhood U of x, there exists an index $n \in \mathbb{N}$ such that $\{x_m \mid m \geq n\} \subseteq U$, so all members of the sequence having an index larger than n are members of U. Now consider the filter $\mathfrak{F}$ generated by the set $\big\{\{x_m \mid m \geq n\} \mid n \in \mathbb{N}\big\}$ of tails. The condition above says exactly that $\mathfrak{U}(x) \subseteq \mathfrak{F}$, if you think a bit about it. This leads to the definition of convergence in terms of filters.

Definition 3.2.1 *Let X be a topological space and $\mathfrak{F}$ a filter on X. Then* $\mathfrak{F}$ converges *to a limit $x \in X$ iff $\mathfrak{U}(x) \subseteq \mathfrak{F}$. This is denoted by $\mathfrak{F} \rightarrow x$.* $\mathfrak{F} \rightarrow x$

Plainly, $\mathfrak{U}(x) \rightarrow x$ for every x. Note that the definition above does not force the limit to be uniquely determined. If two different points x, y share their neighborhood filter, then $\mathfrak{F} \rightarrow x$ iff $\mathfrak{F} \rightarrow y$. Look again at Example 3.1.24. There, all neighborhood filters are contained in $\mathfrak{U}(6)$, so that we have $\mathfrak{U}(6) \rightarrow t$ for $t \in \{1, 2, 3, 6\}$. It may seem that the definition of convergence through a filter is too involved (after all, being a filter should not be taken on a light shoulder!). In fact, sometimes convergence is defined through a *net* as follows. Let $(I, \leq)$ be a *directed set*, i.e., $\leq$ is a partial order such that, given $i, j \in I$, there exists k with $i \leq k$ and $j \leq k$. An I-indexed family $(x_i)_{i \in I}$ is said to converge to a point x iff, given a neighborhood $U \in \mathfrak{U}(x)$, there exists $k \in I$ such that $x_i \in U$ for all $i \geq k$. This generalizes the concept of convergence from sequences to index sets of arbitrary size. But look at this. The sets $\big\{\{x_j \mid j \geq i\} \mid i \in I\big\}$ form a filter base, because $(I, \leq)$ is directed. The corresponding filter converges to x iff the net converges to x. Net

But what about the converse? Take a filter $\mathfrak{F}$ on X; then $F_1 \leq F_2$ iff $F_2 \subseteq F_1$ renders $(\mathfrak{F}, \leq)$ a net. In fact, given $F_1, F_2 \in \mathfrak{F}$, we have $F_1 \leq F_1 \cap F_2$ and $F_2 \leq F_1 \cap F_2$. Now pick $x_F \in F$. Then the net $(x_F)_{F \in \mathfrak{F}}$ converges to x iff $\mathfrak{F} \rightarrow x$. Assume that $\mathfrak{F} \rightarrow x$; take $U \in \mathfrak{U}(x)$, then $U \in \mathfrak{F}$; thus if $F \in \mathfrak{F}$ with $F \geq U$, then $F \subseteq U$; hence $x_F \in U$ for all such x_F. Conversely, if each net $(x_F)_{F \in \mathfrak{F}}$ derived from $\mathfrak{F}$ converges to x, then for a given $U \in \mathfrak{U}(x)$, there exists F_0 such that $x_F \in U$ for $F \subseteq F_0$. Since x_F has been chosen arbitrarily from F, this can only hold if $F \subseteq U$ for $F \subseteq F_0$, so that $U \in \mathfrak{F}$. Because $U \in \mathfrak{U}(x)$ was arbitrary, we conclude $\mathfrak{U}(x) \subseteq \mathfrak{F}$.

Hence we find that filters offer a uniform generalization.

The argument above shows that we may select the elements x_F from a base for $\mathfrak{F}$. If the filter has a countable base, we construct in this way a sequence; conversely, the filter constructed from a sequence has a countable base. Thus the convergence of sequences and the convergence of filters with a countable base are equivalent concepts.

We investigate the characterization of the topological closure in terms of filters. In order to do this, we need to be able to restrict a filter to a set, Trace i.e., looking at the footstep the filter leaves on the set, hence at

$$\mathfrak{F} \cap A := \{F \cap A \mid F \in \mathfrak{F}\}.$$

This is what we will do now.

Lemma 3.2.2 *Let X be a set and $\mathfrak{F}$ be a filter on X. Then $\mathfrak{F} \cap A$ is a filter on A iff $F \cap A \neq \emptyset$ for all $F \in \mathfrak{F}$.*

Proof Since a filter must not contain the empty set, the condition is necessary. But it is also sufficient, because it makes sure that the laws of a filter are satisfied. $\dashv$

Looking at $\mathfrak{F} \cap A$ for an ultrafilter $\mathfrak{F}$, we know that either $A \in \mathfrak{F}$ or $X \setminus A \in \mathfrak{F}$, so if $F \cap A \neq \emptyset$ holds for all $F \in \mathfrak{F}$, then this implies that $A \in \mathfrak{F}$. Thus we obtain:

Corollary 3.2.3 *Let X be a set and $\mathfrak{F}$ be an ultrafilter on X. Then $\mathfrak{F} \cap A$ is a filter iff $A \in \mathfrak{F}$. Moreover, in this case, $\mathfrak{F} \cap A$ is an ultrafilter on A.*

Proof It remains to show that $\mathfrak{F} \cap A$ is an ultrafilter on A, provided $\mathfrak{F} \cap A$ is a filter. Let $B \notin \mathfrak{F} \cap A$ for some subset $B \subseteq A$. Since $A \in \mathfrak{F}$, we conclude $B \notin \mathfrak{F}$; thus $X \setminus B \in \mathfrak{F}$, since $\mathfrak{F}$ is an ultrafilter. Thus $(X \setminus B) \cap A = A \setminus B \in \mathfrak{F} \cap A$, so $\mathfrak{F} \cap A$ is an ultrafilter by Lemma 1.5.21. $\dashv$

From Lemma 3.2.2, we obtain a simple and elegant characterization of the topological closure of a set.

Proposition 3.2.4 *Let X be a topological space, $A \subseteq X$. Then $x \in A^a$ iff $\mathfrak{U}(x) \cap A$ is a filter on A. Thus $x \in A^a$ iff there exists a filter $\mathfrak{F}$ on A with $\mathfrak{F} \to x$.*

Proof We know from the definition of A^a that $x \in A^a$ iff $U \cap A \neq \emptyset$ for all $U \in \mathfrak{U}(x)$. This is by Lemma 3.2.2 equivalent to $\mathfrak{U}(x) \cap A$ being a filter on A. $\dashv$

We know from calculus that continuous functions preserve convergence, i.e., if $x_n \to x$ and f is continuous, then $f(x_n) \to f(x)$. We want to carry this over to the world of filters. For this, we have to define the image of a filter. Let $\mathfrak{F}$ be a filter on a set X and $f : X \to Y$ a map; then

$$f(\mathfrak{F}) := \{B \subseteq Y \mid f^{-1}\big[B\big] \in \mathfrak{F}\}$$

Image of a filter

is a filter on Y. In fact, $\emptyset \notin f(\mathfrak{F})$, and, since f^{-1} preserves the Boolean operations, $f(\mathfrak{F})$ is closed under finite intersections. Let $B \in f(\mathfrak{F})$ and

$B \subseteq B'$. Since $f^{-1}\left[B\right] \in \mathfrak{F}$, and $f^{-1}\left[B\right] \subseteq f^{-1}\left[B'\right]$, we conclude $f^{-1}\left[B'\right] \in \mathfrak{F}$, so that $B' \in f(\mathfrak{F})$. Hence $f(\mathfrak{F})$ is also upper closed, so that it is in fact a filter.

This is an easy representation through the direct image.

Lemma 3.2.5 *Let $f : X \to Y$ be a map, $\mathfrak{F}$ a filter on X, then $f(\mathfrak{F})$ equals the filter generated by $\{f\left[A\right] \mid A \in \mathfrak{F}\}$.*

Proof Because $f\left[A_1 \cap A_2\right] \subseteq f\left[A_1\right] \cap f\left[A_2\right]$, the set $\mathcal{G}_0 := \{f\left[A\right] \mid A \in \mathfrak{F}\}$ is a filter base. Denote by $\mathcal{G}$ the filter generated by $\mathcal{G}_0$.

We claim that $f(\mathfrak{F}) = \mathcal{G}$.

"$\subseteq$": Assume that $B \in f(\mathfrak{F})$; hence $f^{-1}\left[B\right] \in \mathfrak{F}$. Since $f\left[f^{-1}\left[B\right]\right] \subseteq B$, we conclude that B is contained in the filter generated by $\mathcal{G}_0$, hence in $\mathcal{G}$.

"$\supseteq$": If $B \in \mathcal{G}_0$, we find $A \in \mathfrak{F}$ with $B = f\left[A\right]$; hence $A \subseteq f^{-1}\left[f\left[A\right]\right] = f^{-1}\left[B\right] \in \mathfrak{F}$, so that $B \in f(\mathfrak{F})$. This implies the desired inclusion, since $f(\mathfrak{F})$ is a filter.

This establishes the desired equality and proves the claim. ⊣

We see also that not only the filter property is transported through maps, but also the property of being an ultrafilter.

Lemma 3.2.6 *Let $f : X \to Y$ be a map and $\mathfrak{F}$ an ultrafilter on X. Then $f(\mathfrak{F})$ is an ultrafilter on Y.*

Proof It is by Lemma 1.5.21 enough to show that if $f(\mathfrak{F})$ does not contain a set, it will contain its complement, since $f(\mathfrak{F})$ is already known to be a filter. In fact, assume that $H \notin f(\mathfrak{F})$, so that $f^{-1}\left[H\right] \notin \mathfrak{F}$. Since $\mathfrak{F}$ is an ultrafilter, we know that $X \setminus f^{-1}\left[H\right] \in \mathfrak{F}$; but $X \setminus f^{-1}\left[H\right] = f^{-1}\left[Y \setminus H\right]$, so that $Y \setminus H \in f(\mathfrak{F})$. ⊣

Example 3.2.7 Let X be the product of the topological spaces $(X_i)_{i \in I}$ with projections $\pi_i : X \to X_i$. For a filter $\mathfrak{F}$ on X, we have $\pi_j(\mathfrak{F}) = \{A_j \subseteq X_j \mid A_j \times \prod_{i \neq j} X_i \in \mathfrak{F}\}$. ✌

Continuity preserves convergence:

Proposition 3.2.8 *Let X and Y be topological spaces and $f : X \to Y$ a map.*

1. *If f is continuous and $\mathfrak{F}$ a filter on X, then $\mathfrak{F} \to x$ implies $f(\mathfrak{F}) \to f(x)$ for all $x \in X$.*

2. *If $\mathfrak{F} \to x$ implies $f(\mathfrak{F}) \to f(x)$ for all $x \in X$ and all filters $\mathfrak{F}$ on X, then f is continuous.*

Proof Let $V \in \mathfrak{U}(f(x))$; then there exists $U \in \mathfrak{U}(f(x))$ open with $U \subseteq V$. Since $f^{-1}\big[U\big] \in \mathfrak{U}(x) \subseteq \mathfrak{F}$, we conclude $U \in f(\mathfrak{F})$; hence $V \in f(\mathfrak{F})$. Thus $\mathfrak{U}(f(x)) \subseteq f(\mathfrak{F})$, which means that $f(\mathfrak{F}) \to f(x)$ indeed. This establishes the first part.

Now assume that $\mathfrak{F} \to x$ implies $f(\mathfrak{F}) \to f(x)$ for all $x \in X$ and an arbitrary filter $\mathfrak{F}$ on X. Let $V \subseteq Y$ be open. Given $x \in f^{-1}\big[V\big]$, we find an open set U with $x \in U \subseteq f^{-1}\big[V\big]$ in the following way. Because $x \in f^{-1}\big[V\big]$, we know $f(x) \in V$. Since $\mathfrak{U}(x) \to x$, we obtain from the assumption that $f(\mathfrak{U}(x)) \to f(x)$; thus $\mathfrak{U}(f(x)) \subseteq f(\mathfrak{U}(x))$. Because $V \in \mathfrak{U}(f(x))$, it follows $f^{-1}\big[V\big] \in \mathfrak{U}(x)$; hence we find an open set U with $x \in U \subseteq f^{-1}\big[V\big]$. Consequently, $f^{-1}\big[V\big]$ is open in X. $\dashv$

Thus continuity and filters cooperate in a friendly manner.

Proposition 3.2.9 *Assume that X carries the initial topology with respect to a family $(f_i : X \to X_i)_{i \in I}$ of functions. Then $\mathfrak{F} \to x$ iff $f_i(\mathfrak{F}) \to f_i(x)$ for all $i \in I$.*

Proof Proposition 3.2.8 shows that the condition is necessary. Assume that $f_i(\mathfrak{F}) \to f_i(x)$ for every $i \in I$, let τ_i be the topology on X_i. The sets

$$\begin{gathered}\{\{f_{i_1}^{-1}\big[G_{i_1}\big] \cap \ldots \cap f_{i_k}^{-1}\big[G_{i_k}\big]\} \mid i_1, \ldots, i_k \in I,\\ f_{i_1}(x) \in G_{i_1} \in \tau_{i_1}, \ldots, f_{i_k}(x) \in G_{i_k} \in \tau_{i_k}, k \in \mathbb{N}\big\}\end{gathered}$$

form a base for the neighborhood filter for x in the initial topology. Thus, given an open neighborhood U of x, we have $f_{i_1}^{-1}\big[G_{i_1}\big] \cap \ldots \cap f_{i_k}^{-1}\big[G_{i_k}\big] \subseteq U$ for some suitable finite set of indices. Since $f_{i_j}(\mathfrak{F}) \to f_{i_j}(x)$, we infer $G_{i_j} \in f_{i_j}(\mathfrak{F})$; hence $f_{i_j}^{-1}\big[G_{i_j}\big] \in \mathfrak{F}$ for $1 \leq j \leq k$, and thus $U \in \mathfrak{F}$. This means $\mathfrak{U}(x) \subseteq \mathfrak{F}$. Hence $\mathfrak{F} \to x$, as asserted. $\dashv$

We know that in a product, a sequence converges iff its components converge. This is the counterpart for filters:

Corollary 3.2.10 *Let $X = \prod_{i \in I} X_i$ be the product of the topological spaces. Then $\mathfrak{F} \to (x_i)_{i \in I}$ in X iff $\mathfrak{F}_i \to x_i$ in X_i for all $i \in I$, where $\mathfrak{F}_i$ is the i-th projection $\pi_i(\mathfrak{F})$ of $\mathfrak{F}$.* $\dashv$

The next observation further tightens the connection between topological properties and filters. It requires the existence of ultrafilters, so recall that we assume that the axiom of choice holds.

Theorem 3.2.11 *Let X be a topological space. Then X is compact iff each ultrafilter converges.*

Thus we tie compactness, i.e., the possibility to extract from each cover a finite subcover, to the convergence of ultrafilters. Hence an ultrafilter in a compact space cannot but converge. The proof of Alexander's Subbase Theorem 1.5.57 indicates already that there is a fairly close connection between the axiom of choice and topological compactness. This connection is tightened here.

Proof 1. Assume that X is compact but that we find an ultrafilter $\mathfrak{F}$ which fails to converge. Hence we can find for each $x \in X$ an open neighborhood U_x of x which is not contained in $\mathfrak{F}$. Since $\mathfrak{F}$ is an ultrafilter, $X \setminus U_x \in \mathfrak{F}$. Thus $\{X \setminus U_x \mid x \in X\} \subseteq \mathfrak{F}$ is a collection of closed sets with $\bigcap_{x \in X}(X \setminus U_x) = \emptyset$. Since X is compact, we find a finite subset $F \subseteq X$ such that $\bigcap_{x \in F}(X \setminus U_x) = \emptyset$. But $X \setminus U_x \in \mathfrak{F}$, and $\mathfrak{F}$ is closed under finite Intersections; hence $\emptyset \in \mathfrak{F}$. This is a contradiction.

2. Assume that each ultrafilter converges. It is sufficient to show that each family $\mathcal{H}$ of closed sets for which every finite subfamily has a nonempty intersection has a nonempty intersection itself. Now, the set $\{\bigcap \mathcal{H}_0 \mid \mathcal{H}_0 \subseteq \mathcal{H} \text{ finite}\}$ of all finite intersections forms the base for a filter $\mathfrak{F}_0$, which may be extended to an ultrafilter $\mathfrak{F}$ by Theorem 1.5.43. By assumption $\mathfrak{F} \to x$ for some x, hence $\mathfrak{U}(x) \subseteq \mathfrak{F}$. The point x is a candidate for being a member in the intersection. Assume the contrary. Then there exists $H \in \mathcal{H}$ with $x \notin H$, so that $x \in X \setminus H$, which is open. Thus $X \setminus H \in \mathfrak{U}(x) \subseteq \mathfrak{F}$. On the other hand, $H = \bigcap\{H\} \in \mathfrak{F}_0 \subseteq \mathfrak{F}$, so that $\emptyset \in \mathfrak{F}$. Thus we arrive at a contradiction, and $x \in \bigcap \mathcal{H}$. Hence $\bigcap \mathcal{H} \neq \emptyset$. $\dashv$

From Theorem 3.2.11 we obtain Tihonov's celebrated theorem[2] as an easy consequence.

[2] "The Tychonoff Product Theorem concerning the stability of compactness under formation of topological products may well be regarded as the single most important theorem of general topology" according to H. Herrlich and G.E. Strecker, quoted from [Her06, p. 85].

Theorem 3.2.12 *(Tihonov's Theorem) The product $\prod_{i \in I} X_i$ of topological spaces with $X_i \neq \emptyset$ for all $i \in I$ is compact iff each space X_i is compact.*

Proof If the product $X := \prod_{i \in I} X_i$ is compact, then $\pi_i\big[X\big] = X_i$ is compact by Proposition 3.1.17. Let, conversely, be $\mathfrak{F}$ an ultrafilter on X, and assume all X_i are compact. Then $\pi_i(\mathfrak{F})$ is by Lemma 3.2.6 an ultrafilter on X_i for all $i \in I$, which converges to some x_i by Theorem 3.2.11. Hence $\mathfrak{F} \to (x_i)_{i \in I}$ by Corollary 3.2.10. This implies the compactness of X by another application of Theorem 3.2.11. $\dashv$

According to [Eng89, p. 146], Tihonov established the theorem for a product of an arbitrary numbers of closed and bounded intervals of the real line (we know from the Heine–Borel Theorem 1.5.46 that these intervals are compact). Kelley [Kel55, p. 143] gives a proof of the nontrivial implication of the theorem which relies on Alexander's Subbase Theorem 1.5.57. It goes like this. It is sufficient to establish that whenever we have a family of subbase elements, each finite family of which fails to cover X, then the whole family will not cover X. The sets $\{\pi_i^{-1}\big[U\big] \mid U \subseteq X_i \text{ open}, i \in I\}$ form a subbase for the product topology of X. Let $\mathcal{S}$ be a family of sets taken from this subbase such that no finite family of elements of $\mathcal{S}$ covers X. Put $\mathcal{S}_i := \{U \subseteq X_i \mid \pi_i^{-1}\big[U\big] \in \mathcal{S}\}$, then $\mathcal{S}_i$ is a family of open sets in X_i. Suppose $\mathcal{S}_i$ contains sets $U_1, \ldots, U_k$ which cover X_i, then $\pi_i^{-1}\big[U_1\big], \ldots, \pi_i^{-1}\big[U_k\big]$ are elements of $\mathcal{S}$ which cover X; this is impossible, and hence $\mathcal{S}_i$ fails to contain a finite family which covers X_i. Since X_i is compact, there exists a point $x_i \in X_i$ with $x_i \notin \bigcup \mathcal{S}_i$. But then $x := (x_i)_{i \in I}$ cannot be a member of $\bigcup \mathcal{S}$. Hence $\mathcal{S}$ does not cover X. This completes the proof.

Axiom of Choice

Both proofs rely heavily on the axiom of choice, the first one through the existence of an ultrafilter extending a given filter and the second one through Alexander's Subbase Theorem. The relationship of Tihonov's Theorem to the axiom of choice is even closer: It can actually be shown that the theorem and the axiom of choice are equivalent [Her06, Theorem 4.68]; this requires, however, establishing the existence of topological products without any recourse to the infinite Cartesian product as a carrier.

We have defined above the concept of a limit point of a filter. A weaker concept is that of an accumulation point. Talking in terms of sequences, an accumulation point of a sequence has the property that each neighborhood of the point contains infinitely many elements of the sequence. This carries over to filters in the following way.

Definition 3.2.13 *Given a topological space X, the point $x \in X$ is called an* accumulation point *of filter $\mathfrak{F}$ iff $U \cap F \neq \emptyset$ for every $U \in \mathfrak{U}(x)$ and every $F \in \mathfrak{F}$.*

Since $\mathfrak{F} \to x$ iff $\mathfrak{U}(x) \subseteq \mathfrak{F}$, it is clear that x is an accumulation point of $\mathfrak{F}$. But a filter may fail to have an accumulation point at all. Consider the filter $\mathfrak{F}$ over $\mathbb{R}$ which is generated by the filter base $\{]a, \infty[\mid a \in \mathbb{R}\}$; it is immediate that $\mathfrak{F}$ does not have an accumulation point. Let us have a look at a sequence $(x_n)_{n \in \mathbb{N}}$, and the filter $\mathfrak{F}$ generated by the infinite tails $\{\{x_m \mid m \geq n\} \mid n \in \mathbb{N}\}$. If x is an accumulation point of the sequence, $U \cap \{x_m \mid m \geq n\} \neq \emptyset$ for every neighborhood U of x; thus $U \cap F \neq \emptyset$ for all $F \in \mathfrak{F}$ and all such U. Conversely, if x is an accumulation point for filter $\mathfrak{F}$, it is clear that the defining property holds also for the elements of the base for the filter; thus x is an accumulation point for the sequence. Hence we have found the "right" generalization from sequences to filters.

An easy characterization of the set of all accumulation points goes like this:

Lemma 3.2.14 *The set of all accumulation points of filter $\mathfrak{F}$ is exactly $\bigcap_{F \in \mathfrak{F}} F^a$.*

Proof This follows immediately from the observation that $x \in A^a$ iff $U \cap A \neq \emptyset$ for each neighborhood $U \in \mathfrak{U}(x)$. $\dashv$

The lemma has an interesting consequence for the characterization of compact spaces through filters:

Corollary 3.2.15 *X is compact iff each filter on X has an accumulation point.*

Proof Let $\mathfrak{F}$ be a filter in a compact space X, and assume that $\mathfrak{F}$ does not have an accumulation point. Lemma 3.2.14 implies that $\bigcap_{F \in \mathfrak{F}} F^a = \emptyset$. Since X is compact, we find $F_1, \ldots, F_n \in \mathfrak{F}$ with $\bigcap_{i=1}^n F_i{}^a = \emptyset$. Thus $\bigcap_{i=1}^n F_i = \emptyset$. But this set is a member of $\mathfrak{F}$, a contradiction.

Now assume that each filter has an accumulation point. It is by Theorem 3.2.11 enough to show that every ultrafilter $\mathfrak{F}$ converges. An accumulation point x for $\mathfrak{F}$ is a limit: assume that $\mathfrak{F} \not\to x$, then there exists

$V \in \mathfrak{U}(x)$ with $V \notin \mathfrak{F}$; hence $X \setminus V \in \mathfrak{F}$. But $V \cap F \neq \emptyset$ for all $F \in \mathfrak{F}$, since x is an accumulation point. This is a contradiction. $\dashv$

This is a characterization of accumulation points in terms of converging filters.

Lemma 3.2.16 *In a topological space X, the point $x \in X$ is an accumulation point of filter $\mathfrak{F}$ iff there exists a filter $\mathfrak{F}_0$ with $\mathfrak{F} \subseteq \mathfrak{F}_0$ and $\mathfrak{F}_0 \to x$.*

Proof Let x be an accumulation point of $\mathfrak{F}$, then $\{U \cap F \mid U \in \mathfrak{U}(x), F \in \mathfrak{F}\}$ is a filter base. Let $\mathfrak{F}_0$ be the filter generated by this base, then $\mathfrak{F} \subseteq \mathfrak{F}_0$, and certainly $\mathfrak{U}(x) \subseteq \mathfrak{F}_0$, thus $\mathfrak{F}_0 \to x$.

Conversely, let $\mathfrak{F} \subseteq \mathfrak{F}_0 \to x$. Since $\mathfrak{U}(x) \subseteq \mathfrak{F}_0$ follows, we conclude $U \cap F \neq \emptyset$ for all neighborhoods U and all elements $F \in \mathfrak{F}$, for otherwise we would have $\emptyset = U \cap F \in \mathfrak{F}$ for some $U, F \in \mathfrak{F}$, which contradicts $\emptyset \in \mathfrak{F}$. Thus x is indeed an accumulation point of $\mathfrak{F}$. $\dashv$

3.3 Separation Properties

We see from Example 3.1.24 that a filter may converge to more than one point. This may be undesirable. Think of a filter which is based on a sequence, and each element of the sequence indicates an approximation step. Then you want the approximation to converge, but the result of this approximation process should be unique. We will see that this is actually a special case of a separation property.

Proposition 3.3.1 *Given a topological space X, the following properties are equivalent:*

1. *If $x \neq y$ are different points in X, there exists $U \in \mathfrak{U}(x)$ and $V \in \mathfrak{U}(y)$ with $U \cap V = \emptyset$.*
2. *The limit of a converging filter is uniquely determined.*
3. *$\{x\} = \bigcap\{U \mid U \in \mathfrak{U}(x) \text{ is closed}\}$ or all points x.*
4. *The diagonal $\Delta := \{\langle x, x\rangle \mid x \in X\}$ is closed in $X \times X$.*

Proof

1 $\Rightarrow$ 2: If $\mathfrak{F} \to x$ and $\mathfrak{F} \to y$ with $x \neq y$, we have $U \cap V \in \mathfrak{F}$ for all $U \in \mathfrak{U}(x)$ and $V \in \mathfrak{U}(y)$; hence $\emptyset \in \mathfrak{F}$. This is a contradiction.

2 $\Rightarrow$ 3: Let $y \in \bigcap\{U \mid U \in \mathfrak{U}(x) \text{ is closed}\}$; thus y is an accumulation point of $\mathfrak{U}(x)$. Hence there exists a filter $\mathfrak{F}$ with $\mathfrak{U}(x) \subseteq \mathfrak{F} \to y$ by Lemma 3.2.16. Thus $x = y$.

3 $\Rightarrow$ 4: Let $\langle x, y\rangle \notin \Delta$; then there exists a closed neighborhood W of x with $y \notin W$. Let $U \in \mathfrak{U}(x)$ open with $U \subseteq W$, and put $V := X \setminus W$; then $\langle x, y\rangle \in U \times V \cap \Delta = \emptyset$, and $U \times V$ is open in $X \times X$.

4 $\Rightarrow$ 1: If $\langle x, y\rangle \in (X \times X) \setminus \Delta$, there exists open sets $U \in \mathfrak{U}(x)$ and $V \in \mathfrak{U}(y)$ with $U \times V \cap \Delta = \emptyset$; hence $U \cap V = \emptyset$. $\dashv$

Looking at the proposition, we see that having a unique limit for a filter is tantamount to being able to separate two different points through disjoint open neighborhoods. Because these spaces are important, they deserve a special name.

Definition 3.3.2 *A topological space is called a* Hausdorff space *iff any two different points in X can be separated by disjoint open neighborhoods, i.e., iff condition (1) in Proposition 3.3.1 holds. Hausdorff spaces are also called T_2*-spaces.

Example 3.3.3 Let $X := \mathbb{R}$, and define a topology through the base $\big\{[a, b[\mid a, b \in \mathbb{R}, a < b\big\}$. Then this is a Hausdorff space. This space is sometimes called the *Sorgenfrey line*. ✌

Being Hausdorff can be determined from neighborhood filters:

Lemma 3.3.4 *Let X be a topological space. Then X is a Hausdorff space iff each $x \in X$ has a base $\mathfrak{U}_0(x)$ for its neighborhood filters such that for any $x \neq y$, there exists $U \in \mathfrak{U}_0(x)$ and $V \in \mathfrak{U}_0(y)$ with $U \cap V = \emptyset$.* $\dashv$

It follows a first and easy consequence for maps into a Hausdorff space, viz., the set of arguments on which they coincide is closed.

Corollary 3.3.5 *Let X, Y be topological spaces and $f, g : X \to Y$ continuous maps. If Y is a Hausdorff space, then $\{x \in X \mid f(x) = g(x)\}$ is closed.*

Proof The map $t : x \mapsto \langle f(x), g(x)\rangle$ is a continuous map $X \to Y \times Y$. Since $\Delta \subseteq Y \times Y$ is closed by Proposition 3.3.1, the set $t^{-1}\big[\Delta\big]$ is closed. But this is just the set in question. ⊣

The reason for calling a Hausdorff space a T_2 space[3] will become clear once we have discussed other ways of separating points and sets; then T_2 will be a point in a spectrum denoting separation properties. For the moment, we introduce two other separation properties which deal with the possibility of distinguishing two different points through open sets. Let for this X be a topological space.

T_0, T_1

T_0-space: X is called a *T_0-space* iff, given two different points x and y, there exists an open set U which contains exactly one of them.

T_1-space: X is called a *T_1-space* iff, given two different points x and y, there exist open neighborhoods U of x and V of y with $y \notin U$ and $x \notin V$.

The following examples demonstrate these spaces.

Example 3.3.6 Let $X := \mathbb{R}$, and define the topologies on the real numbers through

$$\begin{aligned}\tau_< &:= \{\emptyset, \mathbb{R}\} \cup \big\{]-\infty, a[\ \mid a \in \mathbb{R}\big\},\\ \tau_\le &:= \{\emptyset, \mathbb{R}\} \cup \big\{]-\infty, a]\mid a \in \mathbb{R}\big\}.\end{aligned}$$

Then $\tau_<$ is a T_0-topology. $\tau_\le$ is a T_1-topology which is not T_0. ✌

This is an easy characterization of T_1-spaces.

Proposition 3.3.7 *A topological space X is a T_1-space iff $\{x\}$ is closed for all $x \in X$.*

Proof Let $y \in \{x\}^a$, then y is in every open neighborhood U of x. But this can happen in a T_1-space only if $x = y$. Conversely, if $\{x\}$ is closed, and $y \neq x$, then there exists a neighborhood U of x which does not contain y, and x is not in the open set $X \setminus \{x\}$. ⊣

Example 3.3.8 Let X be a set with at least two points, $x_0 \in X$ be fixed. Put $\emptyset^{\mathbf{c}} := \emptyset$ and for $A^{\mathbf{c}} := A \cup \{x_0\}$ for $A \neq \emptyset$. Then $\cdot^{\mathbf{c}}$ is a closure operator, and we look at the associated topology. Since $\{x\}$ is open for $x \neq x_0$, X is a T_0 space, and since $\{x\}$ is not closed for $x \neq x_0$, X is not T_1. ✌

[3] T stands for German *Trennung*, i.e., separation.

Example 3.3.9 Let $(D, \leq)$ be a partially ordered set. The topology associated with the closure operator for this order according to Example 3.1.20 is T_1 iff $y \leq x \Leftrightarrow x = y$, because this is what $\{x\}^{\mathbf{c}} = \{x\}$ says. ✌

Example 3.3.10 Let $X := \mathbb{N}$, and $\tau := \{A \subseteq \mathbb{N} \mid A \text{ is cofinite}\} \cup \{\emptyset\}$. Recall that a cofinite set is defined as having a finite complement. Then τ is a topology on X such that $X \setminus \{x\}$ is open for each $x \in X$. Hence X is a T_1-space. But X is not a Hausdorff. If $x \neq y$ and U is an open neighborhood of x, then $X \setminus U$ is finite. Thus if V is disjoint from U, we have $V \subseteq X \setminus U$. But then V cannot be an open set with $y \in V$. ✌

While the properties discussed so far deal with the relationship of two different points to each other, the next group of axioms looks at closed sets; given a closed set F, we call an open set U with $F \subseteq U$ an *open neighborhood* of F. Let again X be a topological space.

$T_3, T_{3\frac{1}{2}}, T_4$

T_3-space: X is a T_3-*space* iff given a point x and a closed set F, which does not contain x, there exist disjoint open neighborhoods of x and of F.

$T_{3\frac{1}{2}}$-space: X is a $T_{3\frac{1}{2}}$-*space* iff given a point x and a closed set F with $x \notin F$ there exists a continuous function $f : X \to \mathbb{R}$ with $f(x) = 1$ and $f(y) = 0$ for all $y \in F$.

T_4-space: X is a T_4-*space* iff two disjoint closed sets have disjoint open neighborhoods.

T_3 and T_4 deal with the possibility of separating a closed set from a point resp. another closed set. $T_{3\frac{1}{2}}$ is squeezed in between these axioms. Because $\{x \in X \mid f(x) < 1/2\}$ and $\{x \in X \mid f(x) > 1/2\}$ are disjoint open sets, it is clear that each $T_{3\frac{1}{2}}$-space is a T_3-space. It is also clear that the defining property of $T_{3\frac{1}{2}}$ is a special property of T_4, provided singletons are closed. The relationship and further properties will be explored now.

It might be noted that continuous functions play now an important rôle here in separating objects. $T_{3\frac{1}{2}}$ entails among others that there are "enough" continuous functions. Engelking [Eng89, p. 29 and 2.7.17] mentions that there are spaces which satisfy T_3 but have only constant continuous functions, and comments "they are, however, fairly complicated ..." (p. 29); Kuratowski [Kur66, p. 121] makes a similar remark.

So we will leave it at that and direct the reader, who wants to know more, to these sources and the papers quoted there.

We look at some examples.

Example 3.3.11 Let $X := \{1, 2, 3, 4\}$.

1. With the indiscrete topology $\{\emptyset, X\}$, X is a T_3- space, but it is neither T_2 nor T_1.
2. Take the topology $\big\{\{1\}, \{1, 2\}, \{1, 3\}, \{1, 2, 3\}, X, \emptyset\big\}$; then two closed sets are only disjoint when one of them is empty, because all of them contain the point 4 (with the exception of $\emptyset$, of course). Thus the space is T_4. The point 1 and the closed set $\{4\}$ cannot be separated by open sets; thus the space is not T_3.

✌

The next example displays a space which is T_2 but not T_3.

Example 3.3.12 Let $X := \mathbb{R}$, and put $Z := \{1/n \mid n \in \mathbb{N}\}$. Define in addition for $x \in \mathbb{R}$ and $i \in \mathbb{N}$ the sets $B_i(x) := \,]x-1/i, x+1/i[$. Then $\mathfrak{U}_0(x) := \{B_i(x) \mid i \in \mathbb{N}\}$ for $x \neq 0$, and $\mathfrak{U}_0(0) := \{B_i(0) \setminus Z \mid i \in \mathbb{N}\}$ define neighborhood filters for a Hausdorff space by Lemma 3.3.4. But this is not a T_3-space. One notes first that Z is closed: if $x \notin Z$ and $x \notin [0, 1]$, one certainly finds $i \in \mathbb{N}$ with $B_i(x) \cap Z = \emptyset$, and if $0 < x \leq 1$, there exists k with $1/(k+1) < x < 1/k$, so taking $1/i$ less than the minimal distance of x to $1/k$ and $1/(k+1)$, one has $B_i(x) \cap Z = \emptyset$. If $x = 0$, each neighborhood contains an open set which is disjoint from Z. Now each open set U which contains Z contains also 0, so we cannot separate 0 from Z. ✌

Just one positive message: the reals satisfy $T_{3\frac{1}{2}}$.

Example 3.3.13 Let $F \subseteq \mathbb{R}$ be nonempty, then

$$f(t) := \inf_{y \in F} \frac{|t-y|}{1+|t-y|}$$

defines a continuous function $f : \mathbb{R} \to [0, 1]$ with $z \in F \Leftrightarrow f(z) = 0$. Thus, if $x \notin F$, we have $f(x) > 0$, so that $y \mapsto f(y)/f(x)$ is a continuous function with the desired properties. Thus the reals with the usual topology are a $T_{3\frac{1}{2}}$-space. ✌

The next proposition is a characterization of T_3-spaces in terms of open neighborhoods, motivated by the following observation. Take a point $x \in \mathbb{R}$ and an open set $G \subseteq \mathbb{R}$ with $x \in G$. Then there exists $r > 0$ such that the open interval $]x - r, x + r[$ is entirely contained in G. But we can say more: by making this open interval a little bit smaller, we can actually fit a closed interval around x into the given neighborhood as well, so, for example, $x \in]x - r/2, x + r/2[\subseteq [x - r/2, x + r/2] \subseteq]x - r, x + r[\subseteq G$. Thus we find for the given neighborhood another neighborhood, the closure of which is entirely contained in it.

Proposition 3.3.14 *Let X be a topological space. Then the following are equivalent:*

1. *X is a T_3-space.*
2. *For every point x and every open neighborhood U of x, there exists an open neighborhood V of x with $V^a \subseteq U$.*

Proof 1 $\Rightarrow$ 2: Let U be an open neighborhood of x, then x is not contained in the closed set $X \setminus U$, so by T_3 we find disjoint open sets U_1, U_2 with $x \in U_1$ and $X \setminus U \subseteq U_2$; hence $X \setminus U_2 \subseteq U$. Because $U_1 \subseteq X \setminus U_2 \subseteq U$, and $X \setminus U_2$ is closed, we conclude $U_1^a \subseteq U$.

2 $\Rightarrow$ 1: Assume that we have a point x and a closed set F with $x \notin F$. Then $x \in X \setminus F$, so that $X \setminus F$ is an open neighborhood of x. By assumption, there exists an open neighborhood V of x with $x \in V^a \subseteq X \setminus F$; then V and $X \setminus (V^a)$ are disjoint open neighborhoods of x resp. F. $\dashv$

This characterization can be generalized to T_4-spaces (roughly, by replacing the point through a closed set) in the following way.

Proposition 3.3.15 *Let X be a topological space. Then the following are equivalent:*

1. *X is a T_4-space.*
2. *For every closed set F and every open neighborhood U of F, there exists an open neighborhood V of F with $F \subseteq V \subseteq V^a \subseteq U$.*

The proof of this proposition is actually nearly a copy of the preceding one, *mutatis mutandis*.

Proof 1 $\Rightarrow$ 2: Let U be an open neighborhood of the closed set F, then the closed set $F' := X \setminus U$ is disjoint from F, so that we can find disjoint open neighborhoods U_1 of F and U_2 of F'; thus $U_1 \subseteq X \setminus U_2 \subseteq X \setminus F' = U$, so $V := U_1$ is the open neighborhood we are looking for.

2 $\Rightarrow$ 1: Let F and F' be disjoint closed sets, then $X \setminus F'$ is an open neighborhood for F. Let V be an open neighborhood for F with $F \subseteq V \subseteq V^a \subseteq X \setminus F'$, then V and $U := X \setminus (V^a)$ are disjoint open neighborhoods of F and F'. $\dashv$

We mentioned above that the separation axiom $T_{3\frac{1}{2}}$ makes sure that there are enough continuous functions on the space. Actually, the continuous functions even determine the topology in this case, as the following characterization shows.

Proposition 3.3.16 *Let X be a topological space, then the following statements are equivalent:*

1. *X is a $T_{3\frac{1}{2}}$-space.*
2. *$\beta := \{f^{-1}[U] \mid f : X \to \mathbb{R}$ is continuous, $U \subseteq \mathbb{R}$ is open$\}$ constitutes a basis for the topology of X.*

Proof The elements of β are open sets, since they are comprised of inverse images of open sets under continuous functions.

1 $\Rightarrow$ 2: Let $G \subseteq X$ be an open set with $x \in G$. We show that we can find $B \in \beta$ with $x \in B \subseteq G$. In fact, since X is $T_{3\frac{1}{2}}$, there exists a continuous function $f : X \to \mathbb{R}$ with $f(x) = 1$ and $f(y) = 0$ for $y \in X \setminus G$. Then $B := \{x \in X \mid -\infty < x < 1/2\} = f^{-1}\big[]-\infty, 1/2[\big]$ is a suitable element of β.

2 $\Rightarrow$ 1: Take $x \in X$ and a closed set F with $x \notin F$. Then $U := X \setminus F$ is an open neighborhood x. Then we can find $G \subseteq \mathbb{R}$ open and $f : X \to \mathbb{R}$ continuous with $x \in f^{-1}\big[G\big] \subseteq U$. Since G is the union of open intervals, we find an open interval $I :=]a, b[\subseteq G$ with $f(x) \in I$. Let $G : \mathbb{R} \to \mathbb{R}$ be a continuous with $g(f(x)) = 1$ and $g(t) = 0$, if $t \notin I$; such a function exists since $\mathbb{R}$ is a $T_{3\frac{1}{2}}$-space

(Example 3.3.13). Then $g \circ f$ is a continuous function with the desired properties. Consequently, X is a $T_{3\frac{1}{2}}$-space. $\dashv$

The separation axioms give rise to names for classes of spaces. We will introduce their traditional names now.

Definition 3.3.17 *Let X be a topological space, then X is called:*

- regular *iff X satisfies T_1 and T_3,*
- completely regular *iff X satisfies T_1 and $T_{3\frac{1}{2}}$,*
- normal *iff X satisfies T_1 and T_4.*

The reason T_1 is always included is that one wants to have every singleton as a closed set, which, as the examples above show, is not always the case. Each regular space is a Hausdorff space, each regular space is completely regular, and each normal space is regular. We will obtain as a consequence of Urysohn's Lemma that each normal space is completely regular as well (Corollary 3.3.24).

In a completely regular space, we can separate a point x from a closed set not containing x through a continuous function. It turns out that normal spaces have an analogous property: Given two disjoint closed sets, we can separate these sets through a continuous function. This is what *Urysohn's Lemma* says, a famous result from the beginnings of set-theoretic topology. To be precise:

Theorem 3.3.18 *(Urysohn) Let X be a normal space. Given disjoint closed sets F_0 and F_1, there exists a continuous function $f : X \to \mathbb{R}$ such that $f(x) = 0$ for $x \in F_0$ and $f(x) = 1$ for $x \in F_1$.*

We need some technical preparations for proving Theorem 3.3.18; this gives also the opportunity to introduce the concept of a dense set.

Definition 3.3.19 *A subset $D \subseteq X$ of a topological space X is called* dense *iff $D^a = X$.*

Dense sets are fairly practical when it comes to comparing continuous functions for equality: if it suffices that the functions coincide on a dense set, then they will be equal. Just for the record:

Lemma 3.3.20 *Let $f, g : X \to Y$ be continuous maps with Y Hausdorff, and assume that $D \subseteq X$ is dense. Then $f = g$ iff $f(x) = g(x)$ for all $x \in D$.*

Proof Clearly, if $f = g$, then $f(x) = g(x)$ for all $x \in D$. So we have to establish the other direction.

Because Y is a Hausdorff space, $\Delta_Y := \{\langle y, y\rangle \mid y \in Y\}$ is closed (Proposition 3.3.1), and because $f \times g : X \times X \to Y \times Y$ is continuous, $(f \times g)^{-1}\big[\Delta_Y\big] \subseteq X \times X$ is closed as well. The latter set contains Δ_D, hence its closure Δ_X. $\dashv$

It is immediate that if D is dense, then $U \cap D \neq \emptyset$ for each open set U, so in particular each neighborhood of a point meets the dense set D. To provide an easy example, both $\mathbb{Q}$ and $\mathbb{R} \setminus \mathbb{Q}$ are dense subsets of $\mathbb{R}$ in the usual topology. Note that $\mathbb{Q}$ is countable, so $\mathbb{R}$ has even a countable dense set.

The first lemma has a family of subsets indexed by a dense subset of $\mathbb{R}$ exhaust a given set and provides a useful real function.

Lemma 3.3.21 *Let M be set, $D \subseteq \mathbb{R}_+$ be dense, and $(E_t)_{t \in D}$ be a family of subsets of M with these properties:*

- *if $t < s$, then $E_t \subseteq E_s$,*
- *$M = \bigcup_{t \in D} E_t$.*

Put $f(m) := \inf\{t \in D \mid m \in E_t\}$, then we have for all $s \in \mathbb{R}$:

1. *$\{m \mid f(m) < s\} = \bigcup\{E_t \mid t \in D, t < s\}$,*
2. *$\{m \mid f(m) \leq s\} = \bigcap\{E_t \mid t \in D, t > s\}$.*

Proof 1. Let us work on the first equality. If $f(m) < s$, there exists $t < s$ with $m \in E_t$. Conversely, if $m \in E_t$ for some $t < s$, then $f(m) = \inf\{r \in D \mid m \in E_r\} \leq t < s$.

2. For the second equality, assume $f(m) \leq s$, then we can find for each $r > s$ some $t < r$ with $m \in E_t \subseteq E_r$. To establish the other inclusion, assume that $f(m) \leq t$ for all $t > s$. If $f(m) = r > s$, we can find some $t' \in D$ with $r > t' > s$; hence $f(m) \leq t'$. This is a contradiction; hence $f(m) \leq s$. $\dashv$

This lemma, which does not assume a topology on M by requiring only a plain set, is extended now for the topological scenario in which we will use it. We assume that each set E_t is open, and we assume that E_t contains the closures of its predecessors. Then it will turn out that the function we just have defined is continuous, specifically:

Lemma 3.3.22 *Let X be a topological space and $D \subseteq \mathbb{R}_+$ a dense subset, and assume that $(E_t)_{t \in D}$ is a family of open sets with these properties:*

- *if $t < s$, then $E_t^a \subseteq E_s$,*
- *$X = \bigcup_{t \in D} E_t$.*

Then $f : x \mapsto inf\{t \in D \mid x \in E_t\}$ defines a continuous function on X.

Proof 0. Because a subbase for the topology on $\mathbb{R}$ is comprised of the intervals $]-\infty, x[$ resp. $]x, +\infty[$, we see from Lemma 3.1.9 that it is sufficient to show that for any $s \in \mathbb{R}$ the sets $\{x \in X \mid f(x) < s\}$ and $\{x \in X \mid f(x) > s\}$ are open, since they are the corresponding inverse images under f. For the latter set, we show that its complement $\{x \in X \mid f(x) \leq s\}$ is closed. Fix $s \in \mathbb{R}$.

1. We obtain from Lemma 3.3.21 that $\{x \in X \mid f(x) < s\}$ equals $\bigcup\{E_t \mid t \in D, t < s\}$; since all sets E_t are open, their union is. Hence $\{x \in X \mid f(x) < s\}$ is open.

2. We obtain again from Lemma 3.3.21 that $\{x \in X \mid f(x) \leq s\}$ equals $\bigcap\{E_t \mid t \in D, t > s\}$, so if we can show that $\bigcap\{E_t \mid t \in D, t > s\} = \bigcap\{E_t^a \mid t \in D, t > s\}$, we are done. In fact, the left-hand side is contained in the right-hand side, so assume that x is an element of the right-hand side. If x is not contained in the left-hand side, we find $t' > s$ with $t' \in D$ such that $x \notin E_{t'}$. Because D is dense, we find some r with $s < r < t'$ with $E_r^a \subseteq E_{t'}$. But then $x \notin E_r^a$; hence $x \notin \bigcap\{E_t^a \mid t \in D, t > s\}$, a contradiction. Thus both sets are equal, so that $\{x \in X \mid f(x) \geq s\}$ is closed. $\dashv$

We are now in a position to establish Urysohn's Lemma. The idea of the proof rests on this observation for a T_4-space X: suppose that we have open sets A and B with $A \subseteq A^a \subseteq B$. Then we can find an open set C such that $A^a \subseteq C \subseteq C^a \subseteq B$; see Proposition 3.3.15. Denote just for the proof for open sets A, B the fact that $A^a \subseteq B$ by $A \sqsubseteq^* B$. Then we may express the idea above by saying that $A \sqsubseteq^* B$ implies the existence of an open set C with $A \sqsubseteq^* C \sqsubseteq^* B$, so C may be squeezed in. But now we have $A \sqsubseteq^* C$ and $C \sqsubseteq^* B$, so we find open sets E and F with $A \sqsubseteq^* E \sqsubseteq^* C$ and $C \sqsubseteq^* F \sqsubseteq^* B$, arriving at the chain $A \sqsubseteq^* E \sqsubseteq^* C \sqsubseteq^* F \sqsubseteq^* B$. But why stop here?

Idea of the proof

$A \sqsubseteq^* B$

The proof makes this argument systematic and constructs in this way a continuous function.

Proof (of Theorem 3.3.18) 1. Let $D := \{p/2^q \mid p, q \text{ nonnegative integers}\}$. These are all dyadic numbers, which are dense in $\mathbb{R}_+$. We are about to construct a family $(E_t)_{t\in D}$ of open sets E_t indexed by D in the following way.

2. Put $E_t := X$ for $t > 1$, and let $E_1 := X \setminus F_1$; moreover, let E_0 be an open set containing F_0 which is disjoint from E_1. We now construct open sets $E_{p/2^n}$ by induction on n in the following way. Assume that we have already constructed open sets

$$E_0 \sqsubseteq^* E_{\frac{1}{2^{n-1}}} \sqsubseteq^* E_{\frac{2}{2^{n-1}}} \ldots \sqsubseteq^* E_{\frac{2^{n-1}-1}{2^{n-1}}} \sqsubseteq^* E_1.$$

Let $t = \frac{2m+1}{2^n}$, then we find an open set E_t with $E_{\frac{2m}{n}} \sqsubseteq^* E_t \sqsubseteq^* E_{\frac{2m+2}{2^n}}$; we do this for all m with $0 \leq m \leq 2^{n-1} - 1$.

3. Look as an illustration at the case $n = 3$. We have found already the open sets $E_0 \sqsubseteq^* E_{1/4} \sqsubseteq^* E_{1/2} \sqsubseteq^* E_{3/4} \sqsubseteq^* E_1$. Then the construction goes on with finding open sets $E_{1/8}, E_{3/8}, E_{5/8}$, and $E_{7/8}$ such that after the step is completed, we obtain this chain:

$$\begin{aligned} E_0 \sqsubseteq^* E_{1/8} \sqsubseteq^* E_{1/4} \sqsubseteq^* E_{3/8} \sqsubseteq^* E_{1/2} \sqsubseteq^* E_{5/8} \\ \sqsubseteq^* E_{3/4} \sqsubseteq^* E_{7/8} \sqsubseteq^* E_1. \end{aligned}$$

4. In this way, we construct a family $(E_t)_{t\in D}$ with the properties requested by Lemma 3.3.22. It yields a continuous function $f : X \to \mathbb{R}$ with $f(x) = 0$ for all $x \in F_0$ and $f(1)(x) = 1$ for all $x \in F_1$. $\dashv$

Urysohn's Lemma is used to prove the Tietze extension theorem, which we will only state, but not prove.

Theorem 3.3.23 *Let X be a T_4-space and $f : A \to \mathbb{R}$ be a function which is continuous on a closed subset A of X. Then f can be extended to a continuous function f^* on all of X.* $\dashv$

We obtain as an immediate consequence of Urysohn's Lemma:

Corollary 3.3.24 *A normal space is completely regular.*

We have obtained a hierarchy of spaces through gradually tightening the separation properties and found that continuous functions help with the separation. The question arises, how compactness fits into this hierarchy. It turns out that a compact Hausdorff space is normal; the converse obviously does not hold: the reals with the Euclidean topology are normal, but by no means compact.

We call a subset K in a topological space X compact iff it is compact as a subspace, i.e., a compact topological space in its own right. This is a first and fairly straightforward observation; see Lemma 1.5.55.

Lemma 3.3.25 *A closed subset F of a compact space X is compact.*

Proof Let $(G_i \cap F)_{i \in I}$ be an open cover of F with $G_i \subseteq X$ open; then $\{F\} \cup \{G_i \mid i \in I\}$ is an open cover of X, so we can find a finite subset $J \subseteq I$ such that $\{F\} \cup \{G_j \mid i \in J\}$ covers X; hence $\{G_i \cap F \mid i \in J\}$ covers F. $\dashv$

In a Hausdorff space, the converse holds as well:

Lemma 3.3.26 *Let X be a Hausdorff space and $K \subseteq X$ compact, then:*

1. *Given $x \notin K$, there exist disjoint open neighborhoods U of x and V of K.*
2. *K is closed.*

Proof Given $x \notin K$, we want to find $U \in \mathfrak{U}(x)$ with $U \cap K = \emptyset$ and $V \supseteq K$ open with $U \cap V = \emptyset$.

Let us see how to do that. There exists for x and any element $y \in K$ disjoint open neighborhoods $U_y \in \mathfrak{U}(x)$ and $W_y \in \mathfrak{U}(y)$, because X is Hausdorff. Then $(W_y)_{y \in Y}$ is an open cover of K; hence by compactness, there exists a finite subset $W_0 \subseteq W$ such that $\{W_y \mid y \in W_0\}$ covers K. But then $\bigcap_{y \in W_0} U_y$ is an open neighborhood of x which is disjoint from $V := \bigcup_{y \in W_0} W_y$, hence from K. V is the open neighborhood of K we are looking for. This establishes the first part; the second follows as an immediate consequence. $\dashv$

Look at the reals as an illustrative example.

Corollary 3.3.27 *$A \subseteq \mathbb{R}$ is compact iff it is closed and bounded.*

Proof If $A \subseteq \mathbb{R}$ is compact, then it is closed by Lemma 3.3.26, since $\mathbb{R}$ is a Hausdorff space. Since A is compact, it is also bounded. If, conversely, $A \subseteq \mathbb{R}$ is closed and bounded, then we can find a closed interval $[a, b]$ such that $A \subseteq [a, b]$. We know from the Heine–Borel

Theorem 1.5.46 that this interval is compact, and a closed subset of a compact space is compact by Lemma 3.3.25. $\dashv$

This has yet another, frequently used consequence, viz., that a continuous real-valued function on a compact space assumes its minimal and its maximal value. Just for the record:

Corollary 3.3.28 *Let X be a compact Hausdorff space and $f : X \to \mathbb{R}$ a continuous map. Then there exist $x_*, x^* \in X$ with $f(x_*) = \min f\big[X\big]$ and $f(x^*) = \max f\big[X\big]$.* $\dashv$

But—after traveling an interesting side path—let us return to the problem of establishing that a compact Hausdorff space is normal. We know now that we can separate a point from a compact subset through disjoint open neighborhoods. This is but a small step from establishing the solution to the above problem.

Proposition 3.3.29 *A compact Hausdorff space is normal.*

Proof Let X be compact and A and B disjoint closed subsets. Since X is a Hausdorff, A and B are compact as well. Now the rest is an easy application of Lemma 3.3.26. Given $x \in B$, there exist disjoint open neighborhoods $U_x \in \mathfrak{U}(x)$ of x and V_x of A. Let B_0 be a finite subset of B such that $U := \bigcup\{U(x) \mid b \in B_0\}$ covers B and $V := \bigcap\{V_x \mid x \in B_0\}$ is an open neighborhood of A. U and V are disjoint. $\dashv$

From the point of view of separation, to be compact is a stronger property than being normal for a topological space. The example $\mathbb{R}$ shows that this is a strictly stronger property. We will show now that $\mathbb{R}$ is just one point apart from being compact by investigating locally compact spaces.

3.4 Local Compactness and Compactification

We restrict ourselves in this section to Hausdorff spaces.

Sometimes a space is not compact but has enough compact subsets, because each point has a compact neighborhood. These spaces are called locally compact, and we investigate properties they share with and properties they distinguish them from compact spaces. We show also that a locally compact space misses being compact by just one point. Adding this point will make it compact, so we have an example here where we embed a space into one with a desired property. While we are compactifying spaces, we also provide another one, named after Stone and Čech, which requires the basic space to be completely regular. We establish another classic, the Baire Theorem, which states that in a locally compact T_3-space, the intersection of a countable number of open dense sets is dense again; applications will capitalize on this observation, as we will see.

Definition 3.4.1 *Let X be a Hausdorff space. X is called* locally compact *iff for each $x \in X$ and each open neighborhood $U \in \mathfrak{U}(x)$ there exists a neighborhood $V \in \mathfrak{U}(x)$ such that V^a is compact and $V^a \subseteq U$.*

Thus the compact neighborhoods form a basis for the neighborhood filter for each point. This implies that we can find for each compact subset an open neighborhood with compact closure. The proof of this property gives an indication of how to argue in locally compact spaces.

Proposition 3.4.2 *Let X be a locally compact space and K a compact subset. Then there exists an open neighborhood U of K and a compact set K' with $K \subseteq U \subseteq K'$.*

Proof Let $x \in K$, then we find an open neighborhood $U_x \in \mathfrak{U}(x)$ with U_x^a compact. Then $(U_x)_{x \in K}$ is a cover for K, and there exists a finite subset $K_0 \subseteq K$ such that $(U_x)_{x \in K_0}$ covers K. Put $U := \bigcup_{x \in K_0}$, and note that this open set has a compact closure. $\dashv$

So this is not too bad: We have plenty of compact sets in a locally compact space. Such a space is very nearly compact. We add to X just one point, traditionally called ∞, and define the neighborhood for ∞ in such a way that the resulting space is compact. The obvious way to do that is to make all complements of compact sets a neighborhood of ∞, because it will then be fairly easy to construct a finite subcover from a cover of the new space. This is what the compactification which we discuss now will do for you. We carry out the construction in a sequence of lemmata, just in order to render the process a bit more transparent.

Lemma 3.4.3 *Let X be a Hausdorff space with topology τ and $\infty \notin X$ be a distinguished new point. Put $X^* := X \cup \{\infty\}$, and define* One point extension

$$\tau^* := \{U \subseteq X^* \mid U \cap X \in \tau\} \cup \{U \subseteq X^* \mid \infty \in U, X \setminus U \text{ is compact}\}.$$

Then τ^* is a topology on X^*, and the identity $i_X : X \to X^*$ is τ-τ^*-continuous.

Proof $\emptyset$ and X^* are obviously members of τ^*; note that $X \setminus U$ being compact entails $U \cap X$ being open. Let $U_1, U_2 \in \tau^*$. If $\infty \in U_1 \cap U_2$, then $X \setminus (U_1 \cap U_2)$ is the union of two compact sets in X, hence is compact. If $\infty \notin U_1 \cap U_2$, $X \cap (U_1 \cap U_2)$ is open in X. Thus τ^* is closed under finite intersections. Let $(U_i)_{i \in I}$ be a family of elements of τ^*. The critical case is that $\infty \in U := \bigcup_{i \in I} U_i$, say, $\infty \in U_j$. But then $X \setminus U \subseteq X \subseteq U_j$, which is compact, so that $U \in \tau^*$. Continuity of i_X is now immediate. $\dashv$

We find X in this new construction as a subspace.

Corollary 3.4.4 (X, τ) *is a dense subspace of* (X^*, τ^*).

Proof We have to show that $\tau = \tau^* \cap X$. But this is obvious from the definition of τ^*. $\dashv$

Now we can state and prove the result which has been announced above.

Theorem 3.4.5 *Given a Hausdorff space X, the one point extension X^* is a compact space, in which X is dense. If X is locally compact, X^* is a Hausdorff space.*

Proof It remains to show that X^* is compact and that it is a Hausdorff space, whenever X is locally compact.

Let $(U_i)_{i \in I}$ be an open cover of X^*, then $\infty \in U_j$ for some $j \in I$; thus $X \setminus U_j$ is compact and is covered by $(U_i)_{i \in I, i \neq j}$. Select a finite subset $J \subseteq I$ such that $(U_i)_{i \in J}$ covers $X \setminus U_j$, then—voilà—we have found a finite cover $(U_i)_{i \in J \cup \{j\}}$ of X^*.

Since the given space is Hausdorff, we have to separate the new point ∞ from a given point $x \in X$, provided X is locally compact. But take a compact neighborhood U of x; then $X^* \setminus U$ is an open neighborhood of ∞. $\dashv$

X^* is called the *Alexandrov one-point compactification* of X. The new point is sometimes called the *infinite point*. It is not difficult to show that two different one-point compactifications are homeomorphic, so we may talk about *the* (rather than *a*) one-point compactification.

Looking at the map $i_X : X \to X^*$, which permits looking at elements of X as elements of X^*, we see that i_X is injective and has the property that $i_X\left[G\right]$ is an open set in the image $i_X\left[X\right]$ of X in X^*, whenever $G \subseteq X$ is open. These properties will be used for characterizing compactifications. Let us first define embeddings, which are of interest independently of this process.

Definition 3.4.6 *The continuous map $f : X \to Y$ between the topological spaces X and Y is called an* embedding *iff:*

- *f is injective,*
- *$f\left[G\right]$ is open in $f\left[X\right]$, whenever $G \subseteq X$ is open.*

So if $f : X \to Y$ is an embedding, we may recover a true image of X from its image $f\left[X\right]$, so that $f : X \to f\left[X\right]$ is a homeomorphism.

Consider again the map $[0,1]^N \to [0,1]^M$ which is induced by a map $f : M \to N$ and which we dealt with in Lemma 3.1.16. We will put this map to good use in a moment, so it is helpful to analyze it a bit more closely.

Example 3.4.7 Let $f : M \to N$ be a surjective map. Then $f^* : [0,1]^N \to [0,1]^M$, which sends $g : N \to [0,1]$ to $g \circ f : M \to [0,1]$ is an embedding. We have to show that f^* is injective and that it maps open sets into open sets in the image. This is done in two steps:

f^* is injective: In fact, if $g_1 \neq g_2$, we find $n \in N$ with $g_1(n) \neq g_2(n)$, and because f is onto, we find m with $n = f(m)$; hence $f^*(g_1)(m) = g_1(f(m)) \neq g_2(f(m)) = f^*(g_2)(m)$. Thus $f^*(g_1) \neq f(g_2)$ (an alternative proof is proposed in Lemma 2.1.29 in a more general scenario).

Open sets are mapped to open sets: We know already from Lemma 3.1.16 that f^* is continuous, so we have to show that the image $f\left[G\right]$ of an open set $G \subseteq [0,1]^N$ is open in the subspace $f\left[[0,1]^M\right]$. Let $h \in f\left[G\right]$; hence $h = f^*(g)$ for some $g \in G$. G is open; thus we can find a base element H of the product topology with $g \in H \subseteq G$, say, $H = \bigcap_{i=1}^{k} \pi_{N,n_i}^{-1}\left[H_i\right]$ for some $n_1, \ldots, n_k \in N$ and some open subsets $H_1, \ldots, H_k$ in $[0,1]$. Since f is onto, $n_1 = f(m_1), \ldots, n_k = f(m_k)$ for some $m_1, \ldots, m_k \in M$. Since $h \in \pi_{N,n_i}^{-1}\left[H_i\right]$ iff $f^*(h) \in \pi_{M,m_i}^{-1}\left[H_i\right]$, we obtain

$$\begin{aligned} h &= f^*(g) \in f^*\Big[\bigcap_{i=1}^{k} \pi_{N,n_i}^{-1}\left[H_i\right]\Big] = \Big(\bigcap_{i=1}^{k} \pi_{M,m_i}^{-1}\left[H_i\right]\Big) \\ &\quad \cap f^*\left[[0,1]^N\right] \end{aligned}$$

The latter set is open in the image of $[0,1]^N$ under f^*, so we have shown that the image of an open set is open relative to the subset topology of the image.

These proofs will serve as patterns later on. ✌

Given an embedding, we define the compactification of a space.

Definition 3.4.8 *A pair (e, Y) is said to be a* compactification *of a topological space X iff Y is a compact topological space and if $e : X \to Y$ is an embedding.*

The pair (i_X, X^*) constructed as the Alexandrov one-point compactification is a compactification in the sense of Definition 3.4.8, provided the space X is locally compact. We are about to construct another important compactification for a completely regular space X. Define for X the space βX as follows[4]: Let $F(X)$ be the set of all continuous maps $X \to [0,1]$ and map x to its evaluations from $F(X)$, so construct $e_X : X \ni x \mapsto (f(x))_{f \in F(X)} \in [0,1]^{F(X)}$. Then $\beta X := (e_X\left[X\right])^a$, the closure being taken in the compact space $[0,1]^{F(X)}$. We claim that $(e_X, \beta X)$ is a compactification of X.

Before delving into the proof, we note that we want to have a completely regular space, since in these spaces we have enough continuous functions, e.g., to separate points, as will become clear shortly. We will first show that this is a compactification indeed, and then investigate an interesting property of it.

Proposition 3.4.9 $(e_X, \beta X)$ *is a compactification of the completely regular space X.*

[4]It is a bit unfortunate that there appears to be an ambiguity in notation, since we denote the basis of a topological space by β as well. But tradition demands this compactification to be called βX, and from the context it should be clear what we have in mind.

Proof 1. We take the closure in the Hausdorff space $[0,1]^{F(X)}$, which is compact by Tihonov's Theorem 3.2.12. Hence βX is a compact Hausdorff space by Lemma 3.3.25.

2. e_X is continuous, because we have $\pi_f \circ e_X = f$ for $f \in F(X)$, and each f is continuous. e_X is also injective, because we can find for $x \neq x'$ a map $f \in F(X)$ such that $f(x) \neq f(x')$; this translates into $e_X(x)(f) \neq e_X(x')(f)$; hence $e_X(x) \neq e_X(x')$.

3. The image of an open set in X is open in the image. In fact, let $G \subseteq X$ be open, and take $x \in G$. Since X is completely regular, we find $f \in F(X)$ and an open set $U \subseteq [0,1]$ with $x \in f^{-1}\left[U\right] \subseteq G$; this is so because the inverse images of the open sets in $[0,1]$ under continuous functions form a basis for the topology (Proposition 3.3.16). But $x \in f^{-1}\left[U\right] \subseteq G$ is equivalent to $x \in (\pi_f \circ e_X)^{-1}\left[U\right] \subseteq G$. Because $e_X : X \to e_X\left[X\right]$ is a bijection, this implies $x \in \pi_f^{-1}\left[U\right] \subseteq e_X\left[G\right] \cap e_X\left[X\right] \subseteq e_X\left[G\right] \cap (e_X\left[X\right])^a$. Hence $e_X\left[G\right]$ is open in βX. $\dashv$

If the space we started from is already compact, then we obtain nothing new:

Corollary 3.4.10 *If X is a compact Hausdorff space, $e_X : X \to \beta X$ is a homeomorphism.*

Proof A compact Hausdorff space is normal, hence completely regular by Proposition 3.3.29 and Corollary 3.3.24, so we can construct the space βX for X compact. The assertion then follows from Exercise 3.10. $\dashv$

This kind of compactification is important, so it deserves a name.

Definition 3.4.11 *The compactification $(e_X, \beta X)$ is called the* Stone–Čech compactification *of the regular space X.*

This compactification permits the extension of continuous maps in the following sense: suppose that $f : X \to Y$ is continuous with Y compact, then there exists a continuous extension $\beta X \to Y$. This statement is slightly imprecise, because f is not defined on βX, so we want really to extend $f \circ e_X^{-1} : e_X\left[X\right] \to Y$ —since e_X is a homeomorphism from X onto its image, one tends to identify both spaces.

Theorem 3.4.12 *Let $(e_X, \beta X)$ be the Stone–Čech compactification of the completely regular space X. Then, given a continuous map f :*

$X \to Y$ with Y compact, there exists a continuous extension $f_! : \beta X \to Y$ to $f \circ e_X^{-1}$.

The idea of the proof is to capitalize on the compactness of the target space Y, because Y and βY are homeomorphic. This means that Y has a topologically identical copy in $[0,1]^{F(Y)}$, which may be used in a suitable fashion. The proof is adapted from [Kel55, p. 153]; Kelley calls it a "mildly intricate calculation."

Proof 1. Define $\varphi_f : F(Y) \to F(X)$ through $h \mapsto f \circ h$; then this map induces a map $\varphi_f^* : [0,1]^{F(X)} \to [0,1]^{F(Y)}$ by sending $t : F(X) \to [0,1]$ to $t \circ \varphi_f$. Then φ_f^* is continuous according to Lemma 3.1.16.

2. Consider this diagram:

$$\begin{tikzcd}
e_X[X] \arrow[r, "\subseteq"] & {[0,1]^{F(X)}} \arrow[r, "\varphi_f^*"] & {[0,1]^{F(Y)}} & \beta Y \arrow[l, "\supseteq"'] \\
X \arrow[u, "e_X"] \arrow[rrr, "f"'] & & & Y \arrow[u, "e_Y"']
\end{tikzcd}$$

We claim that $\varphi_f^* \circ e_X = e_Y \circ f$. In fact, take $x \in X$ and $h \in F(Y)$; then

$$\begin{aligned}
(\varphi_f^* \circ e_X)(x)(h) &= (e_X \circ \varphi_f)(h) = e_X(x)(h \circ f) \\
&= (h \circ f)(x) = e_Y(f(x))(h) \\
&= (e_Y \circ f)(x)(h).
\end{aligned}$$

3. Because Y is compact, e_Y is a homeomorphism by Exercise 3.10, and since φ_f^* is continuous, we have

$$\varphi_f^*[\beta X] = \varphi_f^*[e_X[X]^a] \subseteq \left(\varphi_f^*[e_X[X]]\right)^a \subseteq \beta Y.$$

Thus $e_X^{-1} \circ \varphi_f^*$ is a continuous extension to $f \circ e_X$. ⊣

It is immediate from Theorem 3.4.12 that a Stone–Čech compactification is uniquely determined, up to homeomorphism. This justifies the probably a bit prematurely used characterization as *the* Stone–Čech compactification above.

Baire's Theorem, which we will establish now, states a property of locally compact spaces which has a surprising range of applications—it states that the intersection of dense open sets in a locally compact

T_3-space is dense again. This applies of course to compact Hausdorff spaces as well. The theorem has a counterpart for complete pseudometric spaces, as we will see below. For stating and proving the theorem, we lift the assumption of working in a Hausdorff space, because it is really not necessary here.

Theorem 3.4.13 *Let X be a locally compact T_3-space. Then the intersection of dense open sets is dense.*

Proof 0. The idea of the proof is to construct for $\emptyset \neq G$ open a decreasing sequence $(V_n)_{n\in\mathbb{N}}$ of sets with $V_{n+1}^a \subseteq D_n \cap V_n$, where $(D_n)_{n\in\mathbb{N}}$ be a sequence of dense open sets, where V_1 is chosen so that $V_1^a \subseteq D_1 \cap G$. Then we will conclude from the finite intersection property for compact sets that G contains a point in the intersection $\bigcap_{n\in\mathbb{N}} D_n$. Idea of the proof

1. Fix a nonempty open set G; then we have to show that $G \cap \bigcap_{n\in\mathbb{N}} D_n \neq \emptyset$. Now D_1 is dense and open; hence we find an open set V_1 such that V_1^a is compact and $V_1^a \subseteq D_1 \cap G$ by Proposition 3.3.14, since X is a T_3-space. We select inductively in this way a sequence of open sets $(V_n)_{n\in\mathbb{N}}$ with compact closure such that $V_{n+1}^a \subseteq D_n \cap V_n$. This is possible since D_n is open and dense for each $n \in \mathbb{N}$.

2. Hence we have a decreasing sequence $V_2^a \supseteq \ldots V_n^a \supseteq \ldots$ of closed sets in the compact set V_1^a; thus $\bigcap_{n\in\mathbb{N}} V_n^a = \bigcap_{n\in\mathbb{N}} V_n$ is not empty, which entails $G \cap \bigcap_{n\in\mathbb{N}} D_n$ not being empty. ⊣

Just for the record:

Corollary 3.4.14 *The intersection of a sequence of dense open sets in a compact Hausdorff space is dense.*

Proof A compact Hausdorff space is normal by Proposition 3.3.29, hence regular by Proposition 3.3.15; thus the assertion follows from Theorem 3.4.13. ⊣

We give an example from Boolean algebras.

Example 3.4.15 Let B be a Boolean algebra with $\wp_B$ as the set of all prime ideals. Let $X_a := \{I \in \wp_B \mid a \notin I\}$ be all prime ideals which do not contain a given element $a \in B$, then $\{X_a \mid a \in B\}$ is the basis for a compact Hausdorff topology on $\wp_B$, and $a \mapsto X_a$ is a Boolean algebra isomorphism; see the proof of Theorem 1.5.45.

Assume that we have a countable family S of elements of B with $a = \sup S \in B$, then we say that the prime ideal I *preserves the supremum* of S iff $[a]_{\sim_I} = \sup_{s \in S} [s]_{\sim_I}$ holds. Here $\sim_I$ is the equivalence relation induced by I, i.e., $b \sim_I b' \Leftrightarrow b \ominus b' \in I$ with $\ominus$ as the symmetric difference in B (Lemma 1.5.40).

We claim that the set R of all prime ideals, which do *not* preserve the supremum of this family, is closed and has an empty interior. Well, $R = X_a \setminus \bigcup_{k \in K} X_{a_k}$. Because the sets X_a and X_{a_k} are clopen, R is closed. Assume that the interior of R is not empty; then we find $b \in B$ with $X_b \subseteq R$, so that $X_{a_k} \subseteq X_a \setminus X_b = X_{a \wedge -b}$ for all $k \in K$. Since $a \mapsto X_a$ is an isomorphism, this means $a_k \leq a \wedge -b$; hence $\sup_{k \in K} a_k \leq a \wedge -b$ for all $k \in K$; thus $a = a \wedge -b$, and hence $a \leq -b$. But then $X_b \subseteq X_a \subseteq X_{-b}$, which is certainly a contradiction. Consequently, the set of all prime ideal preserving this particular supremum is open and dense in $\wp_B$.

If we are given for each $n \in \mathbb{N}$ a family $S_n \subseteq B$ and $a_0 \in B$ such that:

- $a_0 \neq \top$, the maximal element of B,
- $a_n := \sup_{s \in S_n} s$ is an element of B for each $n \in \mathbb{N}$,

then we claim that there exists a prime ideal I which contains a_0 and which preserves all the suprema of S_n for $n \in \mathbb{N}$.

Let P be the set of all prime ideals which preserve all the suprema of the families above, then

$$P = \bigcap_{n \in \mathbb{N}} P_n,$$

where P_n is the set of all prime ideals which preserve the supremum a_n, which is dense and open by the discussion above. Hence P is dense by Baire's Theorem (Corollary 3.4.14). Since $X_{-a_0} = \wp_B \setminus X_{a_0}$ is open and not empty, we infer that $P \cap X_{-a_0}$ is not empty, because P is dense. Thus we can select an arbitrary prime ideal from this set. ✌

This example, which is taken from [RS50, Sect. 5], will help in establishing Gödel's Completeness Theorem; see Sect. 3.6.1. The approach is typical for an application of Baire's Theorem—it is used to show that a set P, which is obtained from an intersection of countably many open and dense sets in a compact space, is dense and that the object of one's desire is a member of P intersecting an open set; hence this object must exist.

Having been carried away by Baire's Theorem, let us return to the mainstream of the discussion and make some general remarks. We see that local compactness is a somewhat weaker property than compactness. Other notions of compactness have been studied; an incomplete list for Hausdorff space X includes:

countably compact: X is called *countably compact* iff each countable open cover contains a finite subcover.

Lindelöf space: X is a *Lindelöf space* iff each open cover contains a countable subcover.

paracompactness: X is said to be *paracompact* iff each open cover has a locally finite refinement. This explains it:

- An open cover $\mathcal{B}$ is a *refinement* of an open cover $\mathcal{A}$ iff each member of $\mathcal{B}$ is the subset of a member of $\mathcal{A}$.
- An open cover $\mathcal{A}$ is called *locally finite* iff each point has a neighborhood which intersects a finite number of elements of $\mathcal{A}$.

sequentially compact: X is called *sequentially compact* iff each sequence has a convergent subsequence (we will deal with this when discussing compact pseudometric spaces; see Proposition 3.5.31).

The reader is referred to [Eng89, Chap. 3] for a penetrating study.

3.5 Pseudometric and Metric Spaces

We turn to a class of spaces now in which we can determine the distance between any two points numerically. This gives rise to a topology, declaring a set as open iff we can construct for each of its points an open ball which is entirely contained in this set. It is clear that this defines a topology, and it is also clear that having such a metric gives the space some special properties, which are not shared by general topological spaces. It also adds a sense of visual clearness, since an open ball is conceptually easier to visualize that an abstract open set. We will study the topological properties of these spaces starting with pseudometrics, with which we may measure the distance between two objects, but if the distance is zero, we cannot necessarily conclude that the objects are identical. This is a situation which occurs quite frequently when modeling an application, so it is sometimes more adequate to deal with pseudometric rather than metric spaces.

Definition 3.5.1 *A map $d : X \times X \to \mathbb{R}_+$ is called a* pseudometric on *X iff these conditions hold:*

identity: $d(x, x) = 0$ *for all $x \in X$.*

symmetry: $d(x, y) = d(y, x)$ *for all $x, y \in X$,*

triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$ *for all $x, y, z \in X$.*

Then (X, d) is called a pseudometric space. *If, in addition, we have*

$$d(x, y) = 0 \Leftrightarrow x = y,$$

then d is called a metric on *X; accordingly, (X, d) is called a* metric space.

The nonnegative real number $d(x, y)$ is called the distance of the elements x and y in a pseudometric space (X, d). It is clear that one wants point to have distance 0 to itself and that the distance between two points is determined in a symmetric fashion. The triangle inequality is intuitively clear as well:

$$\begin{tikzcd} x \arrow[rr] \arrow[dr] & & y \\ & z \arrow[ur] & \end{tikzcd}$$

Before proceeding, let us have a look at some examples. Some of them will be discussed later on in greater detail.

Example 3.5.2
1. Define for $x, y \in \mathbb{R}$ the distance as $|x - y|$, hence as the absolute value of their difference. Then this defines a metric. Define, similarly,

$$d(x, y) := \frac{|x - y|}{1 + |x - y|},$$

then d defines also a metric on $\mathbb{R}$ (the triangle inequality follows from the observation that $a \leq b \Leftrightarrow a/(1 + a) \leq b/(1 + b)$ holds for nonnegative numbers a and b).

2. Given $x, y \in \mathbb{R}^n$ for $n \in \mathbb{N}$, then

$$d_1(x, y) := \max_{1 \leq i \leq n} |x_i - y_i|,$$
$$d_2(x, y) := \sum_{i=1}^{n} |x_i - y_i|,$$
$$d_3(x, y) := \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

   define all metrics an $\mathbb{R}^n$. Metric d_1 measures the maximal distance between the components, d_2 gives the sum of the distances, and d_3 yields the Euclidean, i.e., the geometric, distance of the given points. The crucial property to be established is in each case the triangle inequality. It follows for d_1 and d_2 from the triangle inequality for the absolute value and for d_3 by direct computation.

3. Given a set X, define

$$d(x, y) := \begin{cases} 0, & \text{if } x = y \\ 1, & \text{otherwise} \end{cases}$$

   Then (X, d) is a metric space, d is called the *discrete metric*. Different points are assigned the distance 1, while each point has distance 0 to itself.

4. Let X be a set, $\mathcal{D}(X)$ be the set of all bounded maps $X \to \mathbb{R}$. Define

$$d(f, g) := \sup_{x \in X} |f(x) - g(x)|.$$

   Then $(\mathcal{D}(X), d)$ is a metric space; the distance between functions f and g is just their maximal difference.

5. Similarly, given a set X, take a set $\mathcal{E} \subseteq \mathcal{D}(X)$ of bounded real-valued functions as a set of evaluations and determine the distance of two points in terms of their evaluations:

$$e(x, y) := \sup_{f \in \mathcal{F}} |f(x) - f(y)|.$$

   So two points are similar if their evaluations on terms of all elements of $\mathcal{F}$ are close. This is a pseudometric on X. It is not a metric if $\mathcal{F}$ does not separate points.

6. Denote by $\mathcal{C}([0,1])$ the set of all continuous real-valued functions $[0,1] \to \mathbb{R}$, and measure the distance between $f, g \in \mathcal{C}([0,1])$ through
$$d(f,g) := \sup_{0 \le x \le 1} |f(x) - g(x)|.$$
Because a continuous function on a compact space is bounded, $d(f,g)$ is always finite, and since for each $x \in [0,1]$ the inequality $|f(x) - g(x)| \le |f(x) - h(x)| + |h(x) - g(x)|$ holds, the triangle inequality is satisfied. Then $(\mathcal{C}([0,1]), d)$ is a metric space, because $\mathcal{C}([0,1])$ separates points.

7. Define for the Borel sets $\mathcal{B}([0,1])$ on the unit interval this distance:
$$d(A,B) := \lambda(A \Delta B)$$
with λ as Lebesgue measure. Then $\lambda(A\Delta B) = \lambda((A\Delta C)\Delta (C\Delta B) \le \lambda(A\Delta C) + \lambda(C\Delta B)$ implies the triangle inequality, so that $(\mathcal{B}([0,1]), d)$ is a pseudometric space. It is no metric space, however, because $\lambda(\mathbb{Q} \cap [0,1]) = 0$; hence $d(\emptyset, \mathbb{Q} \cap [0,1]) = 0$, but the latter set is not empty.

8. Given a nonempty set X and a ranking function $r : X \to \mathbb{N}$, define the closeness $c(A,B)$ of two subset A, B of X as
$$c(A,B) := \begin{cases} +\infty, & \text{if } A = B, \\ \inf\{r(w) \mid w \in A\Delta B\}, & \text{otherwise} \end{cases}$$
If $w \in A\Delta B$, then w can be interpreted as a witness that A and B are different, and the closeness of A and B is just the minimal rank of a witness. We observe these properties:

   - $c(A,A) = +\infty$, and $c(A,B) = +\infty$ iff $A = B$ (because $A = B$ iff $A\Delta B = \emptyset$).
   - $c(A,B) = c(B,A)$,
   - $c(A,C) \ge \min\{c(A,B), c(B,C)\}$. If $A = C$, this is obvious; assume otherwise that $b \in A\Delta C$ is a witness of minimal rank. Since $A\Delta C = (A\Delta B)\Delta(B\Delta C)$, b must be either in $A\Delta B$ or $B\Delta C$, so that $r(b) \ge c(A,C)$ or $r(b) \ge c(B,C)$.

   Now put $d(A,B) := 2^{-c(A,B)}$ (with $2^{-\infty} := 0$). Then d is a metric on $\mathcal{P}(X)$. This metric satisfies even $d(A,B) \le \max\{d(A,C),$

$d(B, C)\}$ for an arbitrary C, and hence d is an *ultrametric*; see Exercise 3.18.

9. A similar construction is possible with a decreasing sequence of equivalence relation on a set X. In fact, let $(\rho_n)_{n\in\mathbb{N}}$ be such a sequence, and put $\rho_0 := X \times X$. Define

$$c(x, y) := \begin{cases} +\infty, & \text{if } \langle x, y\rangle \in \bigcap_{n\in\mathbb{N}} \rho_n \\ \max\{n \in \mathbb{N} \mid \langle x, y\rangle \in \rho_n\}, & \text{otherwise} \end{cases}$$

Then it is immediate that $c(x, y) \geq \min\{c(x, z), c(z, y)\}$. Intuitively, $c(x, y)$ gives the degree of similarity of x and y—the larger this value, the more similar x and y are. Then

$$d(x, y) := \begin{cases} 0, & \text{if } c(x, y) = \infty \\ 2^{-c(x,y)}, & \text{otherwise} \end{cases}$$

defines a pseudometric, which is a metric iff $\bigcap_{n\in\mathbb{N}} \rho_n = \{\langle x, x\rangle \mid x \in X\}$.

✌

Given a pseudometric space (X, d), define for $x \in X$ and $r > 0$ the *open ball* $B(x, r)$ with center x and radius r as $B(x, r)$

$$B(x, r) := \{y \in X \mid d(x, y) < r\}.$$

The *closed ball* $S(x, r)$ is defined similarly as

$$S(x, r) := \{y \in X \mid d(x, y) \leq r\}.$$

If necessary, we indicate the pseudometric explicitly with B and S. Note that $B(x, r)$ is open, and $S(x, r)$ is closed, but that the closure $B(x, r)^a$ of $B(x, r)$ may be properly contained in the closed ball $S(x, r)$ (let d be the discrete metric, then $B(x, 1) = \{x\} = B(x, 1)^a$, but $S(x, 1) = X$, so both closed sets do not coincide if X has more than one point).

Call $G \subseteq X$ open iff we can find for each $x \in G$ some $r > 0$ such that $B(x, r) \subseteq G$. Then this defines the *pseudometric topology* on X. It has the set $\beta := \{B(x, r) \mid x \in X, r > 0\}$ of open balls as a basis. Let us have a look at the properties a base is supposed to have. Assume that $x \in B(x_1, r_1) \cap B(x_2, r_2)$, and select r with $0 < r < \min\{r_1 - d(x, x_1),$ Pseudometric topology

$r_2 - d(x, x_2)\}$. Then $B(x,r) \subseteq B(x_1, r_1) \cap B(x_2, r_2)$, because we have for $z \in B(x,r)$

$$\begin{aligned} d(z, x_1) &\leq d(z, x) + (x, x_1) < r + d(x, x_1) \leq (r_1 - d(x, x_1)) \\ &\quad +d(x, x_1) = r_1, \end{aligned} \tag{3.1}$$

by the triangle inequality; similarly, $d(x, x_2) < r_2$. Thus it follows from Proposition 3.1.1 that β is in fact a base.

Call two pseudometrics on X *equivalent* iff they generate the same topology. An equivalent formulation goes like this. Let τ_i be the topologies generated from pseudometrics d_i for $i = 1, 2$, then d_1 and d_2 are equivalent iff the identity $(X, \tau_1) \to (X, \tau_2)$ is a homeomorphism. These are two common methods to construct equivalent pseudometrics.

Lemma 3.5.3 *Let (X, d) be a pseudometric space. Then*

$$\begin{aligned} d_1(x, y) &:= \max\{d(x, y), 1\}, \\ d_2(x, y) &:= \frac{d(x, y)}{1 + d(x, y)} \end{aligned}$$

both define pseudometrics which are equivalent to d.

Proof It is clear that both d_1 and d_2 are pseudometrics (for d_2, compare Example 3.5.2). Let τ, τ_1, τ_2 be the respective topologies; then it is immediate that (X, τ) and (X, τ_1) are homeomorphic. Since $d_2(x, y) < r$ iff $d(x, y) < r/(1-r)$, provided $0 < r < 1$, we obtain also that (X, τ) and (X, τ_2) are homeomorphic. $\dashv$

These pseudometrics have the advantage that they are bounded, which is sometimes quite practical for establishing topological properties. Just as a point in case:

Proposition 3.5.4 *Let (X_n, d_n) be a pseudometric space with associated topology τ_n. Then the topological product $\prod_{n \in \mathbb{N}}(X_n, \tau_n)$ is a pseudometric space again.*

Proof 1. We may assume that each d_n is bounded by 1; otherwise, we select an equivalent pseudometric with this property (Lemma 3.5.3). Put

$$d\big((x_n)_{n \in \mathbb{N}}, (y_n)_{n \in \mathbb{N}}\big) := \sum_{n \in \mathbb{N}} 2^{-n} \cdot d_n(x_n, y_n).$$

We claim that the product topology is the topology induced by the pseudometric d (it is obvious that d is one).

2. Let $G_i \subseteq X_i$ open for $1 \leq i \leq k$, and assume that $x \in G := G_1 \times \ldots \times G_k \times \prod_{n>k} X_n$. We can find for $x_i \in G_i$ some positive r_i with $B_{d_i}(x_i, r_i) \subseteq G_i$. Put $r := \min\{r_1, \ldots, r_k\}$; then certainly, $B_d(x, r) \subseteq G$. This implies that each element of the base for the product topology is open with respect to d.

3. Given the sequence x and $r > 0$, take $y \in B_d(x, r)$. Put $t := r - d(x, y) > 0$. Select $m \in \mathbb{N}$ with $\sum_{n>m} 2^{-n} < t/2$, and let $G_n := B_{d_n}(y_n, t/2)$ for $n \leq m$. If $z \in U := G_1 \times \ldots \times G_n \times \prod_{k>m} X_k$, then

$$
\begin{aligned}
d(x, z) &\leq d(x, y) + d(y, z) \\
&\leq r - t + \sum_{n=1}^{m} 2^{-n} d_n(y_n, z_n) + \sum_{n>m} 2^{-n} \\
&< r - t + t/2 + t/2 \\
&= r,
\end{aligned}
$$

so that $U \subseteq B_d(x, r)$. Thus each open ball is open in the product topology. $\dashv$

One sees immediately that the pseudometric d constructed above is a metric, provided each d_n is one. Thus:

Corollary 3.5.5 *The countable product of metric spaces is a metric space in the product topology.* $\dashv$

One expects that each pseudometric space can be made a metric space by identifying those elements which cannot be separated by the pseudometric. Let us try:

Proposition 3.5.6 *Let (X, d) be a pseudometric space, and define $x \sim y$ iff $d(x, y) = 0$ for $x, y \in X$. Then the factor space $X/\sim$ is a metric space with metric $D([x]_\sim, [y]_\sim) := d(x, y)$.*

Proof 1. It is clear that $\sim$ is an equivalence relation, since $x \sim y$ and $y \sim z$ imply $d(x, z) \leq d(x, y) + d(y, z) = 0$; hence $x \sim z$ follows.

Because $d(x, x') = 0$ and $d(y, y') = 0$ imply $d(x, y) = d(x', y')$, D is well defined, and it has all the properties of a pseudometric. D is also a metric, since $D([x]_\sim, [y]_\sim) = 0$ is equivalent to $d(x, y) = 0$, hence to $x \sim y$, and thus to $[x]_\sim = [y]_\sim$.

2. The metric topology is the final topology with respect to the factor map $\eta_\sim$. To establish this, take a map $f : X/\sim \to Z$ with a topological space Y. Assume that $(f \circ \eta_\sim)^{-1}\big[G\big]$ is open for $G \subseteq Y$ open. If $[x]_\sim \in f^{-1}\big[G\big]$, we have $x \in (f \circ \eta_\sim)^{-1}\big[G\big]$; thus there exists $r > 0$ with $B_d(x, r) \subseteq \eta_\sim^{-1}\big[f^{-1}\big[G\big]\big]$. But this means that $B_D([x]_\sim, r) \subseteq f^{-1}\big[U\big]$, so that the latter set is open. Thus if $f \circ \eta_\sim$ is continuous, f is. The converse is established in the same way. This implies that the metric topology is final with respect to the factor map $\eta_\sim$, cp. Proposition 3.1.15. $\dashv$

We want to show that a pseudometric space satisfies the T_4-axiom (hence that a metric space is normal). So we take two disjoint closed sets and need to produce two disjoint open sets, each of which containing one of the closed sets. The following construction is helpful.

$d(x, A)$ **Lemma 3.5.7** *Let (X, d) be a pseudometric space. Define the distance of point $x \in X$ to $\emptyset \neq A \subseteq X$ through*

$$d(x, A) := \inf_{y \in A} d(x, y).$$

Then $d(\cdot, A)$ is continuous.

Proof Let $x, z \in X$, and $y \in A$, then $d(x, y) \leq d(x, z) + d(z, y)$. Now take lower bounds on y, then $d(x, A) \leq d(x, z) + d(z, A)$. This yields $d(x, A) - d(z, A) \leq d(x, z)$. Interchanging the rôles of x and z yields $d(z, A) - d(x, A) \leq d(z, x)$; thus $|d(x, A) - d(z, A)| \leq d(x, z)$. This implies continuity of $d(\cdot, A)$. $\dashv$

Given a closed set $A \subseteq X$, we find that $A = \{x \in X \mid d(x, A) = 0\}$; we can say a bit more:

Corollary 3.5.8 *Let X, A be as above, then $A^a = \{x \in X \mid d(x, A) = 0\}$.*

Proof Since $\{x \in X \mid d(x, A) = 0\}$ is closed, we infer that A^a is contained in this set. If, in the other hand, $x \notin A^a$, we find $r > 0$ such that $B(x, r) \cap A = \emptyset$; hence $d(x, A) \geq r$. Thus the other inclusion holds as well. $\dashv$

Armed with this observation, we can establish now

Proposition 3.5.9 *A pseudometric space (X, d) is a T_4-space.*

Proof Let F_1 and F_2 be disjoint closed subsets of X. Define

$$f(x) := \frac{d(x, F_1)}{d(x, F_1) + d(x, F_2)},$$

then Lemma 3.5.7 shows that f is continuous, and Corollary 3.5.8 indicates that the denominator will not vanish, since F_1 and F_2 are disjoint. It is immediate that F_1 is contained in the open set $\{x \mid f(x) < 1/2\}$, that $F_2 \subseteq \{x \mid f(x) > 1/2\}$, and that these open sets are disjoint. $\dashv$

Note that a pseudometric T_1-space is already a metric space (Exercise 3.15).

Define for $r > 0$ the r-neighborhood A^r of set $A \subseteq X$ as A^r

$$A^r := \{x \in X \mid d(x, A) < r\}.$$

This of course makes sense only if $d(x, A)$ is finite. Using the triangle inequality, one calculates $(A^r)^s \subseteq A^{r+s}$. This observation will be helpful when we look at the next example.

Example 3.5.10 Let (X, d) be a pseudometric space, and let

$$\mathfrak{C}(X) := \{C \subseteq X \mid C \text{ is compact and not empty}\}$$

be the set of all compact and not empty subsets of X. Define

$$\delta_H(C, D) := \max\,\{\max_{x \in C}\, d(x, D), \max_{x \in D}\, d(x, C)\}$$

for $C, D \in \mathfrak{C}(X)$. We claim that δ_H is a pseudometric on $\mathfrak{C}(X)$, which is a metric if d is a metric on X. δ_H

One notes first that

$$\delta_H(C, D) = \inf\,\{r > 0 \mid C \subseteq D^r, D \subseteq C^r\}.$$

This follows easily from $C \subseteq D^r$ iff $\max_{x \in C}\, d(x, D) < r$. Hence we obtain that $\delta_H(C, D) \le r$ and $\delta_H(D, E) \le s$ together imply $\delta_H(C, E) \le r + s$, which implies the triangle inequality. The other laws for a pseudometric are obvious. δ_H is called the *Hausdorff pseudometric*.

Now assume that d is a metric, and assume $\delta_H(C, D) = 0$. Thus $C \subseteq \bigcap_{n \in \mathbb{N}} D^{1/n}$ and $D \subseteq \bigcap_{n \in \mathbb{N}} C^{1/n}$. Because C and D are closed, and d is a metric, we obtain $C = D$ from Corollary 3.5.8; thus δ_H is a metric, which is accordingly called the *Hausdorff metric*. ✌

Let us take a magnifying glass and have a look at what happens locally in a point of a pseudometric space. Given $U \in \mathfrak{U}(x)$, we find an open ball $B(x, r)$ which is contained in U; hence we find even a rational number q with $B(x, q) \subseteq B(x, r)$. But this means that the open balls with rational radii form a basis for the neighborhood filter of x. This is sometimes also the case in more general topological spaces, so we define this property and two of its cousins for general topological spaces, rather than pseudometric ones.

Definition 3.5.11 *A topological space:*

1. *satisfies the* first axiom of countability *(and the space is called in this case* first countable*) iff the neighborhood filter of each point has a countable base of open sets,*
2. *satisfies the* second axiom of countability *(the space is called in this case* second countable*) iff the topology has a countable base,*
3. *is* separable *iff it has a countable dense subset.*

The standard example for a separable topological space is of course $\mathbb{R}$, where the rational numbers $\mathbb{Q}$ form a countable dense subset.

This is a trivial consequence of the observation just made.

Proposition 3.5.12 *A pseudometric space is first countable.* $\dashv$

In a pseudometric space, separability and satisfying the second axiom of countability coincide, as the following observation shows.

Proposition 3.5.13 *A pseudometric space (X, d) is second countable iff it has a countable dense subset.*

Proof 1. Let D be a countable dense subset, then

$$\beta := \{B(x, r) \mid x \in D, 0 < r \in \mathbb{Q}\}$$

is a countable base for the topology. For, given $U \subseteq X$ open, there exists $d \in D$ with $d \in U$, hence we can find a rational $r > 0$ with $B(d, r) \subseteq U$. On the other hand, one shows exactly as in the argumentation leading to Eq. (3.1) on page 334 that β is a base.

2. Assume that β is a countable base for the topology; pick from each $B \in \beta$ an element x_B. Then $\{x_B \mid B \in \beta\}$ is dense: given an open U, we find $B \in \beta$ with $B \subseteq U$; hence $x_B \in U$. This argument does

not require X being a pseudometric space (but the axiom of choice). $\dashv$

We know from Exercise 3.8 that a point x in a topological space is in the closure of a set A iff there exists a filter $\mathfrak{F}$ with $i_A(\mathfrak{F}) \to x$ with i_A as the injection $A \to X$. In a first countable space, in particular in a pseudometric space, we can work with sequences rather than filters, which is sometimes more convenient.

Proposition 3.5.14 *Let X be a first countable topological space, $A \subseteq X$. Then $x \in A^a$ iff there exists a sequence $(x_n)_{n\in\mathbb{N}}$ in A with $x_n \to x$.*

Proof If there exists a sequence $(x_n)_{n\in\mathbb{N}}$ which converges to x such that $x_n \in A$ for all $n \in \mathbb{N}$, then the corresponding filter converges to x, so it remains to establish the converse statement.

Now let $(U_n)_{n\in\mathbb{N}}$ be the basis of the neighborhood filter of $x \in A^a$ and $\mathfrak{F}$ be a filter with $i_A(\mathfrak{F}) \to x$. Put $V_n := U_1 \cap \ldots \cap U_n$, then $V_n \cap A \in i_A(\mathfrak{F})$. The sequence $(V_n)_{n\in\mathbb{N}}$ decreases and forms a basis for the neighborhood filter of x. Pick from each V_n an element $x_n \in A$, and take a neighborhood $U \in \mathfrak{U}(x)$. Since there exists n with $V_n \subseteq U$, we infer that $x_m \in U$ for all $m \leq n$, hence $x_n \to x$. $\dashv$

A second countable normal space X permits the following remarkable construction. Let β be a countable base for X, and define $\mathcal{A} := \{\langle U, V\rangle \mid U, V \in \beta, U^a \subseteq V\}$. Then $\mathcal{A}$ is countable as well, and we can find for each pair $\langle U, V\rangle \in \mathcal{A}$ a continuous map $f : X \to [0, 1]$ with $f(x) = 0$ for all $x \in U$ and $f(x) = 1$ for all $x \in X \setminus V$. This is a consequence of Urysohn's Lemma (Theorem 3.3.18). The collection $\mathcal{F}$ of all these functions is countable, because $\mathcal{A}$ is countable. Now define the embedding map

$$e : \begin{cases} X & \to [0,1]^{\mathcal{F}} \\ x & \mapsto (f(x))_{f\in\mathcal{F}} \end{cases}$$

We endow the space $[0, 1]^{\mathcal{F}}$ with the product topology, i.e., with the initial topology with respect to all projections $\pi_f : x \mapsto f(x)$. Then we observe these properties:

1. The map e is continuous. This is so because $\pi_f \circ e = f$ and f is continuous; hence we may infer continuity from Proposition 3.1.15.

2. The map e is injective. This follows from Urysohn's Lemma

(Theorem 3.3.18), since two distinct points constitute two disjoint closed sets.

3. If $G \subseteq X$ is open, $e\left[G\right]$ is open in $e\left[X\right]$. In fact, let $e(x) \in e\left[G\right]$. We find an open neighborhood H of $e(x)$ in $[0,1]^{\mathcal{F}}$ such that $e\left[X\right] \cap H \subseteq e\left[G\right]$ in the following way: from the construction, we infer that we can find a map $f \in \mathcal{F}$ such that $f(x) = 0$ and $f(y) = 1$ for all $y \in X \setminus G$, and hence $f(x) \notin f\left[X \setminus G\right]^a$; hence the set $H := \{y \in [0,1]^{\mathcal{F}} \mid y_f \notin f\left[X \setminus G\right]\}$ is open in $[0,1]^{\mathcal{F}}$, and $H \cap e\left[X\right]$ is contained in $e\left[G\right]$.

4. $[0,1]^{\mathcal{F}}$ is a metric space by Corollary 3.5.5, because the unit interval $[0,1]$ is a metric space and because $\mathcal{F}$ is countable.

Summarizing, X is homeomorphic to a subspace of $[0,1]^{\mathcal{F}}$. This is what *Urysohn's Metrization Theorem* says.

Proposition 3.5.15 *A second countable normal topological space is metrizable.* $\dashv$

The problem of metrization of topological spaces is nontrivial, as one can see from Proposition 3.5.15. The reader who wants to learn more about it may wish to consult Kelley's textbook [Kel55, p. 124 f] or Engelking's treatise [Eng89, 4.5, 5.4].

3.5.1 Completeness

Fix in this section a pseudometric space (X, d). A *Cauchy sequence* $(x_n)_{n \in \mathbb{N}}$ is defined in X just as in $\mathbb{R}$: Given $\epsilon > 0$, there exists an index $n \in \mathbb{N}$ such that $d(x_m, x_{m'}) < \epsilon$ holds for all $m, m' \geq n$.

Thus we have a Cauchy sequence, when we know that eventually the members of the sequence will be arbitrarily close; a converging sequence is evidently a Cauchy sequence. But a sequence which converges requires the knowledge of its limit; this is sometimes a problem in applications. It would be helpful if we could conclude from the fact that we have a Cauchy sequence that we may safely assume that there exists a point to which it converges. Spaces for which this is always guaranteed are called complete; they will be introduced next, and examples show that there are spaces which are not complete; note, however, that we can complete each pseudometric space. This will be considered in some detail later on.

Definition 3.5.16 *The pseudometric space is said to be* complete *iff each Cauchy sequence has a limit.*

It is well known that the rational numbers are not complete, which is usually shown by showing that $\sqrt{2}$ is not rational. Another instructive example proposed by Bourbaki [Bou89, II.3.3] is the following:

Example 3.5.17 The rational numbers $\mathbb{Q}$ are not complete in the usual metric. Take

$$x_n := \sum_{i=0}^{n} 2^{-i\cdot(i+1)/2}.$$

Then $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{Q}$: if $m > n$, then $|x_m - x_n| \leq 2^{-(n+3)/2}$ (this is shown easily through the well-known identity $\sum_{i=0}^{p} i = p\cdot(p+1)/2$). Now assume that the sequence converges to $a/b \in \mathbb{Q}$; then we can find an integer h_n such that

$$\left|\frac{a}{b} - \frac{h_n}{2^{n\cdot(n+1)/2}}\right| \leq \frac{1}{2^{n\cdot(n+3)/2}},$$

yielding

$$|a\cdot 2^{n\cdot(n+1)/2} - b\cdot h_n| \leq \frac{b}{2^n}$$

for all $n \in \mathbb{N}$. The left-hand side of this inequality is a whole number, and the right side is not, once $n > n_0$ with n_0 so large that $b < 2^n$. This means that the left-hand side must be zero, so that $a/b = x_n$ for $n > n_0$. This is a contradiction. ✌

We know that $\mathbb{R}$ is complete with the usual metric, and the rationals are not. But there is a catch: if we change the metric, completeness may be lost.

Example 3.5.18 The half open interval $]0, 1]$ is not complete under the usual metric $d(x, y) := |x - y|$. But take the metric

$$d'(x, y) := \left|\frac{1}{x} - \frac{1}{y}\right|$$

Because $a < x < b$ iff $1/b < 1/x < 1/a$ holds for $0 < a \leq b \leq 1$, the metrics d and d' are equivalent on $]0, 1]$. Let $(x_n)_{n\in\mathbb{N}}$ be a d'-Cauchy sequence, then $(1/x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $(\mathbb{R}, |\cdot|)$; hence it converges, so that $(x_n)_{n\in\mathbb{N}}$ is d'-convergent in $]0, 1]$.

The trick here is to make sure that a Cauchy sequence avoids the region around the critical value 0. ✌

Thus we have to carefully stick to the given metric; changing the metric entails checking completeness properties for the new metric.

Example 3.5.19 Endow the set $\mathcal{C}([0,1])$ of continuous functions on the unit interval with the metric $d(f,g) := \sup_{0\leq x\leq 1} |f(x) - g(x)|$; see Example 3.5.2. We claim that this metric space is complete. In fact, let $(f_n)_{n\in\mathbb{N}}$ be a d-Cauchy sequence in $\mathcal{C}([0,1])$. Because we have for each $x \in [0,1]$ the inequality $|f_n(x) - f_m(x)| \leq d(f_n, f_m)$, we conclude that $\bigl(f_n(x)\bigr)_{n\in\mathbb{N}}$ is a Cauchy sequence for each $x \in [0,1]$, which converges to some $f(x)$, since $\mathbb{R}$ is complete. We have to show that f is continuous and that $d(f, f_n) \to 0$.

Let $\epsilon > 0$ be given; then there exists $n \in \mathbb{N}$ such that $d(f_m, f_{m'}) < \epsilon/2$ for $m, m' \geq n$; hence we have

$$\begin{aligned} |f_m(x) - f_{m'}(x')| &\leq |f_m(x) - f_m(x')| + |f_m(x') - f_{m'}(x')| \\ &\leq |f_m(x) - f_m(x')| + d(f_m, f_{m'}). \end{aligned}$$

Choose $\delta > 0$ so that $|x - x'| < \delta$ implies $|f_m(x) - f_m(x')| < \epsilon/2$, then $|f_m(x) - f_{m'}(x')| < \epsilon$ for $m, m' \geq n$. But this means $|x - x'| < \delta$ implies $|f(x) - f(x')| \leq \epsilon$. Hence f is continuous. Since $\bigl(\{x \in [0,1] \mid |f_n(x) - f(x)| \leq \epsilon\}\bigr)_{n\in\mathbb{N}}$ constitutes an open cover of $[0,1]$, we find a finite cover given by $n_1, \ldots, n_k$; let n' be the smallest of these numbers, then $d(f, f_n) \leq \epsilon$ for all $n \geq n'$, hence $d(f, f_n) \to 0$. ✌

The next example is suggested by an observation in [MPS86].

Example 3.5.20 Let $r : X \to \mathbb{N}$ be a ranking function, and denote the (ultra-)metric on $\mathcal{P}(X)$ constructed from it by d; see Example 3.5.2. Then $(\mathcal{P}(X), d)$ is complete. In fact, let $(A_n)_{n\in\mathbb{N}}$ be a Cauchy sequence; thus we find for each $m \in \mathbb{N}$ an index $n \in \mathbb{N}$ such that $c(A_k, A_\ell) \geq m$, whenever $k, \ell \geq n$. We claim that the sequence converges to

$$A := \bigcup_{n\in\mathbb{N}} \bigcap_{k\geq n} A_k,$$

which is the set of all elements in X which are contained in all but a finite number of sequence elements. Given m, fix n as above; we show that $c(A, A_k) > m$, whenever $k > n$. Take an element $x \in A \Delta A_k$ of minimal rank:

- If $x \in A$, then there exists ℓ such that $x \in A_t$ for all $t \geq \ell$, so take $t \geq \max\{\ell, n\}$; then $x \in A_t \Delta A_k$, and hence $c(A, A_k) = r(x) \geq c(A_t, A_k) > m$.

- If, however, $x \notin A$, we conclude that $x \notin A_t$ for infinitely many t, so $x \notin A_t$ for some $t > n$. But since $x \in A \Delta A_k$, we conclude $x \in A_k$; hence $x \in A_k \Delta A_t$, and thus $c(A, A_k) = r(x) \geq c(A_k, A_t) > m$.

Hence $A_n \to A$ in $(\mathcal{P}(X), d)$. ✌

The following observation is trivial but sometimes helpful.

Lemma 3.5.21 *A closed subset of a complete pseudometric space is complete.* $\dashv$

If we encounter a pseudometric space which is not complete, we may complete it through the following construction. Before discussing it, we need a simple auxiliary statement, which says that we can check completeness already on a dense subset.

Lemma 3.5.22 *Let $D \subseteq X$ be dense. Then the space is complete iff each Cauchy sequence on D converges.*

Proof If each Cauchy sequence from X converges, so does each such sequence from D, so we have to establish the converse. Let $(x_n)_{n \in \mathbb{N}}$ be a Cauchy sequence on X. Given $n \in \mathbb{N}$, there exists for x_n an element $y_n \in D$ such that $d(x_n, y_n) < 1/n$. Because $(x_n)_{n \in \mathbb{N}}$ is a Cauchy sequence, $(y_n)_{n \in \mathbb{N}}$ is one as well, which converges by assumption to some $x \in X$. The triangle inequality shows that $(x_n)_{n \in \mathbb{N}}$ converges to x as well. $\dashv$

This helps in establishing that each pseudometric space can be embedded into a complete pseudometric space. The approach may be described as Charlie Brown's device—"If you can't beat them, join them." So we take all Cauchy sequences as our space into which we embed X, and—intuitively—we flesh out from a Cauchy sequence of these sequences the diagonal sequence, which then will be a Cauchy sequence as well and which will be a limit of the given one. This sounds more complicated than it is, however, because fortunately Lemma 3.5.22 makes life easier, when it comes to establishing completeness. Here we go.

Proposition 3.5.23 *There exists a complete pseudometric space (X^*, d^*) into which (X, d) may be embedded isometrically as a dense subset.*

Fairly direct approach

Proof 0. This is the line of attack: We define X^* and d^*, show that we can embed X isometrically into it as a dense subset, and then we establish completeness with the help of Lemma 3.5.22.

1. Define

$$X^* := \{(x_n)_{n\in\mathbb{N}} \mid (x_n)_{n\in\mathbb{N}} \text{ is a } d\text{-Cauchy sequence in } X\},$$

and put

$$d^*\big((x_n)_{n\in\mathbb{N}}, (y_n)_{n\in\mathbb{N}}\big) := \lim_{n\to\infty} d(x_n, y_n)$$

Before proceeding, we should make sure that the limit in question exists. In fact, given $\epsilon > 0$, there exists $n \in \mathbb{N}$ such that $d(x_{m'}, x_m) < \epsilon/2$ and $d(y_{m'}, y_m) < \epsilon/2$ for $m, m' \geq n$; thus, if $m, m' \geq n$, we obtain

$$\begin{array}{rcl} d(x_m, y_m) & \leq & d(x_m, x_{m'}) + d(x_{m'}, y_{m'}) + d(y_{m'}, y_m) \\ & < & d(x_{m'}, y_{m'}) + \epsilon; \end{array}$$

interchanging the rôles of m and m' yields

$$|d(x_m, y_m) - d(x_{m'}, y_{m'})| < \epsilon$$

for $m, m' \geq n$. Hence $(d(x_n, y_n))_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$, which converges by completeness of $\mathbb{R}$.

2. Given $x \in X$, the sequence $(x)_{n\in\mathbb{N}}$ is a Cauchy sequence, so it offers itself as the image of x; let $e : X \to X^*$ be the corresponding map, which is injective and which preserves the pseudometric. Hence e is continuous. We show that $e\big[X\big]$ is dense in X^*: take a Cauchy sequence $(x_n)_{n\in\mathbb{N}}$ and $\epsilon > 0$. Let $n \in \mathbb{N}$ be selected for ϵ, and assume $m \geq n$. Then

$$D((x_n)_{n\in\mathbb{N}}, e(x_m)) = \lim_{n\to\infty} d(x_n, x_m) < \epsilon.$$

3. The crucial point is completeness. An appeal to Lemma 3.5.22 shows that it is sufficient to show that a Cauchy sequence in $e\big[X\big]$ converges in (X^*, d^*), because $e\big[X\big]$ is dense. But this is trivial. $\dashv$

Having the completion X^* of a pseudometric space X at our disposal, we might be tempted to extend a continuous map $X \to Y$ to a continuous map $X^* \to Y$, e.g., in the case that Y is complete. This is usually not possible; for example, not every continuous function $\mathbb{Q} \to \mathbb{R}$ has a continuous extension. We will deal with this problem when discussing uniform continuity below, but we will state and prove here a condition which is sometimes helpful when one wants to extend a function not to the whole completion but to a domain which is somewhat larger than the given one. Define as in Sect. 1.7.2 the *diameter* $\text{diam}(A)$ of a set A as diam(A)

$$\text{diam}(A) := \sup \{d(x, y) \mid x, y \in A\}$$

(note that the diameter may be infinite). It is easy to see that $\text{diam}(A) = \text{diam}(A^a)$ using Proposition 3.5.14. Now assume that $f : A \to Y$ is given, then we measure the discontinuity of f at point x through the *oscillation* $\varnothing_f(x)$ of f at $x \in A^a$. This is defined as the smallest diameter of the image of an open neighborhood of x, formally, Oscillation $\varnothing_f(x)$

$$\varnothing_f(x) := \inf\{\text{diam}(f\big[A \cap V\big]) \mid x \in V, V \text{ open}\}.$$

If f is continuous on A, we have $\varnothing_f(x) = 0$ for each element x of A. In fact, let $\epsilon > 0$ be given; then there exists $\delta > 0$ such that $\text{diam}(f\big[A \cap V\big]) < \epsilon$, whenever V is a neighborhood of x of diameter less than δ. Thus $\varnothing_f(x) < \epsilon$; since $\epsilon > 0$ was chosen to be arbitrary, the claim follows.

Lemma 3.5.24 *Let Y be a complete metric space and X a pseudometric space, then a continuous map $f : A \to Y$ can be extended to a continuous map $f_* : G \to Y$, where $G := \{x \in A^a \mid \varnothing_f(x) = 0\}$ has these properties:* Extension

1. $A \subseteq G \subseteq A^a$,
2. *G can be written as the intersection of countably many open sets.*

The basic idea for the proof is rather straightforward. Take an element in the closure of A; then there exists a sequence in A converging to this point. If the oscillation at that point is zero, the images of the sequence elements must form a Cauchy sequence, so we extend the map by forming the limit of this sequence. Now we have to show that this map is well defined and continuous. Idea for the proof

Proof 1. We may and do assume that the complete metric d for Y is bounded by 1. Define G as above; then $A \subseteq G \subseteq A^a$, and G can be written as the intersection of a sequence of open sets. In fact, represent G as

$$G = \bigcap_{n \in \mathbb{N}} \{x \in A^a \mid \varnothing_f(x) < \frac{1}{n}\},$$

so we have to show that $\{x \in A^a \mid \o_f(x) < q\}$ is open in A^a for any $q > 0$. But we have

$$\{x \in A^a \mid \o_f(x) < q\} = \bigcup\{V \cap A^a \mid \operatorname{diam}(f\big[V \cap A\big]) < q\}.$$

This is the union of sets open in A^a; hence, it is an open set itself.

2. Now take an element $x \in G \subseteq A^a$. Then there exists a sequence $(x_n)_{n\in\mathbb{N}}$ of elements $x_n \in A$ with $x_n \to x$. Given $\epsilon > 0$, we find a neighborhood V of x with $\operatorname{diam}(f\big[A \cap V\big]) < \epsilon$, since the oscillation of f at x is 0. Because $x_n \to x$, we know that we can find an index $n_\epsilon \in \mathbb{N}$ such that $x_m \in V \cap A$ for all $m > n_\epsilon$. This implies that the sequence $(f(x_n))_{n\in\mathbb{N}}$ is a Cauchy sequence in Y. It converges because Y is complete. Put

$$f_*(x) := \lim_{n\to\infty} f(x_n).$$

3. We have to show now that:

- f_* is well defined.
- f_* extends f.
- f_* is continuous.

Assume that we can find $x \in G$ such that $(x_n)_{n\in\mathbb{N}}$ and $(x'_n)_{n\in\mathbb{N}}$ are sequences in A with $x_n \to x$ and $x'_n \to x$, but $\lim_{n\to\infty} f(x_n) \neq \lim_{n\to\infty} f(x'_n)$. Thus we find some $\eta > 0$ such that $d(f(x_n), f(x'_n)) \geq \eta$ infinitely often. Then the oscillation of f at x is at least $\eta > 0$, a contradiction. This implies that f_* is well defined, and it implies also that f_* extends f. Now let $x \in G$. If $\epsilon > 0$ is given, we find a neighborhood V of x with $\operatorname{diam}(f\big[A \cap V\big]) < \epsilon$. Thus, if $x' \in G \cap V$, then $d(f_*(x), f_*(x')) < \epsilon$. Hence f_* is continuous. ⊣

G_δ-set

We will encounter later on sets which can be written as the countable intersection of open sets. They are called *G_δ-sets*. Rephrasing Lemma 3.5.24, f can be extended from A to a G_δ-set containing A and contained in A^a.

A characterization of complete spaces in terms of sequences of closed sets with decreasing diameters is given below.

Proposition 3.5.25 *These statements are equivalent:*

1. *X is complete.*
2. *For each decreasing sequence $(A_n)_{n\in\mathbb{N}}$ of nonempty closed sets the diameter of which tends to zero, there exists $x \in X$ such that $\bigcap_{n\in\mathbb{N}} A_n = \{x\}^a$.*

In particular, if X is a metric space, then X is complete iff each decreasing sequence of nonempty closed sets the diameter of which tends to zero has exactly one point in common.

Proof The assertion for the metric case follows immediately from the general case, because $\{x\}^a = \{x\}$, and because there can be not more than one element in the intersection.

1 $\Rightarrow$ 2: Let $(A_n)_{n\in\mathbb{N}}$ be a decreasing sequence of nonempty closed sets with $\mathrm{diam}(A_n) \to 0$; then we have to show that $\bigcap_{n\in\mathbb{N}} A_n = \{x\}^a$ for some $x \in X$. Pick from each A_n an element x_n, then $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence which converges to some x, since X is complete. Because the intersection of closed sets is closed again, we conclude $\bigcap_{n\in\mathbb{N}} A_n = \{x\}^a$.

2 $\Rightarrow$ 1: Take a Cauchy sequence $(x_n)_{n\in\mathbb{N}}$; then $A_n := \{x_m \mid m \geq n\}^a$ is a decreasing sequence of closed sets, the diameter of which tends to zero. In fact, given $\epsilon > 0$, there exists $n \in \mathbb{N}$ such that $d(x_m, x_{m'}) < \epsilon$ for all $m, m' \geq n$, hence $\mathrm{diam}(A_n) < \epsilon$, and it follows that this holds also for all $k \geq n$. Then it is obvious that $x_n \to x$ whenever $x \in \bigcap_{n\in\mathbb{N}} A_n$. $\dashv$

We mention all too briefly a property of complete spaces which renders them most attractive, viz., Banach's Fixed-Point Theorem.

Definition 3.5.26 *Call $f : X \to X$ a* contraction *iff there exists γ with $0 < \gamma < 1$ such that $d(f(x), f(y)) \leq \gamma \cdot d(x, y)$ holds for all $x, y \in X$.*

This is the celebrated Banach Fixpoint Theorem.

Theorem 3.5.27 *Let $f : X \to X$ be a contraction with X complete. Then there exists $x \in X$ with $f(x) = x$. If $f(y) = y$ holds as well, then $d(x, y) = 0$. In particular, if X is a metric space, then there exists a unique fixed point for f.* Banach's Fixpoint Theorem

The idea is just to start with an arbitrary element of X and to iterate f on it. This yields a sequence of elements of X. Because the elements become closer and closer, completeness kicks in and makes sure that there exists a limit. This limit is independent of the starting point.

Proof Define the n-th iteration f^n of f through $f^1 := f$ and $f^{n+1} := f^n \circ f$. Now let x_0 be an arbitrary element of X, and define $x_n := f^n(x_0)$. Then $d(x_n, x_{n+m}) \leq \gamma^n \cdot d(x_0, x_m)$, so that $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence which converges to some $x \in X$, and $f(x) = x$. If $f(y) = y$, we have $d(x, y) = d(f(x), f(y)) \leq \gamma \cdot d(x, y)$; thus $d(x, y) = 0$. This implies uniqueness of the fixed point as well. $\dashv$

Google

Banach's Fixed-Point Theorem has a wide range of applications, and it is used for iteratively approximating the solution of equations, e.g., for implicit functions. The following example permits a glance at Google's page rank algorithm; it follows [Rou15] (the linear algebra behind it is explored in, e.g., [LM05, Kee93]).

Example 3.5.28 Let $S := \{\langle x_1, \ldots, x_n\rangle \mid x_i \geq 0, x_1 + \ldots + x_n = 1\}$ be the set of all discrete probability distributions over n objects and $P : \mathbb{R}^n \to \mathbb{R}^n$ be a stochastic matrix; this means that P has nonnegative entries and the rows all add up to 1. The set $\{1, \ldots, n\}$ is usually interpreted as the state space for some random experiment; entry $p_{i,j}$ is then interpreted as the probability for the change of state i to state j. We have in particular $P : S \to S$, so a probability distribution is transformed into another probability distribution. We assume that P has an eigenvector $v_1 \in S$ for the eigenvalue 1 and that the other eigenvalues are in absolute value not greater than 1 (this is what the classic Perron–Frobenius Theorem says; see [LM05, Kee93]); moreover, we assume that we can find a base $\{v_1, \ldots, v_n\}$ of eigenvectors, all of which may be assumed to be in S; let λ_i be the eigenvector for v_i, then $\lambda_1 = 1$ and $|\lambda_i| \leq 1$ for $i \geq 2$. Such a matrix is called a *regular transition matrix*; these matrices are investigated in the context of stability of finite Markov transition chains.

Define for the distributions $p = \sum_{i=1}^{n} p_i \cdot v_i$ and $q = \sum_{i=1}^{n} q_i \cdot v_i$ their distance through

$$d(p, q) := \frac{1}{2} \cdot \sum_{i=1}^{n} |p_i - q_i|.$$

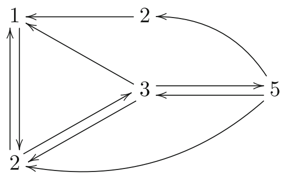
Because $\{v_1, \ldots, v_n\}$ are linearly independent, d is a metric. Because this set forms a basis, hence is given through a bijective linear maps from the base given by the unit vectors, and because the Euclidean metric is complete, d is complete as well.

Now define $f(x) := P \cdot x$; then this is a contraction $S \to S$:

$$\begin{aligned} d(P \cdot x, P \cdot y) &= \frac{1}{2} \cdot \sum_{i=1}^{n} |x_i \cdot P(v_i) - y_i \cdot P(v_i)| \\ &\leq \frac{1}{2} \sum_{i=1}^{n} |\lambda_i \cdot (x_i - y_i)| \leq \frac{1}{2} \cdot d(x, y). \end{aligned}$$

Thus f has a fixed point, which must be v_1 by uniqueness.

Now assume that we have a (very little) Web universe with only five pages. The links are given as in the diagram:



The transitions between pages are at random; the matrix below describes such a random walk:

$$P := \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}$$

It says that we make a transition from state 2 to state 1 with $p_{2,1} = \frac{1}{2}$, also $p_{2,3} = \frac{1}{2}$, the transition from state 2 to state 3. From state 1, one goes with probability one to state 2, because $p_{1,2} = 1$. Iterating P quite a few times will yield a solution which does not change much after 32 steps; one obtains

$$P^{32} = \begin{pmatrix} 0.293 & 0.390 & 0.220 & 0.024 & 0.073 \\ 0.293 & 0.390 & 0.220 & 0.024 & 0.073 \\ 0.293 & 0.390 & 0.220 & 0.024 & 0.073 \\ 0.293 & 0.390 & 0.220 & 0.024 & 0.073 \\ 0.293 & 0.390 & 0.220 & 0.024 & 0.073 \end{pmatrix}$$

The eigenvector p for the eigenvalue 1 looks like this: $p = \langle 0.293, 0.390, 0.220, 0.024, 0.073\rangle$, so this yields a stationary distribution.

Web search

In terms of Web searches, the importance of the pages is ordered according to this stationary distribution as $2, 1, 3, 5, 4$; so this is the ranking one would associate with these pages.

This is the basic idea behind Google's page ranking algorithm. Of course, there are many practical considerations which have been eliminated from this toy example. It may be that the matrix does not follow the assumptions above, so that it has to me modified accordingly in a preprocessing step. Size is a problem, of course, since handling the extremely large matrices occurring in Web searches may become quite intricate. ✌

Compact pseudometric spaces are complete. This will be a byproduct of a more general characterization of compact spaces. We show first that compactness and sequential compactness are the same for these spaces. This is sometimes helpful in those situations in which a sequence is easier to handle than an open cover, or an ultrafilter.

ϵ-net

Before discussing this, we introduce *ϵ-nets* as a cover of X through a *finite* family $\{B(x, \epsilon) \mid x \in A\}$ of open balls of radius ϵ. X may or may not have an ϵ-net for any given $\epsilon > 0$. For example, $\mathbb{R}$ does not have an ϵ-net for any $\epsilon > 0$, in contrast to $[0, 1]$ or $]0, 1]$.

Definition 3.5.29 *The pseudometric space X is* totally bounded *iff there exists for each $\epsilon > 0$ an ϵ-net for X. A subset of a pseudometric space is totally bounded iff it is a totally bounded subspace.*

Thus $A \subseteq X$ is totally bounded iff $A^a \subseteq X$ is totally bounded.

We see immediately

Lemma 3.5.30 *A compact pseudometric space is totally bounded.* ⊣

Now we are in a position to establish this equivalence, which will help characterize compact pseudometric spaces.

Proposition 3.5.31 *The following properties are equivalent for the pseudometric space X:*

1. *X is compact.*
2. *X is sequentially compact, i.e., each sequence has a convergent subsequence.*

Proof 1 $\Rightarrow$ 2: Assume that the sequence $(x_n)_{n\in\mathbb{N}}$ does not have a convergent subsequence, and consider the set $F := \{x_n \mid n \in \mathbb{N}\}$. This set is closed, since, if $y_n \to y$ and $y_n \in F$ for all $n \in \mathbb{N}$, then $y \in F$, because the sequence $(y_n)_{n\in\mathbb{N}}$ is eventually constant. F is also discrete, since, if we could find for some $z \in F$ for each $n \in \mathbb{N}$ an element in $F \cap B(z, 1/n)$ different from z, we would have a convergent subsequence. Hence F is a closed discrete subspace of X which contains infinitely many elements, which is impossible. This contradiction shows that each sequence has a convergent subsequence.

2 $\Rightarrow$ 1: Before we enter into the second and harder part of the proof, we have a look at its plan. Given an open cover for the sequential compact space X, we have to construct a finite cover from it. If we succeed in constructing for each $\epsilon > 0$ a finite net so that we can fit each ball into some element of the cover, we are done, because in this case we may take just these elements of the cover, obtaining a finite cover. That this construction is possible is shown in the first part of the proof. We construct under the assumption that it is not possible a sequence, which has a converging subsequence, and the limit of this subsequence will be used as kind of a flyswatter.

Plan of attack

The second part of the proof is then just a simple application of the net so constructed.

Fix $(U_i)_{i\in I}$ as a cover of X. We claim that we can find for this cover some $\epsilon > 0$ such that, whenever $\text{diam}(A) < \epsilon$, there exists $i \in I$ with $A \subseteq U_i$. Assume that this is wrong; then we find for each $n \in \mathbb{N}$ some $A_n \subseteq X$ which is not contained in one single U_i. Pick from each A_n an element x_n; then $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence, say $(y_n)_{n\in\mathbb{N}}$, with $y_n \to y$. There exists a member U of the cover with $y \in U$, and there exists $r > 0$ with $B(y, r) \subseteq U$. Now we catch the fly. Choose $\ell \in \mathbb{N}$ with $1/\ell < r/2$, then $y_m \in B(y, r/2)$ for $m \geq n_0$ for some suitable chosen $n_0 \in \mathbb{N}$; hence, because $(y_n)_{n\in\mathbb{N}}$ is a subsequence of $(x_n)_{n\in\mathbb{N}}$, there are infinitely many x_k contained in $B(y, r/2)$. But since $\text{diam}(A_\ell) < 1/\ell$, this implies $A_\ell \subseteq B(y, r) \subseteq U$, which is a contradiction.

Now select $\epsilon > 0$ as above for the cover $(U_i)_{i\in I}$, and let the finite set A be the set of centers for an $\epsilon/2$-net, say, $A = \{a_1, \ldots, a_k\}$. Then we can find for each $a_j \in A$ some member U_{i_j} of this cover with $B(a_j, \epsilon/2) \subseteq U_{i_j}$ (note that $\text{diam}(B(x, r) < 2 \cdot r)$. This yields a finite cover $\{U_{i_j} \mid 1 \leq j \leq k\}$ of X. $\dashv$

This proof was conceptually a little complicated, since we had to make the step from a sequence (with a converging subsequence) to a cover (with the goal of finding a finite cover). Both are not immediately related. The missing link turned out to be measuring the size of a set through its diameter and capturing limits through suitable sets.

Using the last equivalence, we are in a position to characterize compact pseudometric spaces.

Theorem 3.5.32 *A pseudometric space is compact iff it is totally bounded and complete.*

Proof 1. Let X be compact. We know already from Lemma 3.5.30 that a compact pseudometric space is totally bounded. Let $(x_n)_{n\in\mathbb{N}}$ be a Cauchy sequence; then we know that it has a converging subsequence, which, being a Cauchy sequence, implies that it converges itself.

2. Assume that X is totally bounded and complete. In view of Proposition 3.5.31 it is enough to show that X is sequentially compact. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X. Since X is totally bounded, we find a subsequence (x_{n_1}) which is entirely contained in an open ball of radius less that 1. Then we may extract from this sequence a subsequence (x_{n_2}) which is contained in an open ball of radius less than $1/2$. Continuing inductively, we find a subsequence $(x_{n_{k+1}})$ of (x_{n_k}) the members of which are completely contained in an open ball of radius less than $2^{-(k+1)}$. Now define $y_n := x_{n_n}$; hence $(y_n)_{n\in\mathbb{N}}$ is the diagonal sequence in this scheme.

We claim that $(y_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. In fact, let $\epsilon > 0$ be given; then there exists $n \in \mathbb{N}$ such that $\sum_{\ell>n} 2^{-\ell} < \epsilon/2$. Then we have for $m > n$

$$d(y_n, y_m) \leq 2 \cdot \sum_{\ell=n}^{m} 2^{-\ell} < \epsilon.$$

By completeness, $y_n \to y$ for some $y \in X$. Hence we have found a converging subsequence of the given sequence $(x_n)_{n\in\mathbb{N}}$, so that X is sequentially compact. ⊣

Shift of emphasis

It might be noteworthy to observe the shift of emphasis between finding a finite cover for a given cover and admitting an ϵ-net for each $\epsilon > 0$. While we have to select a finite cover from an arbitrarily given cover beyond our control, we can construct in the case of a totally bounded space for each $\epsilon > 0$ a cover of a certain size; hence we may be in a position to influence the shape of this special cover. Consequently, the characterization of compact spaces in Theorem 3.5.32 is very helpful and handy, but, alas, it works only in the restricted calls of pseudometric spaces.

We apply this characterization to $(\mathfrak{C}(X), \delta_H)$, the space of all nonempty compact subsets of (X, d) with the Hausdorff metric δ_H; see Example 3.5.10.

Proposition 3.5.33 *$(\mathfrak{C}(X), \delta_H)$ is complete, if X is a complete pseudometric space.*

Proof We fix for the proof a Cauchy sequence $(C_n)_{n\in\mathbb{N}}$ of elements of $\mathfrak{C}(X)$.

0. Let us pause a moment and discuss the approach to the proof first. We show in the first step that $(\bigcup_{n\in\mathbb{N}} C_n)^a$ is compact by showing that it is totally bounded and complete. Completeness is trivial, since the space is complete, and we are dealing with a closed subset, so we focus on showing that the set is totally bounded. Actually, it is sufficient to show that $\bigcup_{n\in\mathbb{N}} C_n$ is totally bounded, because a set is totally bounded iff its closure is. Plan

Then compactness of $(\bigcup_{n\in\mathbb{N}} C_n)^a$ implies that $C := \bigcap_{n\in\mathbb{N}} (\bigcup_{k\geq n} C_k)^a$ is compact as well; moreover, we will argue that C must be nonempty. Then it is shown that $C_n \to C$ in the Hausdorff metric.

1. Let $D := \bigcup_{n\in\mathbb{N}} C_n$, and let $\epsilon > 0$ be given. We will construct an ϵ-net for D. Because $(C_n)_{n\in\mathbb{N}}$ is Cauchy, we find for ϵ an index ℓ so that $\delta_H(C_n, C_m) < \epsilon/2$ for $n, m \geq \ell$. When $n \geq \ell$ is fixed, this means in particular that $C_m \subseteq C_n^{\epsilon/2}$ for all $m \geq \ell$, thus $d(x, C_n) < \epsilon/2$ for all $x \in C_m$ and all $m \geq \ell$. We will use this observation in a moment.

Let $\{x_1, \ldots, x_t\}$ be an $\epsilon/2$-net for $\bigcup_{j=1}^n C_j$; we claim that this is an ϵ-net for D. In fact, let $x \in D$. If $x \in \bigcup_{j=1}^{\ell} C_j$, then there exists some k with $d(x, x_k) < \epsilon/2$. If $x \in C_m$ for some $m > n$, $x \in C_n^{\epsilon/2}$, so that we find $x' \in C_n$ with $d(x, x') < \epsilon/2$, and for x', we find k such that $d(x_k, x') < \epsilon/2$. Hence we have $d(x, x'_k) < \epsilon$, so that we have shown that $\{x_1, \ldots, x_t\}$ is an ϵ-net for D. Thus D^a is totally bounded, hence compact.

2. From the first part it follows that $(\bigcup_{k\geq n} C_k)^a$ is compact for each $n \in \mathbb{N}$. Since these sets form a decreasing sequence of nonempty closed

sets to the compact set given by $n = 1$, their intersection cannot be empty; hence $C := \bigcap_{n\in\mathbb{N}} (\bigcup_{k\geq n} C_k)^a$ is compact and nonempty and hence a member of $\mathfrak{C}(X)$.

We claim that $\delta_H(C_n, C) \to 0$, as $n \to \infty$. Let $\epsilon > 0$ be given; then we find $\ell \in \mathbb{N}$ such that $\delta_H(C_m, C_n) \leq \epsilon/2$, whenever $n, m \geq \ell$. We show that $\delta_H(C_n, C) < \epsilon$ for all $n \geq \ell$. Let $n \geq \ell$. The proof is subdivided into showing that $C \subseteq C_n^\epsilon$ and $C_n \subseteq C^\epsilon$.

- Let us work on the first inclusion. Because $D := (\bigcup_{i\geq n} C_i)^a$ is totally bounded, there exists a $\epsilon/2$-net, say, $\{x_1, \ldots, x_t\}$, for D. If $x \in C \subseteq D$, then there exists j such that $d(x, x_j) < \epsilon/2$, so that we can find $y \in C_n$ with $d(y, x_j) < \epsilon/2$. Consequently, we find for $x \in C$ some $y \in C_n$ with $d(x, y) < \epsilon$. Hence $C \subseteq C_n^\epsilon$.

- Now for the second inclusion. Take $x \in C_n$. Since $\delta_H(C_m, C_n) < \epsilon/2$ for $m \geq \ell$, we have $C_n \subseteq C_m^{\epsilon/2}$; hence find $x_m \in C_m$ with $d(x, x_m) < \epsilon/2$. The sequence $(x_k)_{k\geq m}$ consists of members of the compact set D, so it has a converging subsequence which converges to some $y \in D$. But it actually follows from the construction that $y \in C$, and $d(x, y) \leq d(x, x_m) + d(x_m, y) < \epsilon$ for m taken sufficiently large from the subsequence. This yields $x \in C^\epsilon$.

Taking these inclusions together, they imply $\delta_H(C_n, C) < \epsilon$ for $n > \ell$. This shows that $(\mathfrak{C}(X), \delta_H)$ is a complete pseudometric space, if (X, d) is one. $\dashv$

The topology induced by the Hausdorff metric can be defined in a way which permits a generalization to arbitrary topological spaces, where it is called the *Vietoris topology*. It has been studied with respect to finding continuous selections, e.g., by Michael [Mic51]; see also [JR02, CV77]. The reader is also referred to [Kur66, §33] and to [Eng89, p. 120] for a study of topologies on subsets.

We will introduce uniform continuity now and discuss this concept briefly here. Uniform spaces will turn out to be the proper scenario for the more extended discussion in Sect. 3.6.4. As a motivating example, assume that the pseudometric d on X is bounded, take a subset $A \subseteq X$, and look at the function $x \mapsto d(x, A)$. Since

$$|d(x, A) - d(y, A)| \leq d(x, y),$$

we know that this map is continuous. This means that, given $x \in X$, there exists $\delta > 0$ such that $d(x, x') < \delta$ implies $|d(x, A) - d(x', A)| < \epsilon$. We see from the inequality above that the choice of δ does only *depend on ϵ, but not on x*. Compare this with the function $x \mapsto 1/x$ on $]0, 1]$. This function is continuous as well, but the choice of δ depends on the point x you are considering: whenever $0 < \delta < \epsilon \cdot x^2/(1 + \epsilon \cdot x)$, we may conclude that $|x' - x| \leq \delta$ implies $|1/x' - 1/x| \leq \epsilon$. In fact, we may easily infer from the graph of the function that a uniform choice of δ for a given ϵ is not possible.

This leads to the definition of uniform continuity in a pseudometric space: the choice of δ for a given ϵ does not depend on a particular point, but is rather, well, uniform.

Definition 3.5.34 *The map $f : X \to Y$ into the pseudometric space (Y, d') is called* uniformly continuous *iff given $\epsilon > 0$ there exists $\delta > 0$ such that $d'(f(x), f(x')) < \epsilon$ whenever $d(x, x') < \delta$.*

Doing a game of quantifiers, let us just point out the difference between uniform continuity and continuity:

Continuity vs. uniform continuity

1. Continuity says

$$\begin{aligned} \forall \epsilon &> 0 \underline{\forall x \in X \exists \delta > 0} \forall x' \in X : d(x, x') \\ &< \delta \Rightarrow d'(f(x), f(x')) < \epsilon. \end{aligned}$$

2. Uniform continuity says

$$\begin{aligned} \forall \epsilon > 0 \underline{\exists \delta > 0 \forall x \in X} \forall x' \in X : d(x, x') \\ < \delta \Rightarrow d'(f(x), f(x')) < \epsilon. \end{aligned}$$

The formulation suggests that uniform continuity depends on the chosen metric. In contrast to continuity, which is a property depending on the topology of the underlying spaces, uniform continuity is a property of the underlying uniform space, which will be discussed below. We note that the composition of uniformly continuous maps is uniformly continuous again.

A uniformly continuous map is continuous. The converse is not true, however.

Example 3.5.35 Consider the map $f : x \mapsto x^2$, which is certainly continuous on $\mathbb{R}$. Assume that f is uniformly continuous, and fix $\epsilon > 0$, then there exists $\delta > 0$ such that $|x - y| < \delta$ always implies $|x^2 - y^2| < \epsilon$. Thus we have for all x and for all r with $0 < r \leq \delta$ that $|x^2 - (x+r)^2| = |2 \cdot x \cdot r + r^2| < \epsilon$ after Binomi's celebrated theorem. But this would mean $|2 \cdot x + r| < \epsilon / r$ for all x, which is not possible. In general, a very similar argument shows that polynomials $\sum_{i=1}^{n} a_i \cdot x^i$ with $n > 1$ and $a_n \neq 0$ are not uniformly continuous. ✌

A continuous function on a compact pseudometric space, however, is uniformly continuous. This is established through an argument constructing a cover of the space; compactness will then permit us to extract a finite cover, from which we will infer uniform continuity.

Proposition 3.5.36 *Let $f : X \to Y$ be a continuous map from the compact pseudometric space X to the pseudometric space (Y, d'). Then f is uniformly continuous.*

Proof Given $\epsilon > 0$, there exists for each $x \in X$ a positive δ_x such that $f\big[B(x, \delta_x)\big] \subseteq B_{d'}(f(x), \epsilon/3)$. Since $\{B(x, \delta_x/3) \mid x \in X\}$ is an open cover of X, and since X is compact, we find $x_1, \ldots, x_n \in X$ such that $B(x_1, \delta_{x_1}/3), \ldots, B(x_n, \delta_{x_n}/3)$ cover X. Let δ be the smallest among $\delta_{x_1}, \ldots, \delta_{x_n}$. If $d(x, x') < \delta/3$, then there exist x_i, x_j with $d(x, x_i) < \delta/3$ and $d(x', x_j) < \delta/3$, so that $d(x_i, x_j) \leq d(x_i, x) + d(x, x') + d(x', x_j) < \delta$; hence $d'(f(x_i), f(x_j)) < \epsilon/3$, and thus

$$\begin{aligned} d'(f(x), f(x')) &\leq d'(f(x), f(x_i)) + d'(f(x_i), f(x_j)) \\ &\quad + d'(f(x_j), f(x')) < 3 \cdot \epsilon/3 = \epsilon. \end{aligned}$$

This establishes uniform continuity. ⊣

Idea for a proof

One of the most attractive features of uniform continuity is that it permits certain extensions—given a uniform continuous map $f : D \to Y$ with $D \subseteq X$ dense and Y complete metric, we can extend f to a uniformly continuous map F on the whole space. This extension is necessarily unique (see Lemma 3.3.20). The basic idea is to define $F(x) := \lim_{n\to\infty} f(x_n)$, whenever $x_n \to x$ is a sequence in D which converges to x. This requires that the limit exists and that it is in this case unique; hence it demands the range to be a metric space which is complete.

Proposition 3.5.37 *Let $D \subseteq X$ be a dense subset, and assume that $f : D \to Y$ is uniformly continuous, where (Y, d') is a complete metric space. Then there exists a unique uniformly continuous map $F : X \to Y$ which extends f.*

Proof 0. We have already argued that an extension must be unique, if it exists. So we have to construct it and to show that it is uniformly continuous. We will generalize the argument from above referring to a limit by considering the oscillation at each point. A glimpse at the proof of Lemma 3.5.24 shows indeed that we argue with a limit here and are able to take into view the whole set of points which makes this possible.

Outline—use the oscillation

1. Let us have a look at the oscillation $\varnothing_f(x)$ of f at a point $x \in X$ (see page 345), and we may assume that $x \notin D$. We claim that $\varnothing_f(x) = 0$. In fact, given $\epsilon > 0$, there exists $\delta > 0$ such that $d(x', x'') < \delta$ implies $d'(f(x'), f(x'')) < \epsilon/3$, whenever $x', x'' \in D$. Thus, if $y', y'' \in f\big[D \cap B(x, \delta/2)\big]$, we find $x', x'' \in D$ with $f(x') = y', f(x'') = y''$ and $d(x', x'') \leq d(x, x') + d(x'', x) < \delta$; hence $d'(y', y'') = d'(f(x'), f(y'')) < \epsilon$. This means that $\operatorname{diam}(f\big[D \cap B(x, \delta/2)\big]) < \epsilon$.

2. Lemma 3.5.24 tells us that there exists a continuous extension F of f to the set $\{x \in X \mid \varnothing_f(x) = 0\} = X$. Hence it remains to show that F is *uniformly* continuous. Given $\epsilon > 0$, we choose the same δ as above, which did not depend on the choice of the points we were considering above. Let $x_1, x_2 \in X$ with $d(x_1, x_2) < \delta/2$; then there exists $v_1, v_2 \in D$ such that $d(x_1, v_1) < \delta/4$ with $d'(F(x_1), f(v_1)) \leq \epsilon/3$ and $d(x_2, v_2) < \delta/4$ with $d'(F(x_2), f(v_2)) \leq \epsilon/3$. We see as above that $d(v_1, v_2) < \delta$; thus $d'(f(v_1), f(v_2)) < \epsilon/3$, consequently,

$$\begin{aligned} d'(F(x_1, x_2)) &\leq d'(F(x_1), f(v_1)) \\ &\quad + d'(f(v_1), f(v_2)) + d'(f(v_2), F(x_2)) \\ &< 3 \cdot \epsilon/3 = \epsilon. \end{aligned}$$

But this means that F is uniformly continuous. $\dashv$

Looking at $x \mapsto 1/x$ on $]0, 1]$ shows that uniform continuity is indeed necessary to obtain a continuous extension.

3.5.2 Baire's Theorem and a Banach–Mazur Game

The technique of constructing a shrinking sequence of closed sets with a diameter tending to zero used for establishing Proposition 3.5.25 is helpful in establishing Baire's Theorem 3.4.13 also for complete pseudometric spaces; completeness then makes sure that the intersection is not empty. The proof is essentially a blend of this idea with the proof given above (p. 327). We will then give an interpretation of Baire's Theorem in terms of the game *Angel vs. Demon* introduced in Sect. 1.7. We show that Demon has a winning strategy iff the space is the countable union of nowhere dense sets (the space is then called to be of the *first category*). This is done for a subset of the real line but can be easily generalized.

This is the version of Baire's Theorem in a complete pseudometric space. We mimic the proof of Theorem 3.4.13, having the diameter of a set at our disposal.

Baire's Theorem

Theorem 3.5.38 *Let X be a complete pseudometric space, and then the intersection of a sequence of dense open sets is dense again.*

Proof Let $(D_n)_{n\in\mathbb{N}}$ be a sequence of dense open sets. Fix a nonempty open set G, then we have to show that $G \cap \bigcap_{n\in\mathbb{N}} D_n \neq \emptyset$. Now D_1 is dense and open; hence we find an open set V_1 and $r > 0$ such that $\mathrm{diam}(V_1^a) \leq r$ and $V_1^a \subseteq D_1 \cap G$. We select inductively in this way a sequence of open sets $(V_n)_{n\in\mathbb{N}}$ with $\mathrm{diam}(V_n^a) < r/n$ such that $V_{n+1}^a \subseteq D_n \cap V_n$. This is possible since D_n is open and dense for each $n \in \mathbb{N}$.

Hence we have in the complete space X a decreasing sequence $V_1^a \supseteq \ldots V_n^a \supseteq \ldots$ of closed sets with diameters tending to 0. Thus $\bigcap_{n\in\mathbb{N}} V_n^a = \bigcap_{n\in\mathbb{N}} V_n$ is not empty by Proposition 3.5.25, which entails $G \cap \bigcap_{n\in\mathbb{N}} D_n$ not being empty. $\dashv$

Kelley [Kel55, p. 201] remarks that there is a slight incongruence with this theorem, since the assumption of completeness is non-topological in nature (hence a property which may get lost when switching to another pseudometric; see Example 3.5.18), but we draw a topological conclusion. He suggests that the assumption on space X should be reworded to X being a topological space for which there exists a complete pseudometric. But, alas, the formulation above is the usual one, because it is pretty suggestive after all.

Definition 3.5.39 *Call a set $A \subseteq X$* nowhere dense *iff $A^{oa} = \emptyset$, i.e., the closure of the interior is empty, equivalently, iff the open set $X \setminus A^a$ is dense. The space X is said to be of the* first category *iff it can be written as the countable union of nowhere dense sets.*

Then Baire's Theorem can be reworded that the countable union of nowhere dense sets in a complete pseudometric space is nowhere dense. Cantor's ternary set constitutes an important example for a nowhere dense set:

Example 3.5.40 Cantor's ternary set C from Example 1.6.4 can be written as

Cantor's ternary set

$$C = \{\sum_{i=1}^{\infty} a_i 3^{-i} \mid a_i \in \{0, 2\} \text{ for all } i \in \mathbb{N}\}.$$

This is seen as follows: Define $[a, b]' := [a + (b - a)/3] \cup [a + 2 \cdot (b - a)/3]$ for an interval $[a, b]$, and $(A_1 \cup \ldots \cup A_\ell)' := A_1' \cup \ldots \cup A_\ell'$; then $C = \bigcap_{n \in \mathbb{N}} C_n$ with the inductive definition $C_1 := [0, 1]'$ and $C_{n+1} := C_n'$. It is shown easily by induction that

$$C_n = \{\sum_{i=1}^{\infty} a_i \cdot 3^{-i} \mid a_i \in \{0, 2\} \text{ for } i \leq n \text{ and } a_i \in \{0, 1, 2\} \text{ for } i > n\}.$$

The representation above implies that the interior of C is empty, so that C is in fact nowhere dense in the unit interval. ✌

Cantor's ternary set is a helpful device for investigating the structure of complete metric spaces which have a countable dense subset, i.e., in Polish spaces.

We will give now a game theoretic interpretation of spaces of the first category through a game which is attributed to Banach and Mazur, tying the existence of a winning strategy for Demon to spaces of the first category. For simplicity, we discuss it for a closed interval of the real line. We do not assume that the game is determined; determinacy is not necessary here (and its assumption would bring us into serious difficulties with the assumption of the validity of the axiom of choice; see Sect. 1.7.1).

Let a subset S of a closed interval $L_0 \subseteq \mathbb{R}$ be given; this set is assigned to Angel, and its adversary Demon is assigned its complement $T := L_0 \setminus S$. The game is played in this way:

Banach-Mazur game

- Angel chooses a closed interval $L_1 \subseteq L_0$,
- Demon reacts with choosing a closed interval $L_2 \subseteq L_1$,
- Angel chooses then—knowing the moves L_0 and L_1—a closed interval $L_2 \subseteq L_1$,

- and so on: Demon chooses the intervals with even numbers, and Angel selects the intervals with the odd numbers, each interval is closed and contained in the previous one; both Angel and Demon have complete information about the game's history, when making a move.

Angel wins iff $\bigcap_{n\in\mathbb{N}} L_n \cap S \neq \emptyset$; otherwise, Demon wins.

We focus on Demon's behavior. Its strategy for the n-th move is modeled as a map f_n which is defined on $2 \cdot n$-tuples $\langle L_0, \ldots, L_{2\cdot n-1}\rangle$ of closed intervals with $L_0 \supseteq L_1 \supseteq \ldots \supseteq L_{2\cdot n-1}$, taking a closed interval $L_{2\cdot n}$ as a value with

$$L_{2\cdot n} = f_n(L_0, \ldots, L_{2\cdot n-1}) \subseteq L_{2\cdot n-1}.$$

The sequence $(f_n)_{n\in\mathbb{N}}$ will be a *winning strategy* for Demon iff $\bigcap_{n\in\mathbb{N}} L_n \subseteq T$, when $(L_n)_{n\in\mathbb{N}}$ is chosen according to these rules.

The following theorem relates the existence of a winning strategy for Demon with S being of first category.

Theorem 3.5.41 *There exists a strategy for Demon to win iff S is of the first category.*

We divide the proof into two parts—we show first that we can find a strategy for Demon, if S is of the first category. The converse is technically somewhat more complicated, so we delay it and do the necessary constructions first.

Proof (First part) Assume that S is of the first category, so that we can write $S = \bigcup_{n\in\mathbb{N}} S_n$ with S_n nowhere dense for each $n \in \mathbb{N}$. Angel starts with a closed interval L_1, and then Demon has to choose a closed interval L_2; the choice will be so that $L_2 \subseteq L_1 \setminus S_1$. We have to be sure that such a choice is possible; our assumption implies that $L_1 \cap S_1^a$ is open and dense in L_1; thus it contains an open interval. In the inductive step, assume that Angel has chosen the closed interval $L_{2\cdot n-1}$ such that $L_{2\cdot n-1} \subseteq \ldots \subseteq L_2 \subseteq L_1 \subseteq L_0$. Then Demon will select an interval $L_{2\cdot n} \subseteq L_{2\cdot n-1} \setminus (S_1 \cup \ldots \cup S_n)$. For the same reason as above, the latter set contains an open interval. This constitutes Demon's strategy, and evidently $\bigcap_{n\in\mathbb{N}} L_n \cap S = \emptyset$, so Demon wins. $\dashv$

The proof for the second part requires some technical constructions. We assume that f_n assigns to each $2 \cdot n$-tuple of closed intervals $I_1 \supseteq I_2 \supseteq$

$\ldots \supseteq I_{2\cdot n}$ a closed interval $f_n(I_1, \ldots, I_{2\cdot n}) \subseteq I_{2\cdot n}$, but do not make any further assumptions, for the time being, that is. We are given a closed interval L_0 and a subset $S \subseteq L_0$.

In the first step, we define a sequence $(J_n)_{n\in\mathbb{N}}$ of closed intervals with these properties:

- $J_n \subseteq L_0$ for all $n \in \mathbb{N}$,
- $K_n := f_1(L_0, J_n)$ defines a sequence $(K_n)_{n\in\mathbb{N}}$ of mutually disjoint closed intervals,
- $\bigcup_{n\in\mathbb{N}} K_n^o$ is dense in L_0.

Let us see how to do this. Define $\mathcal{F}$ as the sequence of all closed intervals with rational endpoints that are contained in L_0^o. Take J_1 as the first element of $\mathcal{F}$. Put $K_1 := f_1(L_0, J_1)$; then K_1 is a closed interval with $K_1 \subseteq J_1$ by assumption on f_1. Let J_2 be the first element in $\mathcal{F}$ which is contained in $L_0 \setminus K_1$, and put $K_2 := f_1(L_0, J_2)$. Inductively, select J_{i+1} as the first element of $\mathcal{F}$ which is contained in $L_0 \setminus \bigcup_{t=1}^{i} K_t$, and set $K_{i+1} := f_1(L_0, J_{i+1})$. It is clear from the construction that $(K_n)_{n\in\mathbb{N}}$ forms a sequence of mutually disjoint closed intervals with $K_n \subseteq J_n \subseteq L_0$ for each $n \in \mathbb{N}$. Assume that $\bigcup_{n\in\mathbb{N}} K_n^o$ is not dense in L_0, then we find $x \in L_0$ which is not contained in this union; hence we find an interval T with rational endpoints which contains x but $T \cap \bigcup_{n\in\mathbb{N}} K_n^o = \emptyset$. So T occurs somewhere in $\mathcal{F}$, but it is never the first interval to be considered in the selection process. Since this is impossible, we arrive at a contradiction.

We repeat this process for K_i^o rather than L_0 for some i; hence we will define a sequence $(J_{i,n})_{n\in\mathbb{N}}$ of closed intervals $J_{i,n}$ with these properties:

- $J_{i,n} \subseteq K_i^o$ for all $n \in \mathbb{N}$,
- $K_{i,n} := f_2(L_0, J_i, K_i, J_{i,n})$ defines a sequence $(K_{i,n})_{n\in\mathbb{N}}$ of mutually disjoint closed intervals,
- $\bigcup_{n\in\mathbb{N}} K_{i,n}^o$ is dense in K_i.

It is immediate that $\bigcup_{i,j} K_{i,j}^o$ is dense in L_0.

Continuing inductively, we find for each $\ell \in \mathbb{N}$ two families $J_{i_1,\ldots,i_\ell}$ and $K_{i_1,\ldots,i_\ell}$ of closed intervals with these properties:

- $K_{i_1,\ldots,i_\ell} = f_\ell(L_0, J_{i_1}, K_{i_1}, J_{i_1,i_2}, K_{i_1,i_2}, \ldots, J_{i_1,\ldots,i_\ell})$,
- $J_{i_1,\ldots,i_{\ell+1}} \subseteq K^o_{i_1,\ldots,i_\ell}$,
- the intervals $(K_{i_1,\ldots,i_{\ell-1},i_\ell})_{i_\ell\in\mathbb{N}}$ are mutually disjoint for each $i_1, \ldots, i_{\ell-1}$,
- $\bigcup\{K^o_{i_1,\ldots,i_{\ell-1},i_\ell} \mid \langle i_1, \ldots, i_{\ell-1}, i_\ell\rangle \in \mathbb{N}^\ell\}$ is dense in L_0.

Relax NOW! Note that this sequence depends on the chosen sequence $(f_n)_{n\in\mathbb{N}}$ of functions that represents the strategy for Demon.

Proof (Second part) Now assume that Demon has a winning strategy $(f_n)_{n\in\mathbb{N}}$; hence no matter how Angel plays, Demon will win. For proving the assertion, we have to construct a sequence of nowhere dense subsets, the union of which is S. In the first move, Angel chooses a closed interval $L_1 := J_{i_1} \subseteq L_0$ (we refer here to the enumeration given by $\mathcal{F}$ above, so the interval chosen by Angel has index i_1). Demon's countermove is then

$$L_2 := K_{i_1} := f_1(L_0, L_1) = f_1(L_0, J_{i_1}),$$

as constructed above. In the next step, Angel selects $L_3 := J_{i_1,i_2}$ among those closed intervals which are eligible, i.e., which are contained in $K^o_{i_1}$ and have rational endpoints; Demon's countermove is

$$L_4 := K_{i_1,i_2} := f_2(L_0, L_1, L_2, L_3) = f_2(L_0, J_{i_1}, K_{i_1}, J_{i_1,i_2}).$$

In the n-th step, Angel selects $L_{2\cdot n-1} := J_{1_1,\ldots,i_n}$ and Demon selects $L_{2\cdot n} := K_{i_1,\ldots,i_n}$. Then we see that the sequence $L_0 \supseteq L_1 \ldots \supseteq L_{2\cdot n-1} \supseteq L_{2\cdot n} \ldots$ decreases and $L_{2\cdot n} = f_n(L_0, L_1, \ldots, L_{2\cdot n-1})$ holds, as required.

Put $T := S \setminus L_0$ for convenience, then $\bigcap_{n\in\mathbb{N}} L_n \subseteq T$ by assumption (after all, we assume that Demon wins); put

$$G_n := \bigcup_{\langle i_1,\ldots,i_n\rangle\in\mathbb{N}^n} K^o_{i_1,\ldots,i_n}.$$

Then G_n is open. Let $E := \bigcap_{n\in\mathbb{N}} G_n$. Given $x \in E$, there exists a unique sequence $(i_n)_{n\in\mathbb{N}}$ such that $x \in K_{i_1,\dots,i_n}$ for each $n \in \mathbb{N}$. Hence $x \in \bigcap_{n\in\mathbb{N}} L_n \subseteq T$, so that $E \subseteq T$. But then we can write

$$S = L_0 \setminus T \subseteq L_0 \setminus T = \bigcup_{n\in\mathbb{N}} (L_0 \setminus G_n).$$

Because $\bigcup\{K^o_{i_1,\dots,i_{n-1},i_n} \mid \langle i_1,\dots,i_{n-1},i_n\rangle \in \mathbb{N}^n\}$ is dense in L_0 for each $n \in \mathbb{N}$ by construction, we conclude that $L_0 \setminus G_n$ is nowhere dense, so S is of the first category. $\dashv$

Games are an interesting tool for proofs, as we can see in this example; we have shown already that games may be used for other purposes, e.g., demonstrating the each subset of [0, 1] is Lebesgue measurable under the axiom of determinacy; see Sect. 1.7.2. Further examples for using games to derive properties in a metric space can be found, e.g., in Kechris' book [Kec94].

3.6 A Gallery of Spaces and Techniques

The discussion of the basic properties and techniques suggests that we now have a powerful collection of methods at our disposal. Indeed, we set up a small gallery of showcases, in which we demonstrate some approaches and methods.

We first look at the use of topologies in logics from two different angles. The more conventional one is a direct application of the important Baire Theorem, which permits the construction of a model in a countable language of first-order logic. Here the application of the theorem lies at the heart of the application, which is a proof of Gödel's Completeness Theorem. The other vantage point starts from a calculus of observations and develops the concept of topological systems from it, stressing an order theoretic point of view by perceiving topologies as complete Heyting algebras, when considering them as partially ordered subset of the power set of their carrier. Since partial orders may generate topologies on the set they are based on, this yields an interesting interplay between order and topology, which is reflected here in the Hofmann–Mislove Theorem.

Then we return to the green pastures of classic applications and give a proof of the Stone–Weierstraß Theorem, one of the true classics. It states that a subring of the space of continuous functions on a compact Hausdorff space, which contains the constants and which separates points, is dense in the topology of uniform convergence. We actually give two proofs for this. One is based on a covering argument in a general space; it has a wide range of applications, of course. The second proof is no less interesting. It is essentially based on Weierstraß' original proof and deals with polynomials over $[0, 1]$ only; here concepts like elementary integration and uniform continuity are applied in a very concise and beautiful way.

Finally, we deal with uniform spaces; they are a generalization of pseudometric spaces, but more specific than topological spaces. We argue that the central concept is closeness of points, which is, however, formulated in conceptual rather than quantitative terms. It is shown that many concepts which appear specific to the metric approach like uniform continuity or completeness may be carried into this context. Nevertheless, uniform spaces are topological spaces, but the assumption on having a uniformity available has some consequences for the associated topology.

The reader probably misses Polish spaces in this little gallery. We deal with these spaces in depth, but since most of our applications of them are measure theoretic in nature, I decided to discuss them in the context of a discussion of measures as a kind of natural habitat; see Chap. 4.

3.6.1 Gödel's Completeness Theorem

Gödel's Completeness Theorem states that a set of sentences of first-order logic is consistent iff it has a model. The crucial part is the construction of a model for a consistent set of sentences. This is usually done through Henkin's approach; see, e.g., [Sho67, 4.2], [CK90, Chap. 2] or [Sri08, 5.1]. Rasiowa and Sikorski [RS50] followed a completely different path in their topological proof by making use of Baire's Category Theorem and using the observation that in a compact topological space, the intersection of a sequence of open and dense sets is dense again. The compact space is provided by the clopen sets of a Boolean algebra which in turn is constructed from the formulas of the first-order language upon factoring. The equivalence relation is induced by the consistent set under consideration.

We present the fundamental ideas of their proof in this section, since it is an unexpected application of a combination of the topological version of Stone's Representation Theorem for Boolean algebras and Baire's Theorem, hinted at already in Example 3.4.15. Since we assume that the reader is familiar with the semantics of first-order languages, we do not want to motivate every definition for this area in detail, but we sketch the definitions, indicate the deduction rules, say what a model is, and rather focus on the construction of the model. The references given above may be used to fill in any gaps.

A slightly informal description of the first-order language $\mathfrak{L}$ with identity which we will be working with is given first. For this, we assume that we have a countable set $\{x_n \mid n \in \mathbb{N}\}$ of variables and countably many constants. Moreover, we assume countably many function symbols and countably many predicate symbols. In particular, we have a binary relation $==$, the identity. Each function and each predicate symbol have a positive arity.

These are the components of our language $\mathfrak{L}$.

Terms. A variable is a term and a constant symbol is a term. If f is a function symbol of arity n, and $t_1, \ldots, t_n$ are terms, then $f(t_1, \ldots, t_n)$ is a term. Nothing else is a term.

Atomic Formulas. If t_1 and t_2 are terms, then $t_1 == t_2$ is an atomic formula. If p is a predicate symbol of arity n, and $t_1, \ldots, t_n$ are terms, then $p(t_1, \ldots, t_n)$ is an atomic formula.

Formulas. An atomic formula is a formula. If φ and ψ are formulas, then $\varphi \wedge \psi$ and $\neg\varphi$ are formulas. If x is a variable and φ is a formula, then $\forall x.\varphi$ is a formula. Nothing else is a formula.

Because there are countably many variables resp. constants, the language has countably many formulas.

One usually adds parentheses to the logical symbols, but we do without, using them, however, freely, when necessary. We will use also disjunction [$\varphi \vee \psi$ abbreviates $\neg(\neg\varphi \wedge \neg\psi)$], implication [$\varphi \to \psi$ for $\neg\varphi \vee \psi$], logical equivalence [$\varphi \leftrightarrow \psi$ for $(\varphi \to \psi) \wedge (\psi \to \varphi)$], and existential quantification [$\exists x.\varphi$ for $\neg(\forall x.\neg\varphi)$]. Conjunction and disjunction are associative.

We need logical axioms and inference rules as well. We have four groups of axioms:

Propositional Axioms. Each propositional tautology is an axiom.

Identity Axioms. $x == x$, when x is a variable.

Equality Axioms. $y_1 == z_1 \to \ldots \to y_n == z_n \to f(y_1, \ldots, y_n) == f(z_1, \ldots, z_n)$, whenever f is a function symbol of arity n, and $y_1 == z_1 \to \ldots \to y_n == z_n \to p(y_1, \ldots, y_n) \to p(z_1, \ldots, z_n)$ for a predicate symbol of arity n.

Substitution Axiom. If φ is a formula, $\varphi_x[t]$ is obtained from φ by freely substituting all free occurrences of variable x by term t; then $\varphi_x[t] \to \exists x.\varphi$ is an axiom.

These are the inference rules:

Modus Ponens. From φ and $\varphi \to \psi$, infer ψ.

Generalization Rule. From φ, infer $\forall x.\varphi$.

$\Sigma \vdash \varphi$

A *sentence* is a formula without free variables. Let Σ be a set of sentences and φ a formula; then we denote that φ is deducible from Σ by $\Sigma \vdash \varphi$, i.e., iff there is a proof for φ in Σ. Σ is called *inconsistent* iff $\Sigma \vdash \bot$ or, equivalently, iff each formula can be deduced from Σ. If Σ is not inconsistent, then Σ is called *consistent* or a *theory*.

Fix a theory T, and define

$$\varphi \sim \psi \text{ iff } T \vdash \varphi \leftrightarrow \psi$$

$\mathbf{B}_T$

for formulas φ and ψ; then this defines an equivalence relation on the set of all formulas. Let $\boldsymbol{B}_T$ be the set of all equivalence classes $[\varphi]$, and define

$$\begin{aligned}
[\varphi] \wedge [\psi] &:= [\varphi \wedge \psi] \\
[\varphi] \vee [\psi] &:= [\varphi \vee \psi] \\
-[\varphi] &:= [\neg\varphi].
\end{aligned}$$

Lindenbaum algebra

This defines a Boolean algebra structure on $\boldsymbol{B}_T$, the *Lindenbaum algebra* of T. The maximal element $\top$ of $\boldsymbol{B}_T$ is $\{\varphi \mid T \vdash \varphi\}$, and its minimal element $\bot$ is $\{\varphi \mid T \vdash \neg\varphi\}$. The proof that $\boldsymbol{B}_T$ is a Boolean algebra follows the lines of Lemma 1.5.40 closely; hence it can be safely omitted. It might be noted, however, that the individual steps in the proof require additional properties of $\vdash$; for example, one has to show that $T \vdash \varphi$ and $T \vdash \psi$ together imply $T \vdash \varphi \wedge \psi$. We trust that the reader is in a position to recognize and accomplish this; [Sri08, Chap. 4] provides a comprehensive catalog of useful derivation rules with their proofs.

Let φ be a formula, then denote by $\varphi(k/p)$ the formula obtained in this way:

- all bound occurrences of x_p are replaced by x_ℓ, where x_ℓ is the first variable among $x_1, x_2, \ldots$ which does not occur in φ,
- all free occurrences of x_k are replaced by x_p.

This construction is dependent on the integer ℓ, so the formula $\varphi(k/p)$ is not uniquely determined, but its class is. We have these representations in the Lindenbaum algebra for existentially resp. universally quantified formulas.

Lemma 3.6.1 *Let φ be a formula in $\mathfrak{L}$, then we have for every $k \in \mathbb{N}$:*

1. $\sup_{p\in\mathbb{N}}[\varphi(k/p)] = [\exists x_k.\varphi]$,
2. $\inf_{p\in\mathbb{N}}[\varphi(k/p)] = [\forall x_k.\varphi]$.

Proof 1. Fix $k \in \mathbb{N}$, then we have $T \vdash \varphi(k/p) \to \exists x_k.\varphi$ for each $p \in \mathbb{N}$ by the $\exists$ introduction rule. This implies $[\varphi(k/p)] \leq [\exists x_k.\varphi]$ for all $p \in \mathbb{N}$; hence $\sup_{p\in\mathbb{N}}[\varphi(k/p)] \leq [\exists x_k.\varphi]$, and thus $[\exists x_k.\varphi]$ is an upper bound to $\{[\varphi(k/p)] \mid p \in \mathbb{N}\}$ in the Lindenbaum algebra. We have to show that it is also the least upper bound, so take a formula ψ such that $[\varphi(k/p)] \leq [\psi]$ for all $k \in \mathbb{N}$. Let q be an index such that x_q does not occur free in ψ, then we conclude from $T \vdash \varphi(k/p) \to \psi$ for all p that $\exists x_q.\varphi(k/q) \to \psi$. But $T \vdash \exists x_k.\varphi \leftrightarrow \exists x_q.\varphi(k/q)$; hence $T \vdash \exists x_k.\varphi \to \psi$. This means that $[\exists x_k.\varphi]$ is the least upper bound to $\{[\varphi(k/p)] \mid p \in \mathbb{N}\}$, proving the first equality.

2. The second equality is established in a very similar way. $\dashv$

These representations motivate

Definition 3.6.2 *Let $\mathfrak{F}$ be an ultrafilter on the Lindenbaum algebra $\boldsymbol{B}_T$, $S \subseteq \boldsymbol{B}_T$.*

1. $\mathfrak{F}$ preserves the supremum of S *iff* $\sup S \in \mathfrak{F} \Leftrightarrow s \in \mathfrak{F}$ *for some* $s \in S$.
2. $\mathfrak{F}$ preserves the infimum of S *iff* $\inf S \in \mathfrak{F} \Leftrightarrow s \in \mathfrak{F}$ *for all* $s \in S$.

Preserving the supremum of a set is similar to being inaccessible by joins (see Definition 3.6.29), but inaccessibility refers to directed sets, while we are not making any assumption on S, except, of course, that its supremum exists in the Boolean algebra. Note also that one of the characteristic properties of an ultrafilters is that the join of two elements is in the ultrafilter iff it contains at least one of them. Preserving the supremum of a set strengthens this property *for this particular set only*.

The de Morgan laws and $\mathfrak{F}$ being an ultrafilter make it clear that $\mathfrak{F}$ preserves $\inf S$ iff it preserves $\sup\{-s \mid s \in S\}$, resp. that $\mathfrak{F}$ preserves $\sup S$ iff it preserves $\inf\{-s \mid s \in S\}$. This cuts our work in half.

Proposition 3.6.3 *Let $(S_n)_{n\in\mathbb{N}}$ be a sequence of subsets $S_n \subseteq \boldsymbol{B}_T$ such that $\sup S_n$ exists in $\boldsymbol{B}_T$. Then there exists an ultrafilter $\mathfrak{F}$ such that $\mathfrak{F}$ preserves the supremum of S_n for all $n \in \mathbb{N}$.*

Proof This is an application of Baire's Category Theorem 3.4.13 and is discussed in Example 3.4.15. We find there a prime ideal which does not preserve the supremum for S_n for all $n \in \mathbb{N}$. Since the complement of a prime ideal in a Boolean algebra is an ultrafilter, see Lemmas 1.5.36 and 1.5.37; the assertion follows. $\dashv$

So much for the syntactic side of our language $\mathfrak{L}$. We will leave the ultrafilter $\mathfrak{F}$ alone for a little while and turn to the semantics of the logic.

An *interpretation* of $\mathfrak{L}$ is given by a carrier set A, each constant c is interpreted through an element c_A of A, each function symbol f with arity n is assigned a map $f_A : A^n \to A$, and each n-ary predicate p is interpreted through an n-ary relation $p_A \subseteq A^n$; finally, the binary predicate $==$ is interpreted through equality on A. We also fix a sequence $\{w_n \mid n \in \mathbb{N}\}$ of elements of A for the interpretation of variables, set $\mathcal{A} := (A, \{w_n \mid n \in \mathbb{N}\})$, and call $\mathcal{A}$ a *model* for the first-order language. We then proceed inductively:

Terms. Variable x_i is interpreted by w_i. Assume that the term $f(t_1, \ldots, t_n)$ is given. If the terms $t_1, \ldots, t_n$ are interpreted through the respective elements $t_{A,1}, \ldots, t_{A,n}$ of A, then $f(t_1, \ldots, t_n)$ is interpreted through $f_A(t_{A,1}, \ldots, t_{A,n})$.

Atomic Formulas. The atomic formula $t_1 == t_2$ is interpreted through $t_{A,1} = t_{A,2}$. If the n-ary predicate p is assigned $p_A \subseteq A^n$, then $p(t_1, \ldots, t_n)$ is interpreted as $\langle t_{A,1}, \ldots, t_{A,n}\rangle \in p_A$.

$\mathcal{A} \models \varphi$

We denote by $\mathcal{A} \models \varphi$ that the interpretation of the atomic formula φ yields the value true. We say that φ *holds in* $\mathcal{A}$.

Formulas. Let φ and ψ be formulas, then $\mathcal{A} \models \varphi \wedge \psi$ iff $\mathcal{A} \models \varphi$ and $\mathcal{A} \models \psi$, and $\mathcal{A} \models \neg\varphi$ iff $\mathcal{A} \models \varphi$ is false. Let φ be the formula $\forall x_i.\psi$, then $\mathcal{A} \models \varphi$ iff $\mathcal{A} \models \psi_{x_i|a}$ for every $a \in A$, where $\psi_{x|a}$ is the formula ψ with each free occurrence of x replaced by a.

Construct the ultrafilter $\mathfrak{F}$ constructed in Proposition 3.6.3 for all possible suprema arising from existentially quantified formulas according to Lemma 3.6.1. There are countably many suprema, because the number of formulas is countable. This ultrafilter and the Lindenbaum algebra $\boldsymbol{B}_T$ will be used now for the construction of a *model* $\mathcal{A}$ for T (so that $\mathcal{A} \models \varphi$ holds for all $\varphi \in T$). Model

We will first need to define the carrier set A. Define for the variables x_i and x_j the equivalence relation $\approx$ through $x_i \approx x_j$ iff $[x_i == x_j] \in \mathfrak{F}$; denote by $\hat{x}_i$ the $\approx$-equivalence class of x_i. The carrier set A is defined as $\{\hat{x}_n \mid n \in \mathbb{N}\}$.

Let us take care of the constants now. Given a constant c, we know that $\vdash \exists x_i.c == x_i$ by substitution. Thus $[\exists x_i.c == x_i] = \top \in \mathfrak{F}$. But $[\exists x_i.c == x_i] = \sup_{i\in\mathbb{N}}[c == x_i]$, and $\mathfrak{F}$ preserves suprema, so we conclude that there exists i with $[c == x_i] \in \mathfrak{F}$. We pick this i and define $c_A := \hat{x}_i$. Note that it does not matter which i to choose. Assume that there is more than one. Since $[c == x_i] \in \mathfrak{F}$ and $[c == x_j] \in \mathfrak{F}$ imply $[c == x_i \wedge c == x_j] \in \mathfrak{F}$, we obtain $[x_i == x_j] \in \mathfrak{F}$, so the class is well defined.

Coming to terms, let t be a variable or a constant, so that it has an interpretation already, and assume that f is a unary function. Then $\vdash \exists x_i.f(t) == x_i$, so that $[\exists x_i.f(t) == x_i] \in \mathfrak{F}$; hence there exists i such that $[f(t) == x_i] \in \mathfrak{F}$, then put $f_A(t_A) := \hat{x}_i$. Again, if $[f(t) == x_i] \in \mathfrak{F}$ and $[f(t) == x_j] \in \mathfrak{F}$, then $[x_i == x_j] \in \mathfrak{F}$, so that $f_A(c_A)$ is well defined. The argument for the general case is very similar. Assume that terms $t_1, \dots, t_n$ have their interpretations already and f is a function with arity n; then $\vdash \exists x_i.f(t_1, \dots, t_n) == x_i$, and hence we find j with $[f(t_1, \dots, f(t_n) == x_j] \in \mathfrak{F}$, so put $f_A(t_{A,1}, \dots, t_{A,n}) := \hat{x}_j$. The same argument as above shows that this is well defined.

Having defined the interpretation t_A for each term t, we define for the n-ary relation symbol p the relation $p_A \subseteq A^n$ by

$$\langle t_{A,1}, \ldots, t_{A,n} \rangle \in p_A \Leftrightarrow [p(t_1, \ldots, t_n)] \in \mathfrak{F}$$

Then p_A is well defined by the equality axioms.

Thus $\mathcal{A} \models \varphi$ is defined for each formula φ; hence we know how to interpret each formula in terms of the Lindenbaum algebra of T (and the ultrafilter $\mathfrak{F}$). We can show now that a formula is valid in this model iff its class is contained in ultrafilter $\mathfrak{F}$.

Proposition 3.6.4 *$\mathcal{A} \models \varphi$ iff $[\varphi] \in \mathfrak{F}$ holds for each formula φ of $\mathfrak{L}$.*

Proof The proof is done by induction on the structure of formula φ and is straightforward, using the properties of an ultrafilter. For example,

$$\begin{aligned}
\mathcal{A} \models \varphi \wedge \psi &\Leftrightarrow \mathcal{A} \models \varphi \text{ and } \mathcal{A} \models \psi && \text{(definition)}\\
&\Leftrightarrow [\varphi] \in \mathfrak{F} \text{ and } [\psi] \in \mathfrak{F} && \text{(induction hypothesis)}\\
&\Leftrightarrow [\varphi \wedge \psi] \in \mathfrak{F} && (\mathfrak{F} \text{ is an ultrafilter})
\end{aligned}$$

For establishing the equivalence for universally quantified formulas $\forall x_i.\psi$, assume that x_i is a free variable in ψ such that $\mathcal{A} \models \psi_{x_i|a} \Leftrightarrow [\psi_{x_i|a}] \in \mathfrak{F}$ has been established for all $a \in A$. Then

$$\begin{aligned}
\mathcal{A} \models \forall x_i.\psi &\Leftrightarrow \mathcal{A} \models \psi_{x_i|a} \text{ for all } a \in A && \text{(definition)}\\
&\Leftrightarrow [\psi_{x_i|a}] \in \mathfrak{F} \text{ for all } a \in A && \text{(induction hypothesis)}\\
&\Leftrightarrow \sup_{a \in A}[\psi_{x_i|a}] \in \mathfrak{F} && (\mathfrak{F} \text{ preserves the infimum})\\
&\Leftrightarrow [\forall x_i.\psi] \in \mathfrak{F} && \text{(by Lemma 3.6.1)}
\end{aligned}$$

This completes the proof. $\dashv$

As a consequence, we have established this version of Gödel's Completeness Theorem:

Corollary 3.6.5 *$\mathcal{A}$ is a model for the consistent set T of formulas.* $\dashv$

This approach demonstrates how a topological argument is used at the center of a construction in logic. It should be noted, however, that the argument is only effective since the universe in which we work is countable. This is so because the Baire Theorem, which enables the construction of the ultrafilter, works for a countable family of open and dense sets. If, however, we work in an uncountable language $\mathfrak{L}$, this instrument is no longer available ([CK90, Exercise 2.1.24] points to a possible generalization).

But even in the countable case, one cannot help but note that the construction above depends on the axiom of choice, because we require an ultrafilter. The approach in [CK90, Exercise 2.1.22] resp. [Kop89, Theorem 2.21] suggests to construct a filter without the help of a topology, but, alas, this filter is extended to an ultrafilter, and here the dreaded axiom is needed again.

3.6.2 Topological Systems or Topology via Logic

This section investigates topological systems. They abstract from topologies being sets of subsets and concentrate on the order structure imposed by a topology instead. We focus on the interplay between a topology and the base space by considering these objects separately. A topology is considered a complete Heyting algebra; the carrier set is, well, a set of points; both are related through a validity relation $\models$ which mimics the $\in$ relation between a set and its elements. This leads to the definition of a topological system, and the question is whether this separation really bears fruits. It does; for example, we may replace the point set by the morphisms from the Heyting algebra to the two element algebras $\mathbb{2}$, giving sober spaces, and we show that, e.g., a Hausdorff space is isomorphic to such a structure.

The interplay of the order structure of a topology and its topological obligations will be investigated through the Scott topology on a dcpo, a directed complete partial order, leading to the Hofmann–Mislove Theorem which characterizes compact sets that are represented as the intersection of the open sets containing them in terms of Scott open filters.

Before we enter into a technical discussion, however, we put the following definitions on record.

Definition 3.6.6 *A partially ordered set P is called a* complete Heyting algebra *iff:*

1. *each finite subset S has a join $\bigwedge S$,*
2. *each subset S has a meet $\bigvee S$,*

3. *finite meets distribute over arbitrary joins, i.e.,*

$$a \wedge \bigvee S = \bigvee\{a \wedge s \mid s \in S\}$$

holds for $a \in L$, $S \subseteq L$.

A morphism f *between the complete Heyting algebras* P *and* Q *is a map* $f : P \to Q$ *such that:*

1. $f(\bigwedge S) = \bigwedge f\left[S\right]$ *holds for finite* $S \subseteq P$,
2. $f(\bigvee S) = \bigvee f\left[S\right]$ *holds for arbitrary* $S \subseteq P$.

$\|P, Q\|$ $\|P, Q\|$ *denotes the set of all morphisms* $P \to Q$.

The definition of a complete Heyting algebra is a bit redundant, but never mind. Because the join and the meet of the empty set is a member of such an algebra, it contains a smallest element $\bot$ and a largest element $\top$, and $f(\bot) = \bot$ and $f(\top) = \top$ follow. A topology is a complete Heyting algebra with inclusion as the partial order, as announced already in Exercise 1.27. Sometimes, complete Heyting algebras are called *frames*; but since the structure underlying the interpretation of modal logics are also called frames, we stick here to the longer name.

Example 3.6.7 Call a lattice V *pseudo-complemented* iff given $a, b \in V$, there exists $c \in V$ such that $x \leq c$ iff $x \wedge a \leq b$; c is usually denoted by $a \to b$. A complete Heyting algebra is pseudo-complemented. In fact, let $c := \bigvee\{x \in V \mid x \wedge a \leq b\}$, then

$$c \wedge a = \bigvee\{x \in V \mid x \wedge a \leq b\} \wedge a = \bigvee\{x \wedge a \in V \mid x \wedge a \leq b\} \leq b$$

by the general distributive law, hence $x \leq c$ implies $x \wedge a \leq b$. Conversely, if $x \wedge a \leq b$, then $x \leq c$ follows. ✌

Example 3.6.8 Assume that we have a complete lattice V which is pseudo-complemented. Then the lattice satisfies the general distributive law. In fact, given $a \in V$ and $S \subseteq V$, we have $s \wedge a \leq \bigvee\{a \wedge b \mid b \in S\}$, thus $s \leq a \to \bigvee\{a \wedge b \mid b \in S\}$ for all $s \in S$, from which we obtain $\bigvee S \leq a \wedge \bigvee\{a \wedge b \mid b \in S\}$, which in turn gives $a \wedge \bigvee S \leq a \wedge \bigvee\{a \wedge b \mid b \in S\}$. On the other hand, $\bigvee\{a \wedge b \mid b \in S\} \leq \bigvee S$, and $\bigvee\{a \wedge b \mid b \in S\} \leq a$, so that we obtain $\bigvee\{a \wedge b \mid b \in S\} \leq a \wedge \bigvee S$.

✌

We note:

Corollary 3.6.9 *A complete Heyting algebra is a complete distributive lattice.* $\dashv$

Quite apart from investigating what can be said if open sets are replaced by an element of a complete Heyting algebra, and thus focussing on the order structure, one can argue as follows. Suppose we have observers and events, say, X is the set of observers, and A is the set of events. The observers are not assumed to have any structure; the events have a partial order making them a distributive lattice; an observation may be incomplete, so $a \leq b$ indicates that observing event b contains more information than observing event a. If observer $x \in X$ observes event $a \in A$, we denote this as $x \models a$. The lattice structure should be compatible with the observations, that is, we want to have for $S \subseteq A$ that

$$x \models \bigwedge S \text{ iff } x \models a \text{ for all } a \in S, S \text{ finite},$$
$$x \models \bigvee S \text{ iff } x \models a \text{ for some } a \in S, S \text{ arbitrary}.$$

(recall $\bigwedge \emptyset = \top$ and $\bigvee \emptyset = \bot$). Thus our observations should be closed under finite conjunctions and arbitrary disjunctions; replacing disjunctions by intersections and conjunctions by unions, this shows a somewhat topological face. We define accordingly:

Definition 3.6.10 *A* topological system $(X^\flat, X^\sharp, \models)$ *has a set* $X^\flat$ *of points, a complete Heyting algebra* $X^\sharp$ *of observations, and a satisfaction relation* $\models \subseteq X^\flat \times X^\sharp$ *(written as* $x \models a$ *for* $\langle x, a\rangle \in \models$*) such that we have for all* $x \in X^\sharp$*:*

- *If* $S \subseteq X^\sharp$ *is finite, then* $x \models \bigvee S$ *iff* $x \models a$ *for all* $a \in S$*.*
- *For* $S \subseteq X^\sharp$ *arbitrary,* $x \models \bigvee S$ *iff* $x \models a$ *for some* $a \in S$*.*

The elements of $X^\flat$ *are called* points*, and the elements of* $X^\sharp$ *are called* opens. $X^\flat, X^\sharp$

We will denote a topological system $X = (X^\flat, X^\sharp)$ usually without writing down the satisfaction relation, which is either explicitly defined or understood from the context.

Example 3.6.11 1. The obvious example for a topological system D is a topological space (X, τ) with $D^\flat := X$ and $D^\sharp := \tau$, ordered through inclusion. The satisfaction relation $\models$ is given by the containment relation $\in$, so that we have $x \models G$ iff $x \in G$ for $x \in D^\flat$ and $G \in D^\sharp$.

Extension $(\!| \cdot |\!)$

2. But it works the other way around as well. Given a topological system X, define for the open $a \in X^\sharp$ its *extension*

$$(\!|a|\!) := \{x \in X^\flat \mid x \models a\}.$$

Then $\tau := \{(\!|a|\!) \mid a \in X^\sharp\}$ is a topology on $X^\flat$. In fact, $\emptyset = (\!|\bot|\!)$, $X^\flat = (\!|\top|\!)$, and if $S \subseteq \tau$ is finite, say, $S = \{(\!|a_1|\!), \dots, (\!|a_n|\!)\}$, then $\bigcap S = (\!| \bigwedge_{i=1}^n a_i |\!)$. Similarly, if $S = \{(\!|a_i|\!) \mid i \in I\} \subseteq \tau$ is an arbitrary subset of τ, then $\bigcup S = (\!| \bigvee_{i \in I} a_i |\!)$. This follows directly from the laws of a topological system.

$\mathbb{2}$

3. Put $\mathbb{2} := \{\bot, \top\}$; then this is a complete Heyting algebra. Let $X^\sharp := A$ be another complete Heyting algebra, and put $X^\flat := \|X^\sharp, \mathbb{2}\|$ defining $x \models a$ iff $x(a) = \top$ then yields a topological system. Thus a point in this topological system is a morphism $X^\sharp \to \mathbb{2}$, and a point satisfies the open a iff it assigns $\top$ to it.

✌

Next, we want to define morphisms between topological systems. Before we do that, we have another look at topological spaces and continuous maps. Recall that a map $f : X \to Y$ between topological spaces (X, τ) and (Y, ϑ) is τ-ϑ-continuous iff $f^{-1}\big[H\big] \in \tau$ for all $H \in \vartheta$. Thus f spawns a map $f^{-1} : \vartheta \to \tau$—note the opposite direction. We have $x \in f^{-1}\big[H\big]$ iff $f(x) \in H$, accounting for containment.

This leads to the definition of a morphism as a pair of maps, one working in the opposite direction of the other one, such that the satisfaction relation is maintained, formally:

Definition 3.6.12 *Let X and Y be topological systems. Then $f : X \to Y$ is a* c-morphism *iff:*

1. *f is a pair of maps $f = (f^\flat, f^\sharp)$ with $f^\flat : X^\flat \to Y^\flat$, and $f^\sharp \in \|Y^\sharp, X^\sharp\|$ is a morphism for the underlying algebras.*

$f^\flat, f^\sharp$

2. *$f^\flat(x) \models_Y b$ iff $x \models_X f^\sharp(b)$ for all $x \in X^\flat$ and all $b \in Y^\sharp$.*

We have indicated above for the reader's convenience in which system the satisfaction relation is considered. It is evident that the notion of continuity is copied from topological spaces, taking the slightly different scenario into account.

Example 3.6.13 Let X and Y be topological systems with $f : X \to Y$ a c-morphism. Let $(X^\flat, \tau_{X^\flat})$ and $(Y^\flat, \tau_{Y^\flat})$ be the topological spaces generated from these systems through the extent of the respective opens, as in Example 3.6.11, part 2. Then $f^\flat : X^\flat \to Y^\flat$ is $\tau_{X^\flat}$-$\tau_{Y^\flat}$-continuous. In fact, let $b \in Y^\sharp$, then

$$x \in (f^\flat)^{-1}\big[(\!|b|\!)\big] \Leftrightarrow f^\flat(x) \in (\!|b|\!) \Leftrightarrow f^\flat(x) \models b \Leftrightarrow x \models f^\sharp(b);$$

thus
$$(f^\flat)^{-1}\big[(\!|b|\!)\big] = (\!|f^\sharp(b)|\!) \in \tau_{X^\flat}.$$

This shows that continuity of topological spaces is a special case of c-morphisms between topological systems, in the same way as topological spaces are special cases of topological systems. ✌

Let $f : X \to Y$ and $g : Y \to Z$ be c-morphisms; then their composition is defined as $g \circ f := (g^\flat \circ f^\flat, f^\sharp \circ g^\sharp)$. The identity $id_X : X \to X$ is defined through $id_X := (id_{X^\flat}, id_{X^\sharp})$. If, given the c-morphism $f : X \to Y$, there is a c-morphisms $g : Y \to X$ with $g \circ f = id_X$ and $f \circ g = id_Y$, then f is called a *homeomorphism*.

Corollary 3.6.14 *Topological systems for a category* ***TS****, the objects of which are topological systems, with c-morphisms as morphisms.* $\dashv$

Given a topological system X, the topological space $(X^\flat, \tau_{X^\flat})$ with $\tau_{X^\flat} := \big\{(\!|a|\!) \mid a \in X^\sharp\big\}$ is called the *spatialization* of X and denoted by $\boldsymbol{SP}(X)$. We want to make $\boldsymbol{SP}$ a (covariant) functor $\boldsymbol{TS} \to \boldsymbol{Top}$, the latter one denoting the category of topological spaces with continuous maps as morphisms. Thus we have to define the image $\boldsymbol{SP}(f)$ of a c-morphism $f : X \to Y$. But this is fairly straightforward, since we have shown in Example 3.6.13 that f induces a continuous map $(X^\flat, \tau_{X^\flat}) \to (Y^\flat, \tau_{Y^\flat})$. It is clear now that $\boldsymbol{SP} : \boldsymbol{TS} \to \boldsymbol{Top}$ is a covariant functor. On the other hand, part 1 of Example 3.6.11 shows that we have a forgetful functor $\boldsymbol{V} : \boldsymbol{Top} \to \boldsymbol{TS}$ with $\boldsymbol{V}(X, \tau) := (X^\flat, X^\sharp)$ with $X^\flat := X$ and $X^\sharp := \tau$, and $\boldsymbol{V}(f) := (f, f^{-1})$. These functors are related. ***TS*, *Top*, *SP***

Proposition 3.6.15 *$\boldsymbol{SP}$ is right adjoint to $\boldsymbol{V}$.*

Proof 0. Given a topological space X and a topological system A, we have to find a bijection $\varphi_{X,A} : \hom_{\boldsymbol{TS}}(\boldsymbol{V}(X), A) \to \hom_{\boldsymbol{Top}}(X, \boldsymbol{SP}(A))$ rendering these diagrams commutative:

$$\begin{tikzcd} \hom_{\boldsymbol{TS}}(\boldsymbol{V}(X), A) \arrow[r, "\varphi_{X,A}"] \arrow[d, "F_*"'] & \hom_{\boldsymbol{Top}}(X, \boldsymbol{SP}(A)) \arrow[d, "(\boldsymbol{SP}(F))_*"] \\ \hom_{\boldsymbol{TS}}(\boldsymbol{V}(X), B) \arrow[r, "\varphi_{X,B}"'] & \hom_{\boldsymbol{Top}}(X, \boldsymbol{SP}(B)) \end{tikzcd}$$

and

$$\begin{tikzcd} \hom_{\boldsymbol{TS}}(\boldsymbol{V}(X), A) \arrow[r, "\varphi_{X,A}"] \arrow[d, "(\boldsymbol{V}(G))^*"'] & \hom_{\boldsymbol{Top}}(X, \boldsymbol{SP}(A)) \arrow[d, "G^*"] \\ \hom_{\boldsymbol{TS}}(\boldsymbol{V}(Y), A) \arrow[r, "\varphi_{Y,A}"'] & \hom_{\boldsymbol{Top}}(Y, \boldsymbol{SP}(A)) \end{tikzcd}$$

where $F_* := \hom_{\boldsymbol{TS}}(\boldsymbol{V}(X), F) : f \mapsto F \circ f$ for $F : A \to B$ in $\boldsymbol{TS}$, and $G^* := \hom_{\boldsymbol{Top}}(G, \boldsymbol{SP}(A)) : g \mapsto g \circ G$ for $G : Y \to X$ in $\boldsymbol{Top}$; see Sect. 2.5.

We define $\varphi_{X,A}(f^\flat, f^\sharp) := f^\flat$; hence we focus on the component of a c-morphism which maps points to points.

1. Let us work on the first diagram. Take $f = (f^\flat, f^\sharp) : \boldsymbol{V}(X) \to A$ as a morphism in $\boldsymbol{TS}$, and let $F : A \to B$ be a c-morphism, $F = (F^\flat, F^\sharp)$; then $\varphi_{X,B}(F_*(f)) = \varphi_{X,B}(F \circ f) = F^\flat \circ f^\flat$, and $(\boldsymbol{SP}(F))_*(\varphi_{X,A}(f)) = \boldsymbol{SP}(F) \circ f^\flat = F^\flat \circ f^\flat$.

2. Similarly, chasing f through the second diagram for some continuous map $G : Y \to X$ yields

$$\begin{aligned} \varphi_{Y,A}((\boldsymbol{V}(G))^*(f)) &= \varphi_{Y,A}((f^\flat, f^\sharp) \circ (G, G^{-1})) \\ &= f^\flat \circ G = G^*(f^\flat) = G^*(\varphi_{X,A}(f)). \end{aligned}$$

This completes the proof. $\dashv$

Constructing $\boldsymbol{SP}$, we went from a topological space to its associated topological system by exploiting the observation that a topology τ is a complete Heyting algebra. But we can travel in the other direction as well, as we will show now.

Given a complete Heyting algebra A, we take the elements of A as opens and take all morphisms in $\|A, \mathbb{2}\|$ as points, defining the relation $\models$ which connects the components through

$$x \models a \Leftrightarrow x(a) = \top.$$

This construction was announced already in Example 3.6.11, part 3. In order to extract a functor from this construction, we have to cater for morphisms. In fact, let $\psi \in \|B, A\|$ be a morphism $B \to A$ of the complete Heyting algebras B and A and $p \in \|A, \mathbb{2}\|$ a point of A; then $p \circ \psi \in \|B, \mathbb{2}\|$ is a point in B. Let $\boldsymbol{cHA}$ be the category of all complete Heyting algebras with $\mathrm{hom}_{\boldsymbol{cHA}}(A, B) := \|B, A\|$, then we define the functor $\boldsymbol{Loc} : \boldsymbol{cHA} \to \boldsymbol{TS}$ through $\boldsymbol{Loc}(A) := (\|A, \mathbb{2}\|, A)$, and $\boldsymbol{Loc}(\psi) := (\psi_*, \psi)$ for $\psi \in \mathrm{hom}_{\boldsymbol{cHA}}(A, B)$ with $\psi^*(p) := p \circ \psi$. Thus $\boldsymbol{Loc}(\psi) : \boldsymbol{Loc}(A) \to \boldsymbol{Loc}(B)$, if $\psi : A \to B$ in $\boldsymbol{cHA}$. In fact, let $f := \boldsymbol{Loc}(\psi)$, and $p \in \|A, \mathbb{2}\|$ a point in $\boldsymbol{Loc}(A)$, then we obtain for $b \in B$ *cHA*, *Loc*

$$
\begin{aligned}
f^\flat(p) \models b &\Leftrightarrow f^\flat(p)(b) = \top \\
&\Leftrightarrow (p \circ \psi)(b) = \top && (\text{since } f^\flat = p \circ \psi) \\
&\Leftrightarrow p \models \psi(b) \\
&\Leftrightarrow p \models f^\sharp(b) && (\text{since } f^\sharp = \psi).
\end{aligned}
$$

This shows that $\boldsymbol{Loc}(\psi)$ is a morphism in $\boldsymbol{TS}$. $\boldsymbol{Loc}(A)$ is called the *localization* of the complete Heyting algebra A. The topological system is called *localic* iff it is homeomorphic to the localization of a complete Heyting algebra. Localization

We have also here a forgetful functor $\boldsymbol{V} : \boldsymbol{TS} \to \boldsymbol{cHA}$, and with a proof very similar to the one for Proposition 3.6.15, one shows:

Proposition 3.6.16 *$\boldsymbol{Loc}$ is left adjoint to the forgetful functor $\boldsymbol{V}$.* $\dashv$

In a localic system, the points enjoy as morphisms evidently much more structure than just being flat points without a face, in an abstract set. Before exploiting this wondrous remark, recall these notations, where $(P, \leq)$ is a reflexive and transitive relation:

$$
\begin{aligned}
\uparrow p &:= \{q \in P \mid q \geq p\}, \\
\downarrow p &:= \{q \in P \mid q \leq p\}.
\end{aligned}
$$

The following properties are stated just for the record.

Lemma 3.6.17 *Let $a \in A$ with A a complete Heyting algebra. Then $\uparrow a$ is a filter, and $\downarrow a$ is an ideal in A.* $\dashv$

Definition 3.6.18 *Let A be a complete Heyting algebra.*

1. *$a \in A$ is called a* prime element *iff $\downarrow a$ is a prime ideal.*
2. *The filter $F \subseteq A$ is called* completely prime *iff $\bigvee S \in F$ implies $s \in F$ for some $s \in S$, where $S \subseteq A$.*

Thus $a \in A$ is a prime element iff we may conclude from $\bigwedge S \leq a$ that there exists $s \in S$ with $s \leq a$, provided $S \subseteq A$ is finite. Note that a prime filter has the stipulated property for finite $S \subseteq A$, so a completely prime filter is a prime filter by implication.

Example 3.6.19 Let (X, τ) be a topological space, $x \in X$, then

$$\mathcal{G}_x := \{G \in \tau \mid x \in G\}$$

is a completely prime filter in τ. It is clear that $\mathcal{G}_x$ is a filter in τ, since it is closed under finite intersections, and $G \in \mathcal{G}_x$ and $G \subseteq H$ implies $H \in \mathcal{G}_x$ for $H \in \tau$. Now let $\bigcup_{i \in I} S_i \in \mathcal{G}_x$ with $S_i \in \tau$ for all $i \in I$, then there exists $j \in I$ such that $x \in S_j$, hence $S_j \in \mathcal{G}_x$. ✌

Prime filters in a complete Heyting algebra have this useful property: if we have an element which is not in the filter, then we can find a prime element not in the filter dominating the given one. The proof of this property requires the axiom of choice through Zorn's Lemma.

Proposition 3.6.20 *Let $F \subseteq A$ be a prime filter in the complete Heyting algebra A. Let $a \notin F$; then there exists a prime element $p \in A$ with $a \leq p$ and $p \notin F$.*

Proof Let $Z := \{b \in A \mid a \leq b \text{ and } b \notin F\}$, then $Z \neq \emptyset$, since $a \in Z$. We want to show that Z is inductively ordered. Hence take a chain $C \subseteq Z$, then $c := \sup C \in A$, since A is a complete lattice. Clearly, $a \leq c$; suppose $c \in F$; then, since F is completely prime, we find $c' \in C$ with $c' \in F$, which contradicts the assumption that $C \subseteq Z$. But this means that Z contains a maximal element p by Zorn's Lemma.

Since $p \in Z$, we have $a \leq p$ and $p \notin F$, so we have to show that p is a prime element. Assume that $x \wedge y \leq p$, then either of $x \vee p$ or $y \vee p$ is not in F: if both are in F, we have by distributivity $(x \vee p) \wedge (y \vee p) = (x \wedge y) \vee p = p$, so $p \in F$, since F is a filter; this is a contradiction. Assume that $x \vee p \notin F$, then $a \leq x \vee p$, since $a \leq p$; hence even $x \vee p \in Z$. Since p is maximal, we conclude $x \vee p \leq p$, which entails $x \leq p$. Thus p is a prime element. $\dashv$

The reader might wish to compare this statement to an argument used in the proof of Stone's Representation Theorem; see Sect. 1.5.7. There it is established that in a Boolean algebra, we may find for each ideal a prime ideal which contains it. The argumentation is fairly similar, but, alas, one works there in a Boolean algebra, and not in a complete Heyting algebra, as we do presently.

This is a characterization of completely prime filters and prime elements in a complete Heyting algebra in terms of morphisms into $\mathbb{2}$. We will use this characterization later on.

Lemma 3.6.21 *Let A be a complete Heyting algebra, then:*

1. *$F \subseteq A$ is a completely prime filter iff $F = f^{-1}(\top) := f^{-1}\big[\{\top\}\big]$ for some $f \in \|A, \mathbb{2}\|$.*
2. *$I = f^{-1}(\bot)$ for some $f \in \|A, \mathbb{2}\|$ iff $I =\downarrow p$ for some prime element $p \in A$.*

Proof 1. Let $F \subseteq A$ be a completely prime filter, and define

$$f(a) := \begin{cases} \top, & \text{if } a \in F \\ \bot, & \text{if } a \notin F \end{cases}$$

Then $f : A \to \mathbb{2}$ is a morphism for the complete Heyting algebras A and $\mathbb{2}$. Since F is a filter, we have $f(\bigwedge S) = \bigwedge_{s \in S} f(s)$ for $S \subseteq A$ finite. Let $S \subseteq A$, then

$$\bigvee_{s \in S} f(s) = \top \Leftrightarrow f(s) = \top \text{ for some } s \in S \Leftrightarrow f(\bigvee S) = \top,$$

since F is completely prime. Thus $f \in \|A, \mathbb{2}\|$ and $F = f^{-1}(\top)$. Conversely, given $f \in \|A, \mathbb{2}\|$, the filter $f^{-1}(\top)$ is certainly completely prime.

2. Assume that $I = f^{-1}(\bot)$ for some $f \in \|A, \mathbb{2}\|$, and put

$$p := \bigvee\{a \in A \mid f(a) = \bot\}.$$

Since A is complete, we have $p \in A$, and if $a \leq p$, then $f(a) = \bot$. Conversely, if $f(a) = \bot$, then $a \leq p$, so that $I =\downarrow p$; moreover, I is a prime ideal, for $f(a) \wedge f(b) = \bot$ iff $f(a) = \bot$ or $f(b) = \bot$; thus $a \wedge b \in I$ implies $a \in I$ or $b \in I$. Consequently, p is a prime element.

Let, conversely, the prime element p be given, then one shows as in part 1 that

$$f(a) := \begin{cases} \bot, & \text{if } a \le p \\ \top, & \text{otherwise} \end{cases}$$

defines a member of $\|A, \mathbb{2}\|$ with $\downarrow p = f^{-1}(\bot)$. ⊣

Φ_X Continuing Example 3.6.19, we see that there exists for a topological space $X := (X, \tau)$ for each $x \in X$ an element $f_x \in \|\tau, \mathbb{2}\|$ such that $f_x(G) = \top$ iff $x \in G$. Define the map $\Phi_X : X \to \|\tau, \mathbb{2}\|$ through $\Phi_X(x) := f_x$ (so that $\Phi_X(x) = f_x$ iff $\mathcal{G}_x = f_x^{-1}(\top)$). We will examine Φ_X now in a little greater detail.

Lemma 3.6.22 *Φ_X is injective iff X is a T_0-space.*

Proof Let Φ_X be injective, $x \neq y$, then $\mathcal{G}_x \neq \mathcal{G}_y$. Hence there exists an open set G which contains one of x, y, but not the other. If, conversely, X is a T_0-space, then we have by the same argumentation $\mathcal{G}_x \neq \mathcal{G}_y$ for all x, y with $x \neq y$, so that Φ_X is injective. ⊣

Well, that is not too bad, because the representation of elements into $\|\tau, \mathbb{2}\|$ is reflected by a (very basic) separation axiom. Let us turn to surjectivity. For this, we need to transfer reducibility to the level of open or closed sets; since this is formulated most concisely for closed sets, we use this alternative. A closed set is called irreducible iff each of its covers with closed sets entails its being covered already by one of them (but compare Exercise 1.20), formally:

Definition 3.6.23 *A closed set $F \subseteq X$ is called* irreducible *iff $F \subseteq \bigcup_{i \in I} F_i$ implies $F \subseteq F_i$ for some $i \in I$ for any family $(F_i)_{i \in I}$ of closed sets.*

Thus a closed set F is irreducible iff the open set $X \setminus F$ is a prime element in τ. Let us see: Assume that F is irreducible, and let $\bigcap_{i \in I} G_i \subseteq X \setminus F$ for some open sets $(G_i)_{i \in I}$. Then $F \subseteq \bigcup_{i \in I} X \setminus G_i$ with $X \setminus G_i$ closed; thus there exists $j \in I$ with $F \subseteq X \setminus G_j$, and hence $G_j \subseteq X \setminus F$. Thus $\downarrow (X \setminus F)$ is a prime ideal in τ. One argues in exactly the same way for showing that if $\downarrow (X \setminus F)$ is a prime ideal in τ, then F is irreducible.

Now we have this characterization of surjectivity of our map Φ_X through irreducible closed sets.

Lemma 3.6.24 *Φ_X is onto iff for each irreducible closed set F there exists $x \in X$ such that $F = \{x\}^a$.*

Proof 1. Let Φ_X be onto, $F \subseteq X$ be irreducible. By the argumentation above, $X \setminus F$ is a prime element in τ; thus we find $f \in \|\tau, \mathbb{2}\|$ with $\downarrow (X \setminus F) = f^{-1}(\bot)$. Since Φ_X is into, we find $x \in X$ such that $f = \Phi_X(x)$; hence we have $x \notin G \Leftrightarrow f(x) = \bot$ for all open $G \subseteq X$. It is then elementary to show that $F = \{x\}^a$.

2. Let $f \in \|\tau, \mathbb{2}\|$, then we know that $f^{-1}(\bot) = \downarrow G$ for some prime open G. Put $F := X \setminus G$, then F is irreducible and closed; hence $F = \{x\}^a$ for some $x \in X$. Then we infer $f(H) = \top \Leftrightarrow x \in H$ for each open set H, so we have indeed $f = \Phi_X(x)$. Hence Φ_X is onto. $\dashv$

Thus, if Φ_X is a bijection, we can recover (the topology on) X from the morphisms on the complete Heyting algebra $\|\tau, \mathbb{2}\|$.

Definition 3.6.25 *A topological space (X, τ) is called* sober[5] *iff $\Phi_X : X \to \|\tau, \mathbb{2}\|$ is a bijection.*

Thus we obtain as a consequence this characterization.

Corollary 3.6.26 *Given a topological space X, the following conditions are equivalent:*

- *X is sober.*
- *X is a T_0-space and for each irreducible closed set F there exists $x \in X$ with $F = \{x\}^a$.*

Exercise 3.32 shows that each Hausdorff space is sober. This property is, however, seldom made use of in the context of classic applications of Hausdorff spaces in, say, analysis.

Before continuing, we generalize the Scott topology, which has been defined in Example 3.1.6 for inductively ordered sets. The crucial property is closedness under joins, and we stated this property in a linearly ordered set by saying that, if the supremum of a set S is in a Scott open set G, then we should find an element $s \in S$ with $s \in G$. This will have

[5] The rumors in the domain theory community that a certain *Johann Heinrich-Wilhelm Sober* was a skat partner of Hilbert's gardener at Göttingen could not be confirmed—anyway, what about the third man?

to be relaxed somewhat. Let us analyze the argument why the intersection $G_1 \cap G_2$ of two Scott open sets (old version) G_1 and G_2 is open by taking a set S such that $\bigvee S \in G_1 \cap G_2$. Because G_i is Scott open, we find $s_i \in S$ with $s_i \in G_i$ $(i = 1, 2)$, and because we work in a linear ordered set, we know that either $s_1 \leq s_2$ or $s_2 \leq s_1$. Assuming $s_1 \leq s_2$, we conclude that $s_2 \in G_1$, because open sets are upper closed, so that $G_1 \cap G_2$ is indeed open. The crucial ingredient here is evidently that we can find for two elements of S an element which dominates both, and this is the key to the generalization.

We want to be sure that each directed set has an upper bound; this is the case, e.g., when we are working in a complete Heyting algebra. The structure we are defining now, however, is considerably weaker, but makes sure that we can do what we have in mind.

dcpo

Definition 3.6.27 *A partially ordered set in which every directed subset has an upper bound is called a* directed completed partial ordered set, *abbreviated as* dcpo.

Evidently, complete Heyting algebras are dcpos; in particular topologies are under inclusion. Sober topological spaces with the specialization order induced by the open sets, as introduced in the next example, furnish another example for a dcpo.

Example 3.6.28 Let $X = (X, \tau)$ be a sober topological space. Hence the points in X and the morphisms in $\|\tau, \mathbb{2}\|$ are in a bijective correspondence. Define for $x, x' \in X$ the relation $x \sqsubseteq x'$ iff we have for all open sets $x \models G \Rightarrow x' \models G$ (thus $x \in G$ implies $x' \in G$). If we think that being contained in more open sets means having better information, $x \sqsubseteq x'$ is then interpreted as x' being better informed than x; $\sqsubseteq$ is sometimes called the *specialization order*.

Then $(X, \sqsubseteq)$ is a partially ordered set, antisymmetry following from the observation that a sober space is a T_0-space. But $(X, \sqsubseteq)$ is also a dcpo. Let $S \subseteq X$ be a directed set, then $L := \Phi_X\left[S\right]$ is directed in $\|\tau, \mathbb{2}\|$. Define

$$p(G) := \begin{cases} \top, & \text{if there exists } \ell \in L \text{ with } \ell(G) = \top \\ \bot, & \text{otherwise} \end{cases}$$

We claim that $p \in \|\tau, \mathbb{2}\|$. It is clear that $p(\bigvee W) = \bigvee_{w \in W} p(w)$ for $W \subseteq \tau$. Now let $W \subseteq \tau$ be finite, and assume that $\bigwedge p\left[W\right] = \top$; hence $p(w) = \top$ for all $w \in W$. Thus we find for each $w \in W$ some

$\ell_w \in L$ with $\ell_w(w) = \top$. Because L is directed, and W is finite, we find an upper bound $\ell \in L$ to $\{\ell_w \mid w \in W\}$, hence $\ell(w) = \top$ for all $w \in W$, so that $\ell(\bigwedge W) = \top$, and hence $p(\bigwedge W) = \top$. This implies $\bigwedge p\left[W\right] = p(\bigwedge W)$. Thus $p \in \|\tau, \mathbb{2}\|$, so that there exists $x \in X$ with $x = \Phi_X(p)$. Clearly, x is an upper bound to S. ✌

Definition 3.6.29 *Let* $(P, \leq)$ *be a dcpo, then* $U \subseteq P$ *is called* Scott open *iff:*

1. *U is upper closed.*
2. *If* $\sup S \in U$ *for some directed set* S*, then there exists* $s \in S$ *with* $s \in U$.

The second property can be described as *inaccessibility through directed joins*: If U contains the directed join of a set, it must contain already one of its elements. The following example is taken from [GHK+03, p. 136].

Example 3.6.30 The powerset $\mathcal{P}(X)$ of a set X is a dcpo under inclusion. The sets $\{\mathcal{F} \subseteq \mathcal{P}(X) \mid \mathcal{F}$ is of weakly finite character$\}$ are Scott open ($\mathcal{F} \subseteq \mathcal{P}(X)$ is of *weakly finite character* iff this condition holds: $F \in \mathcal{F}$ iff some finite subset of F is in $\mathcal{F}$). Let $\mathcal{F}$ be of weakly finite character. Then $\mathcal{F}$ is certainly upper closed. Now let $S := \bigcup \mathcal{S} \in \mathcal{F}$ for some directed set $\mathcal{S} \subseteq \mathcal{P}(X)$; thus there exists a finite subset $F \subseteq S$ with $F \in \mathcal{F}$. Because $\mathcal{S}$ is directed, we find $S_0 \in \mathcal{S}$ with $F \subseteq S_0$, so that $S_0 \in \mathcal{F}$. ✌

In a topological space, each compact set gives rise to a Scott open filter as a subset of the topology.

Lemma 3.6.31 *Let* (X, τ) *be a topological space and* $C \subseteq X$ *compact, then*

$$H(C) := \{U \in \tau \mid C \subseteq U\}$$

is a Scott open filter.

Proof Since $H(C)$ is upper closed and a filter, we have to establish that it is not accessible by directed joins. In fact, let $\mathcal{S}$ be a directed subset of τ such that $\bigcup \mathcal{S} \in H(C)$. Then $\mathcal{S}$ forms a cover of the compact set C; hence there exists $\mathcal{S}_0 \subseteq \mathcal{S}$ finite such that $C \subseteq \bigcup \mathcal{S}_0$. But $\mathcal{S}$ is directed, so $\mathcal{S}_0$ has an upper bound $S \in \mathcal{S}$; thus $S \in H(C)$. ⊣

Scott opens form in fact a topology, and continuous functions are characterized in a fashion similar to Example 3.1.11. We just state and prove these properties for completeness, before entering into a discussion of the Hofmann–Mislove Theorem.

Proposition 3.6.32 *Let $(P, \leq)$ be a dcpo:*

1. *$\{U \subseteq P \mid U$ is Scott open$\}$ is a topology on P, the Scott topology of P.*
2. *$F \subseteq P$ is Scott closed iff F is downward closed ($x \leq y$ and $y \in F$ imply $x \in F$) and closed with respect to suprema of directed subsets.*
3. *Given a dcpo $(Q, \leq)$, a map $f : P \to Q$ is continuous with respect to the corresponding Scott topologies iff f preserves directed joins (i.e., if $S \subseteq P$ is directed, then $f\left[S\right] \subseteq Q$ is directed and $\sup\ f\left[S\right] = f(\sup\ S)$).*

Proof 1. Let U_1, U_2 be Scott open, and $\sup\ S \in U_1 \cap U_2$ for the directed set S. Then there exist $s_i \in S$ with $s_i \in G_i$ for $i = 1, 2$. Since S is directed, we find $s \in S$ with $s \geq s_1$ and $s \geq s_2$, and since U_1 and U_2 both are upper closed, we conclude $s \in U_1 \cap U_2$. Because $U_1 \cap U_2$ is plainly upper closed, we conclude that $U_1 \cap U_2$ is Scott open; hence the set of Scott opens is closed under finite intersections. The other properties of a topology are evidently satisfied. This establishes the first part.

2. The characterization of closed sets follows directly from the one for open sets by taking complements.

3. Let $f : P \to Q$ be Scott-continuous. Then f is monotone: if $x \leq x'$, then x' is contained in the closed set $f^{-1}\left[\downarrow f(x')\right]$; thus $x \in f^{-1}\left[\downarrow f(x')\right]$, and hence $f(x) \leq f(x')$. Now let $S \subseteq P$ be directed, then $f\left[S\right] \subseteq Q$ is directed by assumption, and $S \subseteq f^{-1}\left[\downarrow (\sup_{s \in S} f(s))\right]$. Since the latter set is closed, we conclude that it contains $\sup\ S$, hence $f(\sup\ S) \leq \sup\ f\left[S\right]$. On the other hand, since f is monotone, we know that $f(\sup\ S) \geq \sup\ f\left[S\right]$. Thus f preserves directed joins.

Assume that f preserves directed joins, then, if $x \leq x'$,

$$f(x') = f(\sup\ \{x, x'\}) = \sup\ \{f(x), f(x')\}$$

follows; hence f is monotone. Now let $H \subseteq Q$ be Scott open, then $f^{-1}\left[H\right]$ is upper closed. Let $S \subseteq P$ be directed, and assume that

$\sup f\big[S\big] \in H$; then there exists $s \in S$ with $f(s) \in H$, and hence $s \in f^{-1}\big[H\big]$, which therefore is Scott open. Hence f is Scott continuous. $\dashv$

Following [GHK+03, Chap. II-1], we show that in a sober space, there is an order morphism between Scott open filters and certain compact subsets. Preparing for this, we observe that in a sober space, every open subset which contains the intersection of a Scott open filter is already an element of the filter. This will turn out to be a consequence of the existence of prime elements not contained in a prime filter, as stated in Proposition 3.6.20.

Lemma 3.6.33 *Let $\mathcal{F} \subseteq \tau$ be a Scott open filter of open subsets in a sober topological space (X, τ). If $\bigcap \mathcal{F} \subseteq U$ for the open set U, then $U \in \mathcal{F}$.*

Proof 0. The plan of the proof goes like this: Since $\mathcal{F}$ is Scott open, it is a prime filter in τ. We assume that there exists an open set which contains the intersection, but which is not in $\mathcal{F}$. This is exactly the situation in Proposition 3.6.20, so there exists an open set which is maximal with respect to not being a member of $\mathcal{F}$ and which is prime; hence we may represent this set as $f^{-1}(\bot)$ for some $f \in \|\tau, \mathbb{2}\|$. But now sobriety kicks in, and we represent f through an element $x \in X$. This will then lead us to the desired contradiction. Plan

1. Because $\mathcal{F}$ is Scott open, it is a prime filter in τ. Let $G := \bigcap \mathcal{F}$, and assume that U is open with $G \subseteq U$ (note that we do not know whether or not G is empty). Assume that $U \notin \mathcal{F}$, then we obtain from Proposition 3.6.20 a prime open set V which is not in $\mathcal{F}$, which contains U, and which is maximal. Since V is prime, there exists $f \in \|\tau, \mathbb{2}\|$ such that $\{H \in \tau \mid f(H) = \bot\} =\downarrow V$ by Lemma 3.6.21. Since X is sober, we find $x \in X$ such that $\Phi_X(x) = f$; hence $X \setminus V = \{x\}^a$.

2. We claim that $\{x\}^a \subseteq G$. If this is not the case, we have $z \notin H$ for some $H \in \mathcal{F}$ and $z \in \{x\}^a$. Because H is open, this entails $\{x\}^a \cap H = \emptyset$; thus by maximality of V, $H \subseteq V$. Since $\mathcal{F}$ is a filter, this implies $V \in \mathcal{F}$, which is not possible. Thus $\{x\}^a \subseteq G$, hence $G \neq \emptyset$, and $X \setminus V \cap G = \emptyset$. Thus $U \cap G = \emptyset$, contradicting the assumption. $\dashv$

This is a fairly surprising and strong statement, because we usually cannot conclude from $\bigcap \mathcal{F} \subseteq U$ that $U \in \mathcal{F}$ holds, when $\mathcal{F}$ is an arbitrary filter. But we work here under stronger assumptions: the underlying space is sober, so each point is given by a morphism for the underlying complete Heyting algebra *and vice versa*. In addition, we deal with Scott open filters. They have the pleasant property of being inaccessible by directed suprema.

But we may even say more, viz., that the intersection of these filters is compact. For, if we have an open cover of the intersection, the union of this cover is open, thus must be an element of the filter by the previous lemma. We may write the union as a union of a directed set of open sets, which then lets us apply the assumption that the filter is inaccessible.

Corollary 3.6.34 *Let X be sober and $\mathcal{F}$ be a Scott open filter. Then $\bigcap \mathcal{F}$ is compact and nonempty.*

Proof Let $K := \bigcap \mathcal{F}$ and $\mathcal{S}$ be an open cover of K. Thus $U := \bigcup \mathcal{S}$ is open with $K \subseteq S$, hence $U \in \mathcal{F}$ by Lemma 3.6.33. But $\bigcup \mathcal{S} = \bigcup \{\bigcup \mathcal{S}_0 \mid \mathcal{S}_0 \subseteq \mathcal{S} \text{ finite}\}$, and the latter collection is directed, so there exists $\mathcal{S}_0 \subseteq \mathcal{S}$ finite with $\bigcup \mathcal{S}_0 \in \mathcal{F}$. But this means $\mathcal{S}_0$ is a finite subcover of K, which consequently is compact. If K is empty, $\emptyset \in \mathcal{F}$ by Lemma 3.6.33, which is impossible. $\dashv$

This gives a complete characterization of the Scott open filters in a sober space. The characterization involves compact sets which are represented as the intersections of these filters. But we can represent only those compact sets C which are upper sets in the specialization order, i.e., for which holds $x \in C$ and $x \sqsubseteq x'$ implies $x' \in C$. These sets are called *saturated*. Recall that $x \sqsubseteq x'$ means $x \in G \Rightarrow x' \in G$ for all open sets G; hence a set is saturated iff it equals the intersection of all open sets containing it. With this in mind, we state the *Hofmann–Mislove Theorem*.

Theorem 3.6.35 *Let X be a sober space. Then the Scott open filters are in one-to-one and order preserving correspondence with the nonempty saturated compact subsets of X via $\mathcal{F} \mapsto \bigcap \mathcal{F}$.*

Proof We have shown in Corollary 3.6.34 that the intersection of a Scott open filter is compact and nonempty; it is saturated by construction. Conversely, Lemma 3.6.31 shows that we may obtain from a compact and saturated subset C of X a Scott open filter, the intersection of which must be C. It is clear that the correspondence is order preserving. $\dashv$

It is quite important for the proof of Lemma 3.6.33 that the underlying space is sober. Hence it does not come as a surprise that the Theorem of Hofmann–Mislove can be used for a characterization of sober spaces as well [GHK$^+$03, Theorem II-1.21].

Proposition 3.6.36 *Let X be a T_0-space. Then the following statements are equivalent:*

1. *X is sober.*
2. *Each Scott open filter $\mathcal{F}$ on τ consists of all open sets containing $\bigcap \mathcal{F}$.*

Proof 1 $\Rightarrow$ 2: This follows from Lemma 3.6.33.

2 $\Rightarrow$ 1: Corollary 3.6.26 tells us that it is sufficient to show that each irreducible closed set is the closure of one point.

Let $A \subseteq X$ be irreducible and closed. Then $\mathcal{F} := \{G \text{ open} \mid G \cap A \neq \emptyset\}$ is closed under finite intersections, since A is irreducible. In fact, let G and H be open sets with $G \cap A \neq \emptyset$ and $H \cap A \neq \emptyset$. If $A \subseteq (X \setminus G) \cup (X \setminus H)$, then A is a subset of one of these closed sets, say, $X \setminus G$, but then $A \cap H = \emptyset$, which is a contradiction. This implies that $\mathcal{F}$ is a filter, and $\mathcal{F}$ is obviously Scott open.

Assume that A cannot be represented as $\{x\}^a$ for some x. Then $X \setminus \{x\}^a$ is an open set the intersection of which with A is not empty; hence $X \setminus \{x\}^a \in \mathcal{F}$. We obtain from the assumption that $X \setminus A \in \mathcal{F}$, because with $K := \bigcap \mathcal{F} \subseteq \bigcap_{x \in X}(X \setminus \{x\}^a)$, we have $K \subseteq X \setminus A$, and $X \setminus A$ is open. Consequently, $A \cap X \setminus A \neq \emptyset$, which is a contradiction.

Thus there exists $x \in X$ such that $A = \{x\}^a$. Hence X is sober. $\dashv$

These are the first and rather elementary discussions of the interplay between topology and order, considered in a systematic fashion in domain theory. The reader is directed to [GHK$^+$03] or to [AJ94] for further information.

3.6.3 The Stone–Weierstraß Theorem

This section will see the classic Stone–Weierstraß Theorem on the approximation of continuous functions on a compact topological space. We need for this a ring of continuous functions, and show that—under suitable conditions—this ring is dense. This requires some preliminary considerations on the space of continuous functions, because this construction evidently requires a topology.

$\mathcal{C}(X)$ Denote for a topological space X by $\mathcal{C}(X)$ the space of all continuous and bounded functions $f : X \to \mathbb{R}$.

The structure of $\mathcal{C}(X)$ is algebraically fairly rich; just for the record:

Proposition 3.6.37 *$\mathcal{C}(X)$ is a real vector space which is closed under constants, multiplication, and under the lattice operations.* $\dashv$

This describes the algebraic properties, but we need a topology on this space, which is provided by the supremum norm. Define for $f \in \mathcal{C}(X)$

$$\|f\| := \sup_{x \in X} |f(x)|.$$

Then $(\mathcal{C}(X), \|\cdot\|)$ is an example for a normed linear (or vector) space.

Definition 3.6.38 *Let V be a real vector space. A* norm $\|\cdot\| : V \to \mathbb{R}_+$ *assigns to each vector v a nonnegative real number $\|v\|$ with these properties:*

1. $\|v\| \geq 0$, *and* $\|v\| = 0$ *iff* $v = 0$.
2. $\|\alpha \cdot v\| = |\alpha| \cdot \|v\|$ *for all* $\alpha \in \mathbb{R}$ *and all* $v \in V$.
3. $\|x + y\| \leq \|x\| + \|y\|$ *for all* $x, y \in V$.

A vector space with a norm is called a normed space.

It is immediate that a normed space is a metric space, putting $d(v, w) := \|v - w\|$. It is also immediate that $f \mapsto \|f\|$ defines a norm on $\mathcal{C}(X)$. But we can say actually a bit more: with this definition of a metric, $\mathcal{C}(X)$ is a complete metric space; we have established this for the compact interval $[0, 1]$ in Example 3.5.19 already. Let us have a look at the general case.

Lemma 3.6.39 *$\mathcal{C}(X)$ is complete with the metric induced by the supremum norm.*

Proof Let $(f_n)_{n\in\mathbb{N}}$ be a $\|\cdot\|$-Cauchy sequence in $\mathcal{C}(X)$, then $(f_n(x))_{n\in\mathbb{N}}$ is bounded, and $f(x) := \lim_{n\to\infty} f_n(x)$ exists for each $x \in X$. Let $\epsilon > 0$ be given, then we find $n_0 \in \mathbb{N}$ such that $\|f_n - f_m\| < \epsilon$ for all $n, m \geq n_0$; thus $|f(x) - f_n(x)| \geq \epsilon$ for $n \geq n_0$. This inequality holds for each $x \in X$, so that we obtain $\|f - f_n\| \leq \epsilon$ for $n \geq n_0$. It implies also that $f \in \mathcal{C}(X)$. $\dashv$

Normed spaces for which the associated metric space is complete are special, so they deserve their own name.

Definition 3.6.40 *A normed space $(V, \| \cdot \|)$ which is complete in the metric associated with $\| \cdot \|$ is called a* Banach space.

The topology induced by the supremum norm is called the *topology of uniform convergence*, so that we may restate Lemma 3.6.39 by saying the $\mathcal{C}(X)$ is closed under uniform convergence. A helpful example is Dini's Theorem for uniform convergence on $\mathcal{C}(X)$ for compact X. It gives a criterion of uniform convergence, provided we know already that the limit is continuous.

Proposition 3.6.41 *Let X be a compact topological space, and assume that $(f_n)_{n\in\mathbb{N}}$ is a sequence of continuous functions which increases monotonically to a continuous function f. Then $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f.*

Proof We know that $f_n(x) \leq f_{n+1}(x)$ holds for all $n \in \mathbb{N}$ and all $x \in X$ and that $f(x) := \sup_{n\in\mathbb{N}} f_n(x)$ is continuous. Let $\epsilon > 0$ be given, then $F_n := \{x \in X \mid f(x) \geq f_n(x) - \epsilon\}$ defines a closed set with $\bigcap_{n\in\mathbb{N}} F_n = \emptyset$; moreover, the sequence $(F_n)_{n\in\mathbb{N}}$ decreases. Thus we find $n_0 \in \mathbb{N}$ with $F_n = \emptyset$ for $n \geq n_0$; hence $\|f - f_n\| < \epsilon$ for $n \geq n_0$. $\dashv$

The goal of this section is to show that, given the compact topological space X, we can approximate each continuous real function uniformly through elements of a subspace of $\mathcal{C}(X)$. It is plain that this subspace has to satisfy some requirements; it should:

- be a vector space itself,
- contain the constant functions,
- separate points,
- be closed under multiplication.

Hence it is in particular a subring of the ring $\mathcal{C}(X)$. Let A be such a subset, then we want to show that the closure A^a with respect to uniform convergence equals $\mathcal{C}(X)$. We will show first that A^a is closed under the lattice operations, because we will represent an approximating function as the finite supremum of a finite infimum of simpler approximations. So the first goal will be to establish closure under inf and sup. Recall that

$$
\begin{aligned}
f \wedge g &= \frac{1}{2} \cdot (f + g - |f - g|),\\
f \vee g &= \frac{1}{2} \cdot (f + g + |f - g|).
\end{aligned}
$$

Now it is easy to see that A^a is closed under the vector space operations, if A is. Our first step boils down to showing that $|f| \in A^a$ if $f \in A$. Thus, given $f \in A$, we have to find a sequence $(f_n)_{n\in\mathbb{N}}$ such that $|f|$ is the uniform limit of this sequence. It is actually enough to show that $t \mapsto \sqrt{t}$ can be approximated uniformly on the unit interval $[0, 1]$, because we know that $|f| = \sqrt{f^2}$ holds. It suffices to do this on the unit interval, as we will see below.

But wait!

Lemma 3.6.42 *There exists a sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{C}([0, 1])$ which converges uniformly to the function $t \mapsto \sqrt{t}$.*

Proof Define inductively for $t \in [0, 1]$

$$
\begin{aligned}
f_0(t) &:= 0,\\
f_{n+1}(t) &:= f_n(t) + \frac{1}{2} \cdot (t - f_n^2(t)).
\end{aligned}
$$

We show by induction that $f_n(t) \leq \sqrt{t}$ holds. This is clear for $n = 0$. If we know already that the assumption holds for n, then we write

$$
\begin{aligned}
\sqrt{t} - f_{n+1}(t) &= \sqrt{t} - f_n(t) - \frac{1}{2} \cdot (t - f_n^2(t))\\
&= (\sqrt{t} - f_n(t)) \cdot \big((1 - \frac{1}{2} \cdot (\sqrt{t} + f_n(t))\big).
\end{aligned}
$$

Because $t \in [0, 1]$ and from the induction hypothesis, we have $\sqrt{t} + f_n(t) \leq 2 \cdot \sqrt{t} \leq 2$, so that $\sqrt{t} - f_{n+1}(t) \geq 0$.

Thus we infer that $f_n(t) \leq \sqrt{t}$ for all $t \in [0, 1]$, and $\lim_{n\to\infty} f_n(t) = \sqrt{t}$. From Dini's Proposition 3.6.41, we now infer that the convergence is uniform. $\dashv$

This is the desired consequence from this construction.

Corollary 3.6.43 *Let X be a compact topological space, and let $A \subseteq \mathcal{C}(X)$ be a ring of continuous functions which contains the constants and which is closed under uniform convergence. Then A is a lattice.*

Proof It is enough to show that A is closed under taking absolute values. Let $f \in A$, then we may and do assume that $0 \leq f \leq 1$ holds (otherwise consider $(f - \|f\|)/\|f\|$, which is an element of A as well). Because $|f| = \sqrt{f^2}$, and the latter is a uniform limit of elements of A by Lemma 3.6.42, we conclude $|f| \in A$, which entails A being closed under the lattice operations. $\dashv$

We are now in a position to establish the classic Stone–Weierstraß Theorem, which permits to conclude that a ring of bounded continuous functions on a compact topological space X is dense with respect to uniform convergence in $\mathcal{C}(X)$, provided it contains the constants and separates points. The latter condition is obviously necessary, but has not been used in the argumentation so far. It is clear, however, that we cannot do without this condition, because $\mathcal{C}(X)$ separates points, and it is difficult to see how a function which separates points could be approximated from a collection which does not.

The polynomials on a compact interval in the reals are an example for a ring which satisfies all these assumptions. This collection shows also that we cannot extend the result to a non-compact base space like the reals. Take $x \mapsto \sin x$, for example; this function cannot be approximated uniformly over $\mathbb{R}$ by polynomials. For, assume that given $\epsilon > 0$ there exists a polynomial p such that $\sup_{x\in\mathbb{R}} |p(x) - \sin x| < \epsilon$, then we would have $-\epsilon - 1 < p(x) < 1 + \epsilon$ for all $x \in \mathbb{R}$, which is impossible, because a polynomial is unbounded.

Here, then, is the Stone–Weierstraß Theorem for compact topological spaces.

Theorem 3.6.44 *Let X be a compact topological space and $A \subseteq \mathcal{C}(X)$ be a ring of functions which separates points and which contains all constant functions. Then A is dense in $\mathcal{C}(X)$.*

Proof 0. Our goal is to find for some given $f \in \mathcal{C}(X)$ and an arbitrary $\epsilon > 0$ a function $\varphi \in A^a$ such that $\|f - \varphi\| < \epsilon$. Since X is compact, we will find φ through a refined covering argument in the following way. If $a, b \in X$ are given, we find a continuous function $f_{a,b} \in A$ with $f_{a,b} = f(a)$ and $f_{a,b} = f(b)$. From this we construct a cover; using sets like $\{x \mid f_{a,b}(x) < f(x) + \epsilon\}$ and $\{x \mid f_{a,b}(x) > f(x) - \epsilon\}$, extract finite subcovers and construct from the corresponding functions the desired function through suitable lattice operations. Approach

1. Fix $f \in \mathcal{C}(X)$ and $\epsilon > 0$. Given distinct point $a \neq b$, we find a function $h \in A$ with $h(a) \neq h(b)$; thus

$$g(x) := \frac{h(x) - h(a)}{h(b) - h(a)}$$

defines a function $g \in A$ with $g(a) = 0$ and $g(b) = 1$. Then

$$f_{a,b}(x) := (f(b) - f(a)) \cdot g(x) + f(a)$$

is also an element of A with $f_{a,b}(a) = f(a)$ and $f_{a,b}(b) = f(b)$. Now define

$$\begin{aligned} U_{a,b} &:= \{x \in X \mid f_{a,b}(x) < f(x) + \epsilon\}, \\ V_{a,b} &:= \{x \in X \mid f_{a,b}(x) > f(x) - \epsilon\}; \end{aligned}$$

then $U_{a,b}$ and $V_{a,b}$ are open sets containing a and b.

2. Fix b, then $\{U_{a,b} \mid a \in X\}$ is an open cover of X, so we can find points $a_1, \ldots, a_k$ such that $\{U_{a_1,b}, \ldots, U_{a_k,b}\}$ is an open cover of X by compactness. Thus

$$f_b := \bigwedge_{i=1}^{k} f_{a_i,b}$$

defines an element of A^a by Corollary 3.6.43. We have $f_b(x) < f(x) + \epsilon$ for all $x \in X$, and we know that $f_b(x) > f(x) - \epsilon$ for all $x \in V_b := \bigcap_{i=1}^{k} V_{a_i,b}$. The set V_b is an open neighborhood of b, so from the open cover $\{V_b \mid b \in X\}$, we find $b_1, \ldots, b_\ell$ such that X is covered through $\{V_{b_1}, \ldots, V_{b_\ell}\}$. Put

$$\varphi := \bigvee_{i=1}^{\ell} f_{b_i},$$

then $f_\epsilon \in A^a$ and $\|f - \varphi\| < \epsilon$. $\dashv$

This is the example already discussed above.

Example 3.6.45 Let $X := [0, 1]$ be the closed unit interval, and let A consist of all polynomials $\sum_{i=0}^{n} a_i \cdot x^i$ for $n \in \mathbb{N}$ and $a_0, \ldots, a_n \in \mathbb{R}$. Polynomials are continuous, they form a vector space and are closed under multiplication. Moreover, the constants are polynomials. Thus we obtain from the Stone–Weierstraß Theorem 3.6.44 that every continuous function on $[0, 1]$ can be uniformly approximated through a sequence of polynomials. ✌

The polynomials discussed in Example 3.6.45 are given algebraically as the ring generated by the functions $\{1, id_{[0,1]}\}$ in the ring of all continuous functions over $[0, 1]$. In general, given a subset $A \subseteq \mathcal{C}(X)$ which is closed under multiplication, the smallest ring $R(A)$ containing A is

$$R(A) = \left\{\sum_{i=1}^{n} a_i \cdot f_i \mid n \in \mathbb{N}, a_1, \ldots, a_n \in \mathbb{R}, f_1, \ldots, f_n \in A\right\}.$$

This is so because the set above is a ring which must be contained in every ring containing A, and it is a ring itself. We obtain from this consideration:

Corollary 3.6.46 *Let (X, d) be compact metric, then $(\mathcal{C}(X), \|\cdot\|)$ is a separable Banach space.*

Proof Let $\mathcal{G}$ be a countable basis for the topology of X, then the countable set $A_0 := \{d(\cdot, X \setminus G) \mid G \in \mathcal{G}\}$ of continuous functions separates points; hence $R(A_0 \cup \mathbb{R})$ is dense in $\mathcal{C}(X)$ by the Stone–Weierstraß Theorem 3.6.44. Given $f = \sum_{i=1}^{n} a_i \cdot f_i$ and $\epsilon > 0$, there exists $f' = \sum_{i=1}^{n} a_i' \cdot f_i$ with $a_i' \in \mathbb{Q}$ for $1 \leq i \leq n$ and $\|f - f'\| \leq \epsilon$, so that the elements from $R(A)$ with rational coefficients form a countable dense subset as well. $\dashv$

It is said that Oscar Wilde could resist everything but a good temptation. The author concurs. Here is a classic proof of the Weierstraß Approximation Theorem, the original form of Theorem 3.6.44, which deals with polynomials on $[0, 1]$ only, and establishes the statement given in Example 3.6.45. We will give this proof now, based on the discussion in the classic [CH67, §II.4.1]. This proof is elegant and based on the manipulation of specific functions (we are all too often focussed on our pretty little garden of beautiful abstract structures, all too often in danger of loosing the contact to concrete mathematics and to our roots). Oscar Wilde

As a preliminary consideration, we will show that

$$\lim_{n\to\infty} \frac{\int_\delta^1 (1 - v^2)^n \, dv}{\int_0^1 (1 - v^2)^n \, dv} = 0$$

for every $\delta \in \ ]0, 1[$. Define for this

$$J_n := \int_0^1 (1 - v^2)^n \, dv,$$

$$J_n^* := \int_\delta^1 (1 - v^2)^n \, dv.$$

(we will keep these notations for later use). We have

$$J_n > \int_0^1 (1-v)^n \, dv = \frac{1}{n+1}$$

and

$$J_n^* = \int_\delta^1 (1-v^2)^n \, dv < (1-\delta^2)^n \cdot (1-\delta) < (1-\delta^2)^n.$$

Thus

$$\frac{J_n^*}{J_n} < (n+1) \cdot (1-\delta^2)^n \to 0.$$

This establishes the claim.

Let $f : [0,1] \to \mathbb{R}$ be continuous. Given $\epsilon > 0$, there exists $\delta > 0$ such that $|x-y| < \delta$ implies $|f(x) - f(y)| < \epsilon$ for all $x \in [0,1]$, since f is uniformly continuous by Proposition 3.5.36. Thus $0 \le v < \delta$ implies $|f(x+v) - f(x)| < \epsilon$ for all $x \in [0,1]$.

Put

$$Q_n(x) := \int_0^1 f(u) \cdot \left(1 - (u-x)^2\right)^n \, du,$$
$$P_n(x) := \frac{Q_n(x)}{2 \cdot J_n}.$$

We will show that P_n converges to f in the topology of uniform convergence.

We note first that Q_n is a polynomial of degree $2n$. In fact, put

$$A_j := \int_0^1 f(u) \cdot u^j \, du$$

for $j \ge 0$; expanding yields the formidable representation

$$Q_n(x) = \sum_{k=0}^{n} \sum_{j=0}^{2k} \binom{n}{k} \binom{2k}{j} (-1)^{n-k+j} A_j \cdot x^{2k-j}.$$

Let us work on the approximation. We fix $x \in [0,1]$, and note that the inequalities derived below do not depend on the specific choice of x. Hence they will provide a uniform approximation.

Substitute u by $v + x$ in Q_n; this yields

$$\begin{aligned}\int_0^1 f(u)\big(1-(u-x)^2\big)^n\,du &= \int_x^{1-x} f(v+x)(1-v^2)^n\,dv\\ &= I_1 + I_2 + I_3\end{aligned}$$

with

$$\begin{aligned}I_1 &:= \int_{-x}^{-\delta} f(v+x)(1-v^2)^n\,dv,\\ I_2 &:= \int_{-\delta}^{+\delta} f(v+x)(1-v^2)^n\,dv,\\ I_3 &:= \int_{+\delta}^{1-x} f(v+x)(1-v^2)^n\,dv.\end{aligned}$$

We work on these integrals separately. Let $M := \max_{0\le x\le 1}|f(x)|$, then

$$I_1 \le M\int_{-1}^{-\delta}(1-v^2)^n\,dv = M\cdot J_n^*,$$

and

$$I_3 \le M\int_{\delta}^{1}(1-v^2)^n\,dv = M\cdot J_n^*.$$

We can rewrite I_2 as follows:

$$\begin{aligned}I_2 &= f(x)\int_{-\delta}^{+\delta}(1-v^2)^n\,dv + \int_{-\delta}^{+\delta}\big(f(x+v)-f(x)\big)(1-v^2)^n\,dv\\ &= 2f(x)(J_n - J_n^*) + \int_{-\delta}^{+\delta}\big(f(x+v)-f(x)\big)(1-v^2)^n\,dv.\end{aligned}$$

From the choice of δ for ϵ, we obtain

$$\begin{aligned}\Big|\int_{-\delta}^{+\delta}\big(f(x+v)-f(x)\big)(1-v^2)^n\,dv\Big| &\le \epsilon\int_{-\delta}^{+\delta}(1-v^2)^n\,dv\\ &< \epsilon\int_{-1}^{+1}(1-v^2)^n\,dv\\ &= 2\epsilon\cdot J_n\end{aligned}$$

Combining these inequalities, we obtain

$$|P_n(x)-f(x)| < 2M\cdot\frac{J_n^*}{J_n}+\epsilon.$$

Hence the difference can be made arbitrarily small, which means that f can be approximated uniformly through polynomials.

The two approaches presented are structurally very different; it would be difficult to recognize the latter as a precursor of the former. While both make substantial use of uniform continuity, the first one is an existential proof, constructing two covers from which to choose a finite subcover each and deriving from this the existence of an approximating function. It is nonconstructive because it would be difficult to construct an approximating function from it, even if the ring of approximating functions is given by a base for the underlying vector space. The second one, however, starts also from uniform continuity and uses this property to find a suitable bound for the difference of the approximating polynomial and the function proper through integration. The representation of Q_n above shows what the constructing polynomial looks like, and the coefficients of the polynomials may be computed (in principle, at least). And, finally, the abstract situation gives us a greater degree of freedom, since we deal with a ring of continuous functions observing certain properties, while the original proof works for the class of polynomials only.

3.6.4 Uniform Spaces

This section will give a brief introduction to uniform spaces. The objective is to demonstrate that the notion of a metric space can be generalized in meaningful ways without arriving at the full generality of topological spaces but retaining useful properties like completeness or uniform continuity. While pseudometric spaces formulate the concept of two points to be close to each other through a numeric value, and general topological spaces use the concept of an open neighborhood, uniform spaces formulate neighborhoods on the Cartesian product. This concept is truly in the middle: each pseudometric generates neighborhoods, and from a neighborhood, we may obtain the neighborhood filter for a point.

For motivation and illustration, we consider a pseudometric space (X, d) and say that two points are neighbors iff their distance is smaller that r for some fixed $r > 0$; the degree of neighborhood is evidently depending on r. The set

$$V_{d,r} := V_r := \{\langle x, y\rangle \mid d(x, y) < r\}$$

is then the collection of all neighbors. We may obtain from V_r the neighborhood $B(x, r)$ for some point x upon extracting all y such that $\langle x, y\rangle \in V_r$; thus $V_{d,r}; V_r$

$$B(x, r) = V_r[x] := \{y \in X \mid \langle x, y\rangle \in V_r\}.$$

The collection of all these neighborhoods observes these properties:

1. The diagonal $\Delta := \Delta_X$ is contained in V_r for all $r > 0$, because $d(x, x) = 0$.
2. V_r is—as a relation on X—symmetric: $\langle x, y\rangle \in V_r$ iff $\langle y, x\rangle \in V_r$; thus $V_r^{-1} = V_r$. This property reflects the symmetry of d.
3. $V_r \circ V_s \subseteq V_{r+s}$ for $r, s > 0$; this property is inherited from the triangle inequality for d.
4. $V_{r_1} \cap V_{r_2} = V_{\min\{r_1, r_2\}}$; hence this collection is closed under finite intersections.

It is convenient to consider not only these immediate neighborhoods but rather the filter generated by them on $X \times X$ (which is possible because the empty set is not contained in this collection, and the properties above shows that they form the base for a filter indeed). This leads to this definition of a uniformity. It focusses on the properties of the neighborhoods rather than on that of a pseudometric, so we formulate it for a set in general.

Definition 3.6.47 *Let X be a set. A filter $\mathfrak{u}$ on $\mathcal{P}(X \times X)$ is called a* uniformity on X *iff these properties are satisfied:*

1. *$\Delta \subseteq U$ for all $U \in \mathfrak{u}$.*
2. *If $U \in \mathfrak{u}$, then $U^{-1} \in \mathfrak{u}$.*
3. *If $U \in \mathfrak{u}$, there exists $V \in \mathfrak{u}$ such that $V \circ V \subseteq U$.*
4. *$\mathfrak{u}$ is closed under finite intersections.*
5. *If $U \in \mathfrak{u}$ and $U \subseteq W$, then $W \in \mathfrak{u}$.*

The pair $(X, \mathfrak{u})$ is called a uniform space. *The elements of $\mathfrak{u}$ are called* $\mathfrak{u}$-neighborhoods.

The first three properties are gleaned from those of the pseudometric neighborhoods above, the last two are properties of a filter, which have been listed here just for completeness.

$$
\begin{aligned}
U \circ V &:= \{\langle x, z\rangle \mid \exists y : \langle x, y\rangle \in U, \langle y, z\rangle \in V\}\\
U^{-1} &:= \{\langle y, x\rangle \mid \langle x, y\rangle \in U\}\\
U[M] &:= \{y \mid \exists x \in M : \langle x, y\rangle \in U\}\\
U[x] &:= U[\{x\}]\\
U \text{ is symmetric } &:\Leftrightarrow U^{-1} = U\\
(U \circ V) \circ W &= U \circ (V \circ W)\\
(U \circ V)^{-1} &= V^{-1} \circ U^{-1}\\
(U \circ V)[M] &= U[V[M]]\\
V \circ U \circ V &= \bigcup_{\langle x,y\rangle \in U} V[x] \times V[y] \; (V \text{ symmetric})
\end{aligned}
$$

Here $U, V, W \subseteq X \times X$ and $M \subseteq X$.

Figure 3.1: Some relational identities

Neigh-borhood, entourage, Nach-barschaft

We will omit $\mathfrak{u}$ when talking about a uniform space, if this does not yield ambiguities. The term "neighborhood" is used for elements of a uniformity and for the neighborhoods of a point. There should be no ambiguity, because the point is always attached, when talking about neighborhood in the latter, topological sense. Bourbaki uses the term *entourage* for a neighborhood in the uniform sense; the German word for this is *Nachbarschaft* (while the term for a neighborhood of a point is *Umgebung*).

We will need some relational identities; they are listed in Fig. 3.1 for the reader's convenience.

We will proceed here as we do in the case of topologies, where we do not always specify the entire topology, but occasionally make use of the possibility to define it through a base. We have this characterization for the base of a uniformity.

Base

Proposition 3.6.48 *A family $\emptyset \neq \mathfrak{b} \subseteq \mathcal{P}(X \times X)$ is the base for a uniformity iff it has the following properties:*

1. *Each member of $\mathfrak{b}$ contains the diagonal of X.*
2. *For $U \in \mathfrak{b}$, there exists $V \in \mathfrak{b}$ with $V \subseteq U^{-1}$.*
3. *For $U \in \mathfrak{b}$, there exists $V \in \mathfrak{b}$ with $V \circ V \subseteq U$.*
4. *For $U, V \in \mathfrak{b}$, there exists $W \in \mathfrak{b}$ with $W \subseteq U \cap V$.*

Proof Recall that the filter generated by a filter base $\mathfrak{b}$ is defined through $\{F \mid U \subseteq F \text{ for some } U \in \mathfrak{b}\}$. With this in mind, the proof is straightforward. $\dashv$

This permits a description of a uniformity in terms of a base, which is usually easier than giving a uniformity as a whole. Let us look at some examples.

Example 3.6.49

1. The uniformity $\{\Delta, X \times X\}$ is called the *indiscrete uniformity*, and the uniformity $\{A \subseteq X \times X \mid \Delta \subseteq A\}$ is called the *discrete uniformity* on X.

2. Let $V_r := \{\langle x, y\rangle \mid x, y \in \mathbb{R}, |x - y| < r\}$, then $\{V_r \mid r > 0\}$ is a base for a uniformity on $\mathbb{R}$. Since it makes use of the structure of $(\mathbb{R}, +)$ as an additive group, it is called the *additive uniformity* on $\mathbb{R}$.

3. Put $V_E := \{\langle x, y\rangle \in \mathbb{R}^2 \mid x/y \in E\}$ for some neighborhood E of $1 \in \mathbb{R} \setminus \{0\}$. Then the filter generated by $\{V_E \mid E \text{ is a neighborhood of } 1\}$ is a uniformity. This is so because the logarithm function is continuous on $\mathbb{R}_+ \setminus \{0\}$. This uniformity is nourished from the multiplicative group $(\mathbb{R} \setminus \{0\}, \cdot)$, so it is called the *multiplicative uniformity* on $\mathbb{R} \setminus \{0\}$. This is discussed in greater generality in part 9.

4. A *partition* π on a set X is a collection of nonempty and mutually disjoint subsets of X which covers X. It generates an equivalence relation on X by rendering two elements of X equivalent iff they are in the same partition element. Define $V_\pi := \bigcup_{i=1}^n (P_i \times P_i)$ for a finite partition $\pi = \{P_1, \dots, P_k\}$. Then
$$\mathfrak{b} := \{V_\pi \mid \pi \text{ is a finite partition on } X\}$$
is the base for a uniformity. Let π be a finite partition, and denote the equivalence relation generated by π by $|\pi|$; hence $x \; |\pi| \; y$ iff x and y are in the same element of π.
   - $\Delta \subseteq V_\pi$ is obvious, since $|\pi|$ is reflexive.
   - $U^{-1} = U$ for all $U \in V_\pi$, since $|\pi|$ is symmetric.
   - Because $|\pi|$ is transitive, we have $V_\pi \circ V_\pi \subseteq V_\pi$.

- Let π' be another finite partition, then $\{A \cap B \mid A \in \pi, B \in \pi', A \cap B \neq \emptyset\}$ defines a partition π'' such that $V_{\pi''} \subseteq V_\pi \cap V_{\pi'}$.

Thus $\mathfrak{b}$ is the base for a uniformity, which is, you guessed it, called the *uniformity of finite partitions*.

5. Let $\emptyset \neq \mathcal{I} \subseteq \mathcal{P}(X)$ be an ideal (Definition 1.5.31), and define

$$\mathcal{A}_E := \{\langle A, B\rangle \mid A \Delta B \in E\} \text{ for } E \in \mathcal{I},$$
$$\mathfrak{b} := \{\mathcal{A}_E \mid E \in \mathcal{I}\}.$$

Then $\mathfrak{b}$ is a base for a uniformity on $\mathcal{P}(X)$. In fact, it is clear that $\Delta_{\mathcal{P}(X)} \subseteq \mathcal{A}_E$ always holds and that each member of $\mathfrak{b}$ is symmetric. Let $A \Delta B \subseteq E$ and $B \Delta C \subseteq F$, then $A \Delta C = (A \Delta B) \Delta (B \Delta C) \subseteq (A \Delta B) \cup (A \Delta C) \subseteq E \cup F$; thus $\mathcal{A}_E \circ \mathcal{A}_F \subseteq \mathcal{A}_{E \cup F}$, and finally $\mathcal{A}_E \cap \mathcal{A}_F \subseteq \mathcal{A}_{E \cap F}$. Because $\mathcal{I}$ is an ideal, it is closed under finite intersections and finite unions; the assertion follows.

6. Let p be a prime, and put $W_k := \{\langle x, y\rangle \mid x, y \in \mathbb{Z}, p^k \text{ divides } x - y\}$. Then $W_k \circ W_\ell \subseteq W_{\min\{k,\ell\}} = W_k \cap W_\ell$; thus $\mathfrak{b} := \{W_k \mid k \in \mathbb{N}\}$ is the base for a uniformity $\mathfrak{u}_p$ on $\mathbb{Z}$, the *p-adic uniformity*.

7. Let A be a set, $(X, \mathfrak{u})$ a uniform space, and let $F(A, X)$ be the set of all maps $A \to X$. We will define a uniformity on $F(A, X)$; the approach is similar to the one in Example 3.1.4. Define for $U \in \mathfrak{u}$ the set

$$U_F := \{\langle f, g\rangle \in F(A, X) \mid \langle f(x), g(x)\rangle \in U \text{ for all } x \in X\}.$$

Thus two maps are close with respect to U_F iff all their images are close with respect to U. It is immediate that $\{U_F \mid U \in \mathfrak{u}\}$ forms a uniformity on $F(A, X)$ and that $\{U_F \mid U \in \mathfrak{b}\}$ is a base for a uniformity, provided $\mathfrak{b}$ is a base for uniformity $\mathfrak{u}$.

If $X = \mathbb{R}$ is endowed with the additive uniformity, a typical set of the base is given for $\epsilon > 0$ through

$$\{\langle f, g\rangle \in F(A, \mathbb{R}) \mid \sup_{a \in A} |f(a) - g(a)| < \epsilon\};$$

hence the images of f and of g have to be uniformly close to each other.

8. Call a map $f : \mathbb{R} \to \mathbb{R}$ *affine* iff it can be written as $f(x) = a \cdot x + b$ with $a \neq 0$; let $f_{a,b}$ be the affine map characterized by the parameters a and b, and define $X := \{f_{a,b} \mid a, b \in \mathbb{R}, a \neq 0\}$ the set of all affine maps. Note that an affine map is bijective and that its inverse is an affine map again with $f_{a,b}^{-1} = f_{1/a,-b/a}$; the composition of an affine map is an affine map as well, since $f_{a,b} \circ f_{c,d} = f_{ac,ad+b}$. Define for $\epsilon > 0, \delta > 0$ the ϵ, δ-neighborhood $U_{\epsilon,\delta}$ by

$$U_{\epsilon,\delta} := \{f_{a,b} \in X \mid |a-1| < \epsilon, |b| < \delta\}.$$

Put

$$\begin{aligned} U^L_{\epsilon,\delta} &:= \{\langle f_{x,y}, f_{a,b}\rangle \in X \times X \mid f_{x,y} \circ f_{a,b}^{-1} \in U_{\epsilon,\delta}\}, \\ \mathfrak{b}_L &:= \{U^L_{\epsilon,\delta} \mid \epsilon > 0, \delta > 0\}, \\ U^R_{\epsilon,\delta} &:= \{\langle f_{x,y}, f_{a,b}\rangle \in X \times X \mid f_{x,y}^{-1} \circ f_{a,b} \in U_{\epsilon,\delta}\}, \\ \mathfrak{b}_R &:= \{U^R_{\epsilon,\delta} \mid \epsilon > 0, \delta > 0\}. \end{aligned}$$

Then $\mathfrak{b}_L$ resp. $\mathfrak{b}_R$ is the base for a uniformity $\mathfrak{u}_L$ resp. $\mathfrak{u}_R$ on X. Let us check this for $\mathfrak{b}_R$. Given positive ϵ, δ, we want to find positive r, s with $\langle f_{m,n}, f_{p,q}\rangle \in V^R_{r,s}$ implies $\langle f_{p,q}, f_{m,n}\rangle \in U^R_{\epsilon,\delta}$. Now we can find for $\epsilon > 0$ and $\delta > 0$ some $r > 0$ and $s > 0$ so that

$$\begin{aligned} |\frac{p}{m} - 1| < r &\Rightarrow |\frac{m}{p} - 1| < \epsilon \\ |\frac{q}{m} - \frac{n}{m}| < s &\Rightarrow |\frac{n}{p} - \frac{q}{p}| < \delta \end{aligned}$$

holds, which is just what we want, since it translates into $V^R_{r,s} \subseteq \left(U^R_{\epsilon,\delta}\right)^{-1}$. The other properties of a base are easily seen to be satisfied. One argues similarly for $\mathfrak{b}_L$.

Note that $(X, \circ)$ is a topological group with the sets $\{U_{\epsilon,\delta} \mid \epsilon > 0, \delta > 0\}$ as a base for the neighborhood filter of the neutral element $f_{1,0}$ (topological groups are introduced in Example 3.1.25 on page 299).

9. Let, in general, G be a topological group with neutral element e. Define for $U \in \mathfrak{U}(e)$ the sets

$$
\begin{aligned}
U_L &:= \{\langle x, y\rangle \mid xy^{-1} \in U\},\\
U_R &:= \{\langle x, y\rangle \mid x^{-1}y \in U\},\\
U_B &:= U_L \cap U_R.
\end{aligned}
$$

Then $\{U_L \mid U \in \mathfrak{U}(e)\}$, $\{U_R \mid U \in \mathfrak{U}(e)\}$ and $\{U_B \mid U \in \mathfrak{U}(e)\}$ define bases for uniformities on G; it can be shown that they do not necessarily coincide; see Example 3.6.74.

✌

Before we show that a uniformity generates a topology, we derive a sufficient criterion for a family of subsets of $X \times X$ is a subbase for a uniformity.

Subbase

Lemma 3.6.50 *Let $\mathfrak{s} \subseteq \mathcal{P}(X \times X)$, then $\mathfrak{s}$ is the subbase for a uniformity on X, provided the following conditions hold:*

1. *$\Delta \subseteq S$ for each $S \in \mathfrak{s}$.*
2. *Given $U \in \mathfrak{s}$, there exists $V \in \mathfrak{s}$ such that $V \subseteq U^{-1}$.*
3. *For each $U \in \mathfrak{s}$, there exists $V \in \mathfrak{s}$ such that $V \circ V \subseteq U$.*

Proof We have to show that

$$\mathfrak{b} := \{U_1 \cap \ldots \cap U_n \mid U_1, \ldots, U_n \in \mathfrak{s} \text{ for some } n \in \mathbb{N}\}$$

constitutes a base for a uniformity. It is clear that every element of $\mathfrak{b}$ contains the diagonal. Let $U = \bigcap_{i=1}^n U_i \in \mathfrak{b}$ with $U_i \in \mathfrak{s}$ for $i = 1, \ldots, n$, choose $V_i \in \mathfrak{s}$ with $V_i \subseteq U_i^{-1}$ for all i, then $V := \bigcap_{i=1}^n V_i \in \mathfrak{b}$ and $V \subseteq U^{-1}$. If we select $W_i \in \mathfrak{s}$ with $W_i \circ W_i \subseteq U_i$, then $W := \bigcap_{i=1}^n W_i \in \mathfrak{b}$ and $W \circ W \subseteq U$. The last condition of Proposition 3.6.48 is trivially satisfied for $\mathfrak{b}$, since $\mathfrak{b}$ is closed under finite intersections. Thus we conclude that $\mathfrak{b}$ is a base for a uniformity on X by Proposition 3.6.48, which in turn entails that $\mathfrak{s}$ is a subbase. $\dashv$

The Topology Generated by a Uniformity

A pseudometric space (X, d) generates a topology by declaring a set G open iff there exists for $x \in G$ some $r > 0$ with $B(x, r) \subseteq G$; from this we obtain the neighborhood filter $\mathfrak{U}(x)$ for a point x. Note that in the uniformity associated with the pseudometric, the identity

$$B(x, r) = V_r[x]$$

holds. Encouraged by this, we approach the topology for a uniform space in the same way. Given a uniform space $(X, \mathfrak{u})$, a subset $G \subseteq X$ is called open iff we can find for each $x \in G$ some neighborhood $U \in \mathfrak{u}$ such that $U[x] \subseteq G$. The following proposition investigates this construction.

Proposition 3.6.51 *Given a uniform space $(X, \mathfrak{u})$, for each $x \in X$, the family*

From $\mathfrak{u}$ to $\tau_{\mathfrak{u}}$

$$\mathfrak{u}[x] := \{U[x] \mid U \in \mathfrak{u}\}$$

is the base for the neighborhood filter of x for a topology $\tau_{\mathfrak{u}}$, which is called the uniform topology. *The neighborhoods for x in $\tau_{\mathfrak{u}}$ are just $\mathfrak{u}[x]$.*

Proof It follows from Proposition 3.1.22 that $\mathfrak{u}[x]$ defines a topology $\tau_{\mathfrak{u}}$; it remains to show that the neighborhoods of this topology are just $\mathfrak{u}[x]$. We have to show that $U \in \mathfrak{u}$ there exists $V \in \mathfrak{u}$ with $V[x] \subseteq U[x]$ and $V[x] \in \mathfrak{u}[y]$ for all $y \in V[x]$, then the assertion will follow from Corollary 3.1.23. For $U \in \mathfrak{u}$, there exists $V \in \mathfrak{u}$ with $V \circ V \subseteq U$; thus $\langle x, y\rangle \in V$ and $\langle y, z\rangle \in V$ imply $\langle x, z\rangle \in U$. Now let $y \in V[x]$ and $z \in V[y]$; thus $z \in U[x]$, but this means $U[x] \in \mathfrak{u}[y]$ for all $x \in V[y]$. Hence the assertion follows. ⊣

Here are some illustrative example. They indicate also that different uniformities can generate the same topology.

Example 3.6.52

1. The topology obtained from the additive uniformity on $\mathbb{R}$ is the usual topology. The same holds for the multiplicative uniformity on $\mathbb{R} \setminus \{0\}$.
2. The topology induced by the discrete uniformity is the discrete topology, in which each singleton $\{x\}$ is open. Since $\big\{\{x\}, X \setminus \{x\}\big\}$ forms a finite partition of X, the discrete topology is induced also by the uniformity defined by the finite partitions.
3. Let $F(A, \mathbb{R})$ be endowed with the uniformity defined by the sets $\{\langle f, g\rangle \in F(A, \mathbb{R}) \mid \sup_{a \in A} |f(a) - g(a)| < \epsilon\}$; see Example 3.6.49. The corresponding topology yields for each $f \in$

$F(A, \mathbb{R})$ the neighborhood $\{g \in F(A, \mathbb{R}) \mid \sup_{a \in A} |f(a) - g(a)| < \epsilon\}$. This is the topology of uniform convergence.

4. Let $\mathfrak{u}_p$ for a prime p be the p-adic uniformity on $\mathbb{Z}$; see Example 3.6.49, part 6. The corresponding topology τ_p is called the p-adic topology. A basis for the neighborhoods of 0 is given by the sets $V_k := \{x \in \mathbb{Z} \mid p^k \text{ divides } x\}$. Because $p^m \in V_k$ for $m \geq k$, we see that $\lim_{n\to\infty} p^n = 0$ in τ_p, but not in τ_q for $q \neq p, q$ prime. Thus the topologies τ_p and τ_q differ, hence also the uniformities $\mathfrak{u}_p$ and $\mathfrak{u}_q$.

✌

Now that we know that each uniformity yields a topology on the same space, some questions are immediate:

- Do the open resp. the closed sets play a particular role in describing the uniformity?
- Does the topology have particular properties, e.g., in terms of separation axioms?
- What about metric spaces—can we determine from the uniformity that the topology is metrizable?
- Can we find a pseudometric for a given uniformity?
- Is the product topology on $X \times X$ somehow related to $\mathfrak{u}$, which is defined on $X \times X$, after all?

We will give answers to some of these questions; some will be treated only lightly, with an in depth treatment to be found in the ample literature on uniform spaces; see the Bibliographic Notes in Sect. 3.7.

Fix a uniform space X with uniformity $\mathfrak{u}$ and associated topology τ. References to neighborhoods and open sets are always to $\mathfrak{u}$ resp. τ, unless otherwise stated.

This is a first characterization of the interior of an arbitrary set. Recall that in a pseudometric space x is an interior point of A iff $B(x, r) \subseteq A$ for some $r > 0$; the same description applies here as well, *mutatis mutandis* (of course, this "mutatis mutandis" part is the interesting one).

Lemma 3.6.53 *Given $A \subseteq X$, $x \in A^o$ iff there exists a neighborhood U with $U[x] \subseteq A$.*

Proof Assume that $x \in A^o = \bigcup\{G \mid G \text{ open and } G \subseteq A\}$; then it follows from the definition of an open set that we must be able to find an neighborhood U with $U[x] \subseteq A$.

Conversely, we show that the set $B := \{x \in X \mid U[x] \subseteq A \text{ for some neighborhood } U\}$ is open, then this must be the largest open set which is contained in A, hence $B = A^o$. Let $x \in B$, thus $U[x] \subseteq A$, and we should find now a neighborhood V such that $V[y] \subseteq B$ for $y \in V[x]$. But we find a neighborhood V with $V \circ V \subseteq U$. Let us see whether V is suitable: if $y \in V[x]$, then $V[y] \subseteq (V \circ V)[x]$ (this is so because $\langle x, y\rangle \in V$, and if $z \in V[y]$, then $\langle y, z\rangle \in V$; this implies $\langle x, z\rangle \in V \circ V$, hence $z \in (V \circ V)[x]$). But this yields $V[y] \subseteq U[x] \subseteq B$, hence $y \in B$. This means $V[x] \subseteq B$, so that B is open. ⊣

The observation above gives us a handy way of describing the base for a neighborhood filter for a point in X. It states that we may restrict our attention to the members of a base or of a subbase, when we want to work with the neighborhood filter for a particular element.

Corollary 3.6.54 *If $\mathfrak{u}$ has base or subbase $\mathfrak{b}$, then $\{U[x] \mid U \in \mathfrak{b}\}$ is a base resp. subbase for the neighborhood filter for x.*

Proof This follows immediately from Lemma 3.6.53 together with Proposition 3.6.48 resp. Lemma 3.6.50 ⊣

Let us have a look at the topology on $X \times X$ induced by τ. Since the open rectangles generate this topology, and since we can describe the open rectangles in terms of the sets $U[x] \times V[y]$, we can expect that these open sets can also be related to the uniformity proper. In fact:

Proposition 3.6.55 *If $U \in \mathfrak{u}$; then both $U^o \in \mathfrak{u}$ and $U^a \in \mathfrak{u}$.*

Proof 1. Let $G \subseteq X \times X$ be open, then $\langle x, y\rangle \in G$ iff there exist neighborhoods $U, V \in \mathfrak{u}$ with $U[x] \times V[y] \subseteq G$, and because $U \cap V \in \mathfrak{u}$, we may even find some $W \in \mathfrak{u}$ such that $W[x] \times W[y] \subseteq G$. Thus

$$G = \bigcup\{W[x] \times W[y] \mid \langle x, y\rangle \in G, W \in \mathfrak{u}\}.$$

2. Let $W \in \mathfrak{u}$; then there exists a symmetric $V \in \mathfrak{u}$ with $V \circ V \circ V \subseteq W$, and by the identities in Fig. 3.1, we may write

$$V \circ V \circ V = \bigcup_{\langle x,y\rangle \in V} V[x] \times V[y].$$

Hence $\langle x, y\rangle \in W^o$ for every $\langle x, y\rangle \in V$, so $V \subseteq W^o$, and since $V \in \mathfrak{u}$, we conclude $W^o \in \mathfrak{u}$ from $\mathfrak{u}$ being upper closed.

3. Because $\mathfrak{u}$ is a filter, and $U \subseteq U^a$, we infer $U^a \in \mathfrak{u}$. $\dashv$

The closure of a subset of X and the closure of a subset of $X \times X$ may be described as well directly through uniformity $\mathfrak{u}$. These are truly remarkable representations.

Proposition 3.6.56 *$A^a = \bigcap\{U[A] \mid U \in \mathfrak{u}\}$ for $A \subseteq X$, and $M^a = \bigcap\{U \circ M \circ U \mid U \in \mathfrak{u}\}$ for $M \subseteq X \times X$.*

Proof 1. We use the characterization of a point x in the closure through its neighborhood filter from Lemma 1.5.52: $x \in A^a$ iff $U[x] \cap A \neq \emptyset$ for all symmetric $U \in \mathfrak{u}$, because the symmetric neighborhoods form a base for $\mathfrak{u}$. Now $z \in U[x] \cap A$ iff $z \in A$ and $\langle x, z\rangle \in U$ iff $z \in U[z]$ and $z \in A$, hence $U[x] \cap A \neq \emptyset$ iff $x \in U[A]$, because U is symmetric. But this means $A^a = \bigcap\{U[A] \mid U \in \mathfrak{u}\}$.

2. Let $\langle x, y\rangle \in M^a$, then $U[x] \times U[y] \cap M \neq \emptyset$ for all symmetric neighborhoods $U \in \mathfrak{u}$, so that $\langle x, y\rangle \in U \circ M \circ U$ for all symmetric neighborhoods. This accounts for the inclusion from left to right. If $\langle x, y\rangle \in U \circ M \circ U$ for all neighborhoods U, then for every $U \in \mathfrak{u}$ there exists $\langle a, b\rangle \in M$ with $\langle a, b\rangle \in U[x] \times (U^{-1})[y]$, thus $\langle x, y\rangle \in M^a$. $\dashv$

Hence

Corollary 3.6.57 *The closed symmetric neighborhoods form a base for the uniformity.*

Proof Let $U \in \mathfrak{u}$, then there exists a symmetric $V \in \mathfrak{u}$ with $V \circ V \circ V \subseteq U$ with $V \subseteq V^a \subseteq V \circ V \circ V$ by Proposition 3.3.1. Hence $W := V^a \cap (V^a)^{-1})$ is a member of $\mathfrak{u}$ which is contained in U. $\dashv$

Proposition 3.6.56 has also an interesting consequence when looking at the characterization of Hausdorff spaces in Proposition 3.3.1. Putting $M = \Delta$, we obtain $\Delta^a = \bigcap\{U \circ U \mid U \in \mathfrak{u}\}$, so that the associated topological space is Hausdorff iff the intersection of all neighborhoods is the diagonal Δ. Uniform spaces with $\bigcap \mathfrak{u} = \Delta$ are called *separated*.

Separated space

Pseudometrization

We will see shortly that the topology for a separated uniform space is completely regular. First, however, we will show that we can generate pseudometrics from the uniformity by the following idea: suppose that we have a neighborhood V, then there exists a neighborhood V_2 with $V_2 \circ V_2 \circ V_2 \subseteq V_1 := V$; continuing in this fashion, we find for the neighborhood V_n a neighborhood V_{n+1} with $V_{n+1} \circ V_{n+1} \circ V_{n+1} \subseteq V_n$, and finally put $V_0 := X \times X$. Given a pair $\langle x, y \rangle \in X \times X$, this sequence $(V_n)_{n \in \mathbb{N}}$ is now used as a witness to determine how far apart these points are: put $f_V(x, y) := 2^{-n}$, iff $\langle x, y \rangle \in V_n \setminus V_{n-1}$, and $d_V(x, y) := 0$ iff $\langle x, y \rangle \in \bigcap_{n \in \mathbb{N}} V_n$. Then f_V will give rise to a pseudometric d_V, the pseudometric associated with V, as we will show below. d_V

This means that many pseudometric spaces are hidden deep inside a uniform space! Moreover, if we need a pseudometric, we construct one from a neighborhood. These observations will turn out to be fairly practical later on. But before we are in a position to make use of them, we have to do some work.

Proposition 3.6.58 *Assume that $(V_n)_{n \in \mathbb{N}}$ is a sequence of symmetric subsets of $X \times X$ with these properties for all $n \in \mathbb{N}$:*

- $\Delta \subseteq V_n$,
- $V_{n+1} \circ V_{n+1} \circ V_{n+1} \subseteq V_n$.

Put $V_0 := X \times X$. Then there exists a pseudometric d with

$$V_n \subseteq \{\langle x, y \rangle \mid d(x, y) < 2^{-n}\} \subseteq V_{n-1}$$

for all $n \in \mathbb{N}$.

Proof 0. The proof uses the idea outlined above. The main effort will be showing that we can squeeze $\{\langle x, y \rangle \mid d(x, y) < 2^{-n}\}$ between V_n and V_{n-1}.

1. Put $f(x, y) := 2^{-n}$ iff $\langle x, y \rangle \in V_n \setminus V_{n-1}$, and let $f(x, y) := 0$ iff $\langle x, y \rangle \in \bigcap_{n \in \mathbb{N}} V_n$. Then $f(x, x) = 0$, and $f(x, y) = f(y, x)$, because each V_n is symmetric. Define

$$d(x, y) := \inf\{\sum_{i=0}^{k} f(x_i, x_{i+1}) \mid x_0, \ldots, x_{k+1} \in X \text{ with } x_0 = x, x_{k+1} = y, k \in \mathbb{N}\}$$

So we look at all paths leading from x to y, sum the weight of all their edges, and look at their smallest value. Since we may concatenate a path from x to y with a path from y to z to obtain one from x to z, the triangle inequality holds for d, and since $d(x, y) \leq f(x, y)$, we know that $V_n \subseteq \{\langle x, y\rangle \mid d(x, y) < 2^{-n}\}$. The latter set is contained in V_{n-1}; to show this is a bit tricky and requires an intermediary step.

2. We show by induction on n that

$$f(x_0, x_{n+1}) \leq 2 \cdot \sum_{i=0}^{n} f(x_i, x_{i+1}),$$

so if we have a path of length n, then the weight of the edge connecting their endpoints cannot be greater than twice the weight on an arbitrary path. If $n = 1$, there is nothing to show. So assume the assertion is proved for all path with less that n edges. We take a path from x_0 to x_{n+1} with n edges $\langle x_i, x_{i+1}\rangle$. Let w be the weight of the path from x_0 to x_{n+1}, and let k be the largest integer such that the path from x_0 to x_k is at most $w/2$. Then the path from x_{k+1} to x_{n+1} has a weight at most $w/2$ as well. Now $f(x_0, x_k) \leq w$ and $f(x_{k+1}, x_{n+1}) \leq w$ by induction hypothesis, and $f(x_k, x_{k+1}) \leq w$. Let $m \in \mathbb{N}$ the smallest integer with $2^{-m} \leq w$, then we have $\langle x_0, x_k\rangle, \langle x_k, x_{k+1}\rangle, \langle x_{k+1}, x_{n+1}\rangle \in V_m$, thus $\langle x_0, x_{n+1}\rangle \in V_{m-1}$. This implies $f(x_0, x_{n+1}) \leq 2^{-(m-1)} \leq 2 \cdot w = 2 \cdot \sum_{i=0}^{n} f(x_i, x_{i+1})$.

3. Now let $d(x, y) < 2^{-n}$, then $f(x, y) \leq 2^{-(n-1)}$ by part 2., and hence $\langle x, y\rangle \in V_{n-1}$. $\dashv$

This has a—somewhat unexpected—consequence because it permits characterizing those uniformities, which are generated by a pseudometric.

Proposition 3.6.59 *The uniformity* $\mathfrak{u}$ *of* X *is generated by a pseudometric iff* $\mathfrak{u}$ *has a countable base.*

Proof Let $\mathfrak{u}$ be generated by a pseudometric d, then the sets $\{V_{d,r} \mid 0 < r \in \mathbb{Q}\}$ are a countable basis. Let, conversely, $\mathfrak{b} := \{U_n \mid n \in \mathbb{N}\}$ be a countable base for $\mathfrak{u}$. Put $V_0 := X \times X$ and $V_1 := U_1$, and construct inductively the sequence $(V_n)_{n\in\mathbb{N}} \subseteq \mathfrak{b}$ of symmetric base elements with $V_n \circ V_n \circ V_n \subseteq V_{n-1}$ and $V_n \subseteq U_n$ for $n \in \mathbb{N}$. Then $\{V_n \mid n \in \mathbb{N}\}$ is a base for $\mathfrak{u}$. In fact, given $U \in \mathfrak{u}$, there exists $U_n \in \mathfrak{b}$ with $U_n \subseteq U$, hence $V_n \subseteq U$ as well. Construct d for this sequence as above, then we have $V_n \subseteq \{\langle x, y\rangle \mid d(x, y) < 2^{-n}\} \subseteq V_{n-1}$. Thus the sets $V_{d,r}$ are a base for the uniformity. $\dashv$

Note that this does not translate into an observation of the metrizability of the underlying topological space. This space may carry a metric, but the uniform space from which it is derived does not.

Example 3.6.60 Let X be an uncountable set, and let $\mathfrak{u}$ be the uniformity given by the finite partitions; see Example 3.6.49. Then we have seen in Example 3.6.52 that the topology induced by $\mathfrak{u}$ on X is the discrete topology, which is metrizable.

Assume that $\mathfrak{u}$ is generated by a pseudometric, then Proposition 3.6.59 implies that $\mathfrak{u}$ has a countable base; thus given a finite partition π, there exists a finite partition π^* such that $V_{\pi^*} \subseteq V_\pi$, and V_{π^*} is an element of this base. Here $V_{\{P_1,\dots,P_n\}} := \bigcup_{i=1}^n (P_i \times P_i)$ is the basic neighborhood for $\mathfrak{u}$ associated with partition $\{P_1, \dots, P_n\}$. But for any given partition π^* we can only form a finite number of other partitions π with $V_{\pi^*} \subseteq V_\pi$, so that we have only a countable number of partitions on X. ✌

This is another consequence of Proposition 3.6.58: each uniform space satisfies the separation axiom $T_{3\frac{1}{2}}$. For establishing this claim, we take a closed set $F \subseteq X$ and a point $x_0 \notin F$, then we have to produce a continuous function $f : X \to [0, 1]$ with $f(x_0) = 0$ and $f(y) = 1$ for $y \in A$. This is how to do it. Since $X \setminus F$ is open, we find a neighborhood $U \in \mathfrak{u}$ with $U[x_0] \subseteq X \setminus F$. Let d_U be the pseudometric associated with U, then $\{\langle x, y\rangle \mid d_U(x, y) < 1/2\} \subseteq U$. Clearly, $x \mapsto d_U(x, x_0)$ is a continuous function on X; hence

$$f(x) := \max\{0, 1 - 2 \cdot d_U(x, x_0)\}$$

is continuous with $f(x_0) = 1$ and $f(y) = 0$ for $y \in F$, and thus f has the required properties. Thus we have shown

Proposition 3.6.61 *A uniform space is a $T_{3\frac{1}{2}}$-space; a separated uniform space is completely regular.* $\dashv$

Cauchy Filters

We generalize the notion of a Cauchy sequence to uniform spaces now. We do this in order to obtain a notion of convergence which includes convergence in topological spaces and which carries the salient features of a Cauchy sequence with it.

First, we note that filters are a generalization for sequences. So let us have a look at what can be said, when we construct the filter $\mathfrak{F}$ for a Cauchy sequence $(x_n)_{n\in\mathbb{N}}$ in a pseudometric space (X, d). $\mathfrak{F}$ has the sets $\mathfrak{c} := \{B_n \mid n \in \mathbb{N}\}$ with $B_n := \{x_m \mid m \geq n\}$ as a base. Being a Cauchy filter says that for each $\epsilon > 0$, there exists $n \in \mathbb{N}$ such that $B_n \times B_n \subseteq V_{d,\epsilon}$; this inclusion holds then for all B_m with $m \geq n$ as well. Because $\mathfrak{c}$ is the base for $\mathfrak{F}$, and the sets $V_{d,r}$ are a base for the uniformity, we may reformulate that $\mathfrak{F}$ is a Cauchy filter iff for each neighborhood U there exists $B \in \mathfrak{F}$ such that $B \times B \subseteq U$. Now this looks like a property which may be formulated for general uniform spaces.

Small sets

Fix the uniform space $(X, \mathfrak{u})$. Given $U \in \mathfrak{u}$, the set $M \subseteq X$ is called *U-small* iff $M \times M \subseteq U$. A collection $\mathcal{F}$ of sets is said to contain *small sets* iff given $U \in \mathfrak{u}$ there exists $A \in \mathcal{F}$ which is U-small, or, equivalently, given $U \in \mathfrak{u}$ there exists $x \in X$ with $A \subseteq U[x]$.

This helps in formulating the notion of a Cauchy filter.

Definition 3.6.62 *A filter $\mathfrak{F}$ is called a* Cauchy filter *iff it contains small sets.*

In this sense, a Cauchy sequence induces a Cauchy filter. Convergent filters are Cauchy filters as well:

Lemma 3.6.63 *If $\mathfrak{F} \to x$ for some $x \in X$, then $\mathfrak{F}$ is a Cauchy filter.*

Proof Let $U \in \mathfrak{u}$, then there exists a symmetric $V \in \mathfrak{u}$ with $V \circ V \subseteq U$. Because $\mathfrak{U}(x) \subseteq \mathfrak{F}$, we conclude $V[x] \in \mathfrak{F}$, and $V[x] \times V[x] \subseteq U$; thus $V[x]$ is a U-small member of $\mathfrak{F}$. $\dashv$

But the converse does not hold, as the following example shows.

Example 3.6.64 Let $\mathfrak{u}$ be the uniformity induced by the finite partitions with X infinite. We claim that each ultrafilter $\mathfrak{F}$ is a Cauchy filter. In fact, let $\pi = \{A_1, \ldots, A_n\}$ be a finite partition, then $V_\pi = \bigcup_{i=1}^n A_i \times A_i$ is the corresponding neighborhood, then there exists i^* with $A_{i^*} \in \mathfrak{F}$. This is so since, if an ultrafilter contains the finite union of sets, it must contain one of them. A_{i^*} is V-small.

The topology induced by this uniformity is the discrete topology; see Example 3.6.52. This topology is not compact, since X is infinite. By Theorem 3.2.11, there are ultrafilters which do not converge. ✌

If x is an accumulation point of a Cauchy sequence in a pseudometric space, then we know that $x_n \to x$; this is fairly easy to show. A similar observation can be made for Cauchy filters, so that we have a partial converse to Lemma 3.6.63.

Lemma 3.6.65 *Let x be an accumulation point of the Cauchy filter $\mathfrak{F}$, then $\mathfrak{F} \to x$.*

Proof Let $V \in \mathfrak{u}$ be a closed neighborhood; in view of Corollary 3.6.57 is sufficient to show that $V[x] \in \mathfrak{F}$; then it will follow that $\mathfrak{U}(x) \subseteq \mathfrak{F}$. Because $\mathfrak{F}$ is a Cauchy filter, we find $F \in \mathfrak{F}$ with $F \times F \subseteq V$, and because V is closed, we may assume that F is closed as well (otherwise, we replace it by its closure). Because F is closed and x is an accumulation point of $\mathfrak{F}$, we know from Lemma 3.2.14 that $x \in F$; hence $F \subseteq V[x]$. This implies $\mathfrak{U}(x) \subseteq \mathfrak{F}$. $\dashv$

Definition 3.6.66 *The uniform space $(X, \mathfrak{u})$ is called* complete *iff each Cauchy filter converges.*

Each Cauchy sequence converges in a complete uniform space, because the associated filter is a Cauchy filter.

A slight reformulation is given in the following proposition, which is the uniform counterpart to the characterization of complete pseudometric spaces in Proposition 3.5.25. Recall that a collection of sets is said to have the *finite intersection property* iff each finite subfamily has a nonempty intersection.

Proposition 3.6.67 *The uniform space $(X, \mathfrak{u})$ is complete iff each family of closed sets which has the finite intersection property and which contains small sets has a non-void intersection.*

Proof This is essentially a reformulation of the definition, but let us see.

1. Assume that $(X, \mathfrak{u})$ is complete, and let $\mathcal{A}$ be a family of closed sets with the finite intersection property, which contains small sets. Hence $\mathfrak{F}_0 := \{F_1 \cap \ldots \cap F_n \mid n \in \mathbb{N}, F_1, \ldots, F_n \in \mathcal{A}\}$ is a filter base. Let $\mathfrak{F}$ be the corresponding filter, then $\mathfrak{F}$ is a Cauchy filter, for $\mathcal{A}$; hence $\mathfrak{F}_0$ contains small sets. Thus $\mathfrak{F} \to x$, so that $\mathfrak{U}(x) \subseteq \mathfrak{F}$; thus $x \in \bigcap_{F \in \mathfrak{F}} F^a \subseteq \bigcap_{A \in \mathcal{A}} A$.

2. Conversely, let $\mathfrak{F}$ be a Cauchy filter. Since $\{F^a \mid F \in \mathfrak{F}\}$ is a family of closed sets with the finite intersection property which contains small

sets, the assumption says that $\bigcap_{F \in \mathfrak{F}} F^a$ is not empty and contains some x. But then x is an accumulation point of $\mathfrak{F}$ by Lemma 3.2.14, so $\mathfrak{F} \to x$ by Lemma 3.6.65. $\dashv$

As in the case of pseudometric spaces, compact spaces are derived from a complete uniformity.

Lemma 3.6.68 *Let $(X, \mathfrak{u})$ be a uniform space so that the topology associated with the uniformity is compact. Then the uniform space $(X, \mathfrak{u})$ is complete.*

Proof In fact, let $\mathfrak{F}$ be a Cauchy filter on X. Since the topology for X is compact, the filter has an accumulation point x by Corollary 3.2.15. But Lemma 3.6.65 tells us then that $\mathfrak{F} \to x$. Hence each Cauchy filter converges. $\dashv$

The uniform space which is derived from an ideal on the powerset of a set, which has been defined in Example 3.6.49 (part 5) is complete. We establish this first for Cauchy nets as the natural generalization of Cauchy sequences and then translate the proof to Cauchy filters. This will permit an instructive comparison of the handling of these two concepts.

Cauchy net

Example 3.6.69 Recall the definition of a net on page 301. A net $(x_i)_{i \in N}$ in the uniform space X is called a *Cauchy net* iff, given a neighborhood $U \in \mathfrak{u}$, there exists $i \in N$ such that $\langle x_j, x_k \rangle \in U$ for all $j, k \in N$ with $j, k \geq i$. The net converges to x iff given a neighborhood U, there exists $i \in N$ such that $\langle x_j, x \rangle \in U$ for $j \geq i$.

Now assume that $\mathcal{I} \subseteq \mathcal{P}(X)$ is an ideal; part 5 of Example 3.6.49 defines a uniformity $\mathfrak{u}_{\mathcal{I}}$ on $\mathcal{P}(X)$ which has the sets $V_I := \{\langle A, B \rangle \mid A, B \in \mathcal{P}(X), A \Delta B \subseteq I\}$ as a base, as I runs through $\mathcal{I}$. We claim that each Cauchy net $(F_i)_{i \in N}$ converges to $F := \bigcup_{i \in N} \bigcap_{j \geq i} F_j$.

In fact, let a neighborhood U be given; we may assume that $U = V_I$ for some ideal $I \in \mathcal{I}$. Thus there exists $i \in N$ such that $\langle F_j, F_k \rangle \in V_I$ for all $j, k \geq i$; hence $F_j \Delta F_k \subseteq I$ for all these j, k. Let $x \in F \Delta F_j$ for $j \geq i$.

- If $x \in F$, we find $i_0 \in N$ such that $x \in F_k$ for all $k \geq i_0$. Fix $k \in N$ so that $k \geq i$ and $k \geq i_0$, which is possible since N is directed. Then $x \in F_k \Delta F_j \subseteq I$.

- If $x \notin F$, we find for each $i_0 \in N$ some $k \geq i_0$ with $x \notin F_k$. Pick $k \geq i_0$, then $x \notin F_k$; hence $x \in F_\gamma \Delta F_j \subseteq I$

Thus $\langle F, F_j \rangle \in V_I$ for $j \geq i$, and hence the net converges to F. ✌

Now let us investigate convergence of a Cauchy filter. One obvious obstacle in a direct translation seems to be the definition of the limit, because this appears to be bound to the net's indices. But look at this. If $(x_i)_{i \in N}$ is a net, then the sets $\mathcal{B}_i := \{x_j \mid j \geq i\}$ form a filter base $\mathfrak{B}$, as i runs through the directed set N (see the discussion on page 301). Thus we have defined F in terms of this base, viz., $F = \bigcup_{\mathcal{B} \in \mathfrak{B}} \bigcap \mathcal{B}$. This gives an idea for the filter-based case.

Example 3.6.70 Let $\mathfrak{u}_{\mathcal{I}}$ be the uniformity on $\mathcal{P}(X)$ discussed in Example 3.6.69. Then each Cauchy filter $\mathfrak{F}$ converges. In fact, let $\mathfrak{B}$ be a base for $\mathfrak{F}$, then $\mathfrak{F} \to F$ with $F := \bigcup_{\mathcal{B} \in \mathfrak{B}} \bigcap \mathcal{B}$.

Let U be a neighborhood in $\mathfrak{u}_{\mathcal{I}}$, and we may assume that $U = V_I$ for some $I \in \mathcal{I}$. Since $\mathfrak{F}$ is a Cauchy filter, we find $\mathcal{F} \in \mathfrak{F}$ which is V_I-small; hence $F \Delta F' \subseteq I$ for all $F, F' \in \mathcal{F}$. Let $F_0 \in \mathcal{F}$, and consider $x \in F \Delta F_0$; we show that $x \in I$ by distinguishing these cases:

- If $x \in F$, then there exists $\mathcal{B} \in \mathfrak{B}$ such that $x \in \bigcap \mathcal{B}$. Because $\mathcal{B}$ is an element of base $\mathfrak{B}$, and because $\mathfrak{F}$ is a filter, $\mathcal{B} \cap \mathcal{F} \neq \emptyset$, so we find $G \in \mathcal{B}$ with $G \in \mathcal{F}$, in particular $x \in G$. Consequently $x \in G \Delta F_0 \subseteq I$, since $\mathcal{F}$ is V_I-small.
- If $x \notin F$, we find for each $\mathcal{B} \in \mathfrak{B}$ some $G \in \mathcal{B}$ with $x \notin G$. Since $\mathfrak{B}$ is a base for $\mathfrak{F}$, there exists $\mathcal{B} \in \mathfrak{B}$ with $\mathcal{B} \subseteq \mathcal{F}$, so there exists $G \in \mathcal{F}$ with $x \notin G$. Hence $x \in G \Delta F_0 \subseteq I$.

Thus $F \Delta F_0 \subseteq I$, and hence $\langle F, F_0 \rangle \in V_I$. This means $\mathcal{F} \subseteq V_I[F]$, which in turn implies $\mathfrak{U}(F) \subseteq \mathfrak{F}$, or, equivalently, $\mathfrak{F} \to F$. ✌

For further investigations of uniform spaces, we define uniform continuity as the brand of continuity which is adapted to uniform spaces.

Uniform Continuity

Let $f : X \to X'$ be a uniformly continuous map between the pseudometric spaces (X, d) and (X', d'). This means that given $\epsilon > 0$, there exists $\delta > 0$ such that, whenever $d(x, y) < \delta$, $d'(f(x), f(y)) < \epsilon$ follows. In terms of neighborhoods, this means $V_{d,\delta} \subseteq (f \times f)^{-1}\big[V_{d',\epsilon}\big]$, or, equivalently, that $(f \times f)^{-1}\left[V\right]$ is a neighborhood in X, whenever V is a neighborhood in X'. We use this formulation, which is based only on neighborhoods, and not on pseudometrics, for a formulation of uniform continuity.

Definition 3.6.71 *Let* $(X, \mathfrak{u})$ *and* $(Y, \mathfrak{v})$ *be uniform spaces. Then* $f : X \to Y$ *is called* uniformly continuous *iff* $(f \times f)^{-1}\left[V\right] \in \mathfrak{u}$ *for all* $V \in \mathfrak{v}$.

Proposition 3.6.72 *Uniform spaces form a category with uniform continuous maps as morphisms.*

Proof The identity is uniformly continuous, and, since $(g \times g) \circ (f \times f) = (g \circ f) \times (g \circ f)$, the composition of uniformly continuous maps is uniformly continuous again. $\dashv$

We want to know what happens in the underlying topological space. But here nothing unexpected will happen: a uniformly continuous map is continuous with respect to the underlying topologies, formally:

Proposition 3.6.73 *If* $f : (X, \mathfrak{u}) \to (Y, \mathfrak{v})$ *is uniformly continuous, then* $f : (X, \tau_{\mathfrak{u}}) \to (Y, \tau_{\mathfrak{v}})$ *is continuous.*

Proof Let $H \subseteq Y$ be open with $f(x) \in H$. If $x \in f^{-1}\left[H\right]$, there exists a neighborhood $V \in \mathfrak{v}$ such that $V[f(x)] \subseteq H$. Since $U := (f \times f)^{-1}\left[V\right]$ is a neighborhood in X, and $U[x] \subseteq f^{-1}\left[H\right]$, it follows that $f^{-1}\left[H\right]$ is open in X. $\dashv$

The converse is not true, however, as Example 3.5.35 shows.

Before proceeding, we briefly discuss two uniformities on the same topological group which display quite different behaviors, so that the identity is not uniformly continuous.

Example 3.6.74 Let $X := \{f_{a,b} \mid a, b \in \mathbb{R}, a \neq 0\}$ be the set of all affine maps $f_{a,b} : \mathbb{R} \to \mathbb{R}$ with the separated uniformities $\mathfrak{u}_R$ and $\mathfrak{u}_L$, as discussed in Example 3.6.49, part 8.

Let $a_n := d_n := 1/n$, $b_n := -1/n$ and $c_n := n$. Put $g_n := f_{a_n,b_n}$ and $h_n := f_{c_n,d_n}$, $j_n := h_n^{-1}$. Now $g_n \circ h_n = f_{1,1/n^2-1/n} \to f_{1,0}$, $h_n \circ g_n = f_{1,-1+1/n} \to f_{1,-1}$. Now assume that $\mathfrak{u}_R = \mathfrak{u}_L$. Given $U \in \mathfrak{U}(e)$, there exists $V \in \mathfrak{U}(e)$ symmetric such that $V^R \subseteq U^L$. Since $g_n \circ h_n \to f_{1,0}$, there exists for V some n_0 such that $g_n \circ h_n \in V$ for $n \geq n_0$; hence $\langle g_n, j_n\rangle \in V^R$, thus $\langle j_n, g_n\rangle \in V^R \subseteq U^L$, which means that $h_n \circ g_n \in U$ for $n \geq n_0$. Since $U \in \mathfrak{U}(e)$ is arbitrary, this means that $h_n \circ g_n \to e$, which is a contradiction.

Thus we find that the left and the right uniformity on a topological group are different, although they are derived from the same topology. In particular, the identity $(X, \mathfrak{u}_R) \to (X, \mathfrak{u}_L)$ is not uniformly continuous. ✌

We will construct the initial uniformity for a family of maps now. The approach is similar to the one observed for the initial topology (see Definition 3.1.14), but since a uniformity is in particular a filter with certain properties, we have to make sure that the construction can be carried out as intended. Let $\mathcal{F}$ be a family of functions $f : X \to Y_f$, where $(Y_f, \mathfrak{v}_f)$ is a uniform space. We want to construct a uniformity $\mathfrak{u}$ on X rendering all f uniformly continuous, so $\mathfrak{u}$ should contain

$$\mathfrak{s} := \bigcup_{f \in \mathcal{F}} \{(f \times f)^{-1}\big[V\big] \mid V \in \mathfrak{v}_f\},$$

and it should be the smallest uniformity on X with this property. For this to work, it is necessary for $\mathfrak{s}$ to be a subbase. We check this along the properties from Lemma 3.6.50:

1. Let $f \in \mathcal{F}$ and $V \in \mathfrak{v}_f$, then $\Delta_{Y_f} \subseteq V$. Since $\Delta_X = f^{-1}\big[\Delta_{Y_f}\big]$, we conclude $\Delta_X \subseteq (f \times f)^{-1}\big[V\big]$. Thus each element of $\mathfrak{s}$ contains the diagonal of X.
2. Because $\big((f \times f)^{-1}\big[V\big]\big)^{-1} = (f \times f)^{-1}\big[V^{-1}\big]$, we find that, given $U \in \mathfrak{s}$, there exists $V \in \mathfrak{s}$ with $V \subseteq U^{-1}$.
3. Let $U \in \mathfrak{b}$, so that $U = (f \times f)^{-1}\big[V\big]$ for some $f \in \mathcal{F}$ and $V \in \mathfrak{v}_f$. We find $W \in \mathfrak{v}_f$ with $W \circ W \subseteq V$; put $W_0 := (f \times f)^{-1}\big[W\big]$, then $W_0 \circ W_0 \subseteq (f \times f)^{-1}\big[W \circ W\big] \subseteq (f \times f)^{-1}\big[V\big] = U$, so that we find for $U \in \mathfrak{s}$ an element $W_0 \in \mathfrak{s}$ with $W_0 \circ W_0 \subseteq U$.

Thus $\mathfrak{s}$ is the subbase for a uniformity on X, and we have established

Proposition 3.6.75 *Let $\mathcal{F}$ be a family of maps $X \to Y_f$ with $(Y_f, \mathfrak{v}_f)$ a uniform space, then there exists a smallest uniformity $\mathfrak{u}_{\mathcal{F}}$ on X rendering all $f \in \mathcal{F}$ uniformly continuous. $\mathfrak{u}_{\mathcal{F}}$ is called the* initial uniformity *on X with respect to $\mathcal{F}$.*

Proof We know that $\mathfrak{s} := \bigcup_{f \in \mathcal{F}} \{(f \times f)^{-1}\big[V\big] \mid V \in \mathfrak{v}_f\}$ is a subbase for a uniformity $\mathfrak{u}$, which is evidently the smallest uniformity so that each $f \in \mathcal{F}$ is uniformly continuous. So $\mathfrak{u}_f := \mathfrak{u}$ is the uniformity we are looking for. ⊣

Having this tool at our disposal, we can now—in the same way as we did with topologies—define

Product **Product** The product uniformity for the uniform spaces $(X_i, \mathfrak{u}_i)_{i \in I}$ is the initial uniformity on $X := \prod_{i \in I} X_i$ with respect to the projections $\pi_i : X \to X_i$.

Subspace **Subspace** The subspace uniformity $\mathfrak{u}_A$ is the initial uniformity on $A \subseteq X$ with respect to the embedding $i_A : x \mapsto x$.

We can construct dually a final uniformity on Y with respect to a family $\mathcal{F}$ of maps $f : X_f \to Y$ with uniform spaces $(X_f, \mathfrak{u}_f)$, for example, when investigating quotients. The reader is referred to [Bou89, II.2] or to [Eng89, 8.2].

The following is a little finger exercise for the use of a product uniformity. It takes a pseudometric and shows what you would expect: the pseudometric is uniformly continuous iff it generates neighborhoods. The converse holds as well. We do not assume here that d generates $\mathfrak{u}$, rather, it is just an arbitrary pseudometric, of which there may be many.

Proposition 3.6.76 *Let $(X, \mathfrak{u})$ be a uniform space, $d : X \times X \to \mathbb{R}_+$ a pseudometric. Then d is uniformly continuous with respect to the product uniformity on $X \times X$ iff $V_{d,r} \in \mathfrak{u}$ for all $r > 0$.*

Proof 1. Assume first that d is uniformly continuous; thus we find for each $r > 0$ some neighborhood W on $X \times X$ such that $\langle\langle x, u\rangle, \langle y, v\rangle\rangle \in W$ implies $|d(x, y) - d(u, v)| < r$. We find a symmetric neighborhood U on X such that $U_1 \cap U_2 \subseteq W$, where $U_i := (\pi_i \times \pi_i)^{-1}\left[U\right]$ for $i = 1, 2$, and

$$U_1 = \{\langle\langle x, u\rangle, \langle y, v\rangle\rangle \mid \langle x, y\rangle \in U\},$$
$$U_2 = \{\langle\langle x, u\rangle, \langle y, v\rangle\rangle \mid \langle u, v\rangle \in U\}.$$

Thus if $\langle x, y\rangle \in U$, we have $\langle\langle x, y\rangle, \langle y, y\rangle\rangle \in W$; hence $d(x, y) < r$, so that $U \subseteq V_{d,r}$, and thus $V_{d,r} \in \mathfrak{u}$.

2. Assume that $V_{d,r} \in \mathfrak{u}$ for all $r > 0$, and we want to show that d is uniformly continuous in the product. If $\langle x, u\rangle, \langle y, v\rangle \in V_{d,r}$, then

$$\begin{aligned} d(x, y) &\leq d(x, u) + d(u, v) + d(v, y) \\ d(u, v) &\leq d(x, u) + d(x, y) + d(y, v), \end{aligned}$$

hence $|d(x, y) - d(u, v)| < 2 \cdot r$. Thus $(\pi_1 \times \pi_1)^{-1}\left[V_{d,r}\right] \cap (\pi_2 \times \pi_2)^{-1}\left[V_{d,r}\right]$ is a neighborhood on $X \times X$ such that $\langle\langle x, u\rangle, \langle y, v\rangle\rangle \in W$ implies $|d(x, y) - d(u, v)| < 2 \cdot r$. $\dashv$

Combining Proposition 3.6.58 with the observation from Proposition 3.6.76, we have established this characterization of a uniformity through pseudometrics.

Proposition 3.6.77 *The uniformity $\mathfrak{u}$ is the smallest uniformity which is generated by all pseudometrics which are uniformly continuous on $X \times X$, i.e., $\mathfrak{u}$ is the smallest uniformity containing $V_{d,r}$ for all such d and all $r > 0$.* $\dashv$

We fix for the rest of this section the uniform spaces $(X, \mathfrak{u})$ and $(Y, \mathfrak{v})$. Note that for checking uniform continuity it is enough to look at a subbase. The proof is straightforward and hence omitted.

Lemma 3.6.78 *Let $f : X \to Y$ be a map. Then f is uniformly continuous iff $(f \times f)^{-1}\left[V\right] \in \mathfrak{u}$ for all elements of a subbase for $\mathfrak{v}$.* $\dashv$

Cauchy filters are preserved through uniformly continuous maps (the image of a filter is defined on page 303).

Proposition 3.6.79 *Let $f : X \to Y$ be uniformly continuous and $\mathfrak{F}$ a Cauchy filter on X. Then $f(\mathfrak{F})$ is a Cauchy filter.*

Proof Let $V \in \mathfrak{v}$ be a neighborhood in Y, then $U := (f \times f)^{-1}\left[V\right]$ is a neighborhood in X, so that there exists $F \in \mathfrak{F}$ which is U-small; hence $F \times F \subseteq U$, and hence $(f \times f)\left[F \times F\right] = f\left[F\right] \times f\left[F\right] \subseteq V$. Since $f\left[F\right] \in f(\mathfrak{F})$ by Lemma 3.2.5, the image filter contains a V-small member. $\dashv$

A first consequence of Proposition 3.6.79 shows that the subspaces induced by closed sets in a complete uniform space are complete again.

Proposition 3.6.80 *If X is separated, then a complete subspace is closed. Let $A \subseteq X$ be closed and X be complete, then the subspace A is complete.*

Note that the first part does not assume that X is complete and that the second part does not assume that X is separated.

Proof 1. Assume that X is a Hausdorff space and A a complete subspace of X. We show $\partial A \subseteq A$, from which it will follow that A is closed. Let $b \in \partial A$, then $U \cap A \neq \emptyset$ for all open neighborhoods U of b. The trace $\mathfrak{U}(b) \cap A$ of the neighborhood filter $\mathfrak{U}(b)$ on A is a Cauchy filter. In fact, if $W \in \mathfrak{u}$ is a neighborhood for X, which we may choose as symmetric, then $\big((W[b] \cap A) \times (W[b] \cap A)\big) \subseteq W \cap (A \times A)$, which means that $W[b] \cap A$ is $W \cap (A \times A)$- small. Thus $\mathfrak{U}(b) \cap A$ is a Cauchy filter on A, hence it converges to, say, $c \in A$. Thus $\mathfrak{U}(c) \cap A \subseteq \mathfrak{U}(b) \in A$, which means that $b = c$, since X, and hence A, is Hausdorff as a topological space. Thus $b \in A$, and A is closed by Proposition 3.2.4.

2. Now assume that $A \subseteq X$ is closed, and that X is complete. Let $\mathfrak{F}$ be a Cauchy filter on A, then $i_A(\mathfrak{F})$ is a Cauchy filter on X by Proposition 3.6.79. Thus $i_A(\mathfrak{F}) \to x$ for some $x \in X$, and since A is closed, $x \in A$ follows. $\dashv$

We show that a uniformly continuous map on a dense subset into a complete and separated uniform space can be extended uniquely to a uniformly continuous map on the whole space. This was established in Proposition 3.5.37 for pseudometric spaces; having a look at the proof displays the heavy use of pseudometric machinery such as the oscillation and the pseudometric itself. This is not available in the present situation, so we have to restrict ourselves to the tools at our disposal, viz., neighborhoods and filters, in particular Cauchy filters for a complete space. We follow Kelley's elegant proof [Kel55, p. 195].

Theorem 3.6.81 *Let $A \subseteq X$ be a dense subsets of the uniform space $(X, \mathfrak{u})$, and $(Y, \mathfrak{v})$ be a complete and separated uniform space. Then a uniformly continuous map $f : A \to Y$ can be extended uniquely to a uniformly continuous $\varphi : X \to Y$.*

Plan of the proof

Proof 0. The proof starts from the graph $\{\langle a, f(a)\rangle \mid a \in A\}$ of f and investigates the properties of its closure in $X \times Y$. It is shown that the closure is a relation which has $A^a = X$ as its domain, and which is the graph of a map, since the topology of Y is Hausdorff. This map is an extension φ to f, and it is shown that φ is uniformly continuous. We also use the observation that the image of a converging filter under a uniform continuous map is a Cauchy filter, so that completeness of Y kicks in when needed. We do not have to separately establish uniqueness, because this follows directly from Lemma 3.3.20.

1. Let $G_f := \text{graph}(f) = \{\langle a, f(a)\rangle \mid a \in A\}$ be the graph of f. We claim that the closure of the domain of f is the domain of the closure of G_f. Let x be in the domain of the closure of G_f, then there exists $y \in Y$ with $\langle x, y\rangle \in G_f^a$, thus we find a filter $\mathfrak{F}$ on G_f with $\mathfrak{F} \to \langle x, y\rangle$. Thus $\pi_1(\mathfrak{F}) \to x$, so that x is in the closure of the domain of f. Conversely, if x is in the closure of the domain of G_f, we find a filter $\mathfrak{F}$ on the domain of G_f with $\mathfrak{F} \to x$. Since f is uniformly continuous, we know that $f(\mathfrak{F})$ generates a Cauchy filter $\mathfrak{G}$ on Y, which converges to some y. The product filter $\mathfrak{F} \times \mathfrak{G}$ converges to $\langle x, y\rangle$ (see Exercise 3.36), thus x is in the domain of the closure of G_f.

2. Now let $W \in \mathfrak{v}$; we show that there exists a neighborhood $U \in \mathfrak{u}$ with this property: if $\langle x, y\rangle, \langle u, v\rangle \in G_f^a$, then $x \in U[u]$ implies $y \in W[v]$. After having established this, we know

- G_f^a is the graph of a function φ. This is so because Y is separated, hence its topology is Hausdorff. For, assume there exists $x \in X$ some $y_1, y_2 \in Y$ with $y_1 \neq y_2$ and $\langle x, y_1\rangle, \langle x, y_2\rangle \in G_f^a$. Choose $W \in \mathfrak{v}$ with $y_2 \notin W[y_1]$, and consider U as above. Then $x \in U[x]$, hence $y_2 \in W[y_1]$, contradicting the choice of W.

- φ is uniformly continuous. The property above translates to finding for $W \in \mathfrak{v}$ a neighborhood $U \in \mathfrak{u}$ with $U \subseteq (\varphi \times \varphi)^{-1}\big[W\big]$.

So we are done after having established the statement above.

3. Assume that $W \in \mathfrak{v}$ is given, and choose $V \in \mathfrak{v}$ closed and symmetric with $V \circ V \subseteq W$. This is possible by Corollary 3.6.57. There exists $U \in \mathfrak{u}$ open and symmetric with $f\big[U[x]\big] \subseteq V[f(x)]$ for every $x \in A$, since f is uniformly continuous. If $\langle x, y\rangle, \langle u, v\rangle \in G_f^a$ and $x \in U[u]$, then $U[x] \cap U[u]$ is open (since U is open), and there exists $a \in A$ with $x, u \in U[a]$, since A is dense. We claim $y \in \big(f\big[U[a]\big]\big)^a$. Let H be an open neighborhood of y, then, since $U[a]$ is a neighborhood of x, $U[a] \times H$ is a neighborhood of $\langle x, y\rangle$; thus $G_f \cap U[a] \times H \neq \emptyset$. Hence we find $y' \in H$ with $\langle x, y'\rangle \in G_f$, which entails $H \cap f\big[U[a]\big] \neq \emptyset$. Similarly, $z \in \big(f\big[U[a]\big]\big)^a$; note $\big(f\big[U[a]\big]\big)^a \subseteq V[f(a)]$. But now

$\langle y, v\rangle \in V \circ V \subseteq W$, hence $y \in W[v]$. This establishes the claim above, and finishes the proof. $\dashv$

Let us just have a look at the idea lest it gets lost. If $x \in X$, we find a filter $\mathfrak{F}$ on A with $i_A(\mathfrak{F}) \to x$. Then $f(i_a(\mathfrak{F}))$ is a Cauchy filter; hence it converges to some $y \in Y$, which we define as $F(x)$. Then it has to be shown that F is well defined; it clearly extends f. It finally has to be shown that F is uniformly continuous. So there is a lot technical ground which to be covered.

We note on closing that also the completion of pseudometric spaces can be translated into the realm of uniform spaces. Here, naturally, the Cauchy filters defined on the space play an important rôle, and things get very technical. The interplay between compactness and uniformities yields interesting results as well; here the reader is referred to [Bou89, Chap. II] or to [Jam87].

3.7 Bibliographic Notes

The towering references in this area are [Bou89, Eng89, Kur66, Kel55]; the author had the pleasure of taking a course on topology from one of the authors of [Que01], so this text has been an important source, too. The delightful Lecture Note [Her06] by Herrlich has a chapter "Disasters without Choice" which discusses among others the relationship of the axiom of choice and various topological constructions. The discussion of the game related to Baire's Theorem in Sect. 3.5.2 in taken from Oxtoby's textbook [Oxt80, Sect. 6] on the duality between measure and category (*category* in the topological sense introduced on page 358 above); he attributes the game to Banach and Mazur. Other instances of proofs by games for metric spaces are given, e.g., in [Kec94, 8.H, 21]. The uniformity discussed in Example 3.6.70 has been considered in [Dob89] in greater detail. The section on topological systems follows fairly closely the textbook [Vic89] by Vickers, but see also [GHK$^+$03, AJ94], and for the discussion of dualities and the connection to intuitionistic logics [Joh82, Gol06]. The discussion of Gödel's Completeness Theorem in Sect. 3.6.1 is based on the original paper by Rasiowa and Sikorski [RS50] together with occasional glimpses at [RS63], [CK90, Chap. 2.1], [Sri08, Chap. 4] and [Kop89, Chap. 1.2]. Uniform spaces are discussed in [Bou89, Eng89, Kel55, Que01]; special treatises include [Jam87] and [Isb64], the latter one emphasizing an early categorical point of view. A survey on metric techniques in the theory of computation can be found in [SH10].

3.8 Exercises

Exercise 3.1 Formulate and prove an analogue of Proposition 3.1.15 for the final topology for a family of maps.

Exercise 3.2 The Euclidean topology on $\mathbb{R}^n$ is the same as the product topology on $\prod_{i=1}^n \mathbb{R}$.

Exercise 3.3 Recall that the topological space (X, τ) is called *discrete* iff $\tau = \mathcal{P}(X)$. Show that the product $\prod_{i \in I}(\{0,1\}, \mathcal{P}(\{0,1\}))$ is discrete iff the index set I is finite.

Exercise 3.4 Let $L := \{(x_n)_{n\in\mathbb{N}}\} \subseteq \mathbb{R}^{\mathbb{N}} \mid \sum_{n\in\mathbb{N}} |x_n| < \infty\}$ be all sequences of real numbers which are absolutely summable. τ_1 is defined as the trace of the product topology on $\prod_{n\in\mathbb{N}} \mathbb{R}$ on L, τ_2 is defined in the following way: A set G is τ_2-open iff given $x \in G$, there exists $r > 0$ such that $\{y \in L \mid \sum_{n\in\mathbb{N}} |x_n - y_n| < r\} \subseteq G$. Investigate whether the identity maps $(L, \tau_1) \to (L, \tau_2)$ and $(L, \tau_2) \to (L, \tau_1)$ are continuous.

Exercise 3.5 Define for $x, y \in \mathbb{R}$ the equivalence relation $x \sim y$ iff $x - y \in \mathbb{Z}$. Show that $\mathbb{R}/\sim$ is homeomorphic to the unit circle. **Hint:** Example 3.1.19.

Exercise 3.6 Let A be a countable set. Show that a map $q : (A \rightharpoonup B) \to (C \rightharpoonup D)$ is continuous in the topology taken from Example 3.1.5 iff it is continuous, when $A \rightharpoonup B$ as well as $C \rightharpoonup D$ are equipped with the Scott topology.

Exercise 3.7 Let D_{24} be the set of all divisors of 24, including 1, and define an order $\sqsubseteq$ on D_{24} through $x \sqsubseteq y$ iff x divides y. The topology on D_{24} is given through the closure operator as in Example 3.1.20. Write a `Haskell` program listing all closed subsets of D_{24} and determining all filters $\mathfrak{F}$ with $\mathfrak{F} \to 1$. **Hint:** It is helpful to define a type `Set` with appropriate operations first, see [Dob12a, 4.2.2].

Exercise 3.8 Let X be a topological space, $A \subseteq X$, and $i_A : A \to X$ the injection. Show that $x \in A^a$ iff there exists a filter $\mathfrak{F}$ on A such that $i_A(\mathfrak{F}) \to x$.

Exercise 3.9 Show *by expanding Example 3.3.13* that $\mathbb{R}$ with its usual topology is a T_4-space.

Exercise 3.10 Given a continuous bijection $f : X \to Y$ with the Hausdorff spaces X and Y, show that f is a homeomorphism, if X is compact.

Exercise 3.11 Let A be a subspace of a topological space X.

1. If X is a $T_1, T_2, T_3, T_{3\frac{1}{2}}$ space, so is A.
2. If A is closed, and X is a T_4-space, then so is A.

Exercise 3.12 A function $f : X \to \mathbb{R}$ is called *lower semicontinuous* iff for each $c \in \mathbb{R}$ the set $\{x \in X \mid f(x) < c\}$ is open. If $\{x \in X \mid f(x) > c\}$ is open, then f is called *upper semicontinuous*. If X is compact, then a lower semicontinuous map assumes on X its maximum, and an upper semicontinuous map assumes its minimum.

Exercise 3.13 Let $X := \prod_{i \in I} X_i$ be the product of the Hausdorff space $(X_i)_{i \in I}$. Show that X is locally compact in the product topology iff X_i is locally compact for all $i \in I$, and all but a finite number of X_i are compact.

Exercise 3.14 Given $x, y \in \mathbb{R}^2$, define

$$D(x, y) := \begin{cases} |x_2 - y_2|, & \text{if } x_1 = y_1 \\ |x_2| + |y_2| + |x_1 - y_1|, & \text{otherwise.} \end{cases}$$

Show that this defines a metric on the plane $\mathbb{R}^2$. Draw the open ball $\{y \mid D(y, 0) < 1\}$ of radius 1 with the origin as center.

Exercise 3.15 Let (X, d) be a pseudometric space such that the induced topology is T_1. Then d is a metric.

Exercise 3.16 Let X and Y be two first countable topological spaces. Show that a map $f : X \to Y$ is continuous iff $x_n \to x$ implies always $f(x_n) \to f(x)$ for each sequence $(x_n)_{n \in \mathbb{N}}$ in X.

Exercise 3.17 Consider the set $\mathcal{C}([0, 1])$ of all continuous functions on the unit interval, and define

$$e(f, g) := \int_0^1 |f(x) - g(x)| \, dx.$$

Show that

1. e is a metric on $C([0,1])$.
2. $\mathcal{C}([0,1])$ is not complete with this metric.
3. The metrics d on $\mathcal{C}([0,1])$ from Example 3.5.2 and e are not equivalent.

Exercise 3.18 Let (X, d) be an ultrametric space, hence $d(x, z) \leq \max\{d(x, y), d(y, z)\}$ (see Example 3.5.2). Show that

- If $d(x, y) \neq d(y, z)$, then $d(x, z) = \max\{d(x, y), d(y, z)\}$.
- Any open ball $B(x, r)$ is both open and closed, and $B(x, r) = B(y, r)$, whenever $y \in B(x, r)$.
- Any closed ball $S(x, r)$ is both open and closed, and $S(x, r) = S(y, r)$, whenever $y \in S(x, r)$.
- Assume that $B(x, r) \cap B(x', r') \neq \emptyset$, then $B(x, r) \subseteq B(x', r')$ or $B(x', r') \subseteq B(x, r)$.

Exercise 3.19 Let (X, d) be a metric space. Show that X is compact iff each continuous real-valued function on X is bounded.

Exercise 3.20 Show that the set of all nowhere dense sets in a topological space X forms an ideal. Define a set $A \subseteq X$ as *open modulo nowhere dense sets* iff there exists an open set G such that the symmetric difference $A\Delta G$ is nowhere dense (hence both $A \setminus G$ and $G \setminus A$ are nowhere dense). Show that the open sets modulo nowhere dense sets form an σ-algebra.

Exercise 3.21 Consider the game formulated in Sect. 3.5.2; we use the notation from there. Show that there exists a strategy that Angel can win iff $L_1 \cap B$ is of first category for some interval $L_1 \subseteq L_0$.

Exercise 3.22 Let $\mathfrak{u}$ be the additive uniformity on $\mathbb{R}$ from Example 3.6.49. Show that $\{\langle x, y\rangle \mid |x - y| < 1/(1 + |y|)\}$ is not a member of $\mathfrak{u}$.

Exercise 3.23 Show that $V \circ U \circ V = \bigcup_{\langle x,y\rangle \in U} V[x] \times V[y]$ for symmetric $V \subseteq X \times X$ and arbitrary $U \subseteq X \times X$.

Exercise 3.24 Given a base $\mathfrak{b}$ for a uniformity $\mathfrak{u}$, show that

$$\mathfrak{b}' := \{B \cap B^{-1} \mid B \in \mathfrak{b}\},$$
$$\mathfrak{b}'' := \{B^n \mid B \in \mathfrak{b}\}$$

are also bases for $\mathfrak{u}$, when $n \in \mathbb{N}$ (recall $B^1 := B$ and $B^{n+1} := B \circ B^n$).

Exercise 3.25 Show that the uniformities on a set X form a complete lattice with respect to inclusion. Characterize the initial and the final uniformity on X for a family of functions in terms of this lattice.

Exercise 3.26 If two subsets A and B in a uniform space $(X, \mathfrak{u})$ are V-small then $A \cup B$ is $V \circ V$-small, if $A \cap B \neq \emptyset$.

Exercise 3.27 Show that a discrete uniform space is complete. **Hint**: A Cauchy filter is an ultrafilter based on a point.

Exercise 3.28 Let $\mathcal{F}$ be a family of maps $X \to Y_f$ with uniform spaces $(Y_f, \mathfrak{v}_f)$. Show that the initial topology on X with respect to $\mathcal{F}$ is the topology induced by the product uniformity.

Exercise 3.29 Equip the product $X := \prod_{i \in I} X_i$ with the product uniformity for the uniform spaces $\big((X_i, \mathfrak{u}_i)\big)_{i \in I}$, and let $(Y, \mathfrak{v})$ be a uniform space. A map $f : Y \to X$ is uniformly continuous iff $\pi_i \circ f : Y \to X_i$ is uniformly continuous for each $i \in I$.

Exercise 3.30 Let X be a topological system. Show that the following statements are equivalent

1. X is homeomorphic to $\boldsymbol{SP}(Y)$ for some topological system Y.
2. For all $a, b \in X^\sharp$ holds $a = b$, provided we have $x \models a \Leftrightarrow x \models b$ for all $x \in X^\flat$.
3. For all $a, b \in X^\sharp$ holds $a \leq b$, provided we have $x \models a \Rightarrow x \models b$ for all $x \in X^\flat$.

Exercise 3.31 Let L be a dcpo with a smallest element $\perp$, $f : L \to L$ a monotone map. Then f has a least fixed point (i.e., there exists d with $x = f(x)$, and if $y = f(y)$ for some $y \in L$, then $x \leq y$. (Hint: Use transfinite induction.)

Exercise 3.32 Show that a Hausdorff space is sober.

Exercise 3.33 Let X and Y be compact topological spaces with their Banach spaces $\mathcal{C}(X)$ resp. $\mathcal{C}(Y)$ of real continuous maps. Let $f : X \to Y$ be a continuous map, then

$$f^* : \begin{cases} \mathcal{C}(Y) & \to \mathcal{C}(X) \\ g & \mapsto g \circ f \end{cases}$$

defines a continuous map (with respect to the respective norm topologies). f^* is onto iff f is an injection. f is onto iff f^* is an isomorphism of $\mathcal{C}(Y)$ onto a ring $A \subseteq \mathcal{C}(X)$ which contains constants.

Exercise 3.34 Let $\mathfrak{L}$ be a language for propositional logic with constants C and V as the set of propositional variables. Prove that a consistent theory T has a model, hence a map $h : V \to \mathbb{2}$ such that each formula in T is assigned the value $\top$. **Hint:** Fix an ultrafilter on the Lindenbaum algebra of T and consider the corresponding morphism into $\mathbb{2}$.

Exercise 3.35 Let G be a topological group; see Example 3.1.25. Given $F \subseteq G$ closed, show that

1. gF and Fg are closed,
2. F^{-1} is closed,
3. MF and FM are closed, provided M is finite.
4. If $A \subseteq G$, then $A^a = \bigcap_{U \in \mathfrak{U}(e)} AU = \bigcap_{U \in \mathfrak{U}(e)} UA = \bigcap_{U \in \tau} AU = \bigcap_{U \in \tau} UA$.

Exercise 3.36 Let $(X, \mathfrak{u})$ and $(Y, \mathfrak{v})$ be uniform spaces with Cauchy filters $\mathfrak{F}$ and $\mathfrak{G}$ on X resp. Y. Define $\mathfrak{F} \times \mathfrak{G}$ as the smallest filter on $X \times Y$ which contains $\{A \times B \mid A \in \mathfrak{F}, B \in \mathfrak{G}\}$. Show that $\mathfrak{F} \times \mathfrak{G}$ is a Cauchy filter on $X \times Y$ with $\mathfrak{F} \times \mathfrak{G} \to \langle x, y \rangle$ iff $\mathfrak{F} \to x$ and $\mathfrak{G} \to y$.

Chapter 4

Measures for Probabilistic Systems

Markov transition systems are based on transition probabilities on a measurable space. This is a generalization of discrete spaces, declaring certain sets to be measurable. So, in contrast to assuming that we know the probability for the transition between two states, we have to model the probability of a transition going from one state to a set of states: Point-to-point probabilities are no longer available due to working in a comparatively large space. Measurable spaces are the domains of the probabilities involved. This approach has the advantage of being more general than finite or countable spaces, but now one deals with a fairly involved mathematical structure; all of a sudden the dictionary has to be extended with words like "universally measurable" or "sub-σ-algebra." Measure theory becomes an area where one has to find answers to questions which did not appear to be particularly involved before, in the much simpler world of discrete measures (the impression should not arise that I consider discrete measures as kiddie stuff; they are difficult enough to handle. The continuous case, as it is called sometimes, offers questions, however, which simply do not arise in the discrete context). Many arguments in this area are of a measure theoretic nature, and I want to introduce the reader to the necessary tools and techniques.

E.-E. Doberkat, *Special Topics in Mathematics for Computer Scientists*, DOI 10.1007/978-3-319-22750-4_4

It starts off with a discussion of σ-algebras, which have already been met in Sect. 1.6. We look at the structure of σ-algebras, in particular at its generators; it turns out that the underlying space has something to say about it. In particular we will deal with Polish spaces and their brethren. Two aspects deserve to be singled out. The σ-algebra on the base space determines a σ-algebra on the space of all finite measures, and, if this space has a topology, it determines also a topology, the Alexandrov topology. These constructions are studied, since they also affect the applications in logic, and for transition systems, in which measures are vital. Second, we show that we can construct measurable selections, which then enable constructions which are interesting from a categorical point of view.

After having laid the groundwork with a discussion of σ-algebras as the domains of measures, we show that the integral of a measurable function can be constructed through an approximation process, very much in the tradition of the Riemann integral, but with a larger scope. We also go the other way: Given an integral, we construct a measure from it. This is the elegant way P.J. Daniell did propose for constructing measures, and it can be brought to fruit in this context for a direct and elegant proof of the Riesz Representation Theorem on compact metric spaces.

Having all these tools at our disposal, we look at product measures, which can be introduced now through a kind of line sweeping—if you want to measure an area in the plane, you measure the line length as you sweep over the area; this produces a function of the abscissa, which then yields the area through integration. One of the main tools here is Fubini's Theorem. The product measure is not confined to two factors; we discuss the general case. This includes a discussion of projective systems, which may be considered as a generalization of sequences of products. A case study shows that projective systems arise easily in the study of continuous time stochastic logics.

Now that integrals are available, we turn back and have a look at topologies on spaces of measures; one suggests itself—the weak topology which is induced by the continuous functions. This is related to the Alexandrov topology. It is shown that there is a particularly handy metric for the weak topology and that the space of all finite measures is complete with this metric, so that we now have a Polish space. This is capitalized on when discussing selections for set-valued maps into this space, which are helpful in showing that Polish spaces are closed under bisimulations. We use measurable selections for an investigation into the structure of quotients in the Kleisli monad, providing another example for the interplay of arguments from measure theory and categories.

This interplay is stressed also for the investigation of stochastic effectivity functions, which leads to an interpretation of game logics. Since it is known in the world of relations that nondeterministic Kripke models are inadequate for the study of game logics and that effectivity functions serve this purpose well, we develop an approach to stochastic effectivity functions and apply these functions to an interpretation of game logics. This serves as an example for stochastic modeling in logics; it demonstrates the close interaction of algebraic and probabilistic reasoning, indicating the importance of structural arguments, which go beyond the mere discussion of probabilities.

Finally, we take up a true classic: L_p-spaces. We start from Hilbert spaces, apply the representation of linear functionals on L_2 to obtain the Radon–Nikodym Theorem through von Neumann's ingenious proof, and derive from it the representation of the dual spaces. This is applied to disintegration, where we show that a measure on a product can be decomposed into a projection and a transition kernel. On the surface this does not look like an application area for L_p-spaces; the relationship derives from the Radon–Nikodym Theorem.

Because we are driven by applications to Markov transition systems and similar objects, we did not strive for the most general approach to measure and integral. In particular, we usually formulate the results for finite or σ-finite measures, leaving the more general cases outside of our focus. This means also that I do not deal with complex measures (and the associated linear spaces over the complex numbers). Things are discussed rather in the realm of real numbers; we show, however, in which way one could start to deal with complex measures when the occasion arises. Of course, a lot of things had to be left out, among them a careful study of the Borel hierarchy and applications to descriptive set theory, as well as martingales.

4.1 Measurable Sets and Functions

This section contains a systematic study of measurable spaces and measurable functions with a view toward later developments. A brief overview is in order, and a preview indicates why the study of measurable sets is important, nay, fundamental for discussing many of the applications we have in mind.

The measurable structure is lifted to the space of finite measures, which form a measurable set under the weak σ-algebra. This is studied in Sect. 4.1.2. If the underlying space carries a topology, the topological structure is handed down to finite measures through the Alexandrov topology. We will have a look at it in Sect. 4.1.4. The measurable functions from a measurable space to the reals form a vector space, which is also a lattice, and we will show that the step functions, i.e., those functions which take only finite number of values, are dense with respect to pointwise convergence. This mode of convergence is relaxed in the presence of a measure in various ways to almost uniform convergence, convergence almost everywhere, and convergence in measure (Sects. 4.2.1 and 4.2.2), from which also various (pseudo)metrics and norms may be derived.

If the underlying measurable spaces are the Borel sets of a metric space, and if the metric has a countable dense set, then the Borel sets are countably generated as well. But the irritating observation is that being countably generated is not hereditary—a sub-σ-algebra of a countable σ-algebra need not be countably generated. So countably generated σ-algebras deserve a separate look, which is what we will do in Sect. 4.3. The very important class of Polish spaces will be studied in this context as well, and we will show how to manipulate a Polish topology into making certain measurable functions continuous. This is one of the reasons why Polish spaces did not go into the gallery of topological spaces in Sect. 3.6. Polish spaces generalize to analytic spaces in a most natural manner, for example, when taking the factor of a countably generated equivalence relation in a Polish space; we will study the relationship in Sect. 4.3.1. The most important tool here is Souslin's Separation Theorem. This discussion leads quickly to a discussion of the abstract Souslin operation in Sect. 4.5, through which analytic sets may be generated in a Polish space. From there it is but a small step to introducing universally measurable sets in Sect. 4.6, which turn out to be closed under Souslin's operation in general measurable spaces.

Two applications of these techniques are given: Lubin's Theorem extends a measure from a countably generated sub-σ-algebra of the Borel sets of an analytic space to the Borel sets proper; the other application explores the extension a transition kernel to the universal completion (Sects. 4.6.1 and 4.6.2). Lubin's Theorem is established through von Neumann's Selection Theorem, which provides a universally measurable right inverse to a surjective measurable map from an analytic space to a separable measurable space. The topic of selections is taken up in Sect. 4.7, where the selection theorem of Kuratowski and Ryll-Nardzewski is in the center of attention. It gives conditions under which a map which takes values in the closed nonempty subsets of a Polish space has a measurable selector. This is of interest, e.g., when it comes to establishing the existence of bisimulations for Markov transition systems or for identifying the quotient structure of transition kernels.

4.1.1 Measurable Sets

Recall from Example 2.1.12 that a measurable space $(X, \mathcal{A})$ consists of a set X with a σ-algebra $\mathcal{A}$, which is a Boolean algebra of subsets of X that is closed under countable unions (hence countable intersections or countable disjoint unions). If $\mathcal{A}_0$ is a family of subsets of X, then

$$\sigma(\mathcal{A}_0) = \bigcap\{\mathcal{B} \mid \mathcal{B} \text{ is a } \sigma\text{-algebra on } M \text{ with } \mathcal{A}_0 \subseteq \mathcal{A}\}$$

is the smallest σ-algebra on M which contains $\mathcal{A}_0$. This construction works since the power set $\mathcal{P}(X)$ is a σ-algebra on X. Take, for example, as a generator $\mathcal{I}$ all open intervals in the real numbers $\mathbb{R}$; then $\sigma(\mathcal{I})$ is the σ-algebra of real *Borel sets*. These Borel sets are denoted by $\mathcal{B}(\mathbb{R})$, and since each open subset of $\mathbb{R}$ can be represented as a countable union of open intervals, $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra which contains the open sets of $\mathbb{R}$. Unless otherwise stated, the real numbers are equipped with the σ-algebra $\mathcal{B}(\mathbb{R})$.

In general, if (X, τ) is a topological space, the σ-algebra $\mathcal{B}(\tau) := \sigma(\tau)$ is called its *Borel sets*. They will be discussed extensively in the context of Polish spaces. This is, however, not the only σ-algebra of interest on a topological space. $\mathcal{B}(\tau)$

Example 4.1.1 Call $F \subseteq X$ *functionally closed* iff $F = f^{-1}\big[\{0\}\big]$ for some continuous function $f : X \to \mathbb{R}$; $G \subseteq X$ is called functionally open iff $G = X \setminus F$ with F functionally closed. The *Baire sets* $\mathcal{B}a(\tau)$ of (X, τ) are the σ-algebra generated by the functionally closed sets of the space. We write sometimes also $\mathcal{B}a(X)$, if the context is clear. $\mathcal{B}a(\tau)$

Let $F \subseteq X$ be a closed subset of a metric space (X, d); then $d(x, F) := \inf\{d(x, y) \mid y \in F\}$ is the distance of x to F with $x \in F$ iff $d(x, F) = 0$; see Lemma 3.5.7. Moreover, $d(\cdot, F)$ is continuous, and thus $F = d(\cdot, F)^{-1}\big[\{0\}\big]$ is functionally closed; hence the Baire and the Borel sets coincide for metric spaces. ✌

The next example constructs a σ-algebra which comes up quite naturally in the study of stochastic nondeterminism.

Example 4.1.2 Let $\mathcal{A} \subseteq \mathcal{P}(X)$ for some set X, the family of hit sets, and $\mathcal{G}$ a distinguished subsets of $\mathcal{P}(X)$. Define the *hit-σ-algebra* $\mathcal{H}_{\mathcal{A}}(\mathcal{G})$ as the smallest σ-algebra on $\mathcal{G}$ which contains all the sets H_A with $A \in \mathcal{A}$, where H_A is the hit set associated with A, i.e., $H_A := \{B \in \mathcal{G} \mid B \cap A \neq \emptyset\}$. ✌

Rather than working with a closure operation $\sigma(\cdot)$, one sometimes can adjoin additional elements to obtain a σ-algebra from a given one; see also Exercise 4.5. This is demonstrated for a σ-ideal through the following construction, which will be helpful when completing a measure space. Recall that $\mathcal{N} \subseteq \mathcal{P}(X)$ is a σ-ideal iff it is an order ideal which is closed under countable unions (Definition 1.6.10).

Lemma 4.1.3 *Let $\mathcal{A}$ be a σ-algebra on a set X, $\mathcal{N} \subseteq \mathcal{P}(X)$ a σ-ideal. Then*

$$\mathcal{A}_{\mathcal{N}} := \{A \Delta N \mid A \in \mathcal{A}, N \in \mathcal{N}\}$$

is the smallest σ-algebra containing both $\mathcal{A}$ and $\mathcal{N}$.

Proof It is sufficient to demonstrate that $\mathcal{A}_{\mathcal{N}}$ is a σ-algebra. Because

$$X \setminus (A \Delta N) = X \Delta (A \Delta N) = (X \Delta A) \Delta N = (X \setminus A) \Delta N,$$

we see that $\mathcal{A}_{\mathcal{N}}$ is closed under complementation. Now let $\big(A_n \Delta N_n\big)_{n\in\mathbb{N}}$ be a sequence of sets with $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{A}$ and $(N_n)_{n\in\mathbb{N}}$ in $\mathcal{N}$, and we have

$$\bigcup_{n\in\mathbb{N}} (A_n \Delta N_n) = \big(\bigcup_{n\in\mathbb{N}} A_n\big) \Delta N$$

with

$$N = \bigcup_{n\in\mathbb{N}} (A_n \Delta N_n) \Delta \big(\bigcup_{n\in\mathbb{N}} A_n\big) \overset{(\ddagger)}{\subseteq} \bigcup_{n\in\mathbb{N}} \big((A_n \Delta N_n) \Delta A_n\big) = \bigcup_{n\in\mathbb{N}} N_n,$$

using Exercise 4.10 in ($\ddagger$). Because $\mathcal{N}$ is a σ-ideal, we conclude that $N \in \mathcal{N}$. Thus $\mathcal{A}_{\mathcal{N}}$ is also closed under countable unions. Since $\emptyset, X \in \mathcal{A}_{\mathcal{N}}$, it follows that this set is a σ-algebra indeed. ⊣

If $(Y, \mathcal{B})$ is another measurable space, then a map $f : X \to Y$ is called *$\mathcal{A}$-$\mathcal{B}$-measurable* iff the inverse image under f of each set in $\mathcal{B}$ is a member of $\mathcal{A}$, hence iff $f^{-1}\left[G\right] \in \mathcal{A}$ holds for all $G \in \mathcal{B}$; this is discussed in Example 2.1.12.

Checking measurability is made easier by the observation that it suffices for the inverse images of a generator to be measurable sets (see Exercise 2.7).

Lemma 4.1.4 *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and assume that $\mathcal{B} = \sigma(\mathcal{B}_0)$ is generated by a family $\mathcal{B}_0$ of subsets of Y. Then $f : X \to Y$ is $\mathcal{A}$-$\mathcal{B}$-measurable iff $f^{-1}\left[G\right] \in \mathcal{A}$ holds for all $G \in \mathcal{B}_0$.*

Proof Clearly, if f is $\mathcal{A}$-$\mathcal{B}$-measurable, then $f^{-1}\left[G\right] \in \mathcal{A}$ holds for all $G \in \mathcal{B}_0$.

Conversely, suppose $f^{-1}\left[G\right] \in \mathcal{A}$ holds for all $G \in \mathcal{B}_0$, then we need to show that $f^{-1}\left[G\right] \in \mathcal{A}$ for all $G \in \mathcal{B}$. We will use the principle of good sets (see page 86) for the proof. In fact, consider the set $\mathcal{G}$ for which the assertion is true,

$$\mathcal{G} := \{G \in \mathcal{B} \mid f^{-1}\left[G\right] \in \mathcal{A}\}.$$

An elementary calculation shows that the empty set and Y are both members of $\mathcal{G}$, and since $f^{-1}\left[Y \setminus G\right] = X \setminus f^{-1}\left[G\right]$, $\mathcal{G}$ is closed under complementation. Because

$$f^{-1}\left[\bigcup_{i \in I} G_i\right] = \bigcup_{i \in I} f^{-1}\left[G_i\right]$$

holds for any index set I, $\mathcal{G}$ is closed under finite and countable unions. Thus $\mathcal{G}$ is a σ-algebra, so that $\sigma(\mathcal{G}) = \mathcal{G}$ holds. By assumption, $\mathcal{B}_0 \subseteq \mathcal{G}$, so that

$$\mathcal{A} = \sigma(\mathcal{B}_0) \subseteq \sigma(\mathcal{G}) = \mathcal{G} \subseteq \mathcal{A}$$

is inferred. Thus all elements of $\mathcal{B}$ have their inverse image in $\mathcal{A}$. $\dashv$

An example is furnished by a real-valued function $f : X \to \mathbb{R}$ on X which is $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable iff $\{x \in X \mid f(x) \bowtie t\} \in \mathcal{A}$ holds for each $t \in \mathbb{R}$; the relation $\bowtie$ may be taken from $<, \leq, \geq, >$. We infer in particular that a function f from an topological space (X, τ) which is upper or lower semicontinuous (i.e., for which in the *upper semicontinuous* case, the set $\{x \in X \mid f(x) < c\}$ is open, and in the *lower semicontinuous* case, the set $\{x \in X \mid f(x) > c\}$ is open, $c \in \mathbb{R}$

being arbitrary) is Borel measurable. Hence a continuous function is Borel measurable. A continuous function $f : X \to Y$ into a metric space Y is Baire measurable (Exercise 4.2).

These observations will be used frequently.

The proof's strategy is worthwhile repeating, since we will use this strategy over and over again. It consists in having a look at all objects that have the desired property and showing that this *set of good guys* is a σ-algebra. It is similar to showing in a proof by induction that the set of all natural numbers having a certain property is closed under constructing the successor. Then we show that the generator of the σ-algebra is contained in the good guys, which is rather similar to begin the induction. Taking both steps together then yields the desired properties for both cases.

An example is furnished by the equivalence relation induced by a family of sets.

$x \equiv_{\mathcal{C}} x'$

Example 4.1.5 Given a subset $\mathcal{C} \subseteq \mathcal{P}(X)$ for a set X, define the equivalence relation $\equiv_{\mathcal{C}}$ on X upon setting

$$x \equiv_{\mathcal{C}} x' \text{ iff } \forall C \in \mathcal{C} : x \in C \Leftrightarrow x' \in C.$$

Thus $x \equiv_{\mathcal{C}} x'$ iff $\mathcal{C}$ cannot separate the elements x and x'; call $\equiv_{\mathcal{C}}$ the *equivalence relation generated by* $\mathcal{C}$.

Now let $\mathcal{A}$ be a σ-algebra on X with $\mathcal{A} = \sigma(\mathcal{A}_0)$. Then $\mathcal{A}$ and $\mathcal{A}_0$ generate the same equivalence relation, i.e., $\equiv_{\mathcal{A}} = \equiv_{\mathcal{A}_0}$. In fact, define for $x, x' \in X$ with $x \equiv_{\mathcal{A}_0} x'$

$$\mathcal{B} := \{A \in \mathcal{A} \mid x \in A \Leftrightarrow x' \in A\}.$$

Then $\mathcal{B}$ is a σ-algebra with $\mathcal{A}_0 \subseteq \mathcal{B}$; hence $\sigma(\mathcal{A}_0) \subseteq \mathcal{B} \subseteq \mathcal{A}$, so that $\sigma(\mathcal{A}_0) = \mathcal{B}$. Thus $x \equiv_{\mathcal{A}_0} x'$ implies $x \equiv_{\mathcal{A}} x'$; since the reverse implication is obvious, the claim is established. ✌

Let us just briefly discuss initial and final σ-algebras again. The spirit of this is very much similar to defining initial and final topologies; see Sect. 3.1.1 and Definition 2.6.42. If $(X, \mathcal{A})$ is a measurable space and $f : X \to Y$ is a map, then

$$\mathcal{B} := \{D \subseteq Y \mid f^{-1}\left[D\right] \in \mathcal{A}\}$$

is the largest σ-algebra $\mathcal{B}_0$ on Y that renders f $\mathcal{A}$-$\mathcal{B}_0$-measurable; then $\mathcal{B}$ is called the *final* σ-algebra with respect to f. In fact, because the inverse set operator f^{-1} is compatible with the Boolean operations, it is immediate that $\mathcal{B}$ is closed under the operations for a σ-algebra, and a little moment's reflection shows that this is also the largest σ-algebra with this property.

Symmetrically, let $g : P \to X$ be a map; then

$$g^{-1}\left[\mathcal{A}\right] := \{g^{-1}\left[E\right] \mid E \in \mathcal{A}\}$$

is the smallest σ-algebra $\mathcal{P}_0$ on P that renders $g : \mathcal{P}_0 \to \mathcal{A}$-measurable; accordingly, $g^{-1}\left[\mathcal{A}\right]$ is called *initial* with respect to f. It is fairly clear that this is the smallest one with the desired property. In particular, the inclusion $i_Q : Q \to X$ becomes measurable for a subset $Q \subseteq X$ when Q is endowed with the σ-algebra $\{Q \cap B \mid B \in \mathcal{A}\}$. It is called the *trace of $\mathcal{A}$ on Q* and is denoted—in a slight abuse of notation—by $\mathcal{A} \cap Q$.

Initial and final σ-algebras generalize in an obvious way to families of maps. For example, $\sigma\left(\bigcup_{i\in I} g_i^{-1}\left[\mathcal{A}_i\right]\right)$ is the smallest σ-algebra $\mathcal{P}_0$ on P which makes all the maps $g_i : P \to X_i$ $\mathcal{P}_0$-$\mathcal{A}_i$-measurable for a family $\left((X_i, \mathcal{A}_i)\right)_{i\in I}$ of measurable spaces.

This is an intrinsic, universal characterization of the initial σ-algebra for a single map.

Lemma 4.1.6 *Let $(X, \mathcal{A})$ be a measurable space and $f : X \to Y$ be a map. The following conditions are equivalent:*

1. *The σ-algebra $\mathcal{B}$ on Y is final with respect to f.*
2. *If $(P, \mathcal{P})$ is a measurable space, and $g : Y \to P$ is a map, then the $\mathcal{A}$-$\mathcal{P}$-measurability of $g \circ f$ implies the $\mathcal{B}$-$\mathcal{P}$-measurability of g.*

Proof 1. Taking care of 1 $\Rightarrow$ 2, we note that

$$(g \circ f)^{-1}\left[\mathcal{P}\right] = f^{-1}\left[g^{-1}\left[\mathcal{P}\right]\right] \subseteq \mathcal{A}.$$

Consequently, $g^{-1}\left[\mathcal{P}\right]$ is one of the σ-algebras $\mathcal{B}_0$ with $f^{-1}\left[\mathcal{B}_0\right] \subseteq \mathcal{A}$. Since $\mathcal{B}$ is the largest of them, we have $g^{-1}\left[\mathcal{P}\right] \subseteq \mathcal{B}$. Hence g is $\mathcal{B}$-$\mathcal{P}$-measurable.

2. In order to establish the implication 2 $\Rightarrow$ 1, we have to show that $\mathcal{B}_0 \subseteq \mathcal{B}$ whenever $\mathcal{B}_0$ is a σ-algebra on $\mathcal{Y}$ with $f^{-1}\left[\mathcal{B}_0\right] \subseteq \mathcal{A}$. Put $(P, \mathcal{P}) := (Y, \mathcal{B}_0)$, and let g be the identity id_Y. Because $f^{-1}\left[\mathcal{B}_0\right] \subseteq \mathcal{A}$, we see that $id_Y \circ f$ is $\mathcal{B}_0$-$\mathcal{A}$-measurable. Thus id_Y is $\mathcal{B}$-$\mathcal{B}_0$-measurable. But this means $\mathcal{B}_0 \subseteq \mathcal{B}$. $\dashv$

We will use the final σ-algebra mainly for factoring through an equivalence relation. In fact, let α be an equivalence relation on a set X, where $(X, \mathcal{A})$ is a measurable space. Then the factor map

$$\eta_\alpha : \begin{cases} X & \to X/\alpha \\ x & \mapsto [x]_\alpha \end{cases}$$

that maps each element to its class can be made a measurable map by taking the final σ-algebra $\mathcal{A}/\alpha$ with respect to η_α and $\mathcal{A}$ as the σ-algebra on X/α.

Dual to Lemma 4.1.6, the initial σ-algebra is characterized.

Lemma 4.1.7 *Let $(Y, \mathcal{B})$ be a measurable space and $f : X \to Y$ be a map. The following conditions are equivalent:*

1. *The σ-algebra $\mathcal{A}$ on X is initial with respect to f.*
2. *If $(P, \mathcal{P})$ is a measurable space, and $g : P \to X$ is a map, then the $\mathcal{P}$-$\mathcal{B}$-measurability of $f \circ g$ implies the $\mathcal{P}$-$\mathcal{A}$-measurability of g.*

$\dashv$

Let $\big((A_i, \mathcal{A}_i)\big)_{i \in I}$ be a family of measurable spaces; then the product-σ-algebra $\bigotimes_{i \in I} \mathcal{A}_i$ denotes that initial σ-algebra on $\prod_{i \in I} X_i$ for the projections

$$\pi_j : \langle m_i \mid i \in I \rangle \mapsto m_j.$$

It is not difficult to see that $\bigotimes_{i \in I} \mathcal{A}_i = \sigma(\mathcal{Z})$ with

$$\mathcal{Z} := \{\prod_{i \in I} E_i \mid \forall i \in I : E_i \in \mathcal{M}_i, E_i = M_i \text{ for almost all indices}\}$$

as the collection of *cylinder sets* (use Theorem 1.6.30 and the observation that $\mathcal{Z}$ is closed under intersection).

For $I = \{1, 2\}$, the σ-algebra $\mathcal{A}_1 \otimes \mathcal{A}_2$ is generated from the set of *measurable rectangles*

$$\{E_1 \times E_2 \mid E_1 \in \mathcal{A}_1, E_2 \in \mathcal{A}_2\}.$$

This is discussed in Example 2.2.4 as the example of a product in the category of measurable spaces.

Dually, the sum $(X_1 + X_2, \mathcal{A}_1 + \mathcal{A}_2)$ of the measurable spaces $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ is defined through the final σ-algebra on the sum $X_1 + X_2$ for the injections $X_i \to X_1 + X_2$. This is the special case of the co-product $\bigoplus_{i \in I}(X_i, \mathcal{A}_i)$, where the σ-algebra $\bigoplus_{i \in I} \mathcal{A}_i$ is initial with respect to the injections. This is discussed in a general context in Example 2.2.16.

We will construct horizontal and vertical cuts from subsets of a Cartesian product, e.g., when defining the product measure and for investigating measurability properties. We define for $Q \subseteq X \times Y$ the *horizontal cut* Q_x, Q^y

$$Q_x := \{y \in Y \mid \langle x, y \rangle \in Q\}$$

and the *vertical cut*

$$Q^y := \{x \in X \mid \langle x, y \rangle \in Q\}.$$

Lemma 4.1.8 *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. If $Q \in \mathcal{A} \otimes \mathcal{B}$, then $Q_x \in \mathcal{B}$ and $Q^y \in \mathcal{A}$ hold for $x \in X, y \in Y$.*

Proof Take the vertical cut Q_x and consider the set

$$\mathcal{Q} := \{Q \in \mathcal{A} \otimes \mathcal{B} \mid Q_x \in \mathcal{B}\}.$$

Then $A \times B \in \mathcal{Q}$, whenever $A \in \mathcal{A}, B \in \mathcal{B}$; this is so since the set of all measurable rectangles forms a generator for the product σ-algebra which is closed under finite intersections. Because $(X \times Y) \setminus Q)_x = Y \setminus Q_x$, we infer that $\mathcal{Q}$ is closed under complementation, and because $(\bigcup_{n \in \mathbb{N}} Q_n)_x = \bigcup_{n \in \mathbb{N}} Q_{n,x}$, we conclude that $\mathcal{Q}$ is closed under disjoint countable unions. Hence $\mathcal{Q} = \mathcal{A} \otimes \mathcal{B}$ by the π-λ-Theorem 1.6.30. $\dashv$

The converse does not hold. We cannot conclude from the fact for a subset $S \subseteq X \times Y$ that S is product measurable whenever all its cuts are measurable; this follows from Example 4.3.7.

4.1.2 A σ-Algebra on Spaces of Measures

We will now introduce a σ-algebra on the space of all σ-finite measures. It is induced by evaluating measures at fixed events. Note the inversion:

Instead of observing a measure assigning a real number to a set, we take a set and have it act on measures. This approach is fairly natural for many applications.

In addition to $\mathbb{S}$ resp. $\mathbb{P}$, the functors which assign to each measurable space its subprobabilities and its probabilities (see Example 2.3.12), we introduce the space of finite resp. σ-finite measures.

$\mathbb{M}, \mathbb{M}_\sigma$ Denote by $\mathbb{M}(X, \mathcal{A})$ the set of all finite measures on $(X, \mathcal{A})$; the set of all σ-finite measures is denoted by $\mathbb{M}_\sigma(X, \mathcal{A})$. Each set $A \in \mathcal{A}$ gives rise to the evaluation map $ev_A : \mu \mapsto \mu(A)$; the *weak σ-algebra* $\wp(X, \mathcal{A})$ on $\mathbb{M}(X, \mathcal{A})$ is the initial σ-algebra with respect to the family $\{ev_A \mid A \in \mathcal{A}\}$; actually, it suffices to consider a generator $\mathcal{A}_0$ of $\mathcal{A}$;
$\wp(X, \mathcal{A})$ see Exercise 4.1. This is just an extension of the definitions given in Example 2.1.14. It is clear that we have

$$\wp(X, \mathcal{A}) = \sigma(\{\boldsymbol{\beta}_{\mathcal{A}}(A, \bowtie q) \mid A \in \mathcal{A}, q \in \mathbb{R}_+\})$$

when we define

$$\boldsymbol{\beta}_{\mathcal{A}}(A, \bowtie q) := \{\mu \in \mathbb{M}(X, \mathcal{A}) \mid \mu(A) \bowtie q\}.$$

$\boldsymbol{\beta}_{\mathcal{A}}(A, \bowtie q)$ Here $\bowtie$ is one of the relational operators $\leq, <, \geq, >$, and it is apparent that q may be taken from the rationals. We will use the same symbol $\boldsymbol{\beta}_{\mathcal{A}}$ when we refer to probabilities or subprobabilities, if no confusion arises. Thus the base space from which the weak σ-algebra will be constructed should be clear from the context. We will also write $\wp(X)$ or $\wp(\mathcal{A})$ for $\wp(X, \mathcal{A})$, as the situation requires.

Let $(Y, \mathcal{B})$ be another measurable space, and let $f : X \to Y$ be $\mathcal{A}$-$\mathcal{B}$-measurable. Define

$$\mathbb{M}(f)(\mu)(B) := \mu(f^{-1}\big[B\big])$$

$\mathbb{M}(f)$ for $\mu \in \mathbb{M}(X, \mathcal{A})$ and for $B \in \mathcal{B}$, then $\mathbb{M}(f)(\mu) \in \mathbb{M}(Y, \mathcal{B})$; hence $\mathbb{M}(f) : \mathbb{M}(X, \mathcal{A}) \to \mathbb{M}(Y, \mathcal{B})$ is a map, and since

$$(\mathbb{M}(f))^{-1}\big[\boldsymbol{\beta}_{\mathcal{B}}(B, \bowtie q)\big] = \boldsymbol{\beta}_{\mathcal{A}}(f^{-1}\big[B\big], \bowtie q),$$

this map is $\wp(\mathcal{A})$-$\wp(\mathcal{B})$-measurable; see Exercise 2.8. Thus $\mathbb{M}$ is an endofunctor on the category of measurable spaces.

Measurable maps into $\mathbb{M}_\sigma(\cdot)$ deserve special attention.

Definition 4.1.9 *Given measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$, an $\mathcal{A}$-$\wp(\mathcal{B})$-measurable map $K : X \to \mathbb{M}_\sigma(Y, \mathcal{B})$ is called a* transition kernel *and denoted by $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$. A transition kernel with values in $\mathbb{S}(Y, \mathcal{B})$ is also called a* stochastic relation. $X \rightsquigarrow Y$

A transition kernel $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ associates to each $x \in X$ a σ-finite measure $K(x)$ on $(Y, \mathcal{B})$. In a probabilistic setting, this may be interpreted as the probability that a system reacts on input x with $K(x)$ as the probability distribution of its responses. For example, if $(X, \mathcal{A}) = (Y, \mathcal{B})$ is the state space of a probabilistic transition system, then $K(x)(B)$ is often interpreted as the probability that the next state is a member of measurable set B after a transition from x.

This is an immediate characterization of transition kernels.

Lemma 4.1.10 *$K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ is a transition kernel iff these conditions are satisfied:*

1. *$K(x)$ is a σ-finite measure on $(Y, \mathcal{B})$ for each $x \in X$.*
2. *$x \mapsto K(x)(B)$ is a measurable function for each $B \in \mathcal{B}$.*

Proof If $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$, then $K(x)$ is a σ-finite measure on $(Y, \mathcal{B})$, and

$$\{x \in X \mid K(x)(B) > q\} = K^{-1}\big[\boldsymbol{\beta}_{\mathcal{B}}(B, > q)\big] \in \mathcal{A}.$$

Thus $x \mapsto K(x)(B)$ is measurable for all $B \in \mathcal{B}$. Conversely, if $x \mapsto K(x)(B)$ is measurable for $B \in \mathcal{B}$, then the above equation shows that $K^{-1}\big[\boldsymbol{\beta}_{\mathcal{B}}(B, > q)\big] \in \mathcal{A}$, so $K : (X, \mathcal{A}) \to \mathbb{M}_\sigma(Y, \mathcal{B})$ is $\mathcal{A}$-$\wp(\mathcal{B})$-measurable by Lemma 4.1.4. $\dashv$

A special case of transition kernels are *Markov kernels*, sometimes also called *stochastic relations*. These are kernels, the image of which is in $\mathbb{S}$ or in $\mathbb{P}$, whatever the case may be. We encountered these Markov kernels already in Example 2.4.8 as the Kleisli morphisms for the Giry monad. Kleisli

Example 4.1.11 Transition kernels may be used for interpreting modal logics. Consider this grammar for formulas

$$\varphi ::= \top \mid \varphi_1 \wedge \varphi_2 \mid \Diamond_q \varphi$$

with $q \in \mathbb{Q}, q \geq 0$. The informal interpretation in a probabilistic transition system is that $\top$ always holds and that $\Diamond_q \varphi$ holds with probability not smaller than q after a transition in a state in which formula φ holds. Now let $M : (X, \mathcal{A}) \rightsquigarrow (X, \mathcal{A})$ be a transition kernel, and define inductively

$$\begin{aligned} [\![\top]\!]_M &:= X \\ [\![\varphi_1 \wedge \varphi_2]\!]_M &:= [\![\varphi_1]\!]_M \cap [\![\varphi_2]\!]_M \\ [\![\Diamond_q \varphi]\!]_M &:= \{x \in X \mid M(x)([\![\varphi]\!]_M) \geq q\} \\ &= M^{-1}\big[\boldsymbol{\beta}_{\mathcal{A}}(B, \geq q)\big]. \end{aligned}$$

It is easy to show by induction on the structure of the formula that the sets $[\![\varphi]\!]_M$ are measurable, since M is a transition kernel. For a generalization, see Example 4.2.7. ✌

4.1.3 Case Study: Stochastic Effectivity Functions

This section will discuss stochastic effectivity functions as a generalization of stochastic transition kernels resp. stochastic relations. The reasons for this discussion are as follows:

Nondeterminism Stochastic effectivity functions are the probabilistic versions of effectivity functions which are well understood. For a historic overview, the reader is referred to [vdHP07, Sect. 9] or to [Dob14, Sect. 1]. Since these functions model nondeterminism fairly well, their stochastic extensions are also a prime candidate for modeling applications which incorporate stochastic nondeterminism. This argument is studied in depth in [DT41].

Confluence The confluence of categorical and stochastic reasoning can be studied in this instance, for example, when introducing morphisms and congruences. This cooperation is the more interesting as these stochastic effectivity functions may be thought of as the composition of monads, but seem not to be the functorial part of a monad themselves. Hence it becomes mandatory to fine-tune the arguments which are already available in the context of, say, Kleisli morphisms. Extending this argument, the approach proposed here is an exercise in stochastic modeling.

Tool Stochastic effectivity functions are proposed as the tool for interpreting game logic stochastically. It is argued below that the Kripke models are not suitable for interpreting this logic in the usual, non-stochastic realm; hence we have to look for other stochastic methods to interpret this logic. A first proposal in this direction will be discussed in Sect. 4.9.4. Thus the discussion below may also be perceived as sharpening the tools, which later on will be put to use in the context of game logics.

For an interpretation of game logic as formulated in Example 2.7.5, Parikh [Par85] use effectivity functions which are closely related to neighborhood relations [Pau00]; this is sketched in Sect. 2.7.1, in particular in Example 2.7.22. It is shown that the Kripke models are not adequate for interpreting game logics. These arguments should be taken care of for a stochastic interpretation as well, so we need an extension to the stochastic Kripke models, i.e., to Kripke models which are based on stochastic transition systems. Since effectivity functions have been shown to be useful as the formalism underlying neighborhood models, we will formulate here stochastic effectivity functions as a stochastic counterpart and for further development.

When constructing such a probabilistic interpretation, we will take distributions over the state space into account—thus, rather than working with states directly, we will work with probabilities over them. As with the interpretation of modal logics through stochastic Kripke models, it may well be that some information gets lost, so we choose to work with subprobabilities rather than probabilities. Taking into account that Angel may be able to bring about a specific distribution of the new states when playing game γ in state s, we propose that we model Angel's effectivity by a set of distributions; this is remotely similar to the idea of gambling houses in [DS65]. For example, Angel may have a strategy for achieving a normal distribution $\mathcal{N}(s, \sigma^2)$ centered at $s \in \mathbb{R}$ such that the standard distribution varies in an interval I, yielding $\{\mathcal{N}(s, \sigma^2) \mid \sigma \in I\}$ as a set of distributions effective for Angel in that situation.

But we cannot do with just arbitrary subsets of the set of all subprobabilities on state space S. We want also to characterize possible outcomes, i.e., sets of distributions over the state space for composite games. Hence we will want to average over intermediate states. This in turn requires measurability of the functions involved, both for the integrand and for the measure used for integration. Consequently we require measurable sets of subprobabilities as possible outcomes. We also impose a condition for measurability on the interplay between distributions on states and reals for measuring the probabilities of sets of states. This leads to the definition of a stochastic effectivity function.

Modeling all this requires some preparations by fixing the range of a stochastic effectivity function. Put for a measurable space $(S, \mathcal{A})$

$$\boldsymbol{E}(S, \mathcal{A}) := \{V \subseteq \wp(S, \mathcal{A}) \mid V \text{ is upper closed}\};$$

thus if $V \in \boldsymbol{E}(S, \mathcal{A})$, then $A \in V$ and $A \subseteq B$ together imply $B \in V$; note that $\boldsymbol{E} = \boldsymbol{V}' \circ \mathbb{S}$ with $\boldsymbol{V}'$ as the restriction of $\boldsymbol{V}$ to the weakly measurable sets, $\boldsymbol{V}$ being defined in Example 2.3.13. Recall from Sect. 4.1.2 that $\wp(S, \mathcal{A})$ is the weak σ-algebra on the set of finite measures on the measurable space $(X, \mathcal{A})$, i.e., the initial σ-algebra with respect to evaluation.

A measurable map $f : (S, \mathcal{A}) \to (T, \mathcal{B})$ induces a map $\boldsymbol{E}(f) : \boldsymbol{E}(S, \mathcal{A}) \to \boldsymbol{E}(T, \mathcal{B})$ upon setting

$$\boldsymbol{E}(f)(V) := \{W \in \wp(T, \mathcal{B}) \mid (\mathbb{S}f)^{-1}\big[W\big] \in V\}$$

for $V \in \boldsymbol{E}(S)$; then clearly $\boldsymbol{E}(f)(V) \in \boldsymbol{E}(T)$.

Note that $\boldsymbol{E}(S, \mathcal{A})$ has not been equipped with a σ-algebra, so the usual notion of measurability between measurable spaces cannot be applied. In particular, $\boldsymbol{E}$ is not an endofunctor on the category of measurable spaces. We will not discuss functorial aspects of $\boldsymbol{E}$ here, but rather refer the reader to the discussion in Sect. 2.3.1, in particular Example 2.3.13.

We need some measurability properties for dealing with the composition of distributions when discussing composite games. Let $H \subseteq \mathbb{S}(S) \times [0, 1]$ be a measurable subset indicating a quantitative assessment of subprobabilities; a typical example could be the set $\{\langle\mu, q\rangle \mid \mu \in \boldsymbol{\beta}_{\mathcal{A}}(A, > q), q \in C\}$ for some $A \in \mathcal{A}$ and some $C \in \mathcal{B}([0, 1])$. Fix some real q and consider the cut of H at q, viz.,

$$H^q = \{\mu \mid \langle\mu, q\rangle \in H\},$$

for example,

$$\boldsymbol{\beta}_{\mathcal{A}}(A, > q) = \{\langle\mu, r\rangle \mid \mu(A) > r\}^q.$$

We ask for all states s such that this set H^q is effective for s. They should come from a measurable subset of S. It turns out that this is not enough; we also require the real components being captured through a measurable set as well—after all, the real component will be used to be averaged, i.e., integrated, over later on, so it should behave decently. This idea is formulated in the following definition.

Definition 4.1.12 *Call a map $P : S \to \boldsymbol{E}(S)$* t-measurable *iff $\{\langle s, q\rangle \mid H^q \in P(s)\} \in \mathcal{A} \otimes [0,1]$ whenever $H \in \wp(S) \otimes [0,1]$.* t-measurable

Summarizing, we are led to the notion of a stochastic effectivity function.

Definition 4.1.13 *A* stochastic effectivity function *P on a measurable space S is a t-measurable map $P \to \boldsymbol{E}(S)$.*

In order to distinguish between sets of states and sets of state distributions, we call the latter ones *portfolios*; thus $P(s)$ is a set of measurable portfolios. This will render some discussions below easier. By the way, stochastic effectivity functions between measurable spaces S and T could be defined in a similar way, but this added generality is not of interest in the present context. Note that an effectivity function is not given by an endofunctor on the category of measurable spaces, so we will have to assemble what we need from this context without being able to directly refer to coalgebras or similar constructions. Portfolio

We show that a finite transition system can be converted into a stochastic effectivity function.

Example 4.1.14 Let $S := \{1, \dots, n\}$ for some $n \in \mathbb{N}$, and take the power set as a σ-algebra. Then $\mathbb{S}(S)$ can be identified with the compact convex set

$$\Pi_n := \{\langle x_1, \dots, x_n\rangle \mid x_i \geq 0 \text{ for } 1 \leq i \leq n, \sum_{i=1}^n x_i \leq 1\}.$$

Geometrically, Π_n is the positive convex hull of the unit vectors e_i, $1 \leq i \leq n$ and the zero vector; here $e_i(i) = 1$, and $e_i(j) = 0$ if $i \neq j$ is the ith n-dimensional unit vector. The weak σ-algebra $\wp(S)$ is the Borel-σ-algebra $\mathcal{B}(\Pi_n)$ for the Euclidean topology on Π_n.

Assume we have a transition system $\to_S$ on S; hence a relation $\to_S \subseteq S \times S$. Let $succ(s) := \{s' \in S \mid s \to_S s'\}$ be the set of a successor state for state s, and define for $s \in S$ the set of weighted successors

$$\kappa(s) := \left\{\sum_{s' \in succ(s)} \alpha_{s'} \cdot e_{s'} \mid \mathbb{Q} \ni \alpha_{s'} \geq 0 \text{ for } s' \in succ(s), \sum_{s' \in succ(s)} \alpha_{s'} \leq 1\right\}$$

and the upper closed set

$$P(s) := \{A \in \mathcal{B}(\Pi_n) \mid \kappa(s) \subseteq A\}.$$

A set A is in the portfolio for P in state s if A contains all rational subprobability distributions on the successor states of s. We will restrict our attention to these rational distributions.

We claim that P is an effectivity function on S. If $P(s) = \emptyset$, there is nothing to show, so we assume that always $P(s) \neq \emptyset$. Let $H \in \mathcal{B}(\Pi_n) \otimes \mathcal{B}([0,1]) = \mathcal{B}(\Pi_n \otimes [0,1])$, the latter equality holding by Proposition 4.3.16. Then

$$\{\langle s, q\rangle \mid H^q \in P(s)\} = \bigcup_{1 \leq s \leq n} \{s\} \times \{q \in [0,1] \mid H^q \in P(s)\}.$$

Fix $s \in S$, and let $succ(s) = \{s_1, \dots, s_m\}$. Put

$$\Omega_m := \{\langle \alpha_1, \dots, \alpha_m\rangle \in \mathbb{Q}^m \mid \alpha_i \geq 0; \sum_i \alpha_i \leq 1\},$$

hence Ω_m is countable, and

$$\begin{aligned}\{q \in [0,1] \mid H^q \in P(s)\} &= \{q \in [0,1] \mid \kappa(s) \subseteq H^q\} \\ &= \bigcap_{\langle \alpha_1, \dots, \alpha_m\rangle \in \Omega_m} \{q \in [0,1] \mid \sum_i \alpha_i \cdot e_{j_i} \in H^q\}.\end{aligned}$$

Now fix $\alpha := \langle \alpha_1, \dots, \alpha_m\rangle \in \Omega_m$. The map $\zeta_\alpha : [0,1]^{m \cdot n} \to [0,1]^n$ which maps $\langle v_1, \dots, v_m\rangle$ to $\sum_{i=1}^m \alpha_i \cdot v_i$ is continuous, hence measurable, and so is $\xi := \zeta_\alpha \times id_{[0,1]} : [0,1]^{m \cdot n} \times [0,1] \to [0,1]^n \times [0,1]$. Hence $I := \xi^{-1}\big[H\big] \in \mathcal{B}([0,1]^{m \cdot n} \times [0,1])$, and $\sum_{i=1}^m \alpha_i \cdot e_{j_i} \in H^q$ iff $\langle e_{j_1}, \dots, e_{j_m}, q\rangle \in I$. Consequently,

$$\{q \in [0,1] \mid \sum_i \alpha_i \cdot e_{j_i} \in H^q\} = I^{\langle e_{j_1}, \dots, e_{j_m}\rangle} \in \mathcal{B}([0,1]).$$

But this implies that

$$\{q \in [0,1] \mid H^q \in P(s)\} = \bigcap_{\alpha \in \Omega_m} \{q \in [0,1] \mid \sum_i \alpha_i \cdot e_{j_i} \in H^q\} \in \mathcal{B}([0,1])$$

for the fixed state $s \in S$. Collecting states, we obtain

$$\{\langle s, q\rangle \in S \times [0,1] \mid H^q \in P(s)\} \in \mathcal{P}(S) \otimes \mathcal{B}([0,1]).$$

Thus we have converted a finite transition system into a stochastic effectivity function by constructing all subprobabilities over the respective successor sets with rational coefficients.

One might ask whether the restriction to rational coefficients is really necessary. Taking the convex closure with real coefficients might, however, result in loosing measurability; see [Kec94, p. 216]. ✌

The next example shows that a stochastic effectivity function can be used for interpreting a simple modal logic.

Example 4.1.15 Let Φ be a set of atomic propositions, and define the formulas of a logic through this grammar

$$\varphi ::= \top \mid p \mid \varphi_1 \wedge \varphi_2 \mid \Diamond_q \varphi$$

with $p \in \Phi$ an atomic proposition and $q \in [0, 1]$ a threshold value. Intuitively, $\Diamond_q \varphi$ is true in a state s iff there can be a move in s to a state in which φ holds with probability not smaller than q.

This logic is interpreted over the measurable space $(S, \mathcal{A})$; assume that we are given a map $V : \Phi \to \mathcal{A}$, assigning each atomic proposition a measurable set as its validity set. Let P be a stochastic effectivity function on S, then define inductively

$$\begin{aligned}
[\![\top]\!] &:= S,\\
[\![p]\!] &:= V(p), \text{ for } p \in \Phi,\\
[\![\varphi_1 \wedge \varphi_2]\!] &:= [\![\varphi_1]\!] \cap [\![\varphi_2]\!],\\
[\![\Diamond_q \varphi]\!] &:= \{s \in S \mid \boldsymbol{\beta}_{\mathcal{A}}([\![\varphi]\!], > q) \in P(s)\}.
\end{aligned}$$

The interesting line is of course the last one. It assigns to $\Diamond_q \varphi$ all states s such that $\boldsymbol{\beta}_{\mathcal{A}}([\![\varphi]\!], > q)$ is in the portfolio of $P(s)$. These are all states for which the collection of all measures yielding an evaluation on $[\![\varphi]\!]$ greater than q can be achieved.

Then t-measurability of P and the assumption on V make sure that these sets are measurable. This is shown by an easy induction on the structure of the formulas and by observing that $[\![\Diamond_q \varphi]\!] = \{\langle s, r\rangle \in S \times [0, 1] \mid \boldsymbol{\beta}_{\mathcal{A}}([\![\varphi]\!], > r) \in P(s)\}^q$. ✌

We fix for the time being a measurable space $(S, \mathcal{A})$; if a second measurable space enters the discussion, it will be T as carrier with σ-algebra $\mathcal{B}$.

The relationship of stochastic relations and stochastic effectivity functions is of considerable interest; we will discuss stochastic Kripke models and general models for game logic in Sect. 4.9.4. Each stochastic relation $K : S \rightsquigarrow S$ yields a stochastic effectivity function P_K in a natural way upon setting

$$P_K(s) := \{A \in \wp(S) \mid K(s) \in A\}. \tag{4.1}$$

Thus a portfolio in $P_K(s)$ is a measurable subset of $\mathbb{S}(S)$ which contains $K(s)$. We observe

Lemma 4.1.16 *$P_K : S \to \boldsymbol{E}(S)$ is t-measurable, whenever $K : S \rightsquigarrow S$ is a stochastic relation.*

Proof Clearly, $P_K(s)$ is upper closed for each $s \in S$. Put $T_H := \{\langle s, q\rangle \mid H^q \in P_K(s)\}$ for $H \subseteq \mathbb{S}(S) \times [0, 1]$. Thus

$$\langle s, q\rangle \in T_H \Leftrightarrow K(s) \in H^q \Leftrightarrow \langle K(s), q\rangle \in H \Leftrightarrow \langle s, q\rangle \in (K \times id_{[0,1]})^{-1}\left[H\right].$$

Because $K \times id_{[0,1]} : S \times [0, 1] \to \mathbb{S}(S) \times [0, 1]$ is a measurable function, $H \in \wp(S) \otimes [0, 1]$ implies $T_H \in \mathcal{B}(S \otimes [0, 1])$. Hence P_K is t-measurable. ⊣

This argument extends easily to countable families of stochastic relations.

Corollary 4.1.17 *Assume $\mathcal{F} = \{K_n \mid n \in \mathbb{N}\}$ is a countable family of stochastic relations $K_n : S \rightsquigarrow S$; then*

$$s \mapsto \{A \in \wp(S) \mid K_n(s) \in A \textit{ for some } n \in \mathbb{N}\},$$
$$s \mapsto \{A \in \wp(S) \mid K_n(s) \in A \textit{ for all } n \in \mathbb{N}\}$$

define stochastic effectivity functions on S. ⊣

The following example will be of use later on. It shows that we have always a stochastic effectivity function at our disposal, albeit a fairly trivial one.

Example 4.1.18 Let $D : S \rightsquigarrow S$ be the Dirac relation $x \mapsto \delta_x$ with δ_x as the Dirac measure on x (see Example 1.6.13); then $I_D := P_D$ defines an effectivity function, the *Dirac effectivity function*. Consequently we have $W \in I_D(s)$ iff $\delta_s \in W$ for $W \in \wp(S)$. This is akin to assigning each element of a set the ultrafilter based on it. ✌

The Dirac effectivity function will be useful for characterizing the effect of the empty game ϵ, and it will also help in modeling the effects of the test games associated with formula φ; see Sect. 4.9.4.

We will use the construction indicated through (4.1) for generating a model from a stochastic Kripke model, indicating that the models considered here are more general than Kripke models. The converse construction is of course of interest as well: Given a model, can we determine whether or not it comes from a Kripke model? This boils down to the question under which conditions a stochastic effectivity function is generated through a stochastic relation. We will deal with this problem now. The tools for investigating the converse to Lemma 4.1.16 come from the investigation of deduction systems for probabilistic logics.

Characteristic Relations

In fact, we are given a set of portfolios and want to know under which conditions this set is generated from a single subprobability. The situation is roughly similar to the one observed with deduction systems, where a set of formulas is given, and one wants to know whether this set can be constructed as valid under a suitable model. Because of the similarity, we may take some inspiration from the work on deduction systems, and we adapt here the approach proposed by Goldblatt [Gol10]. Goldblatt works with sets of formulas while we are interested foremost in families of sets; this permits a technically somewhat lighter approach in the present scenario.

We first have a look at a relation $R \subseteq [0,1] \times \mathcal{A}$ which models bounding probabilities from below. Intuitively, $\langle r, A\rangle \in R$ is intended to characterize a member of the set $\boldsymbol{\beta}_{\mathcal{A}}(A, \geq r)$, i.e., a measure μ with $\mu(A) \geq r$.

Definition 4.1.19 *$R \subseteq [0,1] \times \mathcal{A}$ is called a* characteristic relation *on S iff these conditions are satisfied :* Characteristic relation

① $\dfrac{\langle r, A\rangle \in R, A \subseteq B}{\langle r, B\rangle \in R}$ ② $\dfrac{\langle r, A\rangle \in R, r \geq s}{\langle s, A\rangle \in R}$

③ $\dfrac{\langle r, A\rangle \notin R, \langle s, B\rangle \notin R, r+s \leq 1}{\langle r+s, A \cup B\rangle \notin R}$ ④ $\dfrac{\langle r, A \cup B\rangle \in R, \langle s, A \cup (S \setminus B)\rangle \in R, r+s \leq 1}{\langle r+s, A\rangle \in R}$

⑤ $\dfrac{\langle r, A\rangle \in R, r+s > 1}{\langle s, S \setminus A\rangle \notin R}$ ⑥ $\dfrac{\langle r, \emptyset\rangle \in R}{r = 0}$

⑦ $\dfrac{A_1 \supseteq A_2 \supseteq \ldots, \forall n \in \mathbb{N} : \langle r, A_n\rangle \in R}{\langle r, \bigcap_{n \geq 1} A_n\rangle \in R}$.

The notation adopted here for convenience and for conciseness is that of deduction systems. For example, condition ① expresses that $\langle r, A\rangle \in R$ and $A \subseteq B$ together imply $\langle r, B\rangle \in R$. The interpretations are as follows. The conditions ① and ② make sure that bounding from below is monotone both in its numeric and in its set-valued component. By ③ and ④, we cater for sub- and superadditivity of the characteristic relation, condition ⑥ sees to the fact that the probability for the impossible event cannot be bounded from below but through 0, and finally ⑦ makes sure that if the members of a decreasing sequence of sets are uniformly bounded below, then so is its intersection.

We show that each characteristic relation defines a subprobability measure; the proof follows *mutatis mutandis*, the proof of [Gol10, Theorem 5.4].

Proposition 4.1.20 *Let $R \subseteq [0,1] \times \mathcal{A}$ be a characteristic relation on S, and define for $A \in \mathcal{A}$*

$$\mu_R(A) := \sup\{r \in [0,1] \mid \langle r, A\rangle \in R\}.$$

Then μ_R is a subprobability measure on $\mathcal{A}$.

Proof 1. ⑥ implies that $\mu_R(\emptyset) = 0$, and μ_R is monotone because of ①. It is also clear that $\mu_R(S) \leq 1$. We obtain from ② that $\langle s, A\rangle \notin R$, whenever $s \geq r$ with $\langle r, A\rangle \notin R$.

2. Let $A_1, A_2 \in \mathcal{A}$ be arbitrary. Then

$$\mu_R(A_1 \cup A_2) \leq \mu_R(A_1) + \mu_R(A_2).$$

In fact, if $\mu_R(A_1)+\mu_R(A_2) < q_1+q_2 \leq \mu_R(A_1 \cup A_2)$ with $\mu_R(A_i) < q_i$ $(i = 1, 2)$, then $\langle q_i, A_i\rangle \notin R$ for $i = 1, 2$. Because $q_1 + q_2 \leq 1$, we obtain from ③ that $\langle q_1 + q_2, A_1 \cup A_2\rangle \notin R$. By ② this yields $\mu_R(a_1 \cup A_2) < q_1 + q_2$, contradicting the assumption.

3. If A_1 and A_2 are disjoint, we observe first that $\mu_R(A_1) + \mu_R(A_2) \leq 1$. Assume otherwise that we can find $q_i \leq \mu_R(A_i)$ for $i = 1, 2$ with $q_1 + q_2 > 1$. Because $\langle q_1, A_1\rangle \in R$, we conclude from ⑤ that $\langle q_2, S \setminus A_2\rangle \notin R$; hence $\langle q_2, A_2\rangle \notin R$ by ①, contradicting $q_2 \leq \mu_R(A_2)$.

This implies that

$$\mu_R(A_1) + \mu_R(A_2) \leq \mu_R(A_1) + \mu_R(A_2).$$

Assuming this to be false, we find $q_1 \leq \mu_R(A_1), q_2 \leq \mu_R(A_2)$ with

$$\mu_R(A_1 \cup A_2) < q_1 + q_2 \leq \mu_R(A_1) + \mu_R(A_2).$$

Because $\langle q_1, A_1\rangle \in R$, we find $\langle q_1, (A_1 \cup A_2) \cap A_1\rangle \in R$, *and* because $\langle q_2, A_2\rangle \in R$, we see that $\langle q_2, (A_1 \cup A_2) \cap (S \setminus A_1)\rangle \in R$ (note that $(A_1 \cup A_2) \cap A_1 = A_1$ and $(A_1 \cup A_2) \cap (S \setminus A_1) = A_2$, since $A_1 \cap A_2 = \emptyset$). From ④ we infer that $\langle q_1 + q_2, A_1 \cup A_2 \rangle \in R$, so that $q_1 + q_2 \leq \mu_R(A_1 \cup A_2)$, which is a contradiction.

Thus we have shown that μ_R is additive.

4. From ⑦ it is obvious that

$$\mu_R(A) = \inf_{n \in \mathbb{N}} \mu_R(A_n),$$

whenever $A = \bigcap_{n \in \mathbb{N}} A_n$ for the decreasing sequence $(A_n)_{n \in \mathbb{N}}$ in $\mathcal{A}$. $\dashv$

Thus we are in a position now to relationally characterize a subprobability on $\mathcal{A}$. But we can even say more, when taking subsets of $\wp(S)$ into account. Note that $A \mapsto \boldsymbol{\beta}_{\mathcal{A}}(A, \geq q)$ is monotone for each q. So if we fix an upper closed subset $Q \subseteq \wp(S)$, then we know that $\boldsymbol{\beta}_{\mathcal{A}}(A, \geq q) \in Q$ implies $\boldsymbol{\beta}_{\mathcal{A}}(B, \geq q) \in Q$, provided $A \subseteq B$. We relate Q to a characteristic relation R on S by comparing $\boldsymbol{\beta}_{\mathcal{A}}(A, \geq q) \in Q$ with $\langle q, A \rangle \in R$ by imposing a syntactic and a semantic condition. They will be shown to be equivalent.

Definition 4.1.21 *The upper closed subset Q of $\wp(S)$ is said to* satisfy *the characteristic relation R on S ($Q \vdash R$) iff we have* $Q \vdash R$

$$\langle q, A \rangle \in R \Leftrightarrow \boldsymbol{\beta}_{\mathcal{A}}(A, \geq q) \in Q$$

for any $q \in [0, 1]$ and any $A \in \mathcal{A}$.

This is a syntactic notion: We look at R and determine from the knowledge of $\langle q, A \rangle \in R$ whether all evaluations on the measurable set A are contained in Q. Its semantic counterpart reads like this.

Definition 4.1.22 *Q is said to* implement *$\mu \in \mathbb{S}(S)$ iff*

$$\mu(A) \geq q \Leftrightarrow \boldsymbol{\beta}_{\mathcal{A}}(A, \geq q) \in Q$$

for any $q \in [0, 1]$ and any $A \in \mathcal{A}$. We write this as $Q \models \mu$. $Q \models \mu$

Thus we actually evaluate μ at A and determine from this value the membership of $\boldsymbol{\beta}_{\mathcal{A}}(A, \geq q)$ in Q.

Note that $Q \models \mu$ and $Q \models \mu'$ implies

$$\forall A \in \mathcal{A} \forall q \geq 0 : \mu(A) \geq q \Leftrightarrow \mu'(A) \geq q.$$

Consequently, $\mu = \mu'$, so that the measure implemented by Q is uniquely determined.

We will show now that syntactic and semantic issues are equivalent: Q satisfies a characteristic relation if and only if it implements the corresponding measure.

Proposition 4.1.23 *$Q \vdash R$ iff $Q \models \mu_R$.*

Proof $Q \vdash R \Rightarrow Q \models \mu_R$: Assume that $Q \vdash R$ holds. It is then immediate that $\mu_R(A) \geq r$ iff $\boldsymbol{\beta}_{\mathcal{A}}(A, \geq r) \in Q$.

$Q \models \mu_R \Rightarrow Q \vdash R$: If $Q \models \mu_R$ for relation $R \subseteq [0,1] \times \mathcal{A}$, we show that the conditions given in Definition 4.1.19 are satisfied.

1. Let $\boldsymbol{\beta}_{\mathcal{A}}(A, \geq r) \in Q$ and $A \subseteq B$; thus $\mu_R(A) \geq r$; hence $\mu_R(B) \geq r$, which in turn implies $\boldsymbol{\beta}_{\mathcal{A}}(A, \geq r) \in Q$. Hence ① holds. ② is established similarly.

2. If $\mu_R(A) < r$ and $\mu_R(B) < s$ with $r + s \leq 1$, then $\mu_R(A \cup B) = \mu_R(A) + \mu(B) - \mu_R(A \cap B) \leq \mu_R(A) + \mu_R(B) < r + s$, which implies ③.

3. If $\mu_R(A \cup B) \geq r$ and $\mu_R(A \cup (S \setminus B)) \geq s$, then $\mu_R(A) = \mu_R(A \cup B) + \mu_R(A \cup (S \setminus B)) \geq r + s$; hence ④.

4. Assume $\mu_R(A) \geq r$ and $r + s > 1$; then $\mu_R(S \setminus A) = \mu_R(S) - \mu_R(A) < p$, and thus ⑤ holds.

5. If $\mu_R(\emptyset) \geq r$, then $r = 0$, yielding ⑥.

6. Finally, if $(A_n)_{n \in \mathbb{N}}$ is decreasing with $\mu_R(A_n) \geq r$ for each $n \in \mathbb{N}$, then it is plain that $\mu_R(\bigcap_{n \in \mathbb{N}} A_n) \geq r$. This implies ⑦.

$\dashv$

This permits a complete characterization of those stochastic effectivity functions which are generated through stochastic relations.

Proposition 4.1.24 *Let P be a stochastic effectivity frame on state space S. Then these conditions are equivalent:*

1. *There exists a stochastic relation $K : S \rightsquigarrow S$ such that $P = P_K$.*

2. *$R(s) := \{\langle r, A \rangle \mid \boldsymbol{\beta}_{\mathcal{A}}(A, \geq r) \in P(s)\}$ defines a characteristic relation on S with $P(s) \vdash R(s)$ for each state $s \in S$.*

Proof 1 $\Rightarrow$ 2: Fix $s \in S$. Because $\boldsymbol{\beta}_{\mathcal{A}}(A, \geq r) \in P_K(s)$ iff $K(s)(A) \geq r$, we see that $P(s) \models K(s)$; hence by Proposition 4.1.23 $P(s) \vdash R(s)$.

2 $\Rightarrow$ 1: Define $K(s) := \mu_{R(s)}$, for $s \in S$; then $K(s)$ is a subprobability measure on $\mathcal{A}$. We show that $K : S \rightsquigarrow S$. Let $G \subseteq \mathbb{S}(S)$ be a measurable set, then $G \times [0,1] \in \mathcal{A} \otimes [0,1]$; hence the measurability condition on P yields that

$$K^{-1}\left[G\right] = \{s \in S \mid K(s) \in G\} = \{s \in S \mid G \in P(s)\}$$

is a measurable subset of S, because

$$\{\langle s, q\rangle \mid (G \times [0,1])^q \in P(s)\} = \{s \in S \mid G \in P(s)\} \times [0,1] \in \mathcal{A} \otimes [0,1].$$

This establishes measurability. $\dashv$

Morphisms and Congruences

Morphisms for stochastic effectivity functions are defined in a way very similar to the definition of morphisms for the functor $\boldsymbol{V}$ in Example 2.3.13, or for the definition of neighborhood frames; see Definition 2.7.33. We have, however, to take into consideration again that we are dealing with the respective subprobabilities as an intermediate layer. Hence, if we consider a measurable map $f : S \to T$, we must use the induced map $\mathbb{S}(f) : \mathbb{S}(S) \to \mathbb{S}(T)$ as an intermediary. This leads to the following:

Definition 4.1.25 *Given stochastic effectivity functions P on S and Q on T, a measurable map $f : S \to T$ is called a* morphism of effectivity functions $f : P \to Q$ *iff $Q \circ f = \boldsymbol{E}(f) \circ P$, hence iff this diagram commutes*

$$\begin{tikzcd} S \arrow[r, "f"] \arrow[d, "P"'] & T \arrow[d, "Q"] \\ \boldsymbol{E}(S) \arrow[r, "\boldsymbol{E}(f)"'] & \boldsymbol{E}(T) \end{tikzcd}$$

Thus we have

$$W \in Q(f(s)) \Leftrightarrow (\mathbb{S}f)^{-1}\left[W\right] \in P(s) \tag{4.2}$$

for all states $s \in S$ and for all $W \in \wp(T)$.

In comparison, recall from Definition 2.6.43 that a measurable map $f : S \to T$ is a *morphism of stochastic relations* $f : K \to L$ for the stochastic relations $K : S \rightsquigarrow S$ and $L : T \rightsquigarrow T$ iff $L \circ f = \mathbb{S}(f) \circ K$, hence iff this diagram commutes

$$\begin{tikzcd}
S \arrow[r, "f"] \arrow[d, "K"'] & T \arrow[d, "L"] \\
\mathbb{S}(S) \arrow[r, "\mathbb{S}(f)"'] & \mathbb{S}(T)
\end{tikzcd}$$

Thus

$$L(f(s))(B) = \mathbb{S}(f)(K(s))(B) = K(s)(f^{-1}\left[B\right]) \tag{4.3}$$

for each state $s \in S$ and each measurable set $B \subseteq T$.

These notions of morphisms are compatible: Each morphism for stochastic relations turns into a morphism for the associated effectivity function.

Proposition 4.1.26 *A morphism $f : K \to L$ for stochastic relations K and L induces a morphism $f : P_K \to P_L$ for the associated stochastic effectivity functions.*

Proof Fix a state $s \in S$. Then $W \in P_L(f(s))$ iff $L(f(s)) \in W$. Because $f : K \to L$ is a morphism, this is equivalent to $\mathbb{S}(f)(K(s)) \in W$, hence to $K(s) \in (\mathbb{S}f)^{-1}\left[W\right]$; thus $(\mathbb{S}f)^{-1}\left[W\right] \in P_K(s)$. $\dashv$

We investigate congruences for stochastic effectivity functions next. Since we have introduced a quantitative component into the argumentation, we want to deal not only with equivalent element of the set S on which the effectivity function is defined, but we have also to take the elements of $[0, 1]$ into account. The most straightforward equivalence relation of $[0, 1]$ is the identity $\Delta := \Delta_{[0,1]}$. Given an equivalence relation ρ on S, define by

$\boldsymbol{\Sigma}_\rho(\mathcal{A})$

$$\boldsymbol{\Sigma}_\rho(\mathcal{A}) := \{A \in \mathcal{A} \mid A \text{ is } \rho\text{-invariant}\}$$

all ρ-*invariant* measurable subsets of S, i.e., all $A \in \mathcal{A}$ which are unions of ρ-equivalence classes.

Definition 4.1.27 *Call the equivalence relation ρ on S* tame *iff*

$$\boldsymbol{\Sigma}_{\rho\times\Delta}(\mathcal{A} \otimes \mathcal{B}([0, 1])) = \boldsymbol{\Sigma}_\rho(\mathcal{A}) \otimes \mathcal{B}([0, 1])$$

holds.

Because $\boldsymbol{\Sigma}_{\Delta}(\mathcal{B}([0,1])) = \mathcal{B}([0,1])$, we rephrase the definition that ρ is tame iff

$$\boldsymbol{\Sigma}_{\rho\times\Delta}(\mathcal{A}\otimes\mathcal{B}([0,1])) = \boldsymbol{\Sigma}_{\rho}(\mathcal{A})\otimes\boldsymbol{\Sigma}_{\Delta}(\mathcal{B}([0,1])).$$

Thus being tame means for an equivalence relation that it cooperates well with the identity on $[0,1]$. From a structural point of view, tameness permits us to find a Borel isomorphism between $S/\rho\otimes[0,1]$ and $S\otimes[0,1]/\rho\times\Delta$, which gives further insight into the cooperation of ρ and Δ on $S\times[0,1]$. Here we go.

Lemma 4.1.28 *Assume that ρ is a tame equivalence relation on S; the measurable spaces $(S\otimes[0,1])/\rho\times\Delta$ and $S/\rho\otimes[0,1]$ are Borel isomorphic.*

Proof 0. We define the Borel isomorphism ϑ through $[\langle x,q\rangle]_{\rho\times\Delta}\mapsto\langle[x]_{\rho},q\rangle$, which is the obvious choice. It is not difficult to see that ϑ is a bijection and that it is measurable. The converse direction is a bit more cumbersome and will be dealt with through the principle of good sets via the π-λ-Theorem and the defining property of tameness for ρ. Outline

1. We show that

$$\vartheta:\begin{cases}(S\otimes[0,1])/\rho\times\Delta & \to S/\rho\otimes[0,1]\\ [\langle s,t\rangle]_{\rho\times\Delta} & \mapsto \langle[s]_{\rho},t\rangle\end{cases}$$

defines a measurable map. Let $A\times B\subseteq S/\rho\otimes[0,1]$ be a measurable rectangle, $\eta_{\rho}^{-1}\big[A\big]\in\mathcal{A}$ and $B\in\mathcal{B}([0,1])$; then $\eta_{\rho\times\Delta}^{-1}\big[\vartheta^{-1}\big[A\times B\big]\big] = \eta_{\rho}^{-1}\big[A\big]\times B\in\mathcal{A}\otimes\mathcal{B}([0,1])$. Hence the inverse image of a generator of the σ-algebra on $S/\rho\otimes[0,1]$ is a measurable subset of the factor space of $\mathcal{A}\otimes\mathcal{B}([0,1])$ under the equivalence relation $\rho\times\Delta$, so that ϑ is measurable by Lemma 4.1.4.

2. For establishing that ϑ^{-1} is measurable, one notes first that $B\in\boldsymbol{\Sigma}_{\rho}(S)\otimes\mathcal{B}([0,1])$ implies that $(\eta_{\rho}\times id)\big[B\big]$ is a measurable subset of $S/\rho\times[0,1]$. This is so because the set of all B for which the assertion is true is closed under complementation and countable disjoint unions, and it contains all measurable rectangles, so the assertion is established by Theorem 1.6.30.

Now take $H\subseteq(S\otimes[0,1])/\rho\times\Delta$ measurable; then

$$\eta_{\rho\times\Delta}^{-1}\big[H\big]\in\boldsymbol{\Sigma}_{\rho\times\Delta}(S\otimes[0,1]) = \boldsymbol{\Sigma}_{\rho}(S)\otimes\mathcal{B}([0,1]),$$

since ρ is tame. Hence $(\eta_\rho \times id)^{-1}\big[\vartheta\big[H\big]\big] \in \boldsymbol{\Sigma}_\rho(S) \otimes \mathcal{B}([0,1])$, so the assertion follows from

$$\vartheta\big[H\big] = (\eta_\rho \times id)\big[(\eta_\rho \times id)^{-1}\big[\vartheta\big[H\big]\big]\big].$$

$\dashv$

Final maps (Definition 2.6.42) provide a rich source for tame relations.

Lemma 4.1.29 *Let $f : S \to T$ be surjective and measurable such that $f \times id : S \times [0,1] \to T \times [0,1]$ is final. Then* $\ker(f)$ *is tame.*

Proof Because $f \times id$ is surjective and final, we infer from Lemma 2.6.44

$$(f \times id)^{-1}\big[\mathcal{B} \otimes \mathcal{B}([0,1])\big] = \boldsymbol{\Sigma}_{\ker(f)\times\Delta}(\mathcal{A} \otimes \mathcal{B}([0,1])).$$

But since $f \times id$ is final, f is final as well; hence $f^{-1}\big[\mathcal{B}\big] = \boldsymbol{\Sigma}_{\ker(f)}(\mathcal{A})$, again by Lemma 2.6.44. Thus

$$(f \times id)^{-1}\big[\mathcal{B} \otimes \mathcal{B}([0,1])\big] = f^{-1}\big[\mathcal{B}\big] \otimes \mathcal{B}([0,1]).$$

This establishes the claim. $\dashv$

Morphisms and congruences are as usual quite closely related, so we are in a position now to characterize congruences through morphisms and factorization as those equivalence relations which are related to the structure of the underlying system. Let P be a stochastic effectivity function on S. Congruences are defined in this way.

Definition 4.1.30 *The equivalence relation ρ on S is called a* congruence for P *iff there exists an effectivity function P_ρ on S/ρ which renders this diagram commutative:*

$$\begin{tikzcd} S \arrow[r, "\eta_\rho"] \arrow[d, "P"'] & S/\rho \arrow[d, "P_\rho"] \\ \boldsymbol{E}(S) \arrow[r, "\boldsymbol{E}(\eta_\rho)"'] & \boldsymbol{E}(S/\rho) \end{tikzcd}$$

Because η_ρ is onto, P_ρ is uniquely determined. The next proposition provides a criterion for an equivalence relation to be a congruence. It requires the equivalence relation to be tame, so that quantitative aspects are being taken care of. The factor space S/ρ will be equipped with the final σ-algebra with respect to the factor map η_ρ, on which we will have to consider the weak σ-algebra $\wp(S/\rho)$.

Proposition 4.1.31 *Let ρ be a tame equivalence relation on S. Then these statements are equivalent:*

1. *ρ is a congruence for P.*
2. *Whenever $s\ \rho\ s'$, we have $(\mathbb{S}\eta_\rho)^{-1}\big[A\big] \in P(s)$ iff $(\mathbb{S}\eta_\rho)^{-1}\big[A\big] \in P(s')$ for every $A \in \wp(S/\rho)$.*

The second property can be read off the diagram above, so one might ask why this property is singled out as characteristic. Well, we have to define a stochastic effectivity function of the factor set, and for this to work, we need t-measurability. The proof shows where the crucial point is. Note

Proof 1 $\Rightarrow$ 2: This follows immediately from the definition of a morphism; see (4.2).

2 $\Rightarrow$ 1: Define for $s \in S$

$$Q([s]_\rho) := \{A \in \wp(S/\rho) \mid (\mathbb{S}\eta_\rho)^{-1}\big[A\big] \in P(s)\},$$

then Q is well defined by the assumption, and it is clear that $Q([s]_\rho)$ is an upper closed set of subsets of $\wp(S/\rho)$ for each $s \in S$. It remains to be shown that Q is a stochastic effectivity function, i.e., that Q is t-measurable. This is the crucial property. In fact, let $H \in \mathcal{B}(\mathbb{S}(S/\rho) \otimes [0,1])$ be a test set, and let $G := (\mathbb{S}(\eta_\rho) \times id_{[0,1]})^{-1}\big[H\big]$ be its inverse image under $\mathbb{S}(\eta_\rho) \times id_{[0,1]}$; then Crucial

$$\{\langle t, q\rangle \in S/\rho \times [0,1] \mid H^q \in Q(t)\} = (\eta_\rho \times id_{[0,1]})\big[Z\big]$$

with $Z := \{\langle s, q\rangle \in S \times [0,1] \mid G^q \in P(s)\}$. By Lemma 4.1.28 it is enough to show that Z is contained in $\boldsymbol{\Sigma}_\rho(S) \otimes \mathcal{B}([0,1])$. Because P is t-measurable, we infer $Z \in \mathcal{B}(S \otimes [0,1])$, and because Z is $(\rho \times \Delta)$-invariant, we conclude that $Z \in \boldsymbol{\Sigma}_{\rho\times\Delta}(S \otimes [0,1])$, the latter σ-algebra being equal to $\boldsymbol{\Sigma}_\rho(S) \otimes \mathcal{B}([0,1])$ by definition of tameness. $\dashv$

The condition on subsets of $\wp(S/\rho)$ imposed above asks for measurable subsets of $\mathbb{S}(S/\rho)$, so factorization is done "behind the curtain" of functor $\mathbb{S}$. It would be more convenient if the space of all subprobabilities could be factored and the corresponding measurable sets gained from the latter space.

The following sketch provides an alternative. Lift equivalence ρ on S to an equivalence $\bar{\rho}$ on $\mathbb{S}(S)$ upon setting

$$\mu\ \bar{\rho}\ \mu' \text{ iff } \forall C \in \boldsymbol{\Sigma}_\rho(S) : \mu(C) = \mu'(C),$$

so measures are considered $\bar{\rho}$-equivalent iff they coincide on the ρ-invariant measurable sets. Define the map ∂_ρ through

$$\begin{cases} \mathbb{S}(S)/\bar{\rho} & \to \mathbb{S}(S/\rho) \\ \partial_\rho([\mu]_{\bar{\rho}}) & \mapsto \lambda G.\mathbb{S}(\eta_\rho)(\mu)(G). \end{cases}$$

Then $\partial_\rho \circ \eta_{\bar{\rho}} = \mathbb{S}(\eta_\rho)$, so that ∂_ρ is measurable by finality of $\eta_{\bar{\rho}}$. If S is a Polish space and ρ is countably generated (thus $\boldsymbol{\Sigma}_\rho(S) = \sigma(\{A_n \mid n \in \mathbb{N}\})$ for some sequence $(A_n)_{n\in\mathbb{N}}$ with measurable A_n), then it can be shown through Souslin's Theorem 4.4.8 that ∂_ρ is an isomorphism. Hence it is in this case sufficient to focus on the sets $\eta_{\bar{\rho}}^{-1}\left[W\right]$ with $W \in \mathcal{B}(\mathbb{S}(S)/\bar{\rho})$. But, as said above, this is a sketch in which many things have to be filled in.

The relationship of morphisms and congruences through the kernel is characterized in the following proposition. It assumes the morphism combined with the identity on [0, 1] to be final. This is a technical condition rendering the kernel of the morphism a tame equivalence relation.

Proposition 4.1.32 *Given a morphism $f : P \to Q$ for the effectivity functions P and Q over the state spaces S resp. T, if $f \times id_{[0,1]} : S \times [0,1] \to T \times [0,1]$ is final, then* $\ker(f)$ *is a congruence for P.*

Idea for the proof

Proof 0. We have to show that the condition in Proposition 4.1.31 is satisfied. The key idea is that we can find for a given $H_0 \in \wp(S/\ker(f))$ a set $H \in \wp(T)$ which helps represent H_0. But we do not take the representation through $\mathbb{S}f$ into account, but factor the space first through $\ker(f)$ and use the corresponding factorization $\tilde{f} \circ \eta_{\ker(f)}$. Using $\tilde{f}$ is helpful because the factor $\tilde{f}$ transports measurable sets in a somewhat advantageous manner. The set H is then used as a stand-in for H_0, and because we know where it comes from, we exploit its properties.

1. We know from Lemma 4.1.29 that $\ker(f)$ is a tame equivalence relation.

2. Given $H_0 \in \wp(S/\ker(f))$, we claim that we can find $H \in \wp(T)$ such that $H_0 = \mathbb{S}(\tilde{f})^{-1}\left[H\right]$, $f = \tilde{f} \circ \eta_{\ker(f)}$ being the decomposition of f according to Exercise 2.25. In fact, put

$$\mathcal{Z} := \left\{H_0 \in \wp(S/\ker(f)) \mid \exists H \in \wp(T) : H_0 = \mathbb{S}(\tilde{f})^{-1}\left[H\right]\right\}.$$

Then $\mathcal{Z}$ is a σ-algebra, because $\emptyset \in \mathcal{Z}$ and it is closed under the countable Boolean operations as well as complementation. Take an element of the basis for the weak σ-algebra, say, $\boldsymbol{\beta}_{S/\ker(f)}(A, \geq q)$, with $A \subseteq S/\ker(f)$ measurable. Because $\tilde{f}^{-1}$ is onto, there exists $B \subseteq T$ with $A = \tilde{f}\left[B\right]$, and we know that $B \in \mathcal{B}(T)$, because f is final. Consequently, $\mu(A) = \mathbb{S}(\tilde{f})(\mu)(B)$ for any $\mu \in \mathbb{S}(S/\ker(f))$. This implies

$$\boldsymbol{\beta}_{S/\ker(f)}(A, \geq q) = \mathbb{S}(\tilde{f})^{-1}\left[\boldsymbol{\beta}_{\mathcal{B}}(B, \geq q)\right] \in \mathcal{Z};$$

consequently,

$$\wp(S/\ker(f)) = \sigma(\{\boldsymbol{\beta}_{S/\ker(f)}(A, \geq q) \mid A \in \wp(S/\ker(f)), q \geq 0\}) \subseteq \mathcal{Z};$$

thus $\wp(S/\ker(f)) = \mathcal{Z}$.

3. Now let $f(s) = f(s')$, take $H_0 \in \wp(S/\ker(f))$, and choose $H \in \wp(T)$ according to part 1 for H_0; then

$$\begin{aligned}\eta_{\ker(f)}^{-1}\left[H_0\right] \in P(s) \quad &\Leftrightarrow f^{-1}\left[H\right] \in P(s) \quad &&\Leftrightarrow H \in Q(f(s)) = Q(f(s')) \\ &\Leftrightarrow f^{-1}\left[H\right] \in P(s') \quad &&\Leftrightarrow \eta_{\ker(f)}^{-1}\left[H_0\right] \in P(s')\end{aligned}$$

because $f : P \to Q$ is a morphism. This is what we want. $\dashv$

Let us turn to the case of stochastic relations and indicate the relationship of congruences and the effectivity functions generated through the factor relation. A *congruence ρ for a stochastic relation $K : S \rightsquigarrow S$* is an equivalence relation with this property: There exists a (unique) stochastic relation $K_\rho : S/\rho \to S/\rho$ such that this diagram commutes:

$$\begin{tikzcd} S \arrow[r, "\eta_\rho"] \arrow[d, "K"'] & S/\rho \arrow[d, "K_\rho"] \\ \mathbb{S}(S) \arrow[r, "\mathbb{S}(\eta_\rho)"'] & \mathbb{S}(S/\rho) \end{tikzcd}$$

This is the direct translation of Definition 2.6.40 in Sect. 2.6.2. The diagram translates to

$$K_\rho([s]_\rho)(B) = K(s)(\eta_\rho^{-1}\left[B\right])$$

for all $s \in S$ and all $B \subseteq S/\rho$ measurable.

We obtain from Proposition 4.1.26

Corollary 4.1.33 *A congruence ρ for a stochastic relation $K : S \rightsquigarrow S$ is also a congruence for the associated effectivity function P_K. Moreover, $P_{K_\rho} = (P_K)_\rho$, so the effectivity function associated with the factor relation K_ρ is the factor relation of P_K with respect to ρ.* $\dashv$

It is noted that we do not require additional assumptions on the congruence for the stochastic relation for being a congruence for the associated effectivity function. This indicates that the condition on tameness captures the general class of effectivity functions, but that subclasses may impose their own conditions. It indicates also that the condition of being a congruence for a stochastic relation itself is a fairly strong one when assessed by the rules pertaining to stochastic effectivity functions.

4.1.4 The Alexandrov Topology on Spaces of Measures

Given a topological space (X, τ), the Borel sets $\mathcal{B}(X) = \sigma(\tau)$ and the Baire sets $\mathcal{B}a(X)$ come for free as measurable structures: $\mathcal{B}(\tau)$ is the smallest σ-algebra on X that contains the open sets; measurability of maps with respect to the Borel sets is referred to as *Borel measurability*. $\mathcal{B}a(X)$ is the smallest σ-algebra on X which contains the functionally closed sets; they provide yet another measurable structure on (X, τ), this time involving the continuous real-valued functions. Since $\mathcal{B}(X) = \mathcal{B}a(X)$ for a metric space X by Example 4.1.1, the distinction between these σ-algebras vanishes, and the Borel sets as the σ-algebra generated by the open sets dominate the scene.

We will now define a topology of spaces of measures on a topological space in a similar way and relate this topology to the weak σ-algebra, for the time being in a special case. Fix a Hausdorff space (X, τ); the space will be specialized as the discussion proceeds. Define for the functionally open set G, the functionally closed sets F and $\epsilon > 0$ for $\mu_0 \in \mathbb{M}(X, \mathcal{B}a(X))$ the sets

$$W_{G,\epsilon}(\mu_0) := \{\mu \in \mathbb{M}(X, \mathcal{B}a(X)) \mid \mu(G) > \mu_0(G) - \epsilon, |\mu(X) - \mu_0(X)| < \epsilon\},$$
$$W_{F,\epsilon}(\mu_0) := \{\mu \in \mathbb{M}(X, \mathcal{B}a(X)) \mid \mu(F) < \mu_0(F) + \epsilon, |\mu(X) - \mu_0(X)| < \epsilon\}.$$

A-topology

The topology which has the sets $W_{G,\epsilon}(\mu_0)$ for functionally open G, equivalently, $W_{F,\epsilon}(\mu_0)$ for F functionally closed, as a subbasis is called the *Alexandrov topology* or *A-topology* [Bog07, 8.10 (iv)]. Thus a generic base element has the shape

$$W_{G_1,\ldots,G_n,\epsilon}(\mu_0) := \bigcap_{1 \le i \le n} W_{G_i,\epsilon}(\mu_0),$$

$$W_{F_1,\ldots,F_n,\epsilon}(\mu_0) := \bigcap_{1 \leq i \leq n} W_{F_i,\epsilon}(\mu_0)$$

with $G_1, \ldots, G_n$ functionally open and $F_1, \ldots, F_n$ functionally closed.

The A-topology is defined in terms of the Baire sets rather than Borel sets of (X, τ). We prefer here the Baire sets, because they take the continuous functions on (X, τ) directly into account. This is in general not the case with the Borel sets, which are defined purely in terms of set-theoretic operations. But the distinction vanishes when we turn to metric spaces, because there each closed set is functionally closed; see Example 4.1.1. Note also that we deal with finite measures here.

Lemma 4.1.34 *The A-topology on $\mathbb{M}(X, \mathcal{B}a(X))$ is Hausdorff.*

Proof The family of functionally closed sets of X is closed under finite intersections; hence if two measures coincide on the functionally closed sets, they must coincide on the Baire sets $\mathcal{B}a(X)$ of X by the π-λ-Theorem 1.6.30. ⊣

Convergence in the A-topology is easily characterized in terms of functionally open or functionally closed sets. Recall that for a sequence $(c_n)_{n\in\mathbb{N}}$ of real numbers, the statement $\limsup_{n\to\infty} c \leq c$ is equivalent to $\inf_{n\in\mathbb{N}} \sup_{k\geq n} c_k \leq c$ which in turn is equivalent to

$$\forall \epsilon > 0 \exists n \in \mathbb{N} \forall k \geq n : c_k < c + \epsilon.$$

Similarly for $\liminf_{n\to\infty} c_n$. This proves

Proposition 4.1.35 *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of measures in $\mathbb{M}(X, \mathcal{B}a(X))$, then the following statements are equivalent:*

1. *$\mu_n \to \mu$ in the A-topology.*
2. *$\limsup_{n\to\infty} \mu_n(F) \leq \mu(F)$ for each functionally closed set F, and $\mu_n(X) \to \mu(X)$.*
3. *$\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$ for each functionally open set G, and $\mu_n(X) \to \mu(X)$.*

⊣

This criterion is sometimes a little impractical, since it deals with inequalities. We could have equality in the limit for all those sets for which the boundary has μ-measure zero, but, alas, the boundary may not be Baire measurable. So we try with an approximation—we approximate a Baire set from within by a functionally open set (corresponding to the interior) and from the outside by a closed set (corresponding to the closure). This is discussed in some detail now.

$\mathcal{R}_\mu$ Given $\mu \in \mathbb{M}(X, \mathcal{B}a(X))$, define by $\mathcal{R}_\mu$ all those Baire sets which have a functional boundary of vanishing μ-measure, formally

$$\begin{aligned}\mathcal{R}_\mu := \{E \in \mathcal{B}a(X) \mid G \subseteq E \subseteq F, \mu(F \setminus G) = 0,\\ G \text{ functionally open}, F \text{ functionally closed}\}.\end{aligned}$$

Hence if X is a metric space, $E \in \mathcal{R}_\mu$ iff $\mu(\partial E) = 0$ for the boundary ∂E of E.

This is another criterion for convergence in the A-topology.

Corollary 4.1.36 *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of the Baire measures. Then $\mu_n \to \mu$ in the A-topology iff $\mu_n(E) \to \mu(E)$ for all $E \in \mathcal{R}_\mu$.*

Proof The condition is necessary by Proposition 4.1.35. Assume, on the other hand, that $\mu_n(E) \to \mu(E)$ for all $E \in \mathcal{R}_\mu$, and take a functionally open set G. We find $f : X \to \mathbb{R}$ continuous such that $G = \{x \in X \mid f(x) > 0\}$. Fix $\epsilon > 0$; then we can find $c > 0$ such that

$$\begin{aligned}\mu(G) < \mu(\{x \in X \mid f(x) > c\}) + \epsilon,\\ \mu(\{x \in X \mid f(x) > c\}) = \mu(\{x \in X \mid f(x) \geq c\}).\end{aligned}$$

Hence $E := \{x \in X \mid f(x) > c\} \in \mathcal{R}_\mu$, since E is open and $F := \{x \in X \mid f(x) \geq c\}$ is closed with $\mu(F \setminus E) = 0$. So $\mu_n(E) \to \mu(E)$, by assumption, and

$$\liminf_{n\to\infty} \mu_n(G) \geq \lim_{n\to\infty} \mu_n(E) = \mu(E) > \mu(G) - \epsilon.$$

Since $\epsilon > 0$ was arbitrary, we infer $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$. Because G was an arbitrary functionally open set, we infer from Proposition 4.1.35 that $(\mu_n)_{n\in\mathbb{N}}$ converges in the A-topology to μ. $\dashv$

The family $\mathcal{R}_\mu$ has some interesting properties, which will be of use later on, because, as we will show in a moment, it contains a base for the topology, if the space is completely regular. This holds whenever there are *enough* continuous functions to separate points from closed sets not containing them. Before we state this property, which will be helpful in the analysis of the A-topology below, we introduce μ-atoms, which are of interest for themselves (we will define later, in Definition 4.3.13, atoms on a strictly order theoretic basis, without reference to measures).

Definition 4.1.37 *A set $A \in \mathcal{A}$ is called a μ-*atom *iff $\mu(A) > 0$ and if $\mu(B) \in \{0, \mu(A)\}$ for every $B \in \mathcal{A}$ with $B \subseteq A$.*

Thus a μ-atom does not permit values other than 0 and $\mu(A)$ for its measurable subsets, so two different μ-atoms A and A' are essentially disjoint, since $\mu(A \cap A') = 0$.

Lemma 4.1.38 *For the finite measure space $(X, \mathcal{A}, \mu)$, there exists an at most countable set $\{A_i \mid i \in I\}$ of atoms such that $X \setminus \bigcup_{i \in I} A_i$ is free of μ-atoms.*

Proof If we do not have any atoms, we are done. Otherwise, let A_1 be an arbitrary atom. This is the beginning. Proceeding inductively, assume that the atoms $A_1, \ldots, A_n$ are already selected, and let $\mathcal{A}_n := \{A \in \mathcal{A} \mid A \subseteq X \setminus \bigcup_{i=1}^n A_i \text{ is an atom}\}$. If $\mathcal{A}_n = \emptyset$, we are done. Otherwise select the atom $A_{n+1} \in \mathcal{A}_n$ with $\mu(A_{n+1}) \geq \frac{1}{2} \cdot \sup_{A \in \mathcal{A}_n} \mu(A)$. Observe that $A_1, \ldots, A_{n+1}$ are mutually disjoint.

Let $\{A_i \mid i \in I\}$ be the set of atoms selected in this way, after the selection has terminated. Assume that $A \subseteq X \setminus \bigcup_{i \in I} A_i$ is an atom; then the index set I must be infinite, and $\mu(A_i) \geq \mu(A)$ for all $i \in I$. But since $\sum_{i \in I} \mu(A_i) \leq \mu(X) < \infty$, we conclude that $\mu(A_i) \to 0$, and consequently, $\mu(A) = 0$; hence A cannot be a μ-atom. ⊣

This is a useful consequence.

Corollary 4.1.39 *Let $f : X \to \mathbb{R}$ be a continuous function. Then there are at most countably many $r \in \mathbb{R}$ such that $\mu(\{x \in X \mid f(x) = r\}) > 0$.*

Proof Consider the image measure $\mathbb{M}(f)(\mu) : B \mapsto \mu(f^{-1}[B])$ on $\mathcal{B}(\mathbb{R})$. If $\mu(\{x \in X \mid f(x) = r\}) > 0$, then $\{r\}$ is a $\mathbb{M}(f)(\mu)$-atom. By Lemma 4.1.38, there are only countably many $\mathbb{M}(f)(\mu)$-atoms. ⊣

Returning to $\mathcal{R}_\mu$, we are now in a position to take a closer look at its structure.

Proposition 4.1.40 *$\mathcal{R}_\mu$ is a Boolean algebra. If (X, τ) is completely regular, then $\mathcal{R}_\mu$ contains a basis for the topology τ.*

Proof It is immediate that $\mathcal{R}_\mu$ is closed under complementation, and it is easy to see that it is closed under finite unions.

Let $f : X \to \mathbb{R}$ be continuous, and define $U(f,r) := \{x \in X \mid f(x) > r\}$; then $U(f,r)$ is open, and $\partial U(f,r) \subseteq \{x \in X \mid f(x) = r\}$; thus $M_f := \{r \in \mathbb{R} \mid \mu(\partial U(f,r)) > 0\}$ is at most countable, such that the sets $U(f,r) \in \mathcal{R}_\mu$, whenever $r \notin M_f$.

Now let $x \in X$ and G be an open neighborhood of x. Because X is completely regular (see Definition 3.3.17), we can find $f : x \to [0,1]$ continuous such that $f(y) = 1$ for all $y \notin G$, and $f(x) = 0$. Hence we can find $r \notin M_f$ such that $x \in U(f,r) \subseteq G$ by Corollary 4.1.39. So $\mathcal{R}_\mu$ is in fact a basis for the topology. $\dashv$

Under the conditions above, $\mathcal{R}_\mu$ contains a base for τ; we lift this base to $\mathbb{M}(X, \mathcal{B}a(X))$ in the hope of obtaining a base for the A-topology. This works, as we will show now.

Corollary 4.1.41 *Let X be a completely regular topological space; then the A-topology has a basis consisting of sets of the form*

$$Q_{A_1,\ldots,A_n,\epsilon}(\mu) := \{\nu \in \mathbb{M}(X, \mathcal{B}a(X)) \mid |\mu(A_i) - \nu(A_i)| < \epsilon \textit{ for } i = 1, \ldots, n\}$$

with $\epsilon > 0$, $n \in \mathbb{N}$ and $A_1, \ldots, A_n \in \mathcal{R}_\mu$.

Proof Let $W_{G_1,\ldots,G_n,\epsilon}(\mu)$ with functionally open sets $G_1, \ldots, G_n$ and $\epsilon > 0$ be given. Select $A_i \in \mathcal{R}_\mu$ functionally open with $A_i \subseteq G_i$ and $\mu(A_i) > \mu(G_i) - \epsilon/2$; then it is easy to see that $Q_{A_1,\ldots,A_n,\epsilon/2}(\mu) \subseteq W_{G_1,\ldots,G_n,\epsilon}(\mu)$. $\dashv$

We will specialize the discussion now to metric spaces. Fix a metric space (X,d), the metric of which we assume to be bounded; otherwise we would switch to the equivalent metric $\langle x, y\rangle \mapsto d(x,y)/(1 - d(x,y))$. Recall that the ϵ-neighborhood B^ϵ of a set $B \subseteq X$ is defined as $B^\epsilon := \{x \in X \mid d(x,B) < \epsilon\}$. Thus B^ϵ is always an open set. Since the Baire and the Borel sets coincide in a metric space by Example 4.1.1, the A-topology is defined on $\mathbb{M}(X, \mathcal{B}(X))$, and we will relate it to a metric now.

B^ϵ

Define the *Lévy–Prohorov distance* $d_P(\mu,\nu)$ of the measures $\mu, \nu \in \mathbb{M}(X, \mathcal{B}(X))$ through

Lévy-Prohorov metric d_P

$$d_P(\mu,\nu) := \inf\{\epsilon > 0 \mid \nu(B) \leq \mu(B^\epsilon) + \epsilon, \mu(B) \leq \nu(B^\epsilon) + \epsilon \text{ for all } B \in \mathcal{B}(X)\}.$$

We note first that d_P defines a metric and that we can find a metrically exact copy of the base space X in the space $\mathbb{M}(X, \mathcal{B}(X))$.

Lemma 4.1.42 *d_P is a metric on $\mathbb{M}(X, \mathcal{B}(X))$. X is isometrically isomorphic to the set $\{\delta_x \mid x \in X\}$ of Dirac measures.*

Proof It is clear that $d_P(\mu, \nu) = d_P(\nu, \mu)$. Let $d_P(\mu, \nu) = 0$; then $\mu(F) \leq \nu(F^{1/n}) + 1/n$ and $\nu(F) \leq \mu(F^{1/n}) + 1/n$ for each closed set $F \subseteq X$; hence $\nu(F) = \mu(F)$ (note that $F^1 \supseteq F^{1/2} \supseteq F^{1/3} \supseteq \dots$ and $F = \bigcap_{n \in \mathbb{N}} F^{1/n}$). Thus $\mu = \nu$. If we have for all $B \in \mathcal{B}(X)$ that

$$\nu(B) \leq \mu(B^\epsilon)+\epsilon, \mu(B) \leq \nu(B^\epsilon)+\epsilon \text{ and } \mu(B) \leq \rho(B^\delta)+\delta, \rho(B) \leq m(B^\delta)+\delta,$$

then

$$\nu(B) \leq \rho(B^{\epsilon+\delta}) + \epsilon + \delta \text{ and } \rho(B) \leq \nu(B^{\epsilon+\delta}) + \epsilon + \delta;$$

thus $d_P(\mu, \nu) \leq d_P(\mu, \rho) + d_P(\rho, \nu)$. We also have $d_P(\delta_x, \delta_y) = d(x, y)$, from which the isometry derives. $\dashv$

We will relate the metric topology to the A-topology now. Without additional assumptions, the following relationship is established:

Proposition 4.1.43 *Each open set in the A-topology is also metrically open; hence A-topology is coarser than the metric topology.*

Proof Let $W_{F_1,\dots,F_n,\epsilon}(\mu)$ be an open basic neighborhood of μ in the A-topology with $F_1, \dots, F_n$ closed. We want to find an open metric neighborhood with center μ which is contained in this A-neighborhood.

Because $(F^{1/n})_{n \in \mathbb{N}}$ is a decreasing sequence with $\inf_{n \in \mathbb{N}} \mu(F_n) = \mu(F)$, whenever F is closed, we find $\delta > 0$ such that $\mu(F_i^\delta) < \mu(F_i) + \epsilon/2$ for $1 \leq i \leq n$ and $0 < \delta < \epsilon/2$. Thus, if $d_P(\mu, \nu) < \delta$, we have for $i = 1, \dots, n$ that $\nu(F_i) < \mu(F_i^\delta) + \delta < \mu(F_i) + \epsilon$. But this means that $\nu \in W_{F_1,\dots,F_n,\epsilon}(\mu)$.

Thus each neighborhood in the A-topology contains in fact an open ball for the d_P-metric. $\dashv$

The converse of Proposition 4.1.43 can only be established under additional conditions, which, however, are met for separable metric spaces. It is a generalization of σ-continuity: While the latter deals with sequences of sets, the concept of τ-regularity deals with the more general notion of directed families of open sets (recall that a family $\mathcal{M}$ of sets is called *directed* iff given $M_1, M_2 \in \mathcal{M}$ there exists $M' \in \mathcal{M}$ with $M_1 \cup M_2 \subseteq M'$).

Definition 4.1.44 *A measure $\mu \in \mathbb{M}(X, \mathcal{B}(X))$ is called* τ-regular *iff*

$$\mu(\bigcup \mathcal{G}) = \sup_{G \in \mathcal{G}} \mu(G)$$

for each directed family $\mathcal{G}$ of open sets.

It is clear that we restrict our attention to open sets, because the union of a directed family of arbitrary measurable sets is not necessarily measurable. It is also clear that the condition above is satisfied for countable increasing sequences of open sets, so that τ-regularity generalizes σ-continuity.

It turns out that finite measures on separable metric spaces are τ-regular. Roughly speaking, this is due to the fact that countably many open sets determine the family of open sets, so that the space cannot be too large when looked at as a measure space.

Lemma 4.1.45 *Let (X, d) be a separable metric space; then each $\mu \in \mathbb{M}(X, \mathcal{B}(X))$ is τ-regular.*

Proof Let $\mathcal{G}_0$ be a countable basis for the metric topology. If $\mathcal{G}$ is a directed family of open sets, we find for each $G \in \mathcal{G}$ a countable cover $(G_i)_{i \in I_G}$ from $\mathcal{G}_0$ with $G = \bigcup_{i \in I_G} G_i$ and $\mu(G) = \sup_{i \in I_G} \mu(G_i)$. Thus

$$\mu(\bigcup \mathcal{G}) = \sup \mu\big(\{\mu(G) \mid G \in \mathcal{G}_0, G \subseteq \bigcup \mathcal{G}\}\big) = \sup_{G \in \mathcal{G}} \mu(G).$$

$\dashv$

As a trivial consequence, it is observed that $\mu(\bigcup \mathcal{G}) = 0$, where $\mathcal{G}$ is the family of all open sets G with $\mu(G) = 0$.

The important observation for our purposes is that a τ-regular measure is supported by a closed set which in terms of μ can be chosen as being as tightly fitting as possible.

Lemma 4.1.46 *Let (X, d) be a separable metric space. Given $\mu \in \mathbb{M}(X, \mathcal{B}(X))$ with $\mu(X) > 0$, there exists a smallest closed set C_μ such that $\mu(C_\mu) = \mu(X)$. C_μ is called the* support of μ *and is denoted by*

supp(μ) supp(μ).

We did use the support already for discrete measures; in this case the support is just the set of points which are assigned positive mass; see Example 2.3.11.

Proof Let $\mathcal{F}$ be the family of all closed sets F with $\mu(F) = \mu(X)$; then $\{X \setminus F \mid F \in \mathcal{F}\}$ is a directed family of open sets of measure zero; hence $\mu(\bigcap \mathcal{F}) = \inf_{F \in \mathcal{F}} \mu(F) = \mu(X)$. Define $\text{supp}(\mu) := \bigcap \mathcal{F}$; then $\text{supp}(\mu)$ is closed with $\mu(\text{supp}(\mu)) = \mu(X)$; if $F \subseteq X$ is a closed set with $\mu(F) = \mu(X)$, then $F \in \mathcal{F}$; hence $\text{supp}(\mu) \subseteq F$. $\dashv$

We may characterize the support of μ also in terms of open sets; this is but a simple consequence of Lemma 4.1.46.

Corollary 4.1.47 *Under the assumptions of Lemma 4.1.46, we have $x \in \text{supp}(\mu)$ iff $\mu(U) > 0$ for each open neighborhood U of x.* $\dashv$

After all these preparations (with some interesting vistas to the landscape of measures), we are in a position to show that the metric topology on $\mathbb{M}(X, \mathcal{B}(X))$ coincides with the A-topology for X separable metric. The following lemma will be the central statement; it is formulated and proved separately, because its proof is somewhat technical. Recall from page 345 that the *diameter* $\text{diam}(Q)$ of $Q \subseteq X$ is defined as diam(Q)

$$\text{diam}(Q) := \sup\{d(x_1, x_2) \mid x_1, x_2 \in Q\}.$$

Lemma 4.1.48 *Every d_P-ball with center $\mu \in \mathbb{M}(X, \mathcal{B}(X))$ contains a neighborhood of μ of the A-topology, if (X, d) is separable metric.*

Proof Fix $\mu \in \mathbb{M}(X, \mathcal{B}(X))$ and $\epsilon > 0$, pick $\delta > 0$ with $4 \cdot \delta < \epsilon$; it is no loss of generality to assume that $\mu(X) = 1$. Because X is separable metric, the support $S := \text{supp}(\mu)$ is defined by Lemma 4.1.46. Because S is closed, we can cover S with a countable number $(V_n)_{n \in \mathbb{N}}$ of open sets, the diameter of which is less than δ and $\mu(\partial V_n) = 0$ by Proposition 4.1.40. Define

$$A_1 := V_1,$$
$$A_n := \bigcup_{i=1}^{n} V_i \setminus \bigcup_{j=1}^{n-1} V_j;$$

then $(A_n)_{n \in \mathbb{N}}$ is a mutually disjoint family of sets which cover S and for which $\mu(\partial A_n) = 0$ holds for all $n \in \mathbb{N}$. We can find an index k such that $\mu(\bigcup_{i=1}^{k}) > 1 - \delta$. Let $T_1, \ldots, T_\ell$ be all sets which are a union of some of the sets $A_1, \ldots, A_k$; then

$$W := W_{T_1,\ldots,T_\ell,\epsilon}(\mu)$$

is a neighborhood of μ in the A-topology by Corollary 4.1.41. We claim that $d_P(\mu, \nu) < \epsilon$ for all $\nu \in W$. In fact, let $B \in \mathcal{B}(X)$ be arbitrary, and put

$$A := \bigcup\{A_i \mid 1 \leq i \leq k, A_i \cap B \neq \emptyset\};$$

then A is among the Ts just constructed, and $B \cap S \subseteq A \cup \bigcup_{i=k+1}^{\infty} A_i$. Moreover, we know that $A \subseteq B^\delta$, because each A_i has a diameter less than δ. This yields

$$\mu(B) = \mu(B \cap S) \leq \mu(A) + \delta < \nu(A) + 2 \cdot \delta \leq \nu(B^\delta) + 2 \cdot \delta.$$

On the other hand, we have

$$\begin{array}{lll} \nu(B) & = \nu(B \cap S) + \nu(B \cap (X \setminus S)) & \leq \nu(A \cap \bigcup_{i=k+1}^{\infty} A_i) + 3 \cdot \delta \\ & \leq \nu(A) + 3 \cdot \delta & \leq \mu(A) + 3 \cdot \delta \\ & \leq \mu(B^\delta) + 4 \cdot \delta. & \end{array}$$

Hence $d_P(\mu, \nu) < 4 \cdot \delta < \epsilon$. Thus W is contained in the open ball centered at μ with radius smaller ϵ. $\dashv$

We have established

Theorem 4.1.49 *The A-topology on $\mathbb{M}(X, \mathcal{B}(X))$ is metrizable by the Lévy-Skohorod metric d_P, provided (X, d) is a separable metric space.* $\dashv$

We will see later that d_P is not the only metric for this topology and that these metric spaces have interesting and useful properties. Some of these properties are best derived through an integral representation, for which a careful study of real-valued functions is required. This is what we are going to investigate in Sect. 4.2. But before doing this, we have a brief and tentative look at the relation between the Borel sets for A-topology and weak σ-algebra.

Lemma 4.1.50 *Let X be a metric space, then the weak σ-algebra is contained in the Borel sets of the A-topology. If the A-topology has a countable basis, both σ-algebras are equal.*

Proof Denote by $\mathcal{C}$ the Borel sets of the A-topology on $\mathbb{M}(X, \mathcal{B}(X))$.

Since X is metric, the Baire sets and the Borel sets coincide. For each closed set F, the evaluation map $ev_F : \mu \mapsto \mu(F)$ is upper semicontinuous by Proposition 4.1.35, so that the set

$$\mathcal{G} := \{A \in \mathcal{B}(X) \mid ev_A \text{ is } \mathcal{C} - \text{measurable}\}$$

contains all closed sets. Because $\mathcal{G}$ is closed under complementation and countable disjoint unions, we conclude that $\mathcal{G}$ contains $\mathcal{B}(X)$. Hence $\wp(\mathcal{B}(X)) \subseteq \mathcal{C}$ by minimality of $\wp(\mathcal{B}(X))$.

2. Assume that the A-topology has a countable basis; then each open set is represented as a countable union of sets of the form $W_{G_1,\ldots,G_n,\epsilon}(\mu_0)$ with $G_1, \ldots, G_n$ open. But $W_{G_1,\ldots,G_n,\epsilon}(\mu_0) \in \wp(\mathcal{B}(X))$, so that each open set is a member of $\wp(\mathcal{B}(X))$. This implies the other inclusion. $\dashv$

We will investigate the A-topology further in Sect. 4.10 and turn to real-valued functions now.

4.2 Real-Valued Functions

In discussing the set of all measurable and bounded functions into the real line, we show first that the set of all these functions is closed under the usual algebraic operations, so that it is a vector space, and that it is also closed under finite infima and suprema, rendering it a distributive lattice; in fact, algebraic operations and order are compatible. Then we show that the measurable step functions are dense with respect to pointwise convergence. This is an important observation, which will help us later on to transfer relevant properties from indicator functions (a.k.a. measurable sets) to general measurable functions. This prepares the stage for discussing convergence of functions in the presence of a measure. We will deal with convergence almost everywhere, which neglects a set of measure zero for the purposes of convergence, and convergence in measure, which is defined in terms of a pseudometric, but surprisingly turns out to be related to convergence almost everywhere through subsequences of subsequences (this sounds a bit mysterious, but carry on).

Lemma 4.2.1 *Let $f, g : X \to \mathbb{R}$ be $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable functions for the measurable space $(X, \mathcal{A})$. Then $f \wedge g$, $f \vee g$ and $\alpha \cdot f + \beta \cdot g$ are $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable for $\alpha, \beta \in \mathbb{R}$.*

Proof If f is measurable, $\alpha \cdot f$ is. This follows immediately from Lemma 4.1.4. From

$$\begin{aligned}&\{x \in X \mid f(x) + g(x) < q\}\\&\quad= \bigcup_{r_1, r_2 \in \mathbb{Q}, r_1 + r_2 \leq q} \big(\{x \mid f(x) < r_1\} \cap \{x \mid g(x) < r_2\}\big),\end{aligned}$$

we conclude that the sum of measurable functions is measurable again. Since

$$\begin{aligned}\{x \in X \mid (f \wedge g)(x) < q\} &= \{x \mid f(x) < q\} \cup \{x \mid g(x) < q\}\\\{x \in X \mid (f \vee g)(x) < q\} &= \{x \mid f(x) < q\} \cap \{x \mid g(x) < q\},\end{aligned}$$

we see that both $f \wedge g$ and $f \vee g$ are measurable. ⊣

Corollary 4.2.2 *If $f : X \to \mathbb{R}$ is $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable, so is $|f|$.*

Proof Write $|f| = f^+ - f^-$ with $f^+ := f \vee 0$ and $f^- := (-f) \vee 0$. ⊣

$\mathcal{F}(X, \mathcal{A})$ The consequence is that for a measurable space $(X, \mathcal{A})$, the set

$$\mathcal{F}(X, \mathcal{A}) := \{f : N \to \mathbb{R} \mid f \text{ is } \mathcal{A} - \mathcal{B}(\mathbb{R}) \text{ measurable}\}$$

is both a vector space and a distributive lattice; in fact, it is what will be called a vector lattice later; see Definition 4.8.10 on page 550. Assume that $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}(X, \mathcal{A})$ is a sequence of bounded measurable functions such that $f : x \mapsto \liminf_{n\to\infty} f_n(x)$ is a bounded function; then $f \in \mathcal{F}(X, \mathcal{A})$. This is so because

$$\begin{aligned}\{x \in X \mid \liminf_{n\to\infty} f_n(x) \leq q\} &= \{x \mid \sup_{n\in\mathbb{N}} \inf_{k\geq n} f_k(x) \leq q\}\\&= \bigcap_{n\in\mathbb{N}} \{x \mid \inf_{k\geq n} f_k(x) \leq q\}\\&= \bigcap_{n\in\mathbb{N}} \bigcap_{\ell\in\mathbb{N}} \{x \mid \inf_{k\geq n} f_k(x) < q + 1/\ell\}\\&= \bigcap_{n\in\mathbb{N}} \bigcap_{\ell\in\mathbb{N}} \bigcup_{k\geq n} \{x \mid f_k(x) < q + 1/\ell\}.\end{aligned}$$

Similarly, if $x \mapsto \limsup_{n\to\infty} f_n(x)$ defines a bounded function, then it is measurable as well. Consequently, if the sequence $(f_n(x))_{n\in\mathbb{N}}$ converges to a bounded function f, then $f \in \mathcal{F}(X, \mathcal{A})$.

Hence we have shown

Proposition 4.2.3 *Let $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}(X, \mathcal{A})$ be a sequence of bounded measurable functions. Then*

- *If $f_*(x) := \liminf_{n\to\infty} f_n(x)$ defines a bounded function, then $f_* \in \mathcal{F}(X, \mathcal{A})$,*
- *If $f^*(x) := \limsup_{n\to\infty} f_n(x)$ defines a bounded function, then $f^* \in \mathcal{F}(X, \mathcal{A})$.*

⊣

We use occasionally the representation of sets through indicator functions. Recall for $A \subseteq X$ its *indicator function*

$$\chi_A(x) := \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A. \end{cases}$$

Clearly, if $\mathcal{A}$ is a σ-algebra on X, then $A \in \mathcal{A}$ iff χ_A is a $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable function. This is so since we have for the inverse image of an interval under χ_A

$$\chi_A^{-1}\big[[0, q]\big] = \begin{cases} \emptyset, & \text{if } q < 0, \\ X \setminus A, & \text{if } 0 \leq q < 1, \\ X, & \text{if } q \geq 1. \end{cases}$$

A measurable *step function* Step function

$$f = \sum_{i=1}^{n} \alpha_i \cdot \chi_{A_i}$$

is a linear combination of indicator functions with $A_i \in \mathcal{A}$. Since $\chi_A \in \mathcal{F}(X, \mathcal{A})$ for $A \in \mathcal{A}$, measurable step functions are indeed measurable functions.

Proposition 4.2.4 *Let $(X, \mathcal{A})$ be a measurable space. Then*

1. *For $f \in \mathcal{F}(X, \mathcal{A})$ with $f \geq 0$, there exists an increasing sequence $(f_n)_{n\in\mathbb{N}}$ of step functions $f_n \in \mathcal{F}(X, \mathcal{A})$ with*

$$f(x) = \sup_{n\in\mathbb{N}} f_n(x)$$

for all $x \in X$.

2. *For $f \in \mathcal{F}(X, \mathcal{A})$, there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of step functions $f_n \in \mathcal{F}(N, \mathcal{A})$ with*

$$f(x) = \lim_{n\to\infty} f_n(x)$$

for all $x \in X$.

Proof 1. Take $f \geq 0$, and assume without loss of generality that $f \leq 1$ (otherwise, if $0 \leq f \leq m$, consider f/m). Put

$$A_{i,n} := \{x \in X \mid i/n \leq f(x) < (i+1)/n\},$$

for $n \in \mathbb{N}$, $0 \leq i \leq n$, then $A_{i,n} \in \mathcal{A}$, since f is measurable. Define

$$f_n(x) := \sum_{0\leq i<2^n} i \cdot 2^{-n} \chi_{A_{i,2^n}}.$$

Then f_n is a measurable step function, and $f_n \leq f$; moreover $(f_n)_{n\in\mathbb{N}}$ is increasing. This is so because given $n \in \mathbb{N}, x \in X$, we can find i such that $x \in A_{i,2^n} = A_{2i,2^{n+1}} \cup A_{2i+1,2^{n+1}}$. If $f(x) < (2i+1)/2^{n+1}$, we have $x \in A_{2i,2^{n+1}}$ with $f_n(x) = f_{n+1}(x)$; if, however, $(2i+1)/2^{n+1} \leq f(x)$, we have $f_n(x) < f_{n+1}(x)$.

Given $\epsilon > 0$, choose $n_0 \in \mathbb{N}$ with $2^{-n} < \epsilon$ for $n \geq n_0$. Let $x \in X, n \geq n_0$, then $x \in A_{i,2^n}$ for some i; hence $|f_n(x) - f(x)| = f(x) - i2^{-n} < 2^{-n} < \epsilon$. Thus $f = \sup_{n\in\mathbb{N}} f_n$.

2. Given $f \in \mathcal{F}(X, \mathcal{A})$, write $f_1 := f \wedge 0$ and $f_2 := f \vee 0$, then $f = f_1 + f_2$ with $f_1 \leq 0$ and $f_2 \geq 0$ as measurable and bounded functions. Hence $f_2 = \sup_{n\in\mathbb{N}} g_n = \lim_{n\to\infty} g_n$ and $-f_1 = -\sup_{n\in\mathbb{N}} h_n = -\lim_{n\to\infty} h_n$ for increasing sequences of step functions $(g_n)_{n\in\mathbb{N}}$ and $(h_n)_{n\in\mathbb{N}}$. Thus $f = \lim_{n\to\infty}(g_n + h_n)$, and $g_n + h_n$ is a step function for each $n \in \mathbb{N}$. $\dashv$

Given $f : X \to \mathbb{R}$ with $f \geq 0$, the set $\{\langle x, q\rangle \in X \times \mathbb{R} \mid 0 \leq f(x) \leq q\}$ can be visualized as the area between the X-axis and the graph of the function. We obtain as a consequence that this set is measurable, provided f is measurable. This gives an example of a product measurable set. To be specific

Corollary 4.2.5 *Let $f : X \to \mathbb{R}$ with $f \geq 0$ be a bounded measurable function for a measurable space $(X, \mathcal{A})$, and define*

$$C_{\bowtie}(f) := \{\langle x, q\rangle \mid q \geq 0 \textit{ and } \bowtie f(x)\} \subseteq X \times \mathbb{R}$$

for the relational operator $\bowtie$ *taken from* $\{\geq, <, =, \neq, >, \geq\}$. *Then* $C_{\bowtie}(f) \in \mathcal{A} \otimes \mathcal{B}(\mathbb{R})$.

Proof We prove the assertion for $C(f) := C_{<}(f)$, from which the other cases may easily be derived, e.g.,

$$C_{\leq}(f) = \bigcap_{k \in \mathbb{N}} \{\langle x, q \rangle \mid f(x) < q + 1/k\} = \bigcap_{k \in \mathbb{N}} C_{<}(f - 1/k).$$

Consider these cases:

- If $f = \chi_A$ with $A \in \mathcal{A}$, then $C(f) = X \setminus A \times \{0\} \cup A \times [0, 1[\in \mathcal{A} \otimes \mathcal{B}(\mathbb{R})$.
- If f is represented as a step function with a finite number of mutually disjoint steps, say, $f = \sum_{i=1}^{k} r_i \cdot \chi_{A_i}$ with $r_i \geq 0$ and all $A_i \in \mathcal{A}$, then

$$C(f) = \big(X \setminus \bigcup_{i=1}^{k} A_i\big) \times \{0\} \cup \bigcup_{i=1}^{k} A_i \times [0, r_i[\ \in \mathcal{A} \otimes \mathcal{B}(\mathbb{R}).$$

- If f is represented as a monotone limit of step function $(f_n)_{n \in \mathbb{N}}$ with $f_n \geq 0$ according to Proposition 4.2.4, then $C(f) = \bigcup_{n \in \mathbb{N}} C(f_n)$; thus $C(f) \in \mathcal{A} \otimes \mathcal{B}(\mathbb{R})$.

$\dashv$

This is a simple first application. We look at the evaluation of a measure at a certain set and want the value to be not smaller than a given threshold. The pairs of measures and corresponding thresholds constitute a product measurable set, to be specific

Example 4.2.6 Given a measurable space $(X, \mathcal{A})$ and the measurable set $A \in \mathcal{A}$, the set $\{\langle \mu, r \rangle \in \mathbb{M}(X, \mathcal{A}) \times \mathbb{R}_+ \mid \mu(A) \bowtie r\}$ is a member of $\wp(X, \mathcal{A}) \otimes \mathcal{B}(\mathbb{R})$. This is so by Corollary 4.2.5, since ev_A is $\wp(X, \mathcal{A})$-$\otimes\mathcal{B}(\mathbb{R})$-measurable. ✌

We obtain also measurability of the validity sets for the simple modal logic discussed above.

Example 4.2.7 Consider the simple modal logic in Example 4.1.11, interpreted through a transition kernel $M : (X, \mathcal{A}) \rightsquigarrow (X, \mathcal{A})$. Given a formula φ, the set $\{\langle x, r \rangle \mid M(x)([\![\varphi]\!]_M) \geq r\}$ is a member of $\mathcal{A} \otimes \mathcal{B}(\mathbb{R})$, from which $[\![\Diamond_q \varphi]\!]_M$ may be extracted through a horizontal cut at q (see page 437). Hence this observation generalizes measurability of $[\![\cdot]\!]_M$, one of the cornerstones for interpreting modal logics probabilistically. ✌

We will turn now to the interplay of measurable functions and measures and have a look at different modes of convergence for sequences of measurable functions in the presence of a (finite) measure.

4.2.1 Essentially Bounded Functions

Fix for this section a finite measure space $(X, \mathcal{A}, \mu)$. We say that a measurable property holds *μ-almost everywhere* (abbreviated as *μ-a.e.*) iff the set on which the property does not hold has μ-measure zero.

μ-a.e.

The measurable function $f \in \mathcal{F}(X, \mathcal{A})$ is called *μ-essentially bounded* iff

$$||f||_\infty^\mu := \inf\{a \in \mathbb{R} \mid |f| \leq_\mu a\} < \infty,$$

where $f \leq_\mu a$ indicates that $f \leq a$ holds μ-a.e. Thus a μ-essentially bounded function may occasionally take arbitrary large values, but the set of these values must be negligible in terms of μ.

The set

$$\mathcal{L}_\infty(\mu) := \mathcal{L}_\infty(X, \mathcal{A}, \mu) := \{f \in \mathcal{F}(X, \mathcal{A}) \mid ||f||_\infty^\mu < \infty\}$$

of all μ-essentially bounded functions is a real vector space, and we have for $||\cdot||_\infty^\mu$ these properties.

Lemma 4.2.8 *Let $f, g \in \mathcal{F}(X, \mathcal{A})$ be essentially bounded, $\alpha, \beta \in \mathbb{R}$, then $||\cdot||_\infty^\mu$ is a* pseudo-norm *on $\mathcal{F}(X, \mathcal{A})$, i.e.,*

1. *If $||f||_\infty^\mu = 0$, then $f =_\mu 0$.*
2. *$||\alpha \cdot f||_\infty^\mu = |\alpha| \cdot ||f||_\infty^\mu$,*
3. *$||f + g||_\infty^\mu \leq ||f||_\infty^\mu + ||g||_\infty^\mu$.*

Proof If $||f||_\infty^\mu = 0$, we have $|f| \leq_\mu 1/n$ for all $n \in \mathbb{N}$, so that

$$\{x \in X \mid |f(x)| \neq 0\} \subseteq \bigcup_{n\in\mathbb{N}} \{x \in X \mid |f(x)| \leq 1/n\};$$

consequently, $f =_\mu 0$. The converse is trivial. The second property follows from $|f| \leq_\mu a$ iff $|\alpha \cdot f| \leq_\mu |\alpha| \cdot a$ and the third one from the observation that $|f| \leq_\mu a$ and $|g| \leq_\mu b$ implies $|f+g| \leq |f|+|g| \leq_\mu a+b$. ⊣

So $||\cdot||_\infty^\mu$ *nearly* a norm, but the crucial property that the norm for a vector is zero only if the vector is zero is missing. We factor $\mathcal{L}_\infty(X, \mathcal{A}, \mu)$ with respect to the equivalence relation $=_\mu$; then the set

$$L_\infty(\mu) := L_\infty(X, \mathcal{A}, \mu) := \{[f] \mid f \in \mathcal{L}_\infty(X, \mathcal{A}, \mu)\}$$

of all equivalence classes $[f]$ of μ-essentially bounded measurable functions is a vector space again. This is so because $f =_\mu g$ and $f' =_\mu g'$ together imply $f + f' =_\mu g + g'$ and $f =_\mu g$ implies $\alpha \cdot f =_\mu \alpha \cdot g$ for all $\alpha \in \mathbb{R}$. Moreover,

$$||\,[f]\,||_\infty^\mu := ||f||_\infty^\mu$$

defines a norm on this space. For easier reading, we will identify in the sequel f with its class $[f]$.

We obtain in this way a normed vector space, which is complete with respect to this norm; see Definition 3.6.40 on page 389.

Proposition 4.2.9 *$(L_\infty(\mu), ||\cdot||_\infty^\mu)$ is a Banach space.*

Proof Let $(f_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $L_\infty(X, \mathcal{A}, \mu)$, and define

$$N := \bigcup_{n_1, n_2 \in \mathbb{N}} \{x \in X \mid |f_{n_1}(x) - f_{n_2}(x)| > ||f_{n_1} - f_{n_2}||_\infty^\mu\};$$

then $\mu(N) = 0$. Put $g_n := \chi_{X\setminus N} \cdot f_n$; then $(g_n)_{n\in\mathbb{N}}$ converges uniformly with respect to the supremum norm $||\cdot||_\infty$ to some element $g \in \mathcal{F}(X, \mathcal{A})$; hence also $||f_n - g||_\infty^\mu \to 0$. Clearly, g is bounded. ⊣

This is the first instance of a vector space intimately connected with a measure space. We will encounter several of these spaces in Sect. 4.11 and discuss them in greater detail, when integration is at our disposal.

The convergence of a sequence of measurable functions into $\mathbb{R}$ in the presence of a finite measure is discussed now. Without a measure, we may use pointwise or uniform convergence for modeling approximations. Recall that *pointwise convergence* of a sequence $(f_n)_{n\in\mathbb{N}}$ of functions to a function f is given by

$$\forall x \in X : \lim_{n\to\infty} f_n(x) = f(x), \tag{4.4}$$

and the stronger form of *uniform convergence* through

$$\lim_{n\to\infty} ||f_n - f||_\infty = 0,$$

with $||\cdot||_\infty$ as the supremum norm, given by

$$||f||_\infty := \sup_{x\in X} |f(x)|.$$

We will weaken the first condition (4.4) to hold not everywhere but *almost* everywhere, so that the set on which it does not hold will be a set of measure zero. This leads to the notion of convergence almost everywhere, which will turn out to be quite close to uniform convergence, as we will see when discussing Egorov's Theorem. Convergence almost everywhere will be weakened to convergence in measure, for which we will define a pseudometric. This in turn gives rise to another Banach space upon factoring.

4.2.2 Convergence *Almost Everywhere* and *in Measure*

$f_n \xrightarrow{a.e.} f$

Recall that we work in a finite measure space $(X, \mathcal{A}, \mu)$. The sequence $(f_n)_{n\in\mathbb{N}}$ of measurable functions $f_n \in \mathcal{F}(X, \mathcal{A})$ is said to *converge almost everywhere* to a function $f \in \mathcal{F}(X, \mathcal{A})$ (written as $f_n \xrightarrow{a.e.} f$) iff the sequence $(f_n(x))_{n\in\mathbb{N}}$ converges pointwise to $f(x)$ for every x outside a set of measure zero. Thus we have $\mu(X \setminus K) = 0$, where $K := \{x \in X \mid f_n(x) \to f(x)\}$. Because

$$K = \bigcap_{n\in\mathbb{N}} \bigcup_{m\in\mathbb{N}} \bigcap_{\ell\geq m} \{x \in X \mid |f_\ell(x) - f(x)| < 1/n\},$$

K is a measurable set. It is clear that $f_n \xrightarrow{a.e.} f$ and $f_n \xrightarrow{a.e.} f'$ imply $f =_\mu f'$.

The next lemma shows that convergence everywhere is compatible with the common algebraic operations on $\mathcal{F}(X, \mathcal{A})$ like addition, scalar multiplication, and the lattice operations. Since these functions can be represented as continuous function of several variables, we formulate this closure property abstractly in terms of compositions with continuous functions.

Lemma 4.2.10 *Let $f_{i,n} \xrightarrow{a.e.} f_i$ for $1 \leq i \leq k$, and assume that $g : \mathbb{R}^k \to \mathbb{R}$ is continuous. Then $g \circ (f_{1,n}, \dots, f_{k,n}) \xrightarrow{a.e.} g \circ (f_1, \dots, f_k)$.*

Proof Put $h_n := g\circ(f_{1,n}, \dots, f_{k,n})$. Since g is continuous, we have

$$\begin{aligned}\{x \in X \mid \bigl(h_n(x)\bigr)_{n\in\mathbb{N}} &\text{ does not converge}\}\\ &\subseteq \bigcup_{j=1}^{k} \{x \in X \mid \bigl(f_{j,n}(x)\bigr)_{n\in\mathbb{N}} \text{ does not converge}\};\end{aligned}$$

hence the set on the left-hand side has measure zero. ⊣

Intuitively, convergence almost everywhere means that the measure of the set

$$\bigcup_{n\geq k}\{x \in X \mid |f_n(x) - f(x)| > \epsilon\}$$

tends to zero, as $k \to \infty$, so we are coming closer and closer to the limit function, albeit on a set the measure of which becomes smaller and smaller. We show that this intuitive understanding yields an adequate model for this kind of convergence.

Lemma 4.2.11 *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions in $\mathcal{F}(X, \mathcal{A})$ and $f \in \mathcal{F}(X, \mathcal{A})$. Then the following conditions are equivalent:*

1. $f_n \xrightarrow{a.e.} f$.
2. $\lim_{k\to\infty} \mu\bigl(\bigcup_{n\geq k}\{x \in X \mid |f_n(x) - f(x)| > \epsilon\}\bigr) = 0$ *for every* $\epsilon > 0$.

Proof 0. Let us first write down what the equality in property 2 really means; then the proof will be nearly straightforward. Plan

1. Let $\epsilon > 0$ be given; then there exists $k \in \mathbb{N}$ with $1/k < \epsilon$, so that

$$\begin{aligned}&\lim_{k\to\infty} \mu\bigl(\bigcup_{n\geq k}\{x \in X \mid |f_n(x) - f(x)| > \epsilon\}\bigr)\\ &\overset{(*)}{=} \mu\bigl(\textstyle\bigcap_{k\in\mathbb{N}} \bigcup_{n\geq k}\{x \in X \mid |f_n(x) - f(x)| > \epsilon\}\bigr)\\ &\leq \mu(\{x \in X \mid (f_n(x))_{n\in\mathbb{N}} \text{ does not converge}\}).\end{aligned}$$

2. Now assume that $f_n \xrightarrow{a.e.} f$; then the implication 1 $\Rightarrow$ 2 is immediate. If, however, $f_n \xrightarrow{a.e.} f$ is false, then we find for each $\epsilon > 0$ so that for all $k \in \mathbb{N}$, there exists $n \geq k$ with $\mu(\{x \in X \mid |f_n(x) - f(x)| \geq \epsilon\}) > 0$. Thus property 2 cannot hold. ⊣

Note that the statement above requires a finite measure space, because the measure of a decreasing sequence of sets is the infimum of the individual measures, used in the equation marked $(*)$. This is not necessarily valid for nonfinite measure space.

The characterization implies that a.e.-Cauchy sequences converge.

Corollary 4.2.12 *An a.e.-Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{F}(X, \mathcal{A})$ converges almost everywhere to some $f \in \mathcal{F}(X, \mathcal{A})$.*

Proof Because $(f_n)_{n\in\mathbb{N}}$ is an a.e.-Cauchy sequence, we have that $\mu(X \setminus K_\epsilon) = 0$ for every $\epsilon > 0$, where

$$K_\epsilon := \bigcap_{k\in\mathbb{N}} \bigcup_{n,m\geq k} \{x \in X \mid |f_n(x) - f_m(x)| > \epsilon\}.$$

Put

$$N := \bigcup_{k\in\mathbb{N}} K_{1/k},$$
$$g_n := f_n \cdot \chi_{X\setminus N};$$

then $(g_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{F}(X, \mathcal{A})$ which converges pointwise to some $f \in \mathcal{F}(X, \mathcal{A})$. Since $\mu(X \setminus N) = 0$, $f_n \xrightarrow{a.e.} f$ follows. $\dashv$

Convergence a.e. is very nearly uniform convergence, where *very nearly* serves to indicate that the set on which uniform convergence is violated is arbitrarily small. To be specific, we find for each threshold a set the complement of which has a measure smaller than this bound, on which convergence is uniform. This is what *Egorov's Theorem* says.

Egorov's Theorem

Proposition 4.2.13 *Let $f_n \xrightarrow{a.e.} f$ for $f_n, f \in \mathcal{F}(X, \mathcal{A})$. Given $\epsilon > 0$, there exists $A \in \mathcal{A}$ such that*

1. $\sup_{x\in A} |f_n(x) - f(x)| \to 0$,
2. $\mu(X \setminus A) < \epsilon$.

Plan of attack

The idea of the proof is that we investigate the set of all x for which uniform convergence is spoiled by $1/k$. This set can be made arbitrarily small in terms of μ, so the countable union of all these sets can be made as small as we want. Outside this set we have uniform convergence. Let us look at a more formal treatment now.

Proof Fix $\epsilon > 0$; then there exists for each $k \in \mathbb{N}$ an index $n_k \in \mathbb{N}$ such that $\mu(B_k) < \epsilon/2^{k+1}$ with

$$B_k := \bigcup_{m \geq n_k} \{x \in X \mid |f_m(x) - f(x)| > 1/k\}.$$

Now put $A := \bigcap_{k \in \mathbb{N}} (X \setminus B_k)$; then

$$\mu(X \setminus A) \leq \sum_{k \in \mathbb{N}} \mu(B_k) \leq \epsilon,$$

and we have for all $k \in \mathbb{N}$

$$\sup_{x \in A} |f_n(x) - f(x)| \leq \sup_{x \notin B_k} |f_n(x) - f(x)| \leq 1/k$$

for $n \geq n_k$. Thus

$$\lim_{n \to \infty} \sup_{x \in A} |f_n(x) - f(x)| = 0,$$

as claimed. $\dashv$

Convergence almost everywhere makes sure that the set on which a sequence of functions does not converge has measure zero, and Egorov's Theorem shows that this is *almost* uniform convergence.

Convergence in measure for a finite measure space $(X, \mathcal{A}, \mu)$ takes another approach: Fix $\epsilon > 0$, and consider the set $\{x \in X \mid |f_n(x) - f(x)| > \epsilon\}$. If the measure of this set (for a fixed, but arbitrary ϵ) tends to zero, as $n \to \infty$, then we say that $(f_n)_{n \in \mathbb{N}}$ *converges in measure* to f, and write $f_n \xrightarrow{i.m.} f$. In order to have a closer look at this notion of convergence, we note that it is invariant against equality almost everywhere: If $f_n =_\mu g_n$ and $f =_\mu g$, then $f_n \xrightarrow{i.m.} f$ implies $g_n \xrightarrow{i.m.} g$, and vice versa. $f_n \xrightarrow{i.m.} f$

We will introduce a pseudometric δ on $\mathcal{F}(X, \mathcal{A})$ first:

$$\delta(f, g) := \inf\bigl\{\epsilon > 0 \mid \mu(\{x \in X \mid |f(x) - g(x)| > \epsilon\} \leq \epsilon\bigr\}.$$

These are some elementary properties of δ:

Lemma 4.2.14 *Let $f, g, h \in \mathcal{F}(X, \mathcal{A})$; then we have*

1. $\delta(f, g) = 0$ *iff* $f =_\mu g$,
2. $\delta(f, g) = \delta(g, f)$,
3. $\delta(f, g) \leq \delta(f, h) + \delta(h, g)$.

Proof If $\delta(f, g) = 0$, but $f \neq_\mu g$, there exists k with $\mu(\{x \in X \mid |f(x) - g(x)| > 1/k\}) > 1/k$. This is a contradiction. The other direction is trivial. Symmetry of δ is also trivial, so the triangle inequality remains to be shown. If $|f(x)-g(x)| > \epsilon_1+\epsilon_2$, then $|f(x)-h(x)| > \epsilon_1$ or $|h(x) - g(x)| > \epsilon_2$; thus

$$\mu(\{x \in X \mid |f(x) - g(x)| > \epsilon_1 + \epsilon_2\}) \leq \mu(\{x \in X \mid |f(x) - h(x)| > \epsilon_1\}) + \mu(\{x \in X \mid |h(x) - g(x)| > \epsilon_2\}).$$

This implies the third property. $\dashv$

This, then, is the formal definition of convergence in measure:

Definition 4.2.15 *The sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{F}(X, \mathcal{A})$ is said to* converge in measure *to $f \in \mathcal{F}(X, \mathcal{A})$ (written as $f_n \xrightarrow{i.m.} f$) iff $\delta(f_n, f) \to 0$, as $n \to \infty$.*

$f_n \xrightarrow{i.m.} f$

We can express convergence in measure in terms of convergence almost everywhere.

Proposition 4.2.16 *$(f_n)_{n\in\mathbb{N}}$ converges in measure to f iff each subsequence of $(f_n)_{n\in\mathbb{N}}$ contains a subsequence $(h_n)_{n\in\mathbb{N}}$ with $h_n \xrightarrow{a.e.} f$.*

Proof The proposal singling out a subsequence from a subsequence rather than from the sequence proper appears strange. The proof will show that we need a subsequence to "prime the pump," i.e., to get going.

"⇒": Assume $f_n \xrightarrow{i.m.} f$, and let $\epsilon > 0$ be arbitrary but fixed. Let $(g_n)_{n\in\mathbb{N}}$ be an arbitrary subsequence of $(f_n)_{n\in\mathbb{N}}$. We find a sequence of indices $n_1 < n_2 < \ldots$ such that

$$\mu(\{x \in X \mid |g_{n_k}(x) - f(x)| > \epsilon\}) < 1/k^2.$$

Let $h_k := g_{n_k}$, then we obtain

$$\mu(\bigcup_{k\geq\ell}\{x \in X \mid |h_k - f| > \epsilon\}) \leq \sum_{k\geq\ell} \frac{1}{k^2} \to 0,$$

as $\ell \to \infty$. Hence $h_k \xrightarrow{a.e.} f$.

"⇐": If $\delta(f_n, f) \not\to 0$, we can find a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ and $r > 0$ such that for all $k \in \mathbb{N}$, $\mu(\{x \in X \mid |f_{n_k}(x) - f(x)| > r\}) > r$ holds. Let $(g_n)_{n\in\mathbb{N}}$ be a subsequence of this subsequence; then

$$\begin{aligned}&\lim_{n\to\infty} \mu(\{x \in X \mid |g_n - f| > r\})\\ &\quad\leq \lim_{n\to\infty} \mu(\bigcup_{m\geq n}\{x \in X \mid |g_m - f| > r\}) = 0\end{aligned}$$

by Lemma 4.2.11. This is a contradiction.

⊣

Hence convergence almost everywhere implies convergence in measure. Just for the record

Corollary 4.2.17 *If $(f_n)_{n\in\mathbb{N}}$ converges almost everywhere to f, then the sequence converges also in measure to f.* ⊣

The converse relationship is a bit more involved. Intuitively, a sequence which converges in measure need not converge almost everywhere.

Example 4.2.18 Let $A_{i,n} := [(i-1)/n, i/n]$ for $n \in \mathbb{N}$ and $1 \leq i \leq n$, and consider the sequence

$$(f_n)_{n\in\mathbb{N}} := \langle \chi_{A_{1,1}}, \chi_{A_{1,2}}, \chi_{A_{2,2}}, \chi_{A_{3,1}}, \chi_{A_{3,2}}, \chi_{A_{3,3}}, \ldots \rangle,$$

so that in general

$$\chi_{A_{1,n}}, \ldots, \chi_{A_{n,n}}$$

is followed by

$$\chi_{A_{1,n+1}}, \ldots, \chi_{A_{n+1,n+1}}.$$

Let μ be the Lebesgue measure λ on $\mathcal{B}([0,1])$. Given $\epsilon > 0$, $\lambda(\{x \in [0,1] \mid f_n(x) > \epsilon\})$ can be made arbitrarily small for any given $\epsilon > 0$; hence $f_n \xrightarrow{i.m.} 0$. On the other hand, $\big(f_n(x)\big)_{n\in\mathbb{N}}$ fails to converge for any $x \in [0,1]$, so $f_n \xrightarrow{a.e.} 0$ is false. ✌

We have, however, this observation, which draws atom into our game.

Proposition 4.2.19 *Let $(A_i)_{i \in I}$ be the at most countable collection of μ-atoms according to Lemma 4.1.38 such that $B := X \setminus \bigcup_{i \in I} A_i$ does not contain any atoms. Then these conditions are equivalent:*

1. *Convergence in measure implies convergence almost everywhere.*
2. *$\mu(B) = 0$.*

Proof 1 $\Rightarrow$ 2: Assume that $\mu(B) > 0$; then we know that for each $k \in \mathbb{N}$ there exist mutually disjoint measurable subsets $B_{1,k}, \ldots, B_{k,k}$ of B such that $\mu(B_{i,k}) = 1/k \cdot \mu(B)$ and $B = \bigcup_{1 \leq i \leq k} B_{i,k}$. This is so because B does not contain any atoms. Define as in Example 4.2.18

$$(f_n)_{n \in \mathbb{N}} := \langle \chi_{B_{1,1}}, \chi_{B_{1,2}}, \chi_{B_{2,2}}, \chi_{B_{3,1}}, \chi_{B_{3,2}}, \chi_{B_{3,3}}, \ldots \rangle,$$

so that in general

$$\chi_{B_{1,n}}, \ldots, \chi_{B_{n,n}}$$

is followed by

$$\chi_{B_{1,n+1}}, \ldots, \chi_{B_{n+1,n+1}}.$$

Because $\mu(\{x \in X \mid f_n(x) > \epsilon\}$ can be made arbitrarily small for any positive ϵ, we find $f_n \xrightarrow{i.m.} 0$. If we assume that convergence in measure implies convergence almost everywhere, we have $f_n \xrightarrow{a.e.} 0$, but this is false, because $\liminf_{n\to\infty} f_n = 0$ and $\limsup_{n\to\infty} f_n = \chi_B$. Thus we arrive at a contradiction.

2 $\Rightarrow$ 1: Let $(f_n)_{n \in \mathbb{N}}$ be a sequence with $f_n \xrightarrow{i.m.} f$. Fix an atom A_i; then $\mu(\{x \in A_i \mid |f_n(x) - f(x)| > 1/k\}) = 0$ for all $n \geq n_k$ with n_k suitably chosen; this is so because A_i is an atom; hence measurable subsets of A_i take only the values 0 and $\mu(A_i)$. Put

$$g := \inf_{n \in \mathbb{N}} \sup_{n_1, n_2 \geq n} |f_{n_1} - f_{n_2}|;$$

then $g(x) \neq 0$ iff $(f_n(x))_{n \in \mathbb{N}}$ does not converge to $f(x)$. We infer $\mu(\{x \in A_i \mid g(x) \geq 2/k\}) = 0$. Because the family $(A_i)_{i \in I}$ is mutually disjoint, we conclude that $\mu(\{x \in X \mid g(x) \geq 2/k\}) = 0$ for all $k \in \mathbb{N}$. But now look at this

$$\mu(\{x \in X \mid \liminf_{n\to\infty} f_n(x) < \limsup_{n\to\infty} f_n(x)\} = \mu(\{x \in X \mid g(x) > 0\}) = 0.$$

Consequently, $f_n \xrightarrow{a.e.} f$. $\dashv$

Again we want to be sure that convergence in measure is preserved by the usual algebraic operations like addition or taking the infimum, so we state as a counterpart to Lemma 4.2.10 now as an easy consequence of Proposition 4.2.16.

Lemma 4.2.20 *Let $f_{i,n} \xrightarrow{i.m.} f_i$ for $1 \leq i \leq k$, and assume that $g : \mathbb{R}^k \to \mathbb{R}$ is continuous. Then $g \circ (f_{1,n}, \ldots, f_{k,n}) \xrightarrow{i.m.} g \circ (f_1, \ldots, f_k)$.*

Proof By iteratively selecting subsequences, we can find subsequences $(h_{i,n})_{n \in \mathbb{N}}$ such that $h_{i,n} \xrightarrow{a.e.} f_i$, as $n \to \infty$ for $1 \leq i \leq k$. Then apply Lemma 4.2.10 and Proposition 4.2.16. $\dashv$

We are in a position now to establish that convergence in measure actually yields a Banach space. But we have to be careful with functions which differ on a set of measure zero, rendering the resulting space non-Hausdorff. Since functions which are equal except on a set of measure zero may be considered to be equal, we simply factor them out, obtaining $F(X, \mathcal{A})$ as the factor space $\mathcal{F}(X, \mathcal{A})/{=_\mu}$ of the space $\mathcal{F}(X, \mathcal{A})$ of all measurable functions with respect to $=_\mu$. Then this is a real vector space again, because the algebraic operations on the equivalence classes are well defined. Note that we have $\delta(f, g) = \delta(f', g')$, provided $f =_\mu g$ and $f' =_\mu g'$. We identify again the class $[f]$ with f. Define $F(X, \mathcal{A})$

$$||f|| := \delta(f, 0)$$

for $f \in F(X, \mathcal{A})$.

Proposition 4.2.21 $\big(F(X, \mathcal{A}), ||\cdot||\big)$ *is a Banach space.*

Proof 1. It follows from Lemma 4.2.14 and the observation $\delta(f, 0) = 0 \Leftrightarrow f =_\mu 0$ that $||\cdot||$ is a norm, so we have to show that $F(X, \mathcal{A})$ is complete with this norm.

2. Let $(f_n)_{n \in \mathbb{N}}$ be a Cauchy sequence in $F(X, \mathcal{A})$; then we can find a strictly increasing sequence $(\ell_n)_{n \in \mathbb{N}}$ of integers such that $\delta(f_{\ell_n}, f_{\ell_{n+1}}) \leq 1/n^2$; hence

$$\mu\big(\{x \in X \mid |f_{\ell_n}(x) - f_{\ell_{n+1}}(x)| > 1/n^2\}\big) \leq 1/n^2.$$

Let $\epsilon > 0$ be given; then there exists $r \in \mathbb{N}$ with $\sum_{n \geq r} 1/n^2 < \epsilon$; hence we have

$$\bigcap_{n \in \mathbb{N}} \bigcup_{m,k \geq n} \{x \in X \mid |f_{\ell_m}(x) - f_{\ell_k}(x)| > \epsilon\}$$
$$\subseteq \bigcup_{n \geq k} \{x \in X \mid |f_{\ell_n}(x) - f_{\ell_{n+1}}(x)| < 1/n^2\},$$

if $k \geq r$. Thus

$$\mu\left(\bigcap_{n \in \mathbb{N}} \bigcup_{m,k \geq n} \{x \in X \mid |f_{\ell_m}(x) - f_{\ell_k}(x)| > \epsilon\}\right) \leq \sum_{n \geq k} 1/n^2 \to 0,$$

as $k \to \infty$. Hence $(f_{\ell_n})_{n \in \mathbb{N}}$ is an a.e.Cauchy sequence which converges a.e. to some $f \in F(X, \mathcal{A})$, which by Proposition 4.2.16 implies that $f_n \xrightarrow{i.m.} f$. $\dashv$

A consequence of $\big(F(X, \mathcal{A}), ||\cdot||\big)$ being a Banach space is that $\mathcal{F}(X, \mathcal{A})$ is complete with respect to convergence in measure for any finite measure μ on $\mathcal{A}$. Thus for any sequence $(f_n)_{n \in \mathbb{N}}$ of functions such that for any given $\epsilon > 0$, there exists n_0 such that $\mu(\{x \in X \mid |f_n(x) - f_m(x)| > \epsilon\}) < \epsilon$ for all $n, m \geq n_0$, we can find $f \in \mathcal{F}(X, \mathcal{A})$ such that $f_n \xrightarrow{i.m.} f$ with respect to μ.

We will deal with measurable real-valued functions again and in greater detail in Sect. 4.11; then we will have integration as a powerful tool at our disposal, and we will know more about Hilbert spaces.

Now we turn to the study of σ-algebras and focus on those which have a countable set as their generator.

4.3 Countably Generated σ-Algebras

Fix a measurable space $(X, \mathcal{A})$. The σ-algebra $\mathcal{A}$ is said to be *countably generated* iff there exists countable $\mathcal{A}_0$ such that $\mathcal{A} = \sigma(\mathcal{A}_0)$.

Example 4.3.1 Let (X, τ) be a topological space with a countable basis. Then $\mathcal{B}(X)$ is countably generated. In fact, if τ_0 is the countable basis for τ, then each open set G can be written as $G = \bigcup_{n \in \mathbb{N}} G_n$ with $(G_n)_{n \in \mathbb{N}} \subseteq \tau_0$, and thus each open set is an element of $\sigma(\tau_0)$; consequently, $\mathcal{B}(X) = \sigma(\tau_0)$. ✌

The observation in Example 4.3.1 implies that the Borel sets for a separable metric space, in particular for the Polish spaces soon to be introduced, are countably generated. Having a countable dense subset for a metric space, we can use the corresponding base for a fairly helpful characterization of the Borel sets. The next lemma says that the Borel sets are in this case generated by a countable collection of open balls.

Lemma 4.3.2 *Let X be a separable metric space with metric d. $B(x, r)$ is the open ball with radius r and center x. Then*

$$\mathcal{B}(X) = \sigma(\{B(x, r) \mid r > 0 \text{ rational}, x \in D\}),$$

where D is countable and dense.

Proof Because an open ball is an open set, we infer that

$$\sigma(\{B(x, r) \mid r > 0 \text{ rational}, x \in D\}) \subseteq \mathcal{B}(X).$$

Conversely, let G be open. Then there exists a sequence $(B_n)_{n\in\mathbb{N}}$ of open balls with rational radii such that $\bigcup_{n\in\mathbb{N}} B_n = G$, accounting for the other inclusion. $\dashv$

Also the characterization of Borel sets in a metric space as the closure of the open (closed) sets under countable unions and countable intersections will be occasionally helpful.

Lemma 4.3.3 *The Borel sets in a metric space X are the smallest collection of sets that contains the open (closed) sets and that are closed under countable unions and countable intersections.*

Proof The smallest collection $\mathcal{G}$ of sets that contains the open sets and that is closed under countable unions and countable intersections is closed under complementation. This is so since each closed set F can be written as the countable intersection $\bigcap_{n\in\mathbb{N}}\{x \in X \mid d(x, F) < 1/n\}$ of open sets; in other words, F is a G_δ-set (see page 346). Thus $\mathcal{B}(X) \subseteq \mathcal{G}$; on the other hand, $\mathcal{G} \subseteq \mathcal{B}(X)$ by construction. $\dashv$

The property of being countably generated is, however, not hereditary for a σ-algebra—a sub-σ-algebra of a countably generated σ-algebra is not necessarily countably generated. This is demonstrated by the following example. Incidentally, we will see in Example 4.4.28 that the intersection of two countably generated σ-algebras need not be countably generated again. This indicates that having a countable generator is a fickle property which has to be observed closely.

Example 4.3.4 Let

$$\mathcal{C} := \{A \subseteq \mathbb{R} \mid A \text{ or } \mathbb{R} \setminus A \text{ is countable}\}.$$

This σ-algebra is usually referred to the *countable–cocountable σ-algebra*. Clearly, $\mathcal{C} \subseteq \mathcal{B}(\mathbb{R})$, and $\mathcal{B}(\mathbb{R})$ is countably generated by Example 4.3.1. But $\mathcal{C}$ is not countably generated. Assume that it is, so let $\mathcal{C}_0$ be a countable generator for $\mathcal{C}$; we may assume that every element of $\mathcal{C}_0$ is countable. Put $A := \bigcup \mathcal{C}_0$; then $A \in \mathcal{C}$, since A is countable. But

$$\mathcal{D} := \{B \subseteq \mathbb{R} \mid B \subseteq A \text{ or } B \subseteq \mathbb{R} \setminus A\}$$

is a σ-algebra, and $\mathcal{D} = \sigma(\mathcal{C}_0)$. On the other hand, there exists $a \in \mathbb{R}$ with $a \notin A$; thus $A \cup \{a\} \in \mathcal{C}$, but $A \cup \{a\} \notin \mathcal{D}$, a contradiction. ✌

Although the entire σ-algebra may not be countably generated, we find for each element of a σ-algebra a countable generator:

Lemma 4.3.5 *Let $\mathcal{A}$ be a σ-algebra on a set X which is generated by family $\mathcal{G}$ of subsets. Then we can find for each $A \in \mathcal{A}$ a countable subset $\mathcal{G}_0 \subseteq \mathcal{G}$ such that $A \in \sigma(\mathcal{G}_0)$.*

Proof Let $\mathcal{D}$ be the set of all $A \in \mathcal{A}$ for which the assertion is true; then $\mathcal{D}$ is closed under complements, and $\mathcal{G} \subseteq \mathcal{A}$. Moreover, $\mathcal{D}$ is closed under countable unions, since the union of a countable family of countable sets is countable again. Hence $\mathcal{D}$ is a σ-algebra which contains $\mathcal{G}$; hence it contains $\mathcal{A} = \sigma(\mathcal{G})$. ⊣

This has a fairly interesting and somewhat unexpected consequence, which will be of use later on. Recall from page 436 or from Example 2.2.4 that $\mathcal{A} \otimes \mathcal{B}$ is the smallest σ-algebra on $X \times Y$ which contains for measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ all measurable rectangles $A \times B$ with $A \in \mathcal{A}$ and $B \in \mathcal{B}$. In particular, $\mathcal{P}(X) \otimes \mathcal{P}(X)$ is generated by $\{A \times B \mid A, B \subseteq X\}$. One may be tempted to assume that this σ-algebra is the same as $\mathcal{P}(X \times X)$, but this is not always the case, because we have

Proposition 4.3.6 *Denote by Δ_X the diagonal $\{\langle x, x\rangle \mid x \in X\}$ for a set X. Then $\Delta_X \in \mathcal{P}(X) \otimes \mathcal{P}(X)$ implies that the cardinality of X does not exceed that of $\mathcal{P}(\mathbb{N})$.*

Proof Assume $\Delta_X \in \mathcal{P}(X) \otimes \mathcal{P}(X)$; then there exists a countable family $\mathcal{C} \subseteq \mathcal{P}(X)$ such that $\Delta_X \in \sigma(\{A \times B \mid A, B \in \mathcal{C}\})$. The map $q : x \mapsto \{C \in \mathcal{C} \mid x \in C\}$ from X to $\mathcal{P}(\mathcal{C})$ is injective. In fact, suppose it is not; then there exists $x \neq x'$ with $x \in C \Leftrightarrow x' \in C$ for all $C \in \mathcal{C}$, so we have for all $C \in \mathcal{C}$ that either $\{x, x'\} \subseteq C$ or $\{x, x'\} \cap C = \emptyset$, so that the pairs $\langle x, x\rangle$ and $\langle x', x'\rangle$ never occur alone in any $A \times B$ with $A, B \in \mathcal{C}$. Hence Δ_X cannot be a member of $\sigma(\{A \times B \mid A, B \in \mathcal{C}\})$, a contradiction. As a consequence, X cannot have more elements that $\mathcal{P}(\mathbb{N})$. $\dashv$

Now we are in a position to show that we cannot conclude from the fact for a subset $S \subseteq X \times Y$ that S is product measurable whenever all its cuts are measurable; see Lemma 4.1.8.

Example 4.3.7 Let X be a set, the cardinality of which is greater than that of $\mathcal{P}(\mathbb{N})$, and let $\Delta := \{\langle x, x\rangle \mid x \in X\} \subseteq X \times X$ be the diagonal of X. Then $\Delta_x = \{x\} = \Delta^x$ for all $x \in X$; thus Δ_x and Δ^x are both members of $\mathcal{P}(X)$, but $\Delta \notin \mathcal{P}(X) \otimes \mathcal{P}(X)$ by Proposition 4.3.6. ✌

Among the countably generated measurable spaces, those are of interest which permit to *separate points*, so that if $x \neq x'$, we can find $A \in \mathcal{C}$ with $x \in A$ and $x' \notin A$; they are called separable. Formally

Definition 4.3.8 *The σ-algebra $\mathcal{A}$ is called* separable *iff it is countably generated and if for any two different elements of X there exists a measurable set $A \in \mathcal{A}$ which contains exactly one of them. The measurable space $(X, \mathcal{A})$ is called separable iff its σ-algebra $\mathcal{A}$ is separable.*

The argumentation from Proposition 4.3.6 yields

Corollary 4.3.9 *Let $\mathcal{A}$ be a separable σ-algebra over the set X with $\mathcal{A} = \sigma(\mathcal{A}_0)$ for $\mathcal{A}_0$ countable. Then $\mathcal{A}_0$ separates points, and $\Delta_X \in \mathcal{A} \otimes \mathcal{A}$.*

Proof Because $\mathcal{A}$ separates points, we obtain from Example 4.1.5 that $\equiv_{\mathcal{A}_0} = \Delta_X$, where $\equiv_{\mathcal{A}_0}$ is the equivalence relation defined by $\mathcal{A}_0$. So $\mathcal{A}_0$ separates points. The representation

$$X \times X \setminus \Delta_X = \bigcup_{A \in \mathcal{A}_0} A \times (X \setminus A) \cup (X \setminus A) \times A.$$

now yields $\Delta_X \in \mathcal{A} \otimes \mathcal{A}$. $\dashv$

In fact, we can say even more.

Proposition 4.3.10 *A separable measurable space $(X, \mathcal{A})$ is isomorphic to $(X, \mathcal{B}(X))$ with the Borel sets coming from a metric d on X such that (X, d) has a separable metric space.*

Proof 1. Let $\mathcal{A}_0 = \{A_n \mid n \in \mathbb{N}\}$ be the countable generator for $\mathcal{A}$ which separates points. Define

$$(M, \mathcal{M}) := \prod_{n\in\mathbb{N}} (\{0,1\}, \mathcal{P}(\{0,1\}))$$

as the product of many countable copies of the discrete space $\big(\{0,1\}, \mathcal{P}(\{0,1\})\big)$. Then $\mathcal{M}$ has as a basis the cylinder sets $\{Z_v \mid v \in \{0,1\}^k$ for some $k \in \mathbb{N}\}$ with $Z_v := \{(t_n)_{n\in\mathbb{N}} \in M \mid \langle m_1, \ldots, m_k\rangle = v\}$ for $v \in \{0,1\}^k$; see page 436. Define $f : X \to M$ through $f(x) := (\chi_{A_n}(x))_{n\in\mathbb{N}}$, then f is injective, because $\mathcal{A}_0$ separates points. Put $Q := f\big[X\big]$, and $\mathcal{Q} := \mathcal{M} \cap Q$, the trace of $\mathcal{M}$ on Q.

Now let $Y_v := Z_v \cap Q$ be an element of the generator for $\mathcal{Q}$ with $v = \langle m_1, \ldots, m_k\rangle$; then $f^{-1}\big[Y_v\big] = \bigcap_{j=1}^k C_j$ with $C_j := A_j$, if $m_j = 1$, and $C_j := X \setminus A_j$ otherwise. Consequently, $f : X \to Q$ is $\mathcal{A}$-$\mathcal{Q}$-measurable.

2. Put for $x, y \in X$

$$d(x, y) := \sum_{n\in\mathbb{N}} 2^{-n} \cdot \big|\chi_{A_n}(x) - \chi_{A_n}(y)\big|;$$

then d is a metric on X which has

$$\mathcal{G} := \big\{\bigcap_{j\in F} B_j \mid B_j \in \mathcal{A}_0 \text{ or } X \setminus B_j \in \mathcal{A}_0, F \subseteq \mathbb{N} \text{ is finite}\big\}$$

as a countable basis. In fact, let $G \subseteq X$ be open; given $x \in G$, there exists $\epsilon > 0$ such that the open ball $B(x, \epsilon) := \{x' \in X \mid d(x, x') < \epsilon\}$ with center x and radius ϵ is contained in G. Now choose k with $2^{-k} < \epsilon$, and put $v := \langle x_1, \ldots, x_k\rangle$; then $x \in \bigcap_{j=1}^k B_j \subseteq B(x, \epsilon)$. This argument shows also that $\mathcal{A} = \mathcal{B}(X)$.

3. Because (X, d) has a countable basis, it is a separable metric space. The map $f : X \to Q$ is a bijection which is measurable, and f^{-1} is measurable as well. This is so because $\{A \in \mathcal{A} \mid f\big[A\big] \in \mathcal{Q}\}$ is a σ-algebra which contains the basis $\mathcal{G}$. $\dashv$

The representation is due to Mackey. It gives the representation of separable measurable spaces as subspaces of the countable product of the discrete space $(\{0,1\}, \mathcal{P}(\{0,1\})$. This space is also a compact metric space, so we may say that a separable measurable space is isomorphic to a subspace of a compact metric space. We will make use of this observation later on.

By the way, this innocently looking statement has some remarkable consequences for our context. Just as an appetizer

Corollary 4.3.11 *Let $(X, \mathcal{A})$ be a separable measurable space. If $f_i : X_i \to X$ is $\mathcal{A}_i$-$\mathcal{A}$-measurable, where $(X_i, \mathcal{A}_i)$ is a measurable space ($i = 1, 2$), then*

$$f_1^{-1}[\mathcal{A}] \otimes f_2^{-1}[\mathcal{A}] = (f_1 \times f_2)^{-1}[\mathcal{A} \otimes \mathcal{A}]$$

holds.

Proof The product σ-algebra $\mathcal{A} \otimes \mathcal{A}$ is generated by the rectangles $B_1 \times B_2$ with B_i taken from some generator $\mathcal{B}_0$ for $\mathcal{B}$, $i = 1, 2$. Since $(f_1 \times f_2)^{-1}[B_1 \times B_2] = f_1^{-1}[B_1] \times f_2^{-1}[B_2]$, we see that $(f_1 \times f_2)^{-1}[\mathcal{B} \otimes \mathcal{B}] \subseteq f_1^{-1}[\mathcal{B}] \otimes f_2^{-1}[\mathcal{B}]$. This is true without the assumption of separability. Now let τ be a second countable metric topology on Y with $\mathcal{B} = \mathcal{B}(\tau)$ and let τ_0 be a countable base for the topology. Then

$$\beta_p := \{T_1 \times T_2 \mid T_1, T_2 \in \tau_0\}$$

is a countable base for the product topology $\tau \otimes \tau$, and (this is the crucial property)

$$\mathcal{B} \otimes \mathcal{B} = \mathcal{B}(Y \times Y, \tau \otimes \tau)$$

holds: Because the projections from $X \times Y$ to X and to Y are measurable, we observe $\mathcal{B} \otimes \mathcal{B} \subseteq \mathcal{B}(Y \times Y, \tau \otimes \tau)$; because β_p is a countable base for the product topology $\tau \otimes \tau$, we infer the other inclusion.

Since for $T_1, T_2 \in \tau_0$ clearly

$$f_1^{-1}[T_1] \times f_2^{-1}[T_2] \in (f_1 \times f_2)^{-1}\left[\tau_p\right] \subseteq (f_1 \times f_2)^{-1}[\mathcal{B} \otimes \mathcal{B}]$$

holds, the nontrivial inclusion is inferred from the fact that the smallest σ-algebra containing $\{f_1^{-1}[T_1] \times f_2^{-1}[T_2] \mid T_1, T_2 \in \tau_0\}$ equals $f_1^{-1}[\mathcal{B}] \otimes f_2^{-1}[\mathcal{B}]$. $\dashv$

Given a measurable function into a separable measurable space, we find that its kernel (defined on page 124) yields a measurable subset in the product of its domain. We will use the kernel for many a construction, so this little observation is quite helpful.

Corollary 4.3.12 *Let $f : X \to Y$ be a $\mathcal{A}$-$\mathcal{B}$-measurable map, where $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ are measurable spaces, the latter being separable. Then the kernel* $\ker(f)$ *of f is a member of $\mathcal{A} \otimes \mathcal{A}$.*

Proof Exercise 4.8. $\dashv$

The observation, made in the proof of Proposition 4.3.6, that it may not always be possible to separate two different elements in a measurable space through a measurable set leads there to a contradiction. Nevertheless it leads also to an interesting notion.

Definition 4.3.13 *The set $A \in \mathcal{A}$ is called an* atom *of $\mathcal{A}$ iff $B \subseteq A$ implies $B = \emptyset$ or $B = A$ for all $B \in \mathcal{A}$.*

For example, each singleton set $\{x\}$ is an atom for the σ-algebra $\mathcal{P}(X)$. Clearly, being an atom depends also on the σ-algebra. If A is an atom, we have alternatively $B \subseteq A$ or $B \cap A = \emptyset$ for all $B \in \mathcal{A}$; this is more radical than being a μ-atom, which merely restricts the values of $\mu(B)$ for measurable $B \subseteq A$ to 0 or $\mu(A)$. Certainly, if A is an atom and $\mu(A) > 0$, then A is a μ-atom.

For a countably generated σ-algebra, atoms are easily identified.

Proposition 4.3.14 *Let $\mathcal{A}_0 = \{A_n \mid n \in \mathbb{N}\}$ be a countable generator of $\mathcal{A}$, and define $A_\alpha := \bigcap_{n \in \mathbb{N}} A_n^{\alpha_n}$ for $\alpha \in \{0,1\}^{\mathbb{N}}$, where $A^0 := A, A^1 := X \setminus A$. Then $\{A_\alpha \mid \alpha \in \{0,1\}^{\mathbb{N}}, A_\alpha \neq \emptyset\}$ is the set of all atoms of $\mathcal{A}$.*

Proof Assume that there exist in $\mathcal{A}$ two different nonempty subsets B_1, B_2 of A_α, and take $y_1 \in B_1, y_2 \in B_2$. Then $y_1 \equiv_{\mathcal{A}_0} y_2$, but $y_1 \not\equiv_{\mathcal{A}} y_2$, contradicting the observation in Example 4.1.5. Hence A_α is an atom. Let $x \in A_\alpha$; then A_α is the equivalence class of x with respect to the equivalence relation $\equiv_{\mathcal{A}_0}$ and hence with respect to $\mathcal{A}$. Thus each atom is given by some A_α. $\dashv$

Incidentally, this gives another proof that the countable–cocountable σ-algebra over $\mathbb{R}$ is not countably generated. Assume it is generated by $\{A_n \mid n \in \mathbb{N}\}$; then

$$H := \bigcap\{A_n \mid A_n \text{ is cocountable}\} \cap \bigcap\{\mathbb{R} \setminus A_n \mid A_n \text{ is countable}\}$$

is an atom, but H is also cocountable. This is a contradiction to H being an atom.

We relate atoms to measurable maps.

Lemma 4.3.15 *Let $f : X \to \mathbb{R}$ be $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable. If $A \in \mathcal{A}$ is an atom of $\mathcal{A}$, then f is constant on A.*

Proof Assume that we can find $x_1, x_2 \in A$ with $f(x_1) \neq f(x_2)$, say, $f(x_1) < c < f(x_2)$. Then $\{x \in A \mid f(x) < c\}$ and $\{x \in A \mid f(x) > c\}$ are two nonempty disjoint measurable subsets of A. This contradicts A being an atom. $\dashv$

We will specialize now our view of measurable spaces to the Borel sets of Polish spaces and their more general cousins, viz., analytic sets. Before we do that, however, we show that for important cases the Borel sets $\mathcal{B}(X \times Y)$ of $X \times Y$ coincide with the product $\mathcal{B}(X) \otimes \mathcal{B}(Y)$. We know from Proposition 4.3.6 that this is not always the case.

Proposition 4.3.16 *Let X and Y be topological spaces such that Y has a countable base. Then $\mathcal{B}(X \times Y) = \mathcal{B}(X) \otimes \mathcal{B}(Y)$.*

Proof 1. We show first that each open set $G \subseteq X \times Y$ is a member of $\mathcal{B}(X) \otimes \mathcal{B}(Y)$. This shows that $\mathcal{B}(X \times Y) \subseteq \mathcal{B}(X) \otimes \mathcal{B}(Y)$. If $\{V_n \mid n \in \mathbb{N}\}$ is a countable base of Y, we can write $G = \bigcup U_\alpha \times V_m$ for suitable open sets $U_\alpha \subseteq X$ and $V_m \subseteq Y$ taken from the base. Now define for fixed $m \in \mathbb{N}$ the set $W_m := \bigcup\{U_\alpha \mid U_\alpha \times V_m \subseteq G\}$, then W_m is open, and $G = \bigcup W_m \times V_m$. This is a countable union of elements from $\mathcal{B}(X) \otimes \mathcal{B}(Y)$, showing that each open set in $X \times Y$ is contained in $\mathcal{B}(X) \otimes \mathcal{B}(Y)$.

2. We claim that $A \times Y \in \mathcal{B}(X \times Y)$ for all $A \in \mathcal{B}(X)$ and that $X \times B \in \mathcal{B}(X \times Y)$ for all $B \in \mathcal{B}(Y)$. Assume that we have established these claims; then we infer that $A \times B = (A \times Y) \cap (X \times B) \in \mathcal{B}(X \times Y)$, from which will follow

$$\mathcal{B}(X) \otimes \mathcal{B}(Y) = \sigma(\{A \times B \mid A \in \mathcal{B}(X), B \in \mathcal{B}(Y)\}) \subseteq \mathcal{B}(X \times Y).$$

We use the principle of good sets for proving the first assertion; the second is established in exactly the same way, so we will not bother with it. In fact, put

$$\mathcal{G} := \{A \in \mathcal{B}(X) \mid A \times Y \in \mathcal{B}(X \times Y)\}.$$

Then $\mathcal{G}$ is a σ-algebra which contains the open sets. This is so because for an open set $H \subseteq X$ the set $H \times Y$ is open in $X \times Y$, thus a Borel set in $X \times Y$. But then $\mathcal{G}$ contains the σ-algebra generated by the open sets, hence $\mathcal{G} = \mathcal{B}(X)$, and we are done. $\dashv$

The proof shows that $\mathcal{B}(X) \otimes \mathcal{B}(Y) \subseteq \mathcal{B}(X \times Y)$ always holds and that it is the converse inclusion which is sometimes critical. We will apply Proposition 4.3.16, for example, when we have a separable metric or even a Polish space as one of the factors.

4.3.1 Borel Sets in Polish and Analytic Spaces

General measurable spaces and even separable metric spaces are sometimes too general for supporting specific structures. We deal with Polish and analytic spaces which are general enough to support interesting applications, but have specific properties which help establish vital properties. We remind the reader first of some basic facts and provide them some helpful tools for working with Polish spaces and their more general cousins, analytic spaces.

An immediate consequence of Lemma 4.1.4 is that continuity implies Borel measurability.

Lemma 4.3.17 *Let (X_1, τ_1) and (X_2, τ_2) be topological spaces. Then $f : X_1 \to X_2$ is $\mathcal{B}(\tau_1)$-$\mathcal{B}(\tau_2)$-measurable, provided f is τ_1-τ_2-continuous.* $\dashv$

We note for later use that the limit of a sequence of measurable functions into a metric space is measurable again; see Exercise 4.13.

Proposition 4.3.18 *Let $(X, \mathcal{A})$ be a measurable, (Y, d) a metric space, and $(f_n)_{n\in\mathbb{N}}$ a sequence of $\mathcal{A}$-$\mathcal{B}(Y)$-measurable functions $f_n : X \to Y$. Then*

- *the set $C := \{x \in X \mid (f_n(x))_{n\in\mathbb{N}}$ exists$\}$ is measurable,*
- *$f(x) := \lim_{n\to\infty} f_n(x)$ defines a $\mathcal{A} \cap C$-$\mathcal{B}(Y)$-measurable map $f : C \to Y$.*

$\dashv$

Neither general topological spaces nor metric spaces offer a structure rich enough for the study of the transition systems that we will enter into. We need to restrict the class of topological spaces to a particularly interesting class of spaces that are traditionally called *Polish*.

As far as notation goes, we will write down a topological or a metric space without its adornment through a topology or a metric, unless this becomes really necessary.

Remember that a metric space (X, d) is called *complete* iff each d-Cauchy sequence has a limit. Recall also that completeness is really a property of the metric rather than the underlying topological space, so a metrizable space may be complete with one metric and incomplete with another one; see Example 3.5.18. In contrast, having a countable base is a topological property which is invariant under the different metrics the topology may admit.

Definition 4.3.19 *A* Polish space *X is a topological space, the topology of which is metrizable through a complete metric, and which has a countable base, or, equivalently, a countable dense subset.* Polish space

Familiar spaces are Polish, as these examples show.

Example 4.3.20 The real $\mathbb{R}$ with their usual topology, which is induced by the open intervals, is a Polish space. The open unit interval $]0, 1[$ with the usual topology induced by the open intervals forms a Polish space.

The latter comes probably as a surprise, because $]0, 1[$ is known not to be complete with the usual metric. But all we need is a dense subset; here we take of course the rationals $\mathbb{Q} \cap]0, 1[$, and a complete metric that generates the topology. Define

$$d(x, y) := \left| \ln \frac{x}{1-x} - \ln \frac{y}{1-y} \right|;$$

then this is a complete metric for $]0, 1[$. This is so since $x \mapsto \ln(x/(1-x))$ is a continuous bijection from $]0, 1[$ to $\mathbb{R}$, and the inverse $y \mapsto e^y/(1+e^y)$ is also a continuous bijection. ✌

Lemma 4.3.21 *Let X be a Polish space, and assume that $F \subseteq X$ is closed; then the subspace F is Polish as well.*

Proof F is complete by Lemma 3.5.21. The topology that F inherits from X has a countable base and is metrizable, so F has a countable dense subset, too. $\dashv$

Lemma 4.3.22 *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of Polish spaces, then their product and their coproduct are Polish spaces.*

Proof Assume that the topology τ_n on X_n is metrized through metric d_n, where it may be assumed that $d_n \leq 1$ holds (otherwise use for

τ_n the complete metric $d_n(x, y)/(1 + d_n(x, y)))$. Then (see Proposition 3.5.4)

$$d\big((x_n)_{n\in\mathbb{N}}, (y_n)_{n\in\mathbb{N}}\big) := \sum_{n\in\mathbb{N}} 2^{-n} d_n(x_n, y_n)$$

is a complete metric for the product topology $\prod_{n\in\mathbb{N}} \tau_n$. For the coproduct, define the complete metric

$$d(x, y) := \begin{cases} 2, & \text{if } x \in X_n, y \in X_m, n \neq m \\ d_n(x, y), & \text{if } x, y \in X_n. \end{cases}$$

All this is established through standard arguments. $\dashv$

Example 4.3.23 The set $\mathbb{N}$ of natural numbers with the discrete topology is a Polish space on account of being the topological sum of its elements. Thus the set $\mathbb{N}^\infty$ of all infinite sequences is a Polish space. The sets

$$\Theta_\alpha := \{\tau \in \mathbb{N}^\infty \mid \alpha \text{ is an initial piece of } \tau\}$$

Θ_α for $\alpha \in \mathbb{N}^*$, the free monoid generated by $\mathbb{N}$, constitute a base for the product topology. ✌

This last example will be discussed in much greater detail later on. It permits occasionally reducing the discussion of properties for general Polish spaces to an investigation of the corresponding properties of $\mathbb{N}^\infty$, the structure of the latter space being more easily accessible than that of a general space. We apply Example 4.3.23 directly to show that *all* open subsets of a metric space X with a countable base can be represented through a *single* open set in $\mathbb{N}^\infty \times X$, similarly for closed sets.

Proposition 4.3.24 *Let X be a separable metric space. Then there exist an open set $U \subseteq \mathbb{N}^\infty \times X$ and a closed set $F \subseteq \mathbb{N}^\infty \times X$ with these properties:*

a. For each open set $G \subseteq X$, there exists $t \in \mathbb{N}^\infty$ such that $G = U_t$.

b. For each closed set $C \subseteq X$, there exists $t \in \mathbb{N}^\infty$ such that $C = F_t$.

Proof 0. It is enough to establish the property for open sets; taking complements will prove it for closed ones.

1. Let $(V_n)_{n\in\mathbb{N}}$ be a basis for the open sets in X with $V_n \neq \emptyset$ for all $n \in \mathbb{N}$. Define

$$U := \{\langle t, x\rangle \mid t \in \mathbb{N}^{\mathbb{N}}, x \in \bigcup_{n\in\mathbb{N}} V_{t_n}\};$$

then $U \subseteq \mathbb{N}^{\infty} \times X$ is open. In fact, let $\langle t, x\rangle \in U$, then there exists $n \in \mathbb{N}$ with $x \in V_n$, thus $\langle t, x\rangle \in \Theta_n \times V_n \subseteq U$, and $\Theta_n \times V_n$ is open in the product.

2. Let $G \subseteq X$ be open. Because $(V_n)_{n\in\mathbb{N}}$ is a basis for the topology, there exists a sequence $t \in \mathbb{N}^{\infty}$ with $G = \bigcup_{n\in\mathbb{N}} V_{t_n} = U_t$. $\dashv$

The set U is usually called a *universally open set*, similar for F, which is universally closed. These universal sets will be used rather heavily when we discuss analytic sets. Universally open

We have seen that a closed subset of a Polish space is a Polish space in its own right; a similar argument shows that an open subset of a Polish space is Polish as well. Both observations turn out to be special cases of the characterization of Polish subspaces through G_δ-sets.

We recall for this characterization an auxiliary statement which permits the extension of a continuous map from a subspace to a G_δ-set containing it—just far enough to be interesting to us. This has been given in Lemma 3.5.24, to which we refer.

This technical Lemma is an important step in establishing a far- reaching characterization of subspaces of Polish spaces that are Polish in their own right. A subset X of a Polish space is a Polish space itself iff it is a G_δ-set. We will present Kuratowski's proof for it. It is not difficult to show that X must be a G_δ-set, using Lemma 3.5.24. This is done in Lemma 4.3.25.

The tricky part, however, is the converse, and at its very center is the following idea: Assume that we have represented $X = \bigcap_{k\in\mathbb{N}} G_k$ with each G_k open, and assume that we have a Cauchy sequence $(x_n)_{n\in\mathbb{N}} \subseteq X$ with $x_n \to x$. How do we prevent x from being outside X? Well, what we will do is to set up Kuratowski's trap, preventing the sequence to wander off. The trap is a new complete and equivalent metric D, which makes it impossible for the sequence to behave in an undesired way. So if x is trapped to be an element of X, we may conclude that X is complete, and the assertion may be established. Kuratowski's trap

Before we begin with the easier half, we fix a Polish space Y and a complete metric d on Y.

Lemma 4.3.25 *If $X \subseteq Y$ is a Polish space, then X is a G_δ-set.*

Proof X is complete and hence closed in Y. The identity $id_X : X \to Y$ can be extended continuously by Lemma 3.5.24 to a G_δ-set G with $X \subseteq G \subseteq X^a$, thus $G = X$, so X is a G_δ-set. ⊣

Now let $X = \bigcap_{k\in\mathbb{N}} G_k$ with G_k open for all $k \in \mathbb{N}$. In order to prepare for Kuratowski's trap, we define

$$f_k(x, x') := \left|\frac{1}{d(x, Y \setminus G_k)} - \frac{1}{d(x', Y \setminus G_k)}\right|$$

for $x, x' \in X$. Because G_k is open, we have $x \in G_k$ iff $d(x, Y \setminus G_k) > 0$, so f_k is a finite and continuous function on $X \times X$. Now let

$$\begin{aligned} F_k(x, x') &:= \frac{f_k(x, x')}{1 + f_k(x, x')}, \\ D(x, x') &:= d(x, x') + \sum_{k\in\mathbb{N}} 2^{-k} \cdot F_k(x, x') \end{aligned}$$

for $x, x' \in X$.

Then D is a metric on X, and the metrics d and D are equivalent on X. Because $d(x, x') \leq D(x, x')$, it is clear that the identity $id : (X, D) \to (X, d)$ is continuous, so it remains to show that $id : (X, d) \to (X, D)$ is continuous. Let $x \in X$ be given, and let $\epsilon > 0$; then we find $\ell \in \mathbb{N}$ such that $\sum_{k>\ell} 2^{-j} \cdot F_k(x, x') < \epsilon/3$ for all $x' \in X$. For $k = 1, \ldots, \ell$ there exists δ_j such that $F_j(x, x') < \epsilon/(3 \cdot \ell)$, whenever $d(x, x') < \delta_j$, since $x \mapsto d(x, Y \setminus G_j)$ is positive and uniformly continuous. Thus define $\delta := \min\{\epsilon/3, \delta_1, \ldots, \delta_\ell\}$; then $d(x, x') < \delta$ implies

$$D(x, x') \leq d(x, x') + \sum_{k=1}^{\ell} 2^{-j} \cdot F_j(x, x') + \frac{\epsilon}{3} < \frac{\epsilon}{3} + \sum_{k=1}^{\ell} \frac{\epsilon}{3 \cdot \ell} + \frac{\epsilon}{3} = \epsilon.$$

Thus (X, d) and (X, D) have in fact the same open sets.

When establishing that (X, D) is complete, we spring Kuratowski's trap. Let $(x_n)_{n\in\mathbb{N}}$ be a D-Cauchy sequence. Then this sequence is also a d-Cauchy sequence, and thus we find $x \in Y$ such that $x_n \to x$, because (Y, d) is complete. We claim that $x \in X$. In fact, if $x \notin X$, we find G_ℓ

with $x \notin G_\ell$, so that we can find for each $\epsilon > 0$ some index $n_\epsilon \in \mathbb{N}$ with $F_\ell(x_n, x_m) \geq 1 - \epsilon$ for $n, m \geq n_\epsilon$. But then $D(x_n, x_m) \geq (1-\epsilon)/2^\ell$ for $n, m \geq n_\epsilon$, so that $(x_n)_{n\in\mathbb{N}}$ cannot be a D-Cauchy sequence. Consequently, X is complete and hence closed.

Thus we have established

Theorem 4.3.26 *Let Y be a Polish space. Then the subspace $X \subseteq Y$ is a Polish space iff X is a G_δ-set.* $\dashv$

In particular, open and closed subsets of Polish spaces are Polish spaces in their subspace topology. Conversely, each Polish space can be represented as a G_δ-set in the *Hilbert cube* $[0,1]^\infty$; this is the famous and very useful characterization of Polish spaces due to Alexandrov [Kur66, III.33.VI].

Theorem 4.3.27 ***(Alexandrov)*** *Let X be a separable metric space; then X is homeomorphic to a subspace of the Hilbert cube. If X is Polish, this subspace is a G_δ.*

Proof 0. The idea is to take a countable and dense subset D of X and to map each element $x \in X$ to the distance it has to each element of D. Since we may assume without loss of generality that the metric is bounded, say, by 1, this yields an embedding into the cube $[0,1]^\infty$, which is compact by Tihonov's Theorem 3.2.12. This map is investigated, and Theorem 4.3.26 is applied. Idea

1. We may and do assume again that the metric d is bounded by 1. Let $(x_n)_{n\in\mathbb{N}}$ be a countable and dense subset of X, and put

$$f(x) := \langle d(x, x_1), d(x, x_2), \ldots \rangle.$$

Then f is injective and continuous. Define $g : f\big[X\big] \to X$ as f^{-1}; then g is continuous as well: Assume that $f(y_m) \to f(y)$ for some y; hence $\lim_{m\to\infty} d(y_m, x_n) = d(y, x_n)$ for each $n \in \mathbb{N}$. Since $(x_n)_{n\in\mathbb{N}}$ is dense, we find for a given $\epsilon > 0$ an index n with $d(y, x_n) < \epsilon$; by construction we find for n an index m_0 with $d(y_m, x_n) < \epsilon$ whenever $m > m_0$. Thus $d(y_m, y) < 2 \cdot \epsilon$ for $m > m_0$, so that $y_m \to y$. This demonstrates that g is continuous; thus f is a homeomorphism.

2. If X is Polish, $f\big[X\big] \subseteq [0,1]^\infty$ is Polish as well. Thus the second assertion follows from Theorem 4.3.26. $\dashv$

Compact metric spaces are Polish. It is inferred from Tihonov's Theorem that the Hilbert cube $[0,1]^\infty$ is compact, because the unit interval

$[0,1]$ is compact by the Heine–Borel Theorem 1.5.46. Thus Alexandrov's Theorem 4.3.27 embeds a Polish space as a G_δ into a compact metric space, the closure of which will be compact.

4.3.2 Manipulating Polish Topologies

We will show now that a Borel map between Polish spaces can be turned into a continuous map. Specifically, we will show that, given a measurable map between Polish spaces, we can find on the domain a finer Polish topology with the same Borel sets which renders the map continuous. This will be established through a sequence of auxiliary statements, each of which will be of interest and of use in its own right.

We fix for the discussion to follow a Polish space X with topology τ. Recall that a set is *clopen* in a topological space iff it is both closed and open.

Lemma 4.3.28 *Let F be a closed set in X. Then there exists a Polish topology τ' such that $\tau \subseteq \tau'$ (hence τ' is finer than τ), F is clopen in τ', and $\mathcal{B}(\tau) = \mathcal{B}(\tau')$.*

Proof Both F and $X \setminus F$ are Polish by Theorem 4.3.26, so the topological sum of these Polish spaces is Polish again by Lemma 4.3.22. The sum topology is the desired topology. $\dashv$

We will now add a sequence of certain Borel sets to the topology; this will happen step by step, so we should know how to manipulate a sequence of Polish topologies. This is explained now.

Lemma 4.3.29 *Let $(\tau_n)_{n\in\mathbb{N}}$ be a sequence of Polish topologies τ_n with $\tau \subseteq \tau_n$.*

1. *The topology τ_∞ generated by $\bigcup_{n\in\mathbb{N}} \tau_n$ is Polish.*
2. *If $\tau_n \subseteq \mathcal{B}(\tau)$, then $\mathcal{B}(\tau_\infty) = \mathcal{B}(\tau)$.*

Proof 1. The product $\prod_{n\in\mathbb{N}}(X_n, \tau_n)$ is by Lemma 4.3.22 a Polish space, where $X_n = X$ for all n. Define the map $f : X \to \prod_{n\in\mathbb{N}} X_n$ through $x \mapsto \langle x, x, \ldots\rangle$; then f is τ_∞-$\prod_{n\in\mathbb{N}} \tau_n$-continuous by construction. One infers that $f\big[X\big]$ is a closed subset of $\prod_{n\in\mathbb{N}} X_n$: If $(x_n)_{n\in\mathbb{N}} \notin f\big[X\big]$, take $x_i \neq x_j$ with $i < j$, and let G_i and G_j be disjoint open neighborhoods of x_i resp. x_j. Then

$$\prod_{\ell<i} X_\ell \times G_i \times \prod_{i<\ell<j} X_\ell \times G_j \times \prod_{\ell>j} X_\ell$$

is an open neighborhood of $(x_n)_{n\in\mathbb{N}}$ that is disjoint from $f\left[X\right]$. By Lemma 4.3.21, the latter set is Polish. On the other hand, f is a homeomorphism between (X, τ_∞) and $f\left[X\right]$, which establishes part 1.

2. τ_n has a countable basis $\{U_{i,n} \mid i \in \mathbb{N}\}$, with $U_{i,n} \in \mathcal{B}(\tau)$, since $\tau_n \subseteq \mathcal{B}(\tau)$. This implies that τ_∞ has $\{U_{i,n} \mid i, n \in \mathbb{N}\}$ as a countable basis, which entails $\mathcal{B}(\tau_\infty) \subseteq \mathcal{B}(\tau)$. The other inclusion is obvious, giving part 2. $\dashv$

As a consequence, we may add a Borel set to a Polish topology as a clopen set without destroying the property of the space to be Polish or changing the Borel sets. This is extended now to sequences of Borel sets, as we will see now.

Proposition 4.3.30 *If $(B_n)_{n\in\mathbb{N}}$ is a sequence of Borel sets in X, then there exists a Polish topology τ_0 on X such that τ_0 is finer than τ, τ and τ_0 have the same Borel sets, and each B_n is clopen in τ_0.*

Proof 1. We show first that we may add just one Borel set to the topology without changing the Borel sets. In fact, call a Borel set $B \in \mathcal{B}(\tau)$ *neat* if there exists a Polish topology τ_B that is finer than τ such that B is clopen with respect to τ_B, and $\mathcal{B}(\tau) = \mathcal{B}(\tau_B)$. Put

$$\mathcal{H} := \{B \in \mathcal{B}(\tau) \mid B \text{ is neat}\}.$$

Then $\tau \subseteq \mathcal{H}$, and each closed set is a member of $\mathcal{H}$ by Lemma 4.3.28. Furthermore, $\mathcal{H}$ is closed under complements by construction and closed under countable unions by Lemma 4.3.29. Thus we may now infer that $\mathcal{H} = \mathcal{B}(\tau)$, so that each Borel set is neat.

2. Now construct inductively Polish topologies τ_n that are finer than τ with $\mathcal{B}(\tau) = \mathcal{B}(\tau_n)$. Start with $\tau_0 := \tau$. Adding B_{n+1} to the Polish topology τ_n according to the first part yields a finer Polish topology τ_{n+1} with the same Borel sets. Thus the assertion follows from Lemma 4.3.29. $\dashv$

We are in a position now which permits turning a Borel map into a continuous one, whenever the domain is Polish and the range is a second countable metric space.

Make Borel f continuous

Proposition 4.3.31 *Let Y be a separable metric space with topology ϑ. If $f : X \to Y$ is a $\mathcal{B}(\tau)$-$\mathcal{B}(\vartheta)$-Borel measurable map, then there exists a Polish topology τ' on X such that τ' is finer than τ, τ and τ' have the same Borel sets, and f is τ'-ϑ continuous.*

Proof The metric topology ϑ is generated from the countable basis $(H_n)_{n\in\mathbb{N}}$. Construct from the Borel sets $f^{-1}\left[H_n\right]$ and from τ a Polish topology τ' according to Proposition 4.3.30. Because $f^{-1}\left[H_n\right] \in \tau'$ for all $n \in \mathbb{N}$, the inverse image of each open set from ϑ is τ'-open; hence f is τ'-ϑ continuous. The construction entails τ and τ' having the same Borel sets. $\dashv$

This property is most useful, because it permits rendering measurable maps continuous, when they go into a second countable metric space (thus in particular into a Polish space).

As a preparation for dealing with analytic sets, we will show now that each Borel subset of the Polish space X is the continuous image of $\mathbb{N}^\infty$. We begin with a reduction of the problem space: It is sufficient to establish this property for closed sets. This is justified by the following observation, the proof of which is sketched by indicating the technical arguments without giving, however, the somewhat laborious but not particularly nutritious or difficult details.

Lemma 4.3.32 *Assume that each closed set in the Polish space X is a continuous image of $\mathbb{N}^\infty$. Then each Borel set of X is a continuous image of $\mathbb{N}^\infty$.*

Plan

Proof (Sketch) 0. The plan of the proof is to extend the assumption that each closed set is the image of $\mathbb{N}^\infty$ to hold for all Borel sets. This assumption is verified in the subsequent Proposition 4.3.33; note that a closed set is a Polish space in its own right by Theorem 4.3.26. The extension is done by applying the principle of good sets.

1. Let

$$\mathcal{G} := \{B \in \mathcal{B}(X) \mid B = f\left[\mathbb{N}^\infty\right] \text{ for } f : \mathbb{N}^\infty \to X \text{ continuous}\}$$

be the set of all good guys. Then $\mathcal{G}$ contains by assumption all closed sets. We show that $\mathcal{G}$ is closed under countable unions and countable intersections. Then the assertion will follow from Lemma 4.3.3.

2. Suppose $B_n = f_n\left[\mathbb{N}^\infty\right]$ for the continuous map f_n; then

$$\mathbb{M} := \{\langle t_1, t_2, \ldots\rangle \mid f_1(t_1) = f_2(t_2) = \ldots\}$$

is a closed subset of $(\mathbb{N}^\infty)^\infty$, and defining $f : \langle t_1, t_2, \ldots\rangle \mapsto f_1(t_1)$ yields a continuous map $f : \mathbb{M} \to X$ with $f\big[\mathbb{M}\big] = \bigcap_{n\in\mathbb{N}} B_n$. $\mathbb{M}$ is homeomorphic to $\mathbb{N}^\infty$. Thus $\mathcal{G}$ is closed under countable intersections.

3. We show that $\mathcal{G}$ is closed also under countable unions. In fact, let $B_n \in \mathcal{G}$ such that $B_n = f_n\big[\mathbb{N}^\mathbb{N}\big]$ with $f_n : \mathbb{N}^\mathbb{N} \to X$ continuous. Define

$$f : \begin{cases} \mathbb{N}^\mathbb{N} & \to X \\ \langle n, t_1, t_2, \ldots, \rangle & \mapsto f_n(t_1, t_2, \ldots). \end{cases}$$

Thus

$$f\big[\mathbb{N}^\mathbb{N}\big] = \bigcup_{n\in\mathbb{N}} f_n\big[\mathbb{N}^\mathbb{N}\big] = \bigcup_{n\in\mathbb{N}} B_n.$$

Moreover, f is continuous. If $G \subseteq X$ is open, we have $f^{-1}\big[G\big] = \bigcup_{n\in\mathbb{N}}\{n\} \times f_n^{-1}\big[G\big]$. Since $f_n^{-1}\big[G\big]$ is open for each $n \in \mathbb{N}$, we conclude that $f^{-1}\big[G\big]$ is open, so that f is indeed continuous. Thus $\mathcal{G}$ is closed under countable unions, and the assertion follows from Lemma 4.3.3. $\dashv$

Thus it is sufficient to show that each closed subset of a Polish space is the continuous image on $\mathbb{N}^\infty$. And since a closed subset of a Polish space is Polish itself by Theorem 4.3.26, we may restrict our attention to Polish spaces proper.

Proposition 4.3.33 *For Polish X there exists a continuous map $f : \mathbb{N}^\infty \to X$ with $f\big[\mathbb{N}^\infty\big] = X$.*

Proof 0. We will define recursively a sequence of closed sets indexed by elements of $\mathbb{N}^*$ that will enable us to define a continuous map on $\mathbb{N}^\infty$.

1. Let d be a metric that makes X complete. Represent X as $\bigcup_{n\in\mathbb{N}} A_n$ with closed sets $A_n \neq \emptyset$ such that the diameter $\text{diam}(A_n) < 1$ for each $n \in \mathbb{N}$. Assume that for a word $\alpha \in \mathbb{N}^*$ of length k the closed set $A_\alpha \neq \emptyset$ is defined, and write $A_\alpha = \bigcup_{n\in\mathbb{N}} A_{\alpha n}$ with closed sets $A_{\alpha n} \neq \emptyset$ such that $\text{diam}(A_{\alpha n}) < 1/(k+1)$ for $n \in \mathbb{N}$. This yields for every $t = \langle n_1, n_2, \ldots\rangle \in \mathbb{N}^\infty$ a sequence of nonempty closed sets $(A_{n_1 n_2 .. n_k})_{k\in\mathbb{N}}$ with diameter $\text{diam}(A_{n_1 n_2 .. n_k}) < 1/k$. Because the

metric is complete, $\bigcap_{k\in\mathbb{N}} A_{n_1n_2..n_k}$ contains exactly one point, which is defined to be $f(t)$.

2. This construction renders $f : \mathbb{N}^\infty \to X$ well defined. We can find for each $x \in X$ an index $n_1' \in \mathbb{N}$ with $x \in A_{n_1'}$, an index n_2' with $x \in A_{n_1'n_2'}$, and so on. The map just defined is onto, so that $f(\langle n_1', n_2', n_3', \ldots\rangle) = x$ for some $t' := \langle n_1', n_2', n_3', \ldots\rangle \in \mathbb{N}^\infty$. Suppose $\epsilon > 0$ is given. Since the diameters of the sets $(A_{n_1n_2\ldots n_k})_{k\in\mathbb{N}}$ tend to 0, we can find $k_0 \in \mathbb{N}$ with $\mathrm{diam}(A_{n_1'n_2'..n_k'}) < \epsilon$ for all $k > k_0$. Put $\alpha' := n_1'n_2'\ldots n_{k_0}'$; then $\Theta_{\alpha'}$ is an open neighborhood of t' with $f\big[\Theta_{\alpha'}\big] \subseteq B_{\epsilon,d}(f(t'))$. Thus we find for an arbitrary open neighborhood V of $f(t')$ an open neighborhood U of t' with $f\big[U\big] \subseteq V$, equivalently, $U \subseteq f^{-1}\big[V\big]$. Hence f is continuous. $\dashv$

Proposition 4.3.33 permits sometimes the transfer of arguments pertaining to Polish spaces from arguments using infinite sequences. Thus a specific space is studied instead of an abstractly given one, the former permitting some rather special constructions. This will be capitalized on for the investigation of some astonishing properties of analytic sets, which we will study now.

4.4 Analytic Sets and Spaces

We will deal now systematically with analytic sets and spaces. One of the core results of this section will be the Lusin Separation Theorem, which permits to separate two disjoint analytic sets through disjoint Borel sets, and its immediate consequence, the Souslin Theorem, which says that a set which is both analytic and co-analytic is Borel. These beautiful results turn out to be very helpful, e.g., in the investigation of Markov transition systems. In addition, they permit to state and prove a weak form of Kuratowski's Isomorphism Theorem, which says that a measurable bijection between two Polish spaces is an isomorphisms (hence its inverse is measurable as well).

But first we give the definition of analytic and co-analytic sets for a Polish space X.

Definition 4.4.1 *An* analytic set *in X is the projection of a Borel subset of $X \times X$. The complement of an analytic set is called a* co-analytic *set.*

One may wonder whether these projections are Borel sets, but we will show in a moment that there are strictly more analytic sets than Borel sets, whenever the underlying Polish space is uncountable. Thus analytic sets are a proper extension to Borel sets. On the other hand, analytic sets arise fairly naturally, for example, from factoring Polish spaces through equivalence relations that are generated from a countable collection of Borel sets. We will see this in Proposition 4.4.22. Consequently it is sometimes more adequate to consider analytic sets rather than their Borel cousins, e.g., when the equivalence of states in a transition system is at stake.

This is a first characterization of analytic sets (using π_X for the projection to X).

Proposition 4.4.2 *Let X be a Polish space. Then the following statements are equivalent for $A \subseteq X$:*

1. *A is analytic.*
2. *There exist a Polish space Y and a Borel set $B \subseteq X \times Y$ with $A = \pi_X\left[B\right]$.*
3. *There exists a continuous map $f : \mathbb{N}^\infty \to X$ with $f\left[\mathbb{N}^\infty\right] = A$.*
4. *$A = \pi_X\left[C\right]$ for a closed subset $C \subseteq X \times \mathbb{N}^\infty$.*

Proof The implication 1 $\Rightarrow$ 2 is trivial, and 2 $\Rightarrow$ 3 follows from Proposition 4.3.33: $B = g\left[\mathbb{N}^\infty\right]$ for some continuous map $g : \mathbb{N}^\infty \to X \times Y$, so put $f := \pi_X \circ g$. We obtain 3 $\Rightarrow$ 4 from the observation that the graph $\{\langle t, f(t)\rangle \mid t \in \mathbb{N}^\infty\}$ of f is a closed subset of $\mathbb{N}^\infty \times X$, the first projection of which equals A. Finally, 4 $\Rightarrow$ 1 is obtained again from Proposition 4.3.33. $\dashv$

As an immediate consequence, we obtain that a Borel set is analytic. Just for the record

Corollary 4.4.3 *Each Borel set in a Polish space is analytic.*

Proof Proposition 4.4.2 together with Proposition 4.3.33. $\dashv$

The converse does not hold, as we will show now. This statement is not only of interest in its own right. Historically it initiated the study of analytic and co-analytic sets as a separate discipline in set theory (what is called now descriptive set theory).

Proposition 4.4.4 *Let X be an uncountable Polish space. Then there exists an analytic set that is not Borel.*

We show as a preparation for the proof of Proposition 4.4.4 that analytic sets are closed under countable unions, intersections, and direct and inverse images of Borel maps. Before doing that, we establish a simple but useful property of the graphs of measurable maps.

Lemma 4.4.5 *Let $(M, \mathcal{M})$ be a measurable space and $f : M \to Z$ be a $\mathcal{M}$-$\mathcal{B}(Z)$-measurable map, where Z is a separable metric space. The graph* $\text{graph}(f)$ *of f is a member if $\mathcal{M} \otimes \mathcal{B}(Z)$.*

Proof Exercise 4.9. ⊣

Analytic sets have closure properties that are similar to those of Borel sets, but not quite the same. Suspiciously missing from the list below is the closure under complementation (which will give rise to Souslin's Theorem). This, of course, is different from Borel sets.

Proposition 4.4.6 *Analytic sets in a Polish space X are closed under countable unions and countable intersections. If Y is another Polish space, with analytic sets $A \subseteq X$ and $B \subseteq Y$ and $f : X \to Y$ is a Borel map, then $f\left[A\right] \subseteq Y$ is analytic in Y, and $f^{-1}\left[B\right]$ is analytic in X.*

Proof 1. Using the characterization of analytic sets in Proposition 4.4.2, it is shown exactly as in the proof to Lemma 4.3.32 that analytic sets are closed under countable unions and under countable intersections. We trust that the reader will reproduce those arguments here.

2. Note first that for $A \subseteq X$ the set $Y \times A$ is analytic in the Polish space $Y \times X$ by Proposition 4.4.2. In fact, $A = \pi_X\left[B\right]$ with $B \subseteq X \times X$ Borel by the first part; hence $Y \times A = \pi_{Y \times X}\left[Y \times B\right]$ with $Y \times B \subseteq Y \times X \times X$ Borel, which is analytic by the second part. Since $y \in f\left[A\right]$ iff $\langle x, y\rangle \in \text{graph}(f)$ for some $x \in A$, we write

$$f\left[A\right] = \pi_Y\left[Y \times A \cap \{\langle y, x\rangle \mid \langle x, y\rangle \in \text{graph}(f)\}\right].$$

The set $\{\langle y, x\rangle \mid \langle x, y\rangle \in \text{graph}(f)\}$ is Borel in $Y \times X$ by Lemma 4.4.5, so the assertion follows for the direct image. The assertion is proved in exactly the same way for the inverse image. ⊣

Again, the proof for Proposition 4.4.4 will be sketched only, delegating the very technical details to Srivastava's book [Sri98, Sect. 2.5]. We give, however, the argument for the case that the space under consideration is our prototypical space $\mathbb{N}^\infty$ through a pretty diagonal argument using a universal set. From this and the structural arguments used so far, the reader has no difficulties filling in the details under the leadership of the text mentioned.

Proof (of Proposition 4.4.4) 1. We will deal with the case $X = \mathbb{N}^\infty$ first and apply a diagonal argument. Let $F \subseteq \mathbb{N}^\infty \times (\mathbb{N}^\infty \times \mathbb{N}^\infty)$ be a universal closed set according to Proposition 4.3.24. Thus each closed set $C \subseteq \mathbb{N}^\infty \times \mathbb{N}^\infty$ can be represented as $C = F_t$ for some $t \in \mathbb{N}^\infty$. Taking first projections, we conclude that there exists a universal analytic set $U \subseteq \mathbb{N}^\infty \times \mathbb{N}^\infty$ such that each analytic set $A \subseteq \mathbb{N}^\infty$ can be represented as U_t for some $t \in \mathbb{N}^\infty$. In fact, we can write $A = \left(\pi'_{\mathbb{N}^\mathbb{N} \times \mathbb{N}^\mathbb{N}}\left[F\right]\right)_t$ with $\pi'_{\mathbb{N}^\mathbb{N} \times \mathbb{N}^\mathbb{N}}$ as the first projection of $(\mathbb{N}^\mathbb{N} \times \mathbb{N}^\mathbb{N}) \times \mathbb{N}^\mathbb{N}$.

Now set

$$A := \{\zeta \mid \langle \zeta, \zeta \rangle \in U\}.$$

Because analytic sets are closed under inverse images of Borel maps by Proposition 4.4.6, A is an analytic set. Suppose that A is a Borel set; then $\mathbb{N}^\infty \setminus A$ is also a Borel set, hence analytic. Thus we find $\xi \in \mathbb{N}^\infty$ such that $\mathbb{N}^\infty \setminus A = U_\xi$. But now

$$\xi \in A \Leftrightarrow \langle \xi, \xi \rangle \in U \Leftrightarrow \xi \in U_\xi \Leftrightarrow \xi \in \mathbb{N}^\infty \setminus A.$$

This is a contradiction.

2. The general case is reduced to the one treated above by observing that an uncountable Polish space contains a homeomorphic copy on $\mathbb{N}^\infty$. But since we are interested mainly in showing that analytic sets are strictly more general than Borel sets, we refrain from a very technical discussion of this case and refer the reader to [Sri98, Remark 2.6.5]. $\dashv$

4.4.1 Souslin's Separation Theorem

The representation of an analytic set through a continuous map on $\mathbb{N}^\infty$ has the remarkable consequence that we can separate two disjoint analytic sets by disjoint Borel sets (Lusin's Separation Theorem). This in turn implies a beautiful characterization of Borel sets due to Souslin which says that an analytic set is Borel iff its complement is analytic as well. Since the latter characterization will be most valuable to us, we will discuss it in greater detail now.

We start with Lusin's Separation Theorem.

Theorem 4.4.7 *Given disjoint analytic sets A and B in a Polish space X, there exist disjoint Borel sets E and F with $A \subseteq E$ and $B \subseteq F$.*

This is the plan

Proof 0. We investigate first what it means for two analytic sets to be separated by Borel sets and show that this property carries over to sequences of analytic sets. From this observation we argue by contradiction what it means that two analytic sets cannot be separated in a way which we want them to. Here the representation of analytic sets as continuous images on $\mathbb{N}^\infty$ is used. We construct in this manner a decreasing sequence of open sets with smaller and smaller diameters, arriving eventually at a contradiction.

1. Call two analytic sets A and B *separated by Borel sets* iff there exist disjoint Borel sets E and F with $A \subseteq E$ and $B \subseteq F$. Observe that if two sequences $(A_n)_{n\in\mathbb{N}}$ and $(B_n)_{n\in\mathbb{N}}$ have the property that A_m and B_n can be separated by Borel sets for all $m, n \in \mathbb{N}$, then $\bigcup_{n\in\mathbb{N}} A_n$ and $\bigcup_{m\in\mathbb{N}} B_m$ can also be separated by Borel sets. In fact, if $E_{m,n}$ and $F_{m,n}$ separate A_n and B_m, then $E := \bigcap_{m\in\mathbb{N}} \bigcup_{n\in\mathbb{N}} E_{m,n}$ and $F := \bigcup_{m\in\mathbb{N}} \bigcap_{n\in\mathbb{N}} F_{m,n}$ separate $\bigcup_{n\in\mathbb{N}} A_n$ and $\bigcup_{m\in\mathbb{N}} B_m$.

2. Now suppose that $A = f\left[\mathbb{N}^\infty\right]$ and $B = g\left[\mathbb{N}^\infty\right]$ cannot be separated by Borel sets, where $f, g : \mathbb{N}^\infty \to X$ are continuous and chosen according to Proposition 4.4.2. Because $\mathbb{N}^\infty = \bigcup_{j\in\mathbb{N}} \Theta_j$, ($\Theta_\alpha$ is defined in Example 4.3.23 for $\alpha \in \mathbb{N}^*$), we find indices k_1 and ℓ_1 such that $f\left[\Theta_{j_1}\right]$ and $g\left[\Theta_{\ell_1}\right]$ cannot be separated by Borel sets. For the same reason, there exist indices k_2 and ℓ_2 such that $f\left[\Theta_{j_1 j_2}\right]$ and $g\left[\Theta_{\ell_1 \ell_2}\right]$ cannot be separated by Borel sets. Continuing in this way, we define infinite sequences $\kappa := \langle k_1, k_2, \ldots\rangle$ and $\lambda := \langle \ell_1, \ell_2, \ldots\rangle$ such that for each $n \in \mathbb{N}$ the sets $f\left[\Theta_{j_1 j_2 \ldots j_n}\right]$ and $g\left[\Theta_{\ell_1 \ell_2 \ldots \ell_n}\right]$ cannot be separated by Borel sets.

Because $f(\kappa) \in A$ and $g(\lambda) \in B$, we know $f(\kappa) \neq g(\lambda)$, so we find $\epsilon > 0$ with $d(f(\kappa), g(\lambda)) < 2\cdot\epsilon$. But we may choose n large enough so that both $f\left[\Theta_{j_1 j_2 \ldots j_n}\right]$ and $g\left[\Theta_{\ell_1 \ell_2 \ldots \ell_n}\right]$ have a diameter smaller than ϵ each. This is a contradiction since we now have separated these sets by open balls. $\dashv$

We obtain as a consequence Souslin's Theorem.

Souslin's Theorem

Theorem 4.4.8 ***(Souslin)*** *Let A be an analytic set in a Polish space. If $X \setminus A$ is analytic, then A is a Borel set.*

Proof Let A and $X \setminus A$ be analytic; then they can be separated by disjoint Borel sets E with $A \subseteq E$ and F with $X \setminus A \subseteq F$ by Lusin's Theorem 4.4.7. Thus $A = E$ is a Borel set. $\dashv$

Souslin's Theorem is important when one wants to show that a set is a Borel set that is given, for example, through the image of another Borel set. A typical scenario for its use is establishing for a Borel set A and a Borel map $f : X \to Y$ that both $C = f\left[A\right]$ and $Y \setminus C = f\left[X \setminus A\right]$ hold. Then one infers from Proposition 4.4.6 that both C and $Y \setminus C$ are analytic and from Souslin's Theorem that A is a Borel set. This is a first simple example (see also Lemma 2.6.44).

Proposition 4.4.9 *Let $f : X \to Y$ be surjective and Borel measurable, where X and Y are Polish. Assume that $A \in \boldsymbol{\Sigma}_{\ker(f)}(\mathcal{B}(X))$; hence $A \in \mathcal{B}(X)$ is* $\ker(f)$*-invariant. Then $f\left[A\right] \in \mathcal{B}(Y)$.*

Proof Put $C := f\left[A\right]$, $D := f\left[X \setminus A\right]$; then both C and D are analytic sets by Proposition 4.4.6. Clearly $Y \setminus C \subseteq D$. For establishing the other inclusion, let $y \in D$; hence there exists $x \notin A$ with $y = f(x)$. But $y \notin C$, for otherwise there exists $x' \in A$ with $y = f(x')$, which implies $x \in A$. Thus $y \in Y \setminus C$; hence $D \subseteq Y \setminus C$, so that we have shown $D = Y \setminus C$. We infer $f\left[A\right] \in \mathcal{B}(Y)$ now from Theorem 4.4.8. $\dashv$

This yields the following observation as an immediate consequence. It will be extended to analytic spaces in Proposition 4.4.13 with essentially the same argument.

Corollary 4.4.10 *Let $f : X \to Y$ be measurable and bijective with X, Y Polish. Then f is a Borel isomorphism.* $\dashv$

We state finally Kuratowski's Isomorphism Theorem.

Theorem 4.4.11 *Any two Borel sets of the same cardinality contained in Polish spaces are Borel isomorphic.* $\dashv$

The proof requires a reduction to the Cantor ternary set, using the tools we have discussed here so far. Since giving the proof would lead us fairly deep into the Wonderland of Descriptive Set Theory, we do not give it here and refer rather to [Sri98, Theorem 3.3.13], [Kec94, Sect. 15.B] or [KM76, p. 442].

We make the properties of analytic sets a bit more widely available by introducing analytic spaces. Roughly, an analytic space is Borel isomorphic to an analytic set in a Polish space; to be more precise

Definition 4.4.12 *A measurable space* $(M, \mathcal{M})$ *is called an* analytic space *iff there exist a Polish space* X *and an analytic set* A *in* X *such that the measurable spaces* $(M, \mathcal{M})$ *and* $(A, \mathcal{B}(X) \cap A)$ *are Borel isomorphic. The elements of* $\mathcal{M}$ *are then called the* Borel sets of M. $\mathcal{M}$ *is denoted by* $\mathcal{B}(M)$.

We will omit the σ-algebra from the notation of an analytic space.

Analytic spaces share many favorable properties with analytic sets and with Polish spaces, but they are a wee bit more general: Whereas an analytic set lives in a Polish space, an analytic space does only require a Polish space to sit in the background somewhere and to be Borel isomorphic to it. This makes life considerably easier, since we are for this reason not obliged to present a Polish space directly when dealing with properties of analytic spaces. We will demonstrate the use (and power) of the structure theorems studied above for investigating properties of analytic spaces and their σ-algebras. The most helpful of these theorems will turn out to be Souslin's Theorem, which can be applied for showing that a set is a Borel set by demonstrating that it is an analytic set and that its complement is analytic as well.

Take a Borel measurable bijection between two Polish spaces. It is not a priori clear whether or not this map is an isomorphism. Souslin's Theorem gives a helpful hand here as well. We will need this property in a moment for a characterization of countably generated sub-σ-algebras of Borel sets, but it appears to be interesting in its own right.

Proposition 4.4.13 *Let* X *and* Y *be analytic spaces and* $f : X \to Y$ *be a bijection that is Borel measurable. Then* f *is a Borel isomorphism.*

Proof 1. It is no loss of generality to assume that we can find Polish spaces P and Q such that X and Y are subsets of P resp. Q. We want to show that $f\left[X \cap B\right]$ is a Borel set in Y, whenever $B \in \mathcal{B}(P)$ is a Borel set. For this we need to find a Borel set $G \in \mathcal{B}(Q)$ such that $f\left[X \cap B\right] = G \cap Q$.

2. Clearly, both $f\left[X \cap B\right]$ and $f\left[X \setminus B\right]$ are analytic sets in Q by Proposition 4.4.6, and because f is injective, they are disjoint. Thus we can find a Borel set $G \in \mathcal{B}(Q)$ with $f\left[X \cap B\right] \subseteq G \cap Y$ and

$f\left[X \setminus B\right] \subseteq Q \setminus (G \cap Y)$. Because f is surjective, we have $f\left[X \cap B\right] \cup f\left[X \setminus B\right] = Y$; thus $f\left[X \cap B\right] = G \cap Y$. $\dashv$

Separable measurable spaces are characterized through subsets of Polish spaces.

Lemma 4.4.14 *The measurable space $(M, \mathcal{M})$ is separable iff there exist a Polish space X and a subset $P \subseteq X$ such that the measurable spaces $(M, \mathcal{M})$ and $(P, \mathcal{B}(X) \cap P)$ are Borel isomorphic.*

It should be noted that we do not assume P to be a measurable subset of X.

Proof 1. Because $\mathcal{B}(X)$ is countably generated for a Polish space X by Lemma 4.3.2, the σ-algebra $\mathcal{B}(X) \cap P$ is countably generated. Since this property is not destroyed by Borel isomorphisms, the condition above is sufficient.

2. It is also necessary by Proposition 4.3.10, because $\prod_{n \in \mathbb{N}}\big(\{0, 1\}, \mathcal{P}(\{0, 1\})\big)$ is a Polish space by Lemma 4.3.22. $\dashv$

Thus analytic spaces are separable measurable spaces; see Definition 4.3.8.

Corollary 4.4.15 *An analytic space is a separable measurable space.* $\dashv$

Let us have a brief look at countably generated sub-σ-algebras of an analytic space. This will help establish, for example, that the factor space for a particularly interesting and important class of equivalence relations is an analytic space. The following statement, which is sometimes referred to as the *Unique Structure Theorem* [Arv76, Theorem 3.3.5], says essentially that the Borel sets of an analytic space are uniquely determined by being countably generated and by separating points. It comes as a consequence of our discussion of Borel isomorphisms.

Proposition 4.4.16 *Let X be an analytic space and $\mathcal{B}_0$ be a countably generated sub-σ-algebra of $\mathcal{B}(X)$ that separates points. Then $\mathcal{B}_0 = \mathcal{B}(X)$.*

Proof 1. $(X, \mathcal{B}_0)$ is a separable measurable space, so there exist a Polish space P and a subset $Y \subseteq P$ of P such that $(X, \mathcal{B}_0)$ is Borel isomorphic to $(Y, \mathcal{B}(P) \cap Y)$ by Lemma 4.4.14. Let f be this isomorphism; then $\mathcal{B}_0 = f^{-1}\left[\mathcal{B}(P) \cap Y\right]$.

2. f is a Borel map from $(X, \mathcal{B}(X))$ to $(Y, \mathcal{B}(P) \cap Y)$; thus Y is an analytic set with $\mathcal{B}(Y) = \mathcal{B}(X) \cap P$ by Proposition 4.4.14. By Proposition 4.4.6, f is an isomorphism; hence $\mathcal{B}(X) = f^{-1}\big[\mathcal{B}(P) \cap Y\big]$. But this establishes the assertion. $\dashv$

This gives an interesting characterization of measurable spaces to be analytic, provided they have a separating sequence of sets. Note that the sequence of sets in the following statement is required to separate points, but we do not assume that it generates the σ-algebra for the underlying space. The statement says that it does, actually.

Lemma 4.4.17 *Let X be analytic and $f : X \to Y$ be $\mathcal{B}(X)$-$\mathcal{B}$-measurable and onto for a measurable space $(Y, \mathcal{B})$, which has a sequence of sets in $\mathcal{B}$ that separates points. Then $(Y, \mathcal{B})$ is analytic.*

Plan

Proof 1. The idea is to show that an arbitrary measurable set is contained in the σ-algebra generated by the sequence in question. Thus let $(B_n)_{n \in \mathbb{N}}$ be the sequence of sets that separates points, take an arbitrary set $N \in \mathcal{B}$, and define the σ-algebra $\mathcal{B}_0 := \sigma(\{B_n \mid n \in \mathbb{N}\} \cup \{N\})$. We want to show that $N \in \sigma(\{B_n \mid n \in \mathbb{N}\})$, and we show this in a roundabout way by showing that $\mathcal{B} = \mathcal{B}(Y) = \mathcal{B}_0$. Here is how.

2. $(Y, \mathcal{B}_0)$ is a separable measurable space, so by Lemma 4.4.14 we can find a Polish space P with $Y \subseteq P$ and $\mathcal{B}_0$ as the trace of $\mathcal{B}(P)$ on Y. Proposition 4.4.6 tells us that $Y = f\big[X\big]$ is analytic with $\mathcal{B}_0 = \mathcal{B}(Y)$, and from Proposition 4.4.16 it follows that $\mathcal{B}(Y) = \sigma(\{B_n \mid n \in \mathbb{N}\})$. Thus $N \in \mathcal{B}(Y)$, and since $N \in \mathcal{B}$ is arbitrary, we conclude $\mathcal{B} \subseteq \mathcal{B}(Y)$; thus $\mathcal{B} \subseteq \mathcal{B}(Y) = \sigma(\{B_n \mid n \in \mathbb{N}\}) \subseteq \mathcal{B}$. $\dashv$

4.4.2 Smooth Equivalence Relations

We will use Lemma 4.4.17 for demonstrating that factoring an analytic space through a smooth equivalence relation yields an analytic space again. This class of relations will be defined now and briefly characterized here. We give a definition in terms of a determining sequence of Borel sets and relate other characterizations of smoothness in Lemma 4.4.21.

Definition 4.4.18 *Let X be an analytic space and ρ an equivalence relation on X. Then ρ is called* smooth *iff there exists a sequence $(A_n)_{n \in \mathbb{N}}$*

of Borel sets such that

Smooth equivalence relation

$$x \,\rho\, x' \Leftrightarrow \forall n \in \mathbb{N} : [x \in A_n \Leftrightarrow x' \in A_n].$$

$(A_n)_{n\in\mathbb{N}}$ *is said to* determine *the relation* ρ*.*

Example 4.4.19 Given an analytic space X, let $M : X \rightsquigarrow X$ be a transition kernel which interprets the modal logic presented in Example 4.1.11. Put for a formula φ and an element of x as usual $M, x \models \varphi$ iff $x \in [\![\varphi]\!]_M$, and thus $M, x \models \varphi$ indicates that formula φ is valid in state x. Define the equivalence relation $\sim$ on X through

$$x \sim x' \iff \forall \varphi : \left[M, x \models \varphi \text{ iff } M, x' \models .\varphi\right]$$

Thus x and x' cannot be separated through a formula of the logic. Because the logic has only countably many formulas, the relation is smooth with the countable set $\{[\![\varphi]\!]_M \mid \varphi \text{ is a formula}\}$ as determining the relation $\sim$. ✌

We obtain immediately from the definition that a smooth equivalence relation—seen as a subset of the Cartesian product—is a Borel set.

Corollary 4.4.20 *Let ρ be a smooth equivalence relation on the analytic space X; then ρ is a Borel subset of $X \times X$.*

Proof Suppose that $(A_n)_{n\in\mathbb{N}}$ determines ρ. Since $x \,\rho\, x'$ is false iff there exists $n \in \mathbb{N}$ with $\langle x, x'\rangle \in (A_n \times (X \setminus A_n)) \cup ((X \setminus A_n) \times A_n)$, we obtain

$$(X \times X) \setminus \rho = \bigcup_{n\in\mathbb{N}} \big(A_n \times (X \setminus A_n)\big) \cup \big((X \setminus A_n) \times A_n\big).$$

This is clearly a Borel set in $X \times X$. $\dashv$

The following characterization of smooth equivalence relations is sometimes helpful and shows that it is not necessary to focus on sequences of sets. It indicates that the kernels of Borel measurable maps and smooth relations are intimately related.

Lemma 4.4.21 *Let ρ be an equivalence relation on an analytic set X. Then these conditions are equivalent:*

1. *ρ is smooth.*
2. *There exists a sequence $(f_n)_{n\in\mathbb{N}}$ of Borel maps $f_n : X \to Z$ into an analytic space Z such that $\rho = \bigcap_{n\in\mathbb{N}} \ker(f_n)$.*

3. *There exists a Borel map $f : X \to Y$ into an analytic space Y with $\rho = \ker(f)$.*

Guide through the proof

Proof The proof is essentially an expansion of the definition of smoothness and the observation that the kernel of a Borel map into an analytic is determined through the inverse images of a countable generator. Here we go.

1 $\Rightarrow$ 2: Let $(A_n)_{n\in\mathbb{N}}$ determine ρ; then

$$\begin{aligned} x \; \rho \; x' \quad &\Leftrightarrow \quad \forall n \in \mathbb{N} : [x \in A_n \Leftrightarrow x' \in A_n] \\ &\Leftrightarrow \quad \forall n \in \mathbb{N} : \chi_{A_n}(x) = \chi_{A_n}(x'). \end{aligned}$$

Thus take $Z = \{0, 1\}$ and $f_n := \chi_{A_n}$.

2 $\Rightarrow$ 3: Put $Y := Z^\infty$. This is an analytic space in the product σ-algebra, and

$$f : \begin{cases} X & \to Y \\ x & \mapsto \big(f_n(x)\big)_{n\in\mathbb{N}} \end{cases}$$

is Borel measurable with $f(x) = f(x')$ iff $\forall n \in \mathbb{N} : f_n(x) = f_n(x')$.

3 $\Rightarrow$ 1: Since Y is analytic, it is separable; hence the Borel sets are generated through a sequence $(B_n)_{n\in\mathbb{N}}$ which separates points. Put $A_n := f^{-1}\big[B_n\big]$; then $(A_n)_{n\in\mathbb{N}}$ is a sequence of Borel sets, because the base sets B_n are Borel in Y and because f is Borel measurable. We claim that $(A_n)_{n\in\mathbb{N}}$ determines ρ:

$$\begin{aligned} f(x) = f(x') \quad &\Leftrightarrow \quad \forall n \in \mathbb{N} : [f(x) \in B_n \Leftrightarrow f(x') \in B_n] \\ &\qquad\; \text{(since } (B_n)_{n\in\mathbb{N}} \text{ separates points in } Z\text{)} \\ &\Leftrightarrow \quad \forall n \in \mathbb{N} : [x \in A_n \Leftrightarrow x' \in A_n]. \end{aligned}$$

Hence $\langle x, x'\rangle \in \ker(f)$ is equivalent to the pair $\langle x, x'\rangle$ being determined by a sequence of measurable sets. $\dashv$

Thus each smooth equivalence relation may be represented as the kernel of a Borel map and vice versa. This is an important property which we will put to use frequently.

The interest in analytic spaces comes from the fact that factoring an analytic space through a smooth equivalence relation will result in an analytic space again. This requires first and foremost the definition of a measurable structure induced by the relation. The natural choice is the structure imposed by the factor map. The final σ-algebra on X/ρ with respect to the Borel sets on X and the factor map η_ρ will be chosen; it is denoted by $\mathcal{B}(X)/\rho$. Recall that $\mathcal{B}(X)/\rho$ is the largest σ-algebra $\mathcal{C}$ on X/ρ rendering η_ρ a $\mathcal{B}(X)$-$\mathcal{C}$-measurable map. Then it turns out that $\mathcal{B}(X/\rho)$ coincides with $\mathcal{B}(X)/\rho$.

Proposition 4.4.22 *Let X be an analytic space, and assume that α is a smooth equivalence relation on X. Then X/α is an analytic space.*

Proof In accordance with the characterization of smooth relations in Lemma 4.4.21, we assume that α is given through a sequence $(f_n)_{n\in\mathbb{N}}$ of measurable maps $f_n : X \to \mathbb{R}$. The factor map is measurable and onto. Put $E_{n,r} := \{[x]_\alpha \mid x \in X, f_n(x) < r\}$; then $\mathcal{E} := \{E_{n,r} \mid n \in \mathbb{N}, r \in \mathbb{Q}\}$ is a countable set of element of the factor σ-algebra that separates points. The assertion now follows without difficulties from Lemma 4.4.17. $\dashv$

Let us have a look at invariant sets for an equivalence relation α. Recall that a subset $A \subseteq X$ is invariant for the equivalence relation α on X iff A is the union of α-equivalence classes; see page 452. Thus $A \subseteq X$ is α-invariant iff $x \in A$ and $x \;\alpha\; x'$ implies $x' \in A$. For example, if $\alpha = \ker(f)$, then A is α-invariant iff A is what we called f-invariant on page 221, i.e., iff $x \in A$ and $f(x) = f(x')$ imply $x' \in A$.

Denote by

$$A^\nabla := \bigcup\{[x]_\alpha \mid x \in A\}$$

the smallest α-invariant set containing A; then we have the representation A^∇

$$A^\nabla = \pi_2\big[\alpha \cap (X \times A)\big],$$

because $x' \in A^\nabla$ iff there exists x with $\langle x', x\rangle \in X \times A$.

An equivalence relation on X is called analytic resp. closed iff it constitutes an analytic resp. closed subset of the Cartesian product $X \times X$.

If X is a Polish space, we know that the smooth equivalence relation $\alpha \subseteq X \times X$ is a Borel subset by Corollary 4.4.20. We want to show that, conversely, each closed equivalence relation $\alpha \subseteq X \times X$ is smooth. This requires the identification of a countable set which generates the relation, and for this we require the following auxiliary statement. It may be called separation through invariant sets.

Lemma 4.4.23 *Let $\rho \subseteq X \times X$ be an analytic equivalence relation on the Polish space X with two disjoint analytic sets A and B. If B is ρ-invariant, then there exists a ρ-invariant Borel set C with $A \subseteq C$ and $B \cap C = \emptyset$.*

Approach

Proof 0. This is the plan for the proof. If D is an analytic set, D^{∇} is; this follows from the representation of D^{∇} above and from Proposition 4.4.6. It is fundamental for the rest of the proof. We construct a sequence $(A_n)_{n\in\mathbb{N}}$ of invariant analytic sets and a sequence $(B_n)_{n\in\mathbb{N}}$ of Borel sets with these properties: $A_n \subseteq B_n \subseteq A_{n+1}$; hence B_n is sandwiched between consecutive elements of the first sequence, $A \subseteq A_1$, and $B \cap B_n = \emptyset$ for all $n \in \mathbb{N}$.

1. Define $A_1 := A^{\nabla}$, then $A \subseteq A_1$, and A_1 is ρ-invariant. Since B is ρ-invariant as well, we conclude $A_1 \cap B = \emptyset$: If $x \in A_1 \cap B$, we find $x' \in A$ with $x \; \rho \; x'$; hence $x' \in B$, a contradiction. Proceeding inductively, assume that we have already chosen A_n and B_n with the properties described above, then put $A_{n+1} := B_n^{\nabla}$, then A_{n+1} is ρ-invariant and analytic, and also $A_{n+1} \cap B = \emptyset$ by the argument above. Hence we can find a Borel set B_{n+1} with $A_{n+1} \subseteq B_{n+1}$ and $B_{n+1} \cap B = \emptyset$.

2. Now put $C := \bigcup_{n\in\mathbb{N}} B_n$. Thus $C \in \mathcal{B}(X)$ and $C \cap B = \emptyset$, so it remains to show that C is ρ-invariant. Let $x \in C$ and $x \; \rho \; x'$. Since $x \in B_n \subseteq B_n^{\nabla} \subseteq B_{n+1}$, we conclude $x' \in B_{n+1} \subseteq C$, and we are done. $\dashv$

We use this observation now for a closed equivalence relation. Note that the assumption on being analytic in the proof above was made use of in order to establish that the invariant hull A^{∇} of an analytic set A is analytic again.

Proposition 4.4.24 *A closed equivalence relation on a Polish space is smooth.*

This is how the proof works

Proof 0. Let X be a Polish space and $\alpha \subseteq X \times X$ be a closed equivalence relation. We have to find a sequence $(A_n)_{n\in\mathbb{N}}$ of Borel sets which determines α. This will be constructed through a countable base for the topology in a somewhat roundabout manner.

1. Since X is Polish, it has a countable basis $\mathcal{G}$. Because α is closed, we can write

$$(X \times X) \setminus \alpha = \bigcup \{U_n \times U_m \mid U_n, U_m \in \mathcal{G}_0, U_n \cap U_m = \emptyset\}$$

for some countable subset $\mathcal{G}_0 \subseteq \mathcal{G}$. Fix U_n and U_m; then also U_n^∇ and U_m are disjoint. Select the invariant Borel set A_n such that $U_n \subseteq A_n$ and $A_n \cap U_m = \emptyset$; this is possible by Lemma 4.4.23.

2. We claim that

$$(X \times X) \setminus \alpha = \bigcup_{n \in \mathbb{N}} (A_n \times (X \setminus A_n)).$$

In fact, if $\langle x, x' \rangle \notin \alpha$, select U_n and U_m with $\langle x, x' \rangle \in U_n \times U_m \subseteq A_n \times (X \setminus A_n)$. If, conversely, $\langle x, x' \rangle \in A_n \times (X \setminus A_n)$, then $\langle x, x' \rangle \in \alpha$ implies by the invariance of A_n that $x' \in A_n$, a contradiction. $\dashv$

The Blackwell–Mackey Theorem analyzes those Borel sets that are unions of $\mathcal{A}$-atoms for a sub-σ-algebra $\mathcal{A} \subseteq \mathcal{B}(X)$. If $\mathcal{A}$ is countably generated by, say, $(A_n)_{n \in \mathbb{N}}$, then it is not difficult to see that an atom in $\mathcal{A}$ can be represented as $\bigcap_{i \in T} A_i \cap \bigcap_{i \in \mathbb{N} \setminus T} (X \setminus A_i)$ for a suitable subset $T \subseteq \mathbb{N}$; see Proposition 4.3.14. It constructs a measurable map f as it goes, so that the set under consideration is $\ker(f)$-invariant, which will be helpful in the application of the Souslin Theorem. But let us see.

Theorem 4.4.25 ***(Blackwell–Mackey)*** *Let X be an analytic space and $\mathcal{A} \subseteq \mathcal{B}(X)$ be a countably generated sub-σ-algebra of the Borel sets of X. If $B \subseteq X$ is a Borel set that is a union of atoms of $\mathcal{A}$, then $B \in \mathcal{A}$.*

The idea of the proof is to show that $f\left[B\right]$ and $f\left[X \setminus B\right]$ are disjoint analytic sets for the measurable map $f : X \to \{0, 1\}^\infty$ and to conclude that $B = f^{-1}\left[C\right]$ for some Borel set C, which will be supplied to us through Souslin's Theorem. Using $\{0, 1\}^\infty$ is suggested through the countable base for the σ-algebra, because we can then use the indicator functions of the base elements. The space $\{0, 1\}^\infty$ is compact and has well-known properties, so it is a pleasant enough choice. Idea of the proof

Proof Let $\mathcal{A}$ be generated by $(A_n)_{n \in \mathbb{N}}$, and define

$$f : X \to \{0, 1\}^\infty$$

through

$$x \mapsto \langle \chi_{A_1}(x), \chi_{A_2}(x), \chi_{A_3}(x), \dots \rangle.$$

Then f is $\mathcal{A}$-$\mathcal{B}(\{0, 1\}^\infty)$-measurable. We claim that $f\left[B\right]$ and $f\left[X \setminus B\right]$ are disjoint. Suppose not; then we find $t \in \{0, 1\}^\infty$ with $t = f(x) = f(x')$ for some $x \in B, x' \in X \setminus B$. Because B is the union of atoms, we find a subset $T \subseteq \mathbb{N}$ with $x \in A_n$, provided $n \in T$, and $x \notin A_n$, provided $n \notin T$. But since $f(x) = f(x')$, the same holds for x' as well, which means that $x' \in B$, contradicting the choice of x'.

Because $f\left[B\right]$ and $f\left[X \setminus B\right]$ are disjoint analytic sets, we find through Souslin's Theorem 4.4.8 a Borel set C with

$$f\left[B\right] \subseteq C, f\left[X \setminus B\right] \cap C = \emptyset.$$

Thus $f\left[B\right] = C$, so that $f^{-1}\left[f\left[B\right]\right] = f^{-1}\left[C\right] \in \mathcal{A}$. We show that $f^{-1}\left[f\left[B\right]\right] = B$. It is clear that $B \subseteq f^{-1}\left[f\left[B\right]\right]$, so assume that $f(b) \in f\left[B\right]$, so $f(b) = f(b')$ for some $b' \in B$. By construction, this means $b \in B$, since B is a union of atoms; hence $f^{-1}\left[f\left[B\right]\right] \subseteq B$. Consequently, $B = f^{-1}\left[C\right] \in \mathcal{A}$. $\dashv$

When investigating modal logics, one wants to be able to identify the σ-algebra which is defined by the validity sets of the formulas. This can be done through the Blackwell–Mackey Theorem and is formulated for general smooth equivalence relations with the proviso of being used for the logics later on.

Proposition 4.4.26 *Let ρ be a smooth equivalence relation on the Polish space X, and assume that $(A_n)_{n\in\mathbb{N}}$ generates ρ. Then*

1. *$\sigma(\{A_n \mid n \in \mathbb{N}\})$ is the σ-algebra $\boldsymbol{\Sigma}_\rho(X)$ of ρ-invariant Borel sets,*
2. *$\mathcal{B}(X/\rho) = \sigma(\{\eta_\rho\left[A_n\right] \mid n \in \mathbb{N}\}$.*

Proof 1. The σ-algebra $\boldsymbol{\Sigma}_\rho(X) = \boldsymbol{\Sigma}_\rho(\mathcal{B}(X))$ of ρ-invariant Borel sets is introduced on page 452. We have to show that $\boldsymbol{\Sigma}_\rho(X) = \sigma(\{A_n \mid n \in \mathbb{N}\}$:

"$\supseteq$": Each A_n is a ρ-invariant Borel set.

"$\subseteq$": Let B be an ρ-invariant Borel set; then $B = \bigcup_{b\in B} [b]_\rho$. Each class $[b]_\rho$ can be written as

$$[b]_\rho = \bigcap_{b\in A_n} A_n \cap \bigcap_{b\notin A_n} (X \setminus A_n);$$

thus $[b]_\rho \in \sigma(\{A_n \mid n \in \mathbb{N}\}$. Moreover, it is easy to see that the classes are the atoms of this σ-algebra, since we cannot find a proper nonempty ρ-invariant subset of an equivalence class. Thus the Blackwell–Mackey Theorem 4.4.25 shows that $B \in \sigma(\{A_n \mid n \in \mathbb{N}\})$.

2. Now let $\mathcal{E} := \sigma(\{\eta_\rho[A_n] \mid n \in \mathbb{N}\}$, and let $g : X/\rho \to P$ be $\mathcal{E}$-$\mathcal{P}$-measurable for an arbitrary measurable space $(P, \mathcal{P})$. Thus we have for all $C \in \mathcal{P}$

$$\begin{aligned} g^{-1}[C] \in \mathcal{E} &\Leftrightarrow \eta_\rho^{-1}[g^{-1}[C]] \\ &\quad \in \sigma(\{A_n \mid n \in \mathbb{N}\}) && \text{(since } A_n = \eta_\rho^{-1}[\eta_\rho[A_n]]\text{)} \\ &\Leftrightarrow \eta_\rho^{-1}[g^{-1}[C]] \in \mathcal{I} && \text{(part 1)} \\ &\Leftrightarrow \eta_\rho^{-1}[g^{-1}[C]] \in \mathcal{B}(X). \end{aligned}$$

Thus $\mathcal{E}$ is the final σ-algebra with respect to η_ρ and hence equals $\mathcal{B}(X/\rho)$. ⊣

The following example shows that the equivalence relation generated by a σ-algebra need not return this σ-algebra as its invariant sets, if the given σ-algebra is not countably generated. Proposition 4.4.26 assures us that this cannot happen in the countably generated case.

Example 4.4.27 Let $\mathcal{C}$ be the countable–cocountable σ-algebra on $\mathbb{R}$. The equivalence relation $\equiv_\mathcal{C}$ generated by $\mathcal{C}$ according to Example 4.1.5 is the identity. Hence it is smooth. The σ-algebra of $\equiv_\mathcal{C}$-invariant Borel sets equals the Borel set $\mathcal{B}(\mathbb{R})$, which is a proper superset of $\mathcal{C}$. ✌

The next example is a somewhat surprising application of the Blackwell–Mackey Theorem, taken from [RR81, Proposition 57]. It shows that the set of countably generated σ-algebras is not closed under finite intersections; hence it fails to be a lattice under inclusion.

Example 4.4.28 There exist two countably generated σ-algebras, the intersection of which is not countably generated. In fact, let $A \subseteq [0, 1]$ be an analytic set which is not Borel; then $\mathcal{B}(A)$ is countably generated by Corollary 4.4.15. Let $f : [0, 1] \to A$ be a bijection, and consider $\mathcal{C} := f^{-1}[\mathcal{B}(A)]$, which is countably generated as well. Then $\mathcal{D} := \mathcal{B}([0, 1]) \cap \mathcal{C}$ is a σ-algebra which has all singletons in $[0, 1]$ as atoms. Assume that $\mathcal{D}$ is countably generated; then $\mathcal{D} = \mathcal{B}([0, 1])$ by the Blackwell–Mackey Theorem 4.4.25. But this means that $\mathcal{C} = \mathcal{B}([0, 1])$, so that $f : [0, 1] \to A$ is a Borel isomorphism; hence A is a Borel set in $[0, 1]$, contradicting the assumption. ✌

Among the consequences of Example 4.4.28 is the observation that the set of smooth equivalence relations of a Polish space does not form a lattice under inclusion, but is usually only a $\cap$-semilattice, as the following example shows. Another consequence is mentioned in Exercise 4.19.

Example 4.4.29 The intersection $\alpha_1 \cap \alpha_2$ of two smooth equivalence relations α_1 and α_2 is smooth again: If α_i is generated by the Borel sets $\{A_{i,n} \mid n \in \mathbb{N}\}$ for $i = 1, 2$, then $\alpha_1 \cap \alpha_2$ is generated by the Borel sets $\{A_{i,n} \mid i = 1, 2, n \in \mathbb{N}\}$. But now take two countably generated σ-algebras $\mathcal{A}_i$, and let α_i be the equivalence relations determined by them; see Example 4.1.5. Then the σ-algebra $\alpha_1 \cup \alpha_2$ is generated by $\mathcal{A}_1 \cap \mathcal{A}_2$, which is by assumption not countably generated. Hence $\alpha_1 \cup \alpha_2$ is not smooth. ✌

We digress briefly and establish tameness (Definition 4.1.27) for smooth equivalence relations.

Proposition 4.4.30 *If ρ is smooth and S is Polish, then ρ is tame.*

Proof 0. There are many σ-algebras around, so let us see what we have to do. We want to show that

Proof outline

$$\boldsymbol{\Sigma}_{\rho}(\mathcal{B}(X)) \otimes \mathcal{B}([0,1]) = \boldsymbol{\Sigma}_{\rho\times\Delta}(\mathcal{B}(X \otimes [0,1])) \tag{4.5}$$

holds. Since X is Polish and ρ is smooth, X/ρ is an analytic space, so we know from Proposition 4.3.16 that $\mathcal{B}(X/\rho \otimes [0,1]) = \mathcal{B}(X/\rho) \otimes \mathcal{B}([0,1])$. In order to establish the inclusion from left to right in Eq. (4.5), we show that each member of the left-hand set is $\rho\times\Delta$-invariant. For the converse direction, we show that each member of the set on the right-hand side can be represented as the inverse image under $\eta_{\rho\times\Delta}$ of a set in the factor space.

1. Let $G \times D \subseteq X \times [0,1]$ be a generator of $\boldsymbol{\Sigma}_{\rho}(\mathcal{B}(X)) \otimes \mathcal{B}([0,1])$ such that $G \in \boldsymbol{\Sigma}_{\rho}(\mathcal{B}(X))$ and $D \in \mathcal{B}([0,1])$. Then $G \times D$ is $\rho \times \Delta$-invariant. But this means

$$\begin{aligned}\boldsymbol{\Sigma}_{\rho}(\mathcal{B}(X)) \otimes \mathcal{B}([0,1]) = \sigma(\{G \times D \mid G \in \boldsymbol{\Sigma}_{\rho}(\mathcal{B}(X)),\\ D \in \mathcal{B}([0,1])\}) \subseteq \boldsymbol{\Sigma}_{\rho\times\Delta}(\mathcal{B}(X \otimes [0,1])).\end{aligned}$$

2. Now let $H \in \boldsymbol{\Sigma}_{\rho\times\Delta}(\mathcal{B}(X) \otimes [0,1])$; then we find $H_0 \in \mathcal{B}((/X \times [0,1])\rho \times \Delta)$ such that $H = \eta_{\rho\times\Delta}^{-1}\big[H_0\big]$. But

$$\mathcal{B}(X \times [0,1]/\rho \times \Delta) = \mathcal{B}(X/\rho \otimes [0,1]) = \mathcal{B}(X/\rho) \otimes \mathcal{B}([0,1]),$$

the first equality following from $(X \times [0,1])/\rho \times \Delta = X/\rho \times [0,1]$ and the second from Proposition 4.3.16. This is so since ρ is smooth; hence

X/ρ is an analytic space, and $[0,1]$ is a Polish space and hence has a countable basis. On the other hand, $\eta_{\rho\times\Delta} = \eta_\rho \times id$; hence we have found that $H = (\eta_\rho \times id)^{-1}\big[H_0\big]$ with $H_0 \in \mathcal{B}(X/\rho) \otimes \mathcal{B}([0,1])$. This establishes the other inclusion. ⊣

This shows that tame equivalence relations constitute a generalization of smooth ones for the case that we do not work in a Polish environment.

Sometimes one starts not with a topological space and its Borel sets but rather with a measurable space: A *standard Borel* space $(X, \mathcal{A})$ is a measurable space such that the σ-algebra $\mathcal{A}$ equals $\mathcal{B}(\tau)$ for some Polish topology τ on X. We will not dwell on this distinction.

4.5 The Souslin Operation

The collection of analytic sets is closed under Souslin's operation $\mathscr{A}$, which we will introduce now. We will also see that complete measure spaces are another important class of measurable spaces which are closed under this operation. Each measurable space can be completed with respect to its finite measures, so that we do not even need a topology for carrying out the constructions ahead.

Let $\mathbb{N}^+$ be the set of all finite and nonempty sequences of natural numbers. Denote for $t = (x_n)_{n\in\mathbb{N}} \in \mathbb{N}^\mathbb{N}$ by $t|k = \langle x_1, \dots, x_k\rangle$ its first k elements. Given a subset $\mathcal{C} \subseteq \mathcal{P}(X)$, put

$$\mathscr{A}(\mathcal{C}) := \{\bigcup_{t\in\mathbb{N}^\mathbb{N}} \bigcap_{k\in\mathbb{N}} A_{t|k} \mid A_v \in \mathcal{C} \text{ for all } v \in \mathbb{N}^+\}.$$

Note that the outer union may be taken of more than countably many sets. A family $(A_v)_{v\in\mathbb{N}^+}$ is called a *Souslin scheme*, which is called *regular* if $A_w \subseteq A_v$ whenever v is an initial piece of w. Because

$$\bigcup_{t\in\mathbb{N}^\mathbb{N}} \bigcap_{k\in\mathbb{N}} A_{t|k} = \bigcup_{t\in\mathbb{N}^\mathbb{N}} \bigcap_{k\in\mathbb{N}} \Big(\bigcap_{1\leq j\leq k} A_{t|j}\Big),$$

we can and will restrict our attention to regular Souslin schemes whenever $\mathcal{C}$ is closed under finite intersections.

We will see now that each analytic set can be represented through a Souslin scheme with a special shape. This has some interesting consequences, among others, that analytic sets are closed under the Souslin operation.

Proposition 4.5.1 *Let X be a Polish space and $(A_v)_{v\in\mathbb{N}^+}$ be a regular Souslin scheme of closed sets such that* $\operatorname{diam}(A_v) \to 0$*, as the length of v goes to infinity. Then*

$$E := \bigcup_{t\in\mathbb{N}^\mathbb{N}} \bigcap_{k\in\mathbb{N}} A_{t|k}$$

is an analytic set in X. Conversely, each analytic set can be represented in this way.

Proof 1. Assume E is given through a Souslin scheme; then we represent $E = f\left[F\right]$ with $F \subseteq \mathbb{N}^\mathbb{N}$ a closed set and $f : F \to X$ continuous.

In fact, put

$$F := \{t \in \mathbb{N}^\mathbb{N} \mid A_{t|k} \neq \emptyset \text{ for all } k\}.$$

Then F is a closed subset of $\mathbb{N}^\mathbb{N}$: Take $s \in \mathbb{N}^\mathbb{N} \setminus F$; then we can find $k' \in \mathbb{N}$ with $A_{s|k'} = \emptyset$, so that $G := \{t \in \mathbb{N}^\mathbb{N} \mid t|k' = s|k'\}$ is open in $\mathbb{N}^\mathbb{N}$, contains s, and is disjoint to F. Now let $t \in F$; then there exists exactly one point $f(t) \in \bigcap_{k\in\mathbb{N}} A_{t|k}$, since X is complete and the diameters of the sets involved tend to zero; see Proposition 3.5.25. Then $E = f\left[F\right]$ follows from this construction, and we show that f is continuous.

Let $t \in F$ and $\epsilon > 0$ be given, take $x := f(t)$, and let B be the ball with center x and radius ϵ. Then we can find an index k such that $A_{t|k'} \subseteq S$ for all $k' \geq k$; hence $U := \{s \in F \mid t|k = s|k\}$ is an open neighborhood of t with $f\left[U\right] \subseteq B$.

2. Let E be an analytic set; then $E = f\left[\mathbb{N}^\mathbb{N}\right]$ with f continuous by Proposition 4.4.2. Define A_v as the closure of the set $f\left[\{t \in \mathbb{N}^\mathbb{N} \mid t|k = v\}\right]$, if the length of v is k. Then clearly

$$E = \bigcup_{t\in\mathbb{N}^\mathbb{N}} \bigcap_{k\in\mathbb{N}} A_{t|k},$$

since f is continuous. It is also clear that $(A_v)_{v\in\mathbb{N}^+}$ is regular with diameter tending to zero. $\dashv$

Before we can enter into the—fairly technical—demonstration that the Souslin operation is idempotent, we need some auxiliary statements.

The first one is readily verified.

Lemma 4.5.2 $b(m,n) := 2^{m-1}(2n-1)$ *defines a bijective map* $\mathbb{N} \times \mathbb{N} \to \mathbb{N}$*. Moreover,* $m \leq b(m,n)$ *and* $n < n'$ *implies* $b(m,n) < b(m,n')$ *for all* $n, n', m \in \mathbb{N}$. $\dashv$

Given $k \in \mathbb{N}$, there exists a unique pair $\langle \ell(k), r(k) \rangle \in \mathbb{N} \times \mathbb{N}$ with $b(\ell(k), r(k)) = k$. We will need the functions $\ell, r : \mathbb{N} \to \mathbb{N}$ later on. The next function is considerably more complicated, since it caters for a more involved set of parameters.

Lemma 4.5.3 *Given* $z = (z_n)_{n\in\mathbb{N}} \in (\mathbb{N}^{\mathbb{N}})^{\mathbb{N}}$ *with* $z_n = (z_{n,m})_{m\in\mathbb{N}}$ *and* $t \in \mathbb{N}^{\mathbb{N}}$*, define the sequence* $B(t,z) \in \mathbb{N}^{\mathbb{N}}$ *through*

$$B(t,z)_k := b(t(k), z_{\ell(k),r(k)})$$

($k \in \mathbb{N}$*). Then* $B : \mathbb{N}^{\mathbb{N}} \times (\mathbb{N}^{\mathbb{N}})^{\mathbb{N}} \to \mathbb{N}^{\mathbb{N}}$ *is a bijection.*

Proof 1. We show first that B is injective. Let $\langle t, z \rangle \neq \langle t', z' \rangle$. If $t \neq t'$, we find k with $t(k) \neq t'(k)$, so that $b(t(k), z_{\ell(k),r(k)}) \neq b(t'(k), z'_{\ell(k),r(k)})$ follows, because b is injective. Now assume that $t = t'$, but $z \neq z'$, so we can find $i, j \in \mathbb{N}$ with $z_{i,j} \neq z'_{i,j}$. Let $k := b(i,j)$, so that $\ell(k) = i$ and $r(k) = j$; hence $\langle t(k), z_{\ell(k),r(k)} \rangle \neq \langle t(k), z'_{\ell(k),r(k)} \rangle$, so that $B(t,z)_k \neq B(t',z')_k$.

2. Now let $s \in \mathbb{N}^{\mathbb{N}}$, and define $t \in \mathbb{N}^{\mathbb{N}}$ and $z \in (\mathbb{N}^{\mathbb{N}})^{\mathbb{N}}$

$$\begin{aligned} t_k &:= \ell(s_k), \\ z_{n,m} &:= r(s_{b(n,m)}). \end{aligned}$$

Then we have for $k \in \mathbb{N}$

$$\begin{aligned} B(t,z)_k &= b(t_k, z_{\ell(k),r(k)}) = b\big(\ell(s_k), r(s_{b(\ell(k),r(k))})\big) \\ &= b\big(\ell(s_k), r(s_k)\big) = s_k. \end{aligned}$$

$\dashv$

We construct maps φ, ψ from the maps b and B now with special properties. They will be made use of in the proof that the Souslin operation is idempotent.

Lemma 4.5.4 *There exist maps $\varphi, \psi : \mathbb{N}^+ \to \mathbb{N}^+$ with this property: Let $w = B(t,z)|b(n,m)$; then $\varphi(w) = t|m$ and $\psi(w) = z_m|n$.*

Proof Fix $v = \langle x_1, \dots, x_k \rangle$, then define for $m := \ell(k)$ and $n := r(k)$

$$\begin{aligned}\varphi(v) &:= \langle \ell(x_1), \dots, \ell(x_m) \rangle,\\ \psi(v) &:= \langle r(x_{b(m,1)}), \dots, r(x_{b(m,n)}) \rangle.\end{aligned}$$

We see from Lemma 4.5.2 that these definitions are possible.

Given $t \in \mathbb{N}^\mathbb{N}$ and $z \in (\mathbb{N}^\mathbb{N})^\mathbb{N}$, we put $k := b(m,n)$ and $v := B(t,z)|k$, then we obtain from the definition of φ resp. ψ

$$\begin{aligned}\varphi(v) &= \langle \ell(v_1), \dots, \ell(v_m) \rangle = t|m,\\ \psi(v) &= \langle r(v_{b(m,0)}), \dots, r(v_{b(m,n)}) \rangle = z_m|n,\end{aligned}$$

as desired. $\dashv$

The construction shows that $\mathscr{A}(\mathcal{C})$ is always closed under countable unions and countable intersections. We are now in a position to prove a much more general observation of the Souslin operation.

Theorem 4.5.5 $\mathscr{A}(\mathscr{A}(\mathcal{C})) = \mathscr{A}(\mathcal{C})$.

Proof It is clear that $\mathcal{C} \subseteq \mathscr{A}(\mathcal{C})$, so we have to establish the other inclusion. Let $\{D_{v,w} \mid w \in \mathbb{N}^+\}$ be a Souslin scheme for each $v \in \mathbb{N}^+$, and put $A_v := \bigcup_{s \in \mathbb{N}^\mathbb{N}} \bigcap_{m \in \mathbb{N}} D_{v,s|m}$. Then we have

$$\begin{aligned}A &:= \bigcup_{t \in \mathbb{N}^\mathbb{N}} \bigcap_{k \in \mathbb{N}} A_{t|k}\\ &= \bigcup_{t \in \mathbb{N}^\mathbb{N}} \bigcap_{k \in \mathbb{N}} \bigcup_{s \in \mathbb{N}^\mathbb{N}} \bigcap_{m \in \mathbb{N}} D_{v,s|m}\\ &= \bigcup_{t \in \mathbb{N}^\mathbb{N}} \bigcup_{(z_n)_{n \in \mathbb{N}} \in (\mathbb{N}^\mathbb{N})^\mathbb{N}} \bigcap_{m \in \mathbb{N}} \bigcap_{k \in \mathbb{N}} D_{t|m, z_m|n}\\ &\overset{(*)}{=} \bigcup_{s \in \mathbb{N}^\mathbb{N}} \bigcap_{k \in \mathbb{N}} C_{s|k}\end{aligned}$$

with

$$C_v := D_{\varphi(v), \psi(v)}$$

for $v \in \mathbb{N}^+$. So we have to establish the equality marked $(*)$.

"$\subseteq$": Given $x \in A$, there exist $t \in \mathbb{N}^{\mathbb{N}}$ and $z \in (\mathbb{N}^{\mathbb{N}})^{\mathbb{N}}$ such that $x \in D_{t|m,z_{m|n}}$. Put $s := B(t,z)$. Let $k \in \mathbb{N}$ be arbitrary; then there exists a pair $\langle m,n\rangle \in \mathbb{N} \times \mathbb{N}$ with $k = b(m,n)$ by Lemma 4.5.2. Thus we have $t|m = \varphi(s|k)$ and $z_m|n = \psi(s|k)$ by Lemma 4.5.4, from which $x \in D_{t|m,z_m|n} = C_{s|k}$ follows.

"$\supseteq$": Let $s \in \mathbb{N}^{\mathbb{N}}$ such that $x \in C_{s|k}$ for all $k \in \mathbb{N}$. We can find by Lemma 4.5.3 some $t \in \mathbb{N}^{\mathbb{N}}$ and $z \in (\mathbb{N}^{\mathbb{N}})^{\mathbb{N}}$ with $B(t,z) = s$. Given k, there exists $m,n \in \mathbb{N}$ with $k = b(m,n)$; hence $C_{s|k} = D_{t|m,z_m|n}$. Thus $x \in A$.

$\dashv$

We obtain as an immediate consequence that analytic sets in a Polish space X are closed under the Souslin operation. This is so because we have seen that the collection of analytic sets is contained in $\mathscr{A}\big(\{F \subseteq X \mid F \text{ is closed}\}\big)$, so an application of Theorem 4.5.5 proves the claim. But we can say even more.

Proposition 4.5.6 *Assume that the complement of each set in $\mathcal{C}$ belongs to $\mathscr{A}(\mathcal{C})$ and $\emptyset \in \mathcal{C}$. Then $\sigma(\mathcal{C}) \subseteq \mathscr{A}(\mathcal{C})$. In particular, analytic sets in a Polish space X are closed under the Souslin operation.*

Proof We apply for establishing the general statement the principle of good sets. Define

$$\mathcal{G} := \{A \in \mathscr{A}(\mathcal{C}) \mid X \setminus A \in \mathscr{A}(\mathcal{C})\}.$$

Then $\mathcal{G}$ is closed under complementation. If $(A_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{G}$, then $\bigcap_{n\in\mathbb{N}} A_n \in \mathcal{G}$, because $\mathscr{A}(\mathcal{C})$ is closed under countable unions. Similarly, $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{G}$. Since $\emptyset \in \mathcal{G}$, we may conclude that $\mathcal{G}$ is a σ-algebra, which contains $\mathcal{C}$ by assumption. Hence $\sigma(\mathcal{C}) \subseteq \sigma(\mathcal{G}) = \mathcal{G} \subseteq \mathscr{A}(\mathcal{C})$. The assertion concerning analytic sets follows from Proposition 4.5.1, because a closed set is a G_δ-set. $\dashv$

With complete measure spaces, we will meet an important class of measurable spaces, which is closed under the Souslin operation. As a preparation for this, we state and prove an interesting criterion for being closed under this operation. This requires the definition of a particular kind of cover.

Definition 4.5.7 *Given a measurable space $(X, \mathcal{A})$ and a subset $A \subseteq X$, we call $A^{\uparrow} \in \mathcal{A}$ an $\mathcal{A}$-*cover *of A iff*

1. $A \subseteq A^{\uparrow}$.
2. *For every $B \in \mathcal{A}$ with $A \subseteq B$, $\mathcal{P}\left(A^{\uparrow} \setminus B\right) \subseteq \mathcal{A}$.*

Thus $A^{\uparrow} \in \mathcal{A}$ covers A in the sense that $A \subseteq A^{\uparrow}$, and if we have another set $B \in \mathcal{A}$ which covers A as well, then *all* the sets which make out the difference between $A^{\uparrow}$ and B are measurable. This last condition indicates that the "breathing space" between A and its cover $A^{\uparrow}$ is an element of $\mathcal{A}$. In a technical sense, this will give us considerable room for maneuvering, when applying this concept.

In addition it follows that if $A \subseteq A' \subseteq A^{\uparrow}$ and $A' \in \mathcal{A}$, then A' is also an $\mathcal{A}$-cover. This concept sounds fairly artificial and somewhat far-fetched, but we will see that it arises in a natural way when completing measure spaces. The surprising observation is that a space is closed under the Souslin operation whenever each subset has an $\mathcal{A}$-cover.

Proposition 4.5.8 *Let $(X, \mathcal{A})$ be a measurable space such that each subset of X has an $\mathcal{A}$-cover. Then $(X, \mathcal{A})$ is closed under the Souslin operation.*

Outline of the proof

Proof 0. The proof is a bit tricky. We first construct from a regular Souslin scheme a sequence of sets which are indexed by the words over $\mathbb{N}$ such that each set B_w can be represented as the union of $\left(B_{wn}\right)_{n \in \mathbb{N}}$. For each set B_w, there exists an $\mathcal{A}$-cover, which due to the properties of B_w can be more easily manipulated than a cover for the sets in the Souslin scheme proper; in particular we can look at the difference between the $\mathcal{A}$-cover for B_n and those for B_{wn}, so that we can move backward, from longer words to shorter ones. It will then turn out that we can represent the set defined by the given Souslin scheme through the $\mathcal{A}$-cover given for the empty word ϵ.

1. Let

$$A := \bigcup_{t \in \mathbb{N}^{\mathbb{N}}} \bigcap_{k \in \mathbb{N}} A_{a|k}$$

with $(A_v)_{v \in \mathbb{N}^+}$ a regular Souslin scheme in $\mathcal{A}$. Define

$$B_w := \bigcup \{\bigcap_{n \in \mathbb{N}} A_{t|n} \mid t \in \mathbb{N}^{\mathbb{N}}, w \text{ is a prefix of } t\}$$

for $w \in \mathbb{N}^* = \mathbb{N}^+ \cup \{\epsilon\}$. Then $B_\epsilon = A$, $B_w = \bigcup_{n\in\mathbb{N}} B_{wn}$, and $B_w \subseteq A_w$ if $w \neq \epsilon$.

By assumption, there exists a $\mathcal{A}$-cover $B_w^\uparrow$ for B_w. We may and do assume that $B_w^\uparrow \subseteq A_w$ and that $(B_w^\uparrow)_{w\in\mathbb{N}^*}$ is regular; otherwise we force this condition by considering the $\mathcal{A}$-cover

$$\left(\bigcap\{B_v^\uparrow \cap A_v \mid v \text{ prefix of } w\}\right)_{w\in\mathbb{N}^*}$$

instead. Now put

$$D_w := B_w^\uparrow \setminus \bigcup_{n\in\mathbb{N}} B_{wn}^\uparrow$$

for $w \in \mathbb{N}^*$. We obtain from this construction

$$B_w \subseteq B_w^\uparrow = \bigcup_{n\in\mathbb{N}} B_{wn}^\uparrow \in \mathcal{A};$$

hence we see that every subset of D_w is in $\mathcal{A}$, since $B_w^\uparrow$ is an $\mathcal{A}$-cover. Thus every subset of

$$D := \bigcup_{w\in\mathbb{N}^*} D_w$$

is in $\mathcal{A}$.

2. We claim that $B_\epsilon^\uparrow \setminus D \subseteq A$. In fact, let $x \in B_\epsilon^\uparrow \setminus D$; then $x \notin D_\epsilon$, so we can find $k_1 \in \mathbb{N}$ with $x \in B_{k_1}^\uparrow$, but $x \notin D_{n_1}$. Since $x \notin D_{k_1}$, we find k_2 with $x \in B_{k_1,k_2}^\uparrow$ such that $x \notin D_{k_1,k_2}$. So we inductively define a sequence $t := (k_n)_{n\in\mathbb{N}}$ so that $x \in B_{t|k}^\uparrow$ for all $k \in \mathbb{N}$. Because $B_{t|k}^\uparrow \subseteq A_{t|k}$, we conclude that $x \in A$.

3. Hence we obtain $B_\epsilon^\uparrow \setminus A \subseteq D$, and since every subset of D is in $\mathcal{A}$, we conclude that $B_\epsilon^\uparrow \setminus A \in \mathcal{A}$, which means that $A = B_\epsilon^\uparrow \setminus (B_\epsilon^\uparrow \setminus A) \in \mathcal{A}$. $\dashv$

The concept of being closed under the Souslin operation will now be applied to universally measurable sets, in particular to analytic sets in a Polish space.

4.6 Universally Measurable Sets

After these technical preparations, we are poised to enter the interesting world of universally measurable sets with the closure operations that are associated with them. We define complete measure spaces and show that an arbitrary (σ-)finite measure space can be completed, uniquely extending the measure as we go. This leads also to completions with respect to families of finite measures, and we show that the resulting measurable spaces are closed under the Souslin operation.

Two applications are discussed. The first one demonstrates that a measure defined on a countably generated sub-σ-algebra of the Borel sets of an analytic space can be extended to the Borel sets, albeit not necessarily in a unique way. This result due to Lubin rests on the important von Neumann Selection Theorem, giving a universally right inverse to a measurable map from an analytic to a separable space. Another application of von Neumann's result is the observation that under suitable topological assumptions for a surjective map f, the lifted map $\mathbb{M}(f)$ is surjective as well. The second application shows that a transition kernel can be extended to the universal closures of the measurable spaces involved, provided the target space is separable. This, however, does not require a selection.

Complete measure

A σ-finite measure space $(X, \mathcal{A}, \mu)$ is called *complete* iff $\mu(A) = 0$ with $A \in \mathcal{A}$ and $B \subseteq A$ implies $B \in \mathcal{A}$. Thus if we have two sets $A, A' \in \mathcal{A}$ with $A \subseteq A'$ and $\mu(A) = \mu(A')$, then we know that each set which can be sandwiched between the two will be measurable as well. This is partly anticipated in the discussion in Sect. 1.6.4, where a similar extension problem is considered, but starting from an outer measure. We will discuss the completion of a measure space and investigate some properties. We first note that it is sufficient to discuss finite measure spaces; in fact, assume that we have a collection of mutually disjoint sets $(G_n)_{n\in\mathbb{N}}$ with $G_n \in \mathcal{A}$ such that $0 < \mu(G_n) < \infty$ and $\bigcup_{n\in\mathbb{N}} G_n = X$, and consider the measure

$$\mu'(B) := \sum_{n\in\mathbb{N}} \frac{\mu(B \cap G_n)}{2^n \cdot \mu(G_n)};$$

then μ is complete iff μ' is complete, and μ' is a probability measure.

We fix for the time being a finite measure μ on a measurable space $(X, \mathcal{A})$.

The outer measure μ^* is defined through

$$\begin{aligned}\mu^*(C) &:= \inf\{\sum_{n\in\mathbb{N}} \mu(A_n) \mid C \subseteq \bigcup_{n\in\mathbb{N}} A_n, A_n \in \mathcal{A} \text{ for all } n \in \mathbb{N}\}\\ &= \inf\{\mu(A) \mid C \subseteq A, A \in \mathcal{A}\}\end{aligned}$$

for any subset C of X; see page 79. μ^*

Definition 4.6.1 *Call $N \subseteq X$ a μ-*null set *iff $\mu^*(N) = 0$. Define $\mathcal{N}_\mu$ as the set of all μ-null sets.*

Because μ^* is countably subadditive by Lemma 1.6.21, we obtain

Lemma 4.6.2 *$\mathcal{N}_\mu$ is a σ-ideal.* $\dashv$

Now assume that we have sets $A, A' \in \mathcal{A}$ and $N, N' \in \mathcal{N}_\mu$ with $A\Delta N = A'\Delta N'$. Then we may infer $\mu(A) = \mu(A')$, because $A\Delta A' = A\Delta\big(A\Delta(N\Delta N')\big) = N\Delta N' \subseteq N \cup N' \in \mathcal{N}_\mu$, and $|\mu(A) - \mu(A')| \leq \mu(A\Delta A')$. Thus we may construct an extension of μ to the σ-algebra generated by $\mathcal{A}$ and $\mathcal{N}_\mu$ in the obvious way. Let us have a look at some properties of this construction.

Proposition 4.6.3 *Define $\mathcal{A}_\mu := \sigma(\mathcal{A} \cup \mathcal{N}_\mu)$ and $\overline{\mu}(A\Delta N) := \mu(A)$ for $A \in \mathcal{A}, N \in \mathcal{N}_\mu$. Then*

1. *$\mathcal{A}_\mu = \{A\Delta N \mid A \in \mathcal{A}, N \in \mathcal{N}_\mu\}$, and $A \in \mathcal{A}_\mu$ iff there exist sets $A', A'' \in \mathcal{A}$ with $A' \subseteq A \subseteq A''$ and $\mu^*(A'' \setminus A') = 0$.*

2. *$\overline{\mu}$ is a finite measure and the unique extension of μ to $\mathcal{A}_\mu$.*

3. *The measure space $(X, \mathcal{A}_\mu, \overline{\mu})$ is complete. It is called the μ-*completion *of $(X, \mathcal{A}, \mu)$.*

Proof 0. The proof is a fairly straightforward check of the properties. It rests essentially on the observation that $\mathcal{N}_\mu$ is a σ-ideal, so that the construction of the σ-algebra under consideration has already been studied.

1. Since $\mathcal{N}_\mu$ is a σ-ideal, we infer from Lemma 4.1.3 that $A \in \mathcal{A}_\mu$ iff there exist $B \in \mathcal{A}$ and $N \in \mathcal{N}_\mu$ with $A = B \Delta N$. Now consider

$$\mathcal{C} := \{A \in \mathcal{A}_\mu \mid \exists A', A'' \in \mathcal{A} : A' \subseteq A \subseteq A'', \mu^*(A'' \setminus A') = 0\}.$$

Then $\mathcal{C}$ is a σ-algebra which contains $\mathcal{A} \cup \mathcal{N}_\mu$; thus $\mathcal{C} = \mathcal{A}_\mu$.

From the observation made just before stating the proposition, it becomes clear that $\overline{\mu}$ is well defined on $\mathcal{A}_\mu$. Since μ^* coincides with $\overline{\mu}$ on $\mathcal{A}_\mu$ and the outer measure is countably subadditive by Lemma 1.6.23, we have to show that $\overline{\mu}$ is additive on $\mathcal{A}_\mu$. This follows immediately from the first part. If ν is another extension to μ on $\mathcal{A}_\mu$, $\mathcal{N}_\nu = \mathcal{N}_\mu$ follows, so that $\overline{\mu}(A \Delta N) = \mu(A) = \nu(A) = \nu(A \Delta N)$ whenever $A \Delta N \in \mathcal{A}_\mu$.

2. Completeness of $(X, \mathcal{A}_\mu, \overline{\mu})$ follows now immediately from the construction. $\dashv$

Surprisingly, we have received more than we have shopped for, since complete measure spaces are closed under the Souslin operation. This is remarkable because the Souslin operation evidently bears no hint at all at measures which are defined on the base space. In addition, measures are defined through countable operations, while the Souslin operation makes use of the uncountable space $\mathbb{N}^\mathbb{N}$.

Proposition 4.6.4 *A complete measure space is closed under the Souslin operation.*

The proof simply combines the pieces we have constructed already into a sensible picture.

Proof Let $(X, \mathcal{A}, \mu)$ be complete; then it is enough to show that each $B \subseteq X$ has an $\mathcal{A}$-cover (Definition 4.5.7); then the assertion will follow from Proposition 4.5.8. In fact, given B, construct $B^* \in \mathcal{A}$ such that $\mu(B^*) = \mu^*(B)$; see Lemma 1.6.35. Whenever $C \in \mathcal{A}$ with $B \subseteq C$, we evidently have every subset of $B^* \setminus C$ in $\mathcal{A}$ by completeness. $\dashv$

These constructions work also for σ-finite measure spaces, as indicated above. Now let M be a nonempty set of σ-finite measures on the measurable space $(X, \mathcal{A})$, then define the *M-completion* $\overline{\mathcal{A}}^M$ and the *universal completion* $\overline{\mathcal{A}}$ of the σ-algebra $\mathcal{A}$ through

$\overline{\mathcal{A}}^M, \overline{\mathcal{A}}$

$$\overline{\mathcal{A}}^M := \bigcap_{\mu \in M} \mathcal{A}_\mu,$$
$$\overline{\mathcal{A}} := \bigcap \{\mathcal{A}_\mu \mid \mu \text{ is a } \sigma\text{-finite measure on } \mathcal{A}\}.$$

As an immediate consequence, this yields that the analytic sets in a Polish space are contained in the universal completion of the Borel sets, specifically

Corollary 4.6.5 *Let X be a Polish space and μ be a finite measure on $\mathcal{B}(X)$. Then all analytic sets are contained in $\overline{\mathcal{B}(X)}$.*

Proof Proposition 4.5.1 shows that each analytic set can be represented through a Souslin scheme based on closed sets, and Proposition 4.6.4 shows that $\overline{\mathcal{B}(X)}$ is closed under the Souslin operation. $\dashv$

Just for the record

Corollary 4.6.6 *The universal closure of a measurable space is closed under the Souslin operation.* $\dashv$

Measurability of maps is preserved when passing to the universal closure.

Lemma 4.6.7 *Let $f : X \to Y$ be $\mathcal{A}$-$\mathcal{B}$ measurable; then f is $\overline{\mathcal{A}}$-$\overline{\mathcal{B}}$ measurable.*

Proof Let $D \in \overline{\mathcal{B}}$ be a universally measurable subset of Y; then we have to show that $E := f^{-1}\left[D\right]$ is universally measurable in X. So we have to show that for every finite measure μ on $\mathcal{A}$, there exists $E', E'' \in \mathcal{A}$ with $E' \subseteq E \subseteq E''$ and $\mu(E' \setminus E'') = 0$.

Define ν as the image of μ under f, so that $\nu(B) = \mu(f^{-1}\left[B\right])$ for each $B \in \mathcal{B}$; then we know that there exists $D', D'' \in \mathcal{B}$ with $D' \subseteq D \subseteq D''$ such that $\nu(D'' \setminus D') = 0$; hence we have for the measurable sets $E' := f^{-1}\left[D'\right]$, $E'' := f^{-1}\left[D''\right]$

$$\mu(E'' \setminus E') = \mu(f^{-1}\left[D'' \setminus D'\right]) = \nu(D'' \setminus D') = 0.$$

Thus $f^{-1}\left[D\right] \in \overline{\mathcal{A}}$. $\dashv$

We will give now two applications of this construction. The first will show that a finite measure on a countably generated sub-σ-algebra of the Borel sets of an analytic space has always an extension to the Borel sets, and the second will construct an extension of a stochastic relation $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ to a stochastic relation $\overline{K} : (X, \overline{\mathcal{A}}) \rightsquigarrow (Y, \overline{\mathcal{B}})$, provided the target space $(Y, \mathcal{B})$ is separable. This first application is derived from von Neumann's Selection Theorem, which is established here as well. It is shown also that a measurable surjection can be lifted to a measurable map between finite measure spaces, provided the target space is a separable metric space.

4.6.1 Lubin's Extension Through von Neumann's Selectors

Let X be an analytic space and $\mathcal{B}$ be a countably generated sub-σ-algebra of $\mathcal{B}(X)$. We will show that each finite measure defined on $\mathcal{B}$ has at least one extension to a measure on $\mathcal{B}(X)$. This is established through a surprising selection argument, as we will see.

As a preparation, we require a universally measurable right inverse of a measurable surjective map $f : X \to Y$. We know from the Axiom of Choice that we can find for each $y \in Y$ some $x \in X$ with $f(x) = y$, because $\{f^{-1}\big[\{y\}\big] \mid y \in Y\}$ is a partition of X into nonempty sets; see Proposition 1.0.1. Set $g(y) := x$. Selecting an inverse image in this way will not guarantee, however, that g has any favorable properties, even if, say, both X and Y are compact metric and f is continuous. Hence we will have to proceed in a more systematic fashion.

We will use the observation that each analytic set in a Polish space can be represented as the continuous image of $\mathbb{N}^{\mathbb{N}}$, as discussed in Proposition 4.4.2. Again, the strategy of using a particular space as a reference point pays off. We move the problem to $\mathbb{N}^{\infty}$, where we can easily define a total order, which is then used for solving the problem. Then we port the solution back to the space from which it originated. We will first formulate a sequence of auxiliary statements that deal with finding for a given surjective map $f : X \to Y$ a map $g : Y \to X$ such that $f \circ g = id_Y$. This map g should have some sufficiently pleasant properties.

In order to make the first step, it turns out to be helpful focusing the attention to analytic sets being the continuous images of $\mathbb{N}^{\mathbb{N}}$. This looks a bit far-fetched, because we want to deal with universally measurable sets, but remember that analytic sets are universally measurable.

We order $\mathbb{N}^{\mathbb{N}}$ lexicographically by saying that $(t_n)_{n\in\mathbb{N}} \preceq (t'_n)_{n\in\mathbb{N}}$ iff there exists $k \in \mathbb{N}$ such that $t_k \leq t'_k$ and $t_j = t'_j$ for all j with $1 \leq j < k$. Then $\preceq$ defines a total order on $\mathbb{N}^{\mathbb{N}}$. We will capitalize on this order, to be more precise, on the interplay between the order and the topology. Let us briefly look into the order structure of $\mathbb{N}^{\mathbb{N}}$. $(t_n) \preceq (t'_n)$

Lemma 4.6.8 *Each nonempty closed set $F \subseteq \mathbb{N}^{\infty}$ has a minimal element in the lexicographic order.*

Proof Let n_1 be the minimal first component of all elements of F, n_2 be the minimal second component of those elements of F that start with n_1, etc. This defines an element $t := \langle n_1, n_2, \ldots\rangle$. We claim that $t \in F$. Let U be an open neighborhood of t; then there exists $k \in \mathbb{N}$ such that $t \in \Theta_{n_1\ldots n_k} \subseteq U$ (Θ_α is defined on page 492). By construction, $\Theta_{n_1\ldots n_k} \cap F \neq \emptyset$; thus each open neighborhood of t contains an element of F. Hence t is an accumulation point of F, and since F is closed, it contains all its accumulation points. Consequently, $t \in F$. $\dashv$

We know for $f : \mathbb{N}^{\mathbb{N}} \to X$ continuous that the inverse images $f^{-1}\left[\{y\}\right]$ with $y \in f\left[\mathbb{N}^{\mathbb{N}}\right]$ are closed. Thus me may pick for each $y \in f\left[\mathbb{N}^{\mathbb{N}}\right]$ this smallest element. This turns out to be a suitable choice, as the following statement shows:

Lemma 4.6.9 *Let X be Polish, $Y \subseteq X$ analytic with $Y = f\left[\mathbb{N}^{\mathbb{N}}\right]$ for some continuous $f : \mathbb{N}^{\mathbb{N}} \to X$. Then there exists $g : Y \to \mathbb{N}^{\mathbb{N}}$ such that*

1. *$f \circ g = id_Y$,*
2. *g is $\overline{\mathcal{B}(Y)}$-$\overline{\mathcal{B}(\mathbb{N}^{\mathbb{N}})}$-measurable.*

Proof 1. Since f is continuous, the inverse image $f^{-1}\left[\{y\}\right]$ for each $y \in Y$ is a closed and nonempty set in $\mathbb{N}^{\infty}$. Thus this set contains a minimal element $g(y)$ in the lexicographic order $\preceq$ by Lemma 4.6.8. It is clear that $f(g(y)) = y$ holds for all $y \in Y$.

2. Denote by $A(t') := \{t \in \mathbb{N}^{\infty} \mid t \prec t'\}$; then $A(t')$ is open: Let $(\ell_n)_{n\in\mathbb{N}} = t \prec t'$ and k be the first component in which t differs from t'; then $\Theta_{\ell_1\ldots\ell_{k-1}}$ is an open neighborhood of t that is entirely contained in $A(t')$. It is easy to see that $\{A(t') \mid t' \in \mathbb{N}^{\infty}\}$ is a generator for the Borel sets of $\mathbb{N}^{\infty}$.

3. We claim that $g^{-1}\left[A(t')\right] = f\left[A(t')\right]$ holds. In fact, let $y \in g^{-1}\left[A(t')\right]$, so that $g(y) \in A(t')$; then $y = f(g(y)) \in f\left[A(t')\right]$.

If, on the other hand, $y = f(t)$ with $t \prec t'$, then by construction $t \in f^{-1}\big[\{y\}\big]$; thus $g(y) \preceq t \prec t'$, settling the other inclusion.

This equality implies that $g^{-1}\big[A(t')\big]$ is an analytic set, because it is the image of an open set under a continuous map. Consequently, $g^{-1}\big[A(t')\big]$ is universally measurable for each $A(t')$ by Corollary 4.6.5. Thus g is a universally measurable map. ⊣

This statement is the work horse for establishing that a right inverse exists for surjective Borel maps between an analytic space and a separable measurable space. All we need to do now is to massage things into a shape that will render this result applicable in the desired context. The following theorem is attributed to von Neumann.

Von Neumann's Selection Theorem

Theorem 4.6.10 *Let X be an analytic space, $(Y, \mathcal{B})$ a separable measurable space, and $f : X \to Y$ a surjective measurable map. Then there exists $g : Y \to X$ with these properties:*

1. $f \circ g = id_Y$,
2. *g is $\overline{\mathcal{B}}$-$\overline{\mathcal{B}(X)}$-measurable.*

Proof 0. Lemma 4.6.9 gives the technical link which permits to use $\mathbb{N}^\infty$ as an intermediary for which we have already a partial solution.

1. We may and do assume by Lemma 4.4.17 that Y is an analytic subset of a Polish space Q and that X is an analytic subset of a Polish space P. $x \mapsto \langle x, f(x)\rangle$ is a bijective Borel map from X to the graph of f, so graph(f) is an analytic set by Proposition 4.4.6. Thus we can find a continuous map $F : \mathbb{N}^\mathbb{N} \to P \times Q$ with $F\big[\mathbb{N}^\mathbb{N}\big] = \text{graph}(f)$. Consequently, $\pi_Q \circ F$ is a continuous map from $\mathbb{N}^\mathbb{N}$ to Q with

$$(\pi_Q \circ F)\big[\mathbb{N}^\mathbb{N}\big] = \pi_Q\big[\text{graph}(f)\big] = Y.$$

Now let $G : Y \to \mathbb{N}^\mathbb{N}$ be chosen according to Lemma 4.6.9 for $\pi_Q \circ F$. Then $g := \pi_P \circ F \circ G : Y \to X$ is the map we are looking for:

- g is universally measurable, because G is, and because $\pi_P \circ F$ are continuous, they are universally measurable as well,
- $f \circ g = f \circ (\pi_P \circ F \circ G) = (f \circ \pi_P) \circ F \circ G = \pi_Q \circ F \circ G = id_Y$, so g is right inverse to f.

⊣

Due to its generality, the von Neumann Selection Theorem has many applications in diverse areas, many of them surprising. The art in applying it is to reformulate the problem in a suitable manner so that the requirements of this selection theorem are satisfied. We pick two applications, viz., showing that the image $\mathbb{M}(f)$ of a surjective Borel map f yields a surjective Borel map again and Lubin's measure extension.

Proposition 4.6.11 *Let X be an analytic space and Y a second countable metric space. If $f : X \to Y$ is a surjective Borel map, so is $\mathbb{M}(f) : \mathbb{M}(X) \to \mathbb{M}(Y)$.*

Proof 1. From Theorem 4.6.10, we find a map $g : Y \to X$ such that $f \circ g = id_Y$ and g is $\overline{\mathcal{B}(Y)} - \overline{\mathcal{B}(X)}$-measurable.

2. Let $\nu \in \mathbb{M}(Y)$, and define $\mu := \mathbb{M}(g)(\nu)$; then $\mu \in \mathbb{M}(X, \overline{\mathcal{B}(X)})$ by construction. Restrict μ to the Borel sets on X, obtaining $\mu_0 \in \mathbb{M}(X, \mathcal{B}(X))$. Since we have for each set $B \subseteq Y$ the equality $g^{-1}\left[f^{-1}[B]\right] = B$, we see that for each $B \in \mathcal{B}(Y)$

$$\begin{aligned}\mathbb{M}(f)(\mu_0)(B) &= \mu_0(f^{-1}\left[B\right]) = \mu(f^{-1}\left[B\right]) \\ &= \nu(g^{-1}\left[f^{-1}\left[B\right]\right]) = \nu(B)\end{aligned}$$

holds. ⊣

This has as a consequence that $\mathbb{M}$ is an endofunctor on the category of Polish or analytic spaces with surjective Borel maps as morphisms; it displays a pretty interaction of reasoning in measurable spaces and arguing in categories.

The following extension theorem due to Lubin shows that one can extend a finite measure from a countably generated sub-σ-algebra to the Borel sets of an analytic space. In contrast to classical extension theorems like Theorem 1.6.29, it does not permit to conclude that the extension is uniquely determined.

Theorem 4.6.12 *Let X be an analytic space and μ be a finite measure on a countably generated sub-σ-algebra $\mathcal{A} \subseteq \mathcal{B}(X)$. Then there exists an extension of μ to a finite measure ν on $\mathcal{B}(X)$.*

Proof Let $(A_n)_{n\in\mathbb{N}}$ be the generator of $\mathcal{A}$, and define the map $f : X \to \{0, 1\}^{\mathbb{N}}$ through $x \mapsto \left(\chi_{A_n}(x)\right)_{n\in\mathbb{N}}$. Then $M := f\left[X\right]$ is an analytic space, and f is $\mathcal{B}(X)$-$\mathcal{B}(M)$-measurable by Propositions 4.3.10 and 4.4.6. Moreover,

$$\mathcal{A} = \{f^{-1}\left[C\right] \mid C \in \mathcal{B}(M)\}. \tag{4.6}$$

By von Neumann's Selection Theorem 4.6.10, there exists $g : M \to X$ with $f \circ g = id_M$ which is $\overline{\mathcal{B}(M)}$-$\overline{\mathcal{B}(X)}$-measurable. Define

$$\nu(B) := \overline{\mu}\big((g \circ f)^{-1}\left[B\right]\big)$$

for $B \in \mathcal{B}(X)$ with $\overline{\mu}$ as the completion of μ on $\overline{\mathcal{A}}$. Since we have $g^{-1}\left[B\right] \in \overline{\mathcal{B}(M)}$ for $B \in \mathcal{B}(X)$, we may conclude from (4.6) that $f^{-1}\left[g^{-1}\left[B\right]\right] \in \overline{\mathcal{A}}$. ν is an extension to μ. In fact, given $A \in \mathcal{A}$, we know that $A = f^{-1}\left[C\right]$ for some $C \in \mathcal{B}(M)$, so that we obtain

$$\begin{aligned} \nu(A) &= \overline{\mu}\big((g \circ f)^{-1}\left[f^{-1}\left[C\right]\right]\big) && = \overline{\mu}\big(f^{-1} \circ g^{-1} \circ f^{-1}[C]\big) \\ &\overset{(*)}{=} \overline{\mu}\big(f^{-1}\left[C\right]\big) && = \overline{\mu}(A) \\ &= \mu(A). \end{aligned}$$

$(*)$ holds, since $f \circ g = id_M$. This completes the proof. $\dashv$

Lubin's Theorem can be rephrased in a slightly different way as follows. The identity $id_{\mathcal{A}} : (X, \mathcal{B}(X)) \to (X, \mathcal{A})$ is measurable, because $\mathcal{A}$ is a sub-σ-algebra of $\mathcal{B}(X)$. Hence it induces a measurable map $\mathbb{S}(id_{\mathcal{A}}) : \mathbb{S}(X, \mathcal{B}(X)) \to \mathbb{S}(X, \mathcal{A})$. Lubin's Theorem then implies that $\mathbb{S}(id_{\mathcal{A}})$ is surjective. This is so since $\mathbb{S}(id_{\mathcal{A}})(\mu)$ is just the restriction of μ to the sub-σ-algebra $\mathcal{A}$ for a given $\mu \in \mathbb{S}(X, \mathcal{B}(X))$.

4.6.2 Completing a Transition Kernel

In some probabilistic models for modal logics, it becomes necessary to assume that the state space is closed under Souslin's operation (see, for example, [Dob12b] or Sect. 4.9.4); on the other hand one may not always assume that a complete measure space is given. Hence one needs to complete it, but it is then also mandatory to extend the transition law to the completion as well. This means that an extension of the transition law to the completion becomes necessary. This problem will be studied now.

The completion of a measure space is described in terms of null sets and using inner and outer approximations; see Proposition 4.6.3. We will use the latter here, fixing measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$. Denote by $\mathcal{S}_X$ the smallest σ-algebra on X which contains $\mathcal{A}$ and which is closed under the Souslin operation; hence $\mathcal{S}_X \subseteq \overline{\mathcal{A}}$ by Corollary 4.6.6. $\mathcal{S}_X$

Fix $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ as a transition kernel, and assume first that $\mathcal{B} = \mathcal{B}(Y)$ is the σ-algebra of Borel sets for a separable metric space. Hence the topology τ of Y has a countable base τ_0, which in turn implies that $G = \bigcup\{H \in \tau_0 \mid H \subseteq G\}$ for each open set $G \in \tau$.

For each $x \in X$, we have a finite measure $K(x)$ through the transition kernel K. We associate to $K(x)$ an outer measure $\big(K(x)\big)^*$ on the power set of X. We want to show that the map

$$x \mapsto \big(K(x)\big)^*(A)$$

is $\mathcal{S}_X$-measurable for each $A \subseteq Y$; define for convenience

$$K^*(x) := \big(K(x)\big)^*.$$

Establishing measurability is broken into a sequence of steps.

We need the following regularity argument (but compare Exercise 4.12 for the nonmetric case):

Lemma 4.6.13 *Let μ be a finite measure on $(Y, \mathcal{B}(Y))$, $B \in \mathcal{B}(Y)$. Then we can find for each $\epsilon > 0$ an open set $G \subseteq Y$ with $B \subseteq G$ and a closed set $F \supseteq B$ such that $\mu(G \setminus F) < \epsilon$.*

Proof 0. Let

$$\mathcal{G} := \{B \in \mathcal{B}(Y) \mid \text{ the assertion is true for } B\}.$$

Outline of the proof

We will use a variant of the principle of good sets by showing that $\mathcal{G}$ has these properties:

- $\mathcal{G}$ is closed under complementation.
- The open sets (and, by implication, the closed sets) are contained in $\mathcal{G}$.
- $\mathcal{G}$ is closed under taking disjoint countable unions.

This will permit applying the π-λ-Theorem, because the open sets are a $\cap$-closed generator of the Borel sets.

1. That $\mathcal{G}$ is closed under complementation trivial. $\mathcal{G}$ contains the open as well as the closed sets. If $F \subseteq Y$ is closed, we can represent $F = \bigcap_{n\in\mathbb{N}} G_n$ with $(G_n)_{n\in\mathbb{N}}$ as a decreasing sequence of open sets; hence $\mu(F) = \inf_{n\in\mathbb{N}} \mu(F_n) = \lim_{n\to\infty} \mu(F_n)$, so that $\mathcal{G}$ also contains the closed sets; one argues similarly for the open sets as increasing unions of open sets.

2. Now let $(B_n)_{n\in\mathbb{N}}$ be a sequence of mutually disjoint sets in $\mathcal{G}$, and select G_n open for B_n and $\epsilon/2^{-(n+1)}$; then $G := \bigcup_{n\in\mathbb{N}} G_n$ is open with $B := \bigcup_{n\in\mathbb{N}} B_n \subseteq G$ and $\mu(G \setminus B) \leq \epsilon$. Similarly, select the sequence $(F_n)_{n\in\mathbb{N}}$ with $F_n \subseteq B_n$ and $\mu(B_n \setminus F_n) < \epsilon/2^{-(n+1)}$ for all $n \in \mathbb{N}$, put $F := \bigcup_{n\in\mathbb{N}} F_n$, and select $m \in \mathbb{N}$ with $\mu(F \setminus \bigcup_{n=1}^{m} F_n) < \epsilon/2$; then $F' := \bigcup_{n=1}^{m} F_n$ is closed, $F' \subseteq B$, and $\mu(B \setminus F') < \epsilon$.

3. Hence $\mathcal{G}$ is closed under complementation as well as countable disjoint unions. This implies $\mathcal{G} = \mathcal{B}(Y)$ by the π-λ Theorem 1.6.30, since $\mathcal{G}$ contains the open sets. $\dashv$

Fix $A \subset Y$ for the moment. We claim that

$$K^*(x)(A) = \inf\{K(x)(G) \mid A \subseteq G \text{ open}\}$$

holds for each $x \in X$. In fact, given $\epsilon > 0$, there exists $A \subseteq A_0 \in \mathcal{B}(Y)$ with $K(x)(A_0) - K^*(x)(A) < \epsilon/2$. Applying Lemma 4.6.13 to $K(x)$, we find an open set $G \supseteq A_0$ with $K(x)(G) - K(x)(A_0) < \epsilon/2$; thus $K(x)(G) - K^*(x)(A) < \epsilon$.

Since τ_0 is a countable base for the open sets, which we may assume to be closed under finite unions (because otherwise $\{G_1 \cup \ldots \cup G_k \mid k \in \mathbb{N}, G_1, \ldots, G_k \in \tau_0\}$ is a countable base which has this property), we obtain

$$K^*(x)(A) = \inf\{\sup_{n\in\mathbb{N}} K(x)(G_n) \mid A \subseteq \bigcup_{n\in\mathbb{N}} G_n, (G_n)_{n\in\mathbb{N}} \subseteq \tau_0 \text{ increases}\}. \tag{4.7}$$

Let

$$\mathcal{G}_A := \{(G_n)_{n\in\mathbb{N}} \subseteq \tau_0 \mid (G_n)_{n\in\mathbb{N}} \text{ increases and } A \subseteq \bigcup_{n\in\mathbb{N}} G_n\}$$

be the set of all increasing sequences from base τ_0 which cover A. Partition $\mathcal{G}_A$ into the sets

$$\begin{aligned}\mathcal{N}_A &:= \{\mathfrak{g} \in \mathcal{G}_A \mid \mathfrak{g} \text{ contains only a finite number of sets}\},\\ \mathcal{M}_A &:= \mathcal{G}_A \setminus \mathcal{N}_A.\end{aligned}$$

Because τ_0 is countable, $\mathcal{N}_A$ is.

We want so show that K^*, suitably restricted, is the extension we are looking for. In order to establish this, we build a tree with basic open sets as nodes. The offsprings of node $G \in \tau_0$ are those open sets $G' \in \tau_0$ which contain G. Thus each node has at most countably many offsprings. This tree will be used to construct a Souslin scheme.

Lemma 4.6.14 *There exists an injective map $\Phi : \mathcal{M}_A \to \mathbb{N}^{\mathbb{N}}$ such that $\mathfrak{g} \mid k = \mathfrak{g}' \mid k$ implies $\Phi(\mathfrak{g}) \mid k = \Phi(\mathfrak{g}') \mid k$ for all $k \in \mathbb{N}$.*

Proof 1. Build an infinite tree in this way: The root is the empty set, a node G at level k has all elements G' from τ_0 with $G \subseteq G'$ as offsprings. Remove from the tree all paths $H_1, H_2, \dots$ such that $A \not\subseteq \bigcup_{n\in\mathbb{N}} H_n$. Call the resulting tree $\mathcal{T}$.

2. Put $G_0 := \emptyset$, and let $\mathcal{T}_{1,G_0}$ be the set of nodes of $\mathcal{T}$ on level 1 (hence just the offsprings of the root G_0); then there exists an injective map $\Phi_{1,G_0} : \mathcal{T}_{1,G_0} \to \mathbb{N}$. If $G_1, \dots, G_k$ is a finite path to inner node G_k in $\mathcal{T}$, denote by $\mathcal{T}_{k+1,G_1,\dots,G_k}$ the set of all offsprings of G_k, and let

$$\Phi_{k+1,G_1,\dots,G_k} : \mathcal{T}_{k+1,G_1,\dots,G_k} \to \mathbb{N}$$

be an injective map. Define

$$\Phi : \begin{cases} \mathcal{M}_A & \to \mathbb{N}^{\mathbb{N}},\\ (G_n)_{n\in\mathbb{N}} & \mapsto \big(\Phi_{n,G_1,\dots,G_{n-1}}(G_n)\big)_{n\in\mathbb{N}}.\end{cases}$$

3. Assume $\Phi(\mathfrak{g}) = \Phi(\mathfrak{g}')$; then an inductive reasoning shows that $\mathfrak{g} = \mathfrak{g}'$. In fact, $G_1 = G_1'$, since $\Phi_{1,\emptyset}$ is injective. If $\mathfrak{g} \mid k = \mathfrak{g}' \mid k$ has already been established, we know that $\Phi_{k+1,G_1,\dots,G_k} = \Phi_{k+1,G_1',\dots,G_k'}$ is injective, so that $G_{k+1} = G_{k+1}'$ follows. A similar inductive argument shows that $\Phi(\mathfrak{g}) \mid k = \Phi(\mathfrak{g}') \mid k$, provided $\mathfrak{g} \mid k = \mathfrak{g}' \mid k$ for each $k \in \mathbb{N}$ holds. $\dashv$

The following lemmata collect some helpful properties:

Lemma 4.6.15 *$\mathfrak{g} = \mathfrak{g}'$ iff $\Phi(\mathfrak{g}) \mid k = \Phi(\mathfrak{g}') \mid k$ for all $k \in \mathbb{N}$, whenever $\mathfrak{g}, \mathfrak{g}' \in \mathcal{M}_A$.* $\dashv$

The argument for establishing the following statement uses the tree and the maps associated with it for constructing a suitable Souslin scheme.

Lemma 4.6.16 *Denote by $J_k := \{\alpha \mid k \mid \alpha \in \Phi\big[\mathcal{M}_A\big]\}$ all initial pieces of sequences in the image of Φ. Then $\alpha \in \Phi\big[\mathcal{M}_A\big]$ iff $\alpha \mid k \in J_k$ for all $k \in \mathbb{N}$.*

Proof Assume that $\alpha = \Phi(\mathfrak{g}) \in \Phi\big[\mathcal{M}_A\big]$ with $\mathfrak{g} = (C_n)_{n\in\mathbb{N}} \in \mathcal{M}_A$ and $\alpha \mid k \in J_k$ for all $k \in \mathbb{N}$, so for given k there exists $\mathfrak{g}^{(k)} = ({C^{(k)}}_n)_{n\in\mathbb{N}} \in \mathcal{M}_A$ with $\alpha \mid k = \Phi(\mathfrak{g}^{(k)}) \mid k$. Because Φ_1 is injective, we obtain $C_1 = C_1^{(1)}$. Assume for the induction step that $G_i = G_i^{(j)}$ has been shown for $1 \leq i, j \leq k$. Then we obtain from $\Phi(\mathfrak{g}) \mid k + 1 = \Phi(\mathfrak{g}^{(k+1)}) \mid k + 1$ that $G_1 = G_1^{(k+1)}, \ldots, G_k = G_k^{(k+1)}$. Since $\Phi_{k+1,G_1,\ldots,G_k}$ is injective, the equality above implies $G_{k+1} = G_{k+1}^{(k+1)}$. Hence $\mathfrak{g} = \mathfrak{g}^{(k)}$ for all $k \in \mathbb{N}$, and $\alpha \in \Phi\big[\mathcal{M}_A\big]$ is established. The reverse implication is trivial. $\dashv$

Lemma 4.6.17 *$E_r := \{x \in X \mid K^*(x)(A) \leq r\} \in \mathcal{S}_X$ for $r \in \mathbb{R}_+$.*

Proof The set E_r can be written as

$$E_r = \bigcup_{\mathfrak{g}\in\mathcal{N}_A} \{x \in X \mid K(x)\big(\bigcup \mathfrak{g}\big) \leq r\} \cup \bigcup_{\mathfrak{g}\in\mathcal{M}_A} \{x \in X \mid K(x)\big(\bigcup \mathfrak{g}\big) \leq r\}.$$

Because $\mathcal{N}_A$ is countable and $K : X \rightsquigarrow Y$ is a transition kernel, we infer

$$\bigcup_{\mathfrak{g}\in\mathcal{N}_A} \{x \in X \mid K(x)\big(\bigcup \mathfrak{g}\big) \leq r\} \in \mathcal{B}(X).$$

Put for $v \in \mathbb{N}^+$

$$D_v := \begin{cases} \emptyset, & \text{if } v \notin \bigcup_{k\in\mathbb{N}} J_k, \\ \{x \in X \mid K(x)(G_n) \leq r\}, & \text{if } v = \Phi\big((G_n)_{n\in\mathbb{N}}\big) \mid n. \end{cases}$$

Lemmata 4.6.15 and 4.6.16 show that $D_v \in \mathcal{B}(X)$ is well defined. Because

$$\bigcup_{\mathfrak{g}\in\mathcal{M}_A} \{x \in X \mid K(x)\big(\bigcup \mathfrak{g}\big) \leq r\} = \bigcup_{\alpha\in\mathbb{N}^{\mathbb{N}}} \bigcap_{n\in\mathbb{N}} D_{\alpha|n} \tag{4.8}$$

and because $\mathcal{S}_X$ is closed under the Souslin operation and contains $\mathcal{B}(X)$, we conclude that $E_r \in \mathcal{S}_X$. ⊣

Proposition 4.6.18 *Let $K : (X; \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ be a transition kernel, and assume that Y is a separable metric space. Let $\mathcal{S}_X$ be the smallest σ-algebra which contains $\mathcal{A}$ and which is closed under the Souslin operation. Then there exists a unique transition kernel*

$$\overline{K} : (X, \mathcal{S}_X) \rightsquigarrow (Y, \overline{\mathcal{B}(Y)}^{\{K(x)|x \in X\}})$$

extending K.

Proof 1. Put $\overline{K}(x)(A) := K^*(x)(A)$for $x \in X$ and $A \in \overline{\mathcal{B}(Y)}^{\{K(x)|x \in X\}}$. Because A is an element of the $K(x)$-completion of $\mathcal{B}(Y)$, we know that $\overline{K}(x) = \overline{K(x)}$ defines a subprobability on $\overline{\mathcal{B}(Y)}^{\{K(x)|x \in X\}}$. It is clear that $\overline{K}(x)$ is the unique extension of $K(x)$ to the latter σ-algebra. It remains to be shown that $\overline{K}$ is a transition kernel.

2. Fix $A \in \overline{\mathcal{B}(Y)}^{\{K(x)|x \in X\}}$ and $q \in [0, 1]$; then

$$\{x \in X \mid K^*(x)(A) < q\} = \bigcup_{\ell \in \mathbb{N}} \bigcup_{\mathfrak{g} \in \mathcal{G}_A} \{x \in X \mid K(x)\big(\bigcup \mathfrak{g}\big) \leq q - \frac{1}{\ell}\}$$

The latter set is a member of $\mathcal{S}_X$ by Lemma 4.6.17. ⊣

Separability of the target space is required because it is this property which makes sure that the measure for each Borel set can be approximated arbitrarily well from within by closed sets and from the outside by open sets (Lemma 4.6.13).

Before discussing consequences, a mild generalization to separable measurable spaces should be mentioned. Proposition 4.6.18 yields as an immediate consequence.

Corollary 4.6.19 *Let $K : (X; \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ be a transition kernel such that $(Y, \mathcal{B})$ is a separable measurable space. Assume that $\mathcal{X}$ is a σ-algebra on X which is closed under the Souslin operation with $\mathcal{S}_X \subseteq \mathcal{X}$ and that $\mathcal{Y}$ is a σ-algebra on X with $\mathcal{B} \subseteq \mathcal{Y} \subseteq \overline{\mathcal{B}}^{\{K(x)|x \in X\}}$. Then there exists a unique extension $(X, \mathcal{X}) \rightsquigarrow (Y, \mathcal{Y})$ to K. In particular K has a unique extension to a transition kernel $\overline{K} : (X, \overline{\mathcal{A}}) \rightsquigarrow (Y, \overline{\mathcal{B}})$.*

Proof This follows from Proposition 4.6.18 and the characterization of separable measurable spaces in Proposition 4.3.10. ⊣

4.7 Measurable Selections

Looking again at von Neumann's Selection Theorem 4.6.10, we found for a given surjection $f : X \to Y$ a universally measurable map $g : Y \to X$ with $f \circ g = id_Y$. This can be rephrased: We have $g(y) \in f^{-1}\big[\{y\}\big]$ for each $y \in Y$, so g may be considered a universally measurable selection for the set-valued map $y \mapsto f^{-1}\big[\{y\}\big]$.

We will consider constructing a selection from a slightly different angle by assuming that $(X, \mathcal{A})$ is a measurable, Y is a Polish space. In addition we are given a set-valued map $F : X \to \mathcal{P}(Y) \setminus \{\emptyset\}$ for which a measurable selection is to be constructed, i.e., a measurable (not merely universally measurable) map $g : X \to Y$ such that $g(x) \in F(x)$ for all $x \in X$. Clearly, the Axiom of Choice guarantees the existence of a map which picks an element from $F(x)$ for each x, but again this is not enough.

We assume that $F(x) \subseteq Y$, $F(x)$ is closed and $F(x) \neq \emptyset$ for all $x \in X$ and that it is measurable. Since F does not necessarily take single values only, we have to define measurability in this case. Denote by $\mathbb{F}(Y)$ the set of all closed and nonempty subsets of Y.

F^w

Definition 4.7.1 *A map $F : X \to \mathbb{F}(Y)$ from a measurable space $(X, \mathcal{A})$ to the closed nonempty subsets of a Polish space Y is called* measurable *(or a* measurable relation*) iff*

$$F^w(G) := \{x \in X \mid F(x) \cap G \neq \emptyset\} \in \mathcal{A}$$

for every open subset $G \subseteq Y$. The map $s : X \to Y$ is called a measurable selector *for F iff s is $\mathcal{A}$-$\mathcal{B}(Y)$-measurable such that $s(x) \in F(x)$ for all $x \in X$.*

Since $\{f(x)\} \cap G \neq \emptyset$ iff $f(x) \in G$, measurability as defined in this definition is a generalization of measurability for point-valued maps $f : X \to Y$.

The selection theorem due to Kuratowski and Ryll-Nardzewski tells us that a measurable selection exists for a measurable closed valued map, provided Y is Polish. To be specific

Kuratowski and Ryll-Nardzewski selection theorem

Theorem 4.7.2 *Given a measurable space $(X, \mathcal{A})$ and a Polish space Y, a measurable map $F : X \to \mathbb{F}(Y)$ has a measurable selector.*

Proof 0. Fix a complete metric d on Y. As usual, $B(y, r)$ is the open ball with center $y \in Y$ and radius $r > 0$. Recall that the distance of an element y to a closed set C is $d(y, C) := \inf\{d(y, y') \mid y' \in C\}$; hence $d(y, C) = 0$ iff $y \in C$. The idea of the proof is to define a sequence $(f_n)_{n\in\mathbb{N}}$ of measurable maps such that $f_n(x)$ comes closer and closer to $F(x)$, measured in terms of the distance $d(f_n(x), F(x))$ of $f_n(x)$ to $F(x)$, and that $\big(f_n(x)\big)_{n\in\mathbb{N}}$ is a Cauchy sequence in the complete space Y for each x. Idea

1. Let $(y_n)_{n\in\mathbb{N}}$ be dense, and define $f_1(x) := y_n$, if n is the smallest index k so that $F(x) \cap B(y_k, 1) \neq \emptyset$. Then $f_1 : X \to Y$ is $\mathcal{A}$-$\mathcal{B}(Y)$-measurable, because the map takes only a countable number of values and

$$\{x \in X \mid f_1(x) = y_n\} = F^w(B(y_n, 1)) \setminus \bigcup_{k=1}^{n-1} F^w(B(y_k, 1)).$$

Proceeding inductively, assume that we have defined measurable maps $f_1, \ldots, f_n$ such that

$$\begin{array}{lll} d(f_j(x), f_{j+1}(x)) & < 2^{-(j-1)}, & 1 \leq j < n, \\ d(f_j(x), F(x)) & < 2^{-j}, & 1 \leq j \leq n. \end{array}$$

Put $X_k := \{x \in X \mid f_n(x) = y_k\}$, and define $f_{k+1}(x) := y_\ell$ for $x \in X_k$, where ℓ is the smallest index m such that $F(x) \cap B(y_k, 2^{-n}) \cap B(y_m, 2^{-(n+1)}) \neq \emptyset$. Moreover, there exists $y' \in B(y_k, 2^{-n}) \cap B(y_m, 2^{-(n+1)})$; thus

$$d(f_n(x), f_{n+1}(x)) \leq d(f_n(x), y') + f(f_{n+1}(x), y') < 2^{-n} + 2^{-(n+1)}.$$

The argumentation from above shows that f_{n+1} takes only countably many values, and we know that $d(f_{n+1}(x), F(x)) < 2^{-(n+1)}$.

2. Thus $(f_n(x))_{n\in\mathbb{N}})$ is a Cauchy sequence for each $x \in X$. Since (Y, d) is complete, the limit $f(x) := \lim_{n\to\infty} f_n(x)$ exists with $d(f(x), F(x)) = 0$; hence $f(x) \in F(x)$, because $F(x)$ is closed. Moreover, as a pointwise limit of a sequence of measurable functions, f is measurable, so f is the desired measurable selector. $\dashv$

It is possible to weaken the conditions on F and on $\mathcal{A}$; see Exercise 4.23. This theorem has an interesting consequence, viz., that we can find a sequence of dense selectors for F.

Corollary 4.7.3 *Under the assumptions of Theorem 4.7.2, a measurable map $F : X \to \mathbb{F}(Y)$ has a sequence $(f_n)_{n\in\mathbb{N}}$ of measurable selectors such that $\{f_n(x) \mid n \in \mathbb{N}\}$ is dense in $F(x)$ for each $x \in X$.*

Proof 1. We use notations from above. Let again $(y_n)_{n\in\mathbb{N}}$ be a dense sequence in Y, and define for $n, m \in \mathbb{N}$ the map

$$F_{n,m}(x) := \begin{cases} F(x) \cap B(y_n, 2^{-m}), & \text{if } x \in F^w(B(y_n, 2^{-m})) \\ F(x), & \text{otherwise.} \end{cases}$$

Denote by $H_{n,m}(x)$ the closure of $F_{n,m}(x)$.

2. $H_{n,m} : X \to \mathbb{F}(Y)$ is measurable. In fact, put $A_1 := F^w(B(y_n, 2^{-m}))$, $A_2 := X \setminus A_1$; then $A_1, A_2 \in \mathcal{A}$, because F is measurable and $B(y_n, 2^{-m})$ is open. But then we have for an open set $G \subseteq Y$

$$\begin{aligned} \{x \in X \mid H_{n,m} \cap G \neq \emptyset\} =& \{x \in X \mid F_{n,m} \cap G \neq \emptyset\} \\ =& \{x \in A_1 \mid F(x) \cap G \cap B(y_n, 2^{-m}) \neq \emptyset\} \\ & \cup \{x \in A_2 \mid F(x) \cap G \neq \emptyset\}; \end{aligned}$$

thus $H^w_{n,m}(G) \in \mathcal{A}$.

3. We can find a measurable selector $s_{n,m}$ for $H_{n,m}$ by Theorem 4.7.2, so we have to show that $\{s_{n,m}(x) \mid n, m \in \mathbb{N}\}$ is dense in $F(x)$ for each $x \in X$. Let $y \in F(x)$. Given $\epsilon > 0$, select m with $2^{-m} < \epsilon/2$; there exists y_n with $d(y, y_n) < 2^{-m}$. Thus $x \in H^w_{n,m}(B(y_n, 2^{-m}))$, and $s_{n,m}(x)$ is a member of the closure of $B(y_n, 2^{-m})$, which means $d(y, s_{n,m}(x)) < \epsilon$. Now arrange $\{s_{n,m}(x) \mid n, m \in \mathbb{N}\}$ as a sequence; then the assertion follows. $\dashv$

This is a first application of measurable selections.

Example 4.7.4 Call a map $h : X \to \mathcal{B}(Y)$ for the Polish space Y *hit-measurable* iff h is measurable with respect to $\mathcal{A}$ and $\mathcal{H}_{\mathcal{G}}(\mathcal{B}(Y))$, where $\mathcal{G}$ is the set of all open sets in Y; see Example 4.1.2. Thus h is hit-measurable iff $\{x \in X \mid h(x) \cap U \neq \emptyset\} \in \mathcal{A}$ for each open set $U \subseteq Y$. If h is image finite (i.e., $h(x)$ is always nonempty and finite), then there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of measurable maps $f_n : X \to Y$ such that $h(x) = \{f_n(x) \mid n \in \mathbb{N}\}$ for each $x \in X$. This is so because $h : X \to \mathbb{F}(Y)$ is measurable; hence Corollary 4.7.3 is applicable. ✌

Transition kernels into Polish spaces induce a measurable closed valued map, for which selectors exist.

Example 4.7.5 Let under the assumptions of Theorem 4.7.2 $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B}(Y))$ be a transition kernel with $K(x)(Y) > 0$ for all $x \in X$. Then there exists a measurable map $f : X \to Y$ such that $K(x)(U) > 0$, whenever U is an open neighborhood of $f(x)$.

In fact, $F : x \mapsto \text{supp}(K(x))$ takes nonempty and closed values by Lemma 4.1.46. If $G \subseteq Y$ is open, then

$$F^w(G) = \{x \in X \mid \text{supp}(K(x)) \cap G \neq \emptyset\} = \{x \in X \mid K(x)(G) > 0\} \in \mathcal{A}.$$

Thus F has a measurable selector f by Theorem 4.7.2. The assertion now follows from Corollary 4.1.47.

Perceiving a stochastic relation $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B}(Y))$ as a probabilistic model for transitions such that $K(x)(B)$ is the probability for making a transition from x to B (with $K(x)(Y) \leq 1$), we may interpret the selection f as one possible deterministic version for a transition: The state $f(x)$ is possible, since $f(x) \in \text{supp}(K(x))$, which entails $K(x)(U) > 0$ for every open neighborhood U of $f(x)$. There exists a sequence $(f_n)_{n\in\mathbb{N}}$ of measurable selectors for F such that $\{f_n(x) \mid n \in \mathbb{N}\}$ is dense in $F(x)$; this may be interpreted as a form of stochastic nondeterminism. ✌

4.8 Integration

After having studied the structure of measurable sets under various conditions on the underlying space with an occasional side glance at real-valued measurable functions, we will discuss integration now. This is a fundamental operation associated with measures. The integral of a function with respect to a measure will be what you expect it to be, viz., for nonnegative functions the area between the curve and the x-axis. This view will be confirmed later on, when Fubini's Theorem will be available for computing measures in Cartesian products. For the time being, we build up the integral in a fairly straightforward way through an approximation by step functions, obtaining a linear map with some favorable properties, for example, the Lebesgue Dominated Convergence Theorem. All the necessary constructions are given in this section, offering more than one occasion to exercise the well-known ϵ-δ-arguments, which are necessary, but not particularly entertaining. But that is life.

The second part of this section offers a complementary view—it starts from a positive linear map with some additional continuity properties and develops a measure from it. This is Daniell's approach, suggesting that measure and integral are really most of the time two sides of the same coin. We show that this duality comes to life especially when we are dealing with a compact metric space: Here the celebrated Riesz Representation Theorem gives a bijection between probability measures on the Borel sets and normed positive linear functions on the continuous real-valued functions. We formulate and prove this theorem here; it should be mentioned that this is not the most general version available, as with most other results discussed here (but probably there is no such thing as a *most general version*, since the development did branch out into wildly different directions).

This section will be fundamental for the discussions and results later in this chapter. Most results are formulated for finite or σ-finite measures, and usually no attempt has been made to find the boundary delineating a development.

4.8.1 From Measure to Integral

We fix a measure space $(X, \mathcal{A}, \mu)$. Denote for the moment by $\mathcal{T}(X, \mathcal{A})$ the set of all measurable step functions, and by $\mathcal{T}_+(X, \mathcal{A})$ the nonnegative step functions; similarly, $\mathcal{F}_+(X, \mathcal{A})$ are the nonnegative measurable functions. Note that $\mathcal{T}(X, \mathcal{A})$ is a vector space under the usual operations and that it is a lattice under finite or countable pointwise suprema and infima. We know from Proposition 4.2.4 that we can approximate each bounded measurable function by a sequence of step functions from below.

$\mathcal{T}(X, \mathcal{A})$

$\mathcal{F}(X, \mathcal{A})$

Define

$$\int_X \sum_{i=1}^n \alpha_i \cdot \chi_{A_i} \, d\mu := \sum_{i=1}^n \alpha_i \cdot \mu(A_i) \tag{4.9}$$

as the *integral with respect to* μ for the step function $\sum_{i=1}^n \alpha_i \cdot \chi_{A_i} \in \mathcal{T}(X, \mathcal{A})$. Exercise 4.24 tells us that the integral is well defined: If $f, g \in \mathcal{T}(X, \mathcal{A})$ with $f = g$, then

$$\sum_{\alpha \in \mathbb{R}} \alpha \cdot \mu(\{x \in X \mid f(x) = \alpha\}) = \sum_{\beta \in \mathbb{R}} \beta \cdot \mu(\{x \in X \mid g(x) = \beta\}).$$

Thus the definition (4.9) yields the same value for the integral. These are some elementary properties of the integral for step functions.

Lemma 4.8.1 *Let $f, g \in \mathcal{T}(X, \mathcal{A})$ be step functions, $\alpha \in \mathbb{R}$. Then*

1. $\int_X \alpha \cdot f \, d\mu = \alpha \cdot \int_X f \, d\mu$,
2. $\int_X (f + g) \, d\mu = \int_X f \, d\mu + \int_X g \, d\mu$,
3. *if* $f \geq 0$, *then* $\int_X f \, d\mu \geq 0$; *in particular, the map* $f \mapsto \int_X f \, d\mu$ *is monotone,*
4. $\int_X \chi_A \, d\mu = \mu(A)$ *for* $A \in \mathcal{A}$,
5. $|\int_X f \, d\mu| \leq \int_X |f| \, d\mu$.

Moreover the map $A \mapsto \int_A f \, d\mu := \int_X f \cdot \chi_A \, d\mu$ *is additive on* $\mathcal{A}$ *whenever* $f \in \mathcal{T}_+(X, \mathcal{A})$. $\dashv$

We know from Proposition 4.2.4 that we can find for $f \in \mathcal{F}_+(X, \mathcal{A})$ a sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{T}_+(X, \mathcal{A})$ such that $f_1 \leq f_2 \leq \ldots$ and $\sup_{n\in\mathbb{N}} f_n = f$. This observation is used for the definition of the integral for f. We define

$$\int_X f \, d\mu := \sup\{\int_X g \, d\mu \mid g \leq f \text{ and } g \in \mathcal{T}_+(X, \mathcal{A})\}.$$

Note that the right-hand side may be infinite; we will discuss this shortly.

The central observation is formulated in Levi's Theorem.

Theorem 4.8.2 *Let $(f_n)_{n\in\mathbb{N}}$ be an increasing sequence of functions in $\mathcal{F}_+(X, \mathcal{A})$ with limit f; then the limit of $\left(\int_X f_n \, d\mu\right)_{n\in\mathbb{N}}$ exists and equals $\int_X f \, d\mu$.* Levi's Theorem

Proof 1. Because the integral is monotone in the integrand by Lemma 4.8.1, the limit

$$\ell := \lim_{n\to\infty} \int_X f_n \, d\mu$$

exists (possibly in $\mathbb{R} \cup \{\infty\}$), and we know from monotonicity that $\ell \leq \int_X f \, d\mu$.

2. Let $f = c > 0$ be a constant, and let $0 < d < c$. Then $\sup_{n\in\mathbb{N}} d \cdot$

$\chi_{\{x\in X\mid f_n(x)\geq d\}} = d$; hence we obtain

$$\begin{aligned}\int_X f\,d\mu \geq \int_X f_n\,d\mu &\geq \int_{\{x\in X\mid f_n(x)\geq d\}} f_n\,d\mu\\ &\geq d\cdot\mu(\{x\in X\mid f_n(x)\geq d\})\end{aligned}$$

for every $n\in\mathbb{N}$; thus

$$\int_X f\,d\mu \geq d\cdot\mu(X).$$

Letting d approach c, we see that

$$\int_X f\,d\mu \geq \lim_{n\to\infty}\int_X f_n\,d\mu \geq c\cdot\mu(X) = \int_X f\,d\mu.$$

This gives the desired equality.

3. If $f = c\cdot\chi_A$ with $A\in\mathcal{A}$, we restrict the measure space to $(A, \mathcal{A}\cap A, \mu)$, so the result is true also for step functions based on one single set.

4. Let $f = \sum_{i=1}^n \alpha_i\cdot\chi_{A_i}$ be a step function; then we may assume that the sets $A_1,\ldots,A_n$ are mutually disjoint. Consider $f_i := f\cdot\chi_{A_i} = \alpha_i\cdot\chi_{A_i}$ and apply the previous step to f_i, taking additivity from Lemma 4.8.1, part 2 into account.

5. Now consider the general case. Select step function $(g_n)_{n\in\mathbb{N}}$ with $g_n\in\mathcal{T}_+(X,\mathcal{A})$ such that $g_n\leq f_n$ and $|\int_X f_n\,d\mu - \int_X g_n\,d\mu| < 1/n$. We may and do assume that $g_1\leq g_2\leq\ldots$, for we otherwise may pass to the step function $h_n := \sup\{g_1,\ldots,g_n\}$. Let $0\leq g\leq f$ be a step function; then $\lim_{n\to\infty}(g_n\wedge g) = g$, so that we obtain from the previous step

$$\int_X g\,d\mu = \lim_{n\to\infty}\int_X g_n\wedge g\,d\mu \leq \lim_{n\to\infty}\int_X g_n\,d\mu \leq \lim_{n\to\infty}\int_X f_n\,d\mu.$$

Because $\int_X g\,d\mu$ may be chosen arbitrarily close to ℓ, we finally obtain

$$\lim_{n\to\infty}\int_X f_n\,d\mu \leq \int_X f\,d\mu \leq \lim_{n\to\infty}\int_X f_n\,d\mu,$$

which implies the assertion for arbitrary $f\in\mathcal{F}_+(X,\mathcal{A})$. $\dashv$

Since we can approximate each nonnegative measurable function from below and from above by step functions (Proposition 4.2.4 and Exercise 4.7), we obtain from Levi's Theorem for $f \in \mathcal{F}_+(X, \mathcal{A})$ the representation

$$\sup\{\int_X g\, d\mu \mid \mathcal{T}_+(X, \mathcal{A}) \ni g \leq f\}$$
$$= \int_X f\, d\mu = \inf\{\int_X g\, d\mu \mid f \leq g \in \mathcal{T}_+(X, \mathcal{A})\}.$$

This strongly resembles—and generalizes—the familiar construction of the Riemann integral for a continuous function f over a bounded interval by sandwiching it between lower and upper sums of step functions.

Compatibility of the integral with scalar multiplication and with addition is now an easy consequence of Levi's Theorem.

Corollary 4.8.3 *Let $a \geq 0$ and $b \geq 0$ be nonnegative real numbers; then*

$$\int_X a \cdot f + b \cdot g\, d\mu = a \cdot \int_X f\, d\mu + b \cdot \int_X g\, d\mu$$

for $f, g \in \mathcal{F}_+(X, \mathcal{A})$.

Proof Let $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ be sequences of step functions which converge monotonically to f resp. g. Then $(a \cdot f_n + b \cdot g_n)_{n\in\mathbb{N}}$ is a sequence of step functions converging monotonically to $a \cdot f + b \cdot g$. Apply Levi's Theorem 4.8.2 and the linearity of the integral on step functions from Lemma 4.8.1 to obtain the assertion. $\dashv$

Given an arbitrary $f \in \mathcal{F}(X, \mathcal{A})$, we can decompose f into a positive and a negative part $f^+ := f \vee 0$ resp. $f^- := (-f) \vee 0$, so that $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

A function $f \in \mathcal{F}(X, \mathcal{A})$ is called *integrable* (with respect to μ) iff Integrable

$$\int_X |f|\, d\mu < \infty;$$

in this case we set

$$\int_X f\, d\mu := \int_X f^+\, d\mu - \int_X f^-\, d\mu.$$

In fact, because $f^+ \leq f$, we obtain from Lemma 4.8.1 that $\int_X f^+ \, d\mu < \infty$; similarly we see that $\int_X f^- \, d\mu < \infty$. The integral is well defined, because if $f = f_1 - f_2$ with $f_1, f_2 \geq 0$, we conclude $f_1 \leq f \leq |f|$, hence $\int_X f_1 \, d\mu < \infty$, and $f_2 \leq |f|$, so that $\int_X f_2 \, d\mu < \infty$, which implies $\int_X f^+ \, d\mu + \int_X f_2 \, d\mu = \int_X f^- \, d\mu + \int_X f_1 \, d\mu$ by Corollary 4.8.3. Thus we obtain in fact $\int_X f^+ \, d\mu - \int_X f^- \, d\mu = \int_X f_1 \, d\mu - \int_X f_2 \, d\mu$.

This special case is also of interest: Let $A \in \mathcal{A}$, define for f integrable

$$\int_A f \, d\mu := \int_X f \cdot \chi_A \, d\mu$$

(note that $|f \cdot \chi_A| \leq |f|$). We emphasize occasionally the integration variable by writing $\int_X f(x) \, d\mu(x)$ instead of $\int_X f \, d\mu$.

Collecting some useful and a.e. used properties, we state

Proposition 4.8.4 *Let $f, g \in \mathcal{F}(X, \mathcal{A})$ be measurable functions; then*

1. *If $f \geq_\mu 0$, then $\int_X f \, d\mu = 0$ iff $f =_\mu 0$.*
2. *If f is integrable and $|g| \leq_\mu |f|$, then g is integrable.*
3. *If f and g are integrable, then so are $a \cdot f + b \cdot g$ for all $a, b \in \mathbb{R}$, and $\int_X a \cdot f + b \cdot g \, d\mu = a \cdot \int_X f \, d\mu + b \cdot \int_X g \, d\mu$.*
4. *If f and g are integrable and $f \leq_\mu g$, then $\int_X g \, d\mu \leq \int_X f \, d\mu$.*
5. *If f is integrable, then $|\int_X f \, d\mu| \leq \int_X |f| \, d\mu$.*

$\dashv$

We now state and prove some statements which relate sequences of functions to their integrals. The first one is traditionally called *Fatou's Lemma*.

Fatou's Lemma

Proposition 4.8.5 *Let $(f_n)_{n \in \mathbb{N}}$ be a sequence in $\mathcal{F}_+(X, \mathcal{A})$. Then*

$$\int_X \liminf_{n \to \infty} f_n \, d\mu \leq \liminf_{n \to \infty} \int_X f_n \, d\mu.$$

Proof Since $(\inf_{m \geq n} f_m)_{n \in \mathbb{N}}$ is an increasing sequence of measurable functions in $\mathcal{F}_+(X, \mathcal{A})$, we obtain from Levi's Theorem 4.8.2

$$\int_X f \, d\mu = \lim_{n \to \infty} \int_X \inf_{m \geq n} f_m \, d\mu = \sup_{n \in \mathbb{N}} \int_X \inf_{m \geq n} f_m \, d\mu.$$

Because we plainly have by monotonicity $\int_X \inf_{m\geq n} f_m \, d\mu \leq \inf_{m\geq n} \int_X f_m \, d\mu$, the assertion follows. ⊣

The *Lebesgue Dominated Convergence Theorem* is a very important and eagerly used tool; it can be derived now easily from Fatou's Lemma.

Theorem 4.8.6 *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions with $f_n \xrightarrow{a.e.} f$ for some measurable function f and $|f_n| \leq_\mu g$ for all $n \in \mathbb{N}$ and an integrable function g. Then f_n and f are integrable, and*

Lebesgue Dominated Convergence Theorem

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f \, d\mu \text{ and } \lim_{n\to\infty} \int_X |f_n - f| \, d\mu = 0.$$

Proof 1. It is no loss of generality to assume that $f_n \to f$ and $\forall n \in \mathbb{N} : f_n \leq g$ pointwise (otherwise modify the f_n, f, and g on a set of μ-measure zero). Because $|f_n| \leq g$, we conclude from Proposition 4.8.4 that f_n is integrable, and since $f \leq g$ holds as well, we infer that f is also integrable.

2. Put $g_n := |f| + g - |f_n - f|$, then $g_n \geq 0$, and g_n is integrable for all $n \in \mathbb{N}$. We obtain from Fatou's Lemma

$$\begin{aligned}
\int_X |f| + g \, d\mu &= \int_X \liminf_{n\to\infty} g_n \, d\mu \\
&\leq \liminf_{n\to\infty} \int_X g_n \, d\mu \\
&= \int_X |f| + g \, d\mu - \limsup_{n\to\infty} \int_X |f_n - f| \, d\mu.
\end{aligned}$$

Hence we obtain $\limsup_{n\to\infty} \int_X |f_n - f| \, d\mu = 0$, thus $\lim_{n\to\infty} \int_X |f_n - f| \, d\mu = 0$.

3. We finally note that

$$\left|\int_X f_n \, d\mu - \int_X f \, d\mu\right| = \left|\int_X (f_n - f) \, d\mu\right| \leq \int_X |f_n - f| \, d\mu,$$

which completes the proof. ⊣

The following is an immediate consequence of the Lebesgue Theorem. We know from Calculus that interchanging integration and infinite summation may be dangerous, so we gain a good criterion here permitting this operation.

Corollary 4.8.7 *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions, g integrable, such that $|\sum_{k=1}^{n} f_k| \leq_\mu g$ for all $n \in \mathbb{N}$. Then all f_n and $f := \sum_{n\in\mathbb{N}} f_n$ are integrable, and $\int_X f\ d\mu = \sum_{n\in\mathbb{N}} \int_X f_n\ d\mu$.* $\dashv$

Moreover, we conclude that each nonnegative measurable function begets a finite measure. This observation will be fruitful for the discussion of L_p-spaces in Sect. 4.11.

Corollary 4.8.8 *Let $f \geq_\mu 0$ be an integrable function; then $A \mapsto \int_A f\ d\mu$ defines a finite measure on $\mathcal{A}$.*

Proof All the properties of a measure are immediate, and σ-additivity follows from Corollary 4.8.7. $\dashv$

Integration with respect to an image measure is also available right away. It yields the fairly helpful *change of variables formula* for image measures.

Change of variables

Corollary 4.8.9 *Let $(Y, \mathcal{B})$ a measurable space and $g : X \to Y$ be $\mathcal{A}$-$\mathcal{B}$-measurable. Then $h \in \mathcal{F}(Y, \mathcal{B})$ is $\mathbb{M}(g)(\mu)$ integrable iff $g \circ h$ is μ-integrable, and in this case we have*

$$\int_Y h\ d\mathbb{M}(g)(\mu) = \int_X h \circ g\ d\mu. \tag{4.10}$$

Proof We show first that formula (4.10) is true for step functions. In fact, if $h = \chi_B$ with a measurable set B, then we obtain from the definition

$$\int_Y \chi_B\ d\mathbb{M}(g)(\mu) = \mathbb{M}(g)(\mu)(B) = \mu(g^{-1}\left[B\right]) = \int_X \chi_B \circ g\ d\mu$$

(since $\chi_B(g(x)) = 1$ iff $x \in g^{-1}\left[B\right]$). This observation extends by linearity to step functions, so that we obtain for $h = \sum_{i=1}^{n} b_i \cdot \chi_{B_i}$

$$\int_Y h\ d\mathbb{M}(g)(\mu) = \sum_{i=1}^{n} b_i \cdot \int_X \chi_{B_i} \circ g\ d\mu = \int_X h \circ g\ d\mu.$$

Thus the assertion now follows from Levi's Theorem 4.8.2. $\dashv$

The reader is probably familiar with the change of variables formula in classical calculus. It deals with k-dimensional Lebesgue measure λ^k and a differentiable and injective map $T : V \to W$ from an open set $V \subseteq \mathbb{R}^k$ to a bounded set $W \subseteq \mathbb{R}^k$. T is assumed to have a continuous inverse. Then the integral of a measurable and bounded function $f : T\big[V\big] \to \mathbb{R}$ can be expressed in terms of the integral over V of $f \circ T$ and the Jacobian J_T of T. To be specific

$$\int_{T\left[V\right]} f \, d\lambda^k = \int_V (f \circ T) \cdot |J_T| \, d\lambda^k.$$

Recall that the Jacobian J_T of T is the determinant of the partial derivatives of T, i.e.,

$$J_T(x) = \det\big(\big(\frac{\partial T_i(x)}{\partial x_j}\big)\big).$$

This representation can be derived from the representation for the integral with respect to the image measure from Corollary 4.8.9 and from the Radon–Nikodym Theorem 4.11.26 through a somewhat lengthy application of results from fairly elementary linear algebra. We do not want to develop this apparatus in the present presentation; we will, however, provide a glimpse at the one-dimensional situation in Proposition 4.11.29. The reader is referred for the general case rather to Rudin's exposition [Rud74, pp. 181–188] or to Stromberg's more elementary discussion in [Str81, pp. 385–392]; if you read German, Elstrodt's derivation [Els99, §V.4] should not be missed.

4.8.2 The Daniell Integral and Riesz Representation Theorem

The previous section developed the integral from a finite or σ-finite measure; the result was a linear functional on a subspace of measurable functions, which will be investigated in greater detail later on. This section will demonstrate that it is possible to obtain a measure from a linear functional on a well- behaved space of functions. This approach was proposed by P. J. Daniell ca. 1920; it is called in his honor the *Daniell integral*. It is useful when a linear functional is given, and one wants to show that this functional is actually defined by a measure, which then permits putting the machinery of measure theory into action. We will encounter such a situation, e.g., when studying linear functionals on spaces of integrable functions. Specifically, we derive the Riesz Representation Theorem, which shows that there is a one-to-one correspondence between probability measures and normed positive linear functionals on the vector lattice of continuous real-valued functions on a compact metric space.

Let us fix a set X throughout. We will also fix a set $\mathcal{F}$ of functions $X \to \mathbb{R}$ which is assumed to be a vector space (as always, over the reals) with a special property.

Definition 4.8.10 *A vector space $\mathcal{F} \subseteq \mathbb{R}^X$ is called a* vector lattice *iff $|f| \in \mathcal{F}$ whenever $f \in \mathcal{F}$.*

Now fix the vector lattice $\mathcal{F}$. Each vector lattice is indeed a lattice: Define

$$
\begin{aligned}
f \vee g &:= (|f - g| + f + g)/2,\\
f \wedge g &:= -\big((-f) \vee (-g)\big)\\
f \leq g &\Leftrightarrow f \vee g = g\\
&\Leftrightarrow f \wedge g = f.
\end{aligned}
$$

Thus $\mathcal{F}$ contains f and g also $f \wedge g$ and $f \vee g$, and it is easy to see that $\leq$ defines a partial order on $\mathcal{F}$ such that $\sup\{f, g\} = f \vee g$ and $\inf\{f, g\} = f \wedge g$. Note that we have $\max\{\alpha, \beta\} = (|\alpha - \beta| + \alpha + \beta)/2$ for $\alpha, \beta \in \mathbb{R}$; thus we conclude that $f \leq g$ iff $f(x) \leq g(x)$ for all $x \in \mathbb{R}$.

We will find these properties helpful; they will be used silently below.

Lemma 4.8.11 *If $0 \leq \alpha \leq \beta \in \mathbb{R}$ and $f \in \mathcal{F}$ with $f \geq 0$, then $\alpha \cdot f \leq \beta \cdot f$. If $f, g \in \mathcal{F}$ with $f \leq g$, then $f + h \leq g + h$ for all $h \in \mathcal{F}$. Also, $f \wedge g + f \vee g = f + g$.*

Proof Because $f \geq 0$, we obtain from $\alpha \leq \beta$

$$2 \cdot \big((\alpha \cdot f) \vee (\beta \cdot f)\big) = (|\alpha - \beta| + \alpha + \beta) \cdot f = 2 \cdot (\alpha \vee \beta) \cdot f = 2 \cdot \beta \cdot f.$$

This establishes the first claim. The second one follows from

$$2 \cdot \big((f + h) \vee (g + h)\big) = |f - g| + f + g + 2 \cdot h = 2 \cdot (g + h).$$

The third one is established through the observation that it holds pointwise and from the observation that $f \leq g$ iff $f(x) \leq g(x)$ for all $x \in X$. $\dashv$

We assume that $1 \in \mathcal{F}$ and that a function $L : \mathcal{F} \to \mathbb{R}$ is given, which has these properties:

- $L(\alpha \cdot f + \beta \cdot g) = \alpha \cdot L(f) + \beta \cdot L(g)$, so that L is linear,

- if $f \geq 0$, then $L(f) \geq 0$, so that L is positive,
- $L(1) = 1$, so that L is normed,
- If $(f_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{F}$ which decreases monotonically to 0, then $\lim_{n\to\infty} L(f_n) = 0$, so that L is continuous from above at 0.

These are some immediate consequences from the properties of L.

Lemma 4.8.12 *If $f, g \in \mathcal{F}$, then $L(f \wedge g) + L(f \vee g) = L(f) + L(g)$. If $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ are increasing sequences of nonnegative functions in $\mathcal{F}$ with $\lim_{n\to\infty} f_n \leq \lim_{n\to\infty} g_n$, then $\lim_{n\to\infty} L(f_n) \leq \lim_{n\to\infty} L(g_n)$.*

Proof The first property follows from the linearity of L. For the second one, we observe that $\lim_{k\to\infty}(f_n \wedge g_k) = f_n \in \mathcal{F}$, the latter sequence being increasing. Consequently, we have

$$L(f_n) \leq \lim_{k\to\infty} L(f_n \wedge g_k) \leq \lim_{k\to\infty} L(g_k)$$

for all $n \in \mathbb{N}$, which implies the assertion. $\dashv$

$\mathcal{F}$ determines a σ-algebra $\mathcal{A}$ on X, viz., the smallest σ-algebra which renders each $f \in \mathcal{F}$ measurable. We will show now that L determines a unique probability measure on $\mathcal{A}$ such that

$$L(f) = \int_X f \, d\mu$$

holds for all $f \in \mathcal{F}$.

This will be done in a sequence of steps. A brief outline looks like this: We will first show that L can be extended to the set $\mathcal{L}^+$ of all bounded monotone limits from the nonnegative elements of $\mathcal{F}$ and that the extension respects monotone limits. From $\mathcal{L}^+$ we extract via indicator functions an algebra of sets and from the extension of L an outer measure. This will then turn out to yield the desired probability. Outline

Define

$$\mathcal{L}^+ := \{f : X \to \mathbb{R} \mid f \text{ is bounded; there exists } 0 \leq f_n \in \mathcal{F} \text{ increasing with } f = \lim_{n\to\infty} f_n\}.$$

Define $L(f) := \lim_{n\to\infty} L(f_n)$ for $f \in \mathcal{L}^+$, whenever $f = \lim_{n\to\infty} f_n$ for the increasing sequence $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}$. Then we obtain from Lemma 4.8.12 that this extension L on $\mathcal{L}^+$ is well defined, and it is clear that $L(f) \geq 0$ and that $L(\alpha\cdot f + \beta\cdot g) = \alpha\cdot L(f) + \beta\cdot L(g)$, whenever $f, g \in \mathcal{L}^+$ and $\alpha, \beta \in \mathbb{R}_+$. We see also that $f, g \in \mathcal{L}^+$ implies that $f \wedge g, f \vee g \in \mathcal{L}^+$ with $L(f \wedge g) + L(f \vee g) = L(f) + L(g)$. It turns out that L also respects the limits of increasing sequences.

Lemma 4.8.13 *Let $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{L}^+$ be an increasing and uniformly bounded sequence; then $L(\lim_{n\to\infty} f_n) = \lim_{n\to\infty} L(f_n)$.*

Proof Because $f_n \in \mathcal{L}^+$, we know that there exists for each $n \in \mathbb{N}$ an increasing sequence $(f_{m,n})_{m\in\mathbb{N}}$ of elements $f_{m,n} \in \mathcal{F}$ such that $f_n = \lim_{m\to\infty} f_{m,n}$. Define

$$g_m := \sup_{n\leq m} f_{m,n}.$$

Then $(g_m)_{m\in\mathbb{N}}$ is an increasing sequence in $\mathcal{F}$ with $f_{m,n} \leq g_m$, and $g_m \leq f_1 \vee f_2 \vee \ldots \vee f_m = f_m$, so that g_m is sandwiched between $f_{m,n}$ and f_m for all $m \in \mathbb{N}$ and $n \leq m$. This yields $L(f_{m,n}) \leq L(g_m) \leq L(f_m)$ for these n, m. Thus $\lim_{n\to\infty} f_n = \lim_{m\to\infty} g_m$, and hence

$$\lim_{n\to\infty} L(f_n) = \lim_{m\to\infty} L(g_m) = L(\lim_{m\to\infty} g_m) = L(\lim_{n\to\infty} f_n).$$

Thus we have shown that $\lim_{n\to\infty} f_n$ can be obtained as the limit of an increasing sequence of functions from $\mathcal{F}$; because $(f_n)_{n\in\mathbb{N}}$ is uniformly bounded, this limit is an element of $\mathcal{L}^+$. $\dashv$

Now define

$$\mathcal{G} := \{G \subseteq X \mid \chi_G \in \mathcal{L}^+\},$$
$$\mu(G) := L(\chi_G) \text{ for } G \in \mathcal{G}.$$

Then $\mathcal{G}$ is closed under finite intersections and finite unions by the remarks made before Lemma 4.8.13. Moreover, $\mathcal{G}$ is closed under countable unions with $\mu(\bigcup_{n\in\mathbb{N}} G_n) = \lim_{n\to\infty} \mu(G_n)$, if $(G_n)_{n\in\mathbb{N}}$ is an increasing sequence in $\mathcal{G}$. Also $\mu(X) = 1$. Now define, as in the Carathéodory approach in Sect. 1.6.3,

$$\mu^*(A) := \inf\{\mu(G) \mid G \in \mathcal{G}, A \subseteq G\},$$
$$\mathcal{B} := \{B \subseteq X \mid \mu^*(B) + \mu^*(X \setminus B) = 1\}.$$

We obtain from the Carathéodory extension process (Theorem 1.6.29)

Proposition 4.8.14 *$\mathcal{B}$ is a σ-algebra, and μ^* is countably additive on $\mathcal{B}$.*$\dashv$

Put $\mu(B) := \mu^*(B)$ for $B \in \mathcal{B}$, then $(X, \mathcal{B}, \mu)$ is a measure space, and μ is a probability measure on $(X, \mathcal{B})$.

In order to carry out the program sketched above, we need a σ-algebra. We have on one hand the σ-algebra $\mathcal{A}$ generated by $\mathcal{F}$ and on the other hand $\mathcal{B}$ gleaned from the Carathéodory extension. It is not immediately clear how these σ-algebras are related to each other. And then we also have $\mathcal{G}$ as an intermediate family of sets, obtained from $\mathcal{L}^+$. This diagram shows the objects we will discuss, together with a shorthand indication of the respective relationships:

A bird's eye view

$$\begin{tikzcd}
\mathcal{F} \arrow[rr, "\sigma(\cdot)"] \arrow[d] & & \mathcal{A} \arrow[d, dash, "="] \\
\mathcal{L}^+ \arrow[r, "\chi"'] & \mathcal{G} \arrow[ur, "\sigma(\cdot)="] \arrow[r, "\subseteq", "\text{Carathéodory}"'] & \mathcal{B}
\end{tikzcd}$$

We investigate the relationship of $\mathcal{A}$ and $\mathcal{G}$ first.

Lemma 4.8.15 $\mathcal{A} = \sigma(\mathcal{G})$.

Proof 1. Because $\mathcal{A}$ is the smallest σ-algebra rendering all elements of $\mathcal{F}$ measurable and because each element of $\mathcal{L}^+$ is the limit of a sequence of elements of $\mathcal{F}$, we obtain $\mathcal{A}$-measurability for each element of $\mathcal{L}^+$. Thus $\mathcal{G} \subseteq \mathcal{A}$.

2. Let $f \in \mathcal{L}^+$ and $c \in \mathbb{R}_+$; then $f_n := 1 \wedge n \cdot \sup\{f - c, 0\} \in \mathcal{L}^+$, and $\chi_{\{x \in X \mid f(x) > c\}} = \lim_{n\to\infty} f_n$. This is a monotone limit. Hence $\{x \in X \mid f(x) > c\} \in \mathcal{G}$; thus in particular each element of $\mathcal{F}$ is $\sigma(\mathcal{G})$-measurable. This implies that $\mathcal{A} \subseteq \sigma(\mathcal{G})$ holds. $\dashv$

The relationship between $\mathcal{B}$ and $\mathcal{G}$ is a bit more difficult to establish.

Lemma 4.8.16 $\mathcal{G} \subseteq \mathcal{B}$.

Proof We have to show that $\mu^*(G) + \mu^*(X \setminus G) = 1$ for all $G \in \mathcal{G}$. Fix $G \in \mathcal{G}$. We obtain from additivity that $\mu(G) + \mu(H) = \mu(G \cap H) + \mu(G \cup H) \geq \mu(X) = 1$ holds for any $H \in \mathcal{G}$ with $X \setminus G \subseteq H$, so that $\mu^*(G) + \mu^*(X \setminus G) \leq 1$ remains to be shown. The idea is to approximate χ_G for $G \in \mathcal{G}$ by a suitable sequence from $\mathcal{F}$ and to manipulate this sequence accordingly. Idea

Because $G \in \mathcal{G}$, there exists an increasing sequence $(f_n)_{n\in\mathbb{N}}$ of elements in $\mathcal{F}$ such that $\chi_G = \sup_{n\in\mathbb{N}} f_n$; consequently, $\chi_{X\setminus G} = \inf_{n\in\mathbb{N}} (1 - f_n)$. Now let $n \in \mathbb{N}$, and $0 < c \leq 1$; then $X \setminus G \subseteq U_{n,c} := \{x \in X \mid 1 - f_n(x) > c\}$ with $U_{n,c} \in \mathcal{G}$. Because $\chi_{U_{n,c}} \leq (1 - f_n)/c$, we obtain $\mu^*(X \setminus G) \leq L(1 - f_n)/c$; this inequality holds for all c and all $n \in \mathbb{N}$. Letting $c \to 1$ and $n \to \infty$, this yields $\mu^*(X \setminus G) \leq 1 - \mu^*(G)$.

Consequently, $\mu^*(G)+\mu^*(X\setminus G) = 1$ for all $G \in \mathcal{G}$, which establishes the claim. ⊣

This yields the desired relationship of $\mathcal{A}$, the σ-algebra generated by the functions in $\mathcal{F}$, and $\mathcal{B}$, the σ-algebra obtained from the extension process.

Corollary 4.8.17 *$\mathcal{A} \subseteq \mathcal{B}$, and each element of $\mathcal{L}^+$ is $\mathcal{B}$-measurable.*

Proof We have seen that $\mathcal{A} = \sigma(\mathcal{G})$ and that $\mathcal{G} \subseteq \mathcal{B}$, so the first assertion follows from Proposition 4.8.14. The second assertion is immediate from the first one. ⊣

Because μ is countably additive, hence a probability measure on $\mathcal{B}$, and because each element of $\mathcal{F}$ is $\mathcal{B}$-measurable, the integral $\int_X f\ d\mu$ is defined, and we are done.

Theorem 4.8.18 *Let $\mathcal{F}$ be a vector lattice of functions $X \to \mathbb{R}$ with $1 \in \mathcal{F}$, $L : \mathcal{F} \to \mathbb{R}$ be a linear and monotone functional on $\mathcal{F}$ such that $L(1) = 1$, and $L(f_n) \to 0$, whenever $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}$ decreases to 0. Then there exists a unique probability measure μ on the σ-algebra $\mathcal{A}$ generated by $\mathcal{F}$ such that*

$$L(f) = \int_X f\ d\mu$$

holds for all $f \in \mathcal{F}$.

Proof Let $\mathcal{G}$ and $\mathcal{B}$ be constructed as above.

Existence: Because $\mathcal{A} \subseteq \mathcal{B}$, we may restrict μ to $\mathcal{A}$, obtaining a probability measure. Fix $f \in \mathcal{F}$, then f is $\mathcal{B}$-measurable, and hence $\int_X f\ d\mu$ is defined. Assume first that $0 \leq f \leq 1$; hence $f \in \mathcal{L}^+$. We can write $f = \lim_{n\to\infty} f_n$ with step functions f_n, the contributing sets being members of $\mathcal{G}$. Hence $L(f_n) = \int_X f_n\ d\mu$, since $L(\chi_G) = \mu(G)$ by construction. Consequently, we obtain from Lemma 4.8.13 and Lebesgue's Dominated Convergence

Theorem 4.8.6

$$L(f) = L(\lim_{n\to\infty} f_n) = \lim_{n\to\infty} L(f_n) = \lim_{n\to\infty} \int_X f_n \, d\mu$$
$$= \int_X \lim_{n\to\infty} f_n \, d\mu = \int_X f \, d\mu.$$

This implies the assertion also for bounded $f \in \mathcal{F}$ with $f \geq 0$. If $0 \leq f$ is unbounded, write $f = \sup_{n\in\mathbb{N}}(f \wedge n)$ and apply Levi's Theorem 4.8.2. In the general case, decompose $f = f^+ - f^-$ with $f^+ := f \vee 0$ and $f^- := (-f) \vee 0$, and apply the foregoing.

Uniqueness: Assume that there exists a probability measure ν on $\mathcal{A}$ with $L(f) = \int_X f \, d\nu$ for all $f \in \mathcal{F}$; then the construction shows that $\mu(G) = L(\chi_G) = \nu(G)$ for all $G \in \mathcal{G}$. Since $\mathcal{G}$ is closed under finite intersections and since $\mathcal{A} = \sigma(\mathcal{G})$, we conclude that $\nu(A) = \mu(A)$ for all $A \in \mathcal{A}$.

This establishes the claim. ⊣

We obtain as a consequence the famous *Riesz Representation Theorem*, which we state and formulate for the metric case. Recall from Sect. 3.6.3 that $\mathcal{C}(X)$ is the linear space of all bounded continuous functions $X \to \mathbb{R}$ on a topological X. We state the result first for metric spaces and for bounded continuous functions, specializing the result subsequently to the compact metric case.

The reason for not formulating the Riesz Representation Theorem immediately for general topological spaces is that Theorem 4.8.18 works with the σ-algebra generated—in this case—by $\mathcal{C}(X)$; this is in general the σ-algebra of the Baire sets, which in turn may be properly contained in the Borel sets. Thus one obtains in the general case a Baire measure which then would have to be extended uniquely to a Borel measure. This is discussed in detail in [Bog07, Sect. 7.3].

$\mathcal{B}a(X)$ vs. $\mathcal{B}(X)$

Corollary 4.8.19 *Let X be a metric space, and let $L : \mathcal{C}(X) \to \mathbb{R}$ be a positive linear function with $\lim_{n\to\infty} L(f_n) = 0$ for each sequence $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{C}(X)$ which decreases monotonically to 0. Then there exists a unique finite Borel measure μ such that*

$$L(f) = \int_X f \, d\mu$$

holds for all $f \in \mathcal{C}(X)$.

Proof It is clear that $\mathcal{C}(X)$ is a vector lattice with $1 \in \mathcal{C}(X)$. We may and do assume that $L(1) = 1$. The result follows immediately from Theorem 4.8.18 now. $\dashv$

If we take a compact metric space, then each continuous map $X \to \mathbb{R}$ is bounded. We show that the assumption on L's continuity follows from compactness (the latter is usually referred to as *Dini's Theorem*; see Proposition 3.6.41).

Riesz Representation Theorem

Theorem 4.8.20 *Let X be a compact metric space. Given a positive linear functional $L : \mathcal{C}(X) \to \mathbb{R}$, there exists a unique finite Borel measure μ such that*

$$L(f) = \int_X f \, d\mu$$

holds for all $f \in \mathcal{C}(X)$.

Proof obligation

Proof It is clear that $\mathcal{C}(X)$ is a vector lattice which contains 1. Again, we assume that $L(1) = 1$ holds. In order to apply Theorem 4.8.18, we have to show that $\lim_{n\to\infty} L(f_n) = 0$, whenever $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{C}(X)$ decreases monotonically to 0. But since X is compact, we claim that $\sup_{x\in X} f_n(x) \to 0$, as $n \to \infty$.

This is so because $\{x \in X \mid f_n \geq c\}$ is a family of closed sets with empty intersection for any $c > 0$, so we find by compactness a finite subfamily with empty intersection. Hence the assumption that $\sup_{x\in X} f_n(x) \geq c > 0$ for all $n \in \mathbb{N}$ would lead to a contradiction (note that this is a variation of the argument in the proof of Proposition 3.6.41). Thus the assertion follows from Theorem 4.8.18. $\dashv$

Because $f \mapsto \int_X f \, d\mu$ defines for each Borel measure μ a positive linear functional on $\mathcal{C}(X)$ and because a measure on a metric space is uniquely determined by its integral on the bounded continuous functions, we obtain

Corollary 4.8.21 *For a compact metric space X, there is a bijection between positive linear functionals on $\mathcal{C}(X)$ and finite Borel measures.* $\dashv$

A typical scenario for the application of the Riesz Theorem runs like this: One starts with a probability measure on a metric space X. This space can be embedded into a compact metric space X'; one knows that the integral on the bounded continuous functions on X extends to a positive linear map on the continuous functions on X'. Then the Riesz

Representation Theorem kicks in and gives a probability measure on X'. We will see a situation like this when investigating the weak topology on the space of all finite measures on a Polish space in Sect. 4.10.

4.9 Product Measures

As a first application of integration, we show that the product of two finite measures yields a measure again. This will lead to Fubini's Theorem on product integration, which evaluates a product integrable function on a product along its vertical or its horizontal cuts (in this sense it may be compared to a line sweeping algorithm—you traverse the Cartesian product, and in each instance you measure the cut).

We apply this to infinite products, first with a countable index set, then for an arbitrary one. Infinite products are a special case of projective systems, which may be described as sequences of probabilities which are related through projections. We show that such a projective system has a projective limit, i.e., a measure on the set of all sequences such that the elements of the projective system proper are obtained through projections. This construction is, however, only feasible in a Polish space, since here a compactness argument is available which ascertains that the measure we are looking for is σ-additive.

A small step leads to projective limits for stochastic relations. We demonstrate an application for projective limits through the interpretation for the logic `CSL`. An interpretation for game logic, i.e., a modal logic, the modalities of which are given by games, is discussed as well, since now all tools for this quest are provided.

Fix for the time being two finite measure spaces $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$. The Cartesian product $X \times Y$ is endowed with the product σ-algebra $\mathcal{A} \otimes \mathcal{B}$ which is the smallest σ-algebra containing all measurable rectangles $A \times B$ with $A \in \mathcal{A}$ and $B \in \mathcal{B}$; see Sect. 4.1.1.

Take $Q \in \mathcal{A} \otimes \mathcal{B}$; then we know from Lemma 4.1.8 that Q_x and Q^y are measurable sets for $x \in X$ and $y \in Y$. We need measurability urgently here because otherwise functions like $x \mapsto \nu(Q_x)$ and $y \mapsto \mu(Q^y)$ would not be defined. In fact, we can say more about these functions.

Lemma 4.9.1 *Let $Q \in \mathcal{A} \otimes \mathcal{B}$ be a measurable set; then both $\varphi(x) := \nu(Q_x)$ and $\psi(y) := \mu(Q^y)$ define bounded measurable functions with*

$$\int_X \nu(Q_x)\, d\mu(x) = \int_Y \mu(Q^y)\, d\nu(y).$$

Proof We use the same argument as in the proof of Lemma 4.1.8 for establishing that both φ and ψ are measurable functions, noting that $\nu((A \times B)_x) = \chi_A(x) \cdot \nu(B)$, similarly $\mu((A \times B)^y) = \chi_B(y) \cdot \mu(A)$. In the next step, the set of all $Q \in \mathcal{A} \otimes \mathcal{B}$ is shown to satisfy the assumptions of the π-λ-Theorem 1.6.30.

In the same way, the equality of the integrals is finally established, noting that

$$\int_X \nu((A \times B)_x)\, d\mu(x) = \mu(A) \cdot \nu(B) = \int_X \mu((A \times B)^y)\, d\nu(y).$$

$\dashv$

Thus it does not matter which cut to take—integrating with the other measure will yield in any case the same result. Visualize this in the Cartesian plane. You have a geometric figure $F \subseteq \mathbb{R}^2$, say, for simplicity, F is compact. For each $x \in \mathbb{R}$, F_x is a vertical line, probably broken, the length $\ell(x)$ of which you can determine. Then $\int_{\mathbb{R}} \ell(x)\, dx$ yields the area A of F. But you may obtain A also by measuring the—also probably broken—horizontal line F^y with length $r(y)$ and integrating $\int_{\mathbb{R}} r(y)\, dy$.

Lemma 4.9.1 yields without much ado.

Theorem 4.9.2 *Given the finite measure spaces $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$, there exists a unique finite measure $\mu \otimes \nu$ on $\mathcal{A} \otimes \mathcal{B}$ such that $(\mu \otimes \nu)(A \times B) = \mu(A) \cdot \nu(B)$ for $A \in \mathcal{A}$, $B \in \mathcal{B}$. Moreover,*

$$(\mu \otimes \nu)(Q) = \int_X \nu(Q_x)\, d\mu(x) = \int_Y \mu(Q^y)\, d\nu(y)$$

holds for all $Q \in \mathcal{A} \otimes \mathcal{B}$.

Proof 1. We establish the existence of $\mu \otimes \nu$ by an appeal to Lemma 4.9.1 and to the properties of the integral according to Proposition 4.8.4. Define

$$(\mu \otimes \nu)(Q) := \int_X \nu(Q_x)\, d\mu(x);$$

then this defines a finite measure on $\mathcal{A} \otimes \mathcal{B}$:

Monotonicity: Let $Q \subseteq Q'$, then $Q_x \subseteq Q'_x$ for all $x \in X$, and hence $\int_X \nu(Q_x)\, d\mu(x) \leq \int_X \nu(Q'_x)\, d\mu(x)$. Thus $\mu \otimes \nu$ is monotone.

Additivity: If Q and Q' are disjoint, then $Q_x \cap Q'_x = (Q \cap Q')_x = \emptyset$ for all $x \in X$. Thus $\mu \otimes \nu$ is additive.

σ-Additivity: Let $(Q_n)_{n\in\mathbb{N}}$ be a sequence of disjoint measurable sets; then $(Q_{n,x})_{n\in\mathbb{N}}$ is disjoint for all $x \in X$, and

$$\begin{aligned}\int_X \nu\Big(\bigcup_{n\in\mathbb{N}} Q_{n,x}\Big)\, d\mu(x) &= \int_X \sum_{n\in\mathbb{N}} \nu(Q_{n,x})\, d\mu(x)\\ &= \sum_{n\in\mathbb{N}} \int_X \nu(Q_{n,x})\, d\mu(x)\end{aligned}$$

by Corollary 4.8.7. Thus $\mu \otimes \nu$ is σ-additive.

2. It remains to establish uniqueness. Here we repeat essentially the argumentation from Lemma 1.6.31 on page 86. Suppose that ρ is a finite measure on $\mathcal{A} \otimes \mathcal{B}$ with $\rho(A \times B) = \mu(A) \cdot \nu(B)$ for all $A \in \mathcal{A}$ and all $B \in \mathcal{B}$. Then

$$\mathcal{G} := \{Q \in \mathcal{A} \otimes \mathcal{B} \mid \rho(Q) = (\mu \otimes \nu)(Q)\}$$

contains the generator $\{A \times B \mid A \in \mathcal{A}, B \in \mathcal{B}\}$ of $\mathcal{A} \otimes \mathcal{B}$, which is closed under finite intersections. Because both ρ and $\mu \otimes \nu$ are measures, $\mathcal{G}$ is closed under countable disjoint unions, and because both contenders are finite, $\mathcal{G}$ is also closed under complementation. The π-λ-Theorem 1.6.30 shows that $\mathcal{G} = \mathcal{A} \otimes \mathcal{B}$. Thus $\mu \otimes \nu$ is uniquely determined. $\dashv$

Theorem 4.9.2 holds also for σ-finite measures. In fact, assume that the contributing measure spaces are σ-finite, and let $(X_n)_{n\in\mathbb{N}}$ resp. $(Y_n)_{n\in\mathbb{N}}$ be increasing sequences in $\mathcal{A}$ resp. $\mathcal{B}$ such that $\mu(X_n) < \infty$ and $\nu(Y_n) < \infty$ for all $n \in \mathbb{N}$, and $\bigcup_{n\in\mathbb{N}} X_n = X$ and $\bigcup_{n\in\mathbb{N}} Y_n = Y$. Localize μ and ν to X_n resp. Y_n by defining $\mu_n(A) := \mu(A \cap X_n)$, similarly, $\nu_n(B) := \nu(B \cap Y_n)$; since these measures are finite, we can extend them uniquely to a measure $\mu_n \otimes \nu_n$ on $\mathcal{A} \otimes \mathcal{B}$. Since $\bigcup_{n\in\mathbb{N}} X_n \times Y_n = X \times Y$ with the increasing sequence $(X_n \times Y_n)_{n\in\mathbb{N}}$, we set

$$(\mu \otimes \nu)(Q) := \sup_{n\in\mathbb{N}} (\mu_n \otimes \nu_n)(Q).$$

Then $\mu \otimes \nu$ is a σ-finite measure on $\mathcal{A} \otimes \mathcal{B}$. Now assume that we have another σ-finite measure ρ on $\mathcal{A} \otimes \mathcal{B}$ with $\rho(A \times B) = \mu(A) \cdot \nu(B)$

for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$. Define $\rho_n(Q) := \rho(Q \cap (X_n \times Y_n))$; hence $\rho_n = \mu_n \otimes \nu_n$ by uniqueness of the extension to μ_n and ν_n, so that we obtain

$$\rho(Q) = \sup_{n \in \mathbb{N}} \rho_n(Q) = \sup_{n \in \mathbb{N}} (\mu_n \otimes \nu_n)(Q) = (\mu \otimes \nu)(Q)$$

for all $Q \in \mathcal{A} \otimes \mathcal{B}$. Thus we have shown

Corollary 4.9.3 *Given two σ-finite measure spaces $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$, there exists a unique σ-finite measure $\mu \otimes \nu$ on $\mathcal{A} \otimes \mathcal{B}$ such that $(\mu \otimes \nu)(A \times B) = \mu(A) \cdot \nu(B)$. We have*

$$(\mu \otimes \nu)(Q) = \int_X \nu(Q_x)\, d\mu(x) = \int_Y \mu(Q^y)\, d\nu(y).$$

$\dashv$

The construction of the product measure has been done here through integration of cuts. An alternative would have been the canonical approach. It would have investigated the map $\langle A, B \rangle \mapsto \mu(A) \cdot \nu(B)$ on the set of all rectangles, and then put the extension machinery developed through the Carathéodory approach into action. It is a matter of taste which approach to prefer.

The following example displays a slight generalization (a finite measure is but a constant transition kernel).

Example 4.9.4 Let $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ be a transition kernel (see Definition 4.1.9) such that the map $x \mapsto K(x)(Y)$ is integrable with respect to the finite measure μ. Then

$$(\mu \otimes K)(Q) := \int_X K(x)(Q_x)\, d\mu(x)$$

defines a finite measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$. The π-λ-Theorem 1.6.30 tells us that this measure is uniquely determined by the condition $(\mu \otimes K)(A \times B) = \int_A K(x)(B)\, d\mu(x)$ for $A \in \mathcal{A}$, $B \in \mathcal{B}$.

Interpret in a probabilistic setting $K(x)(B)$ as the probability that an input $x \in X$ yields an output in $B \in \mathcal{B}$, and assume that μ gives the initial probability with which the system starts; then $\mu \otimes K$ gives the probability of all pairings, i.e., $(\mu \otimes K)(Q)$ is the probability that a pair $\langle x, y \rangle$ consisting of an input value $x \in X$ and an output value $y \in Y$ will be a member of $Q \in \mathcal{A} \otimes \mathcal{B}$. ✌

This may be further extended, replacing the measure on K's domain by a transition kernel as well.

Example 4.9.5 Consider the scenario of Example 4.9.4 again, but take a third measurable space $(Z, \mathcal{C})$ with a transition kernel $L : (Z, \mathcal{C}) \rightsquigarrow (X, \mathcal{A})$ into account; assume furthermore that $x \mapsto K(x)(Y)$ is integrable for each $L(z)$. Then $L(z) \otimes K$ defines a finite measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ for each $z \in Z$ according to Example 4.9.4. We claim that this defines a transition kernel $(Z, \mathcal{C}) \rightsquigarrow (X \times Y, \mathcal{A} \otimes \mathcal{B})$. For this to be true, we have to show that $z \mapsto \int_X K(x)(Q_x)\, dL(z)(x)$ is measurable for each $Q \in \mathcal{A} \otimes \mathcal{B}$. This is a typical application for the principle of good sets through the π-λ-Theorem.

In fact, consider

$$\mathcal{Q} := \{Q \in \mathcal{A} \otimes \mathcal{B} \mid \text{ the assertion is true for } Q\}.$$

Then $\mathcal{Q}$ is closed under complementation. It is also closed under countable disjoint unions by Corollary 4.8.7. If $Q = A \times B$ is a measurable rectangle, we have $\int_X K(x)(Q_x)\, dL(z)(x) = \int_A K(x)(B)\, dL(z)(x)$. Then Exercise 4.14 shows that this is a measurable function $Z \to \mathbb{R}$. Thus $\mathcal{Q}$ contains all measurable rectangles, so $\mathcal{Q} = \mathcal{A} \otimes \mathcal{B}$ by the π-λ-Theorem 1.6.30. This establishes measurability of $z \mapsto \int_X K(x)(Q_x)\, dL(z)(x)$ and shows that it defines a transition kernel. ✌

As a slight modification, the next example shows the composition of transition kernels, usually called *convolution*.

Example 4.9.6 Let $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ and $L : (Y, \mathcal{B}) \rightsquigarrow (Z, \mathcal{C})$ be transition kernels, and assume that the map $y \mapsto L(y)(Z)$ is integrable with respect to measures $K(x)$ for an arbitrary $x \in X$. Define for $x \in X$ and $C \in \mathcal{C}$

$$(L * K)(x)(C) := \int_X L(y)(C)\, dK(x)(y).$$

Then $L * K : (X, \mathcal{A}) \rightsquigarrow (Z, \mathcal{C})$ is a transition kernel. In fact, $(L * K)(x)$ is for fixed $x \in X$ a finite measure on $\mathcal{C}$ according to Corollary 4.8.7. From Exercise 4.14, we infer that $x \mapsto \int_X L(y)(C)\, dK(x)(y)$ is a measurable function, since $y \mapsto L(y)(C)$ is measurable for all $C \in \mathcal{C}$.

Because transition kernels are the Kleisli morphisms for the endofunctor $\mathbb{M}$ on the category of measurable spaces (Example 2.4.8), it is not difficult to see that this defines Kleisli composition; in particular it follows that this composition is associative. ✌

Example 4.9.7 Let $f \in \mathcal{F}_+(X, \mathcal{A})$; then we know that "the area under the graph," viz.,

$$C_{\leq}(f) := \{\langle x, r\rangle \mid x \in X, 0 \leq r \leq f(x)\}$$

is a member of $\mathcal{A} \otimes \mathcal{B}(\mathbb{R})$. This was shown in Corollary 4.2.5. Then Corollary 4.9.3 tells us that

$$(\mu \otimes \lambda)(C_{\leq}(f)) = \int_X \lambda\big((C_{\leq}(f))_x\big)\, d\mu(x),$$

where λ is the Lebesgue measure on $\mathcal{B}(\mathbb{R})$. Because

$$\lambda\big((C_{\leq}(f))_x\big) = \lambda(\{r \mid 0 \leq r \leq f(x)\}) = f(x),$$

we obtain

$$(\mu \otimes \lambda)(C_{\leq}(f)) = \int_X f\, d\mu.$$

On the other hand,

$$(\mu \otimes \lambda)(C_{\leq}(f)) = \int_{\mathbb{R}_+} \mu\big((C_{\leq}(f)_r\big)\, d\lambda(r),$$

and this gives the integration formula

$$\int_X f\, d\mu = \int_0^\infty \mu(\{x \in X \mid f(x) \geq r\})\, dr. \tag{4.11}$$

In this way, the integral of a nonnegative function may in fact be interpreted as measuring the area under its graph. ✌

With a similar technique, we show that $\mu \mapsto \int_0^1 \mu(B^r)\, dr$ defines a measurable map for $B \in \mathcal{A} \otimes \mathcal{B}([0,1])$; this may be perceived as the average evaluation of a set $B \subseteq X \times [0,1]$. Moreover the set of all these evaluations is a measurable subset of $\mathbb{S}(X, \mathcal{A}) \times [0,1]$.

Lemma 4.9.8 *Let $(X, \mathcal{A})$ be a measurable space; then*

$$\Lambda_B(\mu) := \int_0^1 \mu(B^r)\, dr$$

defines a $\wp(X, \mathcal{A})$-$\mathcal{B}(\mathbb{R})$-measurable map on $\mathbb{S}(X, \mathcal{A})$, whenever $B \in \mathcal{A} \otimes \mathcal{B}(\mathbb{R})$.

Proof This is established through the principle of good sets: Put

$$\mathcal{G} := \{B \in \mathcal{A} \otimes \mathcal{B}([0,1]) \mid \Lambda_B \text{ is } \wp(X,\mathcal{A}) - \mathcal{B}([0,1]) \text{ measurable}\}.$$

Then $\mathcal{G}$ contains all sets $A \times D$ with $A \in \mathcal{A}$, $B \in \mathcal{B}([0,1])$. This is so because

$$\mu\big((A \times D)^r\big) = \begin{cases} \mu(A), & \text{if } r \in D, \\ 0, & \text{otherwise.} \end{cases}$$

Hence $\Lambda_{A\times D}(\mu) = \int_D \mu(A)\, d\lambda = \mu(A) \cdot \lambda(D)$. This is certainly a measurable map on $\mathbb{S}(X,\mathcal{A})$. Thus $\mathcal{G}$ contains the generator $\{A \times D \mid A \in \mathcal{A}, D \in \mathcal{B}([0,1])\}$ of $\mathcal{A} \otimes \mathcal{B}([0,1])$, which in turn is closed under finite intersections. It is clear that $\mathcal{G}$ is closed under complementation and under countable disjoint unions, the latter by Corollary 4.8.7. Hence we obtain from the π-λ-Theorem 1.6.30 that $\mathcal{G} = \mathcal{A} \otimes \mathcal{B}([0,1])$.

We obtain from Corollary 4.2.5 that the set

$$\{\langle \mu, q\rangle \in \mathbb{S}(X,\mathcal{A}) \times [0,1] \mid \int_0^1 \mu(B^r)\, dr \bowtie q\}$$

is a member of $\wp(X,\mathcal{A}) \otimes \mathcal{B}([0,1])$, whenever $B \subseteq X \times [0,1]$ is a member of $\mathcal{A} \otimes \mathcal{B}([0,1])$. $\dashv$

This provides us with a measurable subset of $\mathbb{S}(X,\mathcal{A}) \times [0,1]$; it will be useful later on, when discussing the semantics of game logic in Sect. 4.9.4.

Corollary 4.9.9 *Let $B \in \mathcal{A} \otimes \mathcal{B}([0,1])$; then*

$$\{\langle \mu, q\rangle \in \mathbb{S}(X,\mathcal{A}) \times [0,1] \mid \int_0^1 \mu(B^r)\, dr \bowtie q\}$$

is a measurable subset of $\mathbb{S}(X,\mathcal{A}) \times [0,1]$.

Proof 1. We claim that the map $\mu \mapsto \int_0^1 \mu(B^r)\, dr$ is $\wp(X,\mathcal{A})$-$\mathcal{B}([0,1])$-measurable. The claim is established in a stepwise manner: First we establish that B^r is a measurable set for each r, then we show that $r \mapsto \mu(B^r)$ gives a measurable function for every μ, and then we are nearly done. The assertion follows from Corollary 4.2.5. Plan

2. Lemma 4.1.8 tells us that $B^r \in \mathcal{A}$ for all $B \in \mathcal{A} \otimes \mathcal{B}([0,1])$, so that $\mu(B^r)$ is defined for each $r \in [0,1]$. The map $r \mapsto \mu(B^r)$ defines a measurable map $[0,1] \to \mathbb{R}_+$; this is shown in Lemma 4.9.1; hence

$\int_0^1 \mu(B^r)\, dr$ is defined. We finally show that the map in question is measurable by applying the principle of good sets. Let

$$\mathcal{G} := \{B \in \mathcal{A} \otimes \mathcal{B}([0,1]) \mid \text{the assertion is true for } B\}.$$

Then $B = A \times C \in \mathcal{G}$ for $A \in \mathcal{A}$ and $C \in \mathcal{B}([0,1])$, since $\int_0^1 \mu(B^r)\, dr = \mu(A) \cdot \lambda(C)$, which is defined by a $\wp(X,\mathcal{A})$-$\mathcal{B}([0,1])$-measurable function, λ being the Lebesgue measure on $[0,1]$. $\mathcal{G}$ is closed under complementation and under disjoint countable unions by Corollary 4.8.7. Hence we obtain $\mathcal{G} = \mathcal{A} \otimes \mathcal{B}([0,1])$ from the π-λ-Theorem 1.6.30, because the set of all measurable rectangles generates this σ-algebra and is closed under finite intersection. $\dashv$

4.9.1 Fubini's Theorem

In order to discuss integration with respect to a product measure, we introduce the cuts of a function $f : X \times Y \to \mathbb{R}$, defining $f_x := \lambda y. f(x,y)$ and $f^y := \lambda x. f(x,y)$. Thus we have $f(x,y) = f_x(y) = f^y(x)$, the first equality resembling currying.

$\widetilde{\mathbb{R}}$, Admit $+\infty, -\infty$

For the discussion to follow, we will admit also the values $\{-\infty, +\infty\}$ as function values. So define $\widetilde{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$, and let $B \subseteq \widetilde{\mathbb{R}}$ be a Borel set iff $B \cap \mathbb{R} \in \mathcal{B}(\mathbb{R})$. Measurability of functions extends accordingly: If $f : X \to \widetilde{\mathbb{R}}$ is measurable, then in particular $\{x \in X \mid f(x) \in \mathbb{R}\} \in \mathcal{A}$, and the set of values on which f takes the values $+\infty$ or $-\infty$ is a member of $\mathcal{A}$. Denote by $\widetilde{\mathcal{F}}(X,\mathcal{A})$ the set of measurable functions with values in $\widetilde{\mathbb{R}}$ and by $\widetilde{\mathcal{F}}_+(X,\mathcal{A})$ those which take nonnegative values. The integral $\int_X f\, d\mu$ and integrability are defined in the same way as above for $f \in \widetilde{\mathcal{F}}_+(X,\mathcal{A})$. Then it is clear that $f \in \widetilde{\mathcal{F}}_+(X,\mathcal{A})$ is integrable iff $f \cdot \chi_{\{x \in X \mid f(x) \in \mathbb{R}\}}$ is integrable and $\mu(\{x \in X \mid f(x) = \infty\}) = 0$.

$\widetilde{\mathcal{F}}(X,\mathcal{A})$

With this in mind, we tackle the integration of a measurable function $f : X \times Y \to \widetilde{\mathbb{R}}$ for the finite measure spaces $(X,\mathcal{A},\mu)$ and $(Y,\mathcal{B},\nu)$.

Proposition 4.9.10 *Let $f \in \widetilde{\mathcal{F}}_+(X \times Y, \mathcal{A} \otimes \mathcal{B})$; then*

1. *$\lambda x. \int_Y f_x\, d\nu$ and $\lambda y. \int_X f^y\, d\mu$ are measurable functions $X \to \widetilde{\mathbb{R}}$ resp. $Y \to \widetilde{\mathbb{R}}$.*

2. *We have*

$$\begin{aligned}\int_{X\times Y} f\ d\mu\otimes\nu &= \int_X \big(\int_Y f_x\ d\nu\big)\ d\mu(x)\\ &= \int_Y \big(\int_X f^y\ d\mu\big)\ d\nu(y).\end{aligned}$$

Proof 0. The proof is a bit longish, but, after all, there is a lot to show. It is an application of the machinery developed so far and shows that it is well oiled. We start off by establishing the claim for step functions, then we go on to apply Levi's Theorem for settling the general case. Line of attack

1. Let $f = \sum_{i=1}^{n} a_i \cdot \chi_{Q_i}$ be a step function with $a_i \geq 0$ and $Q_i \in \mathcal{A}\otimes\mathcal{B}$ for $i = 1, \ldots, n$. Then

$$\int_Y f_x\ d\nu = \sum_{i=1}^{n} a_i \cdot \nu(Q_{i,x}).$$

This is a measurable function $X \to \mathbb{R}$ by Lemma 4.9.1. We obtain

$$\begin{aligned}\int_{X\times Y} f\ d\mu\otimes\nu &= \sum_{i=1}^{n} a_i \cdot (\mu\otimes\nu)(Q_i)\\ &= \sum_{i=1}^{n} a_i \cdot \int_X \nu(Q_{i,x})\ d\mu(x)\\ &= \int_X \sum_{i=1}^{n} a_i \cdot \nu(Q_{i,x})\ d\mu(x)\\ &= \int_X \big(\int_Y f_x\ d\nu\big)\ d\mu(x).\end{aligned}$$

Interchanging the rôles of μ and ν, we obtain the representation of $\lambda y.\int_{X\times Y} f\ d\mu\otimes\nu$ in terms of $\lambda y.\int_X f^y\ d\mu$ and ν. Thus the assertion is true for step functions.

2. In the general case, we know that we can find an increasing sequence $(f_n)_{n\in\mathbb{N}}$ of step functions with $f = \sup_{n\in\mathbb{N}} f_n$. Given $x \in X$, we infer that $f_x = \sup_{n\in\mathbb{N}} f_{x,n}$, so that

$$\int_Y f_x\ d\nu = \sup_{n\in\mathbb{N}} \int_X f_{n,x}\ d\nu$$

by Levi's Theorem 4.8.2. This implies measurability. Applying Levi's Theorem again to the results from part 1, we have

$$\begin{aligned}\int_{X\times Y} f\ d\mu\otimes\nu &= \sup_{n\in\mathbb{N}}\int_{X\times Y} f_n\ d\mu\otimes\nu\\ &= \sup_{n\in\mathbb{N}}\int_X \bigl(\int_Y f_{n,x}\ d\nu\bigr)\ d\mu(x)\\ &= \int_X \bigl(\sup_{n\in\mathbb{N}}\int_Y f_{n,x}\ d\nu\bigr)\ d\mu(x)\\ &= \int_{X\times Y}\bigl(\int_Y f_x\ d\nu\bigr)\ d\mu(x).\end{aligned}$$

Again, interchanging rôles yields the symmetric equality. ⊣

This yields as an immediate consequence that the cuts of a product integrable function are integrable almost everywhere, to be specific

Corollary 4.9.11 *Let $f : X \times Y \to \mathbb{R}$ be $\mu\otimes\nu$ integrable, and put*

$$\begin{aligned}A &:= \{x\in X \mid f_x \text{ is not } \nu\text{-integrable}\},\\ B &:= \{y\in Y \mid f^y \text{ is not } \mu\text{-integrable}\}.\end{aligned}$$

Then $A\in\mathcal{A}$, $B\in\mathcal{B}$, and $\mu(A)=\nu(B)=0$.

Proof Because $A=\{x\in X \mid \int_Y |f_x|\ d\nu=\infty\}$, we see that $A\in\mathcal{A}$. By the additivity of the integral, we have

$$\begin{aligned}\int_{X\times Y}|f|\ d\mu\otimes\nu = &\int_{X\setminus A}\bigl(\int_Y |f_x|\ d\nu\bigr)\ d\mu(x)\\ &+\int_A\bigl(\int_Y |f_x|\ d\nu\bigr)\ d\mu(x)<\infty;\end{aligned}$$

hence $\mu(A)=0$. B is treated in the same way. ⊣

It is helpful to extend our integral in a minor way. Assume that $\int_X |f|\ d\mu<\infty$ for $f : X\to\widetilde{\mathbb{R}}$ measurable and that $\mu(A)=0$ with $A:=\{x\in X \mid |f(x)|=\infty\}$. Change f on A to a finite value, obtaining a measurable function $f_* : X\to\mathbb{R}$, and define

$$\int_X f\ d\mu := \int_X f_*\ d\mu.$$

Thus $f\mapsto\int_X f\ d\mu$ *does not notice* this change on a set of measure zero. In this way, we assume always that an integrable function takes finite values, even if we have to convince it to do so on a set of measure zero.

With this in mind, we obtain

Corollary 4.9.12 *Let $f : X \times Y \to \mathbb{R}$ be integrable, then $\lambda x. \int_Y f_x \, d\nu$ and $\lambda y. \int_X f^y \, d\nu$ are integrable with respect to μ resp. ν, and*

$$\int_{X \times Y} f \, d\mu \otimes \nu = \int_X \Big(\int_Y f_x \, d\nu\Big) \, d\mu(x) = \int_Y \Big(\int_X f^y \, d\mu\Big) \, d\nu(y).$$

Proof After the modification on a set of μ-measure zero, we know that

$$\Big|\int_X f_x \, d\nu\Big| \leq \int_Y |f_x| \, d\nu < \infty$$

for all $x \in X$, so that $\lambda x. \int_Y f_x \, d\nu$ is integrable with respect to μ; similarly, $\lambda y. \int_X f^y \, d\nu$ is integrable with respect to ν for all $y \in Y$. We obtain from Proposition 4.9.10 and the linearity of the integral

$$\begin{aligned}
\int_{X \times Y} f \, d\mu \otimes \nu &= \int_{X \times Y} f^+ \, d\mu \otimes \nu - \int_{X \times Y} f^- \, d\mu \otimes \nu \\
&= \int_X \Big(\int_Y f_x^+ \, d\nu\Big) \, d\mu(x) - \int_X \Big(\int_Y f_x^- \, d\nu\Big) \, d\mu(x) \\
&= \int_X \Big(\int_Y f_x^+ \, d\nu - \int_Y f_x^- \, d\nu\Big) \, d\mu(x) \\
&= \int_X \Big(\int_Y f_x \, d\nu\Big) \, d\mu(x).
\end{aligned}$$

The second equation is treated in exactly the same way. $\dashv$

Now we know how to treat a function which is integrable, but we do not yet have a criterion for integrability. The elegance of Fubini's Theorem shines through the observation that the existence of the iterated integrals yields integrability for the product integral. To be specific

Theorem 4.9.13 *Let $f : X \times Y \to \mathbb{R}$ be measurable. Then these statements are equivalent:*

1. $\int_{X \times Y} |f| \, d\mu \otimes \nu < \infty$.
2. $\int_X \big(\int_Y |f_x| \, d\nu\big) \, d\mu(x) < \infty$.
3. $\int_Y \big(\int_X |f^y| \, d\mu\big) \, d\nu(y) < \infty$.

Fubini's Theorem

Under one of these conditions, f is $\mu \otimes \nu$-integrable, and

$$\int_{X \times Y} f \, d\mu \otimes \nu = \int_X \Big(\int_Y f_x \, d\nu \Big) \, d\mu(x) = \int_Y \Big(\int_X f^y \, d\mu \Big) \, d\nu(y). \tag{4.12}$$

Proof We discuss only 1 $\Rightarrow$ 2; the other implications are proved similarly. From Proposition 4.9.10 it is inferred that $|f|$ is integrable, so 2. holds by Corollary 4.9.12, from which we also obtain representation (4.12). $\dashv$

Product integration and Fubini's Theorem are an essential extension to our tool kit.

4.9.2 Infinite Products and Projective Limits

We will discuss as a first application of product integration now the existence of infinite products for probability measures. Then we will go on to establish the existence of projective limits. These limits will be useful for interpreting a logic which works with continuous time.

Corollary 4.9.3 extends to a finite number of σ-finite measure spaces in a natural way. Let $(X_i, \mathcal{A}_i, \mu_i)$ be σ-finite measure spaces for $1 \leq i \leq n$, the uniquely determined product measure on $\mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n$ is denoted by $\mu_1 \otimes \ldots \otimes \mu_n$, and we infer from Corollary 4.9.3 that we may write

$$\begin{aligned}
(\mu_1 \otimes \ldots \otimes \mu_n)(Q) \\
&= \int_{X_2 \times \ldots \times X_n} \mu_1(Q^{x_2, \ldots, x_n}) \, d(\mu_2 \otimes \ldots \otimes \mu_n)(x_2, \ldots, x_n), \\
&= \int_{X_1 \times \ldots \times X_{n-1}} \mu_n(Q_{x_1, \ldots, x_{n-1}}) \, d(\mu_1 \otimes \ldots \otimes \mu_{n-1})(x_1, \ldots, x_{n-1})
\end{aligned}$$

whenever $Q \in \mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n$. This is fairly straightforward.

Infinite Products

We will have a closer look now at infinite products, where we restrict ourselves to probability measures, and here we consider the countable case first. So let $(X_n, \mathcal{A}_n, \varpi_n)$ be a measure space with a probability measure ϖ_n on $\mathcal{A}_n$ for $n \in \mathbb{N}$. We want to construct the infinite product of this sequence.

Let us fix some notations first and then describe the goal in greater detail. Put

$$\begin{aligned} X^{(n)} &:= \prod_{k \geq n} X_k, \\ \mathcal{A}^{(n)} &:= \{A \times X^{(n+\ell)} \mid A \in \mathcal{A}_n \otimes \ldots \otimes \mathcal{A}_{n+\ell-1} \text{ for some } \ell \in \mathbb{N}\}. \end{aligned}$$

The elements of $\mathcal{A}^{(n)}$ are the *cylinder sets* for $X^{(n)}$. Thus $X^{(1)} = \prod_{n \in \mathbb{N}} X_n$, and $\bigotimes_{n \in \mathbb{N}} \mathcal{A}_n = \sigma(\mathcal{A}^{(1)})$. Given $A \in \mathcal{A}^{(n)}$, we can write A as $A = C \times X^{n+\ell}$ with $C \in \mathcal{A}_n \otimes \ldots \otimes \mathcal{A}_{n+\ell-1}$. So if we set Cylinder sets

$$\varpi^{(n)}(A) := \varpi_n \otimes \ldots \otimes \varpi_{n+\ell-1}(C),$$

then $\varpi^{(n)}$ is well defined on $\mathcal{A}^{(n)}$, and it is readily verified that it is monotone and additive with $\varpi^{(n)}(\emptyset) = 0$ and $\varpi^{(n)}(X^{(n)}) = 1$. Moreover, we infer from Theorem 4.9.2 that

$$\varpi^{(n)}(C) = \int_{X_{n+1} \times \ldots \times X_{n+m}} \varpi_n(C^{x_{n+1},\ldots,x_{n+m}})\, d(\varpi_{n+1} \otimes \ldots \otimes \varpi_{n+m}) \\ (x_{n+1}, \ldots, x_{n+m})$$

for all $C \in \mathcal{A}^{(n)}$.

The goal is to show that there exists a unique probability measure ϖ on $\bigotimes_{n \in \mathbb{N}}(X_n, \mathcal{A}_n)$ such that $\varpi\left(A \times X^{(n+1)}\right) = (\varpi_1 \otimes \ldots \otimes \varpi_n)(A)$ whenever $A \in \mathcal{A}_n \otimes \mathcal{A}^{(n+1)}$. If we can show that $\varpi^{(1)}$ is σ-additive on $\mathcal{A}^{(1)}$, then we can extend $\varpi^{(1)}$ to the desired σ-algebra by Theorem 1.6.29. For this it is sufficient to show that $\inf_{n \in \mathbb{N}} \varpi^{(1)}(A_n) > \epsilon > 0$ implies $\bigcap_{n \in \mathbb{N}} A_n \neq \emptyset$ for any decreasing sequence $(A_n)_{n \in \mathbb{N}}$ in $\mathcal{A}^{(1)}$. Goal $\overset{!}{\Longleftarrow}$

The basic idea is to construct a sequence $(x_n)_{n \in \mathbb{N}} \in \bigcap_{n \in \mathbb{N}} A_n$. We do this step by step: Basic idea

- First we determine an element $x_1 \in X_1$ such that we can expand the—admittedly very short—partial sequence x_1 to a sequence which is contained in all A_n; this means that we have to have $A_n^{x_1} \neq \emptyset$ for all $n \in \mathbb{N}$, because $A_n^{x_1}$ contains all possible continuations of x_1 into A_n. We conclude that these sets are nonempty, because their measure is strictly positive.

- If we have such an x_1, we start working on the second element of the sequence, so we have a look at some $x_2 \in X_2$ such that we can expand x_1, x_2 to a sequence which is contained in all A_n, so we have to have $A_n^{x_1,x_2} \neq \emptyset$ for all $n \in \mathbb{N}$. Again, we look for x_2 so that the measure of $A_n^{x_1,x_2}$ is strictly positive for each n.
- Continuing in this fashion, we obtain the desired sequence, which then has to be an element of $\bigcap_{n\in\mathbb{N}} A_n$ by construction.

This is the plan. Let us explore finding x_1. Put

$$E_1^{(n)} := \{x_1 \in X_1 \mid \varpi^{(2)}(A_n^{x_1}) > \epsilon/2\}.$$

Because

$$\varpi^{(1)}(A_n) = \int_{X_1} \varpi^{(2)}(A_n^{x_1})\, d\varpi_1(x_1),$$

we have

$$\begin{aligned}
0 < \epsilon < \varpi^{(1)}(A_n) &= \int_{E_1^{(n)}} \varpi^{(2)}(A_n^{x_1})\, d\varpi_1(x_1) \\
&\quad + \int_{X_1 \setminus E_1^{(n)}} \varpi^{(2)}(A_n^{x_1})\, d\varpi_1(x_1) \\
&\leq \varpi_1(E_1^{(n)}) + \epsilon/2 \cdot \varpi^{(1)}(X_1 \setminus E_1^{(n)}) \\
&\leq \varpi_1(E_1^{(n)}) + \epsilon/2.
\end{aligned}$$

Thus $\varpi_1(E_1^{(n)}) \geq \epsilon/2$ for all $n \in \mathbb{N}$. Since $A_1 \supseteq A_2 \supseteq \ldots$, we have also $E_1^{(1)} \supseteq E_1^{(2)} \supseteq \ldots$, so let $E_1 := \bigcap_{n\in\mathbb{N}} E_1^{(n)}$, then $E_1 \in \mathcal{A}_1$ with $\varpi_1(E_1) \geq \epsilon/2 > 0$. In particular, $E_1 \neq \emptyset$. Pick and fix $x_1 \in E_1$. Then $A_n^{x_1} \in \mathcal{A}^{(2)}$, and $\varpi^{(2)}(A_n^{x_1}) > \epsilon/2$ for all $n \in \mathbb{N}$.

Let us have a look at how to find the second element; this is but a small variation of the idea just presented. Put $E_2^{(n)} := \{x_2 \in X_2 \mid \varpi^{(3)}(A_n^{x_1,x_2}) > \epsilon/4\}$ for $n \in \mathbb{N}$. Because

$$\varpi^{(2)}(A_n^{x_1}) = \int_{X_2} \varpi^{(3)}(A_n^{x_1,x_2})\, d\varpi_2(x_2),$$

we obtain similarly $\varpi_2(E_2^{(n)}) \geq \epsilon/4$ for all $n \in \mathbb{N}$. Again, we have a decreasing sequence, and putting $E_2 := \bigcap_{n\in\mathbb{N}} E_2^{(n)}$, we have $\varpi_2(E_2)$

$\geq \epsilon/4$, so that $E_2 \neq \emptyset$. Pick $x_2 \in E_2$; then $A_n^{x_1,x_2} \in \mathcal{A}^{(3)}$ and $\varpi^{(3)}(A_n^{x_1,x_2}) > \epsilon/4$ for all $n \in \mathbb{N}$.

In this manner we determine inductively for each $k \in \mathbb{N}$ the finite sequence $\langle x_1, \ldots, x_k \rangle \in X_1 \times \ldots \times X_k$ such that $\varpi^{(k+1)}(A_n^{x_1,\ldots,x_k}) > \epsilon/2^k$ for all $n \in \mathbb{N}$. Consider now the sequence $(x_n)_{n\in\mathbb{N}}$. From the construction it is clear that $\langle x_1, x_2, \ldots, x_k, \ldots \rangle \in \bigcap_{n\in\mathbb{N}} A_n$. This shows that $\bigcap_{n\in\mathbb{N}} A_n \neq \emptyset$, and it implies by the argumentation above that $\varpi^{(1)}$ is σ-additive on the algebra $\mathcal{A}^{(1)}$.

Hence we have established

Theorem 4.9.14 *Let $(X_n, \mathcal{A}_n, \varpi_n)$ be probability spaces for all $n \in \mathbb{N}$. Then there exists a unique probability measure ϖ on $\bigotimes_{n\in\mathbb{N}}(X_n, \mathcal{A}_n)$ such that*

$$\varpi(A \times \prod_{k>n} X_k) = (\varpi_1 \otimes \ldots \otimes \varpi_n)(A)$$

for all $A \in \bigotimes_{i=1}^n \mathcal{A}_i$. $\dashv$

Define the projection $\pi_n^\infty : (x_n)_{n\in\mathbb{N}} \mapsto \langle x_1, \ldots, x_n \rangle$ from $\prod_{n\in\mathbb{N}} X_n$ to $\prod_{i=1}^n X_i$. In terms of image measures, the theorem states that there exists a unique probability measure ϖ on the infinite product such that $\mathbb{S}(\pi_n^\infty)(\varpi) = \varpi_1 \otimes \ldots \otimes \varpi_n$.

Non-countable Index Sets. Now let us have a look at the general case, in which the index set is not necessarily countable. Let $(X_i, \mathcal{A}_i, \mu_i)$ be a family of probability spaces for $i \in I$, put $X := \prod_{i\in I} X_i$ and $\mathcal{A} := \bigotimes_{i\in I} \mathcal{A}_i$. Given $J \subseteq I$, define $\pi_J : (x_i)_{i\in I} \mapsto (x_i)_{i\in J}$ as the projection $X \to \prod_{i\in J} X_i$. Put $\mathcal{A}_J := \pi_J^{-1}\left[\bigotimes_{j\in J} \mathcal{A}_j\right]$.

Although the index set I may be large, the measurable sets in $\mathcal{A}$ are always determined by a countable subset of the index set.

Lemma 4.9.15 *Given $A \in \mathcal{A}$, there exists a countable subset $J \subseteq I$ such that $\chi_A(x) = \chi_A(x')$, whenever $\pi_J(x) = \pi_J(x')$.*

Proof Let $\mathcal{G}$ be the set of all $A \in \mathcal{A}$ for which the assertion is true. Then $\mathcal{G}$ is a σ-algebra which contains $\pi_{\{i\}}^{-1}\left[\mathcal{A}_i\right]$ for every $i \in I$; hence $\mathcal{G} = \mathcal{A}$. $\dashv$

This yields as an immediate consequence.

Corollary 4.9.16 $\mathcal{A} = \bigcup\{\mathcal{A}_J \mid J \subseteq I$ *is countable*$\}$.

Proof It is enough to show that the set on the right-hand side is a σ-algebra. This follows easily from Lemma 4.9.15. $\dashv$

We obtain from this observation and from the previous result for the countable case that arbitrary products exist.

Theorem 4.9.17 *Let $(X_i, \mathcal{A}_i, \varpi_i)$ be a family of probability spaces for $i \in I$. Then there exists a unique probability measure ϖ on $\prod_{i \in I} (X_i, \mathcal{A}_i)$ such that*

$$\varpi\left(\pi^{-1}_{\{i_1,\ldots,i_k\}}[C]\right) = (\varpi_{i_1} \otimes \ldots \otimes \varpi_{i_k})(C) \tag{4.13}$$

for all $C \in \bigotimes_{j=1}^{k} \mathcal{A}_{i_j}$ and all $i_1, \ldots, i_k \in I$.

Proof Let $A \in \mathcal{A}$; then there exists a countable subset $J \subseteq I$ such that $A \in \mathcal{A}_J$. Let ϖ_J be the corresponding product measure on $\mathcal{A}_J$. Define $\varpi(A) := \varpi_J(A)$; then it easy to see that ϖ is a well-defined measure on $\mathcal{A}$, since the extension to countable products is unique. From the construction it follows also that the desired property (4.13) is satisfied. $\dashv$

Projective Limits

For the interpretation of some logics, the projective limit of a projective family of stochastic relations is helpful; this is the natural extension of
X^∞ a product. It will be discussed now. Denote by $X^\infty := \prod_{k \in \mathbb{N}} X$ the countable product of X with itself; recall that $\mathbb{P}$ is the probability functor, assigning to each measurable space its probability measures.

Definition 4.9.18 *Let X be a Polish space, and $(\mu_n)_{n \in \mathbb{N}}$ a sequence of probability measures $\mu_n \in \mathbb{P}(X^n)$. This sequence is called a* projective system *iff*

$$\mu_n(A) = \mu_{n+1}(A \times X)$$

for all $n \in \mathbb{N}$ and all Borel sets $A \in \mathcal{B}(X^n)$. A probability measure $\mu_\infty \in \mathbb{P}(X^\infty)$ is called the projective limit *of the projective system $(\mu_n)_{n \in \mathbb{N}}$ iff*

$$\mu_n(A) = \mu_\infty(A \times \prod_{j>n} X)$$

for all $n \in \mathbb{N}$ and $A \in \mathcal{B}(X^n)$.

Defining the projections

$$\pi_n^{n+1} : \langle x_1, \ldots, x_{n+1}\rangle \mapsto \langle x_1, \ldots, x_n\rangle,$$
$$\pi_n^{\infty} : \langle x_1, \ldots, \rangle \mapsto \langle x_1, \ldots, x_n\rangle,$$

the projectivity condition on $(\mu_n)_{n\in\mathbb{N}}$ can be rewritten as $\mu_n = \mathbb{P}\left(\pi_n^{n+1}\right)(\mu_{n+1})$ for all $n \in \mathbb{N}$ and the condition on μ_∞ to be a projective limit as $\mu_n = \mathbb{P}\left(\pi_n^{\infty}\right)(\mu_\infty)$ for all $n \in \mathbb{N}$. Thus a sequence of measures is a projective system iff each measure is the projection of the next one; its projective limit is characterized through the property that its values on cylinder sets coincide with the value of a member of the sequence, after taking projections. A special case is given by product measures. Assume that $\mu_n = \varpi_1 \otimes \ldots \otimes \varpi_n$, where $(\varpi_n)_{n\in\mathbb{N}}$ is a sequence of probability measures on X. Then the condition on projectivity is satisfied, and the projective limit is the infinite product constructed above. It should be noted, however, that the projectivity condition does not express $\mu_{n+1}(A \times B)$ in terms of $\mu_n(A)$ for an arbitrary measurable set $B \subseteq X$, as the product measure does; it only says that $\mu_n(A) = \mu_{n+1}(A \times X)$ holds.

It is not immediately obvious that a projective limit exists in general, given the rather weak dependency of the measures. In general, it will not, and this is why. The basic idea for the construction of the infinite product has been to define the limit on the cylinder sets and then to extend this set function—but it has to be established that it is indeed σ-additive, and this is difficult in general. The crucial property in the proof above has been that $\mu_{n_k}(A_k) \to 0$ whenever $(A_n)_{n\in\mathbb{N}}$ is a sequence of cylinder set A_k (with at most n_k components that do not equal X) that decreases to $\emptyset$. This property has been established above for the case of the infinite product through Fubini's Theorem, but this is not available in the general setting considered here, because we do not deal with an infinite product of measures. We will see, however, that a topological argument will be helpful. This is why we did postulate the base space X to be Polish. Crucial property

We start with an even stronger topological condition, viz., that the space under consideration is compact and metric. The central statement is

Proposition 4.9.19 *Let X be a compact metric space. Then the projective system $(\mu_n)_{n\in\mathbb{N}}$ has a unique projective limit μ_∞.*

Proof 0. Let $A = A'_k \times \prod_{j>k} X$ be a cylinder set with $A'_k \in \mathcal{B}(X^k)$. Define $\mu^\bullet(A) := \mu_k(A'_k)$. Then $\mu^\bullet$ is well defined on the cylinder

Idea of the proof

sets, since the sequence forms a projective system. In order to show that $\mu^\bullet$ is σ-additive on the cylinder sets, we take a decreasing sequence $(A_n)_{n\in\mathbb{N}}$ of cylinder sets with $\bigcap_{n\in\mathbb{N}} A_n = \emptyset$ and show that $\inf_{n\in\mathbb{N}} \mu^\bullet(A_n) = 0$. In fact, suppose that $(A_n)_{n\in\mathbb{N}}$ is decreasing with $\mu^\bullet(A_n) > \delta$ for all $n \in \mathbb{N}$; then we show that $\bigcap_{n\in\mathbb{N}} A_n \neq \emptyset$ by approximating each A_n from within through a compact set, which is given to us through Lemma 4.6.13. Then we are in a position to apply the observation that compact sets have the finite intersection property: A decreasing sequence of nonempty compact sets cannot have an empty intersection.

1. We can write $A_n = A'_n \times \prod_{j>k_n} X$ for some $A'_n \in \mathcal{B}(X^{k_n})$. From Lemma 4.6.13 we obtain for each n a closed, hence compact set $K'_n \subseteq A'_n$ such that $\mu_{k_n}(A'_n \setminus K'_n) < \delta/2^n$. Because X^∞ is compact by Tihonov's Theorem 3.2.12, $K''_n := K'_n \times \prod_{j>k_n} X$ is a compact set, and $K_n := \bigcap_{j=1}^n K''_j \subseteq A_n$ is compact as well, with

$$\begin{aligned}\mu^\bullet(A_n \setminus K_n) &\leq \mu^\bullet(\bigcup_{j=1}^n A''_n \setminus K''_j) \leq \sum_{j=i}^n \mu^\bullet(A''_j \setminus K''_j)\\ &= \sum_{j=1}^n \mu_{k_j}(A'_j \setminus K'_j) \leq \sum_{j=1}^\infty \delta/2^j = \delta.\end{aligned}$$

Thus $(K_n)_{n\in\mathbb{N}}$ is a decreasing sequence of nonempty compact sets. Consequently, by the finite intersection property for compact sets,

$$\emptyset \neq \bigcap_{n\in\mathbb{N}} K_n \subseteq \bigcap_{n\in\mathbb{N}} A_n.$$

2. Since the cylinder sets generate the Borel sets of X^∞ and since $\mu^\bullet$ is σ-additive, we know that there exists a unique extension $\mu_\infty \in \mathbb{P}(X^\infty)$ to it. Clearly, if $A \subseteq X^n$ is a Borel set, then

$$\mu_\infty(A \times \textstyle\prod_{j>n} X) = \mu^\bullet(A \times \prod_{j>n} X) = \mu_n(A),$$

Hurray!

so we have constructed a projective limit.

3. Uniqueness is established by an appeal to the π-λ-Theorem. Suppose that μ' is another probability measure in $\mathbb{P}(X^\infty)$ that has the desired property. Consider

$$\mathcal{D} := \{D \in \mathcal{B}(X^\infty) \mid \mu_\infty(D) = \mu'(D)\}.$$

It is clear that $\mathcal{D}$ contains all cylinder sets and that it is closed under complements and under countable disjoint unions. By the π-λ-Theorem 1.6.30, $\mathcal{D}$ contains the σ-algebra generated by the cylinder sets, hence all Borel subset of X^{∞}. This establishes uniqueness of the extension. $\dashv$

The proof makes critical use of the observation that we can approximate the measure of a Borel set arbitrarily well by compact sets from within; see Lemma 4.6.13. It is also important to observe that compact sets have the finite intersection property: If each finite intersection of a family of compact sets is nonempty, the intersection of the entire family cannot be empty. Consequently the proof given above works in general Hausdorff spaces, provided the measures under consideration have the approximation property mentioned above.

We free ourselves from the restrictive assumption of having a compact metric space using the Alexandrov embedding of a Polish space into a compact metric space.

Proposition 4.9.20 *Let X be a Polish space, $(\mu_n)_{n\in\mathbb{N}}$ be a projective system on X. Then there exists a unique projective limit $\mu_\infty \in \mathbb{P}(X^{\infty})$ for $(\mu_n)_{n\in\mathbb{N}}$.*

Proof X is a dense measurable subset of a compact metric space $\vec{X}$ by Alexandrov's Theorem 4.3.27. Defining $\vec{\mu}_n(B) := \mu_n(B \cap X^n)$ for the Borel set $B \subseteq \vec{X}^n$ yields a projective system $\left(\vec{\mu}_n\right)_{n\in\mathbb{N}}$ on $\vec{X}$ with a projective limit $\vec{\mu}_\infty$ by Proposition 4.9.19. Since by construction $\vec{\mu}_\infty(X^{\infty}) = 1$, restrict $\vec{\mu}_\infty$ to the Borel sets of X^{∞}, then the assertion follows. $\dashv$

An interesting application of this construction arises through stochastic relations that form a projective system. We will show now that there exists a kernel which may be perceived as a (pointwise) projective limit.

Corollary 4.9.21 *Let X and Y be Polish spaces, and assume that $J^{(n)} : X \rightsquigarrow Y^n$ is a stochastic relation for each $n \in \mathbb{N}$ such that the sequence $(J^{(n)}(x))_{n\in\mathbb{N}}$ forms a projective system on Y for each $x \in X$, in particular $J^{(n)}(x)(Y^n) = 1$ for all $x \in X$. Then there exists a unique stochastic relation J_∞ on X and Y^{∞} such that $J_\infty(x)$ is the projective limit of $(J^{(n)}(x))_{n\in\mathbb{N}}$ for each $x \in X$.*

Plan

Proof 0. Let for x fixed $J_\infty(x)$ be the projective limit of the projective system $(J^{(n)}(x))_{n\in\mathbb{N}}$. By the definition of a stochastic relation, we need to show that the map $x \mapsto J_\infty(x)(B)$ is measurable for every $B \in \mathcal{B}(Y^\infty)$.

1. We apply the principle of good sets, considering

$$\mathcal{D} := \{B \in \mathcal{B}(Y^\infty) \mid x \mapsto J_\infty(x)(B) \text{ is measurable}\}.$$

The general properties of measurable functions imply that $\mathcal{D}$ is a σ-algebra on Y^∞. Take a cylinder set $B = B_0 \times \prod_{j>k} Y$ with $B_0 \in \mathcal{B}(Y^k)$ for some $k \in \mathbb{N}$; then, by the properties of the projective limit, we have $J_\infty(x)(B) = J^{(k)}(x)(B_0)$. But $x \mapsto J^{(k)}(x)(B_0)$ constitutes a measurable function on X. Consequently, $B \in \mathcal{D}$, and so $\mathcal{D}$ contains the cylinder sets which generate $\mathcal{B}(Y^\infty)$. Thus measurability is established for each Borel set $B \subseteq Y^\infty$, arguing with the π-λ-Theorem 1.6.30 as in the last part of the proof for Proposition 4.9.19. $\dashv$

4.9.3 Case Study: Continuous Time Stochastic Logic

We illustrate the construction of the projective limit through the interpretation of a path logic over infinite paths; the logic is called `CSL`—*continuous time stochastic logic*. Since the discussion of this application requires some preparations, some of which are of independent interest, we develop the example in a series of steps.

We introduce `CSL` now and describe it informally first.

Fix P as a countable set of atomic propositions. We define recursively state formulas and path formulas for `CSL`:

State formulas are defined through the syntax

$$\varphi ::= \top \mid a \mid \neg\varphi \mid \varphi \wedge \varphi' \mid \mathcal{S}_{\bowtie p}(\varphi) \mid \mathcal{P}_{\bowtie p}(\psi).$$

Here $a \in P$ is an atomic proposition, ψ is a path formula, $\bowtie$ is one of the relational operators $<, \leq, \geq, >$, and $p \in [0,1]$ is a rational number.

Path formulas are defined through

$$\psi ::= \mathcal{X}^I \varphi \mid \varphi\, \mathcal{U}^I \varphi'$$

with φ, φ' as state formulas, $I \subseteq \mathbb{R}_+$ a closed interval of the real numbers with rational bounds (including $I = \mathbb{R}_+$).

The operator $\mathcal{S}_{\bowtie p}(\varphi)$ gives the *steady-state probability* for φ to hold with the boundary condition $\bowtie p$; the formula $\mathcal{P}$ replaces quantification: The *path-quantifier* formula $\mathcal{P}_{\bowtie p}(\psi)$ holds in a state s iff the probability of all paths starting in s and satisfying ψ is specified by $\bowtie p$. Thus ψ holds on almost all paths starting from s iff s satisfies $\mathcal{P}_{\geq 1}(\psi)$, a path being an alternating infinite sequence $\sigma = \langle s_0, t_0, s_1, t_1, \ldots \rangle$ of states x_i and of times t_i. Note that the time is being made explicit here. The *next operator* $\mathcal{X}^I\ \varphi$ is assumed to hold on path σ iff s_1 satisfies φ and $t_0 \in I$ holds. Finally, the *until operator* $\varphi_1\ \mathcal{U}^I\ \varphi_2$ holds on path σ iff we can find a point in time $t \in I$ such that the state $\sigma@t$ which σ occupies at time t satisfies φ_2, and for all times t' before that, $\sigma@t'$ satisfies φ_1. $\sigma@t$

A Polish state space S is fixed; this space is used for modeling a transition system that takes also time into account. We are not only interested in the next state of a transition but also in the time after which to make a transition. So the basic probabilistic data will be a stochastic relation $M : S \rightsquigarrow \mathbb{R}_+ \times S$; if we are in state s, we will do a transition to a new state s' after we did wait some specified time t; $M(s)(D)$ will give the probability that the pair $\langle t, s' \rangle \in D$. We assume that $M(s)(\mathbb{R}_+ \times S) = 1$ holds for all $s \in S$.

A *path* σ is an element of the set $(S \times \mathbb{R}_+)^\infty$. Path $\sigma = \langle s_0, t_0, s_1, t_1, \ldots \rangle$ may be written as $s_0 \xrightarrow{t_0} s_1 \xrightarrow{t_1} \ldots$ with the interpretation that t_i is the time spent in state s_i. Given $i \in \mathbb{N}$, denote s_i by $\sigma[i]$ as the $(i+1)$-st state of σ, and let $\delta(\sigma, i) := t_i$. Let for $t \in \mathbb{R}_+$ the index j be the smallest index k such that $t < \sum_{i=0}^{k} t_i$, and put $\sigma@t := \sigma[j]$, if j is defined; set $\sigma@t := \#$, otherwise (here # is a new symbol not in $S \cup \mathbb{R}_+$). $S_\#$ denotes $S \cup \{\#\}$; this is a Polish space when endowed with the sum σ-algebra. The definition of $\sigma@t$ makes sure that for any time t we can find a rational time t' with $\sigma@t = \sigma@t'$. Path

We will deal only with infinite paths. This is no loss of generality because events that happen at a certain time with probability 0 will have the effect that the corresponding infinite paths occur only with probability 0. Thus we do not prune the path; this makes the notation somewhat easier to handle.

The Borel sets $\mathcal{B}((S\times\mathbb{R}_+)^\infty)$ are the smallest σ-algebra which contains all the cylinder sets

$$\{\prod_{j=1}^{n}(B_j \times I_j) \times \prod_{j>n}(S \times \mathbb{R}_+) \mid n \in \mathbb{N}, I_1, \dots, I_n \text{ rational intervals},$$
$$B_1, \dots, B_n \in \mathcal{B}(S)\}.$$

Thus a cylinder set is an infinite product that is determined through the finite product of an interval with a Borel set in S. It will be helpful to remember that the intersection of two cylinder sets is again a cylinder set.

Given $M : S \rightsquigarrow \mathbb{R}_+ \times S$ with Polish S, define inductively $M_1 := M$, and

$$M_{n+1}(s_0)(D) := \int_{(\mathbb{R}_+\times S)^n} M(s_n)(D_{t_0,s_1,\dots,t_{n-1},s_n})\, dM_n(s_0)$$
$$(t_0, s_1, \dots, t_{n-1}, s_n)$$

for the Borel set $D \subseteq (\mathbb{R}_+ \times S)^{n+1}$. Let us illustrate this for $n = 1$. Given $D \in \mathcal{B}((\mathbb{R}_+ \times S)^2)$ and $s_0 \in S$ as a state to start from, we want to calculate the probability $M_2(s_0)(D)$ that $\langle t_0, s_1, t_1, s_2\rangle \in D$. This is the probability for the initial path $\langle s_0, t_0, s_1, t_1, s_2\rangle$ (a path*let*), given the initial state s_0. Since $\langle t_0, s_1\rangle$ is taken care of in the first step, we fix it and calculate $M(s_1)(\{\langle t_1, s_2\rangle \mid \langle t_0, s_1, t_1, s_2\rangle \in D\}) = M(s_1)(D_{t_0,s_1})$, by averaging, using the probability provided by $M(s_0)$, so that we obtain

$$M_2(s_0)(D) = \int_{\mathbb{R}_+\times S} M(s_1)(D_{t_0,s_1})\, dM(s_0)(t_0, s_1).$$

We obtain for the general case $M_{n+1}(s_0)(D)$ as the probability for $\langle s_0, t_0, \dots, s_n, t_n, s_{n+1}\rangle$ as the initial piece of an infinite path to be a member of D. This probability indicates that we start in s_0, remain in this state for t_0 units of time, then enter state s_1, remain there for t_1 time units, etc., and finally leave state s_n after t_n time units, entering s_{n+1}, all this happening within D.

We claim that $\big(M_n(s)\big)_{n\in\mathbb{N}}$ is a projective system. We first see from Example 4.9.5 that $M_n : S \rightsquigarrow (\mathbb{R}_+ \times S)^n$ defines a transition kernel for each $n \in \mathbb{N}$. Let $D = A \times (\mathbb{R}_+ \times S)$ with $A \in \mathcal{B}((\mathbb{R}_+ \times$

$S)^n$), then $M(s_n)(D_{t_0,s_1,\ldots,t_{n-1},s_n}) = M(s_n)(\mathbb{R}_+ \times S) = 1$ for all $\langle t_0, s_1, \ldots, t_{n-1}, s_n\rangle \in A$, so that we obtain $M_{n+1}(s)(A \times (\mathbb{R}_+ \times S)) = M_n(s)(A)$. Consequently, the condition on projectivity is satisfied. Hence there exists a unique projective limit; thus a transition kernel

$$M_\infty : S \rightsquigarrow (\mathbb{R}_+ \times S)^\infty$$

with

$$M_n(s)(A) = M_\infty(s)\bigl(A \times \prod_{k>n} (\mathbb{R}_+ \times S)\bigr)$$

for all $s \in S$ and for all $A \in \mathcal{B}((\mathbb{R}_+ \times S)^n)$.

The projective limit displays indeed limiting behavior: Suppose B is an infinite measurable cube $\prod_{n\in\mathbb{N}} B_n$ with $B_n \in \mathcal{B}(\mathbb{R}_+ \times S)$ as Borel sets. Because

$$B = \bigcap_{n\in\mathbb{N}} \Bigl(\prod_{1\le j\le n} B_j \times \prod_{j>n}(\mathbb{R}_+ \times S)\Bigr)$$

represents B as the intersection of a monotonically decreasing sequence, we have for all $s \in S$

$$\begin{aligned} M_\infty(s)(B) &= \lim_{n\to\infty} M_\infty(s)\bigl(\prod_{1\le j\le n} B_j \times \prod_{j>n}(\mathbb{R}_+ \times S)\bigr) \\ &= \lim_{n\to\infty} M_n(s)\bigl(\prod_{1\le j\le n} B_j\bigr). \end{aligned}$$

Hence $M_\infty(s)(B)$ is the limit of the probabilities $M_n(s)(B_n)$ at step n.

In this way models based on a Polish state space S yield stochastic relations $S \rightsquigarrow (\mathbb{R}_+ \times S)^\infty$ through projective limits. Without this limit it would be difficult to model the transition behavior on infinite paths; the assumption that we work in Polish spaces makes sure that these limits in fact do exist. To get started, we need to assume that given a state $s \in S$, there is always a state to change into after a finite amount of time.

We obtain—as an aside—a recursive formulation for the transition law $M : X \rightsquigarrow (\mathbb{R}_+ \times S)^\infty$ as a first consequence of the construction for the projective limit. Interestingly, it reflects the domain equation $(\mathbb{R}_+ \times S)^\infty = (\mathbb{R}_+ \times S) \times (\mathbb{R}_+ \times X)^\infty$.

Lemma 4.9.22 *If $D \in \mathcal{B}((\mathbb{R}_+ \times S)^\infty)$, then*

$$M_\infty(s)(D) = \int_{\mathbb{R}_+ \times S} M_\infty(s')(D_{\langle t,s'\rangle})\, M_1(s)(d\langle t, s'\rangle)$$

holds for all $s \in S$.

Proof Recall that $D_{\langle t,s'\rangle} = \{\tau \mid \langle t, s', \tau\rangle \in D\}$. Let

$$D = (H_1 \times \ldots \times H_{n+1}) \times \prod_{j>n} (\mathbb{R}_+ \times S)$$

be a cylinder set with $H_i \in \mathcal{B}(\mathbb{R}_+ \times S), 1 \leq i \leq n+1$. The equation in question in this case boils down to

$$M_{n+1}(s)(H_1 \times \ldots \times H_{n+1}) = \int_{H_1} M_n(s')(H_2 \times \ldots \times H_{n+1}) M_1(s)(d\langle t, s'\rangle).$$

This may easily be derived from the definition of the projective sequence. Consequently, the equation in question holds for all cylinder sets, and thus the π-λ-Theorem 1.6.30 implies that it holds for all Borel subsets of $(\mathbb{R}_+ \times S)^\infty$. $\dashv$

This decomposition indicates that we may first select in state s a new state and a transition time; with these data the system then works just as if the selected new state would have been the initial state. The system does not have a memory but reacts depending on its current state, no matter how it arrived there. Lemma 4.9.22 may accordingly be interpreted as a Markov property for a process, the behavior of which is independent of the specific step that is undertaken.

Markov property

We need some information about the @-operator before continuing.

Lemma 4.9.23 *$\langle \sigma, t\rangle \mapsto \sigma@t$ is a Borel measurable map from $(S \times \mathbb{R}_+)^\infty \times \mathbb{R}_+$ to $S_\#$. In particular, the set $\{\langle \sigma, t\rangle \mid \sigma@t \in S\}$ is a measurable subset of $(S \times \mathbb{R}_+)^\infty \times \mathbb{R}_+$.*

Proof 0. Note that we claim joint measurability in both components, which is strictly stronger than measurability in each component. Thus we have to show that $\{\langle \sigma, t\rangle \mid \sigma@t \in A\}$ is a measurable subset of $(S \times \mathbb{R}_+)^\infty \times \mathbb{R}_+$, whenever $A \subseteq S_\#$ is Borel.

1. Because the map $\sigma \mapsto \delta(\sigma, i)$ is a projection for fixed $i \in \mathbb{N}$, $\delta(\cdot, i)$ is measurable; hence $\sigma \mapsto \sum_{i=0}^{j} \delta(\sigma, i)$ is. Consequently,

$$\begin{aligned}\{\langle \sigma, t\rangle \mid \sigma@t = \#\} &= \{\langle \sigma, t\rangle \mid \forall j : t \geq \sum_{i=0}^{j} \delta(\sigma, i)\} \\ &= \bigcap_{j \geq 0} \{\langle \sigma, t\rangle \mid t \geq \sum_{i=0}^{j} \delta(\sigma, i)\}.\end{aligned}$$

This is clearly a measurable set.

2. Put $stop(\sigma, t) := \inf\{k \geq 0 \mid t < \sum_{i=0}^{k} \delta(\sigma, i)\}$; thus $stop(\sigma, t)$ is the smallest index for which the accumulated waiting times exceed t.

$$\begin{aligned} X_k := \{\langle \sigma, t\rangle \mid stop(\sigma, t) = k\} = \{\langle \sigma, t\rangle \mid \sum_{i=0}^{k-1} \delta(\sigma, i) \\ \leq t < \sum_{i=0}^{k} \delta(\sigma, i)\} \end{aligned}$$

is a measurable set by Corollary 4.2.5. Now let $B \in \mathcal{B}(S)$ be a Borel set; then

$$\begin{aligned} &\{\langle \sigma, t\rangle \mid \sigma @ t \in B\} \\ &\quad = \bigcup_{k \geq 0} \{\langle \sigma, t\rangle \mid \sigma @ t \in B, stop(\sigma, t) = k\} \\ &\quad = \bigcup_{k \geq 0} \{\langle \sigma, t\rangle \mid \sigma[k] \in B, stop(\sigma, t) = k\} \\ &\quad = \bigcup_{k \in \mathbb{N}} \bigl(X_k \cap \bigl(\prod_{i<k} (S \times \mathbb{R}_+) \times (B \times \mathbb{R}_+) \times \prod_{i>k} (S \times \mathbb{R}_+)\bigr)\bigr). \end{aligned}$$

Because X_k is measurable, the latter set is measurable. This establishes measurability of the @-map. $\dashv$

As a consequence, we establish that some sets and maps, which will be important for the later development, are actually measurable. A notational convention for improving readability is proposed: The letter σ will always denote a generic element of $(S \times \mathbb{R}_+)^\infty$, and the letter ϑ always a generic element of $\mathbb{R}_+ \times (S \times \mathbb{R}_+)^\infty$. Convention

Proposition 4.9.24 *We observe the following properties:*

1. $\{\langle \sigma, t\rangle \mid \lim_{i \to \infty} \delta(\sigma, i) = t\}$ *is a measurable subset of* $(S \times \mathbb{R}_+)^\infty \times \mathbb{R}_+$.

2. *Let* $N_\infty : S \rightsquigarrow (\mathbb{R}_+ \times S)^\infty$ *be a stochastic relation; then*

$$\begin{aligned} s &\mapsto \liminf_{t \to \infty} N_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle @ t \in A\}), \\ s &\mapsto \limsup_{t \to \infty} N_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle @ t \in A\}) \end{aligned}$$

constitute measurable maps $X \to \mathbb{R}_+$ *for each Borel set* $A \subseteq S$.

Proof 0. The proof makes crucial use of the fact that the real line is a complete metric space and that the rational numbers are a dense and countable set.

1. In order to establish part 1, write

$$\begin{aligned}
&\{\langle\sigma, t\rangle \mid \lim_{i\to\infty} \delta(\sigma, i) = t\} \\
&\quad = \bigcap_{\mathbb{Q}\ni\epsilon>0} \bigcup_{n\in\mathbb{N}} \bigcap_{m\geq n} \{\langle\sigma, t\rangle \mid \left|\delta(\sigma, m) - t\right| < \epsilon\}.
\end{aligned}$$

By Lemma 4.9.23, the set

$$\begin{aligned}
\{\langle\sigma, t\rangle \mid \left|\delta(\sigma, m) - t\right| < \epsilon\} = {} & \{\langle\sigma, t\rangle \mid \delta(\sigma, m) > t - \epsilon\} \\
& \cap \{\langle\sigma, t\rangle \mid \delta(\sigma, m) < t + \epsilon\}
\end{aligned}$$

is a measurable subset of $(S \times \mathbb{R}_+)^\infty \times \mathbb{R}_+$, and since the union and the intersections are countable, measurability is inferred.

2. From the definition of the @-operator, it is immediate that given an infinite path σ and a time $t \in \mathbb{R}_+$, there exists a rational t' with $\sigma@t = \sigma@t'$. Thus we obtain, for an arbitrary real number x, arbitrary Borel sets $A \subseteq S$ and $s \in S$

$$\begin{aligned}
&\liminf_{t\to\infty} N_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle@t \in A\}) \leq x \\
&\quad\Leftrightarrow \sup_{t\geq 0} \inf_{r\geq t} N_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle@r \in A\}) \leq x \\
&\quad\Leftrightarrow \sup_{\mathbb{Q}\ni t\geq 0} \inf_{\mathbb{Q}\ni r\geq t} N_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle@r \in A\}) \leq x \\
&\quad\Leftrightarrow s \in \bigcap_{\mathbb{Q}\ni t\geq 0} \bigcup_{\mathbb{Q}\ni r\geq t} A_{r,x}
\end{aligned}$$

with

$$A_{r,x} := \big\{s' \mid N_\infty(s')\big(\{\vartheta \mid \langle s', \vartheta\rangle@r \in A\}\big) \leq x\big\}.$$

We infer that $A_{r,x}$ is a measurable subset of S from the fact that N_∞ is a stochastic relation. Since a map $f : W \to \mathbb{R}$ is measurable iff each of the sets $\{w \in W \mid f(w) \leq s\}$ is a measurable subset of W, the assertion follows for the first map. The second part is established in exactly the same way, using that $f : W \to \mathbb{R}$ is measurable iff $\{w \in W \mid f(w) \geq s\}$ is a measurable subset of W and observing

$$\begin{aligned}
&\limsup_{t\to\infty} N_\infty(s)(\{\vartheta \mid \langle x, \vartheta\rangle@t \in A\}) \geq x \\
&\quad\Leftrightarrow \inf_{\mathbb{Q}\ni t\geq 0} \sup_{\mathbb{Q}\ni r\geq t} N_\infty(x)(\{\vartheta \mid \langle s, \vartheta\rangle@r \in A\}) \geq x.
\end{aligned}$$

$\dashv$

This has some consequences which will come in useful for the interpretation of CSL. Before stating them, it is noted that the statement above (and the consequences below) does not make use of N_∞ being a projective limit; in fact, we assume $N_\infty : S \rightsquigarrow (\mathbb{R}_+ \times S)^\infty$ to be an arbitrary stochastic relation. A glimpse at the proof shows that these statements even hold for finite transition kernels, but since we will use it for the probabilistic case, we stick to stochastic relations.

Now for the consequences. As a first consequence, we obtain that the set on which the asymptotic behavior of the transition times is reasonable (in the sense that it tends probabilistically to a limit) is well behaved in terms of measurability.

Corollary 4.9.25 *Let $A \subseteq X$ be a Borel set, and assume that $N_\infty : S \rightsquigarrow (\mathbb{R}_+ \times S)^\infty$ is a stochastic relation. Then*

1. *the set $Q_A := \{s \in S \mid \lim_{t\to\infty} N_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle@t \in A\})$ exists$\}$ on which the limit exists is a Borel subset of S,*
2. *$s \mapsto \lim_{t\to\infty} N_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle@t \in A\}$ is a measurable map $Q_A \to \mathbb{R}_+$.*

Proof Since $s \in Q_A$ iff

$$\liminf_{t\to\infty} N_\infty(x)(\{\vartheta \mid \langle s, \vartheta\rangle@t \in A\}) = \limsup_{t\to\infty} N_\infty(x)(\{\vartheta \mid \langle s, \vartheta\rangle@t \in A\}),$$

and since the set on which two Borel measurable maps coincide is a Borel set itself, the first assertion follows from Proposition 4.9.24, part 2. This implies the second assertion as well. ⊣

When dealing with the semantics of the until operator below, we will also need to establish measurability of certain sets. Preparing for that, we state

Lemma 4.9.26 *Assume that A_1 and A_2 are Borel subsets of S, and let $I \subseteq \mathbb{R}_+$ be an Interval; then*

$$U(I, A_1, A_2) := \{\sigma \mid \exists t \in I : \sigma@t \in A_2 \text{ and } \forall t' \in [0, t[: \sigma@t' \in A_1\}$$

is a measurable set of paths, and thus $U(I, A_1, A_2) \in \mathcal{B}((S \times \mathbb{R}_+)^\infty)$.

Proof 0. Remember that, given a path σ and a time $t \in \mathbb{R}_+$, there exists a rational time $t_r \leq t$ with $\sigma@t = \sigma@t_r$. Consequently,

$$U(I, A_1, A_2) = \bigcup_{t \in \mathbb{Q} \cap I} \left(\{\sigma \mid \sigma@t \in A_2\} \cap \bigcap_{t' \in \mathbb{Q} \cap [0,t]} \{\sigma \mid \sigma@t' \in A_1\} \right).$$

The inner intersection is countable and is performed over measurable sets by Lemma 4.9.23, thus forming a measurable set of paths. Intersecting it with a measurable set and forming a countable union yield a measurable set again. ⊣

Interpretation of `CSL` Now that a description for the behavior of paths is available, we are ready for a probabilistic interpretation of `CSL`. We did start from the assumption that the one-step behavior is governed through a stochastic relation $M : S \rightsquigarrow \mathbb{R}_+ \times S$ with $M(s)(\mathbb{R}_+ \times S) = 1$ for all $s \in S$ from which the stochastic relation $M_\infty : S \rightsquigarrow \mathbb{R}_+ \times (S \times \mathbb{R}_+)^\infty$ has been constructed. The interpretations for the formulas can be established now, and we show that the sets of states resp. paths on which formulas are valid are Borel measurable.

$V : P \to \mathcal{B}(S)$

To get started on the formal definition of the semantics, we assume that we know for each atomic proposition which state it is satisfied in. Thus we fix a map V that maps P to $\mathcal{B}(S)$, assigning each atomic proposition a Borel set of states.

The semantics is described as usual recursively through relation $\models$ between states resp. paths and formulas. Hence $s \models \varphi$ means that state formula φ holds in state s, and $\sigma \models \psi$ means that path formula ψ is true on path σ.

Here we go:

$s \models \varphi$

1. $s \models \top$ is true for all $s \in S$.
2. $s \models a$ iff $s \in V(a)$.
3. $s \models \varphi_1 \wedge \varphi_2$ iff $s \models \varphi_1$ and $s \models \varphi_2$.
4. $s \models \neg\varphi$ iff $s \models \varphi$ is false.

5. $s \models \mathcal{S}_{\bowtie p}(\varphi)$ iff $\lim_{t\to\infty} M_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle @t \models \varphi\})$ exists and is $\bowtie p$.

6. $s \models \mathcal{P}_{\bowtie p}(\psi)$ iff $M_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle \models \psi\}) \bowtie p$.

7. $\sigma \models \mathcal{X}^I \varphi$ iff $\sigma[1] \models \varphi$ and $\delta(\sigma, 0) \in I$. $\sigma \models \psi$

8. $\sigma \models \varphi_1 \mathcal{U}^I \varphi_2$ iff $\exists t \in I : \sigma@t \models \varphi_2$ and $\forall t' \in [0, t[: \sigma@t' \models \varphi_1$.

Most interpretations should be obvious. Given a state s, we say that $s \models \mathcal{S}_{\bowtie p}(\varphi)$ iff the asymptotic behavior of the paths starting at s gets eventually stable with a limiting probability given by $\bowtie p$. Similarly, $s \models \mathcal{P}_{\bowtie p}(\psi)$ holds iff the probability that path formula ψ holds for all s-paths is specified through $\bowtie p$. For $\langle s_0, t_0, s_1, \dots, \rangle \models \mathcal{X}^I \varphi$ to hold, we require $s_1 \models \varphi$ after a waiting time t_0 for the transition to be a member of interval I. Finally, $\sigma \models \varphi_1 \mathcal{U}^I \varphi_2$ holds iff we can find a time point t in the interval I such that the corresponding state $\sigma@t$ satisfies φ_2, and for all states on that path before t, formula φ_1 is assumed to hold. The kinship to `CTL*` is obvious; see Example 2.7.65.

Denote by $[\![\varphi]\!]$ and $[\![\psi]\!]$ the set of all states for which the state formula φ holds, resp. the set of all paths for which the path formula φ is valid. We do not distinguish notationally between these sets, as far as the basic domains are concerned, since it should always be clear whether we describe a state formula or a path formula. $[\![\varphi]\!], [\![\psi]\!]$

We show that we are dealing with measurable sets. Most of the work for establishing this has been done already. What remains to be done is to fit in the patterns that we have set up in Proposition 4.9.24 and its corollaries.

Proposition 4.9.27 *The set $[\![\xi]\!]$ is Borel whenever ξ is a state formula or a path formula.*

Proof 0. The proof proceeds by induction on the structure of the formula ξ. The induction starts with the formula $\top$, for which the assertion is true, and with the atomic propositions, for which the assertion follows from the assumption on V: $[\![a]\!] = V(a) \in \mathcal{B}(S)$. We assume for the induction step that we have established that $[\![\varphi]\!]$, $[\![\varphi_1]\!]$ and $[\![\varphi_2]\!]$ are Borel measurable.

1. For the next operator, we write

$$[\![\mathcal{X}^I \varphi]\!] = \{\sigma \mid \sigma[1] \in [\![\varphi]\!] \text{ and } \delta(\sigma, 0) \in I\}.$$

This is the cylinder set $(S \times I \times [\![\varphi]\!] \times \mathbb{R}_+) \times (S \times \mathbb{R}_+)^\infty$; hence is a Borel set.

2. The until operator may be represented through

$$[\![\varphi_1 \,\mathcal{U}^I\, \varphi_2]\!] = U(I, [\![\varphi_1]\!], [\![\varphi_2]\!]),$$

which is a Borel set by Lemma 4.9.26.

3. Since $M_\infty : S \rightsquigarrow (\mathbb{R}_+ \times S)^\infty$ is a stochastic relation, we know that

$$[\![\mathcal{P}_{\bowtie p}(\psi)]\!] = \{s \in S \mid M_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle \in [\![\varphi]\!]\}) \bowtie p\}$$

is a Borel set.

4. We know from Corollary 4.9.25 that the set

$$Q_{[\![\varphi]\!]} := \{s \in S \mid \lim_{t\to\infty} M_\infty(s)(\{\vartheta \mid \langle s, \vartheta\rangle @t \in [\![\varphi]\!]\}) \text{ exists}\}$$

is a Borel set and that

$$J_\varphi : Q_{[\![\varphi]\!]} \ni s \mapsto \lim_{t\to\infty} M_\infty(x)\,(\{\vartheta \mid \langle s, \vartheta\rangle @t \in [\![\varphi]\!]\}) \in [0, 1]$$

is a Borel measurable function. Consequently,

$$[\![\mathcal{S}_{\bowtie p}(\varphi)]\!] = \{s \in Q_{[\![\varphi]\!]} \mid J_\varphi(s) \bowtie p\}$$

is a Borel set. $\dashv$

Measurability of the sets on which a given formula is valid constitutes of course a prerequisite for computing interesting properties. So we can compute, e.g.,

$$\mathcal{P}_{\geq 0.5}((\neg down)\,\mathcal{U}^{[10,20]}\,\mathcal{S}_{\geq 0.8}(up_2 \vee up_3)))$$

as the set of all states that with probability at least 0.5 will reach a state between 10 and 20 time units so that the system is operational ($up_2, up_3 \in P$) in a steady state with a probability of at least 0.8; prior to reaching this state, the system must be operational continuously ($down \in P$).

The description of the semantics is just the basis for entering into the investigation of expressivity of the models associated with M and with V. We leave `CSL` here, however, and note that the construction of the projective limit is the basic and fundamental ingredient for further investigations.

4.9.4 Case Study: A Stochastic Interpretation of Game Logic

Game logic is a modal logic, the modalities of which are given by games. The grammar for games has been presented and discussed in Example 2.7.5; here it is again:

$$g ::= \gamma \mid g_1 \cup g_2 \mid g_1 \cap g_2 \mid g_1; g_2 \mid g^d \mid g^* \mid g^\times \mid \varphi?$$

with $\gamma \in \Gamma$ and φ a formula of the underlying logic; the set Γ is the collection of primitive games, from which compound games are constructed.

This section will deal with a stochastic interpretation of game logic, based on the discussion from Sect. 2.7, which indicates that Kripke models are not fully adequate for this task. It results in an interpretation of game logics through neighborhood models, which in turn are defined through effectivity functions. Since we have stochastic effectivity functions at our disposal, we will base a probabilistic interpretation based on them. We refer also to the discussion above concerning the observation that the games we are discussing here are different from the Banach–Mazur games, which are introduced in Chap. 1 and put to good use in Sects. 1.7 and 3.5.2.

The games which we consider here are assumed to be determined, so if one player does not have a winning strategy, the other player has one (again, we do not formalize the notion of a strategy here); determinedness serves here the purpose of relating the players' behavior to each other. In this sense, determinedness has the consequence that if we model Angel's portfolio for a game $\gamma \in \Gamma$ through the effectivity function P_γ, then Demon's portfolio in state s is given by $\{A \mid S \setminus A \notin P_\gamma(s)\}$, which defines also an effectivity function.

We make as in Example 2.7.5 the following assumptions (when writing down games, we assume for simplicity that composition binds tighter than angelic or demonic choice):

❶ $(g^d)^d$ is identical to g.

❷ Demonic choice can be represented through angelic choice: The game $g_1 \cap g_2$ coincides with the game $(g_1^d \cup g_2^d)^d$.

❸ Similarly, demonic iteration can be represented through its angelic counterpart: $(g^\times)^d$ is equal to $(g^d)^*$.

❹ Composition is right distributive with respect to angelic choice: Making a decision to play g_1 or g_2 and then playing g should be the same as deciding to play $g_1; g$ or $g_2; g$; thus $(g_1 \cup g_2); g$ equals $g_1; g \cup g_2; g$.

Note that left distributivity would mean that a choice between $g; g_1$ and $g; g_2$ is the same as playing first g, then $g_1 \cup g_2$, as discussed in Example 2.7.22. This is a somewhat restrictive assumption, since the choice of playing g_1 or g_2 may be a decision made by Angel only after g is completed. Thus we do not assume this in general. It will turn out, however, that in Kripke generated models these choices are in fact equivalent; see Proposition 4.9.40.

❺ We assume similarly that $g^*; g_0$ equals $g_0 \cup g^*; g; g_0$. Hence when playing $g^*; g_0$, Angel may decide to play g not at all and to continue with g_0 right away or to play g^* followed by $g; g_0$. Thus $g^*; g_0$ expands to $g_0 \cup g; g_0 \cup g; g; g_0 \cup \ldots$.

❻ $(g_1; g_2)^d$ is the same as $g_1^d; g_2^d$.

❼ The binary operations angelic and demonic choice are commutative and associative, and composition is associative as well.

We do not discuss for the time being the test operator, since its semantics depends on a model; we will fit in the operator later in (p. 603), when all the necessary operations are available.

This is the stage on which we play our game.

Definition 4.9.28 *A game frame* $\mathcal{G} = \big((S, \mathcal{A}), (P_\gamma)_{\gamma \in \Gamma}\big)$ *has a measurable space* $(S, \mathcal{A})$ *of states and a t-measurable map* $P_\gamma : S \to \boldsymbol{E}(S, \mathcal{A})$ *for each primitive game* $\gamma \in \Gamma$.

($\boldsymbol{E}(S, \mathcal{A})$ is defined on page 442). As usual, we will omit the reference to the σ-algebra of the state space S, unless we really need to write it down. Fix a game frame $\mathcal{G} = \big(S, (P_\gamma)_{\gamma \in \Gamma}\big)$, the set of primitive games is extended by the empty game ϵ, and set $P_\epsilon := D$ with D as the Dirac effectivity function according to Example 4.1.18. We assume in the sequel that $\epsilon \in \Gamma$.

We define now recursively the set-valued function $\Omega_{\mathcal{G}}(g \mid A, q)$ with the informal meaning that this set describes the set of states so that Angel has a strategy of reaching a state in set A with probability greater than q

upon playing game g. Assume that $A \in \mathcal{A}$ is a measurable subset of S, and $0 \le q < 1$, and define for $0 \le k \le \infty$

$$Q^{(k)}(q) := \{\langle a_1, \dots, a_k \rangle \in \mathbb{Q}^k \mid a_i \ge 0 \text{ and } \sum_{i=1}^{k} a_i \le q\}$$

as the set of all nonnegative rational k-tuples, the sum of which does not exceed q.

① Let $\gamma \in \Gamma$; then put

$$\Omega_{\mathcal{G}}(\gamma \mid A, q) := \{s \in S \mid \boldsymbol{\beta}_{\mathcal{A}}(A, > q) \in P_\gamma(s)\},$$

in particular $\Omega_{\mathcal{G}}(\epsilon \mid A, q) = \{s \in S \mid \delta_s(A) > q\} = A$. Thus $s \in \Omega_{\mathcal{G}}(\gamma \mid A, q)$ iff Angel has $\boldsymbol{\beta}_{\mathcal{A}}(A, > q)$ in its portfolio when playing γ in state s. This implies that the set of all state distributions which evaluate at A with a probability greater than q can be effected by Angel in this situation. If Angel does not play at all, hence if the game γ equals ϵ, nothing is about to change, which means $\Omega_{\mathcal{G}}(\epsilon \mid A, q) = \{s \mid \delta_s \in \boldsymbol{\beta}_{\mathcal{A}}(A, > q)\} = A$.

② Let g be a game; then

$$\Omega_{\mathcal{G}}(g^d \mid A, q) := S \setminus \Omega_{\mathcal{G}}(g \mid S \setminus A, q).$$

The game is determined; thus Demon can reach a set of states iff Angel does not have a strategy for reaching the complement. Consequently, upon playing g in state s, Demon can reach a state in A with probability greater than q iff Angel cannot reach a state in $S \setminus A$ with probability greater q.

Illustrating, let us assume for the moment that $P_\gamma = P_{K_\gamma}$, i.e., that the effectivity function for $\gamma \in \Gamma$ is generated from a stochastic relation K_γ; see Lemma 4.1.16. Then

$$s \in \Omega_{\mathcal{G}}(\gamma^d \mid A, q) \Leftrightarrow s \notin \Omega_{\mathcal{G}}(\gamma \mid S \setminus A, q) \Leftrightarrow K_\gamma(s)(S \setminus A) \le q.$$

In general, $s \in \Omega_{\mathcal{G}}(\gamma^d \mid A, q)$ iff $\boldsymbol{\beta}_{\mathcal{A}}(S \setminus A, > q) \notin P_\gamma(s)$ for $\gamma \in \Gamma$. This is exactly what one would expect in a determined game.

③ Assume s is a state such that Angel has a strategy for reaching a state in A when playing the game $g_1 \cup g_2$ with probability not greater than q. Then Angel should have a strategy in s for reaching a state in A when playing game g_1 with probability not greater than a_1 and playing game g_2 with probability not greater than a_2 such that $a_1 + a_2 \leq q$. Thus

$$\Omega_{\mathcal{G}}(g_1 \cup g_2 \mid A, q) := \bigcap_{a \in Q^{(2)}(q)} \big(\Omega_{\mathcal{G}}(A \mid g_1, a_1) \cup \Omega_{\mathcal{G}}(A \mid g_2, a_2)\big).$$

④ Right distributivity of composition over angelic choice translates to this equation:

$$\Omega_{\mathcal{G}}((g_1 \cup g_2); g \mid A, q) := \Omega_{\mathcal{G}}(g_1; g \cup g_2; g \mid A, q).$$

⑤ If $\gamma \in \Gamma$, put

$$\Omega_{\mathcal{G}}(\gamma; g \mid A, q) := \{s \in S \mid G_g(A, q) \in P_\gamma(s)\},$$

where

$$G_g(A, q) := \{\mu \in \mathbb{S}(S) \mid \int_0^1 \mu(\Omega_{\mathcal{G}}(g \mid A, r))\, dr > q\}. \tag{4.14}$$

Suppose that $\Omega_{\mathcal{G}}(g \mid A, r)$ is already defined for each r as the set of states for which Angel has a strategy to effect a state in A through playing g with probability greater than r. Given a distribution μ over the states, the integral $\int_0^1 \mu(\Omega_{\mathcal{G}}(g \mid A, r))\, dr$ is the expected value for entering a state in A through playing g for μ. The set $G_g(A, q)$ collects all distributions, the expected value of which is greater than q. We ask for all states such that Angel has this set in its portfolio when playing γ in this state. Being able to select this set from the portfolio means that, when playing γ and subsequently g, a state in A may be reached with probability greater than q.

⑥ This is just the translation of assumption ❺ above with a repeated application of the rule ③ for angelic choice:

$$\Omega_{\mathcal{G}}(g^*; g_0 \mid A, q) := \bigcap_{a \in Q^{(q)}(\infty)} \bigcup_{n \geq 0} \Omega_{\mathcal{G}}(g^n; g_0 \mid A, a_{n+1})$$

with $g^0 := \epsilon$ and $g^n := g; \ldots; g$ (n times).

One observes that $q \mapsto \Omega_{\mathcal{G}}(\gamma \mid A, q)$ is a monotonically decreasing function for $\gamma \in \Gamma$, because $q_1 \geq q_2$ implies $\boldsymbol{\beta}_{\mathcal{A}}(> q_1, A) \subseteq \boldsymbol{\beta}_{\mathcal{A}}(> q_2, A)$, so that $\boldsymbol{\beta}_{\mathcal{A}}(> q_1, A) \in P_\gamma(s)$ implies $\boldsymbol{\beta}_{\mathcal{A}}(> q_2, A) \in P_\gamma(s)$. The situation changes, however, when Demon enters the stage, because then $q \mapsto \Omega_{\mathcal{G}}(\gamma^d \mid A, Q)$ is increasing. So $q \mapsto \Omega_{\mathcal{G}}(g \mid A, q)$ can be guaranteed to be monotonically decreasing only if g belongs to the PDL fragment (shattering the hope to simplify some arguments due to monotonicity.

It has to be established for every game g that $\Omega_{\mathcal{G}}(g \mid A, q) \in \mathcal{A}$, provided $A \in \mathcal{A}$. We look at different cases, but we do have some preparations first. They address the embedded integration which is found in case ⑤. The technical property on measurability established in Corollary 4.9.9 has an immediate consequence.

Corollary 4.9.29 *Let $P : S \to \boldsymbol{E}(S)$ be a stochastic effectivity function, and assume $B \in \mathcal{A} \otimes \mathcal{B}([0,1])$. Put*

$$G := \{\langle \mu, q\rangle \in \mathbb{S}(S) \times [0,1] \mid \int_0^1 \mu(B^r)\, dr \bowtie q\}.$$

Then $\{\langle s, q\rangle \in S \times [0,1] \mid G^q \in P(s)\} \in \mathcal{A} \otimes \mathcal{B}([0,1])$.

Proof $G \in \wp(S) \otimes \mathcal{B}([0,1])$ by Corollary 4.9.9, so the assertion follows from t-measurability. ⊣

Perceiving the set $\Omega_{\mathcal{G}}(g \mid A, r)$ as the result of transforming set A through game g, we may say something about the effect of the transformation which is induced by game $\gamma; g$ with $\gamma \in \Gamma$. We need this technical property for making sure that we do not leave the kingdom of measurable sets while playing our games.

Lemma 4.9.30 *Let g be a game such that $\{\langle s, r\rangle \in S \times [0,1] \mid s \in \Omega_{\mathcal{G}}(g \mid A, r)\}$ is a measurable subset of $S \times [0,1]$, and assume that $\gamma \in \Gamma$. Then*

$$\{\langle s, r\rangle \in S \times [0,1] \mid s \in \Omega_{\mathcal{G}}(\gamma; g \mid A, r)\}$$

is a measurable subset of $S \times [0,1]$.

Proof This follows from Corollary 4.9.29, because P_γ is t-measurable. ⊣

The transformation associated with the indefinite iteration in ⑥ above involves an uncountable intersection, since for $q > 0$ the set $Q^{(\infty)}(q)$

has the cardinality of the continuum. Since σ-algebras are closed only under countable operations, we might in this way generate a set which is not measurable at all, unless we do take cautionary measures into account.

If the state space is closed under the *Souslin operation*, which is discussed in Sect. 4.5, then the resulting set will still be measurable. A fairly popular class of spaces closed under this operation is the class of universally complete measurable spaces; see Proposition 4.6.4. The class of analytic sets in a Polish space is closed under the Souslin operation as well; see Proposition 4.5.1 together with Theorem 4.5.5 (but we have to be careful here, because the class of analytic sets in *not* closed under complementation and demonic argumentation introduces complements).

In order to deal with indefinite iteration in ⑥, we establish first that we can encode the elements of $Q^{(\infty)}(q)$ conveniently through infinite sequences over $\mathbb{N}_0$. This will then serve as an encoding, so that we obtain $\Omega_{\mathcal{G}}(g^*; g_0 \mid A, q)$ as the result of a Souslin scheme, as defined on page 517.

Recall that $\cdot|n$ is the truncation operator which takes the first n elements of an infinite sequence or a sequence of length greater than n.

Lemma 4.9.31 *There exists for $q > 0$ a bijection $f : \mathbb{N}_0^{\mathbb{N}} \to Q^{(\infty)}(q)$ such that $\alpha|n = \alpha'|n$ implies $f(\alpha)|n = f(\alpha')|n$ for all $n \in \mathbb{N}$ and all $\alpha, \alpha' \in \mathbb{N}_0^{\mathbb{N}}$.*

Idea for a proof

Proof 0. Let us look at the idea first. Define $D_n := \{a|n \mid a \in Q^{(\infty)}(q)\}$ as the set of all truncated sequences in $Q^{(\infty)}(q)$ of length n. Since $q > 0$, the set $D_1 = [0, q] \cap \mathbb{Q}$ is not empty and countable; hence we can find a bijection $\ell_1 : D_1 \to \mathbb{N}_0$. Let us see what happens when we consider D_2. If $\langle a_1, a_2 \rangle \in D_2$, we consider two cases: $a_1 + a_2 = q$ and $a_1 + a_2 < q$. In the former case, we put $\ell_2(a_1, a_2) := \langle \ell_1(a_1), 0 \rangle$, and in the latter case we know that $[0, q - (a_1 + a_2)] \cap \mathbb{Q} \neq \emptyset$ is countable; thus we find a bijection $h_{a_1,a_2} : [0, q - (a_1 + a_2)] \cap \mathbb{Q} \to \mathbb{N}_0$, so we put $\ell_2(a_1, a_2) := \langle \ell_1(a_1), h_{a_1,a_2}(a_2) \rangle$. This yields a bijection, and the projection of ℓ_2 equals ℓ_1. The inductive step is done through the same argumentation. From this construction we piece together a bijection $Q^{(\infty)}(q) \to \mathbb{N}_0^{\mathbb{N}}$, which will be inverted to give the function we are looking for.

1. We construct first inductively for each $n \in \mathbb{N}$ a bijection $\ell_n : D_n \to \mathbb{N}_0^n$ such that $\ell_{n+1}(w)|n = \ell_n(w|n)$ holds for all $w \in D_{n+1}$. $\ell_1 : D_1 = [0,q] \cap \mathbb{Q} \to \mathbb{N}_0$ is an arbitrary bijection, and if ℓ_n is already defined, put for $w = \langle w_1, \ldots, w_{n+1}\rangle \in D_{n+1}$

$$\ell_{n+1}(w) := \begin{cases} \langle \ell_n(w_1, \ldots, w_n), 0\rangle, & \text{if } w_1 + \ldots + w_n = q, \\ \langle \ell_n(w_1, \ldots, w_n), \\ \quad h_{w_1,\ldots,w_n}(w_{n+1})\rangle, & \text{otherwise,} \end{cases}$$

where $h_{w_1,\ldots,w_n} : [0, q - (w_1 + \ldots + w_n)] \cap \mathbb{Q} \to \mathbb{N}$ is a bijection. An easy inductive argument shows that ℓ_{n+1} is bijective, if ℓ_n is, and that the projectivity condition $\ell_{n+1}(w)|n = \ell_n(w|n)$ holds for each $w \in D_{n+1}$.

2. Define $\ell_\infty(a)|n := \ell_n(a|n)$ for $a \in Q^{(\infty)}(q)$ and $n \in \mathbb{N}$; then the projectivity condition makes sure that $\ell_\infty : Q^{(\infty)}(q) \to \mathbb{N}_0^{\mathbb{N}}$ is well defined. Because each ℓ_n is a bijection, ℓ_∞ is. Assume that $\alpha|n = \alpha'|n$ holds for $\alpha, \alpha' \in \mathbb{N}_0^{\mathbb{N}}$ and for some $n \in \mathbb{N}$, then $\alpha = \ell_\infty(a), \alpha' = \ell_\infty(a')$ for some $a, a' \in Q^{(\infty)}(q)$. Thus $\alpha|n = \ell_\infty(a)|n = \ell_n(a|n)$; hence $a|n = a'|n$, since ℓ_n is injective.

3. Now let f be the inverse to ℓ_∞, then the assertion follows. $\dashv$

After this technical preparation, we are in a position to establish this important property.

Proposition 4.9.32 *Let g and g_0 be games such that $\Omega_{\mathcal{G}}(g^n; g_0 \mid A, r)$ is a measurable subset of S for each $n \in \mathbb{N}$ and each rational $r \in [0,1]$. Assume that the measurable space S is closed under the Souslin operation. Then $\Omega_{\mathcal{G}}(g^*; g_0 \mid A, q)$ is a measurable subset of S for each $n \in \mathbb{N}$ and each rational $q \in [0,1]$.*

Proof 0. The proof just encodes the representation of $\Omega_{\mathcal{G}}(g^*; g_0 \mid A, q)$ so that a Souslin scheme arises.

1. Let $f : \mathbb{N}_0^{\mathbb{N}} \to Q^{(\infty)}(q)$ define the encoding Lemma 4.9.31. Write

$$\begin{aligned}\Omega_{\mathcal{G}}(g^*; g_0 \mid A, q) &= \bigcap_{a \in Q^{(q)}(\infty)} \bigcup_{n \geq 0} \Omega_{\mathcal{G}}(g^n; g_0 \mid A, a_{n+1}) \\ &= \bigcap_{\alpha \in \mathbb{N}_0^{\mathbb{N}}} \bigcup_{n \geq 0} C_{\alpha|n}\end{aligned}$$

with $C_{\alpha|n} := \Omega_{\mathcal{G}}(g^n; g_0 \mid A, f(\alpha)_{n+1})$. $\dashv$

Thus we know that the construction above does not lead us outside the realm of measurable sets, provided the space is closed under the Souslin operation. We may think of $A \mapsto \Omega_{\mathcal{G}}(g \mid A, q)$ as a set transformation. But this applies for the time being only to those games the syntactic pattern is given by one of the cases above, and we have to investigate now whether we can interpret each game that way.

Interpretable game

Call the game g *interpretable* iff $\Omega_{\mathcal{G}}(g \mid A, q)$ is defined for each $A \in \mathcal{A}$, $q \in [0, 1]$ rational so that the set

$$Gr(g, A) := \{\langle s, q\rangle \in S \times [0, 1] \mid s \in \Omega_{\mathcal{G}}(g \mid A, q)\}$$

is a measurable subset of $S \times [0, 1]$. Note that we exclude for the time being the test operator.

Lemma 4.9.33 *Each game g is interpretable, provided the state space is closed under the Souslin operation.*

Strategy

Proof 0. The proof employs a kind of inductive strategy. We show first that all primitive games are interpretable and that the dual of an interpretable game is interpretable again. Then we go through the different cases and have a look at what happens.

1. Let

$$J := \{g \mid g \text{ is interpretable}\}.$$

We show that J contains all games.

2. Let $A \in \mathcal{A}$; then

$$\begin{aligned}\{\langle \mu, q\rangle \in \mathbb{S}(S) \times [0, 1] \mid \mu \in \boldsymbol{\beta}_{\mathcal{A}}(A, > q)\}\\ = \{\langle \mu, q\rangle \in \mathbb{S}(S) \times [0, 1] \mid \mu(A) > q\}\end{aligned}$$

is a measurable subset of $\mathbb{S}(S) \times [0, 1]$; see Corollary 4.2.5. Thus, if $\gamma \in \Gamma$, we have

$$Gr(\gamma, A) = \{\langle s, q\rangle \in S \times [0, 1] \mid \boldsymbol{\beta}_{\mathcal{A}}(A, > q)\},$$

which is a measurable subset of $S \times [0, 1]$, because P_γ is t-measurable. Consequently, we obtain $\Gamma \subseteq J$.

3. Clearly, J is closed under demonization and angelic choice and hence under demonic choice as well. Now let

$$L := \{g \mid g; g_1 \text{ is interpretable for all interpretable } g_1\}.$$

Then Lemma 4.9.30 implies that $\Gamma \cup \{\gamma^d \mid \gamma \in \Gamma\} \subseteq L$. Moreover, because angelic choice distributes from the left over composition, L is closed under angelic choice. It is also closed under demonization: Let $g \in L$ and take an interpretable game g_1; then $g; g_1^d$ is interpretable, and thus the interpretation of $(g; g_1^d)^d$ is defined; hence $g^d; g_1$ is interpretable. Clearly, L is closed under composition. Thus $g \in L$ implies $g^* \in L$ as well as $g^\times \in L$, so that L is the set of all games.

This implies that J is closed under composition, and hence both are under angelic and demonic iteration. Consequently, J contains all games. $\dashv$

Summarizing, we obtain

Proposition 4.9.34 *Assume that the state space is closed under the Souslin operation; then we have $\Omega_{\mathcal{G}}(g \mid A, q) \in \mathcal{A}$ for all games g, $A \in \mathcal{A}$ and $0 \leq q \leq 1$.*

Proof We infer for each game g from Lemma 4.9.33 that the set $Gr(g, A)$ is a measurable subset of $S \times [0, 1]$ for $A \in \mathcal{A}$. But $\Omega_{\mathcal{G}}(g \mid A, q) = Gr(g, A)^q$. $\dashv$

We relate game frames to each other through morphisms. Suppose that $\mathcal{H} = \big(T, (Q_\gamma)_{\gamma \in \Gamma}\big)$ is another game frame; then $f : \mathcal{G} \to \mathcal{H}$ is a *game frame morphism* iff $f : P_\gamma \to Q_\gamma$ is a morphism for the associated effectivity functions for all $\gamma \in \Gamma$; these morphisms are defined in Definition 4.1.25. Game frame morphism

The transformations above are compatible with frame morphisms.

Proposition 4.9.35 *Let $f : \mathcal{G} \to \mathcal{H}$ be a game frame morphism, and assume that $\Omega_{\mathcal{G}}(g \mid \cdot, q)$ and $\Omega_{\mathcal{H}}(g \mid \cdot, q)$ always transform measurable sets into measurable sets for all games g and all q. Then we have*

$$f^{-1}\big[\Omega_{\mathcal{H}}(g \mid B, q)\big] = \Omega_{\mathcal{G}}(g \mid f^{-1}\big[B\big], q)$$

for all games g, all measurable sets $B \in \mathcal{B}(T)$, and all q.

Proof 0. The proof proceeds by induction on g. Because f is a morphism, the assertion is true for $g = \gamma \in \Gamma$. Because f^{-1} is compatible with the Boolean operations on sets, it is sufficient to consider the case $g = \gamma; g_1$ in detail. It wants essentially the computation of an image measure, and it looks worse than it really is.

1. Assume that the assertion is true for game g_1, fix $B \in \mathcal{B}, q \geq 0$. Then

$$\begin{aligned}
&G_{g_1,\mathcal{G}}(f^{-1}\big[B\big], q)\\
&\quad= \{\mu \in \mathbb{S}(S) \mid \int_0^1 \mu\big(\Omega_{\mathcal{G}}(g_1 \mid f^{-1}\big[B\big], r)\big)\, dr > q\}\\
&\quad\overset{(*)}{=} \{\mu \in \mathbb{S}(S) \mid \int_0^1 \mu\big(f^{-1}\big[\Omega_{\mathcal{G}}(g_1 \mid B, r)\big]\big)\, dr > q\}\\
&\quad\overset{(\oplus)}{=} \{\mu \in \mathbb{S}(S) \mid \int_0^1 \mathbb{S}f(\mu)\big(\Omega_{\mathcal{H}}(g_1 \mid B, r)\big)\, dr > q\}\\
&\quad= (\mathbb{S}f)^{-1}\big[\{\nu \in \mathbb{S}(T) \mid \int_0^1 \nu\big(\Omega_{\mathcal{H}}(g_1 \mid f^{-1}\big[B\big], r)\big)\, dr > q\}\big]\\
&\quad= (\mathbb{S}f)^{-1}\big[G_{g_1,\mathcal{H}}(B, q)\big].
\end{aligned}$$

The equation $(*)$ derives from the induction hypothesis, and $(\oplus)$ from the definition of $(\mathbb{S}f)(\mu)$.

2. Because $f : P_\gamma \to Q_\gamma$ is a morphism, we obtain from the first part

$$\begin{aligned}
\Omega_{\mathcal{G}}(\gamma; g_1 \mid f^{-1}\big[B\big], q) &= \{s \in S \mid G_{g_1,\mathcal{G}}(f^{-1}\big[B\big], q) \in P_\gamma(s)\}\\
&= \{s \in S \mid (\mathbb{S}f)^{-1}\big[G_{g_1,\mathcal{H}}(B, q)\big] \in P_\gamma(s)\}\\
&= \{s \in S \mid G_{g_1,\mathcal{H}}(B, q) \in Q_\gamma(f(s))\}\\
&= f^{-1}\big[\Omega_{\mathcal{H}}(\gamma; g_1 \mid B, q)\big].
\end{aligned}$$

This shows that the assertion is also true for $g = \gamma; g_1$. $\dashv$

Let us briefly interpret Proposition 4.9.35 in terms of natural transformations, providing some coalgebraic *sfumatura*.

Example 4.9.36 We extract in this example the measurable sets $\mathcal{A}$ from a measurable space $(S, \mathcal{A})$ which serves as a state space for our discussions. Define $\boldsymbol{T}(S) := \mathcal{A}$; then $\boldsymbol{T}$ acts as a contravariant functor from the category of measurable spaces satisfying the Souslin condition to the category of sets, where the measurable map $f : S \to T$ is mapped to $\boldsymbol{T}f : \boldsymbol{T}(T) \to \boldsymbol{T}(S)$ by $\boldsymbol{T}f := f^{-1}$.

Fix a game g and a real $q \in [0, 1]$, then $\Omega_{\mathcal{G}}(g \mid \cdot, q) : \boldsymbol{T}(S) \to \boldsymbol{T}(S)$ by Proposition 4.9.34, provided S satisfies the Souslin condition. Then

$\Omega_{\mathcal{G}}(g \mid \cdot, q)$ induces a natural transformation $\boldsymbol{T} \to \boldsymbol{T}$, because this diagram commutes by Proposition 4.9.35:

$$\begin{tikzcd} S \arrow[d, "f"'] & & \boldsymbol{T}(T) \arrow[r, "f^{-1}"] \arrow[d, "\Omega_{\mathcal{H}}(g|\cdot,q)"'] & \boldsymbol{T}(S) \arrow[d, "\Omega_{\mathcal{G}}(g|\cdot,q)"] \\ T & & \boldsymbol{T}(T) \arrow[r, "f^{-1}"'] & \boldsymbol{T}(S) \end{tikzcd}$$

✌

Kripke Generated Frames

A *stochastic Kripke frame* $\mathcal{K} = \big(S, (K_\gamma)_{\gamma \in \Gamma}\big)$ is a measurable state space S, and each primitive game $\gamma \in \Gamma$ is associated with a stochastic relation $K_\gamma : S \rightsquigarrow S$; see Example 4.2.7. Morphisms carry over in the obvious fashion from stochastic relations to stochastic Kripke frames by applying the defining condition to the stochastic relation associated with each primitive game; see Sect. 4.1.3.

We associate with $\mathcal{K}$ a game frame $\mathcal{G}_{\mathcal{K}} := \big(S, (P_{K_\gamma})_{\gamma \in \Gamma}\big)$. Thus the transformations associated with games considered above are also applicable to Kripke models. We will discuss this shortly. An application of Proposition 4.1.24 for each $\gamma \in \Gamma$ states under which conditions a game frame is generated by a stochastic Kripke frame. Just for the record

Proposition 4.9.37 *Let $\mathcal{G} = \big(S, (P_\gamma)_{\gamma \in \Gamma}\big)$ be a game frame. Then these conditions are equivalent:*

1. *There exists a stochastic game frame $\mathcal{K}$ with $\mathcal{G} = \mathcal{G}_{\mathcal{K}}$.*
2. *$R_\gamma(s) := \{\langle r, A\rangle \mid \boldsymbol{\beta}_{\mathcal{A}}(A, \geq r) \in P_\gamma(s)\}$ defines a characteristic relation on S such that $P_\gamma(s) \vdash R_\gamma(s)$ for each state $s \in S, \gamma \in \Gamma$.* ⊣

Let $\mathcal{K} = (S, (K_\gamma)_{\gamma \in \Gamma})$ be a Kripke frame with associated game frame $\mathcal{G}_{\mathcal{K}}$. $K_\gamma : S \rightsquigarrow S$ are Kleisli morphisms; their product—also known as convolution (see Example 4.9.6)—is defined through

$$(K_{\gamma_1} * K_{\gamma_2})(s)(A) := \int_S K_{\gamma_2}(t)(A)\ K_{\gamma_1}(s)(dt);$$

see Example 4.9.6. Intuitively, this gives the probability of reaching a state in $A \in \mathcal{A}$, provided we start with game γ_1 in state s and continue with game γ_2, averaging over intermediate states (here $\gamma_1, \gamma_2 \in \Gamma$). The observation that composing stochastic relations models the composition of modalities is one of the cornerstones for the interpretation of modal logics through Kripke models [Pan09, Dob09].

Let $\Omega_{\mathcal{G}}(A \mid g, q)$ be defined as above when working in the game frame associated with Kripke frame $\mathcal{K}$. It turns out that $\Omega_{\mathcal{G}}(A \mid \gamma_1; \ldots; \gamma_k, q)$ can be described in terms of the Kleisli product for $K_{\gamma_1}, \ldots, K_{\gamma_k}$, provided $\gamma_1, \ldots, \gamma_k \in \Gamma$ are primitive games.

Proposition 4.9.38 *Assume that $\gamma_1, \ldots, \gamma_k \in \Gamma$; then this equality holds in the game frame $\mathcal{G}$ associated with the Kripke frame $\mathcal{K}$:*

$$\Omega_{\mathcal{G}}(A \mid \gamma_1; \ldots \gamma_k, q) = \{s \in S \mid (K_{\gamma_1} * \ldots * K_{\gamma_k})(s)(A) > q\}$$

for all $A \in \mathcal{A}, 0 \leq q < 1$.

Proof 1. The proof proceeds by induction on k. If $k = 1$, we have

$$s \in \Omega_{\mathcal{G}}(A \mid \gamma_1, q) \Leftrightarrow \boldsymbol{\beta}_{\mathcal{A}}(A, > q) \in P_{\mathcal{K},\gamma_1}(s) \Leftrightarrow K_{\gamma_1}(s)(A) > q.$$

2. Assume that the claim is established for k, and let $\gamma_0 \in \Gamma$. Then, borrowing the notation from above,

$$\begin{aligned}
&s \in \Omega_{\mathcal{G}}(A \mid \gamma_0; \gamma_1; \ldots; \gamma_k, q) \\
&\quad \Leftrightarrow G_{\gamma_1;\ldots;\gamma_k}(A, q) \in P_{\mathcal{K},\gamma_0}(s) \\
&\quad \Leftrightarrow K_{\gamma_0}(s) \in G_{\gamma_1;\ldots;\gamma_k}(A, q) \\
&\quad \Leftrightarrow \int_0^1 K_{\gamma_0}(s)\big(\Omega_{\mathcal{G}}(A \mid \gamma_1; \ldots; \gamma_k, r)\big)\, dr > q \\
&\quad \overset{(\dagger)}{\Leftrightarrow} \int_0^1 K_{\gamma_0}(s)\big(\{t \mid (K_{\gamma_1} * \ldots * K_{\gamma_k})(t)(A) > r\}\big)\, dr > q \\
&\quad \overset{(\ddagger)}{\Leftrightarrow} \int_S (K_{\gamma_1} * \ldots * K_{\gamma_k})(t)(A)\; K_{\gamma_0}(s)(dt) > q \\
&\quad \overset{(\|)}{\Leftrightarrow} (K_{\gamma_0} * K_{\gamma_1} * \ldots * K_{\gamma_k})(s)(A) > q.
\end{aligned}$$

Here (†) marks the induction hypothesis, (‡) is just the integral representation given in Eq. (4.11) in Example 4.9.7, and (‖) is the definition of the Kleisli product. This establishes the claim for $k + 1$. ⊣

We note as a consequence that the respective definitions of state transformations through the games under consideration coincide for game frames generated by Kripke frames. On the other hand, it is noted that the definition of these transformations for general frames based on effectivity functions extends the one which has been used for Kripke frames for general modal logics.

Distributivity in the PDL Fragment

The games which are described through the grammar from Example 2.7.4

$$t ::= \psi \mid t_1 \cup t_2 \mid t_1; t_2 \mid t^* \mid \varphi?$$

with $\psi \in \Psi := \Gamma$ as the set of atomic programs and φ a formula of the underlying modal logic define the *PDL fragment* of game logic. The corresponding games are called *programs* for simplicity. We will show now that in this fragment

$$\Omega_{\mathcal{G}}(\cdot \mid g_1; (g_2 \cup g_3), \cdot) = \Omega_{\mathcal{G}}(\cdot \mid g_1; g_2 \cup g_1; g_3), \cdot)$$

holds, provided frame $\mathcal{G}$ is generated by a stochastic Kripke frame $\mathcal{K}$.

Recall that $\mathbb{M}_\sigma(S, \mathcal{A})$ is the set of σ-finite measures on $(S, \mathcal{A})$ to the extended nonnegative reals. $\mathbb{M}_\sigma(S)$ is closed under addition and under multiplication with nonnegative reals (we omit the σ-algebra in the sequel). The set is also closed under countable sums: Given $(\mu_n)_{n\in\mathbb{N}}$ with $\mu_n \in \mathbb{M}_\sigma(S)$, put

$$\bigl(\textstyle\sum_{n\in\mathbb{N}} \mu_n\bigr)(A) := \sup_{n\in\mathbb{N}} \sum_{i\leq n} \mu_i(A).$$

Then $\sum_{n\in\mathbb{N}} \mu_n$ is monotone and σ-additive with $\bigl(\sum_{n\in\mathbb{N}} \mu_n\bigr)(\emptyset) = 0$, hence a member of $\mathbb{M}_\sigma(S)$.

Call a map $N : S \to \mathbb{M}_\sigma(S)$ an *extended kernel* iff for each $A \in \mathcal{A}$ the map $s \mapsto N(s)(A)$ is measurable with the usual conventions regarding measurability to the extended reals $\widetilde{\mathbb{R}}$; see Sect. 4.9.1. Extended kernels are closed under convolution: Put

$$(N_1 * N_2)(s)(A) := \int_S N_2(t)(A)\ N_1(s)(dt);$$

then $N_1 * N_2$ is an extended kernel again. This is but the Kleisli composition applied to extended kernels. Thus $\mathbb{M}_\sigma(S)$ is closed under convolution which distributes both from the left and from the right under addition and under scalar multiplication. Note that the countable sum of extended kernels is an extended kernel as well.

Define recursively for the stochastic relations in the Kripke frame $\mathcal{K}$

$$\begin{aligned} K_{g_1 \cup g_2} &:= K_{g_1} + K_{g_2}, \\ K_{g_1;g_2} &:= K_{g_1} * K_{g_2}, \\ K_{g^*} &:= \sum_{n\geq 0} K_{g^n}. \end{aligned}$$

This defines K_g for each g in the PDL fragment of game logic, similar to the proposal in [Koz85].

Define

$$\mathcal{L}(A \mid g, r) := S \setminus \Omega_{\mathcal{G}}(A \mid g, r),$$

where $\mathcal{G}$ is the game frame associated with the Kripke frame $\mathcal{K}$ over state space S; $A \in \mathcal{A}$ is a measurable set; g is a program, i.e., a member of the PDL fragment; and $r \in [0, 1]$. It is more convenient to work with these complements, as we will see in a moment.

Lemma 4.9.39 $\mathcal{L}(A \mid g, r) = \{s \in S \mid K_g(s)(A) \leq r\}$ *holds for all programs g, all measurable sets $A \in \mathcal{A}$, and all $r \in [0, 1]$.*

Proof 1. The proof is fairly straightforward and proceeds by induction on g. Assume that $g = \gamma \in \Gamma$ is a primitive program; then

$$\begin{aligned} K_\gamma(s)(A) \leq r \Leftrightarrow K_\gamma(s) \notin \boldsymbol{\beta}_{\mathcal{A}}(A, > r) \Leftrightarrow \boldsymbol{\beta}_{\mathcal{A}}(A, > r) \notin P_\gamma(s) \\ \Leftrightarrow s \notin \Omega_{\mathcal{G}}(A \mid g, r). \end{aligned}$$

2. Assume that the assertion is true for g_1 and g_2; then

$$\begin{aligned} &\mathcal{L}(A \mid g_1 \cup g_2, r) \\ &\quad = \bigcap_{\langle a_1, a_2\rangle \in Q^{(k)}(r)} \big(\mathcal{L}(A \mid g_1, a_1) \cap \mathcal{L}(A \mid g_2, a_2)\big) \\ &\quad = \bigcap_{\langle a_1, a_2\rangle \in Q^{(k)}(r)} \big(\{s \mid K_{g_1}(s)(A) \leq a_1\} \cap \{s \mid K_{g_2}(s)(A) \leq a_2\}\big) \\ &\quad = \{s \in S \mid (K_{g_1} + K_{g_2})(s)(A) \leq r\} \\ &\quad = \{s \in S \mid K_{g_1 \cup g_2}(s)(A) \leq r\}. \end{aligned}$$

3. The proof for angelic iteration g^* is very similar, observing that $\sum_{n\geq 0} K_{g^n}(s)(A) \leq r$ iff there exists a sequence $(a_n)_{n\in\mathbb{N}} \in Q^{(\infty)}(r)$ with $K_{g^n}(s)(A) \leq a_n$ for all $n \in \mathbb{N}$.

4. Finally, assume that the assertion is true for program g, and take $\gamma \in \Gamma$. Then, borrowing the notation from (4.14)

$$\begin{aligned}
G_g(A, r) \notin P_\gamma(s) &\Leftrightarrow K_\gamma(s) \notin G_g(A, r)\\
&\Leftrightarrow \int_0^1 K_\gamma(s)(\Omega_{\mathcal{G}}(A \mid g, t))\, dt \leq r\\
&\overset{(\dagger)}{\Leftrightarrow} \int_0^1 K_\gamma(s)(\{x \in S \mid K_g(x)(A) > t\})\, dt \leq r\\
&\overset{(\ddagger)}{\Leftrightarrow} \int_S K_g(t)(A)\; K_\gamma(s)(dt) \leq r\\
&\overset{(*)}{\Leftrightarrow} K_{\gamma;g}(s)(A) \leq r.
\end{aligned}$$

Here (†) is the induction hypothesis, (‡) derives from Eq. (4.11) in Example 4.9.7, and (∗) comes from the definition of the convolution. ⊣

It follows from this representation that for each program g the set $\mathcal{L}(A \mid g^*, r)$ is a measurable subset of S, provided $A \in \mathcal{A}$. This holds even without the assumption that the state space S is closed under the Souslin operation.

Proposition 4.9.40 *If games g_1, g_2, g_3 are in the PDL fragment and the game frame $\mathcal{G}$ is generated by a Kripke frame, then*

$$\Omega_{\mathcal{G}}(A \mid g_1; (g_2 \cup g_3), r) = \Omega_{\mathcal{G}}(A \mid g_1; g_2 \cup g_1; g_3, r) \tag{4.15}$$

$$\Omega_{\mathcal{G}}(A \mid (g_1 \cup g_2); g_3, r) = \Omega_{\mathcal{G}}(A \mid g_1; g_3 \cup g_2; g_3, r) \tag{4.16}$$

for all $A \in \mathcal{A}, r \geq 0$.

Proof Right distributivity (4.16) is a basic assumption, which is given here for the sake of completeness. It remains to establish left distributivity (4.15). Here it suffices to prove the equality for the respective complements. But this is easily established through Lemma 4.9.39 and the observation that $K_{g_1;(g_2 \cup g_3)} = K_{g_1;g_2} + K_{g_1;g_3}$ holds, because integration of nonnegative functions is additive. ⊣

Game Models

We discuss briefly game models, because we need a model for the discussion of the test operator, which has been delayed so far. A game model describes the semantics of a game logic, which in turn is a modal logic where the modalities are given through games. It is defined through grammar

$$\varphi = \top \mid p \mid \varphi_1 \wedge \varphi_2 \mid \langle g \rangle_q \varphi;$$

see Example 4.1.11. Here $p \in \Psi$ is an atomic proposition, g is a game, and $q \in [0, 1]$ is a real number. Intuitively, formula $\langle g \rangle_q \varphi$ is true in state s if playing game g in state s will result in a state in which formula φ holds with a probability greater than q.

Definition 4.9.41 *A game model* $\mathcal{G} = \big((S, \mathcal{A}), (P_\gamma)_{\gamma\in\Gamma}, (V_p)_{p\in\Psi}\big)$ *over the measurable space* S *is given by a game frame* $\big((S, \mathcal{A}),$ $(P_\gamma)_{\gamma\in\Gamma}\big)$ *and by a family* $(V_p)_{p\in\Psi} \subseteq \mathcal{A}$ *of sets which assigns to each atomic statement a measurable set of state space* S*. We denote the underlying game frame by* $\mathcal{G}$ *as well.*

$[\![\varphi]\!]_{\mathcal{G}}$ Define the validity sets for each formula recursively as follows:

$$\begin{aligned}
[\![\top]\!]_{\mathcal{G}} &:= S \\
[\![p]\!]_{\mathcal{G}} &:= V_p, \text{ if } p \in \Psi \\
[\![\varphi_1 \wedge \varphi_2]\!]_{\mathcal{G}} &:= [\![\varphi_1]\!]_{\mathcal{G}} \cap [\![\varphi_2]\!]_{\mathcal{G}} \\
[\![\langle g \rangle_q \varphi]\!]_{\mathcal{G}} &:= \Omega_{\mathcal{G}}([\![\varphi]\!]_{\mathcal{G}} \mid g, q).
\end{aligned}$$

Accordingly, we say that formula φ holds in state s, in symbols $\mathcal{G}, s \models$ $\mathcal{G}, s \models \varphi$ φ, iff $s \in [\![\varphi]\!]_{\mathcal{G}}$.

The definition of $[\![\langle g \rangle_q \varphi]\!]_{\mathcal{G}}$ has a coalgebraic flavor. Coalgebraic logics define the validity of modal formulas through special *predicate liftings* associated with the modalities; see Sect. 2.7.3. This connection becomes manifest through Example 4.9.36 where $\Omega_{\mathcal{G}}(\cdot \mid g, q)$ is shown to be a natural transformation.

Proposition 4.9.42 *If state space* S *is closed under the Souslin operation,* $[\![\varphi]\!]_{\mathcal{G}}$ *is a measurable subset for all formulas* φ*. Moreover,* $\{\langle s, r\rangle \mid s \in [\![\langle g \rangle_r \varphi]\!]_{\mathcal{G}}\} \in \mathcal{A} \otimes [0, 1]$.

Proof The proof proceeds by induction on the formula φ. If $\varphi = p \in \Psi$ is an atomic proposition, then the assertion follows from $V_p \in \mathcal{A}$. The straightforward induction step uses Lemma 4.9.30. $\dashv$

Stochastic Kripke models are defined similarly to game models: $\mathcal{K} = \big((S, \mathcal{A}), (K_\gamma)_{\gamma\in\Gamma}, (V_p)_{p\in\Psi}\big)$ is called a *stochastic Kripke model* iff $\big((S, \mathcal{A}), (K_\gamma)_{\gamma\in\Gamma}\big)$ is a stochastic Kripke frame with $V_p \in \mathcal{A}$ for each atomic proposition $p \in \Psi$. The validity of a formula in the state of a stochastic Kripke model is defined as validity in the associated game model. Thus we know that for primitive games $\gamma_1, \ldots, \gamma_n \in \Gamma$

$$\begin{aligned} s \in [\![\langle \gamma_1; \ldots; \gamma_n \rangle_q \varphi]\!]_{\mathcal{G}} &\Leftrightarrow \mathcal{K}, s \models \langle \gamma_1; \ldots; \gamma_n \rangle_q \varphi \\ &\Leftrightarrow \big(K_{\gamma_1} * \ldots * K_{\gamma_n}\big)(s)([\![\varphi]\!]_{\mathcal{K}}) \geq q \end{aligned} \tag{4.17}$$

holds (Proposition 4.9.38) and that games are semantically equivalent to their distributive counterparts (Proposition 4.9.40).

Let $\mathcal{H} = \big(T, (Q_\gamma)_{\gamma \in \Gamma}, (W_p)_{p \in \Psi}\big)$ be a second game model; then a measurable map $f : S \to T$ which is also a frame morphism $f : (S, (P_\gamma)_{\gamma \in \Gamma}) \to (T, (Q_\gamma)_{\gamma \in \Gamma})$ is called a *model morphism* $f : \mathcal{G} \to \mathcal{H}$ iff $f^{-1}\big[W_p\big] = V_p$ holds for all atomic propositions, i.e., if $f(s) \in W_p$ iff $s \in V_p$ always holds. Model morphisms are compatible with validity. Model morphism

Proposition 4.9.43 *Let φ be a formula of game logic and $f : \mathcal{G} \to \mathcal{H}$ be a model morphism. Then*

$$\mathcal{G}, s \models \varphi \textit{ iff } \mathcal{H}, f(s) \models \varphi.$$

Proof The claim is equivalent to saying that

$$[\![\varphi]\!]_{\mathcal{G}} = f^{-1}\big[[\![\varphi]\!]_{\mathcal{H}}\big]$$

for all formulas φ. This is established through induction on the formula φ. Because f is a model morphism, the assertion holds for atomic proposition. The induction step is established through Proposition 4.9.35. $\dashv$

We assume from now on that the respective state spaces of our models are closed under the Souslin operation.

The Test Operator

The test operator $\varphi?$ may be incorporated now. Given a formula φ, Angel may test whether or not the formula is satisfied; this yields the two games $\varphi?$ and $\varphi¿$. Game $\varphi?$ checks whether formula φ is satisfied in the current state; if it is, Angel continues with the next game; if it is not, Angel loses. Similarly for $\varphi¿$, Angel checks whether formula φ is not satisfied. Note that we do not have negation in our logic, so we $\varphi?, \varphi¿$

cannot test directly for $\neg\varphi$. We can test, however, whether a state state s does not satisfy a formula φ by evaluating $s \in S \setminus [\![\varphi]\!]_{\mathcal{G}}$, since complementation is available in the underlying σ-algebra. So we actually extend our considerations by incorporating two test operators, but talk for convenience usually about "the" test operator.

In order to seamlessly integrate these testing games into our models, we define for each formula two effectivity functions for positive and for negative testing, resp. The following technical observations will be helpful. It helps transporting stochastic effectivity functions along measurable maps.

Lemma 4.9.44 *Let P be a stochastic effectivity function on S, and assume that $F : \mathbb{S}(S) \to \mathbb{S}(S)$ is measurable; then $P'(s) := \{W \in \wp(S) \mid F^{-1}\big[W\big] \in P(s)\}$ defines a stochastic effectivity function on S.*

Proof $P'(s)$ is upper closed, since $P(s)$ is, so t-measurability has to be established. Let $H \in \wp(S) \otimes \mathcal{B}([0,1])$ be a test set; then $H^q \in P'(s) \Leftrightarrow \big((F \times id_{[0,1]})^{-1}\big[H\big]\big)^q \in P(s)$. Since F is measurable, $F \times id_{[0,1]} : \mathbb{S}(S) \times [0,1] \to \mathbb{S}(S) \times [0,1]$ is; hence $(F \times id_{[0,1]})^{-1}\big[H\big]$ is a member of $\wp(S) \otimes \mathcal{B}([0,1])$. Because P is t-measurable, we conclude $\{\langle s, q\rangle \mid H^q \in P'(s)\} \in \mathcal{A} \otimes [0,1]$. ⊣

Localization

Lemma 4.9.45 *Define for $A \in \mathcal{A}$, $\mu \in \mathbb{S}(S)$, and $B \in \mathcal{A}$ the* localization *to A as $F_A(\mu)(B) := \mu(A \cap B)$. Then $F_A : \mathbb{S}(S) \to \mathbb{S}(S)$ is measurable.*

Proof This follows from $F_A^{-1}\big[\boldsymbol{\beta}_{\mathcal{A}}(C, \bowtie q)\big] = \boldsymbol{\beta}_{\mathcal{A}}(A \cap C, \bowtie q)$. ⊣

F_A localizes measures to the set A, because everything outside A is discarded. Now define for state s and formula φ

$$
\begin{aligned}
P_{\varphi ?}(s) &:= \{W \in \wp(S) \mid F^{-1}_{[\![\varphi]\!]_{\mathcal{G}}}\big[W\big] \in I_D(s)\},\\
P_{\varphi ¿}(s) &:= \{W \in \wp(S) \mid F^{-1}_{S\setminus[\![\varphi]\!]_{\mathcal{G}}}\big[W\big] \in I_D(s)\},
\end{aligned}
$$

where I_D is the Dirac function defined in Example 4.1.18. Let us decode the definition for $[\![\varphi]\!]_{\mathcal{G}}$. We have $W \in P_{\varphi ?}(s)$ iff $\delta_s \in F^{-1}_{[\![\varphi]\!]_{\mathcal{G}}}\big[W\big]$ and thus iff $F_{[\![\varphi]\!]_{\mathcal{G}}}(\delta_s) \in W$. Specializing to $W = \boldsymbol{\beta}_{\mathcal{A}}(A, > q)$ translates the latter condition to $F_{[\![\varphi]\!]_{\mathcal{G}}}(\delta_s)(A) > q$ and hence to $\delta_s([\![\varphi]\!]_{\mathcal{G}} \cap A) > q$, which in turn is equivalent to $\mathcal{G}, s \models \varphi$ and $s \in A$. Thus we see

$$\begin{aligned} \boldsymbol{\beta}_{\mathcal{A}}(A, > q) \in P_{\varphi ?}(s) &\Leftrightarrow \mathcal{G}, s \models \varphi \text{ and } s \in A, \\ \boldsymbol{\beta}_{\mathcal{A}}(A, > q) \in P_{\varphi ¿}(s) &\Leftrightarrow \mathcal{G}, s \not\models \varphi \text{ and } s \in A, \end{aligned}$$

the argument for $P_{\varphi ¿}$ being completely analogous. We obtain

Proposition 4.9.46 *$P_{\varphi ?}$ and $P_{\varphi ¿}$ define a stochastic effectivity function for each formula φ.*

Proof From Proposition 4.9.42, we infer that $[\![\varphi]\!]_{\mathcal{G}} \in \mathcal{A}$; consequently, $F_{[\![\varphi]\!]_{\mathcal{G}}}$ and $F_{S \setminus [\![\varphi]\!]_{\mathcal{G}}}$ are measurable functions $\mathbb{S}(S) \to \mathbb{S}(S)$ by Lemma 4.9.45. Thus the assertion follows from Lemma 4.9.44. $\dashv$

Given a stochastic Kripke model, test operators may be defined as well; they serve also for the integration of the test operators into PDL in a similar way. The definitions for the associated stochastic relations $K_{\varphi ?} : S \rightsquigarrow S$ and $K_{\varphi ¿} : S \rightsquigarrow S$ read

$$\begin{aligned} K_{\varphi ?}(s) &:= F_{[\![\varphi]\!]_{\mathcal{G}}}(D(s)), \\ K_{\varphi ¿}(s) &:= F_{S \setminus [\![\varphi]\!]_{\mathcal{G}}}(D(s)). \end{aligned}$$

These relations can be defined for formulas of game logic as well, when they are interpreted through a Kripke model. Thus we have, e.g.,

$$K_{\varphi ?}(s)(B) = \begin{cases} 1, & \text{if } s \in B \text{ and } \mathcal{G}, s \models \varphi \\ 0, & \text{otherwise.} \end{cases}$$

This is but a special case, since $P_{\varphi ?}$ and $P_{\varphi ¿}$ are generated by these stochastic relations.

Lemma 4.9.47 *Let φ be a formula of game logic; then $P_{\varphi ?} = P_{K_{\varphi ?}}$ and $P_{\varphi ¿} = P_{K_{\varphi ¿}}$.*

Proof The assertions follow from expanding the definitions. $\dashv$

This extension integrates well into the scenario, because it is compatible with morphisms. We will establish this now. Because a model morphism is given by a morphism for the underlying game frame and a morphism for game frames is determined by morphisms for the underlying effectivity functions, it is enough to show that a morphism $f : P_\gamma \to Q_\gamma$ for all $\gamma \in \Gamma$ is also a morphism $f : P_{\varphi ?} \to Q_{\varphi ?}$ for all formulas φ, similarly for $\varphi ¿$.

Proposition 4.9.48 *Let $\mathcal{G}$ and $\mathcal{H}$ be game models over state spaces S and T, resp. Assume that $f : \mathcal{G} \to \mathcal{H}$ is a morphism of game models; define for each formula φ the effectivity functions $P_{\varphi ?}$ and $P_{\varphi ¿}$ for $\mathcal{G}$ resp. $Q_{\varphi ?}$ and $Q_{\varphi ¿}$ for $\mathcal{H}$. Then f is a morphism $P_{\varphi ?} \to Q_{\varphi ?}$ and $P_{\varphi ¿} \to Q_{\varphi ¿}$ for each formula φ.*

Proof 0. Fix formula φ; we will prove the assertion only for $\varphi ?$, and the proof for $\varphi ¿$ is the same. The notation is fairly overloaded. We will use primed quantities when referring to $\mathcal{H}$ and state space T and unprimed ones when referring to model $\mathcal{G}$ with state space S.

The plan

The plan for the proof is as follows: We first show that $\mathbb{S}f$ transforms $F_{\varphi ?} \circ D$ into $F'_{\varphi ?} \circ D'$; here we use the assumption that f is a morphism, so the validity sets are respects by the inverse image of f. But once we have shown this, the proof proper is just a matter of showing that the corresponding diagram commutes by comparing $\boldsymbol{E}(f)(P_{\varphi ?}(s))$ against $Q_{\varphi ?}(f(s))$.

Note first that $\mathbb{S}(f)(F_{\varphi ?}(D(s))) = F'_{\varphi ?}(D'(f(s)))$, because we have for each $G \in \mathcal{B}$

$$\begin{aligned} &\mathbb{S}f\big(F_{\varphi ?}(D(s))\big)(G) \\ &\quad = F_{\varphi ?}(D(s))(f^{-1}\big[G\big]) && = D(s)\big([\![\varphi]\!]_{\mathcal{G}} \cap f^{-1}\big[G\big]\big) \\ &\quad \stackrel{(\dagger)}{=} D(s)\big(f^{-1}\big[[\![\varphi]\!]_{\mathcal{H}}\big] \cap f^{-1}\big[G\big]\big) && = D'(f(s))\big([\![\varphi]\!]_{\mathcal{H}} \cap G\big) \\ &\quad = F'_{\varphi ?}\big(D'(f(s))\big)(G). \end{aligned}$$

We have used $f^{-1}\big[[\![\varphi]\!]_{\mathcal{H}}\big] = [\![\varphi]\!]_{\mathcal{G}}$ in ($\dagger$), since f is a morphism; see Proposition 4.9.43. But now we may conclude

$$\begin{aligned} W \in \boldsymbol{E}(f)(P_{\varphi ?}(s)) &\Leftrightarrow \mathbb{S}f(F_{\varphi ?}(D(s))) \in W \Leftrightarrow F'_{\varphi ?}(D'(f(s))) \in W \\ &\Leftrightarrow W \in Q_{\varphi ?}(f(s)); \end{aligned}$$

hence $\boldsymbol{E}(f) \circ P_{\varphi ?} = Q_{\varphi ?} \circ f$ is established. $\dashv$

Example 4.9.49 We compute $[\![\langle p?; g\rangle_q \varphi]\!]_{\mathcal{G}}$ and $[\![\langle p¿; g\rangle_q \varphi]\!]_{\mathcal{G}}$ for a primitive formula $p \in \Psi$ and an arbitrary game g for the sake of illustration.

First a technical remark: Let λ be Lebesgue measure on the unit interval; then

$$\mathcal{G}, s \models \langle g\rangle_q \varphi \Leftrightarrow \lambda(\{r \in [0,1] \mid \mathcal{G}, s \models \langle g\rangle_r \varphi\}) > q. \tag{4.18}$$

In fact, the map $r \mapsto [\![\langle g\rangle_r\varphi]\!]_{\mathcal{G}}$ is monotone and decreasing; this is intuitively clear: If Angel will have a strategy for reaching a state in which formula φ holds with probability at least $q > q'$, it will have a strategy for reaching such a state with probability at least q'. But this entails that the set $\{r \in [0,1] \mid \mathcal{G}, s \models \langle g\rangle_r\varphi\}$ constitutes an interval which contains 0 if it is not empty. This interval is longer than q (i.e., its Lebesgue measure is greater than q) iff q is contained in it. From this (4.18) follows.

Now assume $\mathcal{G}, s \models \langle p?; g\rangle_q\varphi$. Thus

$$\begin{aligned} G_g([\![\varphi]\!]_{\mathcal{G}}, q) \in P_{p?}(s) &\Leftrightarrow F_{V_p}(D(s)) \in G_g([\![\varphi]\!]_{\mathcal{G}}, q) \\ &\Leftrightarrow \int_0^1 D(s)(V_p \cap [\![\langle g\rangle_r\varphi]\!]_{\mathcal{G}})\, dr > q, \end{aligned}$$

which means

$$D(s)(V_p) \cdot \int_0^1 D(s)([\![\langle g\rangle_r\varphi]\!]_{\mathcal{G}})\, dr > q.$$

This implies

$$D(s)(V_p) = 1 \text{ and } \int_0^1 D(s)([\![\langle g\rangle_r\varphi]\!]_{\mathcal{G}})\, dr > q,$$

the latter integral being equal to $\lambda(\{r \in [0,1] \mid \mathcal{G}, s \models \langle g\rangle_r\varphi\})$. Hence by (4.18) it follows that $\mathcal{G}, s \models \langle g\rangle_q\varphi$. Thus we have found

$$\mathcal{G}, s \models \langle p?; g\rangle_q\varphi \Leftrightarrow \mathcal{G}, s \models p \wedge \langle g\rangle_q\varphi.$$

Replacing in the above argumentation V_p by $S \setminus V_p$, we see that $\mathcal{G}, s \models \langle p\text{¿}; g\rangle_q\varphi$ is equivalent to

$$D(s)(S \setminus V_p) = 1 \text{ and } \int_0^1 D(s)([\![\langle g\rangle_r\varphi]\!]_{\mathcal{G}})\, dr > q.$$

Because we do not have negation in our logic, we obtain

$$\mathcal{G}, s \models \langle p\text{¿}; g\rangle_q\varphi \Leftrightarrow \mathcal{G}, s \not\models p \text{ and } \mathcal{G}, s \models \langle g\rangle_q\varphi.$$

✌

We leave this logic now and return to the discussion of topological properties of the space of all finite measures.

4.10 The Weak Topology

We will look again at topological issues for the space of finite measures. Because we have integration now at our disposal, we can use it for additional characterizations. We fix in this section (X, d) as a metric space. Recall that $\mathcal{C}(X)$ is the space of all bounded continuous functions $X \to \mathbb{R}$. This space induces the weak topology on the space $\mathbb{M}(X) = \mathbb{M}(X, \mathcal{B}(X))$ of all finite Borel measures on $(X, \mathcal{B}(X))$. This is the smallest topology which renders the evaluation map

$$ev_f : \mu \mapsto \int_X f\, d\mu$$

continuous for every continuous and bounded map $f : X \to \mathbb{R}$, i.e., the initial topology with respect to $(ev_f)_{f \in \mathcal{C}(X)}$; see Definition 3.1.14. This topology is fairly natural, and it is related to the topologies on $\mathbb{M}(X)$ considered so far, the Alexandrov topology, and the topology given by the Levy–Prokhorov metric, which are discussed in Sect. 4.1.4. We will show that these topologies are the same, provided the underlying space is Polish, and we will demonstrate that $\mathbb{M}(X)$ is a Polish space itself for this case. Somewhat weaker results may be obtained if the base space is only separable metric, and it turns out that tightness, i.e., inner approximability through compact sets, is the property which sets Polish spaces apart. We introduce also a very handy metric for the weak topology due to Hutchinson. Two case studies on bisimulations of Markov transition systems and on quotients for stochastic relations demonstrate the interplay of topological considerations with selection arguments, which become available on $\mathbb{M}(X)$ once this space is identified as Polish.

Define as the basis for the topology the sets

$$U_{f_1,\ldots,f_n,\epsilon}(\mu) := \{\nu \in \mathbb{M}(X) \mid \left|\int_X f_i\, d\nu - \int_X f_i\, d\mu\right| < \epsilon \text{ for } 1 \leq i \leq n\}$$

with $\epsilon > 0$ and $f_1, \ldots, f_n \in \mathcal{C}(X)$. Call the topology the *weak topology* on $\mathbb{M}(X)$.

With respect to convergence, we have this characterization, which indicates the relationship between the weak topology and the Alexandrov topology investigated in Sect. 4.1.4.

Theorem 4.10.1 *The following statements are equivalent for a sequence $(\mu_n)_{n \in \mathbb{N}} \subseteq \mathbb{M}(X)$:*

1. *$\mu_n \to \mu$ in the weak topology.*
2. *$\int_X f \, d\mu_n \to \int_X f \, d\mu$ for all $f \in \mathcal{C}(X)$.*
3. *$\int_X f \, d\mu_n \to \int_X f \, d\mu$ for all bounded and uniformly continuous $f : X \to \mathbb{R}$.*
4. *$\mu_n \to \mu$ in the A-topology.*

Proof The implications 1 $\Rightarrow$ 2 and 2 $\Rightarrow$ 3 are trivial.

3 $\Rightarrow$ 4: Let $G \subseteq X$ be open, then $f_k(x) := 1 \wedge k \cdot d(x, X \setminus G)$ defines a uniformly continuous map, and $0 \leq f_1 \leq f_2 \leq \ldots$ with $\lim_{k\to\infty} f_k = \chi_G$. Hence $\int_X f_k \, d\mu \leq \int_X \chi_G \, d\mu = \mu(G)$, and by monotone convergence $\int_X f_k \, d\mu \to \mu(G)$. From the assumption we know that $\int_X f_k \, d\mu_n \to \int_X f_k \, d\mu$, as $n \to \infty$, so that we obtain for all $k \in \mathbb{N}$

$$\lim_{n\to\infty} \int_X f_k \, d\mu_n \leq \liminf_{n\to\infty} \mu_n(G),$$

which in turn implies $\mu(G) \leq \liminf_{n\to\infty} \mu_n(G)$.

4 $\Rightarrow$ 2 We may assume that $f \geq 0$, because the integral is linear. By Example 4.9.7, Eq. (4.11), we can represent the integral through

$$\int_X f \, d\nu = \int_0^\infty \nu(\{x \in X \mid f(x) > t\}) \, dt.$$

Since f is continuous, the set $\{x \in X \mid f(x) > t\}$ is open. By Fatou's Lemma (Proposition 4.8.5), we obtain from the assumption

$$\begin{aligned} \liminf_{n\to\infty} \int_X f \, d\mu_n &= \liminf_{n\to\infty} \int_0^\infty \mu_n(\{x \in X \mid f(x) > t\}) \, dt \\ &\geq \int_0^\infty \liminf_{n\to\infty} \mu_n(\{x \in X \mid f(x) > t\}) \, dt \\ &\geq \int_0^\infty \mu(\{x \in X \mid f(x) > t\}) \, dt \\ &= \int_X f \, d\mu. \end{aligned}$$

Because $f \geq 0$ is bounded, we find $T \in \mathbb{R}$ such that $f(x) \leq T$ for all $x \in X$; hence $g(x) := T - f(x)$ defines a nonnegative and bounded function. Then by the preceding argument $\liminf_{n\to\infty} \int_X g\ d\mu_n \geq \int_X g\ d\mu$. Since $\mu_n(X) \to \mu(X)$, we infer

$$\limsup_{n\to\infty} \int_X f\ d\mu_n \leq \int_X f\ d\mu,$$

which implies the desired equality. ⊣

Let X be separable; then the A-topology is metrized by the Prohorov metric (Theorem 4.1.49). Thus we have established that the metric topology and the topology of weak convergence are the same for separable metric spaces. Just for the record

Theorem 4.10.2 *Let X be a separable metric space, then the Prohorov metric is a metric for the topology of weak convergence.* ⊣

It is now easy to find a dense subset in $\mathbb{M}(X)$. As one might expect, the measures living on discrete subsets are dense. Before stating and proving the corresponding statement, we have a brief look at the embedding of X into $\mathbb{M}(X)$.

Example 4.10.3 The base space X is embedded into $\mathbb{M}(X)$ as a closed subset through $x \mapsto \delta_x$. In fact, let $(\delta_{x_n})_{n\in\mathbb{N}}$ be a sequence which converges weakly to $\mu \in \mathbb{M}(X)$. We have in particular $\mu(X) = \lim_{n\to\infty} \delta_{x_n}(X) = 1$; hence $\mu \in \mathbb{P}(X)$. Now assume that $(x_n)_{n\in\mathbb{N}}$ does not converge; hence it does not have a convergent subsequence in X. Then the set $S := \{x_n \mid n \in \mathbb{N}\}$ is closed in X, so are all subsets of S. Take an infinite subset $C \subseteq S$ with an infinite complement $S \setminus C$; then $\mu(C) \geq \limsup_{n\to\infty} \delta_{x_n}(C) = 1$, and with the same argument, $\mu(S \setminus C) = 1$. This contradicts $\mu(X) = 1$. Thus we can find $x \in X$ with $x_n \to x$; hence $\delta_{x_n} \to \delta_x$, so that the image of X in $\mathbb{M}(X)$ is closed. ✌

This is what one would expect: The discrete measures form a dense set in the topology of weak convergence.

Proposition 4.10.4 *Let X be a separable metric space. The set*

$$\left\{\sum_{k\in\mathbb{N}} r_k \cdot \delta_{x_k} \mid x_k \in X, r_k \geq 0\right\}$$

of discrete measures is dense in the topology of weak convergence.

Plan for the proof

Proof 0. The plan for the proof goes like this: We cover the space with Borel sets of small diameter, and then take a uniformly continuous function as a witness. Uniform continuity then makes for uniform deviations on these sets, which establishes the claim.

1. Fix $\mu \in \mathbb{M}(X)$. Cover X for each $k \in \mathbb{N}$ with open sets $(G_{n,k})_{n \in \mathbb{N}}$, each of which has a diameter not less than $1/k$. Convert the cover through the first entrance trick to a cover of mutually disjoint Borel sets $A_{n,k} \subseteq G_{n,k}$, eliminating all empty sets arising from this process. Select an arbitrary $x_{n,k} \in A_{n,k}$. We claim that

$$\mu_n := \sum_{k \in \mathbb{N}} \mu(A_{n,k}) \cdot \delta_{x_{n,k}}$$

converges weakly to μ.

2. In fact, let $f : X \to \mathbb{R}$ be a uniformly continuous and bounded map. Since f is uniformly continuous,

$$\eta_n := \sup_{k \in \mathbb{N}} \bigl(\sup_{x \in A_{n,k}} f(x) - \inf_{x \in A_{n,k}} f(x) \bigr)$$

tends to 0, as $n \to \infty$. Thus

$$\begin{aligned} \left| \int_X f \, d\mu_n - \int_X f \, d\mu \right| &= |\sum_{k \in \mathbb{N}} (\int_{A_{n,k}} f \, d\mu_n - \int_{A_{n,k}} f \, d\mu)| \\ &\leq \eta_n \cdot \sum_{k \in \mathbb{N}} \mu(A_{n,k}) \\ &\leq \eta_n \cdot \mu(X) \\ &\to 0. \end{aligned}$$

This establishes the claim and proves the assertion. $\dashv$

We may arrange the cover in the proof in such a way that the points are taken from a dense set. Hence we obtain immediately

Corollary 4.10.5 *If X is a separable metric space, then $\mathbb{M}(X)$ is a separable metric space in the topology of weak convergence.*

Proof Because $\sum_{k=1}^{n} r_k \cdot \delta_{x_k} \to \sum_{k \in \mathbb{N}} r_k \cdot \delta_{x_k}$, as $n \to \infty$ in the weak topology and because the rationals $\mathbb{Q}$ are dense in the reals, we obtain from Proposition 4.10.4 that $\bigl\{\sum_{k=1}^{n} r_k \cdot \delta_{x_k} \mid x_k \in D, 0 \leq r_k \in \mathbb{Q}, n \in \mathbb{N}\bigr\}$ is a countable and dense subset of $\mathbb{M}(X)$, whenever $D \subseteq X$ is a countable and dense subset of X. $\dashv$

Another immediate consequence refers to the weak σ-algebra. We obtain from Lemma 4.1.50 together with Corollary 4.10.5

Corollary 4.10.6 *Let X be a metric space, then the weak σ-algebra is the Borel sets of the A-topology.* $\dashv$

We will show now that $\mathbb{M}(X)$ is a Polish space, provided X is one; thus applying the $\mathbb{M}$-functor to a Polish space does not leave the realm of Polish spaces.

We know from Alexandrov's Theorem 4.3.27 that a separable metrizable space is Polish iff it can be embedded as a G_δ-set into the Hilbert cube. We show first that for compact metric X, the space $\mathbb{S}(X)$ of all subprobability measures with the topology of weak convergence is itself a compact metric space. This is established by embedding it as a closed subspace into $[-1, +1]^\infty$. But there is nothing special about taking $\mathbb{S}$; the important property is that all measures are uniformly bounded (by 1, in this case). Any other bound would also do.

We require for this the Stone–Weierstraß Theorem, which implies that the unit ball in the space of all bounded continuous functions on a compact metric space is separable itself (Corollary 3.6.46). The idea of the embedding is to take a countable dense sequence $(g_n)_{n\in\mathbb{N}}$ of this unit ball. Since we are dealing with probability measures and since we know that each g_n maps X into the interval $[-1, 1]$, we know that $-1 \leq \int_X g_n \, d\mu \leq 1$ for each μ. This then spawns the desired map, which, together with its inverse, is shown to be continuous through the Riesz Representation Theorem 4.8.20.

Well, this is the plan of attack for establishing

Proposition 4.10.7 *Let X be a compact metric space. Then $\mathbb{S}(X)$ is a compact metric space.*

Proof 1. The space $\mathcal{C}(X)$ of continuous maps into the reals is for compact metric X a separable Banach space under the sup-norm $\|\cdot\|_\infty$ by Corollary 3.6.46. The closed unit ball

$$\mathcal{C}_1 := \{f \in \mathcal{C}(X) \mid \|f\|_\infty \leq 1\}$$

is a separable metric space in its own right, because it is Polish by Theorem 4.3.26. Let $(g_n)_{n\in\mathbb{N}}$ be a countable sense subset in $\mathcal{C}_1$, and define

$$\Omega : \mathbb{S}(X) \ni \nu \mapsto \langle \int_X g_1 \, d\nu, \int_X g_2 \, d\nu, \ldots \rangle \in [-1, 1]^\infty.$$

Then Ω is injective, because the sequence $(g_n)_{n\in\mathbb{N}}$ is dense.

2. Also, Ω^{-1} is continuous. In fact, let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{S}(X)$ such that $\big(\Omega(\mu_n)\big)_{n\in\mathbb{N}}$ converges in $[-1, 1]^\infty$; put $\alpha_i := \lim_{n\to\infty} \int_X g_i \, d\mu_n$. For each $f \in \mathcal{C}_1$, there exists a subsequence $(g_{n_k})_{k\in\mathbb{N}}$ such that $\|f - g_{n_k}\|_\infty \to 0$ as $k \to \infty$, because $(g_n)_{n\in\mathbb{N}}$ is dense in $\mathcal{C}_1$. Thus

$$L(f) := \lim_{n\to\infty} \int_X f \, d\mu_n$$

exists. Define $L(\alpha \cdot f) := \alpha \cdot L(f)$, for $\alpha \in \mathbb{R}$; then it is immediate that $L : \mathcal{C}(X) \to \mathbb{R}$ is linear and that $L(f) \geq 0$, provided $f \geq 0$. The Riesz Representation Theorem 4.8.20 now gives a unique $\mu \in \mathbb{S}(X)$ with

$$L(f) = \int_X f \, d\mu,$$

and the construction shows that

$$\lim_{n\to\infty} \Omega(\mu_n) = \langle \int_X g_1 \, d\mu, \int_X g_2 \, d\mu, \ldots \rangle.$$

3. Consequently, $\Omega : \mathbb{S}(X) \to \Omega\big[\mathbb{S}(X)\big]$ is a homeomorphism, and $\Omega\big[\mathbb{S}(X)\big]$ is closed, hence compact. Thus $\mathbb{S}(X)$ is compact. ⊣

We obtain as a first consequence

Proposition 4.10.8 *X is compact iff $\mathbb{S}(X)$ is, whenever X is a Polish space.*

Proof It remains to show that X is compact, provided $\mathbb{S}(X)$ is. Choose a complete metric d for X. Thus X is isometrically embedded into $\mathbb{S}(X)$ by $x \mapsto \delta_x$ with $A := \{\delta_x \mid x \in X\}$ being closed. We could appeal to Example 4.10.3, but a direct argument is available as well. In fact, if $\delta_{x_n} \to \mu$ in the weak topology, then $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in X on account of the isometry. Since (X, d) is complete, $x_n \to x$ for some $x \in X$, hence $\mu = \delta_x$, and thus A is closed, hence compact. ⊣

The next step for showing that $\mathbb{M}(X)$ is Polish is nearly canonical. If X is a Polish space, it may be embedded as a G_δ-set into a compact space $\tilde{X}$, the subprobabilities of which are topologically a closed subset of $[-1, +1]^\infty$, as we have just seen. We will show now that $\mathbb{M}(X)$ is a G_δ in $\mathbb{M}(\tilde{X})$ as well.

Proposition 4.10.9 *Let X be a Polish space. Then $\mathbb{M}(X)$ is a Polish space in the topology of weak convergence.*

Proof 1. Embed X as a G_δ-subset into a compact metric space $\tilde{X}$; hence $X \in \mathcal{B}(\tilde{X})$. Put

$$\mathbb{M}_0 := \{\mu \in \mathbb{M}(\tilde{X}) \mid \mu(\tilde{X} \setminus X) = 0\},$$

so $\mathbb{M}_0$ contains exactly those finite measures on $\tilde{X}$ that are concentrated on X. Then $\mathbb{M}_0$ is homeomorphic to $\mathbb{M}(X)$.

2. Write X as $X = \bigcap_{n \in \mathbb{N}} G_n$, where $(G_n)_{n \in \mathbb{N}}$ is a sequence of open sets in $\tilde{X}$. Given $r > 0$, the set

$$\Gamma_{k,r} := \{\mu \in \mathbb{M}(\tilde{X}) \mid \mu(\tilde{X} \setminus G_k) < r\}$$

is open in $\mathbb{M}(\tilde{X})$. In fact, if $\mu_n \notin \Gamma_{k,r}$ converges to μ_0 in the weak topology, then

$$\mu_0(\tilde{X} \setminus G_k) \geq \limsup_{n \to \infty} \mu_n(\tilde{X} \setminus G_k) \geq r$$

by Theorem 4.10.1, since $\tilde{X} \setminus G_k$ is closed. Consequently, $\mu_0 \notin \Gamma_{k,r}$. This shows that $\Gamma_{k,r}$ is open, because its complement is closed. Thus

$$\mathbb{M}_0 = \bigcap_{n \in \mathbb{N}} \bigcap_{k \in \mathbb{N}} \Gamma_{n,1/k}$$

is a G_δ-set, and the assertion follows. ⊣

Thus we obtain as a consequence

Proposition 4.10.10 *$\mathbb{M}(X)$ is a Polish space in the topology of weak convergence iff X is.*

Proof Let $\mathbb{M}(X)$ be Polish. The base space X is embedded into $\mathbb{M}(X)$ as a closed subset by Example 4.10.3; hence is a Polish space by Theorem 4.3.26. ⊣

Let $\mu \in \mathbb{M}(X)$ with X Polish. Since X has a countable basis, we know from Lemma 4.1.46 that μ is supported by a closed set, since μ is τ-regular. But in the presence of a complete metric, we can say a bit more, viz., that the value of $\mu(A)$ may be approximated from within by compact sets to arbitrary precision.

Definition 4.10.11 *A finite Borel measure μ is called* tight *iff*

$$\mu(A) = \sup\{\mu(K) \mid K \subseteq A \textit{ compact}\}$$

holds for all $A \in \mathcal{B}(X)$.

Thus tightness means for μ that we can find for any $\epsilon > 0$ and for any Borel set $A \subseteq X$ a compact set $K \subseteq A$ with $\mu(A \setminus K) < \epsilon$. Because a finite measure on a separable metric space is regular, i.e., $\mu(A)$ can be approximated from within A by closed sets (Lemma 4.6.13), it suffices in this case to consider tightness at X and hence to postulate that there exists for any $\epsilon > 0$ a compact set $K \subseteq X$ with $\mu(X \setminus K) < \epsilon$. We know in addition that each finite measure is τ-regular by Lemma 4.1.45; hence the union of a directed family of open sets has the supremum of these open sets as its measure. Capitalizing on this and on completeness, we find

Proposition 4.10.12 *Each finite Borel measure on a Polish space X is tight.*

We cover the space with open sets which are constructed as the open neighborhood of finite sets. From this a directed cover of open sets is easily constructed, and since we know that the measure is τ-regular, we extract suitable finite sets, from which a compact set is manufactured. This set is then shown to be suitable for our purposes. The important property here is τ-regularity and the observation that a complete bounded set in a metric space is compact. Line of attack

Proof 1. We show first that we can find for each $\epsilon > 0$ a compact set $K \subseteq X$ with $\mu(X \setminus K) < \epsilon$. In fact, given a complete metric d, consider

$$\mathcal{G} := \big\{\{x \in X \mid d(x, M) < 1/n\} \mid M \subseteq X \text{ is finite}\big\}.$$

Then $\mathcal{G}$ is a directed collection of open sets with $\bigcup \mathcal{G} = X$; thus we know from τ-regularity of μ that $\mu(X) = \sup\{\mu(G) \mid G \in \mathcal{G}\}$. Consequently, given $\epsilon > 0$, there exists for each $n \in \mathbb{N}$ a finite set $M_n \subseteq X$ with $\mu(\{x \in X \mid d(x, M_n) < 1/n\}) > \mu(X) - \epsilon/2^n$. Now define $K := \bigcap_{n\in\mathbb{N}}\{x \in X \mid d(x, M_n) \leq 1/n\}$.

Then K is closed and complete (since (X, d) is complete). Because each M_n is finite, K is totally bounded. Thus K is compact by Theorem 3.5.32. We obtain

$$\mu(X \setminus K) \leq \sum_{n\in\mathbb{N}} \mu(\{x \in X \mid d(x, M_n) \geq 1/n\}) \leq \sum_{n\in\mathbb{N}} \epsilon \cdot 2^{-n} = \epsilon.$$

2. Now let $A \in \mathcal{B}(X)$, then for $\epsilon > 0$ there exists $F \subseteq A$ closed with $\mu(A \setminus F) < \epsilon/2$, and choose $K \subseteq X$ compact with $\mu(X \setminus K) < \epsilon/2$. Then $K \cap F \subseteq A$ is compact with $\mu(A \setminus (F \cap K)) < \epsilon$. $\dashv$

Tightness is sometimes an essential ingredient when arguing about measures on a Polish space. The discussion on the Hutchinson metric in the next section provides an example; it shows that at a crucial point tightness kicks in and saves the day.

4.10.1 The Hutchinson Metric

We will explore now another approach to the weak topology for Polish spaces through the Hutchinson metric. Given a fixed metric d on X and a fixed real $\gamma > 0$, define

$$V_\gamma := \{f : X \to \mathbb{R} \mid |f(x) - f(y)| \le d(x, y) \text{ and } |f(x)| \le \gamma \text{ for all } x, y \in X\}.$$

Thus f is a member of V_γ iff f is non-expanding (hence has a Lipschitz constant 1) and iff its supremum norm $\|f\|_\infty$ is bounded by γ. Trivially, all elements of V_γ are uniformly continuous. Note the explicit dependence of the elements of V_γ on the metric d. The *Hutchinson distance* $H_\gamma(\mu, \nu)$ between $\mu, \nu \in \mathbb{M}(X)$ is defined as

$$H_\gamma(\mu, \nu) := \sup_{f \in V_\gamma} \Big(\int_X f\, d\mu - \int_X f\, d\nu\Big).$$

Then H_γ is easily seen to be a metric on $\mathbb{M}(X)$. H_γ is called the *Hutchinson metric* (sometimes also Hutchinson–Monge–Kantorovicz metric).

Discussion and plan

The relationship between this metric and the topology of weak convergence is stated in Proposition 4.10.13, the proof of which follows [Edg98, Theorem 2.5.17]. The program goes as follows. We first show that convergence in the Hutchinson metric implies convergence in the weak topology. This is a straightforward approximation argument on closed sets through suitable continuous functions. The converse is more complicated and relies on tightness. We find for the target measure μ of a converging sequence a good approximating compact set, which can be covered by a finite number of open sets, the boundaries of which vanish for μ. From this we construct a suitable approximation in the Hutchinson metric; clearly, uniform boundedness will be used heavily here.

Proposition 4.10.13 *Let X be a Polish space. Then H_γ is a metric for the topology of weak convergence on $\mathbb{M}(X)$ for any $\gamma > 0$.*

Proof 1. We may and do assume that $\gamma = 1$; otherwise we scale accordingly. Now let $H_1(\mu_n, \mu) \to 0$ as $n \to \infty$; then $\lim_{n\to\infty} \mu_n(X) = \mu(X)$. Let $F \subseteq X$ be closed; then we can find for given $\epsilon > 0$ a function $f \in V_1$ such that $f(x) = 1$ for $x \in F$, and $\int_X f\, dm \leq \mu(F) + \epsilon$. This gives

$$\limsup_{n\to\infty} \mu_n(F) \leq \lim_{n\to\infty} \int_X f\, d\mu_n = \int_X f\, d\mu \leq \mu(F) + \epsilon.$$

Thus convergence in the Hutchinson metric implies convergence in the A-topology and hence in the topology of weak convergence, by Proposition 4.1.35.

2. Now assume that $\mu_n \to \mu$ in the topology of weak convergence; thus $\mu_n(A) \to \mu(A)$ for all $A \in \mathcal{B}(X)$ with $\mu(\partial A) = 0$ by Corollary 4.1.36; we assume that μ_n and μ are probability measures; otherwise we scale again. Because X is Polish, μ is tight by Proposition 4.10.12.

Fix $\epsilon > 0$; then there exists a compact set $K \subseteq X$ with

$$\mu(X \setminus K) < \frac{\epsilon}{5 \cdot \gamma}.$$

Given $x \in K$, there exists an open ball $B_r(x)$ with center x and radius r such that $0 < r < \epsilon/10$ such that $\mu(\partial B_r(x)) = 0$; see Corollary 4.1.39. Because K is compact, a finite number of these balls will suffice; thus $K \subseteq B_{r_1}(x_1) \cup \ldots \cup B_{r_p}(x_p)$. Transform this cover into a disjoint cover by setting

$$\begin{aligned}
E_1 &:= B_{r_1}(x_1),\\
E_2 &:= B_{r_2}(x_2) \setminus E_1,\\
&\ldots\\
E_p &:= B_{r_p}(x_p) \setminus (E_1 \cup \ldots \cup E_{p-1}),\\
E_0 &:= S \setminus (E_1 \cup \ldots \cup E_p).
\end{aligned}$$

We observe these properties:

1. For $i = 1, \ldots, p$, the diameter of each E_i is not greater than $2 \cdot r_i$, hence smaller than $\epsilon/5$,

2. For $i = 1, \ldots, p$, $\partial E_i \subseteq \partial\big(B_{r_1}(x_1) \cup \ldots \cup B_{r_p}(x_p)\big)$, thus $\partial E_i \subseteq (\partial B_{r_1}(x_1)) \cup \ldots \cup (\partial B_{r_p}(x_p))$, and hence $\mu(\partial E_i) = 0$.

3. Because the boundary of a set is also the boundary of its complement, we conclude $\mu(\partial E_0) = 0$ as well. Moreover, $\mu(E_0) < \epsilon/(5 \cdot \gamma)$, since $E_0 \subseteq X \setminus K$.

Eliminate all E_i which are empty. Select $\eta > 0$ such that $p \cdot \eta < \epsilon/5$, and determine $n_0 \in \mathbb{N}$ so that $|\mu_n(E_i) - \mu(E_i)| < \eta$ for $i = 0, \dots, p$ and $n \geq n_0$.

We have to show that

$$\sup_{f \in V_\gamma} \Big(\int_X f \, d\mu_n - \int_X f \, d\mu \Big) \to 0, \text{ as } n \to \infty.$$

So take $f \in V_\gamma$ and fix $n \geq n_0$. Let $i = 1, \dots, p$, and pick an arbitrary $e_i \in E_i$; because each E_i has a diameter not greater than $\epsilon/5$, we know that $|f(x) - f(e_i)| < \epsilon/5$ for each $x \in E_i$. If $x \in E_0$, we have $|f(x)| \leq \gamma$. Now we are getting somewhere: Let $n \geq n_0$; then we obtain

$$\begin{aligned}
\int_X f \, d\mu_n &= \sum_{i=0}^{p} \int_{E_i} f \, d\mu_n \\
&\leq \gamma \cdot \mu_n(E_0) + \sum_{i=1}^{p} \big(f(t_i) + \frac{\epsilon}{5} \big) \cdot \mu_n(E_i) \\
&\leq \gamma \cdot (\mu(E_0) + \eta) + \sum_{i=1}^{p} \big(f(t_i) + \frac{\epsilon}{5} \big) \cdot (\mu(E_i) + \eta) \\
&\leq \gamma \cdot (\frac{\epsilon}{5 \cdot \gamma} + \eta) + \sum_{i=1}^{p} (f(t_i) - \frac{\epsilon}{5}) \cdot \mu(E_i) \\
&\quad + \frac{2 \cdot \epsilon}{5} \sum_{i=1}^{p} \mu(E_i) + \frac{p \cdot \epsilon \cdot \eta}{5} \\
&\leq \int_X f \, d\mu + \epsilon.
\end{aligned}$$

Recall that

$$\sum_{i=1}^{p} \mu(E_i) \leq \sum_{i=0}^{p} \mu(E_i) = \mu(X) = 1$$

and that

$$\int_{E_i} f \, d\mu \geq \mu(E_i) \cdot (f(t_i) - \epsilon/5).$$

In a similar fashion, we obtain $\int_X f \, d\mu_n \geq \int_X f \, d\mu - \epsilon$, so that we have established

$$|\int_X f \, d\mu - \int_X f \, d\mu_n| < \epsilon$$

for $n \geq n_0$. Since $f \in V_\gamma$ was arbitrary, we have shown that $H_\gamma(\mu_n, \mu) \to 0$. $\dashv$

The Hutchinson metric is occasionally easier to use than the Prohorov metric, because integrals may sometimes be easily manipulated in convergence arguments than ϵ-neighborhoods of sets.

4.10.2 Case Study: Eilenberg–Moore Algebras for the Giry Monad

We will study the Eilenberg–Moore algebras for the Giry monad again, but this time for the non-discrete case. Theorem 2.5.23 contains a complete characterization of these algebras for the discrete probability functor $\boldsymbol{D}$ as the positive convex structures on X. We will derive a complete characterization for the probability functor on Polish spaces from this.

We work in the category of Polish spaces with continuous maps as morphisms and the Borel sets as the σ-algebra. The Giry monad $(\mathbb{P}, e, m)$ is introduced in Example 2.4.8; its functorial part is the subprobability functor $\mathbb{P}$, and the unit e and the multiplication m are for a Polish space X defined through

$$\begin{aligned} e_X(x) &:= \delta_x, \\ m_X(M)(A) &:= \int_X \vartheta(A) \, M(d\vartheta) \end{aligned}$$

for $x \in X$, $M \in \mathbb{P}^2 X$, and $A \in \mathcal{B}(X)$. An Eilenberg–Moore algebra $h : \mathbb{P}X \to X$ is a morphism so that these diagrams from page 189 commute

$$\begin{tikzcd} \mathbb{P}^2 X \arrow[r, "\mathbb{P}h"] \arrow[d, "m_X"'] & \mathbb{P}X \arrow[d, "h"] \\ \mathbb{P}X \arrow[r, "h"'] & X \end{tikzcd} \qquad \begin{tikzcd} X \arrow[r, "e_X"] \arrow[dr, "id_X"'] & \mathbb{P}X \arrow[d, "h"] \\ & x \end{tikzcd}$$

We note first that unit and multiplication of the monad are compatible with the weak topology.

Lemma 4.10.14 *Given a Polish space X, the unit $e_X : X \to \mathbb{P}X$ and the multiplication $m_X : \mathbb{P}^2 X \to \mathbb{P}X$ are continuous in the respective weak topologies.*

Proof We know this already from Lemma 4.1.42 for the unit, so we have to establish the claim for the multiplication. Let $M \in \mathbb{P}^2(X)$ and $f \in \mathcal{C}(X)$ be a bounded continuous function; then

$$\int_X f \, dm_X(M) = \int_{\mathbb{P}X} \Big(\int_X f \, d\vartheta \Big) \, dM(\vartheta) \tag{4.19}$$

holds. Granted that we have shown this, we argue then as follows: By the definition of the weak convergence on $\mathbb{P}X$, the map $E_f : \vartheta \mapsto \int_X f \, d\vartheta$ is continuous, whenever $f \in \mathcal{C}(X)$; thus $M \mapsto \int_{\mathbb{P}X} E_f \, dM$ is continuous by the definition of the weak topology on $\mathbb{P}^2 X$. But this is just m_X.

So it remains to establish Eq. (4.19). If $f = \chi_A$ for $A \in \mathcal{B}(X)$, this is just the definition of m_X, so the equation holds in this case. Since the integral is linear, Eq. (4.19) holds for bounded step functions f. If $f \geq 0$ is bounded and measurable, Levi's Theorem 4.8.2 in combination with Lebesgue's Dominated Convergence Theorem 4.8.6 shows that the equation holds. The general case now follows from decomposing $f = f^+ - f^-$ with $f^+ \geq 0$ and $f^- \geq 0$. $\dashv$

Now fix a Polish space X and a complete metric d; the Hutchinson metric H_γ for some $\gamma > 0$ is assumed as a metric for the topology of weak convergence; see Proposition 4.10.13. Put as in Sect. 2.5.2

$$\Omega := \{\langle \alpha_1, \ldots, \alpha_k \rangle \mid k \in \mathbb{N}, \alpha_i \geq 0, \textstyle\sum_{i=1}^k \alpha_i \leq 1\}.$$

Given an Eilenberg–Moore algebra $\langle X, h \rangle$ for $\mathbb{P}$, define for $\alpha = \langle \alpha_1, \ldots, \alpha_n \rangle \in \Omega$ the map

$$\langle \alpha_1, \ldots, \alpha_n \rangle_h (x_1, \ldots, x_n) := h\big(\textstyle\sum_{i=1}^n \alpha_i \cdot \delta_{x_i}\big).$$

Then the lengthy computation in Lemma 2.5.20 shows that $\alpha \mapsto \alpha_h$ defines a positive convex structure. Let, conversely, a positive convex structure $\mathfrak{p}$ be given; then Lemma 2.5.21 shows that

$$h_{\mathfrak{p}}\big(\textstyle\sum_{i=1}^n \alpha_i \cdot \delta_{x_i}\big) := \textstyle\sum_{1 \leq i \leq n}^{\mathfrak{p}} \alpha_i \cdot x_i$$

for $\langle \alpha_1, \ldots, \alpha_n \rangle \in \Omega$, and $x_1, \ldots, x_n \in X$ defines an algebra $h_{\mathfrak{p}}$ for the discrete probability functor $\boldsymbol{D}X$. Since the discrete probability measures are dense in $\mathbb{P}X$ by Proposition 4.10.4, we look for a continuous extension from $\boldsymbol{D}X$ to $\mathbb{P}X$. In fact, this is possible provided $\mathfrak{p}$ is regular in the following sense:

Definition 4.10.15 *The positive convex structure $\mathfrak{p}$ is said to be* regular *iff the oscillation $\varnothing_{h_\mathfrak{p}}(\mu)$ vanishes for every $\mu \in \mathbb{P}X$.*

Thus we measure the oscillation

$$\varnothing_{h_\mathfrak{p}}(\mu) = \inf\big\{\mathrm{diam}\big(h_\mathfrak{p}\big[\boldsymbol{D}X \cap B(\mu, r)\big]\big) \mid r > 0\big\}$$

for $h_\mathfrak{p}$ at every probability measure $\mu \in \mathbb{P}X$ with $B(\mu, r)$ as the ball at center μ and radius r for the Hutchinson metric H_γ. The oscillation of a function is discussed on page 345. Thus $\mathfrak{p}$ is regular iff we know for all $\mu \in \mathcal{P}X$ that, given $\epsilon > 0$, there exists $r > 0$ such that

$$d\left(\sum\nolimits^{\mathfrak{p}}_{1 \le i \le n} \alpha_i \cdot x_i, \sum\nolimits^{\mathfrak{p}}_{1 \le i \le m} \beta_j \cdot y_j\right) < \epsilon,$$

whenever

$\left|\sum_{i=1}^{n} \alpha_i \cdot f(x_i) - \int_X f\ d\mu\right| < r$ and $\left|\sum_{j=1}^{m} \beta_i \cdot f(y_i) - \int_X f\ d\mu\right| < r$ for all $f \in V_\gamma$.

This gives the following characterization of the Eilenberg–Moore algebras for the non-discrete case, generalizing the "discrete" Theorem 2.5.23 to the non-discrete case.

Theorem 4.10.16 *The Eilenberg–Moore algebras for the Giry monad for Polish spaces are exactly the regular positive convex structures.* Complete characterization

Proof 1. Let $\langle X, h\rangle$ be an Eilenberg–Moore algebra; then $h : \mathbb{P}X \to X$ is continuous, and hence the restriction of h to $\boldsymbol{D}X$ has oscillation 0 at each $\mu \in \mathbb{P}X$. But this means that the corresponding positive convex structure is regular.

2. Conversely, given a regular positive convex structure $\mathfrak{p}$, the associated map $h_\mathfrak{p} : \boldsymbol{D}X \to X$ has oscillation zero at each $\mu \in \mathbb{P}X$. Thus there exists a unique continuous extension $h'_\mathfrak{p} : \mathbb{P}X \to X$ by Lemma 3.5.24. It remains to show that $h'_\mathfrak{p}$ satisfies the laws of an Eilenberg–Moore algebra. This follows easily from Lemma 4.10.14, since $h_\mathfrak{p}$ satisfies the corresponding equations. $\dashv$

We now have a complete characterization of the Eilenberg–Moore algebras for the probability monad over Polish spaces. This closes the gap we had to leave in Sect. 2.5.2 because we did not yet have the necessary tools at our disposal. It displays an interesting cooperation between arguments from categories, topology, and measure theory.

4.10.3 Case Study: Bisimulation

Bisimilarity is an important notion in the theory of concurrent systems, introduced originally by Milner for transition systems; see Sect. 2.6.1 for a general discussion. We will show in this section that the methods developed so far may be used in the investigation of bisimilarity for stochastic systems. We will first show that the category of stochastic relations has semi-pullbacks and use this information for a construction of bisimulations for these systems.

If we are in a general category $\boldsymbol{K}$, then the *semi-pullback* for two morphisms $f : a \to c$ and $g : b \to c$ with common range c consists of an object x and of morphisms $p_a : x \to a$ and $p_b : x \to b$ such that $f \circ p_a = g \circ p_b$, i.e., such that this diagram commutes in $\boldsymbol{K}$:

$$\begin{tikzcd} x \arrow[r, "p_b"] \arrow[d, "p_a"'] & b \arrow[d, "g"] \\ a \arrow[r, "f"'] & c \end{tikzcd}$$

We want to show that semi-pullbacks exist for stochastic relations over Polish spaces. This requires some preparations, provided through selection arguments.

The next statement appears to be interesting in its own right; it shows that a measurable selection for weakly continuous stochastic relations exist.

Proposition 4.10.17 *Let X_i, Y_i be Polish spaces, $K_i : X_i \rightsquigarrow Y_i$ be a weakly continuous stochastic relation, and $i = 1, 2$. Let $A \subseteq X_1 \times X_2$ and $B \subseteq Y_1 \times Y_2$ be closed subsets of the respective Cartesian products with projections equal to the base spaces, and assume that for $\langle x_1, x_2 \rangle \in A$ the set*

$$\Gamma(x_1, x_2) := \{ \mu \in \mathbb{S}(B) \mid \mathbb{S}(\beta_i)(\mu) = K_i(x_i),\ i = 1, 2 \}$$

is not empty, $\beta_i : B \to Y_i$ denoting the projections. Then there exists a stochastic relation $M : A \rightsquigarrow B$ such that $M(x_1, x_2) \in \Gamma(x_1, x_2)$ for all $\langle x_1, x_2 \rangle \in A$.

Outline of the proof

Let us have a look at the flow of the proof, before diving into it. It proceeds as follows. First, Y_i is embedded into its Alexandrov compactification $\overline{Y_i}$ with the purpose of obtaining a selector from the Kuratowski and Ryll-Nardzewski Selection Theorem. But we have to make sure that the set-valued map remains concentrated on the originally given space. This can be established; we obtain a selection for the embedding and adjust the selection accordingly.

Proof 1. Let $\overline{Y_i}$ for $i = 1, 2$ be the Alexandrov compactification of Y_i and $\overline{B}$ the closure of B in $\overline{Y}_1 \times \overline{Y}_2$. Then $\overline{B}$ is compact and contains the embedding of B into $\overline{Y}_1 \times \overline{Y}_2$, which we identify with B as a Borel subset. This is so since Y_i is a Borel subset in its compactification. The projections $\overline{\beta}_i : \overline{B} \to \overline{Y}_i$ are the continuous extensions to the projections $\beta_i : B \to Y_i$.

2. The map $r_i : \mathbb{S}(Y_i) \to \mathbb{S}(\overline{Y}_i)$ with $r_i(\mu)(G) := \mu(G \cap Y_i)$ for $G \in \mathcal{B}(\overline{Y}_i)$ is continuous; in fact, it is an isometry with respect to the respective Hutchinson metrics, once we have fixed metrics for the underlying spaces. Define for $\langle x_1, x_2 \rangle \in A$ the set

$$\Gamma_0(x_1, x_2) := \{ \mu \in \mathbb{S}(\overline{B}) \mid \mathbb{S}(\overline{\beta}_i)(\mu) = (r_i \circ K_i)(x_i),\ i = 1, 2 \}.$$

Thus Γ_0 maps A to the nonempty closed subsets of $\mathbb{S}(\overline{B})$, since $\mathbb{S}(\overline{\beta}_i)$ and $r_i \circ K_i$ are continuous for $i = 1, 2$. If $\mu \in \Gamma_0(x_1, x_2)$, then

$$\begin{aligned}
\mu(\overline{B} \setminus B) &\leq \mu\big(\overline{B} \cap (\overline{Y}_1 \setminus Y_1 \times \overline{Y}_2) \cup (\overline{Y}_1 \times \overline{Y}_2 \setminus Y_2)\big) \\
&= \mathbb{S}(\overline{\beta}_1)(\mu)(\overline{Y}_1 \setminus Y_1) + \mathbb{S}(\overline{\beta}_2)(\mu)(\overline{Y}_2 \setminus Y_2) \\
&= \big(r_1 \circ K_1\big)(x_1)(\overline{Y}_1 \setminus Y_1) + \big(r_2 \circ K_2\big)(x_2)(\overline{Y}_2 \setminus Y_2) \\
&= 0.
\end{aligned}$$

Hence all members of $\Gamma_0(x_1, x_2)$ are concentrated on B.

3. Let $C \subseteq \mathbb{S}(\overline{B})$ be compact, and assume that $(t_n)_{n \in \mathbb{N}}$ is a converging sequence in A with $t_n \in \Gamma_0^w(C)$ for all $n \in \mathbb{N}$ such that $t_n \to t_0 \in A$. Then there exists some $\mu_n \in C \cap \Gamma_0(t_n)$ for each $n \in \mathbb{N}$. Since C is compact, there exists by Proposition 3.5.31 a converging subsequence, which we assume to be the sequence itself, so $\mu_n \to \mu$ for some $\mu \in C$ in the topology of weak convergence. Continuity of $\mathbb{S}(\overline{\beta}_i)$ and of $K_i(x_i)$ for $i = 1, 2$ implies $\mu \in \Gamma_0$. Consequently, $\Gamma_0^w(C)$ is a closed subset of A.

4. Since $\mathbb{S}(\overline{B})$ is compact, we may represent each open set G as a countable union of compact sets $(C_n)_{n \in \mathbb{N}}$, so that

$$\Gamma_0^w(G) = \bigcup_{n \in \mathbb{N}} \Gamma_0^w(C_n);$$

hence $\Gamma_0^w(G)$ is a Borel set in A. The Kuratowski and Ryll-Nardzewski Selection Theorem 4.7.2 together with Lemma 4.1.10 gives us a stochastic relation $M_0 : A \rightsquigarrow \overline{B}$ with $M_0(x_1, x_2) \in \Gamma_0(x_1, x_2)$ for all $\langle x_1, x_2\rangle \in A$. Define $M(x_1, x_2)$ as the restriction of $M_0(x_1, x_2)$ to the Borel sets of B; then $M : A \rightsquigarrow B$ is the desired relation, because $M_0(x_1, x_2)(\overline{B} \setminus B) = 0$. ⊣

For the construction we are about to undertake, we will put to work the selection machinery just developed; this requires us to show that the set from which we want to select is nonempty. The following technical argument will be of assistance.

Assume that we have Polish spaces X_1, X_2 and a separable measure space $(Z, \mathcal{C})$ with surjective and measurable maps $f_i : X_i \to Z$ for $i = 1, 2$. We also have subprobability measures $\mu_i \in \mathbb{S}(X_i)$. Since $(Z, \mathcal{C})$ is separable, we may assume that $\mathcal{C}$ constitutes the Borel sets for some separable metric space (Z, d); see Proposition 4.3.10. Proposition 4.3.31 then tells us that we may assume that f_1 and f_2 are continuous. Now define

$$\begin{aligned} S &:= \{\langle x_1, x_2\rangle \in X_1 \times X_2 \mid f_1(x_1) = f_2(x_2)\} \\ \mathcal{A} &:= S \cap (f_1 \times f_2)^{-1}\left[\mathcal{C} \otimes \mathcal{C}\right]. \end{aligned}$$

Since $\Delta_Z := \{\langle z, z\rangle \mid z \in Z\}$ is a closed subset of $Z \times Z$ and since f_1 and f_2 are continuous, $S = (f_1 \times f_2)^{-1}\left[\Delta_Z\right]$ is a closed subset of the Polish space $X_1 \times X_2$ and hence a Polish space itself by Lemma 4.3.21. Assume that we have a finite measure ϑ on $\mathcal{A}$ such that $\mathbb{S}(\pi_i)(\vartheta)(E_i) = \mu_i(E_i)$ for all $E_i \in f_i^{-1}\left[\mathcal{C}\right]$, $i = 1, 2$ with $\pi_1 : X_1 \to Z$ and $\pi_2 : X_2 \to Z$ as the projections. Now $\mathcal{A} \subseteq \mathcal{B}(S)$ is usually not the σ-algebra of Borel sets for some Polish topology on S, which, however, will be needed. Here Lubin's construction steps in.

Lemma 4.10.18 *In the notation above, there exists a measure ϑ^+ on the Borel sets of S extending ϑ such that $\mathbb{S}(\pi_i)(\vartheta^+)(E_i) = \mu_i(E_i)$ holds for all $E_i \in \mathcal{B}(S)$.*

Proof Because $\mathcal{C}$ is countably generated, $\mathcal{C} \otimes \mathcal{C}$ is, so $\mathcal{A}$ is a countably generated σ-algebra. By Lubin's Theorem 4.6.12, there exists an extension ϑ^+ to ϑ. ⊣

So much for the technical preparations; we will now turn to bisimulations. A bisimulation relates two transition systems which are connected through a mediating system. In order to define this for the present context, we need morphisms. Recall from Example 2.1.14 that a morphism $m = (f, g) : K_1 \to K_2$ for stochastic relations $K_i : (X_i, \mathcal{A}_i) \rightsquigarrow (Y_i, \mathcal{B}_i)$ $(i = 1, 2)$ over general measurable spaces is given through the measurable maps $f : X_1 \to X_2$ and $g : Y_1 \to Y_2$ such that this diagram of measurable maps commutes

$$\begin{tikzcd}
(X_1, \mathcal{A}_1) \arrow[r, "f"] \arrow[d, "K_1"'] & (X_2, \mathcal{A}_2) \arrow[d, "K_2"] \\
\mathbb{S}(Y_1, \mathcal{B}_1) \arrow[r, "\mathbb{S}(g)"'] & \mathbb{S}(Y_2, \mathcal{B}_2)
\end{tikzcd}$$

Equivalently, $K_2(f(x_1)) = \mathbb{S}(g)(K_1(x_1))$, which translates to $K_2(f(x_1))(B) = K_1(x_1)(g^{-1}\left[B\right])$ for all $B \in \mathcal{B}_2$.

Definition 4.10.19 *The stochastic relations $K_i : (X_i, \mathcal{A}_i) \rightsquigarrow (Y_i, \mathcal{B}_i)$ $(i = 1, 2)$ are called* bisimilar *iff there exist a stochastic relation $M : (A, \mathcal{X}) \rightsquigarrow (B, \mathcal{Y})$ and surjective morphisms $m_i = (f_i, g_i) : M \to K_i$ such that the σ-algebra $g_1^{-1}\left[\mathcal{B}_1\right] \cap g_2^{-1}\left[\mathcal{B}_2\right]$ is nontrivial, i.e., contains not only $\emptyset$ and B. The relation M is called* mediating. Bisimilarity

The first condition on bisimilarity is in accordance with the general definition of bisimilarity of coalgebras; it requests that m_1 and m_2 form a span of morphisms

$$K_1 \xleftarrow{m_1} M \xrightarrow{m_2} K_2$$

Hence, the following diagram of measurable maps is supposed to commute with $m_i = (f_i, g_i)$ for $i = 1, 2$:

$$\begin{tikzcd}
(X_1, \mathcal{A}_1) \arrow[d, "K_1"'] & (A, \mathcal{X}) \arrow[l, "f_1"'] \arrow[r, "f_2"] \arrow[d, "M"] & (X_2, \mathcal{A}_2) \arrow[d, "K_2"] \\
\mathbb{S}(Y_1, \mathcal{B}_1) & \mathbb{S}(B, \mathcal{Y}) \arrow[l, "\mathbb{S}(g_1)"] \arrow[r, "\mathbb{S}(g_2)"'] & \mathbb{S}(Y_2, \mathcal{B}_2)
\end{tikzcd}$$

Thus, for each $a \in A$, $D \in \mathcal{B}_1$, $E \in \mathcal{B}_2$, the equalities

$$K_1\big(f_1(a)\big)(D) = \big(\mathbb{S}(g_1) \circ M\big)(a)(D) = M(a)\big(g_1^{-1}\left[D\right]\big)$$
$$K_2\big(f_2(a)\big)(E) = \big(\mathbb{S}(g_2) \circ M\big)(a)(E) = M(a)\big(g_2^{-1}\left[E\right]\big)$$

should be satisfied. The second condition, however, is special; it states that we can find an event $C^* \in \mathcal{Y}$ which is common to both K_1 and K_2 in the sense that

$$g_1^{-1}\left[B_1\right] = C^* = g_2^{-1}\left[B_2\right]$$

for some $B_1 \in \mathcal{B}_1$ and $B_2 \in \mathcal{B}_2$ such that both $C^* \neq \emptyset$ and $C^* \neq B$ hold (note that for $C^* = \emptyset$ or $C^* = B$, we can always take the empty and the full set, respectively). Given such a C^* with B_1, B_2 from above, we get for each $a \in A$

$$\begin{aligned} K_1(f_1(a))(B_1) = M(a)(g_1^{-1}\left[B_1\right]) &= M(a)(C^*) \\ &= M(a)(g_2^{-1}\left[B_2\right]) = K_2(g_2(a))(B_2); \end{aligned}$$

thus the event C^* ties K_1 and K_2 together. Loosely speaking, $g_1^{-1}\left[\mathcal{B}_1\right] \cap g_2^{-1}\left[\mathcal{B}_2\right]$ can be described as the σ-algebra of common events, which is required to be nontrivial.

Note that without the second condition, two relations K_1 and K_2 which are strictly probabilistic (i.e., for which the entire space is always assigned probability 1) would always be bisimilar: Put $A := X_1 \times X_2$, $B := Y_1 \times Y_2$ and set for $\langle x_1, x_2 \rangle \in A$ as the mediating relation $M(x_1, x_2) := K_1(x_1) \otimes K_2(x_2)$; that is, define M pointwise to be the product measure of K_1 and K_2. Then the projections will make the diagram commutative. But although this notion of bisimilarity is sometimes suggested, it is way too weak, because bisimulations relate transition systems, and it does not promise particularly interesting insights when two arbitrary systems can be related. It is also clear that using products for mediation does not work for the subprobabilistic case.

We will show now that we can construct a bisimulation for stochastic relations which are linked through a co-span $K_1 \longleftarrow K \longrightarrow K_2$ The center K of this co-span should be defined over second countable metric spaces, K_1 and K_2, over Polish spaces. This situation is sometimes easy to obtain, e.g., when factoring Kripke models over Polish spaces through a suitable logic; then K is defined over analytic spaces, which are separable metric. This is described in greater detail in Example 4.10.21.

Proposition 4.10.20 *Let $K_i : X_i \rightsquigarrow Y_i$ be stochastic relations over Polish spaces, and assume that $K : X \rightsquigarrow Y$ is a stochastic relation, where X, Y are second countable metric spaces. Assume that we have a co-span of morphisms $m_i : K_i \to K, i = 1, 2$; then there exist a stochastic relation M and morphisms $m_i^+ : M \rightsquigarrow K_i, i = 1, 2$*

rendering this diagram commutative:

$$\begin{tikzcd} M \arrow[r, "m_1^+"] \arrow[d, "m_2^+"'] & K_2 \arrow[d, "m_2"] \\ K_1 \arrow[r, "m_1"'] & K \end{tikzcd}$$

The stochastic relation M is defined over Polish spaces.

Proof 1. Assume $K_i = (X_i, Y_i, K_i)$ with $m_i = (f_i, g_i)$, $i = 1, 2$. Because of Proposition 4.3.31, we may assume that the respective σ-algebras on X_1 and X_2 are obtained from Polish topologies which render f_1 and K_1 as well as f_2 and K_2 continuous. These topologies are fixed for the proof. Put

$$A := \{\langle x_1, x_2\rangle \in X_1 \times X_2 \mid f_1(x_1) = f_2(x_2)\},$$
$$B := \{\langle y_1, y_2\rangle \in Y_1 \times Y_2 \mid g_1(y_1) = g_2(y_2)\};$$

then both A and B are closed, hence Polish. $\alpha_i : A \to X_i$ and $\beta_i : B \to Y_i$ are the projections, $i = 1, 2$. The diagrams

$$\begin{tikzcd} X_1 \arrow[r, "f_1"] \arrow[d, "K_1"'] & X \arrow[d, "K"'] & X_2 \arrow[l, "f_2"'] \arrow[d, "K_2"] \\ \mathbb{S}(Y_1) \arrow[r, "\mathbb{S}(g_1)"'] & \mathbb{S}(Y) & \mathbb{S}(Y_2) \arrow[l, "\mathbb{S}(g_2)"] \end{tikzcd}$$

are commutative by assumption; thus we know that for $x_i \in X_i$

$$K(f_1(x_1)) = \mathbb{S}(g_1)(K_1(x_1)) \text{ and } K(f_2(x_2)) = \mathbb{S}(g_2)(K_2(x_2))$$

holds. The construction implies that $(g_1 \circ \beta_1)(y_1, y_2) = (g_2 \circ \beta_2)(y_1, y_2)$ is true for $\langle y_1, y_2\rangle \in B$, and $g_1 \circ \beta_1 : B \to Y$ is surjective.

2. Fix $\langle x_1, x_2\rangle \in A$. Separability of the target spaces now enters: We know that the image of a surjective map under $\mathbb{S}$ is onto again by Proposition 4.6.11, so that there exists $\mu_0 \in \mathbb{S}(B)$ with $\mathbb{S}(g_1 \circ \beta_1)(\mu_0) = K(f_1(x_1))$, consequently, $\mathbb{S}(g_i \circ \beta_i)(\mu_0) = \mathbb{S}(g_i)(K_i(x_i))$ $(i = 1, 2)$. But this means for $i = 1, 2$

$$\forall E_i \in g_i^{-1}\left[\mathcal{B}(Y)\right] : \mathbb{S}(\beta_i)(\mu_0)(E_i) = K_i(x_i)(E_i).$$

Put

$$\begin{aligned}&\Gamma(x_1, x_2)\\&\quad := \{\mu \in \mathbb{S}(B) \mid \mathbb{S}(\beta_1)(\mu) = K_1(x_1) \text{ and } \mathbb{S}(\beta_2)(\mu) = K_2(x_2)\};\end{aligned}$$

then Lemma 4.10.18 shows that $\Gamma(x_1, x_2) \neq \emptyset$.

3. The set

$$\Gamma^w(C) = \{\langle x_1, x_2\rangle \in A \mid \Gamma(x_1, x_2) \cap C \neq \emptyset\}$$

is closed in A for compact $C \subseteq \mathbb{S}(B)$. This is shown exactly as in the second part of the proof for Proposition 4.10.17, from which now is inferred that there exists a measurable map $M : A \to \mathbb{S}(B)$ with $M(x_1, x_2) \in \Gamma(x_1, x_2)$ for every $\langle x_1, x_2\rangle \in A$. Thus $M : A \rightsquigarrow B$ is a stochastic relation with

$$K_1 \circ \alpha_1 = \mathbb{S}(\beta_1) \circ M \text{ and } K_2 \circ \alpha_2 = \mathbb{S}(\beta_2) \circ M.$$

Thus M with $m_1^+ := (\alpha_1, \beta_1)$ and $m_2^+ := (\alpha_2, \beta_2)$ is the desired semi-pullback. $\dashv$

Now we know that we may construct from a co-span of stochastic relations a span. Let us have a look at a typical situation in which such a co-span may occur.

Example 4.10.21 Consider the modal logic from Example 4.1.11 again, and interpret the logic through stochastic relations $K : S \rightsquigarrow S$ and $L : T \rightsquigarrow T$ over the Polish spaces S and T. The equivalence relations $\sim_K$ and $\sim_L$ are defined as in Example 4.4.19. Because we have only countably many formulas, these relations are smooth. For readability, denote the equivalence class associated with $\sim_K$ by $[\cdot]_K$, similar for $[\cdot]_L$. Because $\sim_K$ and $\sim_L$ are smooth, the factor spaces S/K resp. T/L are analytic spaces, when equipped with the final σ-algebra with respect to η_K resp. η_L by Proposition 4.4.22. The factor relation $K_F : S/K \rightsquigarrow S/K$ is then the unique relation which makes this diagram commutative:

$$\begin{tikzcd} S \arrow[r, "\eta_K"] \arrow[d, "K"'] & S/K \arrow[d, "K_F"] \\ \mathbb{S}(S) \arrow[r, "\mathbb{S}(\eta_K)"'] & \mathbb{S}(S/K) \end{tikzcd}$$

This translates to $K(s)(\eta_K^{-1}\big[B\big]) = K_F([s]_K)(B)$ for all $B \in \mathcal{B}(S/K)$ and all $s \in X$.

Associate with each formula φ its validity sets $[\![\varphi]\!]_K$ resp. $[\![\varphi]\!]_L$, and call $s \in S$ *logically equivalent* to $t \in T$ iff we have for each formula φ

$$s \in [\![\varphi]\!]_K \Leftrightarrow t \in [\![\varphi]\!]_L.$$

Hence s and t are logically equivalent iff no formula can distinguish state s from state t. Call the stochastic relations K and L *logically equivalent* iff given $s \in S$ there exists $t \in T$ such that s and t are logically equivalent and vice versa.

Logical equivalence

Now assume that K and L are logically equivalent, and consider

$$\Phi := \big\{\langle [s]_K, [t]_L\rangle \mid s \in S \text{ and } t \in T \text{ are logically equivalent}\big\}.$$

Then Φ is the graph of a bijective map; this is easy to see. Denote the map by Φ as well. Since $\Phi^{-1}\big[\eta_L\big[[\![\varphi]\!]_L\big]\big] = \eta_K\big[[\![\varphi]\!]_K\big]$ and since the set $\{\eta_L\big[[\![\varphi]\!]_L\big] \mid \varphi \text{ is a formula}\}$ generates $\mathcal{B}(T/L)$ by Proposition 4.4.26, $\Phi : S/K \to T/L$ is Borel measurable; interchanging the rôles of K and L yields measurability of Φ^{-1}.

Hence we have this picture for logical equivalent K and L:

$$\begin{tikzcd}
 & L \arrow[d, "\eta_L"] \\
K \arrow[r, "\Phi\circ\eta_K"'] & L_F
\end{tikzcd}$$

✌

This example can be generalized to the case that the relations operate on two spaces rather than only on one. Let $K : X \rightsquigarrow Y$ be a transition kernel over the Polish spaces X and Y. Then the pair (κ, λ) of smooth equivalence relations κ on X and λ on Y is called a *congruence* for K iff there exists a transition kernel $K_{\kappa,\lambda} : X/\kappa \rightsquigarrow Y/\lambda$ rendering the diagram commutative:

Congruence

$$\begin{tikzcd}
X \arrow[r, "\eta_\kappa"] \arrow[d, "K"'] & X/\kappa \arrow[d, "K_{\kappa,\lambda}"] \\
Y \arrow[r, "\mathbb{S}(\eta_\lambda)"'] & Y/\lambda
\end{tikzcd}$$

Because η_κ is an epimorphism, $K_{\kappa,\lambda}$ is uniquely determined, if it exists. For a discussion of congruences for stochastic coalgebras, see Sect. 2.6.2. Commutativity of the diagram translates to

$$K(x)(\eta_\lambda^{-1}\left[B\right]) = K_{\kappa,\lambda}([x]_\kappa)(B)$$

Logical equivalence through factors

for all $x \in X$ and all $B \in \mathcal{B}(Y/\lambda)$. Call in analogy to Example 4.10.21 the transition kernels $K_1 : X_1 \rightsquigarrow Y_1$ and $K_2 : X_2 \rightsquigarrow Y_2$ *logically equivalent* iff there exist congruences (κ_1, λ_1) for K_1 and (κ_2, λ_2) for K_2 such that the factor relations K_{κ_1,λ_1} and K_{κ_2,λ_2} are isomorphic.

In the spirit of this discussion, we obtain from Proposition 4.10.20

Theorem 4.10.22 *Logically equivalent stochastic relations over Polish spaces are bisimilar.*

Proof 1. The proof applies Proposition 4.10.20; first it has to show how to satisfy the assumptions of that statement. Let $K_i : X_i \rightsquigarrow Y_i$ be stochastic relations over Polish spaces for $i = 1, 2$. We assume that K_1 is logically equivalent to K_2; hence there exist congruences (κ_i, λ_i) for K_i such that the associated stochastic relations $K_{\kappa_i,\lambda_i} : X_i/\kappa_i \rightsquigarrow Y_i/\lambda_i$ are isomorphic. Denote this isomorphism by (φ, ψ), so $\varphi : X_1/\kappa_1 \to X_2/\kappa_2$ and $\psi : Y_1/\lambda_1 \to Y_2/\lambda_2$ are in particular measurable bijections, so are their inverses.

2. Let $\eta_2 := (\eta_{\kappa_2}, \eta_{\lambda_2})$ be the factor morphisms $\eta_2 : K_2 \to K_{\kappa_2,\lambda_2}$, and put $\eta_1 := (\varphi \circ \eta_{\kappa_1}, \psi \circ \eta_{\lambda_1})$; thus we obtain this co-span of morphisms

$$K_1 \xrightarrow{\eta_1} K_{\kappa_2,\lambda_2} \xleftarrow{\eta_2} K_2$$

Because both X_2/κ_2 and Y_2/λ_2 are analytic spaces on account of κ_2 and λ_2 being smooth (see Proposition 4.4.22), we apply Proposition 4.10.20 and obtain a mediating relation $M : A \rightsquigarrow B$ with Polish A and B such that the projections $\alpha_i : A \to X_i$ and $\beta_i : B \to Y_i$ are morphisms for $i = 1, 2$. Here

$$\begin{aligned} A &:= \{\langle x_1, x_2\rangle \mid \varphi([x_1]_{\kappa_1}) = [x_2]_{\kappa_2}\}, \\ B &:= \{\langle y_1, y_2\rangle \mid \varphi([y_1]_{\lambda_1}) = [y_2]_{\lambda_2}\}. \end{aligned}$$

It remains to be demonstrated that the σ-algebra of common events, viz., the intersection $\beta_1^{-1}\left[\mathcal{B}(Y_1)\right] \cap \beta_2^{-1}\left[\mathcal{B}(Y_2)\right]$ is not trivial.

3. Let $U_2 \in \mathcal{B}(Y_2)$ be λ_2-invariant. Then $\eta_{\lambda_2}\big[U_2\big] \in \mathcal{B}(Y_2/\lambda_2)$, because $U_2 = \eta_{\lambda_2}^{-1}\big[\eta_{\lambda_2}\big[U_2\big]\big]$ on account of U_2 being λ_2-invariant. Thus $U_1 := \eta_{\lambda_1}^{-1}\big[\psi^{-1}\big[\eta_{\lambda_2}\big[U_2\big]\big]\big]$ is an λ_1-invariant Borel set in Y_1 with

$$\begin{aligned}\langle y_1, y_2\rangle \in (Y_1 \times U_2) \cap B &\Leftrightarrow y_2 \in U_2 \text{ and } \psi([y_1]_{\lambda_1}) = [y_2]_{\lambda_2}\\ &\Leftrightarrow \langle y_1, y_2\rangle \in (U_1 \times U_2) \cap B.\end{aligned}$$

One shows in exactly the same way

$$\langle y_1, y_2\rangle \in (U_1 \times Y_2) \cap B \Leftrightarrow \langle y_1, y_2\rangle \in (U_1 \times U_2) \cap B.$$

Consequently, $(U_1 \times U_2) \cap B$ belongs to both $\beta_1^{-1}\big[\mathcal{B}(Y_1)\big]$ and $\beta_2^{-1}\big[\mathcal{B}(Y_2)\big]$, so that this intersection is not trivial. ⊣

Call a class $\mathfrak{A}$ of spaces *closed under bisimulations* if the mediating relation for stochastic relations over spaces from $\mathfrak{A}$ is again defined over spaces from $\mathfrak{A}$. Then the result above shows that Polish spaces are closed under bisimulations. This generalizes a result by Edalat [Eda99] and Desharnais et al. [DEP02] which demonstrates—through a completely different approach—that analytic spaces are closed under bisimulations; Sánchez Terraf [ST11] has shown that general measurable spaces are not closed under bisimulations. In view of von Neumann's Selection Theorem 4.6.10, it might be interesting to see whether complete measurable spaces are closed.

We present finally a situation in which no semi-pullback exists. A first example in this direction was presented in [ST11, Theorem 12]. It is based on the extension of Lebesgue measure to a σ-algebra which does contain the Borel sets of [0, 1] augmented by a nonmeasurable set, and it shows that one can construct Markov transition systems which do not have a semi-pullback. The example below extends this by showing that one does not have to consider transition systems, but that a look at the measures on which they are based suffices.

Example 4.10.23 A morphism $f : (X, \mathcal{A}, \mu) \to (Y, \mathcal{B}, \nu)$ of measure spaces is an $\mathcal{A}$-$\mathcal{B}$-measurable map $f : X \to Y$ such that $\nu = \mathbb{M}(f)(\mu)$. Since each finite measure can be viewed as a transition kernel, this is a special case of morphisms for transition kernels. If $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{A}$ with μ an extension to ν, then the identity is a morphism $(X, \mathcal{A}, \mu) \to (X, \mathcal{B}, \nu)$.

Denote Lebesgue measure on $([0,1], \mathcal{B}([0,1]))$ by λ. Assuming the Axiom of Choice, we know that there exists $W \subseteq [0,1]$ with $\lambda_*(W) = 0$ and $\lambda^*(W) = 1$ by Lemma 1.7.7. Denote by $\mathcal{A}_W := \sigma(\mathcal{B}([0,1]) \cup \{W\})$ the smallest σ-algebra containing the Borel sets of $[0,1]$ and W. Then we know from Exercise 4.6 that we can find for each $\alpha \in [0,1]$ a measure μ_α on $\mathcal{A}_W$ which extends λ such that $\mu_\alpha(W) = \alpha$.

Hence by the remark just made, the identity yields a morphism $f_\alpha : ([0,1], \mathcal{A}_W, \mu_\alpha) \to ([0,1], \mathcal{B}([0,1]), \lambda)$. Now let $\alpha \neq \beta$; then

$$([0,1], \mathcal{A}_W, \mu_\alpha) \xrightarrow{f_\alpha} ([0,1], \mathcal{B}([0,1]), \lambda) \xleftarrow{f_\beta} ([0,1], \mathcal{A}_W, \mu_\beta)$$

is a co-span of morphisms.

We claim that this co-span does not have a semi-pullpack. In fact, assume that $(P, \mathcal{P}, \rho)$ with morphisms π_α and π_β is a semi-pullback; then $f_\alpha \circ \pi_\alpha = f_\beta \circ \pi_\beta$, so that $\pi_\alpha = \pi_\beta$, and $\pi_\alpha^{-1}\big[W\big] = \pi_\beta^{-1}\big[W\big] \in \mathcal{P}$. But then

$$\alpha = \mu_\alpha(W) = \rho(\pi_\alpha^{-1}\big[W\big]) = \rho(\pi_\beta^{-1}\big[W\big]) = \mu_\beta(W) = \beta.$$

This contradicts the assumption that $\alpha \neq \beta$. ✌

This example shows that the topological assumptions imposed above are indeed necessary. It assumes the Axiom of Choice, so one might ask what happens if this axiom is replaced by the Axiom of Determinacy. We know that the latter one implies that each subset of the unit interval is λ-measurable by Theorem 1.7.14, so $\lambda_*(W) = \lambda^*(W)$ holds for each $W \subseteq [0,1]$. Then at least the construction above does not work. On the other hand, we made free use of Tihonov's Theorem, which is known to be equivalent to the Axiom of Choice [Her06, Theorem 4.68], so there is probably no escape from the Axiom of Choice.

4.10.4 Case Study: Quotients for Stochastic Relations

As Monty Python used to say, "And now for something completely different!" We will deal now with quotients for stochastic relations, perceived as morphisms in the Kleisli category over the monad which is given by the subprobability functor. We will first have a look at surjective maps as epimorphisms in the category of sets, explaining the problem there, show that a straightforward approach gleaned from the category of sets does not appear promising, and show then that measurable selections are the appropriate tool for tackling the problem. The example demonstrates also that constructions which are fairly straightforward for sets may become somewhat involved in the category of stochastic relations.

For motivation, we start with surjective maps on a fixed set M, serving as a domain. Let $f : M \to X$ and $g : M \to Y$ be onto, and define the partial order $f \leq g$ iff $f = \zeta \circ g$ for some $\zeta : Y \to X$. Clearly, $\leq$ is reflexive and transitive; the equivalence relation $\sim$ defines through $f \sim g$ iff $f \leq g$ and $g \leq f$ are of interest here. Thus $f = \zeta \circ g$ and $g = \xi \circ f$ for suitable $\zeta : Y \to X$ and $\xi : X \to Y$. Because surjective maps are epimorphisms in the category of sets with maps as morphisms, we obtain $\zeta \circ \xi = id_X$ and $\xi \circ \zeta = id_Y$. Hence ζ and ξ are bijections. The surjections f and g, both with domain M, are equivalent iff there exists a bijection β with $f = \beta \circ g$. This is called a *quotient object* for M. We know that the surjection $f : M \to Y$ can be factored as $f = \tilde{f} \circ \eta_{\ker(f)}$ with $\tilde{f} : [x]_{\ker(f)} \mapsto f(x)$ as the bijection; see Proposition 2.1.26. Thus for maps, the quotient objects for M may be identified through the quotient maps $\eta_{\ker(f)}$; in a similar way, the quotient objects in the category of groups can be identified through normal subgroups; see [ML97, V.7] for a discussion. Quotients are algebraically of interest. Quotient object in ***Set***

We turn to stochastic relations. The subprobability functor on the category of measurable spaces is the functorial part of the Giry monad, and the stochastic relations are just the Kleisli morphism for this monad; see the discussion in Example 2.4.8 on page 173. Let $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ be a stochastic relation; then Exercise 4.14 shows that

$$\overline{K}(\mu) : B \mapsto \int_X K(x)(B) \, d\mu(x)$$

defines a $\wp(X, \mathcal{A})$-$\wp(Y, \mathcal{B})$-measurable map $\mathbb{S}(X, \mathcal{A}) \to \mathbb{S}(Y, \mathcal{B})$; $\overline{K}$ is the *Kleisli map* associated with the Kleisli morphism K (it should not be confused with the completion of K as discussed in Sect. 4.6.2). It is clear that $K \mapsto \overline{K}$ is injective, because $\overline{K}(\delta_x) = K(x)$.

It will be helpful to evaluate the integral with respect to $\overline{K}(\mu)$: Let $g : Y \to \mathbb{R}$ be bounded and measurable; then

$$\int_Y g \, dK(\mu) = \int_X \int_Y g(y) \, dK(x)(y) \, d\mu(x). \tag{4.20}$$

This has been anticipated in Example 2.4.8. In order to establish this, assume first that $g = \chi_B$ for $B \in \mathcal{B}$, then both sides evaluate to $K(\mu)(B)$, so the representation is valid for indicator functions. Linearity of the integral yields the representation for step functions. Since we may find for general g a sequence $(g_n)_{n\in\mathbb{N}}$ of step functions with $\lim_{n\to\infty} g_n(y) = g(y)$ for all $y \in Y$ and since g is bounded, hence integrable with respect to all finite measures, we obtain from Lebesgue's Dominated Convergence Theorem 4.8.6 that

$$\begin{aligned}
\int_Y g \, d\overline{K}(\mu) &= \lim_{n\to\infty} \int_Y g_n \, d\overline{K}K(\mu) \\
&= \lim_{n\to\infty} \int_X \int_Y g_n(y) \, dK(x)(y) \, d\mu(x) \\
&= \int_X \lim_{n\to\infty} \int_Y g_n(y) \, dK(x)(y) \, d\mu(x) \\
&= \int_X \int_Y g(y) \, dK(x)(y) \, d\mu(x).
\end{aligned}$$

This gives the desired representation.

The Kleisli map is related to the convolution operation defined in Example 4.9.6.

Lemma 4.10.24 *Let* $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ *and* $L : (Y, \mathcal{B}) \rightsquigarrow (Z, \mathcal{C})$*; then* $\overline{L * K} = \overline{L} \circ \overline{K}$.

Proof Evaluate both the left-hand and the right-hand sides for $\mu \in \mathbb{S}(X, \mathcal{A})$ and $C \in \mathcal{C}$:

$$\begin{aligned}
\overline{L * K}(\mu)(C) &= \int_X \int_Y L(y)(C) \, dK(x)(y) \, d\mu(x) \\
&= \int_Y L(y)(C) \, d\overline{K}(\mu)(y) && \text{(by (4.20))} \\
&= \overline{L}(\overline{K})(\mu)(C).
\end{aligned}$$

This implies the desired equality. $\dashv$

Associate with each measurable $f : Y \to Z$ a stochastic relation $\delta_f : Y \rightsquigarrow Z$ through $\delta_f(y)(C) := \delta_y(f^{-1}[C])$, then $\delta_f = \mathbb{S}(f) \circ \delta$, and a direct computation shows $\delta_f * K = \mathbb{S}(f) \circ K$. In fact,

$$\begin{aligned}
\big(\delta_f * K\big)(x)(C) &= \int_Y \delta_f(y)(C)\ K(x)(dy) \\
&= \int_Y \chi_{f^{-1}\left[C\right]}(y)\ K(x)(dy) \\
&= K(x)(f^{-1}\left[C\right]) \\
&= \big(\mathbb{S}(f) \circ K\big)(x)(C).
\end{aligned}$$

On the other hand, if $f : W \to X$ is measurable, then

$$\big(K * \delta_f\big)(w)(B) = \int_X K(x)(B)\, \delta_f(w)(dx) = (K \circ f)(w)(B).$$

In particular, it follows that $e_X := \mathbb{S}(id_X)$ is the neutral element: $K = e_X * K = K * e_X = K$. Recall that K is an epimorphism in the Kleisli category iff $L_1 * K = L_2 * K$ implies $L_1 = L_2$ for any stochastic relations $L_1, L_2 : (Y, \mathcal{B}) \rightsquigarrow (Z, \mathcal{C})$. Lemma 4.10.24 tells us that if the Kleisli map $\overline{K}$ is onto, then K is an epimorphism. Now let $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ and $L : (X, \mathcal{A}) \rightsquigarrow (Z, \mathcal{C})$ be stochastic relations, and assume that both K and L are epis. Define as above

$$\begin{aligned}
&K \le L \Leftrightarrow K = J * L \text{ for some } J : (Z, \mathcal{C}) \rightsquigarrow (Y, \mathcal{B}) \\
&K \approx L \Leftrightarrow K \le L \text{ and } L \le K.
\end{aligned}$$

Hence we can find in case $K \le L$ a stochastic relation J such that

$$K(x)(B) = \int_Z J(z)(B)\, dL(x)(z)$$

for $x \in X$ and $B \in \mathcal{B}$.

We will deal for the rest of this section with Polish spaces. Fix X as a Polish space. For identifying the quotients with respect to Kleisli morphisms, one could be tempted to mimic the approach observed for the sets as outlined above. This is studied in the next example.

Example 4.10.25 Let $K : X \rightsquigarrow Y$ be a stochastic relation with Polish Y which is an epi. $X/\ker(K)$ is an analytic space, since $K : X \to \mathbb{S}(Y)$ is a measurable map into the Polish space $\mathbb{S}(Y)$ by Proposition 4.10.10, so that $\ker(K)$ is smooth. Define the map $E_K : X \to \mathbb{S}(X/\ker(K))$

through $E_K(x) := \delta_{[x]_{\ker(K)}}$; hence we obtain for each $x \in X$ and each Borel set $G \in \mathcal{B}(X/\ker(K))$

$$E_K(x)(G) = \delta_{[x]_{\ker(K)}}(G) = \delta_x(\eta_{\ker(K)}^{-1}\left[G\right]) = \mathbb{S}(\eta_{\ker(K)})(\delta_x)(G).$$

Thus E_K is an epi as well: Take $\mu \in \mathbb{S}(X)$ and $G \in \mathcal{B}(X/\ker(K))$; then

$$\begin{aligned} \overline{E}_K(\mu)(G) &= \int_X E_K(x)(G)\, d\mu(x) &&= \int_X \delta_x(\eta_{\ker(K)}^{-1}\left[G\right])\, d\mu(x) \\ &= \mu(\eta_{\ker(K)}^{-1}\left[G\right]) &&= \mathbb{S}(\eta_{\ker(K)})(\mu)(G), \end{aligned}$$

so that $\overline{E}_K = \mathbb{S}(\eta_{\ker(K)})$; since the image of a surjective map under $\mathbb{S}$ is surjective again by Proposition 4.6.11, we conclude that E_K is an epi. Now define for $x \in X$ the map

$$\tilde{K}([x]_{\ker(K)}) := K(x);$$

then the construction of the final σ-algebra on $X/\ker(K)$ shows that $\tilde{K}$ is well defined and constitutes a stochastic relation $\tilde{K} : X/\ker(K) \rightsquigarrow Y$. Moreover we obtain for $x \in X, H \in \mathcal{B}(Y)$ by the change of variables formula in Corollary 4.8.9

$$\begin{aligned} (\tilde{K} * E_K)(x)(H) &= \int_{X/\ker(K)} \tilde{K}(t)(H)\, dE_K(x)(t) \\ &= \int_{X/\ker(K)} \tilde{K}(t)(H)\, d\mathbb{S}(\eta_{\ker(K)})(\delta_x)(t) \\ &= \int_X \tilde{K}([w]_{\ker(K)})(H)\, d\delta_x(w) \\ &= \int_X K(w)(H)\, d\delta_x(w) \\ &= K(x)(H). \end{aligned}$$

Consequently, K can be factored as $K = \tilde{K} * E_K$ with the epi E_K. But there is no reason why in general $\tilde{K}$ should be invertible; for this to hold, the map $\overline{\tilde{K}} : \mathbb{S}(X/\ker(K)) \to \mathbb{S}(Y)$ is required to be injective. Hence $K \approx E_K$ holds only in very special cases. ✌

This last example indicates that a characterization of quotients for the Kleisli category at least for the Giry monad cannot be derived directly by translating a characterization for the underlying category from the category of sets.

For the rest of the section, we discuss the Kleisli category for the Giry monad over Polish spaces; hence we deal with stochastic relations. Let X, Y, and Z be Polish, and fix $K : X \rightsquigarrow Y$ and $L : X \rightsquigarrow Z$ so that $K \approx L$. Hence there exists $J : Y \rightsquigarrow Z$ with inverse $H : Z \rightsquigarrow Y$ and $L = J * K$ and $K = H * L$. Because both K and L are epis, we obtain these simultaneous equations

$$H * J = e_Y \text{ and } J * H = e_Z.$$

They entail

$$\int_Z H(z)(B)\, dJ(y)(z) = \delta_y(B) \text{ and } \int_Y J(y)(C)\, dH(z)(y) = \delta_z(C)$$

for all $y \in Y, z \in Z$ and $B \in \mathcal{B}(Y), C \in \mathcal{B}(Z)$. Because singletons are Borel sets, these equalities imply

$$\int_Z H(z)(\{y\})\, dJ(y)(z) = 1 \text{ and } \int_Y J(y)(\{z\})\, dH(z)(y) = 1.$$

Consequently, we obtain

$$\forall y \in Y : J(y)(\{z \in Z \mid H(z)(\{y\}) = 1\}) = 1,$$
$$\forall z \in Z : H(z)(\{y \in Y \mid J(y)(\{z\}) = 1\}) = 1.$$

Proposition 4.10.26 *There exist Borel maps $f : Y \to Z$ and $g : Z \to Y$ such that $H\big(f(y)\big)(\{y\}) = 1$ and $J\big(g(z)\big)(\{z\}) = 1$ for all $y \in Y, z \in Z$.*

Proof 1. Define $P := \{\langle y, z\rangle \in Y \times Z \mid H(z)(\{y\}) = 1\}$ and $Q := \{\langle z, y\rangle \in Z \times Y \mid J(y)(\{z\}) = 1\}$; then P and Q are Borel sets. We establish this for P; the argumentation for Q is very similar.

2. With a view toward Proposition 4.3.31, we may and do assume that $H : Z \to \mathbb{S}(Y)$ is continuous. Let $\big(\langle y_n, z_n\rangle\big)_{n\in\mathbb{N}}$ be a sequence in P with $\langle y_n, z_n\rangle \to \langle y, z\rangle$; hence the sequence $\big(H(z_n)\big)_{n\in\mathbb{N}}$ converges weakly $H(z)$. Given $m \in \mathbb{N}$, there exists $n_0 \in \mathbb{N}$ such that $y_n \in V_{1/m}(y)$ for all $n_0 \geq n$, where $V_{1/m}(y)$ is the closed ball of radius $1/m$ around y. Since H is weakly continuous, we obtain

$$\limsup_{n\to\infty} H(z_n)\big(V_{1/m}(y)\big) \leq H(z)\big(V_{1/m}(y)\big)$$

from Proposition 4.1.35; hence

$$H(z)\big(V_{1/m}(y)\big) = 1.$$

Because

$$\bigcap_{m \in \mathbb{N}} V_{1/m}(y) = \{y\},$$

we conclude $H(z)(\{y\}) = 1$; thus $\langle y, z \rangle \in P$. Consequently, P is a closed subset of $Y \times Z$, hence a Borel set.

3. Since P is closed, the cut P_y at y is closed as well, and we have

$$J(y)(P_y) = J(y)(\{z \in Z \mid H(z)(\{y\}) = 1\} = 1;$$

thus we obtain $\text{supp}(J(y)) \subseteq P_y$, because the support $\text{supp}(J(y))$ is the smallest closed set C with $J(y)(C) = 1$. Since $y \mapsto \text{supp}(J(y))$ is measurable, as we have seen in Example 4.7.5, we obtain from Theorem 4.7.2 a measurable map $f : Y \to Z$ with $f(y) \in \text{supp}(J(y)) \subseteq P_y$ for all $y \in Y$; thus $H(f(y))(\{y\}) = 1$ for all $y \in Y$.

4. In the same way, we obtain measurable $g : Z \to Y$ with the desired properties. $\dashv$

Discussing the maps f, g obtained above from H and J, we see that

$$H \circ f = e_Y \text{ and } J \circ g = e_Z,$$

and we calculate through the change of variables formula in Corollary 4.8.9 for each $z_0 \in Y$ and each $H \in \mathcal{B}(Z)$

$$\begin{aligned}
\big(H*(\mathbb{S}(f) \circ H)\big)(z_0)(H) &= \int_Z H(z)(H)\ (\mathbb{S}(f) \circ H)(z_0)(dz) \\
&= \int_Y H(f(y))(H)\ H(z_0)(dy) \\
&= \int_Y \delta_y(H)\ H(z_0)(dy) \\
&= H(z_0)(H).
\end{aligned}$$

Thus $H*(\mathbb{S}(f) \circ H) = H$, and because H is a mono, we infer that $\mathbb{S}(f) \circ H = e_Z$. Since

$$\mathbb{S}(f) \circ H = (e_Z \circ f)*H = J*H,$$

we infer on account of H being an epi that $J = e_Z \circ f$. Similarly we see that $H = e_Y \circ g$.

Lemma 4.10.27 *Given stochastic relations $J : Y \rightsquigarrow Z$ and $H : Z \rightsquigarrow Y$ with $H * J = e_Y$ and $J * H = e_Z$, there exist Borel isomorphisms $f : Y \to Z$ and $g : Z \to Y$ with $J = e_Z \circ f$ and $H = e_Y \circ g$.*

Proof We infer for $y \in Y$ from

$$\begin{aligned}\delta_y(G) &= e_Y(y)(G)\\ &= (H * J)(y)(G)\\ &= \int_Z H(z)(G)\, dJ(y)(z)\\ &= \delta_{f(y)}(g^{-1}[G])\\ &= \delta_y(f^{-1}[g^{-1}[G]])\end{aligned}$$

for all Borel sets $G \in \mathcal{B}(Y)$ that $g \circ f = id_Y$; similarly, $f \circ g = id_Z$ is inferred. Hence the Borel maps f and g are bijections and thus Borel isomorphisms. ⊣

This yields a characterization of the quotient equivalence relation in the Kleisli category for the Giry monad.

Proposition 4.10.28 *Assume the stochastic relations $K : X \rightsquigarrow Y$ and $L : X \rightsquigarrow Z$ are both epimorphisms with respect to Kleisli composition; then these conditions are equivalent:*

1. $K \approx L$.
2. $L = \mathbb{S}(f) \circ K$ *for a Borel isomorphism* $f : Y \to Z$.

Proof 1 $\Rightarrow$ 2: Because $K \approx L$, there exists an invertible $J : Y \rightsquigarrow Z$ with inverse $H : Z \rightsquigarrow Y$ and $L = J * K$. We infer from Lemma 4.10.27 the existence of a Borel isomorphism $f : Y \to Z$ such that $J = \eta_Z \circ f$. Consequently, we have for $x \in X$ and the Borel set $H \in \mathcal{B}(Z)$

$$\begin{aligned}L(x)(H) &= \int_Y J(y)(H)\, dK(x)(y)\\ &= \int_Y \delta_{f(y)}(H)\, dK(x)(y)\\ &= K(x)(f^{-1}[H])\\ &= \big(\mathbb{S}(f) \circ K\big)(x)(H).\end{aligned}$$

2 $\Rightarrow$ 1: If $L = \mathbb{S}(f) \circ K = (\eta_Z \circ f) * K$ for the Borel isomorphism $f : Y \to Z$, then $K = (\eta_Y \circ g) * L$ with $g : Z \to Y$ as the inverse to f. ⊣

Consequently, given the epimorphisms $K : X \rightsquigarrow Y$ and $L : X \rightsquigarrow Z$, the relation $K \approx L$ entails their base spaces Y and Z being Borel isomorphic and vice versa. Hence the Borel isomorphism classes are the quotient objects for this relation.

This classification should be complemented by a characterization of epimorphic Kleisli morphisms for this monad. This seems to be an open question.

4.11 L_p-Spaces

We will construct for a measure space $(X, \mathcal{A}, \mu)$ a family $\{L_p\mu \mid 1 \leq p \leq \infty\}$ of Banach spaces. Some properties of these spaces are discussed now; in particular we will identify their dual spaces. The case $p = 2$ gives the particularly interesting space $L_2(\mu)$, which is a Hilbert space under the inner product $\langle f, g\rangle \mapsto \int_X f \cdot g \, d\mu$. Hilbert spaces have some properties which will turn out to be helpful and which will be exploited for the underlying measure spaces. For example, von Neumann obtained from a representation of their continuous linear maps both the Lebesgue decomposition and the Radon–Nikodym Theorem derivative in one step! We join Rudin's exposition [Rud74, Sect. 6] in giving the truly ravishing proof here. But we are jumping ahead. After investigating the basic properties of Hilbert spaces including the closest approximation property and the identification of continuous linear functions, we move to a discussion of the more general L_p-spaces and investigate the positive linear functionals on them.

Some important developments like the definition of signed measures are briefly touched, while some are not. The topics which had to be omitted here include the weak topology induced by L_q on L_p for conjugate pairs p, q; this would have required some investigations into convexity, which would have led into a wonderful, wondrous but unfortunately faraway country.

The last section deals with disintegration as an application of both the Radon–Nikodym derivative and the measure extension theorem. It deals with the problem of decomposing a finite measure on a product into its projection onto the first component and an associated transition kernel.

4.11.1 A Spoonful Hilbert Space Theory

Let H be a real vector space. A map $(\cdot, \cdot) : H \times H \to \mathbb{R}$ is said to be an *inner product* iff these conditions hold for all $x, y, z \in H$ and all $\alpha, \beta \in \mathbb{R}$: Inner product

1. $(x, y) = (y, x)$, so the inner product is commutative.
2. $(\alpha \cdot x + \beta \cdot z, y) = \alpha \cdot (x, y) + \beta \cdot (z, y)$, so the inner product is linear in the first and hence also in the second component.
3. $(x, x) \geq 0$, and $(x, x) = 0$ iff $x = 0$.

We confine ourselves to real vector spaces. Hence the laws for the inner product are somewhat simplified in comparison to vector spaces over the complex number. There one would, e.g., postulate that (y, x) is the complex conjugate for (x, y).

The inner product is the natural generalization of the scalar product in Euclidean spaces

$$(\langle x_1, \dots, x_n \rangle, \langle y_1, \dots, y_n \rangle) := \sum_{i=1}^{n} x_i \cdot y_i,$$

which satisfies these laws, as one verifies readily.

We fix an inner product $(\cdot, \cdot)$ on H. Define the norm of $x \in H$ through

$$\|x\| := \sqrt{(x, x)};$$

this is possible because $(x, x) \geq 0$. We show that this yields a normed space indeed.

The map $\| \cdot \|$ has a very appealing geometric property, which is known as the *parallelogram law*: The sum of the squares of the diagonals is the sum of the squares of the sides in a parallelogram. Parallelogram law

$$\|x + y\|^2 + \|x - y\|^2 = 2 \cdot \|x\|^2 + 2 \cdot \|y\|^2$$

holds for all $x, y \in H$; see Exercise 4.33.

Before investigating $\| \cdot \|$ in detail, we need the *Schwarz inequality* as a tool. It relates the norm to the inner product of two elements. Here it is.

Schwarz inequality

Lemma 4.11.1 $|(x, y)| \leq \|x\| \cdot \|y\|$.

Proof Let $a := \|x\|^2$, $b := \|y\|^2$, and $c := |(x, y)|$. Then $c = t \cdot (x, y)$ with $t \in \{-1, +1\}$. We have for each real r

$$0 \leq (x - r \cdot t \cdot y, x - r \cdot t \cdot y) = (x, x) - 2 \cdot r \cdot t \cdot (x, y) + r^2 \cdot (y, y);$$

thus $a - 2 \cdot r \cdot c + r^2 \cdot b \geq 0$. If $b = 0$, we must also have $c = 0$; otherwise the inequality would be false for large positive r. Hence the inequality is true in this case. So we may assume that $b \neq 0$. Put $r := c/b$, so that $a \geq c^2/b$, so that $a \cdot b \geq c^2$, from which the desired inequality follows. $\dashv$

Schwarz's inequality will help in establishing that a vector space with an inner product is a normed space, as introduced in Definition 3.6.38.

Proposition 4.11.2 *Let H be a real vector space with an inner product; then $(H, \|\cdot\|)$ is a normed space.*

Proof It is clear from the definition of the inner product that $\|\alpha \cdot x\| = |\alpha| \cdot \|x\|$ and that $\|x\| = 0$ iff $x = 0$; the crucial point is the triangle inequality. We have

$$\begin{aligned}
\|x + y\|^2 = (x + y, x + y) &= \|x\|^2 + \|y\|^2 \\
&\quad + 2 \cdot (x, y) \\
&\leq \|x\|^2 + 2 \cdot \|x\| \cdot \|y\| + \|y\|^2 \qquad \text{(by Lemma 4.11.1)} \\
&= (\|x\| + \|y\|)^2.
\end{aligned}$$

$\dashv$

Thus each inner product space yields a normed space; consequently it spawns a metric space through $\langle x, y \rangle \mapsto \|x - y\|$. Finite dimensional vector spaces $\mathbb{R}^n$ are Hilbert spaces under the inner product mentioned above. It produces for $\mathbb{R}^n$ the familiar Euclidean distance

$$\|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

We will meet square integrable functions as another class of Hilbert spaces, but before discussing them, we need some preparations.

Corollary 4.11.3 *The maps $x \mapsto \|x\|$ and $x \mapsto (x, y)$ with fixed $y \in H$ are continuous.*

Proof We obtain from $\|x\| \leq \|y\| + \|x - y\|$ and $\|y\| \leq \|x\| + \|x - y\|$ that $\big|\|x\| - \|y\|\big| \leq \|x - y\|$; hence the norm is continuous. From Schwarz's inequality we see that $|(x, y) - (x', y)| = |(x - x', y)| \leq \|x - x'\| \cdot \|y\|$, which shows that $(\cdot, y)$ is continuous. ⊣

From the properties of the inner product, it is apparent that $x \mapsto (x, y)$ is a continuous linear functional in the following sense:

Definition 4.11.4 *Let H be an inner product space with norm $\|\cdot\|$. A linear map $L : H \to \mathbb{R}$ which is continuous in the norm topology is called a* continuous linear functional *on H.*

We saw linear functionals already in Sect. 1.5.4, where the domination through a sublinear map was concentrated on, leading to an extension. This, however, is not the focus in the present discussion.

If $L : H \to \mathbb{R}$ is a continuous linear functional, then its *kernel* $Kern(L)$

$$Kern(L) := \{x \in H \mid L(x) = 0\}$$

is a closed linear subspace of H, i.e., is a real vector space in its own right. Note that

$$\langle x, y\rangle \in \ker(L) \text{ iff } x - y \in Kern(L),$$

so that both versions of kernels are related in an obvious way.

Say that $x \in H$ is *orthogonal* to $y \in H$ iff $(x, y) = 0$, and denote this by $x \bot y$. This is the generalization of the familiar concept of orthogonality in Euclidean spaces, which is formulated also in terms of the inner product. Given a linear subspace $M \subseteq H$, define the *orthogonal complement* $M^\bot$ of M as $M^\bot$

$$M^\bot := \{y \in H \mid x \bot y \text{ for all } x \in M\}.$$

The orthogonal complement is a linear subspace as well, and it is closed by Corollary 4.11.3, since $M = \bigcap_{x \in M}\{y \in H \mid (x, y) = 0\}$. Then $M \cap M^\bot = \{0\}$, since a vector $z \in M \cap M^\bot$ is orthogonal to itself; hence $(z, z) = 0$, which implies $z = 0$.

Hilbert spaces are introduced now as those linear spaces for which this metric is complete. Our goal is to show that continuous linear functionals on a Hilbert space H are given exactly through the inner product.

Definition 4.11.5 *A* Hilbert space *is a real vector space with an inner product so that the induced metric is complete.*

Note that we fix the metric for which the space is complete, noting that completeness is not a property of the underlying topological space but rather of a specific metric. It is also worth noting that a Hilbert space is a topological group, as discussed in Example 3.1.25, and hence that it is a complete uniform space.

Convex

Recall that a subset $C \subseteq H$ is called *convex* iff it contains with two points also the straight line between them and thus iff $\alpha \cdot x + (1-\alpha) \cdot y \in C$, whenever $x, y \in C$ and $0 \leq \alpha \leq 1$.

A key tool for our development is the observation that a closed convex subset of a Hilbert space has a unique element of smallest norm. This property is familiar from Euclidean spaces. Visualize a compact convex set in $\mathbb{R}^3$; then this set has a unique point which is closest to the origin. The statement below is more general, because it refers to closed and convex sets.

Proposition 4.11.6 *Let $C \subseteq H$ be a closed and convex subset of the Hilbert space H. Then there exists a unique $y \in C$ such that $\|y\| = \inf_{z \in C} \|z\|$.*

Plan

Proof 0. We construct a sequence $(x_n)_{n \in \mathbb{N}}$ in C, the norms of which converge against the infimum of the vectors' length in C. Using the parallelogram law, we find that the vectors themselves converge to a point of minimal length, which by convexity must belong to the C.

1. Put $r := \inf_{z \in C} \|z\|$, and let $x, y \in C$; hence by convexity $(x + y)/2 \in C$ as well. The parallelogram law gives

$$\begin{aligned}\|x - y\|^2 &= 2 \cdot \|x\|^2 + 2 \cdot \|y\|^2 - 4 \cdot \|(x + y)/2\|^2 \\ &\leq 2 \cdot \|x\|^2 + 2 \cdot \|y\|^2 - 4 \cdot r^2.\end{aligned}$$

Hence if we have two vectors $x \in C$ and $y \in C$ of minimal norm, we obtain $x = y$. Thus, if such a vector exists, it must be unique.

2. Let $(x_n)_{n \in \mathbb{N}}$ be a sequence in C such that $\lim_{n \to \infty} \|x_n\| = r$. At this point, we have only information about the sequence $\big(\|x_x\|\big)_{n \in \mathbb{N}}$ of real numbers, but we can actually show that the sequence proper is a Cauchy sequence. It works like this. We obtain, again from the parallelogram law, the estimate

$$\|x_n - x_m\| \leq 2 \cdot (\|x_n\|^2 + \|x_m\|^2 - 2 \cdot r^2),$$

so that for each $\epsilon > 0$ we find n_0 such that $\|x_n - x_m\| < \epsilon$ if $n, m \geq n_0$. Hence $(x_n)_{n\in\mathbb{N}}$ is actually a Cauchy sequence, and since H is complete, we find some x such that $\lim_{n\to\infty} x_n = x$. Clearly, $\|x\| = r$, and since C is closed, we infer that $x \in C$. $\dashv$

Note how the geometric properties of an inner product space, formulated through the parallelogram law, and the metric property of being complete cooperate.

This unique approximation property has two remarkable consequences. The first one establishes for each element $x \in H$ a unique representation as $x = x_1 + x_2$ with $x_1 \in M$ and $x_2 \in M^\perp$ for a closed linear subspace M of H, and the second one shows that the only continuous linear maps on the Hilbert space H are given by the maps $\lambda x.(x, y)$ for $y \in H$. We need the first one for establishing the second one, so both find their place in this somewhat minimal discussion of Hilbert spaces.

Proposition 4.11.7 *Let H be a Hilbert space, $M \subseteq H$ a closed linear subspace. Each $x \in H$ has a unique representation $x = x_1 + x_2$ with $x_1 \in M$ and $x_2 \in M^\perp$.*

Proof 1. If such a representation exists, it must be unique. In fact, assume that $x_1+x_2 = x = y_1+y_2$ with $x_1, y_1 \in M$ and $x_2, y_2 \in M^\perp$; then $x_1-y_1 = y_2-x_2 \in M\cap M^\perp$, which implies $x_1 = y_1$ and $x_2 = y_2$ by the remark above.

2. Fix $x \in H$, we may and do assume that $x \notin M$, and define $C := \{x - y \mid y \in M\}$, then C is convex, and, because M is closed, it is closed as well. Thus we find an element in C which is of smallest norm, say, $x - x_1$ with $x_1 \in M$. Put $x_2 := x - x_1$, and we have to show that $x_2 \in M^\perp$ and hence that $(x_2, y) = 0$ for any $y \in M$. Let $y \in M$, $y \neq 0$ and choose $\alpha \in \mathbb{R}$ arbitrarily (for the moment, we will fix it later). Then $x_2 - \alpha \cdot y = x - (x_1 + \alpha \cdot y) \in C$, and thus $\|x_2 - \alpha \cdot y\|^2 \geq \|x_2\|^2$. Expanding, we obtain

$$(x_2 - \alpha \cdot y, x_2 - \alpha \cdot y) = (x_2, x_2) - 2 \cdot \alpha \cdot (x_2, y) + \alpha^2 \cdot (y, y) \geq (x_2, x_2).$$

Now put $\alpha := (x_2, y)/(y, y)$; then the above inequality yields

$$-2 \cdot \frac{(x_2, y)^2}{(y, y)} + \frac{(x_2, y)^2}{(y, y)} \geq 0,$$

which implies $-(x_2, y)^2 \geq 0$; hence $(x_2, y) = 0$. Thus $x_2 \in M^\perp$. $\dashv$

Thus H is decomposed into M and $M^{\perp}$ for any closed linear subspace M of H in the sense that each element of H can be written as a sum of elements of M and of $M^{\perp}$, and, even better, this decomposition is unique. These elements are perceived as the projections to the subspaces. In the case that we can represent M as the kernel $\{x \in H \mid L(x) = 0\}$ of a continuous linear map $L : H \to \mathbb{R}$ with $L \neq 0$, we can say actually more.

Lemma 4.11.8 *Given Hilbert space H, let $L : H \to \mathbb{R}$ be a continuous linear functional with $L \neq 0$. Then $Kern(L)^{\perp}$ is isomorphic to $\mathbb{R}$.*

Proof Define $\varphi(y) := L(y)$ for $y \in Kern(L)^{\perp}$. Then $\varphi(\alpha \cdot y + \beta \cdot y') = \alpha \cdot \varphi(y) + \beta \cdot \varphi(y')$ follows from the linearity of L. If $\varphi(y) = \varphi(y')$, then $y - y' \in Kern(L) \cap Kern(L)^{\perp}$, so that $y = y'$; hence φ is one to one. Given $t \in \mathbb{R}$, we find $x \in H$ with $L(x) = t$; decompose x as $x_1 + x_2$ with $x_1 \in Kern(L)$ and $x_2 \in Kern(L)^{\perp}$, then $\varphi(x_2) = L(x - x_1) = t$. Thus φ is onto. Hence we have found a linear and bijective map $Kern(L)^{\perp} \to \mathbb{R}$. $\dashv$

Returning to the decomposition of an element $x \in H$, we fix an arbitrary $y \in Kern(L)\setminus\{0\}$. Then we may write $x = x_1 + \alpha \cdot y$, where $\alpha \in \mathbb{R}$. This follows immediately from Lemma 4.11.8, and it has the consequence we are aiming at.

Theorem 4.11.9 *Let H be a Hilbert space and $L : H \to \mathbb{R}$ be a continuous linear functional. Then there exists $y \in H$ with $L(x) = (x, y)$ for all $x \in H$.*

Proof If $L = 0$, this is trivial. Hence we assume that $L \neq 0$. Thus we can find $z \in Kern(L)^{\perp}$ with $L(z) = 1$; put $y = \gamma \cdot z$ so that $L(y) = (y, y)$. Each $x \in H$ can be written as $x = x_1 + \alpha \cdot y$ with $x_1 \in Kern(L)$. Hence

$$L(x) = L(x_1 + \alpha \cdot y) = \alpha \cdot L(y) = \alpha \cdot (y, y) = (x_1 + \alpha \cdot y, y) = (x, y).$$

Thus $L = \lambda x.(x, y)$ is established. $\dashv$

Thus a Hilbert space does not only determine the space of all continuous linear maps on it, but it *is* actually this space. Wonderful world of Hilbert spaces! This property sets Hilbert spaces apart, making them particularly interesting for applications, e.g., in quantum computing.

The rather abstract view of Hilbert spaces discussed in this section will be put to use now to the more specific case of integrable functions.

4.11.2 The L_p-Spaces are Banach Spaces

We will investigate now the structure of integrable functions for a fixed σ-finite measure space $(X, \mathcal{A}, \mu)$. We will obtain a family of Banach spaces, all of which have some interesting properties. In the course of investigations, we will usually not distinguish between functions which differ only on a set of measure zero (because the measure will not be aware of the differences). For this, we introduced above the equivalence relation $=_\mu$ ("equal μ-almost everywhere") with $f =_\mu g$ iff $\mu(\{x \in X \mid f(x) \neq g(x)\}) = 0$; see Sect. 4.2.1 on page 472. In those cases where we will need to look at the value of a function at certain points, we will make sure that we will point out the difference.

Let us see how this works in practice. Define $\mathcal{L}_1(\mu)$

$$\mathcal{L}_1(\mu) := \{f \in \mathcal{F}(X, \mathcal{A}) \mid \int_X |f| \, d\mu < \infty\};$$

thus $f \in \mathcal{L}_1(\mu)$ iff $f : X \to \mathbb{R}$ is measurable and has a finite μ-integral.

Then this space defines a vector space, which closed with respect to $|\cdot|$; hence we have immediately

Proposition 4.11.10 *$\mathcal{L}_1(\mu)$ is a vector lattice.* $\dashv$

Now put $L_1(\mu)$

$$L_1(\mu) := \{[f] \mid f \in L_1(\mu)\};$$

then we have to explain how to perform the algebraic operations on the equivalence classes (note that we write $[f]$ rather than $[f]_\mu$, which we will do when more than one measure has to be involved). Since the set of all null sets is a σ-ideal, these operations are easily shown to be well defined:

$$\begin{aligned} [f] + [g] &:= [f + g], \\ [f] \cdot [g] &:= [f \cdot g], \\ \alpha \cdot [f] &:= [\alpha \cdot f]. \end{aligned}$$

Thus we obtain

Proposition 4.11.11 *$L_1(\mu)$ is a vector lattice.* $\dashv$

$\|\cdot\|_1$

Let $f \in L_1(\mu)$; then we define

$$\|f\|_1 := \int_X |f| \, d\mu$$

as the L_1-norm for f. Let us have a look at the properties which a decent norm should have. First, we have $\|f\|_1 \geq 0$, and $\|\alpha \cdot f\|_1 = |\alpha| \cdot \|f\|_1$; this is immediate. Because $|f + g| \leq |f| + |g|$, the triangle inequality holds. Finally, let $\|f\|_1 = 0$, and thus $\int_X |f| \, d\mu = 0$; consequently, $f =_\mu 0$, which means $f = [0]$.

Conjugate numbers

This will be a basis for the definition of a whole family of linear spaces of integrable functions. Call the positive real numbers p and q *conjugate* iff they satisfy

$$\frac{1}{p} + \frac{1}{q} = 1$$

(for example, 2 is conjugate to itself). This may be extended to $p = 0$, so that we also consider 0 and ∞ as conjugate numbers, but using this pair will be made explicit.

The first step for extending the definition of L_1 will be *Hölder's inequality*, which is based on this simple geometric fact.

Lemma 4.11.12 *Let a, b be positive real numbers and $p > 0$ conjugate to q; then*

$$a \cdot b \leq \frac{a^p}{p} + \frac{b^q}{q},$$

equality holding iff $b = a^{p-1}$.

Proof The exponential function is convex, i.e., we have

$$e^{(1-\alpha)\cdot x - \alpha \cdot y} \leq (1 - \alpha) \cdot e^x + \alpha \cdot e^y$$

for all $x, y \in \mathbb{R}$ and $0 \leq \alpha \leq 1$. Because both $a > 0$ and $b > 0$, we find r, s such that $a = e^{r/p}$ and $b = e^{s/q}$. Since p and q are conjugate, we obtain from $1/p = 1 - 1/q$

$$a \cdot b = e^{r/p + s/q} \leq \frac{e^s}{p} + \frac{e^q}{q} = \frac{a^p}{p} + \frac{b^q}{q}.$$

$\dashv$

This betrays one of the secrets of conjugate p and q, viz., that they give rise to a convex combination.

We are ready to formulate and prove *Hölder's inequality*, arguably one of the most frequently used inequalities in integration (as we will see as well); the proof follows the one given for [Rud74, Theorem 3.5].

Proposition 4.11.13 *Let $p > 0$ and $q > 0$ be conjugate and f and g be nonnegative measurable functions on X. Then* Hölder's inequality

$$\int_X f \cdot g \, d\mu \leq \Big(\int_X f^p \, d\mu\Big)^{1/p} \cdot \Big(\int_X g^q \, d\mu\Big)^{1/q}.$$

Proof Put for simplicity

$$A := \Big(\int_X f^p \, d\mu\Big)^{1/p} \text{ and } B := \Big(\int_X g^q \, d\mu\Big)^{1/q}.$$

If $A = 0$, we may conclude from $f =_\mu 0$ that $f \cdot g =_\mu 0$, so there is nothing to prove. If $A > 0$ and $B = \infty$, the inequality is trivial, so we assume that $0 < A < \infty, 0 < B < \infty$. Put

$$F := \frac{f}{A}, G := \frac{g}{B};$$

thus we obtain

$$\int_X F^p \, d\mu = \int_X G^q \, d\mu = 1.$$

We obtain $F(x) \cdot G(x) \leq F(x)^p/p + G(x)^q/q$ for every $x \in X$ from Lemma 4.11.12; hence

$$\int_X F \cdot G \, d\mu \leq \frac{1}{p} \cdot \int_X F^p \, d\mu + \frac{1}{q} \cdot \int_X G^q \, d\mu \leq \frac{1}{p} + \frac{1}{q} = 1.$$

Multiplying both sides with $A \cdot B > 0$ now yields the desired result. $\dashv$

This gives *Minkowski's inequality* as a consequence. Put for $f : X \to \mathbb{R}$ measurable and for $p \geq 1$ Minkowski's inequality

$$\|f\|_p := \Big(\int_X |f|^p \, d\mu\Big)^{1/p}.$$

Proposition 4.11.14 *Let $1 \leq p < \infty$ and let f and g be nonnegative measurable functions on X. Then*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Proof The inequality follows for $p = 1$ from the triangle inequality for $|\cdot|$, so we may assume that $p > 1$. We may also assume that $f, g \geq 0$. Then we obtain from Hölder's inequality with q conjugate to p

$$\begin{aligned}\|f+g\|_p^p &= \int_X (f+g)^{p-1} \cdot f \, d\mu + \int_X (f+g)^{p-1} \cdot g \, d\mu \\ &\leq \|f+g\|_p^{p/q} \cdot \big(\|f\|_p + \|g\|_p\big).\end{aligned}$$

Now assume that $\|f+g\|_p = \infty$, we may divide by the factor $\|f+g\|_p^{p/q}$, and we obtain the desired inequality from $p - p/q = p \cdot (1 - 1/q) = 1$. If, however, the left-hand side is infinite, then the inequality

$$(f+g)^p \leq 2^p \cdot max\{f^p, g^p\} \leq 2^p \cdot (f^p + g^p)$$

shows that the right-hand side is infinite as well. ⊣

Given $1 \leq p < \infty$, define

$$\mathcal{L}_p(\mu) := \{f \in \mathcal{F}(X, \mathcal{A}) \mid \|f\|_p < \infty\}$$

$\mathcal{L}_p(\mu)$, $L_p(\mu)$

with $L_p(\mu)$ as the corresponding set of $=_\mu$-equivalence classes. An immediate consequence from Minkowski's inequality is

Proposition 4.11.15 *$\mathcal{L}_p(\mu)$ is a linear space over $\mathbb{R}$, and $\|\cdot\|_p$ is a pseudo-norm on it. $L_p(\mu)$ is a normed space.*

Proof It is immediate from Proposition 4.11.14 that $f + g \in \mathcal{L}_p(\mu)$ whenever $f, g \in \mathcal{L}_p(\mu)$, and $\mathcal{L}_p(\mu)$ is closed under scalar multiplication as well. That $\|\cdot\|_p$ is a pseudo-norm is also immediate. Because scalar multiplication and addition are compatible with forming equivalence classes, the set $L_p(\mu)$ of classes is a real vector space as well. As usual, we will identify f with its class, unless otherwise stated. Now $f \in L_p(\mu)$ with $\|f\|_p = 0$, then $|f| =_\mu 0$, hence $f =_\mu 0$, and thus $f = 0$. So $\|\cdot\|_p$ is a norm on $L_p(\mu)$. ⊣

In Sect. 4.2.1 the vector spaces $\mathcal{L}_\infty(\mu)$ and $L_\infty(\mu)$ are introduced, so we have now a family $\big(\mathcal{L}_p(\mu)\big)_{1 \leq p \leq \infty}$ of vector spaces together with their associated spaces $\big(L_p(\mu)\big)_{1 \leq p \leq \infty}$ of μ-equivalence classes, which are normed spaces. They share the property of being Banach spaces.

Proposition 4.11.16 *$L_p(\mu)$ is a Banach space for $1 \leq p \leq \infty$.*

Proof 1. Let us first assume that the measure is finite. We know already from Proposition 4.2.9 that $\mathcal{L}_\infty(\mu)$ is a Banach space, so we may assume that $p < \infty$.

Given $(f_n)_{n\in\mathbb{N}}$ as a Cauchy sequence in $L_p(\mu)$, then we obtain

$$\epsilon^p \cdot \mu(\{x \in X \mid |f_n - f_m| \geq \epsilon\}) \leq \int_X |f_n - f_m|^p \, d\mu$$

for $\epsilon > 0$. Thus $(f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence for convergence in measure, so we can find $f \in \mathcal{F}(X, \mathcal{A})$ such that $f_n \xrightarrow{i.m.} f$ by Proposition 4.2.21. Proposition 4.2.16 tells us that we can find a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ such that $f_{n_k} \xrightarrow{a.e.} f$. But we do not yet know that $f \in \mathcal{L}_p(\mu)$. We infer $\lim_{k\to\infty} |f_{n_k} - f|^p = 0$ outside a set of measure zero. Thus we obtain from Fatou's Lemma (Proposition 4.8.5) for every $n \in \mathbb{N}$

$$\int_X |f - f_n|^p \, d\mu \leq \liminf_{k\to\infty} \int_X |f_{n_k} - f_n|^p \, d\mu.$$

Thus $f - f_n \in \mathcal{L}_p(\mu)$ for all $n \in \mathbb{N}$, and from $f = (f - f_n) + f_n$, we infer $f \in \mathcal{L}_p(\mu)$, since $\mathcal{L}_p(\mu)$ is closed under addition. We see also that $\|f - f_n\|_p \to 0$, as $n \to \infty$.

2. If the measure space is σ-finite, we may write $\int_X f \, d\mu$ as $\lim_{n\to\infty} \int_{A_n} f \, d\mu$, where $\mu(A_n) < \infty$ for an increasing sequence $(A_n)_{n\in\mathbb{N}}$ of measurable sets with $\bigcup_{n\in\mathbb{N}} A_n = X$. Since the restriction to each A_n yields a finite measure space, where the result holds, it is not difficult to see that completeness holds for the whole space as well. Specifically, given $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ so that for all $n, m \geq n_0$

$$\|f_n - f_m\|_p \leq \|f_n - f_m\|_p^{(n)} + \epsilon$$

holds, with $\|g\|_p^{(n)} := \left(\int_X |g|^p \, d\mu_n\right)^{1/p}$ and $\mu_n : B \mapsto \mu(B \cap A_n)$ as the measure μ localized to A_n. Then $\|f_n - f\|_p^{(n)} \to 0$, from which we obtain $\|f_n - f\|_p \to 0$. Hence completeness is also valid for the σ-finite case. ⊣

Example 4.11.17 Let $|\cdot|$ be the counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$; then this is a σ-finite measure space. Define

$$\begin{aligned} \ell_p &:= L_p(|\cdot|), 1 \leq p < \infty, \\ \ell_\infty &:= L_\infty(|\cdot|). \end{aligned}$$

Then ℓ_p is the set of all real sequences $(x_n)_{n\in\mathbb{N}}$ with $\sum_{n\in\mathbb{N}} |x_n|^p < \infty$ and $(x_n)_{n\in\mathbb{N}} \in \ell_\infty$ iff $\sup_{n\in\mathbb{N}} |x_n| < \infty$. Note that we do not need to pass to equivalence classes, since $|A| = 0$ iff $A = \emptyset$. These spaces are well known and well studied; here they make their only appearance. ✌

The case $p = 2$ deserves particular attention, since the norm in this case is obtained from the inner product

$$(f, g) := \int_X f \cdot g \, d\mu.$$

In fact, linearity of the integral shows that

$$(\alpha \cdot f + \beta \cdot g, h) = \alpha \cdot (f, h) + \beta \cdot (g, h)$$

holds, commutativity of multiplications yields $(f, g) = (g, f)$, and finally it is clear that $(f, f) \geq 0$ always holds. If we have $f \in \mathcal{L}_2(\mu)$ with $f =_\mu 0$, then we know that also $(f, f) = 0$; thus $(f, f) = 0$ iff $f = 0$ in $L_2(\mu)$.

Thus we obtain from Proposition 4.11.16

Corollary 4.11.18 *$L_2(\mu)$ is a Hilbert space with the inner product $(f, g) := \int_X f \cdot g \, d\mu$.* $\dashv$

This will have some interesting consequences, which we will explore in Sect. 4.11.3.

Before doing so, we show that the step functions belonging to L_p are dense.

Corollary 4.11.19 *Given $1 \leq p < \infty$, the set $\{f \in \mathcal{T}(X, \mathcal{A}) \mid \mu(\{x \in X \mid f(x) \neq 0\}) < \infty\}$ is dense in $L_p(\mu)$ with respect to $\|\cdot\|_p$.*

Proof The proof makes use of the fact that the step functions are dense with respect to pointwise convergence: We will just have to mark those functions which are in $L_p(\mu)$. Assume that $f \in \mathcal{L}_p(\mu)$ with $f \geq 0$; then there exists by Proposition 4.2.4 an increasing sequence $(g_n)_{n\in\mathbb{N}}$ of step functions with $f(x) = \lim_{n\to\infty} f_n(x)$. Because $0 \leq g_n \leq f$, we conclude that g_n belongs to the set under consideration, and we know from Lebesgue's Dominated Convergence Theorem 4.8.6 that $\|f - g_n\|_p \to 0$. Thus every nonnegative element of $\mathcal{L}_p(\mu)$ can be approximated through elements of this set in the $\|\cdot\|_p$-norm. In the general case, decompose $f = f^+ - f^-$ and apply the argument to both summands separately. $\dashv$

Because the rationals form a countable and dense subset of the reals, we take all step functions with rational coefficients, and obtain

Corollary 4.11.20 *$L_p(\mu)$ is a separable Banach space for $1 \leq p < \infty$.* $\dashv$

Note that we did exclude the case $p = \infty$; in fact, $L_\infty(\mu)$ is usually not a separable Banach space, as this example shows.

Example 4.11.21 Let λ be Lebesgue measure on the Borel sets of the unit interval $[0, 1]$. Put $f_t := \chi_{[0,t]}$ for $0 \leq t \leq 1$, then $f_t \in L_\infty(\lambda)$ for all t, and we have $||f_s - f_t||_\infty^\lambda = 1$ for $0 < s < t < 1$. Let

$$K_t := \{f \in L_\infty(\lambda) \mid ||f - f_t||_\infty^\lambda < 1/2\}.$$

Then $K_s \cap K_t = \emptyset$ for $s \neq t$. In fact, if $g \in K_s \cap K_t$, then $||f_s - f_t||_\infty^\lambda \leq ||g - f_t||_\infty^\lambda + ||f_s - g||_\infty^\lambda < 1$. On the other hand, each K_t is open, so if we have a countable subset $D \subseteq L_\infty(\lambda)$, then $K_t \cap D = \emptyset$ for uncountably many t. Thus D cannot be dense. But this means that $L_\infty(\lambda)$ is not separable. ✌

This is the first installment on the properties of L_p-spaces. We will be back with a general discussion in Sect. 4.11.4 after having explored the Lebesgue–Radon–Nikodym Theorem as a valuable tool in general and for our discussion in particular.

4.11.3 The Lebesgue–Radon–Nikodym Theorem

The Hilbert space structure of the L_2-spaces will now be used for decomposing a measure into an absolutely continuous and a singular part with respect to another measure and for constructing a density. This construction requires a more general study of the relationship between two measures.

We even go a bit beyond that and define absolute continuity and singularity as a relationship of two arbitrary additive set functions. This will be specialized fairly quickly to a relationship between finite measures, but this added generality will turn out to be beneficial nevertheless, as we will see.

Definition 4.11.22 *Let $(X, \mathcal{A})$ be a measurable space with two additive set functions $\rho, \zeta : \mathcal{A} \to \mathbb{R}$.*

1. *ρ is said to be* absolutely continuous *with respect to ζ ($\rho \ll \zeta$) iff $\rho(E) = 0$ for every $E \in \mathcal{A}$ for which $\zeta(A) = 0$.* $\rho \ll \zeta$
2. *ρ is said to be* concentrated *on $A \in \mathcal{A}$ iff $\rho(E) = \rho(E \cap A)$ for all $E \in \mathcal{A}$.*

$\rho \perp \zeta$

3. *ρ and ζ are called* mutually singular *($\rho \perp \zeta$) iff there exists a pair of disjoint sets A and B such that ρ is concentrated on A and ζ is concentrated on B.*

If two additive set functions are mutually singular, they live on disjoint measurable sets in the same measurable space. These are elementary properties.

Lemma 4.11.23 *Let $\rho_1, \rho_2, \zeta : \mathcal{A} \to \mathbb{R}$ be additive set functions; then we have for $a_1, a_2 \in \mathbb{R}$*

1. *If $\rho_1 \perp \zeta$ and $\rho_2 \perp \zeta$, then $a_1 \cdot \rho_1 + a_2 \cdot \rho_2 \perp \zeta$.*
2. *If $\rho_1 \ll \zeta$ and $\rho_2 \ll \zeta$, then $a_1 \cdot \rho_1 + a_2 \cdot \rho_2 \ll \zeta$.*
3. *If $\rho_1 \ll \zeta$ and $\rho_2 \perp \zeta$, then $\rho_1 \perp \rho_2$.*
4. *If $\rho \ll \zeta$ and $\rho \perp \zeta$, then $\rho = 0$.*

Proof 1. For proving 1, note that we can find a measurable set B and sets $A_1, A_2 \in \mathcal{A}$ with $B \cap (A_1 \cup A_2) = \emptyset$ with $\zeta(E) = \zeta(E \cap B)$ and $\rho_i(E) = \rho_i(E \cap A_i)$ for $i = 1, 2$. By additivity, we obtain $(a_1 \cdot \rho_1 + a_2 \cdot \rho_2)(E) = (a_1 \cdot \rho_1 + a_2 \cdot \rho_2)(E \cap (A_1 \cup A_2))$. Property 2 is obvious.

2. ρ_2 is concentrated on A_2, ζ is concentrated on B with $A \cap B = \emptyset$, hence $\zeta(E \cap A_2) = 0$, and thus $\rho_1(E \cap A_2) = 0$ for all $E \in \mathcal{A}$. Additivity implies $\rho_1(E) = \rho_1\big(E \cap (X \setminus A_2)\big)$, so ρ_1 is concentrated on $X \setminus A_2$. This proves 3. For proving 4, note that $\rho \ll \zeta$ and $\rho \perp \zeta$ imply $\rho \perp \rho$ by property 3, which implies $\rho = 0$. $\dashv$

We specialize these relations now to finite measures on $\mathcal{A}$. Absolute continuity can be expressed in a different way, which makes the concept more transparent. Specifically, absolute continuity could have been defined akin to the well-known ϵ-δ definition of continuity for real functions. Then the name becomes a bit more descriptive.

Lemma 4.11.24 *Given finite measures μ and ν on a measurable space $(X, \mathcal{A})$, these conditions are equivalent:*

1. *$\mu \ll \nu$.*
2. *For every $\epsilon > 0$, there exists $\delta > 0$ such that $\nu(A) < \delta$ implies $\mu(A) < \epsilon$ for all measurable sets $A \in \mathcal{A}$.*

Proof 1 $\Rightarrow$ 2: Assume that we can find $\epsilon > 0$ so that there exist sets $A_n \in \mathcal{A}$ with $\nu(A_n) < 2^{-n}$ but $\mu(A_n) \geq \epsilon$. Then we have $\mu(\bigcup_{k \geq n} A_k) \geq \epsilon$ for all $n \in \mathbb{N}$; consequently, by monotone convergence, also $\mu\big(\bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} A_k\big) \geq \epsilon$. On the other hand, $\nu(\bigcup_{k \geq n} A_k) \leq \sum_{k \geq n} 2^{-k} = 2^{-n+1}$ for all $n \in \mathbb{N}$, so by monotone convergence again, $\nu\big(\bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} A_k\big) = 0$. Thus $\mu \ll \nu$ does not hold.

2 $\Rightarrow$ 1: Let $\nu(A) = 0$, then $\mu(A) \leq \epsilon$ for every $\epsilon > 0$; hence $\mu \ll \nu$ is true. $\dashv$

Given two finite measures μ and ν, one, say μ, can be decomposed uniquely as a sum $\mu_a + \mu_s$ such that $\mu_a \ll \nu$ and $\mu_s \perp \nu$; additionally $\mu_s \perp \mu_a$ holds. This is stated and proved in the following theorem, which actually shows much more, viz., that there exists a *density* h of μ_a with respect to ν. This means that $\mu_a(A) = \int_A h \, d\nu$ holds for all $A \in \mathcal{A}$. Density

Densities are familiar from probability distributions, for example, the normal distribution $N(0, 1)$ has the density $e^{-x^2/2}/\sqrt{2 \cdot \pi}$. This means that a random variable which is distributed according to $N(0, 1)$ takes values in the Borel set $A \in \mathcal{B}(\mathbb{R})$ with probability $\int_A e^{-x^2/2}/\sqrt{2 \cdot \pi} \, dx$.

This is but a special case. What a density is used for in our context will be described now also in greater detail. Before entering into formalities, it is noted that the decomposition is usually called the *Lebesgue decomposition* of μ with respect to ν and that the density h is usually called the *Radon–Nikodym derivative* of μ_a with respect to ν and denoted by $d\mu/d\nu$.

The proof both for the existence of Lebesgue decomposition and of the Radon–Nikodym derivative is done in one step. The beautiful proof given below was proposed by von Neumann; see [Rud74, 6.9]. Here we go.

Theorem 4.11.25 *Let μ and ν be finite measures on $(X, \mathcal{A})$.*

1. *There exists a unique pair μ_a and μ_s of finite measures on $(X, \mathcal{A})$ such that $\mu = \mu_a + \mu_s$ with $\mu_a \ll \nu$, $\mu_a \perp \nu$. In addition, $\mu_a \perp \mu_s$ holds.*

2. *There exists a unique $h \in L_1(\nu)$ such that*
$$\mu_a(A) = \int_A h \, d\nu$$
for all $A \in \mathcal{A}$.

Overview of the proof

The line of attack will be as follows: We show that $f \mapsto \int_X f \, d\mu$ is a continuous linear functional on the Hilbert space $L_2(\mu + \nu)$. We can express this functional by the representation for these functionals on Hilbert spaces through some function $g \in \mathcal{L}_2(\mu + \nu)$; hence

$$\int_X f \, d\mu = \int_X f \cdot g \, d(\mu + \nu);$$

note the way the measures μ and $\mu + \nu$ interact by exploiting the integral with respect to μ as a linear functional on $L_2(\mu)$. A closer investigation of g will then yield the sets we need for the decomposition and permit constructing the density h.

Proof 1. Define the finite measure $\varphi := \mu + \nu$ on $\mathcal{A}$; note that

$$\int_X f \, d\varphi = \int_X f \, d\mu + \int_X f \, d\nu$$

holds for all measurable f for which the sum on the right-hand side is defined; this follows from Levi's Theorem 4.8.2 (for $f \geq 0$) and from additivity (for general f). We show first that $L : f \mapsto \int_X f \, d\mu$ is a continuous linear operator on $L_2(\varphi)$. In fact,

$$\left|\int_X f \, d\mu\right| \leq \int_X |f| \, d\varphi = \int_X |f| \cdot 1 \, d\varphi \leq \left(\int_X d|f|^2\right)^{1/2} \cdot \sqrt{\varphi(X)}$$

by Schwarz's inequality (Lemma 4.11.1). Thus

$$\sup_{\|f\|_2 \leq 1} |L(f)| \leq \sqrt{\varphi(X)} < \infty.$$

Hence L is continuous (Exercise 4.29); thus by Theorem 4.11.9 there exists $g \in \mathcal{L}_2(\mu)$ such that

$$L(f) = \int_X f \cdot g \, d\varphi \tag{4.21}$$

for all $f \in L_2(\mu)$.

2. Let $f = \chi_A$ for $A \in \mathcal{A}$; then we obtain

$$\int_A g \, d\varphi = \mu(A) \leq \varphi(A)$$

from (4.21). This yields $0 \leq g \leq 1$ φ-a.e.; we can change g on a set of φ-measure 0 to the effect that $0 \leq g(x) \leq 1$ holds for all $x \in X$. This will not affect the representation in (4.21).

We know that

$$\int_X (1-g)\cdot f\,d\mu = \int_X f\cdot g\,d\nu \tag{4.22}$$

holds for all $f \in L_2(\varphi)$. Put

$$A := \{x \in X \mid 0 \le g(x) < 1\},$$
$$B := \{x \in X \mid g(x) = 1\},$$

then $A, B \in \mathcal{A}$, and we define for $E \in \mathcal{A}$

$$\mu_a(E) := \mu(E \cap A),$$
$$\mu_s(E) := \mu(E \cap B).$$

If $f = \chi_B$, then we obtain $\nu(B) = \int_B g\,d\nu = \int_B 0\,d\mu = 0$ from (4.22), and thus $\nu(B) = 0$, so that $\mu_s \perp \nu$.

3. Replace for a fixed $E \in \mathcal{A}$ in (4.22) the function f by $(1+g+\ldots+g^n)\cdot\chi_E$; then we have

$$\int_E (1-g^{n+1})\,d\mu = \int_E g\cdot(1+g+\ldots+g^n)\,d\nu.$$

Look at the integrand on the right-hand side: It equals zero on B and increases monotonically to 1 on A; hence $\lim_{n\to\infty}\int_E (1-g^{n+1})\,d\mu = \mu(E\cap A) = \mu_a(E)$. This provides a bound for the left-hand side for all $n \in \mathbb{N}$. The integrand on the left-hand side converges monotonically to some function $0 \le h \in \mathcal{L}_1(\nu)$ with $\lim_{n\to\infty}\int_E g\cdot(1+g+\ldots+g^n)\,d\nu = \int_E h\,d\nu$ by Levi's Theorem 4.8.2. Hence we have

$$\int_E h\,d\nu = \mu_a(E)$$

for all $E \in \mathcal{A}$, in particular $\mu_a \ll \nu$.

4. Assume that we can find another pair μ_a' and μ_s' with $\mu_a' \ll \nu$ and $\mu_s' \perp \nu$ and $\mu = \mu_a' + \mu_s'$. Then we have $\mu_a - \mu_a' = \mu_s' - \mu_s$ with $\mu_a - \mu_a' \ll \nu$ and $\mu_s' - \mu_s \perp \nu$ by Lemma 4.11.23; hence $\mu_s - \mu_s' = 0$, again by Lemma 4.11.23, which implies $\mu_a - \mu_a' = 0$. So the decomposition is unique. From this, the uniqueness of the density h is inferred. ⊣

We obtain as a consequence the well-known Radon–Nikodym Theorem.

Radon-Nikodym Theorem

Theorem 4.11.26 *Let μ and ν be finite measures on $(X, \mathcal{A})$ with $\mu \ll \nu$. Then there exists a unique $h \in L_1(\mu)$ with $\mu(A) = \int_A h\, d\nu$ for all $A \in \mathcal{A}$. Moreover, $f \in L_1(\mu)$ iff $f \cdot h \in L_1(\nu)$; in this case*

$$\int_X f\, d\mu = \int_X f \cdot h\, d\nu.$$

$d\mu/d\nu$

h is called the Radon–Nikodym derivative *of μ with respect to ν and sometimes denoted by $d\mu/d\nu$.*

Proof Write $m = \mu_a + \mu_s$, where μ_a and μ_s are the Lebesgue decomposition of μ with respect to ν by Theorem 4.11.25. Since $\mu_s \perp \nu$, we find $\mu_s = 0$, so that $\mu_a = \mu$. Then apply the second part of Theorem 4.11.25 to μ. This accounts for the first part. The second part follows from this by an approximation through step functions according to Corollary 4.11.19. $\dashv$

Note that the Radon–Nikodym Theorem gives a one-to-one correspondence of finite measures μ such that $\mu \ll \nu$ and the Banach space $L_1(\nu)$.

Theorem 4.11.25 can be extended to complex measures; we will comment on this after the Jordan decomposition has been established in Proposition 4.11.32.

Both constructions have, as one might expect, a plethora of applications. We will not discuss the Lebesgue decomposition further, but rather focus on the Radon–Nikodym Theorem and discuss two applications, viz., identifying the dual space of the L_p-spaces for $p < \infty$ and disintegrating a measure on a product space.

But this is a place to have a look at integration by substitution, a technique well known from Calculus. The multidimensional case has been mentioned at the end of Sect. 4.8.1 on page 549; we deal here with the one-dimensional case. The approach displays a pretty interplay of integrating with respect to an image measure and the Radon–Nikodym Theorem, which should not be missed.

We prepare the stage with an auxiliary statement, which is of interest of its own. Recall that ρ_* denotes the inner measure (p. 89) with respect to measure ρ.

Lemma 4.11.27 *Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \rho)$ be finite measure spaces and $\psi : X \to Y$ be measurable and onto such that $\rho_*(\psi\left[A\right]) = 0$,* *whenever $\mu(A) = 0$. Put $\nu := \mathbb{M}(\psi)(\mu)$. Then there exists a measurable function $w : X \to \mathbb{R}_+$ such that*

1. *$f \in L_1(\rho)$ iff $(f \circ g) \cdot w \in L_1(\mu)$.*
2. *$\int_Y f(y)\, d\rho(y) = \int_X (f \circ \psi)(x) \cdot g(x)\, d\mu(x)$ for all $f \in L_1(\rho)$.*

Proof We show first that $\rho \ll \nu$, from which we obtain a derivative. This is used then through the change of variables formula from Corollary 4.8.9 for obtaining the desired result. Plan

In fact, assume that $\nu(B) = 0$ for some $B \in \mathcal{B}$ and, equivalently, $\mu(\psi^{-1}\big[B\big]) = 0$. By assumption $0 = \rho_*(\psi\big[\psi^{-1}\big[B\big]\big]) = \rho(B)$, since $B = \psi\big[\psi^{-1}\big[B\big]\big]$ due to ψ being onto. Thus we find $g_1 : Y \to \mathbb{R}_+$ such that $f \in L_1(\rho)$ iff $f \cdot g_1 \in L_1(\nu)$ and $\int_Y f\, d\rho = \int_Y f \cdot g_1\, d\nu$. Since $\nu = \mathbb{M}(\psi)(\mu)$, we obtain from Corollary 4.8.9 that

$$\int_Y f\, d\rho = \int_X (f \circ \psi) \cdot (g_1 \circ \psi)\, d\mu$$

holds. Putting $g := g_1 \circ \psi$, the assertion follows. $\dashv$

The rôle of ν as the image measure is interesting here. It just serves as a kind of facilitator, but it remains in the background. Only the measures ρ and μ are acting, and the image measure is used only for obtaining the Radon–Nikodym derivative and for converting its integral to an integral with respect to its preimage through change of variables.

We specialize things now to intervals on the real line and make restrictive assumptions on ψ. Then—voilà—the well-known formula on integration by substitution will result.

But first a more general consequence of Lemma 4.11.27 is to be presented. We will be working with Lebesgue measure on intervals of the reals. Here we assume that $\psi : [\alpha, \beta] \to [a, b]$ is continuous with the additional property that $\lambda(A) = 0$ implies $\lambda_*(\psi\big[A\big]) = 0$ for all $A \subseteq \mathcal{B}([\alpha, \beta])$. This class of functions is generally known as *absolutely continuous* and discussed in great detail in [HS65, Sect. 18, Theorem (18.25)]. We obtain from Lemma 4.11.27

Corollary 4.11.28 *Let $[\alpha, \beta] \subseteq \mathbb{R}$ be a closed interval and $\psi : [\alpha, \beta] \to [a, b]$ be a surjective and absolutely continuous function. Then there exists a Borel measurable function $w : [\alpha, \beta] \to \mathbb{R}$ such that*

1. *$f \in L_1([a, b], \lambda)$ iff $(f \circ \psi) \cdot w \in L_1([\alpha, \beta], \lambda)$*

2. $\int_a^b f(x)\,dx = \int_\alpha^\beta (f(\psi(t)) \cdot w(t)\,dt$.

Proof The assertion follows from Lemma 4.11.27 by specializing μ and ρ to λ. $\dashv$

If we restrict ψ further, we obtain even more specific information about the function w. The following proof shows how we exploit the properties of ψ, viz., being monotone and having a continuous first derivative, through the definition of the integral as a limit of approximations on a system on subintervals which get smaller and smaller. The subdivisions in the domain are then related to the one in the range of ψ; the relationship is done through Lagrange's Theorem which brings in the derivative. But see for yourself.

Proposition 4.11.29 *Assume that $\psi : [\alpha, \beta] \to [a, b]$ is continuous and monotone with a continuous first derivative such that $\psi(\alpha) = a$ and $\psi(\beta) = b$. Then f is Lebesgue integrable over $[a, b]$ iff $(f \circ \psi) \cdot \psi'$ is Lebesgue integrable over $[\alpha, \beta]$, and*

$$\int_a^b f(x)\,dx = \int_\alpha^\beta f(\psi(z)) \cdot \psi'(z)\,dz$$

holds.

Basic idea

We follow [Fic64, Nr. 316] in his proof. The basic idea is to approximate the integral through step functions, which are obtained by subdividing the interval $[\alpha, \beta]$ into subintervals, and to refine the subdivisions, using uniform continuity both of ψ and ψ' on its compact domain. So this is a fairly classical proof.

Proof 0. We may assume that $f \geq 0$; otherwise we decompose $f = f^+ - f^-$ with $f^+, f^- \geq 0$. Also we assume that f is bounded by some constant L; otherwise we establish the property for $f \wedge n$ with $n \in \mathbb{N}$, letting $n \to \infty$ appeal to Levi's Theorem 4.8.2. Moreover we assume that ψ is increasing.

1. The interval $[\alpha, \beta]$ is subdivided through $\alpha = z_0 < z_1 < \ldots < z_n = \beta$; put $x_i := \psi(z_i)$; then $a = x_0 \leq x_1 \leq \ldots \leq x_n = b$, and $\Delta z_i := z_{i+1} - z_i$, and $\Delta x_i := x_{i+1} - x_i$. Let $\ell := \max_{i=1,\ldots,n-1} \Delta z_i$; then if $\ell \to 0$, the maximal difference $\max_{i=1,\ldots,n-1} \Delta x_i$ tends to 0 as well, because ψ is uniformly continuous. This is so since the interval $[\alpha, \beta]$ is compact.

For approximating the integral $\int_\alpha^\beta f(\psi(z))\cdot\psi'(z)\,dz$, we select ζ_i from each interval $[z_i, z_{i+1}]$ and write

$$S := \sum_i f(\psi(\zeta_i)) \cdot \psi'(\zeta_i) \cdot \Delta z_i.$$

Put $\xi_i := \psi(\zeta_i)$; hence $x_i \leq \xi_i \leq x_{i+1}$. By Lagrange's Formula[1] there exists $\tau_i \in [z_i, z_{i+1}]$ such that $\Delta x_i = \psi'(\tau_i)\cdot\Delta z_i$, so that we can write as an approximation to the integral $\int_a^b f(x)\,dx$ the sum

$$\begin{aligned} s &:= \sum_i f(\xi_i) \cdot \Delta z_i \\ &= \sum_i f(\xi_i) \cdot \psi(\tau_i) \cdot \Delta z_i \\ &= \sum_i f(\psi(\zeta_i)) \cdot \psi'(\tau_i) \cdot \Delta z_i. \end{aligned}$$

If $\ell \to 0$, we know that $s \to \int_a^b f(x)\,dx$ and $S \to \int_\alpha^\beta f(\psi(z)) \cdot \psi'(z)\,dz$, so that we have to get a handle at the difference $|S - s|$. We claim that this difference tends to zero, as $\ell \to 0$. Given $\epsilon > 0$, we find $\delta > 0$ such that $|\psi'(\zeta_i) - \psi'(\tau_i)| < \epsilon$, provided $\ell < \delta$. This is so because ψ' is continuous, hence uniformly continuous. But then we obtain by telescoping

$$|S - s| \leq \sum_i |f(\psi(\zeta_i))| \cdot |\psi'(\zeta_i) - \psi'(\tau_i)| \cdot \Delta z_i < L \cdot (\beta - \alpha) \cdot \epsilon.$$

Thus the difference vanishes, and we obtain indeed the equality claimed above. $\dashv$

4.11.4 Continuous Linear Functionals on L_p

After all these preparations and an excursion into classical Calculus, we will investigate now continuous linear functionals on the L_p-spaces and show that the map $f \mapsto \int_X f\,d\mu$ plays an important rôle in identifying them. For full generality with respect to the functional concern, we introduce signed measures here and show that they may be obtained

[1] Recall that *Lagrange's Formula* says the following: Assume that g is continuous on the interval $[c, d]$ with a continuous derivative g' on the open interval $]c, d[$. Then there exists $t \in\ ]c, d[$ such that $g(d) - g(c) = g'(t) \cdot (d - c)$.

in a fairly specific way from the (unsigned) measures considered so far.

But before entering into this discussion, we make some general remarks. If V is a real vector space with a norm $\|\cdot\|$, then a map $\Lambda : V \to \mathbb{R}$ is a *linear functional* on V iff it is compatible with the vector space structure, i.e., iff $\Lambda(\alpha \cdot x + \beta \cdot y) = \alpha \cdot \Lambda(x) + \beta \cdot \Lambda(y)$ holds for all $x, y \in V$ and all $\alpha, \beta \in \mathbb{R}$. If $\Lambda \neq 0$, the range of Λ is unbounded, so $\sup_{x \in V} |\Lambda(x)| = \infty$. Consequently it is difficult to assign to Λ something like the sup-norm for characterizing continuity. It turns out, however, that we may investigate continuity through the behavior of Λ on the unit ball of V, so we define

$$\|\Lambda\| := \sup_{\|x\| \leq 1} |\Lambda(x)|.$$

Call Λ *bounded* iff $\|\Lambda\| < \infty$. Then Λ is continuous iff Λ is bounded; see Exercise 4.29.

Now let μ be a finite measure with p and q conjugate to each other; see page 648. Define for $g \in L_q(\mu)$ the linear functional

$$\Lambda_g(f) := \int_X f \cdot g \, d\mu$$

on $L_p(\mu)$; then we know from Hölder's inequality in Proposition 4.11.13 that

$$\|\Lambda_g\| \leq \sup_{\|f\|_p \leq 1} \int_X |f \cdot g| \, d\mu \leq \|g\|_q.$$

That was easy. But what about the converse? Given a bounded linear functional Λ on $L_p(\mu)$, does there exist $g \in L_q(\mu)$ with $\Lambda = \Lambda_g$? It is immediate that this will not work in general, since $\Lambda_g(f) \geq 0$, provided $f \geq 0$. So we have to assume that Λ maps positive functions to a nonnegative value. Call Λ *positive* iff this is the case.

Summarizing, we consider maps $\Lambda : \mathcal{L}_p(\mu) \to \mathbb{R}$ with these properties:

Linearity: $\Lambda(\alpha \cdot x + \beta \cdot y) = \alpha \cdot \Lambda(x) + \beta \cdot \Lambda(y)$ holds for all $x, y \in V$ and all $\alpha, \beta \in \mathbb{R}$.

Boundedness: $\|\Lambda\| := \sup_{\|f\|_p = 1} |\Lambda(f)| \leq \infty$ (hence $|\Lambda(f)| \leq \|\Lambda\| \cdot \|f\|_p$ for all f).

Positiveness: $f \geq 0 \Rightarrow \Lambda(f) \geq 0$ (note that $f \geq 0$ means $f' \geq 0$ almost everywhere with respect to μ for each representative f' of f by our convention).

We will first work on this restricted problem, and then we will expand the answer. This will require a slight generalization: We will talk about signed measures rather than about measures.

Let us jump right in.

Theorem 4.11.30 *Assume that μ is a finite measure on $(X, \mathcal{A})$, $1 \leq p < \infty$, and that Λ is a bounded positive linear functional on $L_p(\mu)$. Then there exists a unique $g \in L_q(\mu)$ such that*

$$\Lambda(f) = \int_X f \cdot g \, d\mu$$

holds for each $f \in L_p(\mu)$. In addition, $\|\Lambda\| = \|g\|_q$.

This is our line of attack: We will first see that we obtain from Λ a finite measure ν on $\mathcal{A}$ such that $\nu \ll \mu$. The Radon–Nikodym Theorem will then give us a density $g := d\nu/d\mu$ which will turn out to be the function we are looking for. This is shown by separating the cases $p = 1$ and $p > 1$.

Line of attack

Proof 1. Define for $A \in \mathcal{A}$

$$\nu(A) := \Lambda(\chi_A).$$

Then $A \subseteq B$ implies $\chi_A \leq \chi_B$; hence $\Lambda(\chi_A) \leq \Lambda(\chi_B)$. Because Λ is monotone, ν is monotone as well. Since Λ is linear, we have $\nu(\emptyset) = 0$, and ν is additive. Let $(A_n)_{n\in\mathbb{N}}$ be an increasing sequence of measurable sets with $A := \bigcup_{n\in\mathbb{N}} A_n$, then $\chi_{A\setminus A_n} \to 0$, and thus

$$\nu(A) - \nu(A_n) = \|\chi_{A\setminus A_n}\|_p^p = \Lambda(\chi_{A\setminus A_n})^p \to 0,$$

since Λ is continuous. Thus Λ is a finite measure on $\mathcal{A}$ (note $\nu(X) = \Lambda(1) < \infty$). If $\mu(A) = 0$, we see that $\chi_A =_\mu 0$; thus $\Lambda(\chi_A) = 0$, because we are dealing with the $=_\mu$-class of χ_A, so that $\nu(A) = 0$. Thus $\nu \ll \mu$, and the Radon–Nikodym Theorem 4.11.26 provides us with $g \in L_1(\mu)$ with

$$\Lambda(\chi_A) = \nu(A) = \int_A g \, d\mu$$

for all $A \in \mathcal{A}$. Since the integral and Λ are linear, we obtain from this

$$\Lambda(f) = \int_X f \cdot g \, d\mu$$

for all step functions f.

2. We have to show that $g \in L_q(\mu)$. Consider these cases.

Case $p = 1$: We have for each $A \in \mathcal{A}$

$$\left|\int_A g \, d\mu\right| \leq |\Lambda(\chi_A)| \leq \|\Lambda\| \cdot \|\chi_A\|_1 = \|\Lambda\| \cdot \mu(A).$$

But this implies $|g(x)| \leq_\mu \|\Lambda\|$; thus $\|g\|_\infty \leq \|\Lambda\|$.

Case $1 < p < \infty$: Let

$$t := \chi_{\{x \in X \mid g(x) \geq 0\}} - \chi_{\{x \in X \mid g(x) < 0\}},$$

then $|g| = t \cdot g$, and t is measurable, since g is. Define $A_n := \{x \in X \mid |g(x)| \leq n\}$, and put $f := \chi_{A_n} \cdot |g|^{q-1} \cdot t$. Then

$$\begin{aligned} |f|^p \cdot \chi_{A_n} &= |g|^{(q-1)\cdot p} \cdot \chi_{A_n} \\ &= |g|^q \cdot \chi_{A_n}, \\ \chi_{A_n} \cdot (f \cdot g) &= \chi_{A_n} \cdot |g|^{q-1} \cdot t \cdot g \\ &= \chi_{A_n} \cdot |g|^q \cdot t; \end{aligned}$$

thus

$$\int_{A_n} |g|^q \, d\mu = \int_{A_n} f \cdot g \, d\mu = \Lambda(f) \leq \|\Lambda\| \cdot \Bigl(\int_{A_n} |g|^q \, d\mu\Bigr)^{1/p}.$$

Since $1 - 1/p = 1/q$, dividing by the factor $\|\Lambda\|$ and raising the result by q yield

$$\int_{E_n} |g|^q \, d\mu \leq \|\Lambda\|^q.$$

By Lebesgue's Dominated Convergence Theorem 4.8.6, we obtain $\|g\|_q \leq \|\Lambda\|$; hence $g \in L_q(\mu)$, and $\|g\|_q = \|\Lambda\|$.

The proof is completed now by the observation that $\Lambda(f) = \int_X f \cdot g \, d\mu$ holds for all step functions f. Since both sides of this equation represent continuous functions and since the step functions are dense in $L_p(\mu)$ by Corollary 4.11.19, the equality holds on all of $L_p(\mu)$. $\dashv$

This representation says something only for positive linear functions; what about the rest? It turns out that we need to extend our notion of measures to signed measures and that a very similar statement holds for signed measures. Of course we will have to explain what the integral of a signed measure is, but this will work out very smoothly. So what we will do next is to define signed measures and to relate them to the measures with which we have worked until now. We follow essentially Halmos' exposition [Hal50, §29].

Definition 4.11.31 *A map $\mu : \mathcal{A} \to \mathbb{R}$ is said to be a* signed measure *iff μ is σ-additive, i.e., iff $\mu(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$, whenever $(A_n)_{n\in\mathbb{N}}$ is a sequence of mutually disjoint sets in $\mathcal{A}$.*

Clearly, $\mu(\emptyset) = 0$, since a signed measure μ is finite, so the distinguishing feature is the absence of monotonicity. It turns out, however, that we can partition the whole space X into a positive and a negative part, that restricting μ to these parts will yield a measure each, and that μ can be written in this way as the difference of two measures.

Fix a signed measure μ. Call $N \in \mathcal{A}$ a *negative set* iff $\mu(A \cap N) \leq 0$ for all $A \in \mathcal{A}$; a positive set is defined accordingly. It is immediate that the difference of two negative sets is a negative set again and that the union of a disjoint sequence of negative sets is a negative set as well. Thus the union of a sequence of negative sets is negative again.

Proposition 4.11.32 *Let μ be a signed measure on $\mathcal{A}$. Then there exists a pair X^+ and X^- of disjoint measurable sets such that X^+ is a positive set and X^- is a negative set. Then $\mu^+(B) := \mu(B \cap X^+)$ and $\mu^-(B) := -\mu(B \cap X^-)$ are finite measures on $\mathcal{A}$ such that $\mu = \mu^+ - \mu^-$. The pair μ^+ and μ^- is called the* Jordan Decomposition *of the signed measure μ.* Jordan decomposition μ^+, μ^-

Proof 1. Define

$$\alpha := \inf\{\mu(A) \mid A \in \mathcal{A} \text{ is negative}\} > -\infty.$$

Assume that $(A_n)_{n\in\mathbb{N}}$ is a sequence of measurable sets with $\mu(A_n) \to \alpha$; then we know that $A := \bigcup_{n\in\mathbb{N}} A_n$ is negative again with $\alpha = \mu(A)$. In fact, put $B_1 := A_1$, $B_{n+1} := A_{n+1} \setminus B_n$, then each B_n is negative, and we have

$$\mu(A) = \mu(\bigcup_{n\in\mathbb{N}} B_n) = \sum_{n\in\mathbb{N}} \mu(B_n) = \lim_{n\to\infty} \mu(A_n)$$

by telescoping.

2. We claim that

$$X^+ := X \setminus A$$

is a positive set. In fact, assume that this is not true—now this is truly the tricky part—then there exists $E_0 \subseteq X^+$ with $\mu(E_0) < 0$. E_0 cannot be a negative set, because otherwise $A \cup E_0$ would be a negative set with $\mu(A \cup E_0) = \mu(A) + \mu(E_0) < \alpha$, which contradicts the construction of A. Let k_1 be the smallest positive integer such that E_0 contains a measurable set E_1 with $\mu(E_1) \geq 1/k_1$. Now look at $E_0 \setminus E_1$. We have

$$\mu(E_0 \setminus E_1) = \mu(E_0) - \mu(E_1) \leq \mu(E_0) - \mu(E_1) \leq \mu(E_0) - 1/k_1 < 0.$$

We may repeat the same consideration now for $E_0 \setminus E_1$; let k_2 be the smallest positive integer such that $E_0 \setminus E_1$ contains a measurable set E_2 with $\mu(E_2) \geq 1/k_2$. This produces a sequence of disjoint measurable sets $(E_n)_{n\in\mathbb{N}}$ with

$$E_{n+1} \subseteq E_0 \setminus (E_1 \cup \ldots \cup E_n),$$

and since $\sum_{n\in\mathbb{N}} \mu(E_n)$ is finite (because $\bigcup_{n\in\mathbb{N}} E_n \in \mathcal{A}$ and μ takes only finite values), we infer that $\lim_{n\to\infty} 1/k_n = 0$.

3. Let $F \subseteq F_0 := E_0 \setminus \bigcup_{n\in\mathbb{N}} E_n$, and assume that $\mu(F) \geq 0$. Let ℓ be the largest positive integer with $\mu(F) \geq 1/\ell$. Since $k_n \to 0$, as $n \to \infty$, we find $m \in \mathbb{N}$ with $1/\ell \geq 1/k_m$. Since $F \subseteq E_0 \setminus (E_1 \cup \ldots \cup E_m)$, this yields a contradiction. But F_0 is disjoint from A, and since

$$\mu(F_0) = \mu(E_0) - \sum_{n\in\mathbb{N}} \mu(E_n) \leq \mu(E_0) < 0,$$

we have arrived at a contradiction. Thus $\mu(E_0) \geq 0$.

4. Now define μ^+ and μ^- as the traces of μ on X^+ and $X^- := A$, resp., then the assertion follows. ⊣

It should be noted that the decomposition of X into X^+ and X^- is not unique, but the decomposition of μ into μ^+ and μ^- is. Assume that X_1^+ with X_1^- and X_2^+ with X_2^- are two such decompositions. Let $A \in \mathcal{A}$, then we have $A \cap (X_1^+ \setminus X_2^+) \subseteq A \cap X_1^+$, and hence $\mu(A \cap (X_1^+ \setminus X_2^+) \geq 0$; on the other hand, $A \cap (X_1^+ \setminus X_2^+) \subseteq A \cap X_2^-$, and thus $\mu(A \cap (X_1^+ \setminus X_2^+) \leq 0$, so that we have $\mu(A \cap (X_1^+ \setminus X_2^+) = 0$, which implies $\mu(A \cap X_1^+) = \mu(A \cap X_2^+)$. Thus uniqueness of μ^+ and μ^- follows.

Given a signed measure μ with a Jordan decomposition μ^+ and μ^-, we define a (positive) measure $|\mu| := \mu^+ + \mu^-$; $|\mu|$ is called the *total variation* of μ. It is clear that $|\mu|$ is a finite measure on $\mathcal{A}$. A set $A \in \mathcal{A}$ is called a *μ-nullset* iff $\mu(B) = 0$ for every $B \in \mathcal{A}$ with $B \subseteq A$; hence A is a μ-null set iff A is a $|\mu|$-null set iff $|\mu|(A) = 0$. In this way, we can define that a property holds μ-everywhere also for signed measures, viz., by saying that it holds $|\mu|$ everywhere (in the traditional sense). Also the relation $\mu \ll \nu$ of absolute continuity between the signed measure μ and the positive measure ν can be redefined as saying that each ν-null set is a μ-null set. Thus $\mu \ll \nu$ is equivalent to $|\mu| \ll \nu$ and to both $\mu^+ \ll \nu$ and $\mu^- \ll \nu$. For the derivatives, it is easy to see that

Total variation $|\mu|$

$$\frac{d\mu}{d\nu} = \frac{d\mu^+}{d\nu} - \frac{d\mu^-}{d\nu},$$
$$\frac{d|\mu|}{d\nu} = \frac{d\mu^+}{d\nu} + \frac{d\mu^-}{d\nu}$$

hold.

We define integrability of a measurable function through $|\mu|$ by putting

$$\mathcal{L}_p(|\mu|) := \mathcal{L}_p(\mu^+) \cap \mathcal{L}_p(\mu^-),$$

and define $L_p(\mu)$ again as the set of equivalence classes.

These observations provide a convenient entry point into discussing complex measures. Call $\mu : \mathcal{A} \to \mathbb{C}$ a *(complex) measure* iff μ is σ-additive, i.e., iff $\mu(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$ for each sequence $(A_n)_{n\in\mathbb{N}}$ of mutually disjoint sets in $\mathcal{A}$. Then it can be easily shown that μ can be written as $\mu = \mu_r + i \cdot \mu_c$ with (real) signed measures μ_r and μ_c, which in turn have a Jordan decomposition and consequently a total variation each. In this way the L_p-spaces can be defined also for complex measures and complex measurable functions; the reader is referred to [Rud74] or [HS65] for further information.

Complex measure

Returning to the main current of the discussion, we are able to state the general representation of continuous linear functionals on an $L_p(\mu)$-space. We need only to sketch the proof, mutatis mutandis, since the main work has already been done in the proof of Theorem 4.11.30 for the real-valued and nonnegative case.

Theorem 4.11.33 *Assume that μ is a finite measure on $(X, \mathcal{A})$, $1 \leq p < \infty$ and that Λ is a bounded linear functional on $L_p(\mu)$. Then* *there exists a unique* $g \in L_q(\mu)$ *such that*

$$\Lambda(f) = \int_X f \cdot g \, d\mu$$

holds for each $f \in L_p(\mu)$*. In addition,* $\|\Lambda\| = \|g\|_q$*.*

Proof $\nu(A) := \Lambda(\chi_A)$ defines a signed measure on $\mathcal{A}$ with $\nu \ll \mu$. Let h be the Radon–Nikodym derivative of ν with respect to μ; then $h \in L_q(\mu)$ and

$$\Lambda(f) = \int_X f \cdot h \, d\mu$$

are shown as above. ⊣

It should be noted that Theorem 4.11.33 holds also for σ-finite measures and that it is true for $1 < p < \infty$ in the case of general (positive) measures; see, e.g., [Els99, §VII.3] for a discussion.

The case of continuous linear functionals for the space $L_\infty(\mu)$ is considerably more involved. Example 4.11.21 indicates already that these spaces play a special rôle. Looking back at the discussion above, we found that for $p < \infty$ the map $A \mapsto \int_A |f|^p \, d\mu$ yields a measure, and this measure was instrumental through the Radon–Nikodym Theorem for providing the factor which could be chosen to represent the linear functional. This argument, however, is not available for the case $p = \infty$, since we are not working there with a norm which is derived from an integral. It can be shown, however, that continuous linear functional has an integral representation with respect to finitely additive set functions; in fact, [HS65, Theorem 20.35] or [DS57, Theorem IV.8.16] shows that the continuous linear functionals on $L_\infty(\mu)$ are in a one-to-one correspondence with all finitely additive set functions ξ such that $\xi \ll \mu$. Note that this requires an extension of integration to not necessarily σ-additive set functions.

4.11.5 Disintegration

We provide another application of the Radon–Nikodym Theorem.

One encounters occasionally the situation of needing to decompose a measure on a product of two spaces. Consider this scenario. Given a measurable space $(X, \mathcal{A})$ as an input and $(Y, \mathcal{B})$ as an output space, let

$$(\mu \otimes K)(B) = \int_X K(x)(D_x)\, d\mu(x)$$

be the probability for $\langle x_1, x_2\rangle \in B \in \mathcal{A}\otimes\mathcal{B}$ with μ as the initial distribution and $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ as the transition law (see Example 4.9.4; think of an epidemic which is set off according to μ and propagates according to K). Assume that you want to reverse the process: Given $F \in \mathcal{B}$, you put

$$\nu(F) := \mathbb{S}(\pi_Y(\mu \otimes K))(F) = (\mu \otimes K)(X \times F),$$

so this is the probability that your process hits an element of F. Can you find a stochastic relation $L : (Y, \mathcal{B}) \rightsquigarrow (X, \mathcal{A})$ such that

$$(\mu \otimes K)(B) = \int_X L(x)(B^y)\, d\nu(y)$$

holds? The relation L is the *converse* of K given μ. It is probably not particularly important that the measure on the product has the shape $\mu\otimes K$, so we state the problem in such a way that we are given a measure on a product of two measurable spaces, and the question is whether we can decompose it into the product of a projection onto one space, and a stochastic relation between the spaces.

This problem is of course easiest dealt with when one can deduce that the measure is the product of measures on the coordinate spaces; probabilistically, this would correspond to the distribution of two independent random variables. But sometimes one is not so lucky, and there is some hidden dependence, or one simply cannot assess the degree of independence. Then one has to live with a somewhat weaker result: In this case one can decompose the measure into a measure on one component and a transition probability. This will be made specific in the discussion to follow.

Because it will not cost substantially more attention, we will treat the question a bit more generally. Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ be measurable spaces, assume that $\mu \in \mathbb{S}(X, \mathcal{A})$, and let $f : X \to Y$ and $g : X \to Z$ be measurable maps. Then $\mu_f := \mathbb{S}(f)(\mu)$ and $\mu_g := \mathbb{S}(g)(\mu)$ define subprobabilities on $(Y, \mathcal{B})$ resp. $(Z, \mathcal{C})$. μ_f and μ_g can be interpreted as the distribution of f resp. g under μ.

We will show that we can represent the joint distribution as

$$\mu(\{x \in X \mid f(x) \in B, g(x) \in C\}) = \int_B K(y)(C)\, d\mu_f(y),$$

where $K : (Y, \mathcal{B}) \rightsquigarrow (Z, \mathcal{C})$ is a suitable stochastic relation. This will require Z to be a Polish space with $\mathcal{C} = \mathcal{B}(Z)$.

Let us see how this corresponds to the initially stated problem. Suppose $X := Y \times Z$ with $\mathcal{A} = \mathcal{B} \otimes \mathcal{C}$, and let $f := \pi_Y$, $g := \pi_Z$; then

$$\begin{aligned} \mu_f(B) &= \mu(B \times Z), \\ \mu_g(C) &= \mu(Y \times Z), \\ \mu(B \times C) &= \mu(\{x \in X \mid f(x) \in B, g(x) \in C\}). \end{aligned}$$

Granted that we have established the decomposition, we can then write

$$\mu(B \times C) = \int_B K(y)(C)\, d\mu_f(y);$$

thus we have decomposed the probability on the product into a probability on the first component and, conditioned on the value the first component may take, a probability on the second factor.

Definition 4.11.34 *Using the notation from above, K is called a* regular conditional distribution of g given f *iff*

$$\mu(\{x \in X \mid f(x) \in B, g(x) \in C\}) = \int_B K(y)(C)\, \mu_f(dy)$$

holds for each $B \in \mathcal{B}, C \in \mathcal{C}$, where $K : (Y, \mathcal{B}) \rightsquigarrow (C, \mathcal{C})$ is a stochastic relation on $(X, \mathcal{A})$ and $(Z, \mathcal{C})$. If only $y \mapsto K(y)(C)$ is $\mathcal{B}$-measurable for all $C \in \mathcal{C}$, then it will be called a conditional distribution of g given f.

The existence of a regular conditional distribution will be established, provided Z is Polish with $\mathcal{C} = \mathcal{B}(Z)$. This will be accomplished in several steps: First the existence of a conditional distribution will be shown using the Radon–Nikodym Theorem. The latter construction will then be examined further. It will turn out that there exists a set of measure zero outside of which the conditional distribution behaves like a regular one, but at first sight only on an algebra of sets, not on the entire σ-algebra. But do not worry; the second step will apply a classical extension argument and yield a regular conditional distribution on the Borel sets, just as we want it. The proofs are actually a kind of a round trip through the important techniques from measure theory, with the Radon–Nikodym Theorem together in the driver's seat. It displays also some nice and helpful proof techniques.

We fix $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ as measurable spaces, assume that $\mu \in \mathbb{S}(X, \mathcal{A})$, and take $f : X \to Y$ and $g : X \to Z$ to be measurable maps. The measures $\mu_f := \mathbb{S}(f)(\mu)$ and $\mu_g := \mathbb{S}(g)(\mu)$ are defined as above as the distribution of f resp. g under μ.

The existence of a conditional distribution of g given f is established first, and it is shown that it is essentially unique.

Lemma 4.11.35 *Using the notation from above, then*

1. *there exists a conditional distribution K_0 of g given f,*
2. *if there is another conditional distribution K_0' of g given f, then there exists for any $C \in \mathcal{C}$ a set $N_C \in \mathcal{B}$ with $\mu_f(N_C) = 0$ such that $K_0(y)(C) = K_0'(C)$ for all $y \notin C$.*

Proof 1. Fix $C \in \mathcal{C}$; then

$$\varpi_C(B) := \mu(f^{-1}[B] \cap g^{-1}[C])$$

defines a subprobability measure ϖ_C on $\mathcal{B}$ which is absolutely continuous with respect to μ_g, because $\mu_g(B) = 0$ implies $\varpi_C(B) = 0$. The Radon–Nikodym Theorem 4.11.26 now gives a density $h_C \in \mathcal{F}(Y, \mathcal{B})$ with

$$\varpi_C(B) = \int_B h_C \, d\mu_f$$

for all $B \in \mathcal{B}$. Setting $K_0(y)(C) := h_C(y)$ yields the desired conditional distribution.

2. Suppose K_0' is another conditional distribution of g given f; then we have

$$\forall B \in \mathcal{B} : \int_B K_0(y)(C) \, d\mu_f(y) = \int_B K_0(y)(C) \, d\mu_f(y),$$

for all $C \in \mathcal{C}$, which implies that the set on which $K_0(\cdot)(C)$ disagrees with $K_0'(\cdot)(C)$ is μ_f-null. $\dashv$

Essential uniqueness may be strengthened if the σ-algebra $\mathcal{C}$ is countably generated and if the conditional distribution is regular.

Lemma 4.11.36 *Assume that K and K' are regular conditional distributions of g given f and that $\mathcal{C}$ has a countable generator. Then there exists a set $N \in \mathcal{B}$ with $\mu_f(N) = 0$ such that $K(y)(C) = K'(y)(C)$ for all $C \in \mathcal{C}$ and all $y \notin N$.*

Proof If $\mathcal{C}_0$ is a countable generator of $\mathcal{C}$, then

$$\mathcal{C}_f := \{\bigcap \mathcal{E} \mid \mathcal{E} \subseteq \mathcal{C}_0 \text{ is finite}\}$$

is a countable generator of $\mathcal{C}$ as well, and $\mathcal{C}_f$ is closed under finite intersections; note that $Z \in \mathcal{C}_f$. Construct for $D \in \mathcal{C}_f$ the set $N_D \in \mathcal{B}$ outside of which $K(\cdot)(D)$ and $K'(\cdot)(D)$ coincide, and define

$$N := \bigcup_{D \in \mathcal{C}_f} N_D \in \mathcal{B}.$$

Evidently, $\mu_f(N) = 0$. We claim that $K(y)(C) = K'(y)(C)$ holds for all $C \in \mathcal{C}$, whenever $y \notin N$. In fact, fix $y \notin N$, and let

$$\mathcal{C}_1 := \{C \in \mathcal{C} \mid K(y)(C) = K'(y)(C)\};$$

then $\mathcal{C}_1$ contains $\mathcal{C}_f$ by construction and is closed under complements and countable disjoint unions. Thus $\mathcal{C} = \sigma(\mathcal{C}_f) \subseteq \mathcal{C}_1$, by the π-λ-Theorem 1.6.30, and we are done. ⊣

Steps for the proof

We will show now that a regular conditional distribution of g given f exists. This will be done through several steps, given the construction of a conditional distribution K_0:

① A set $N_a \in \mathcal{B}$ is constructed with $\mu_f(N_a) = 0$ such that $K_0(y)$ is additive on a countable generator $\mathcal{C}_z$ for $\mathcal{C}$.

② We construct a set $N_z \in B$ with $\mu_f(N_z) = 0$ such that $K_0(y)(Z) \leq 1$ for $y \notin N_z$.

③ For each element G of $\mathcal{C}_z$, we will find a set $N_G \in \mathcal{B}$ with $\mu_f(N_G) = 0$ such that $K_0(y)(G)$ can be approximated from inside through compact sets, whenever $y \notin N_G$.

④ Then we will combine all these sets of μ_f-measure zero to produce a set $N \in \mathcal{B}$ with $\mu_f(N) = 0$ outside of which $K_0(y)$ is σ-additive on the generator $\mathcal{C}_z$ and hence can be extended to a measure on all of C.

Well, this looks like a full program, so let us get on with it.

Theorem 4.11.37 *Given measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$, a Polish space Z, a subprobability $\mu \in \mathbb{S}(X, \mathcal{A})$, and measurable maps $f : X \to Y$, $g : X \to Z$, there exists a regular conditional distribution K of g given f. K is uniquely determined up to a set of μ_f-measure zero.*

Proof 0. Since Z is a Polish space, its topology has a countable base. We infer from Lemma 4.3.2 that $\mathcal{B}(Z)$ has a countable generator $\mathcal{C}$. Then the Boolean algebra $\mathcal{C}_1$ generated by $\mathcal{C}$ is also a countable generator of $\mathcal{B}(Z)$.

1. Given $C_n \in \mathcal{C}_1$, we find by Proposition 4.10.12 a sequence $(E_{n,k})_{k\in\mathbb{N}}$ of compact sets in Z with

$$E_{n,1} \subseteq E_{n,2} \subseteq E_{n,3} \ldots \subseteq C_n$$

such that

$$\mu_g(C_n) = \sup_{k\in\mathbb{N}} \mu_g(E_{n,k}).$$

Then the Boolean algebra $\mathcal{C}_z$ generated by $\mathcal{C} \cup \{E_{n,k} \mid n, k \in \mathbb{N}\}$ is also a countable generator of $\mathcal{B}(Z)$.

2. From the construction of the conditional distribution of g given f, we infer that for disjoint $C_1, C_2 \in \mathcal{C}_z$

$$\begin{aligned}
&\int_Y K_0(y)(C_1 \cup C_2)\, d\mu_f(y) \\
&\quad= \mu(\{x \in X \mid f(x) \in B, g(x) \in C_1 \cup C_2\}) \\
&\quad= \mu(\{x \in X \mid f(x) \in B, g(x) \in C_1\}) + \\
&\qquad\qquad \mu(\{x \in X \mid f(x) \in B, g(x) \in C_2\}) \\
&\quad= \int_Y K_0(y)(C_1)\, d\mu_f(y) + \int_Y K_0(y)(C_2)\, d\mu_f(y).
\end{aligned}$$

Thus there exists $N_{C_1,C_2} \in \mathcal{B}$ with $\mu_f(N_{C_1,C_2}) = 0$ such that

$$K_0(y)(C_1 \cup C_2) = K_0(y)(C_1) + K_0(y)(C_2)$$

for $y \notin N_{C_1,C_2}$. Because $\mathcal{C}_z$ is countable, we may deduce (by taking the union of N_{C_1,C_2} over all pairs C_1, C_2) that there exists a set $N_a \in \mathcal{B}$ such that K_0 is additive outside N_a and $\mu_f(N_a) = 0$. This accounts for part ① in the plan above. ① ✓

3. By the previous arguments, it is easy to construct a set $N_z \in \mathcal{B}$ with $\mu_f(N_z) = 0$ such that $K_0(y)(Z) \leq 1$ for $y \notin N_z$ (part ②). ② ✓

4. Because

$$\begin{aligned}
&\int_Y K_0(y)(C_n)\, d\mu_f(y) \\
&\quad = \mu(f^{-1}\left[Y\right] \cap g^{-1}\left[C_n\right]) \\
&\quad = \mu_g(C_n) \\
&\quad = \sup_{k\in\mathbb{N}} \mu_g(E_{n,k}) \\
&\quad = \sup_{k\in\mathbb{N}} \int_Y K_0(y)(E_{n,k})\, \mu_f(dy) \qquad \text{(Levi's Theorem 4.8.2)} \\
&\quad = \int_Y \sup_{k\in\mathbb{N}} K_0(y)(E_{n,k})\, d\mu_f(y),
\end{aligned}$$

we find for each $n \in \mathbb{N}$ a set $N_n \in \mathcal{B}$ with

$$\forall y \notin N_n : K_0(y)(C_n) = \sup_{k\in\mathbb{N}} K_0(y)(E_{n,k})$$

③ ✓ and $\mu_f(N_n) = 0$. This accounts for part ③.

5. Now we may begin to work on part ④. Put

$$N := N_a \cup N_z \cup \bigcup_{n\in\mathbb{N}} N_n;$$

then $N \in B$ with $\mu_f(N) = 0$. We claim that $K_0(y)$ is a premeasure on $\mathcal{C}_z$ for each $y \notin N$. It is clear that $K_0(y)$ is additive on $\mathcal{C}_z$, hence monotone, so only σ-additivity has to be demonstrated: Let $(D_\ell)_{\ell\in\mathbb{N}}$ be a sequence in $\mathcal{C}_z$ that is monotonically decreasing with

$$\eta := \inf_{\ell\in\mathbb{N}} K_0(y)(D_\ell) > 0;$$

then we have to show that

$$\bigcap_{\ell\in\mathbb{N}} D_\ell \neq \emptyset.$$

We approximate the sets D_ℓ now by compact sets, so we assume that $D_\ell = C_{n_\ell}$ for some n_ℓ (otherwise the sets are compact themselves). By construction we find for each $\ell \in \mathbb{N}$ a compact set $E_{n_\ell,k_\ell} \subseteq C_\ell$ with

$$K_0(y)(C_{n_\ell} \setminus E_{n_\ell,k_\ell}) < \eta \cdot 2^{\ell+1},$$

then

$$E_r := \bigcap_{i=\ell}^{r} E_{n_\ell, k_\ell} \subseteq C_{n_r} = D_r$$

defines a decreasing sequence of compact sets with

$$K_0(y)(E_r) \geq K_0(y)(C_{n_r}) - \sum_{i=\ell}^{r} K_0(y)(E_{n_\ell, k_\ell}) > \eta/2, \mathit{and}$$

thus $E_r \neq \emptyset$. Since E_r is compact and decreasing, we know that the sequence has a nonempty intersection (otherwise one of the E_r would already be empty). We may infer

$$\bigcap_{\ell \in \mathbb{N}} D_\ell \supseteq \bigcap_{r \in \mathbb{N}} E_r \neq \emptyset.$$

6. The Extension Theorem 1.6.29 now tells us that there exists a unique extension of $K_0(y)$ from $\mathcal{C}_z$ to a measure $K(y)$ on $\sigma(\mathcal{C}_z) = \mathcal{B}(Z)$, whenever $y \notin N$. If, however, $y \in N$, then we define $K(y) := \nu$, where $\nu \in \mathbb{S}(Z)$ is arbitrary. Because

$$\begin{aligned}\int_B K(y)(C)\, d\mu_f(y) = \int_B K_0(y)(C)\, d\mu_f(y)\\ = \mu(\{x \in X \mid f(x) \in B, g(x) \in C\})\end{aligned}$$

holds for $C \in \mathcal{C}_z$, the π-λ-Theorem 1.6.30 asserts that this equality is valid for all $C \in \mathcal{B}(Z)$ as well.

Measurability of $y \mapsto K(y)(C)$ needs to be shown, and then we are done. We do this through the principle of good sets: Put

$$\mathcal{E} := \{C \in \mathcal{B}(Z) \mid y \mapsto K(y)(C) \text{ is } \mathcal{B} - \text{measurable}\}.$$

Then $\mathcal{E}$ is a σ-algebra, and $\mathcal{E}$ contains the generator $\mathcal{C}_z$ by construction; thus $\mathcal{E} = \mathcal{B}(Z)$. $\dashv$ ④ ✓

The scenario in which the space $X = Y \times Z$ with a measurable space $(Y, \mathcal{B})$ and a Polish space Z with $\mathcal{A} = \mathcal{B} \otimes \mathcal{B}(Z)$ with f and g as projections deserves particular attention. In this case we decompose a measure on A into its projection onto Z and a conditional distribution for the projection onto Z given the projection onto Y. This is sometimes called the *disintegration* of a measure $\mu \in \mathbb{S}(Y \times Z)$.

We state the corresponding proposition explicitly, since one needs it usually in this specialized form.

Proposition 4.11.38 *Given a measurable space $(Y, \mathcal{B})$ and a Polish space Z, there exists for every subprobability $\mu \in \mathbb{S}(Y \times Z, \mathcal{B} \otimes \mathcal{B}(Z))$ a regular conditional distribution of π_Z given π_Y, that is, a stochastic relation $K : (Y, \mathcal{B}) \rightsquigarrow (Z, \mathcal{B}(Z))$ such that*

$$\mu(E) = \int_Y K(y)(E_y)\, d\mathbb{S}(\pi_Y)(\mu)(y)$$

for all $E \in \mathcal{B} \otimes \mathcal{B}(Z)$. $\dashv$

The construction requires a Polish as one of the factors. The proof shows that it is indeed tightness which saves the day. Otherwise it would be difficult to make sure that the conditional distribution constructed above is σ-additive. We know from Proposition 4.10.12 that finite measures on a Polish space are tight. In fact, examples show that this assumption is in fact necessary: [Kel72] constructs a product measure on spaces which fail to be Polish, for which no disintegration exists.

4.12 Bibliographic Notes

Most topics of this chapter are fairly standard; hence there are plenty of sources to mention. One of my favorite texts is the rich compendium written by Bogachev [Bog07], not to forget Fremlin's massive [Fre08] or Halmos' classic [Hal50]. The discussion on Souslin's operation $\mathscr{A}(\mathcal{A})$ on a σ-algebra $\mathcal{A}$ is heavily influenced by Srivastava's representation [Sri98] of this topic, but see also [Par67, Arv76, Kel72]. The measure extension is taken from [Lub74], following a suggestion by S.M. Srivastava; the extension of a stochastic relation is from [Dob12b]. The approach to integration centering around B. Levi's Theorem is taken mostly from the elegant representation by Doob [Doo94, Chap. VI]; see also [Els99, Kapitel IV]. The introduction of the Daniell integral follows essentially [Bog07, Sect. 7.8]; see also [Kel72]. The logic `CSL` is defined and investigated in terms of model checking in [BHHK03], and the stochastic interpretation is taken from [Dob07]; see also [DP03]. The Hutchinson metric is discussed in detail in Edgar's monograph [Edg98], from which the present proof of Proposition 4.10.13 is taken. There are many fine books on Banach spaces, Hilbert spaces, and the application to L^p spaces; my sources are [Doo94, Hal50, Rud74, DS57, Loo53, Sch70]. The exposition of projective limits and of disintegration follows basically [Par67, Chap. V] with an occasional glimpse at [Bog07].

4.13 Exercises

Exercise 4.1 Assume that $\mathcal{A} = \sigma(\mathcal{A}_0)$. Show that the weak σ-algebra $\wp(\mathcal{A})$ on $\mathbb{M}(X, \mathcal{A})$ is the initial σ-algebra with respect to $\{ev_A \mid A \in \mathcal{A}_0\}$.

Show also that both $\mathbb{S}(X, \mathcal{A})$ and $\mathbb{P}(X, \mathcal{A})$ are measurable subsets of $\mathbb{M}(X, \mathcal{A})$.

Exercise 4.2 Let (X, τ) be a topological and (Y, d) a metric space. Each continuous function $X \to Y$ is also Baire measurable.

Exercise 4.3 Let (X, d) be a separable metric space, $\mu \in \mathbb{M}(X, \mathcal{B}(X))$. Show that $x \in \text{supp}(\mu)$ iff $\mu(U) > 0$ for each open neighborhood U of x.

Exercise 4.4 Let $(X, \mathcal{A}, \mu)$ be a finite measure space. Show that norm convergence in $L_\infty(X, \mathcal{A}, \mu)$ implies convergence almost everywhere ($f_n \xrightarrow{a.e.} f$, provided $||f_n - f||_\infty^\mu \to 0$). Give an example showing that the converse is false.

Exercise 4.5 If $\mathcal{A}$ is a σ-algebra on X and $B \subseteq X$ with $A \notin \mathcal{A}$, then

$$\{(A_1 \cap B) \cup (A_2 \cap (X \setminus B)) \mid A_1, A_2 \in \mathcal{A}\}$$

is the smallest σ-algebra $\sigma(\mathcal{A} \cup \{B\})$ on X containing $\mathcal{A}$ and B. If τ is a topology on X with $H \notin \tau$, then

$$\{G_1 \cup (G_2 \cap H) \mid G_1, G_2 \in \tau\}$$

is the smallest topology τ_H on X containing τ and H. Show that $\mathcal{B}(\tau_H) = \sigma(\mathcal{A} \cup \{H\})$

Exercise 4.6 Let $(X, \mathcal{A}, \mu)$ be a finite measure space, $B \notin \mathcal{A}$, and $\beta := \alpha \cdot \mu_*(B) + (1 - \alpha) \cdot \mu^*(B)$ with $0 \leq \alpha \leq 1$. Then there exists a measure ν on $\sigma(\mathcal{A} \cup \{B\})$ which extends μ such that $\nu(B) = \beta$. (Hint: Exercise 4.5).

Exercise 4.7 Given the measurable space $(X, \mathcal{A})$ and $f \in \mathcal{F}(X, \mathcal{A})$ with $f \geq 0$, show that there exists a decreasing sequence $(f_n)_{n \in \mathbb{N}}$ of step functions $f_n \in \mathcal{F}(X, \mathcal{A})$ with

$$f(x) = \inf_{n \in \mathbb{N}} f_n(x)$$

for all $x \in X$.

Exercise 4.8 Let $f_i : X_i \to Y_i$ be $\mathcal{A}_i$-$\mathcal{B}_i$-measurable maps for $i \in I$. Show that

$$f : \begin{cases} \prod_{i \in I} X_i & \to \prod_{i \in I} Y_i \\ (x_i)_{i \in I} & \mapsto (f_i(x_i))_{i \in I} \end{cases}$$

is $\bigotimes_{i \in I} \mathcal{A}_i$-$\bigotimes_{i \in I} \mathcal{B}_i$-measurable. Conclude that the kernel of f

$$\ker(f) := \{\langle x, x' \rangle \mid f(x) = f(x')\}$$

is a measurable subset of $Y \times Y$, whenever $f : (X, \mathcal{A}) \to (Y, \mathcal{B})$ is measurable and $\mathcal{B}$ is separable.

Exercise 4.9 Let $f : X \to Y$ be $\mathcal{A}$-$\mathcal{B}$ measurable, and assume that $\mathcal{B}$ is separable. Show that the graph $\text{graph}(f) := \{\langle x, f(x) \rangle \mid x \in X\}$ of f is a measurable subset of $\mathcal{A} \otimes \mathcal{B}$.

Exercise 4.10 Let χ_A be the indicator function of set A. Show that

1. $A \subseteq B$ iff $\chi_A \leq \chi_B$,
2. $\chi_{\bigcup_{n \in \mathbb{N}} A_n} = \sup_{n \in \mathbb{N}} \chi_{A_n}$ and $\chi_{\bigcap_{n \in \mathbb{N}} A_n} = \inf_{n \in \mathbb{N}} \chi_{A_n}$
3. $\chi_{A \Delta B} = |\chi_A - \chi_B| = \chi_A + \chi_B \pmod 2$. Conclude that the power set $(\mathcal{P}(X), \Delta)$ is a commutative group with $A \Delta A = \emptyset$.
4. $\left(\bigcup_{n \in \mathbb{N}} A_n\right) \Delta \left(\bigcup_{n \in \mathbb{N}} B_n\right) \subseteq \bigcup_{n \in \mathbb{N}} (A_n \Delta B_n)$

Exercise 4.11 Let $(X, \mathcal{A}, \mu)$ be a finite measure space, and put $d(A, B) := \mu(A \Delta B)$ for $A, B \in \mathcal{A}$. Show that $(\mathcal{A}, d)$ is a complete pseudometric space.

Exercise 4.12 (This Exercise draws heavily on Exercises 4.5 and 4.6). Let $X := [0, 1]$ with λ as the Lebesgue measure on the Borel set of X. There exists a set $B \subseteq X$ with $\lambda_*(B) = 0$ and $\lambda^*(B) = 1$ by Lemma 1.7.7, so that $B \notin \mathcal{B}(X)$.

1. Show that (X, τ_B) is a Hausdorff space with a countable base, where τ_B is the smallest topology containing the interval topology on $[0, 1]$ and B (see Exercise 4.5).
2. Extend λ to a measure μ with $\alpha = 1/2$ in Exercise 4.6.
3. Show that $\inf\{\mu(G) \mid G \supseteq X \setminus B \text{ and } G \text{ is } \tau_B\text{-open}\} = 1$, but $\mu(X \setminus B) = 1/2$. Thus μ is not regular (since (X, τ_B) is not a metric space).

Exercise 4.13 Prove Proposition 4.3.18.

Exercise 4.14 Let $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ be a transition kernel.

1. Assume that $f \in \mathcal{F}_+(Y, \mathcal{B})$ is integrable with respect to $K(x)$ for all $x \in X$. Show that
$$K(f)(x) := \int_X f \, dK(x)$$
defines a measurable function $K(f) : X \to \mathbb{R}_+$.

2. Assume that $x \mapsto K(x)(Y)$ is bounded. Define for $B \in \mathcal{B}$
$$\overline{K}(\mu)(B) := \int_X K(x)(B) \, d\mu(x).$$
Show that $\overline{K}: \mathbb{S}(X, \mathcal{A}) \to \mathbb{S}(Y, \mathcal{B})$ is $\wp(X, \mathcal{A})$-$\wp(Y, \mathcal{B})$-measurable (see Example 2.4.8).

Exercise 4.15 Let $\mu \in \mathbb{S}(X, \mathcal{A})$ be s subprobability measure on $(X, \mathcal{A})$, and let $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ be a stochastic relation. Assume that $f : X \times Y \to \mathbb{R}$ is bounded and measurable. Show that
$$\int_{X \times Y} f \, d\mu \otimes K = \int_X \big(\int_Y f_x \, dK(x) \big) \, d\mu(x)$$
($\mu \otimes K$ is defined in Example 4.9.4 on page 560).

Exercise 4.16 Let $K : (X, \mathcal{A}) \rightsquigarrow (Y, \mathcal{B})$ and $L : (Y, \mathcal{B}) \rightsquigarrow (Z, \mathcal{C})$ be stochastic relations. Then the convolution $L * K$ can be represented as $(L * K)(x) = \mathbb{S}(\pi_Z)(K(x) \otimes L)$.

Exercise 4.17 Let $S := \{1, \ldots, n\}$ for some $n \in \mathbb{N}$. Show that the weak topology on $\mathbb{M}(S, \mathcal{P}(S))$ can be identified with the Euclidean topology on $(\mathbb{R}_+)^n$.

Exercise 4.18 Let $(S, \mathcal{A})$ be a measurable space, and assume that $\mathcal{A}$ is countably generated. Show that a stochastic effectivity function $P : S \to \boldsymbol{E}(S)$ is $\mathcal{A}$-$\mathcal{B}(\tau)$-measurable, where τ is the Priestley topology on $\wp(S, \mathcal{A})$. This topology is defined in Example 1.5.58 on page 63.

Exercise 4.19 Show that the category of analytic spaces with measurable maps is not closed under taking pushouts. **Hint**: Show that the pushout of X/α_1 and X/α_2 is $X/(\alpha_1 \cup \alpha_2)$ for equivalence relations α_1 and α_2 on a Polish space X. Then use Proposition 4.4.22 and Example 4.4.29.

Exercise 4.20 Let X and Y be Polish spaces with a transition kernel $K : X \rightsquigarrow Y$. The equivalence relations α on X and β on Y are assumed to be smooth with determining sequences $(A_n)_{n\in\mathbb{N}}$ resp. $(B_n)_{n\in\mathbb{N}}$ of Borel sets. Put $\mathcal{I}_\alpha := \sigma(\{A_n \mid n \in \mathbb{N}\}$ and $\mathcal{J}_\beta := \sigma(\{B_n \mid n \in \mathbb{N}\}$. Show that the following statements are equivalent:

1. $K : (X, \mathcal{I}_\alpha) \rightsquigarrow (Y, \mathcal{J}_\beta)$ is a transition kernel.

2. (α, β) is a congruence for K.

3. $\alpha \subseteq \ker\big(\mathbb{S}(\eta_\beta \circ K)\big)$.

4. There exists a transition kernel $K' : (X, \mathcal{I}_\alpha) \rightsquigarrow (Y, \mathcal{J}_\beta)$ such that $(i_\alpha, j_\beta) : K \to K'$ is a morphism, where the measurable maps $i_\alpha : (X, \mathcal{B}(X)) \to (X, \mathcal{I}_\alpha)$ and $j_\beta : (Y, \mathcal{B}(Y)) \to (Y, \mathcal{I}_\beta)$ are given by the respective identities.

Exercise 4.21 Let $\mathtt{S}_X$ be the set of all smooth equivalence relations on the Polish space X, which is ordered by inclusion. Then $\mathtt{S}_X$ is closed under countable infima, and $\Delta_X \subseteq \rho \subseteq \nabla_X$, where $\nabla_X := X \times X$ is the universal relation.

1. $\rho \mapsto \{A \in \mathcal{B}(X) \mid A \text{ is } \rho - \text{invariant}\}$ is an order reversing bijection between $\mathtt{S}_X$ and the countably generated sub-σ-algebras of $\mathcal{B}(X)$ such that $\Delta_X \mapsto \mathcal{B}(X)$ and $\nabla_X \mapsto \{\emptyset, X\}$.

2. Define for $x, x' \in X$ with $x \neq x'$ the equivalence relation $\vartheta_{x,x'} := \Delta_X \cup \{\langle x, x'\rangle, \langle x', x\rangle\}$. Then $\vartheta_{x,x'}$ is an atom of $\mathtt{S}_X$. Describe the σ-algebra of $\vartheta_{x,x'}$-invariant Borel sets.

3. Define for the Borel set B with $\emptyset \neq B \neq X$ the equivalence relation τ_B through $x \; \tau_B \; x'$ iff $\{x, x'\} \subseteq B$ or $\{x, x'\} \cap B = \emptyset$ for all $x, x' \in X$. Then τ_B is an anti-atom in $\mathtt{S}_X$ (i.e., an atom in the reverse order). Describe the σ-algebra of τ_B-invariant Borel sets.

4. Show that for each $\rho \in \mathtt{S}_X$, there exists a countable family $(\beta_n)_{n\in\mathbb{N}}$ of anti-atoms with $\rho = \bigwedge_{n\in\mathbb{N}} \beta_n$.

5. Show that $\tau_B \wedge \vartheta_{x,x'} = \Delta_X$ and $\tau_B \vee \vartheta_{x,x'} = \nabla_X$, whenever B is a Borel set with $\emptyset \neq B \neq X$ and $x \in B$, $x' \notin B$.

Exercise 4.22 Let α and β be smooth equivalence relations on the Polish spaces X resp. Y, and assume that we have an injective map f :

$X/\alpha \to Y/\beta$. Define $f^* : \boldsymbol{\Sigma}_\alpha(X) \to \boldsymbol{\Sigma}_\beta(Y)$ through $f^*(A) := \bigcup\{f([x]_\alpha) \mid [x]_\alpha \in A\}$. Show that f^* is an isomorphism.

Exercise 4.23 Let Y be a Polish space, $F : X \to \mathbb{F}(Y)$ be a map, and $\mathcal{L}$ be an algebra of sets on X. We assume that $F^w(G) \in \mathcal{L}_\sigma$ for each open $G \subseteq Y$ (F^w is defined on page 538). Show that there exists a map $s : X \to Y$ such that $s(x) \in F(x)$ for all $x \in X$ such that $s^{-1}\left[B\right] \in \mathcal{L}_\sigma$ for each $B \in \mathcal{B}(Y)$. **Hint**: Modify the proof for Theorem 4.7.2 suitably.

Exercise 4.24 Given a finite measure space $(X, \mathcal{A}, \mu)$, let $f = \sum_{i=1}^n \alpha_i \cdot \chi_{A_i}$ be a step function with $A_1, \ldots, A_n \in \mathcal{A}$ and coefficients $\alpha_1, \ldots,$ α_n. Show that

$$\sum_{i=1}^n \alpha_i \cdot \mu(A_i) = \sum_{\gamma>0} \gamma \cdot \mu(\{x \in X \mid f(x) = \gamma\}).$$

Exercise 4.25 Let $(Y, \mathcal{B})$ be a measurable space and assume that X is compact metric; $\mathfrak{C}(X)$ is the set of all nonempty compact subsets of X endowed with the Hausdorff metric δ_H; see Example 3.5.10. Show that $F : Y \to \mathfrak{C}(X)$ is $\mathcal{B}$-$\mathcal{B}(\mathfrak{C}(X))$ measurable iff F is measurable as a relation (in the sense of Definition 4.7.1 on page 538).

Exercise 4.26 Show that $(\mathfrak{C}(X), \delta_H)$ is second countable iff (X, d) is.

Exercise 4.27 Let $(X, \mathcal{A})$ be a measurable space with two effectivity functions P and Q on it.

1. Define for $A \in \mathcal{A}$ and $0 \leq q \leq 1$

$$P^+(A, q) := \{x \in X \mid \boldsymbol{\beta}_\mathcal{A}(A, > q) \in P(x)\}.$$

   Show that $P^+ : \mathcal{A} \times [0, 1] \to \mathcal{A}$ such that $A \mapsto P^+(A, q)$ is monotone for each q.

2. Put

$$G_Q(A, q) := \{\nu \in \mathbb{S}(X, \mathcal{A}) \mid \int_0^1 \nu(Q^+(A, r)\, dr \geq q\}.$$

   Show that $(P^+ * Q^+)(A, q) := \{x \in X \mid G_Q(A, q) \in P(s)\}$ defines a map $\mathcal{A} \times [0, 1] \to \mathcal{A}$ such that $A \mapsto (P^+ * Q^+)(A, q)$ is monotone for each q.

This construction serves as a stand-in for the Kleisli product in the interpretation of game logic in Sect. 4.9.4.

Exercise 4.28 Given the plane $E := \{\langle x_1, x_2, x_3\rangle \in \mathbb{R}^3 \mid 2 \cdot x_1 + 4 \cdot x_2 - 7 \cdot x_3 = 12\}$, determine the point in E which is closest to $\langle 4, 2, 0\rangle$ in the Euclidean distance.

Exercise 4.29 Let $(V, \|\cdot\|)$ be a real normed space and $L : V \to \mathbb{R}$ be linear. Show that L is continuous iff L is bounded, i.e., iff $\sup_{\|v\|\leq 1} |L(v)| < \infty$.

Exercise 4.30 Let $(V, \|\cdot\|)$ be a real normed space, and define

$$V^* := \{L : V \to \mathbb{R} \mid L \text{ is linear and continuous}\},$$

the *dual space* of V. Then V^* is a vector space. Show that

$$\|L\| := \sup_{\|v\|\leq 1} |L(v)|$$

defines a norm on V^* with which $(V^*, \|\cdot\|)$ is a Banach space.

Exercise 4.31 Let H be a Hilbert space, then H^* is isometrically isomorphic to H.

Exercise 4.32 Let $(V, \|\cdot\|)$ be a real normed space, and define

$$\pi(x)(L) := L(x)$$

for $x \in V$ and $L \in V^*$.

1. Show that $\pi(x) \in V^{**}$ and that $x \mapsto \pi(x)$ defines a continuous map $V \to V^{**}$.

2. Given $x \in V$ with $x \neq 0$, there exists $L \in V^*$ with $\|L\| = 1$ and $L(x) = \|x\|$ (use the Hahn–Banach Theorem 1.5.14).

3. Show that π is an isometry (thus a normed space can be embedded isometrically into its bidual).

Exercise 4.33 Given a real vector space V:

1. Let $(\cdot, \cdot)$ be an inner product on V. Show that the parallelogram law

$$\|x + y\|^2 + \|x - y\|^2 = 2 \cdot \|x\|^2 + 2 \cdot \|y\|^2$$

always holds (see page 641).

2. Assume, conversely, that $\| \cdot \|$ is a norm for which the parallelogram law holds. Show that

$$(x, y) := \frac{\|x + y\|^2 - \|x - y\|^2}{4}$$

defines an inner product on V.

Exercise 4.34 Let H be a Hilbert space and $L : H \to \mathbb{R}$ be a continuous linear map with $L \neq 0$. Relating *Kern*(L) and ker (L), show that $H/\mathit{Kern}(L)$ and $\mathbb{R}$ are isomorphic as vector spaces.

List of Examples

This is a list of most examples together with a short description, ordered by chapters. It should help to find an example quickly.

The Axiom of Choice and Some of Its Equivalents

E.-E. Doberkat, *Special Topics in Mathematics for Computer Scientists*, DOI 10.1007/978-3-319-22750-4

Categories

Topological Spaces

Measures for Probabilistic Systems

Bibliography

[Acz00] A. D. Aczel. *The Mystery of the Aleph - Mathematics, the Kabbalah and the Search for Infinity*. Pocket Books, New York, 2000. ⟨75⟩

[AJ94] S. Abramsky and A. Jung. Domain theory. In A. Abramsky, D. M. Gabbay, and T. S. E. Maibaum, editors, *Handbook of Logic in Computer Science*, volume 3 — Semantic Structures, pages 1–168. Oxford University Press, Oxford, 1994. ⟨282, 306⟩

[Arv76] W. Arveson. *An Invitation to C*-Algebra*. Number 39 in Graduate Texts in Mathematics. Springer-Verlag, New York, Berlin, Heidelberg, 1976. ⟨369, 492⟩

[Aum54] G. Aumann. *Reelle Funktionen*. Springer-Verlag, Berlin, Göttingen, Heidelberg, 1954. ⟨52, 75⟩

[Awo10] S. Awodey. *Category Theory*. Number 52 in Oxford Logic Guides. Oxford University Press, Oxford, 2nd edition, 2010. ⟨197⟩

[Bar77] J. Barwise, editor. *Handbook of Mathematical Logic*. North-Holland Publishing Company, Amsterdam, 1977. ⟨25⟩

[Bar01] L. M. Barbosa. *Components as Coalgebras*. PhD thesis, Universidade do Minho, 2001. ⟨79⟩

[BdRV01] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Number 53 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge, UK, 2001. ⟨84, 197⟩

E.-E. Doberkat, *Special Topics in Mathematics for Computer Scientists*, DOI 10.1007/978-3-319-22750-4

[BHHK03] C. Baier, B. Haverkort, H. Hermanns, and J.-P. Katoen. Model-checking algorithms for continuous time Markov chains. *IEEE Trans. Softw. Eng.*, 29(6):524–541, June 2003. $\langle 195, 492 \rangle$

[Bil95] P. Billingsley. *Probability and Measure*. John Wiley & Sons, New York, third edition, 1995. $\langle 75 \rangle$

[Bir67] G. Birkhoff. *Lattice Theory*, volume 25 of *Colloquium Publications*. American Mathematical Society, Providence, RI, 1967. $\langle 75 \rangle$

[Blo08] W. B. Bloch. *The Unimaginable Mathematics of Borges' Library of Babel*. Oxford University Press, Oxford, 2008. $\langle x \rangle$

[BN98] F. Baader and T. Nipkow. *Term rewriting and* all that. Cambridge University Press, Cambridge, UK, 1998. $\langle 75 \rangle$

[Bog07] V. I. Bogachev. *Measure Theory*. Springer-Verlag, 2007. $\langle 60, 75, 334, 404, 492 \rangle$

[Bor94a] F. Borceux. *Handbook of Categorical Algebra 1: Basic Category Theory*, volume 50 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, Cambridge, UK, 1994. $\langle 197 \rangle$

[Bor94b] F. Borceux. *Handbook of Categorical Algebra 2: Categories and Structures*, volume 51 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, Cambridge, UK, 1994. $\langle 197 \rangle$

[Bou89] N. Bourbaki. *General Topology*. Elements of Mathematics. Springer-Verlag, Berlin, Heidelberg, New York, 1989. $\langle 249, 303, 306 \rangle$

[Bro08] T. J. Bromwich. *In Introduction to the Theory of Infinite Series*. MacMillan and Co., London, 1908. $\langle 75 \rangle$

[BW99] M. Barr and C. Wells. *Category Theory for Computing Science*. Les Publications CRM, Montreal, 1999. $\langle 118, 197 \rangle$

[CFO89] D. Cantone, A. Ferro, and E. G. Omodeo. *Computable Set Theory*, volume 6 of *International Series of Monographs on Computer Science*. Oxford University Press, Oxford, 1989. ⟨1⟩

[CGP99] E. M. Clarke, O. Grumberg, and D. A. Peled. *Model Checking*. The MIT Press, Cambridge, MA, 1999. ⟨195⟩

[CH67] R. Courant and D. Hilbert. *Methoden der Mathematischen Physik I*, volume 30 of *Heidelberger Taschenbücher*. Springer-Verlag, third edition, 1967. ⟨286⟩

[Che89] B. F. Chellas. *Modal Logic*. Cambridge University Press, Cambridge, UK, 1989. ⟨146⟩

[Chr64] G. Chrystal. *Textbook of Algebra I/II*. Chelsea Publishing Company (Reprint), New York, 1964. ⟨75⟩

[CK90] C. C. Chang and H. J. Keisler. *Model Theory*, volume 73 of *Studies in Logic and the Foundations of Mathematics*. Elsevier, Amsterdam, 1990. ⟨266, 270, 271, 306⟩

[CLR92] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *An Introduction to Algorithms*. The MIT Press, Cambridge MA, 1992. ⟨25⟩

[COP01] D. Cantone, E. G. Omodeo, and A. Policriti. *Set Theory for Computing*. Springer-Verlag, Berlin, Heidelberg, New York, 2001. ⟨1, 75⟩

[CV77] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*. Number 580 in Lect. Notes Math. Springer-Verlag, Berlin, Heidelberg, New York, 1977. ⟨259⟩

[Dav00] M. Davis. *Engines of Logic - Mathematics and the Origin of the Computer*. W. W. Norton, New York, London, 2000. ⟨75⟩

[DEP02] J. Desharnais, A. Edalat, and P. Panangaden. Bisimulation of labelled Markov processes. *Information and Computation*, 179(2):163–193, 2002. ⟨459⟩

[DF89] E.-E. Doberkat and D. Fox. *Software Prototyping mit SETL*. Leitfäden und Monographien der Informatik. Teubner-Verlag, Stuttgart, 1989. ⟨1⟩

[Dob89] E.-E. Doberkat. Topological completeness in an ideal model for recursive polymorphic types. *SIAM J. Computing*, 18(5):977 – 991, 1989. ⟨306⟩

[Dob03] E.-E. Doberkat. Pipelines: Modelling a software architecture through relations. *Acta Informatica*, 40:37–79, 2003. ⟨79, 197⟩

[Dob07] E.-E. Doberkat. The Hennessy-Milner equivalence for continuous-times stochastic logic with mu-operator. *J. Appl. Logic*, 35:519–544, 2007. ⟨195, 492⟩

[Dob09] E.-E. Doberkat. *Stochastic Coalgebraic Logic*. EATCS Monographs in Theoretical Computer Science. Springer-Verlag, Berlin, 2009. ⟨197, 434⟩

[Dob10] E.-E. Doberkat. A note on the coalgebraic interpretation of game logic. *Rendiconti Ist. di Mat. Univ. di Trieste*, 42:191 – 204, 2010. ⟨197⟩

[Dob12a] E.-E. Doberkat. *Haskell für Objektorientierte*. Oldenbourg-Verlag, München, 2012. ⟨197, 307⟩

[Dob12b] E.-E. Doberkat. A stochastic interpretation of propositional dynamic logic: Expressivity. *J. Symb. Logic*, 77(2):687–716, 2012. ⟨387, 492⟩

[Dob14] E.-E. Doberkat. Algebraic properties of stochastic effectivity functions. *J. Logic and Algebraic Progr.*, 83:339–358, 2014. ⟨320⟩

[Doo94] J. L. Doob. *Measure Theory*. Number 143 in Graduate Texts in Mathematics. Springer Verlag, 1994. ⟨492⟩

[DP02] B.A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, Cambridge, UK, 2002. ⟨75⟩

[DP03] J. Desharnais and P. Panangaden. Continuous stochastic logic characterizes bisimulation of continuous-time Markov processes. *J. Log. Alg. Programming*, 56(1-2):99–115, 2003. ⟨492⟩

[DPP09] A. Doxiadis, C. Papadimitriou, and A. Papadatos. *Logicomix: An Epic Search for Truth*. Bloomsbury, London, 2009. ⟨76⟩

[DS57] N. Dunford and J. T. Schwartz. *Linear Operators*, volume I. Interscience Publishers, 1957. ⟨486, 492⟩

[DS65] L. E. Dubins and L. J. Savage. *How to Gamble if you Must: Inequalities for Stochastic Processes*. McGraw-Hill, New York, 1965. ⟨321⟩

[DS11] E.-E. Doberkat and Ch. Schubert. Coalgebraic logic over general measurable spaces - a survey. *Math. Struct. Comp. Science*, 21:175–234, 2011. Special issue on coalgebraic logic. ⟨193, 197⟩

[DT41] E.-E. Doberkat and P. Sànchez Terraf. Stochastic nondeterminism and effectivity functions. *J. Logic and Computation*, in print, 2015 (arxiv: 1405.7141). ⟨320⟩

[Eda99] A. Edalat. Semi-pullbacks and bisimulations in categories of Markov processes. *Math. Struct. Comp. Science*, 9(5):523–543, 1999. ⟨459⟩

[Edg98] G. A. Edgar. *Integral, Probability, and Fractal Measures*. Springer-Verlag, New York, 1998. ⟨448, 492⟩

[Els99] J. Elstrodt. *Maß- und Integrationstheorie*. Springer-Verlag, Berlin-Heidelberg-New York, 2 edition, 1999. ⟨62, 75, 399, 486, 492⟩

[Eng89] R. Engelking. *General Topology*, volume 6 of *Sigma Series in Pure Mathematics*. Heldermann-Verlag, Berlin, revised and completed edition, 1989. ⟨225, 228, 241, 249, 259, 303, 306⟩

[Fia05] J. L. Fiadeiro. *Categories for Software Engineering*. Springer-Verlag, Berlin, Heidelberg, 2005. ⟨79⟩

[Fic64] G. M. Fichtenholz. *Differential- und Integralrechnung I-III*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1964. ⟨41, 480⟩

[Fre08] D. H. Fremlin. *Measure Theory*, volume 1–5. Torres Fremlin, Colchester, 2000–2008. ⟨492⟩

[GHK+03] G. Gierz, K. H. Hofmann, K. Keimel, J. D. Lawson, M. W. Mislove, and D. S. Scott. *Continuous Lattices and Domains*. Number 93 in Encyclopaedia of Mathematics and its Applications. Cambridge University Press, Cambridge, UK, 2003. ⟨279, 280, 282, 306⟩

[Gir81] M. Giry. A categorical approach to probability theory. In *Categorical Aspects of Topology and Analysis*, number 915 in Lect. Notes Math., pages 68–85, Berlin, 1981. Springer-Verlag. ⟨197⟩

[GKP89] R. E. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, MA, 1989. ⟨ix⟩

[Gol96] D. Goldrei. *Classic Set Theory*. Chapman and Hall/CRC, Boca Raton, 1996. ⟨75⟩

[Gol06] R. Goldblatt. *Topoi — The Categorical Analysis of Logic*. Dover Publications, New York, 2006. ⟨306⟩

[Gol10] R. Goldblatt. Deduction systems for coalgebras over measurable spaces. *Journal of Logic and Computation*, 20(5):1069–1100, 2010. ⟨325, 326⟩

[Gol12] R. Goldblatt. Topological proofs of some Rasiowa-Sikorski lemmas. *Studia Logica*, 100:1–18, 2012. ⟨45⟩

[Hal50] P. R. Halmos. *Measure Theory*. Van Nostrand Reinhold, New York, 1950. ⟨483, 492⟩

[Her06] H. Herrlich. *Axiom of Choice*. Number 1876 in Lect. Notes Math. Springer-Verlag, Berlin, Heidelberg, New York, 2006. ⟨6, 11, 75, 76, 224, 225, 306, 459⟩

[HS65] E. Hewitt and K. R. Stromberg. *Real and Abstract Analysis*. Springer-Verlag, Berlin, Heidelberg, New York, 1965. ⟨44, 480, 485, 486⟩

[Isb64] J. R. Isbell. *Uniform Spaces*. Number 12 in Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 1964. ⟨306⟩

[Jam87] I. M. James. *Topological and Uniform Spaces*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, Berlin, Heidelberg, 1987. ⟨306⟩

[Jec73] T. Jech. *The Axiom of Choice*, volume 75 of *Studies in Logic and the Foundations of Mathematics*. North-Holland Publishing Company, New York, 1973. ⟨75⟩

[Jec06] T. Jech. *Set Theory*. Springer-Verlag (The Third Millennium Edition), Berlin, Heidelberg, New York, 2006. ⟨37, 53, 69, 75⟩

[Joh82] P. T. Johnstone. *Stone Spaces*. Cambridge University Press, Cambridge, UK, 1982. ⟨306⟩

[JR02] J. E. Jayne and C. A. Rogers. *Selectors*. Princeton University Press, Princeton, N. J., 2002. ⟨259⟩

[Kec94] A. S. Kechris. *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. Springer-Verlag, Berlin, Heidelberg, New York, 1994. ⟨265, 306, 324, 368⟩

[Kee93] J. P. Keener. The Perron-Frobenius theorem and the ranking of football teams. *SIAM Review*, 35(1):80–93, 1993. ⟨254, 255⟩

[Kel55] J. L. Kelley. *General Topology*. Number 27 in Graduate Texts in Mathematics. Springer Verlag, New York, Berlin, Heidelberg, 1955. ⟨208, 225, 238, 249, 262, 305, 306⟩

[Kel72] H. G. Kellerer. Topologische Maßtheorie. Lecture notes, Mathematisches Institut, Ruhr-Universität Bochum, Sommersemester 1972. ⟨492⟩

[KM76] K. Kuratowski and A. Mostowski. *Set Theory*, volume 86 of *Studies in Logic and the Foundations of Mathematics*. North-Holland and PWN, Polish Scientific Publishers, Amsterdam and Warzawa, 1976. ⟨11, 15, 75, 368⟩

[Knu73] D. E. Knuth. *The Art of Computer Programming. Vol. I, Fundamental Algorithms*. Addison-Wesley, Reading, Mass., 2 edition, 1973. ⟨25⟩

[Kop89] S. Koppelberg. *Handbook of Boolean Algebras*, volume 1. North Holland, Amsterdam, 1989. ⟨271, 306⟩

[Koz85] D. Kozen. A probabilistic PDL. *J. Comp. Syst. Sci.*, 30(2):162–178, 1985. ⟨436⟩

[Kur66] K. Kuratowski. *Topology*, volume I. PWN – Polish Scientific Publishers and Academic Press, Warsaw and New York, 1966. ⟨229, 259, 306, 360⟩

[Lip11] M. Lipovača. *Learn You a Haskell for Great Good!* no starch press, San Francisco, 2011. ⟨197⟩

[LM05] A. M. Langville and C. D. Mayer. A survey of eigenvector methods for Web information retrieval. *SIAM Review*, 47(1):135–161, 2005. ⟨254, 255⟩

[Loo53] L. H. Loomis. *An Introduction to Abstract Harmonic Analysis*. D. van Nostrand Company, Princeton, N. J., 1953. ⟨492⟩

[Lub74] A. Lubin. Extensions of measures and the von Neumann selection theorem. *Proc. Amer. Math. Soc.*, 43(1):118–122, 1974. ⟨492⟩

[Mic51] E. Michael. Topologies on spaces of subsets. *Trans. Am. Math. Soc.*, 71(2):152–182, 1951. ⟨259⟩

[ML97] S. Mac Lane. *Categories for the Working Mathematician*. Graduate Texts in Mathematics. Springer-Verlag, Berlin, 1997. ⟨115, 118, 197, 460⟩

[Mog89] E. Moggi. An abstract view of programming languages. Lecture Notes, Stanford University, June 1989. ⟨120⟩

[Mog91] E. Moggi. Notions of computation and monads. *Information and Computation*, 93:55–92, 1991. ⟨120, 197⟩

[Mos06] Y. Moschovakis. *Notes on Set Theory*. Undergraduate Texts in Mathematics. Springer Verlag, 2nd edition, 2006. ⟨207⟩

[MPS86] D. MacQueen, G. Plotkin, and R. Sethi. An ideal model for recursive polymorphic types. *Information and Control*, 71:95–130, 1986. ⟨250⟩

[MS64] J. Mycielski and S. Swierczkowski. On the Lebesgue measurability and the axiom of determinateness. *Fund. Math.*, 54:67–71, 1964. ⟨75⟩

[OGS09] B. O'Sullivan, J. Goerzen, and D. Stewart. *Real World Haskell*. O'Reilly, Sebastopol, CA, 2009. ⟨1, 197⟩

[Oxt80] J. C. Oxtoby. *Measure and Category*. Number 2 in Graduate Texts in Mathematics. Springer Verlag, New York, Berlin, Heidelberg, 2nd edition, 1980. ⟨306⟩

[Pan09] P. Panangaden. *Labelled Markov Processes*. World Scientific Pub Co, 2009. ⟨434⟩

[Par67] K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, New York, 1967. ⟨492⟩

[Par85] R. Parikh. The logic of games and its applications. In M. Karpinski and J. van Leeuwen, editors, *Topics in the Theory of Computation*, volume 24, pages 111–140. Elsevier, 1985. ⟨197, 321⟩

[Pat04] D. Pattinson. Expressive logics for coalgebras via terminal sequence induction. *Notre Dame J. Formal Logic*, 45(1):19–33, 2004. ⟨193, 197⟩

[Pau00] M. Pauly. Game logic for game theorists. Technical Report INS-R0017, CWI, Amsterdam, 2000. ⟨321⟩

[PP03] M. Pauly and R. Parikh. Game logic — an overview. *Studia Logica*, 75:165–182, 2003. ⟨175, 197⟩

[Pum99] D. Pumplün. *Elemente der Kategorientheorie*. Spektrum Akademischer Verlag, Heidelberg, 1999. ⟨112, 118, 197⟩

[Pum03] D. Pumplün. Positively convex modules and ordered normed linear spaces. *J. Convex Analysis*, 10(1):109–127, 2003. ⟨141⟩

[Que01] B. v. Querenburg. *Mengentheoretische Topologie*. Springer -Lehrbuch. Springer-Verlag, Berlin, 3rd edition, 2001. ⟨306⟩

[Rou15] Ch. Rousseau. How Google works. `www.kleinproject.org`, August 2010 (last visited January 10, 2015). ⟨254⟩

[RR81] K. P. S. Bhaskara Rao and B. V. Rao. Borel spaces. *Dissertationes Mathematicae*, 190:1–63, 1981. ⟨375⟩

[RS50] H. Rasiowa and R. Sikorski. A proof of the completeness theorem of Gödel. *Fund. Math.*, 37:193–200, 1950. ⟨240, 266, 306⟩

[RS63] H. Rasiowa and R. Sikorski. *The Mathematics of Metamathematics*, volume 41 of *Monografie Matematyczne*. PWN, Warsaw, 1963. ⟨306⟩

[Rud74] W. Rudin. *Real and Complex Analysis*. Tata McGraw-Hill, 2nd edition, 1974. ⟨399, 465, 472, 477, 485, 492⟩

[Rut00] J. J. M. M. Rutten. Universal coalgebra: a theory of systems. *Theor. Comp. Sci.*, 249(1):3–80, 2000. Special issue on modern algebra and its applications. ⟨197⟩

[Sch70] H. H. Schaefer. *Topological Vector Spaces*. Number 3 in Graduate Texts in Mathematics. Springer Verlag, New York, Heidelberg, Berlin, 1970. ⟨492⟩

[SDDS86] J. T. Schwartz, R. B. K. Dewar, E. Dubinsky, and E. Schonberg. *Programming with Sets — An Introduction to SETL*. Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1986. ⟨1⟩

[SH10] A. K. Seda and P. Hitzler. Generalized distance functions in the theory of computations. *The Computer Journal*, 53(4):443–464, 2010. ⟨306⟩

[Sho67] J. R. Shoenfield. *Mathematical Logic*. Addison-Wesley, Reading, MA, 1967. ⟨266⟩

[Smy92] M. B. Smyth. Topology. In S. Abramsky, D. M. Gabbay, and T. S. E. Maibaum, editors, *Handbook of Logic in Computer Science*, volume 1 — Background: Mathematical Structures, pages 641–761. Oxford University Press, Oxford, 1992. ⟨207⟩

[Sok05] A. Sokolova. *Coalgebraic Analysis of Probabilistic Systems*. PhD thesis, Department of Computer Science, University of Eindhoven, 2005. ⟨197⟩

[Sri98] S. M. Srivastava. *A Course on Borel Sets*. Graduate Texts in Mathematics. Springer-Verlag, Berlin, 1998. ⟨366, 368, 492⟩

[Sri08] S. M. Srivastava. *A Course on Mathematical Logic*. Universitext. Springer Verlag, 2008. ⟨266, 268, 306⟩

[ST11] P. Sánchez Terraf. Unprovability of the logical characterization of bisimulation. *Information and Computation*, 209(7):1048–1056, 2011. ⟨459⟩

[Sta97] R. Stanley. *Enumerative Combinatorics*, volume 1 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, UK, 1997. ⟨75⟩

[Str81] K. R. Stromberg. *An Introduction to Classical Real Analysis*. Wadsworth International Group, Belmont, 1981. ⟨399⟩

[Tay99] P. Taylor. *Practical Foundations of Mathematics*, volume 59 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. ⟨75⟩

[Thi65] C. Thiel. *Sinn und Bedeutung in der Logik Gottlob Freges*. Verlag Anton Hain, Meisenheim am Glan, 1965. ⟨4⟩

[vdHP07] W. van der Hoek and M. Pauly. Modal logic for games and information. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, volume 3 of *Studies in Logic and Practical Reasoning*, pages 1077–1148. Elsevier, Amsterdam, 2007. ⟨320⟩

[Ven07] Y. Venema. Algebras and co-algebras. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, volume 3 of *Studies in Logic and Practical Reasoning*, pages 331–426. Elsevier, Amsterdam, 2007. ⟨71, 146⟩

[Vic89] S. Vickers. *Topology via Logic*. Number 5 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge, UK, 1989. ⟨207, 306⟩

Index

E.-E. Doberkat, *Special Topics in Mathematics for Computer Scientists*, DOI 10.1007/978-3-319-22750-4

U